Facultad de Ciencias Sociales y Jurídicas

Grado en Estadística y Empresa

Universidad de Jaén

Facultad de Ciencias Sociales y Jurídicas

Trabajo Fin de Grado

# ANÁLISIS DEL NÚMERO DE BARES, RESTAURANTES Y CAFETERÍAS EN LOS MUNICIPIOS DE LA PROVINCIA DE JAÉN MEDIANTE UN MODELO DE REGRESIÓN DE POISSON

**Alumno: Antonio Gómez Gallego**

**Mayo, 2019**

# Índice



# 1. Resumen

En este trabajo de fin de grado se introducen los modelos lineales generalizados como alternativa a los modelos lineales clásicos. Más concretamente se estudia el modelo de regresión de Poisson y se aplica a este modelo tratando explicar el comportamiento que tienen el número de bares, restaurantes y cafeterías en los municipios de la provincia de Jaén en función de otras covariables. Se tendrá en cuenta la población como una variable que hay que corregir dentro del modelo (offset). El análisis de los residuos se apoya en mapas a nivel municipal de la provincia de Jaén. Finalmente, se ha implementado una aplicación web a través de la librería Shiny del software R con la que el usuario puede interactuar con el modelo especificando la variable explicativa a demanda, proporcionando resultados tanto numéricos como gráficos.

## 1.1. Abstract

In this project, generalized linear models are introduced as an alternative to classical linear models. More specifically, the Poisson regression model was studied and applied to this model trying to explain the behavior of the number of bars, restaurants and coffee shops in the municipalities of the province of Jaén based on other covariates. The population will be taken into account as a variable that must be corrected within the model (offset). The analysis of the residuals is based on maps at the municipal level of the province of Jaén. Finally, a web application has been implemented through the Shiny library of the R software with which the user can interact with the model by specifying the explanatory variable on demand, providing both numerical and graphic results.

# 2. Introducción y objetivos

## 2.1. Motivación

Bares, restaurantes y cafeterías son ese tipo de establecimientos que no dejan indiferente a nadie. Hay quien puede llegar a odiarlos y a quererlos por igual, dependiendo de la situación personal y vital. Pueden llegar a ser un elemento molesto en cuanto al descanso del vecindario, pero pueden también ser parte del motor de la economía para muchas familias y una gran opción de ocio, barata y de calidad en muchos casos, especialmente en la provincia de Jaén.

La percepción, personal y subjetiva, del autor de este trabajo es que en su localidad, Linares, estos establecimientos tradicionalmente han funcionado muy bien. Son lugar de encuentro de aquellos amigos que vuelven a

casa por Navidad, sitio de paso para aquellos que en Semana Santa desean descansar durante un rato, lugar al que ir después de la jornada laboral y compartir un momento más distendido con tus compañeros. En general, no cabe duda que pueden ser un gran indicador de la economía o del nivel de riqueza de la zona; incluso pueden llegar a ayudar a entender el comportamiento de los habitantes de una localidad.

Además, a la oferta más tradicional en los últimos años se le ha añadido la de los gastrobares o cafés de autor, que junto a la recuperación de la economía española después de la crisis han impulsado a dichas empresas a un éxito que diariamente se ve latente en la calle.

### 2.2. Objetivos

El objetivo principal de este trabajo es analizar, en el contexto de los municipios de la provincia de Jaén, cómo se comportan estos establecimientos en relación con otras covariables que describen aspectos sociales o económicos.

Como segundo objetivo, nos planteamos un análisis de los residuos en el modelo de regresión de Poisson (que analizaremos más adelante) que arroje luz sobre las causas de la pérdida de la equidispersión en aquellos municipios que más destaquen por un exceso de variabilidad con respecto al modelo de Poisson, interpretando los resultados.

### 2.3. Estructura de la memoria

Una vez que se ha introducido la motivación de este trabajo y los objetivos del mismo, se va a detallar de forma resumida lo que se trata en los siguientes capítulos.

En el apartado de Metodología se explicará, en primer lugar, el origen de la muestra, es decir, dónde y cómo se han obtenido los datos utilizados en el modelo. Así mismo, se describen las covariables junto con la variable de interés utilizadas en el modelo de regresión de Poisson. Finalmente, se detalla cómo se ha realizado el análisis estadístico en torno a dicho modelo.

En el apartado del Estado del arte se estudian los modelos teóricos utilizados, así como su interpretación y los supuestos e hipótesis que deben cumplir, es decir, se detallan todas las herramientas utilizadas para el análisis del modelo explicadas en la Metodología.

El modelo se implementa mediante funciones del software estadístico R que se explican en el apartado de Implementación en R y, además, se introduce la librería Shiny que permite al usuario interactuar con el modelo implementado.

En el apartado de Resultados se aplica el modelo de regresión de Poisson a los datos y se desarrolla la aplicación web comentada anteriormente.

Para finalizar, el apartado de Conclusiones y vías futuras recoge los hallazgos más relevantes y algunas ideas que podrían desarrollarse en trabajos posteriores.

Como Anexo se presentan todos los scripts creados para la elaboración de este trabajo, así como la Bibliografía utilizada en el mismo.

## 3. Metodología

### 3.1. Origen de los datos

Los datos se han obtenido del Sistema de información Multiterritorial de Andalucía (SIMA) disponible dentro de la página web Instituto de Estadística y Cartografía de Andalucía (IECA). En un primer momento, se encontró la variable dependiente “Restaurantes y cafeterías” dentro de la sección 4.5.5., donde se diferenciaban restaurantes y cafeterías por categoria y número de plazas de cada tipo de establecimiento. El principal

inconveniente era el hecho de que sólo había datos históricos hasta el periodo 2009, por lo que no se podría llegar a explicar el comportamiento de bares, restaurantes y cafeterías en un momento más reciente.

Ante tal escollo, se continuó la investigación y se consultaron los códigos de la Clasificación Nacional de Actividades Económicas (CNAE), en los que se asigna un código a cada actividad económica de las que se pueden realizar. Generalmente este código (que suele ser de 5 dígitos) se utiliza en muchos formularios e impresos, tanto oficiales como a nivel de empresa. Por otra parte, se consultó la página web del diario Expansión, en su apartado de empresa para confirmar que bares y cafeterías se clasifican bajo el código 5630 (Establecimientos de bebidas), mientras que los restaurantes lo hacen bajo el código 5610 (Restaurantes y puestos de comidas). Ello permitió volver al SIMA, realizando una consulta relativa a estos dos códigos en la sección 4.11. referente a actividades económicas. Como resultado se ha obtenido el número de bares, restaurantes y cafeterías a 1 de enero de 2017, añadiendo a esta variable otras de tipo socioecnómico general como covariables que se describen posteriormente.

## 3.2. Análisis estadístico

Se ha utilizado un modelo de regresión de Poisson considerando el número de bares, restaurantes y cafeterías como variable dependiente o de respuesta y las siguientes posibles variables explicativas:

1. Población total según padrón municipal
2. Número de plazas hoteleras por habitante
3. Existencia de centro histórico en el municipio
4. Número de oficinas de entidades de crédito por habitante. (Entidades de depósito)
5. Parque de turismos con mas de 2.000 cc por habitante
6. Renta neta declarada por habitante
7. Número de pensionistas
8. Pensión media
9. Número de hoteles por habitante

Para valorar la significación de las covariables se han realizado contrastes t-Student. Igualmente, se ha utilizado un contraste para la hipótesis nula de equidispersión frente a la alternativa de sobredispersión (Cameron and Trivedi, 1990), junto con un test $\chi^2$ sobre la deviance para valorar la bondad del ajuste. También se ha optado por el coeficiente de determinación $R^2$ como medida de bondad de ajuste.

Para la diagnosis del modelo se han utilizado cuatro tipos de gráficos: gráfico de residuos frente a valores del predictor lineal, gráfico de probabilidad normal de los residuos estudentizados, gráfico de la raíz cuadrada de la desviación de los residuos frente al valores del predictor lineal y gráfico de los residuos de Pearson frente al leverage y distancia de Cook.

En relación a los residuos, se ha creado una representaciones de los municipios de la provincia en un mapa donde dos colores diferencian si el municipio tiene un residuo positivo o negativo.

Finalmente, hemos utilizado el test de Razón de Verosimilitud para valorar la significación de una covariable en un modelo ampliado con respecto a aquel que no la incluye.

## 3.3. Implementación en R

### 3.3.1. Ajuste, descripción y diagnosis del modelo

Se ha utilizado el software estadístico R (R Core Team, 2018) para el ajuste del modelo. Concretamente, la función glm del paquete stats (R Core Team, 2018) y su métodos summary. También hemos utilizado el

paquete ggfortify (Tang et al., 2016), concretamente su método autoplot, para visualizar el modelo y sus gráficos de diagnosis. Este a su vez requiere del paquete ggplot2 (Hadley Wickham, 2016). Las funciones rsq del paquete rsq (Zhang, 2018) y dispersiontest del paquete AER (Achim Zeileis, 2008) se utilizan para calcular el coeficiente de determinación y realizar el test de sobredispersión, respectivamente.

Además, se han utilizado los paquetes readxl (Wickham et al., 2019) y tidyverse (Wickham, 2017) para exportar los datos de archivos con formato Excel y para su posterior limpieza, el paquete ggplot2 comentado anteriormente, ggrepel (Slowikowsk, 2018) y RColorBrewer (Neuwirth, 2014) para realizar representaciones gráficas y el paquete stargazer (Hlavac, 2018) para darle formato de tabla a los resultados obtenidos de la función glm. También se ha utilizado el paquete sf (Pebesma, 2018) para realizar mapas visuales de la provincia de Jaén.

### 3.3.2. Aplicación Shiny

Shiny (Chang et al., 2018) es una librería de R que permite la creación de aplicaciones web interactivas con un tipo especial de sintaxis. Se basa en la relación automática *reactiva* entre entradas y salidas y amplios widgets precompilados que hacen posible crear aplicaciones fáciles y potentes con el mínimo esfuerzo.

En este trabajo se proporciona una implementación del código a través de Shiny que permite a un usuario, a través de la web, obtener los resultados que se describen (reproducibilidad) u otros similares basados en el catálogo amplio de posibles covariables.

# 4. Estado del arte

Según Manuel Ato García (1996, pág. 79), "*el modelado estadístico se propone la búsqueda del modelo más simple que sea capaz de explicar los datos con el mínimo error posible*". En base a esto, Judd et al. (2009, pag. 3) proponían la ecuación

DATOS = MODELO + ERROR,

donde los datos son la respuesta de la variable dependiente y el modelo está formado por un conjunto de variables que intenta explicar los datos. A todo modelo se le añade una parte de error, que es la diferencia entre el modelo teórico propuesto y los datos empíricos obtenidos.

Atendiendo a las fases del modelado y el proceso que se lleva a cabo para obtenerlo, debemos diferenciar entre las siguientes:

- Selección del modelo. En esta etapa se realiza un estudio descriptivo de aquellas variables a incluir en el modelo para comprobar la posibilidad de que hubiera datos faltantes o su escala de medida. Además se debe tener en cuenta el tipo de relación que hay entre las variables (lineal cuadrática, exponencial, etc) para especificar correctamente el componente sistemático que estudiaremos más adelante.
- Ajuste del modelo. En esta fase se procede a la estimación de los parámetros, es decir, aquellos coeficientes que acompañan a los regresores y que nos permitirán realizar las estimaciones de la variable respuesta. Lo más adecuado en este caso es seguir el principio de parsimonia, es decir, a igualdad de modelos nos quedaremos con aquel que tenga menos parámetros a estimar, ya que se ganará en grados de libertad. También es adecuado realizar una diagnosis del modelo a través de sus residuos, es decir, analizar la discrepancia que hay entre valores esperados por el modelo y valores empíricos u observados.
- Evaluación del modelo. En este caso se debe comprobar que se cumplen los supuestos del modelo: normalidad, homogeneidad, linealidad, independencia, etc. También se debe comprobar la capacidad predictiva del modelo mediante los test de bondad de ajuste que permitan cuantificar qué proporción de la variabilidad de los datos es explicada por el modelo o si el modelo es adecuado o no.
- Interpretación del modelo. Se interpretan cómo afectan los regresores del modelo a la variable respuesta de forma aislada al resto de regresores y se realizan predicciones.

## 4.1. Modelos lineales generalizados

Cuando en los modelos lineales no se cumplen alguna de las hipótesis (linealidad, homocedasticidad, independencia o normalidad), normalmente se realizan transformaciones en la variable respuesta que complican la interpretación de los resultados obtenidos.

En el caso de datos de recuento, como los que motivan este trabajo, lo más usual es que no se cumplean estas hipótesis por la propia naturaleza de los datos. Es por ello que se opta por una solución más simple. Esta solución consiste en utilizar los Modelos Lineales Generalizados, que basan su existencia en aplicar distribuciones de la familia exponencial para modelar la variable respuesta ajustándola según esta familia de distribuciones.

Siguiendo a Ato et al. (2005), el Modelo Lineal Generalizado tiene los siguientes componentes:

1. El vector de la respuesta media: $g(\mu_i) = \eta_i$
2. El vector del predictor lineal: $\beta_j X_{ij}$
3. La componente aleatoria (error): $\varepsilon_i$

## 4.2. Regresión de Poisson

### 4.2.1. Modelo teórico. La distribución de Poisson

La distribución de Poisson se utiliza como modelo para el número de ocurrencias de un evento en un intervalo de tiempo concreto. Si $Y \sim P(\mu)$, la función de probabilidad viene dada por

$$P(Y = y) = \frac{e^{-\mu}\mu^y}{y!}$$

con $y = 0, 1, 2, \ldots$ y parámetro $\mu > 0$, donde la media y la varianza vienen dadas por $E(Y) = \mu = Var(Y)$.

Dado que la distribución de Poisson pertenece a la familia exponencial, el modelo de regresión basado en ella pertenece a la familia de Modelos Lineales Generalizados.

### 4.2.2. El modelo de regresión de Poisson

En el modelo de regresión de Poisson las covariables se introducen en la ecuación de la media de forma log-lineal, es decir,

$$E[y_i] = \mu_i = e^{x_i' \times \beta}$$

para $i = 1, 2, , n$. De esta forma,

$$\log(\mu_i) = \eta_i = x_i'\beta = \beta_0 + \beta_1 x_{1i} + \ldots + \beta_k x_{ki}$$

para $i = 1, 2, \ldots, n$.

En cuanto a las partes del modelo, tenemos:

- Componente sistemático: $\eta_i = x_i'\beta$
- Componente aleatorio: $y_i \sim P(\mu_i)$
- Función de enlace: $g(\mu_i) = \eta_i = log(\mu_i)$

### 4.2.3. Estimación de los parámetros del modelo

A diferencia de los modelos lineales, que estiman los parámetros mediante estimadores de mínimos cuadrados, los estimadores del modelo de regresión de Poisson se estiman habitualmente mediante máxima verosimilitud. En el caso del modelo de regresión de Poisson, la función de log-verosimilitud a optimizar viene definida por

$$\log L(\beta; Y, X) = \log \prod_{i=1}^{n} f(y_i \mid x_i; \beta) = \sum_{i=1}^{n} \log[f(y_i \mid x_i; \beta)] = \sum_{i=1}^{n} (-\exp(x_i'\beta) + y_i x_i'\beta - log(y_i!))$$

## 4.3. Interpretación de los parámetros

En un modelo log-lineal las estimaciones de los coeficientes se interpretan del siguiente modo:

1. $e^{\hat{\beta}_0}$ es el valor esperado de la variable dependiente cuando el resto de variables explicativas son todas nulas o cero.
2. Cada valor de $\beta_j$ se interpreta en función de si este valor es mayor o menor que 0. Así, por ejemplo, si $\beta_j > 0$ se produce un incremento porcentual de $(e^{\hat{\beta}_j} - 1) \times 100$ de la variable respuesta cuando se produce un incremento unitario de la covariable o variable explicativa correspondiente, permaneciendo constantes el resto de predictores. De la misma forma si $\beta_j < 0$ se produce un decremento porcentual de la variable explicativa cuando se produce un incremento de la variable explicativa y el resto de predictores son constantes.

## 4.4. Equidispersión

En un modelo de Poisson, la media es $E(Y) = \mu$ y la varianza es $Var(Y) = \mu$. En el paquete AER de R se encuentra la función dispersiontest() implementada según el Test de Sobredispersión por Cameron y Trivedi (1990). En este test, la hipótesis nula asume que $Var(Y) = \mu + cf(\mu)$ donde la constante $c > 0$ significa infradispersión, mientras que $c > 0$ significa sobredispersión. $f(\mu)$ es una función monótona (lineal o cuadrática). Las hipótesis del contraste son

$$H_0 : c = 0,$$

que implica equidispersión, frente a

$$H_1 : c \neq 0,$$

y se basa en un estadístico $t - Student$.

Cuando trabajamos con recuentos reales la equidispersión no suele ser cierta. Una forma de cuantificar el habitual exceso de variabilidad es el estadístico de Pearson dividido por sus grados de libertad o la función de desviación, también dividida por sus grados de libertad.

Si se rechaza la hipótesis de igualdad de media y varianza y $c > 0$, podemos decir que existe sobredispersión (la varianza es mayor que la media). Esta sobredispersión suele estar asociada a la heterogeneidad que pudiera haber entre las observaciones, lo que se puede interpretar como una mezcla de distribuciones de Poisson. Esto no es un problema cuando la variable dependiente sigue una distribución normal, ya que la normal tiene un parámetro específico que modeliza la variabilidad. Otras causas por las que puede existir la sobredispersión son que los datos no sean de Poisson o una mala elección de la función de enlace.

## 4.5. Bondad de ajuste

Las medidas o estadísticos de bondad de ajuste son útiles para comprobar o verificar que un modelo se ajusta bien a los datos y, por tanto, afirmar que este modelo es válido.

Las hipótesis de contraste para la bondad del ajuste del modelo son que el modelo ajusta bien los datos frente a que no lo hace.

Se han utilizado dos criterios que comentamos a continuación.

### 4.5.1. Estadístico basado en la Deviance

En un ajuste de regresión de Poisson, la Deviance es igual a:

$$D = 2\sum_{i=1}^{n}[Y_i \log(Y_i \mid \mu_i) - (Y_i - \mu_i)]$$

donde

$$\mu_i = \exp(\hat{\beta}_0 + \hat{\beta}_1 X_1 + \ldots + \hat{\beta}_k X_k)$$

Esta desviación residual (Deviance) se entiende como la diferencia entre la desviación del modelo que tenemos y la desviación del modelo ideal en el que los valores predichos son iguales a los observados, es decir, la desviación que no es capaz de explicar el modelo ajustado. Si realmente esta diferencia es pequeña y, por tanto, nuestro modelo es adecuado, entonces la desviación residual sigue aproximadamente una distribución $\chi^2$ con los grados de libertad correspondientes a la desviación residual.

Según el criterio de Bartlett (2014), se debe tener cuidado con el uso de esta medida ya que lo considera adecuado para tamaños de muestra muy grandes.

### 4.5.2. Coeficiente de determinación $R^2$

El coeficiente de determinación sirve para valorar los efectos del modelo. $R^2$ nos muestra el porcentaje de variabilidad que explica el modelo de los datos.

## 4.6. Diagnosis del modelo

En cuanto a los gráficos de diagnosis tenemos cuatro, resultantes de la función autoplot():

- Residuos frente a valores del predictor lineal. Este gráfico se utiliza para comprobar la linealidad de los predictores, es decir, permite visualizar si el predictor lineal es adecuado, si falta algún predictor en el modelo o si hay que hacer algún tipo de transformación a los predictores. También permite visualizar si se cumple la hipótesis de independencia de los residuos propia de los modelos lineales. Las observaciones deben de posicionarse entorno a la linea horizontal.
- Gráfico de probabilidad normal de residuos estudentizados para comprobar si los residuos se distribuyen según una Normal.
- Raiz cuadrada de la desviación de los residuos frente al valores del predictor lineal. Útil para comprobar que se cumple la hipótesis de homocedasticidad. La forma de "altavoz" en los residuos podría ser un indicativo de heterocedasticidad.
- Residuos de Pearson frente al leverage y distancia de Cook para visualizar si existen observaciones influyentes.

# 5. Resultados

Como se ha mencionado en los apartados de introducción y metodología, lo que se pretende es explicar el número de restaurantes y cafeterías de los municipios de la provincia de Jaén en función de una serie de covariables.

## 5.1. Modelo nulo

El modelo de partida (modelo nulo) tan sólo cuenta con la población del municipio en la ecuación de la media, considerada como un offset, es decir, con coeficiente de regresión igual a uno. Con ello se está forzando a que el modelo considere que el número de habitantes tiene un efecto multiplicativo sobre el promedio de la variable dependiente. Es decir, y como ejemplo, a doble de población, doble de bares, restaurantes y cafeterías.

El valor del coeficiente $R^2$ es de 0.9769. El test de dispersión ofrece un p-valor de 0.0144. Por su parte, el p-valor del test $\chi^2$ de bondad del ajuste es de 0.

Por tanto, en cuanto al supuesto de equidispersión, debe rechazarse, en favor de la alternativa de que existe sobredispersión. Esta sobredispersión se debe, muy probablemente, a la heterogeneidad entre las observaciones.

En cuanto a la bondad del ajuste, el resultado muestra que el ajuste obtenido no puede considerarse como adecuado ya que su p-valor es muy pequeño.

Sin embargo, el coeficiente de determinación $R^2$ para valorar los efectos del modelo del 97.69 % indica que un elevado porcentaje de la variabilidad de la variable dependiente (número de bares, restaurantes y cafeterías [1]) es explicada sólo por el offset (la población). El gráfico $Y \sim X$ de la Figura 1 muestra esta elevada capacidad predictiva del modelo.

En cuanto a la diagnosis (Figura 2), observamos:

- El gráfico de residuos frente a valores del predictor lineal (gráfico superior izquierdo) muestra un patrón que indica inadecuación del modelo.
- El gráfico de probabilidad normal (gráfico superior derecha) de los residuos estudentizados muestra un ajuste a la normal bastante bueno, salvo en algunas observaciones.
- El gráfico de la raíz cuadrada de la desviación de los residuos frente al valores del predictor lineal (gráfico inferior izquierda), útil para comprobar que se cumple la hipótesis de homocedasticidad, no muestra ningún patrón.
- Los residuos de Pearson frente al leverage y distancia de Cook (gráfico inferior derecha), para ver si existen observaciones influyentes, indica la presencia de dos de ellas, Cazorla (25) y Úbeda (84).

De forma global, la diagnosis muestra que Cazorla y Úbeda presentan discrepancias en el modelo en cuanto al número esperado y observado de bares. Más concretamente, al tener unos residuos positivos, podemos decir que estos municipios tienen más bares que los ajustados o esperados por el modelo. De la misma forma, Andújar tiene menos bares de lo esperado. Además, existe una observación influyente: Jaén (capital) (46).

Dentro de la provincia de Jaén, Úbeda y Cazorla son dos poblaciones eminentemente turísticas, la primera por su patrimonio histórico y la segunda por pertenecer al parque natural de Sierra Cazorla, Segura y Las Villas. Esto sugiere introducir una variable relacionada con el sector turístico para ayudar al modelo a explicar la variabilidad que en el modelo nulo permanece no explicada y que puede estar provocando el exceso de dispersión con respecto al modelo de Poisson. Por ello, hemos considerado el número de plazas hoteleras por habitante en estos municipios. De hecho, cabe pensar que el que existan más bares de lo esperado es debido a que estos bares pueden alimentarse de aquellos turistas que visitan la ciudad y se alojan en dichas plazas.

[1] De ahora en adelante, por brevedad, hablaremos simplemente de bares.

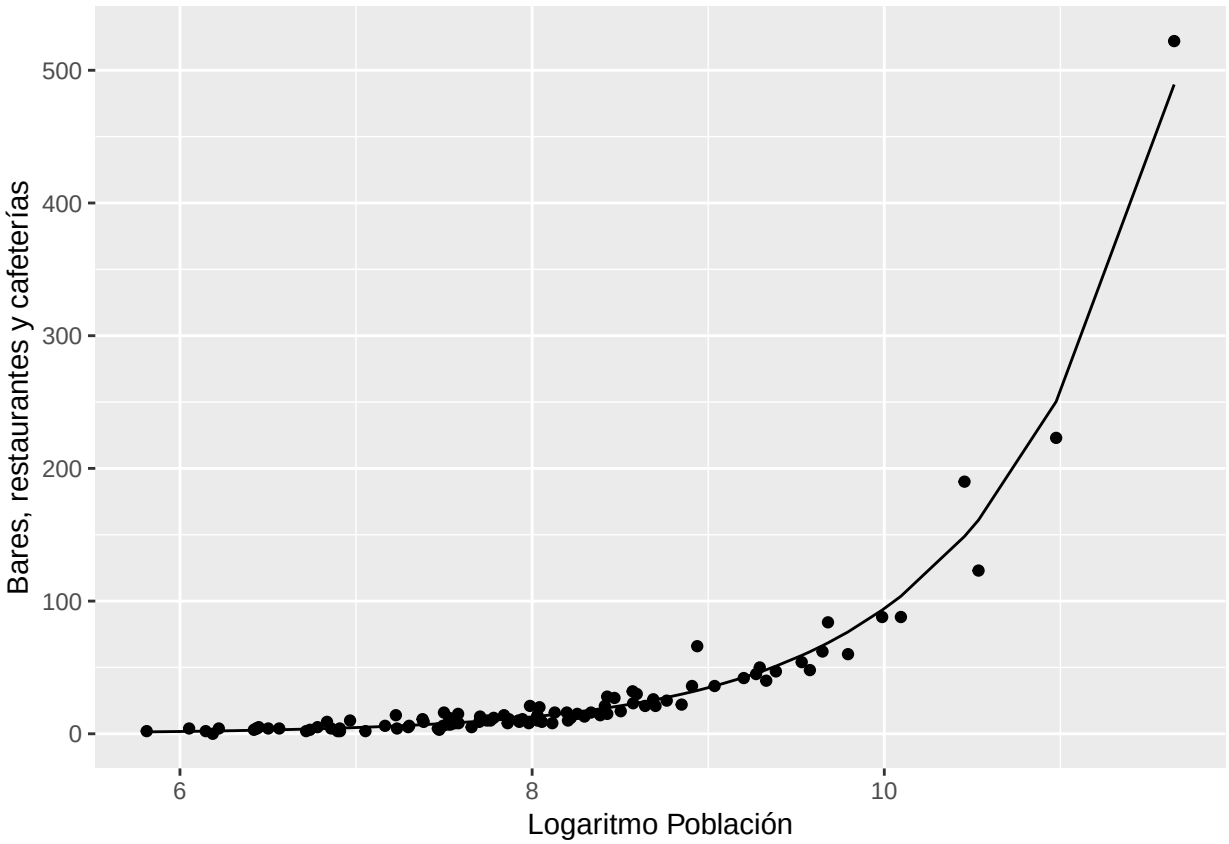


Figura 1: Capacidad predictiva Modelo nulo

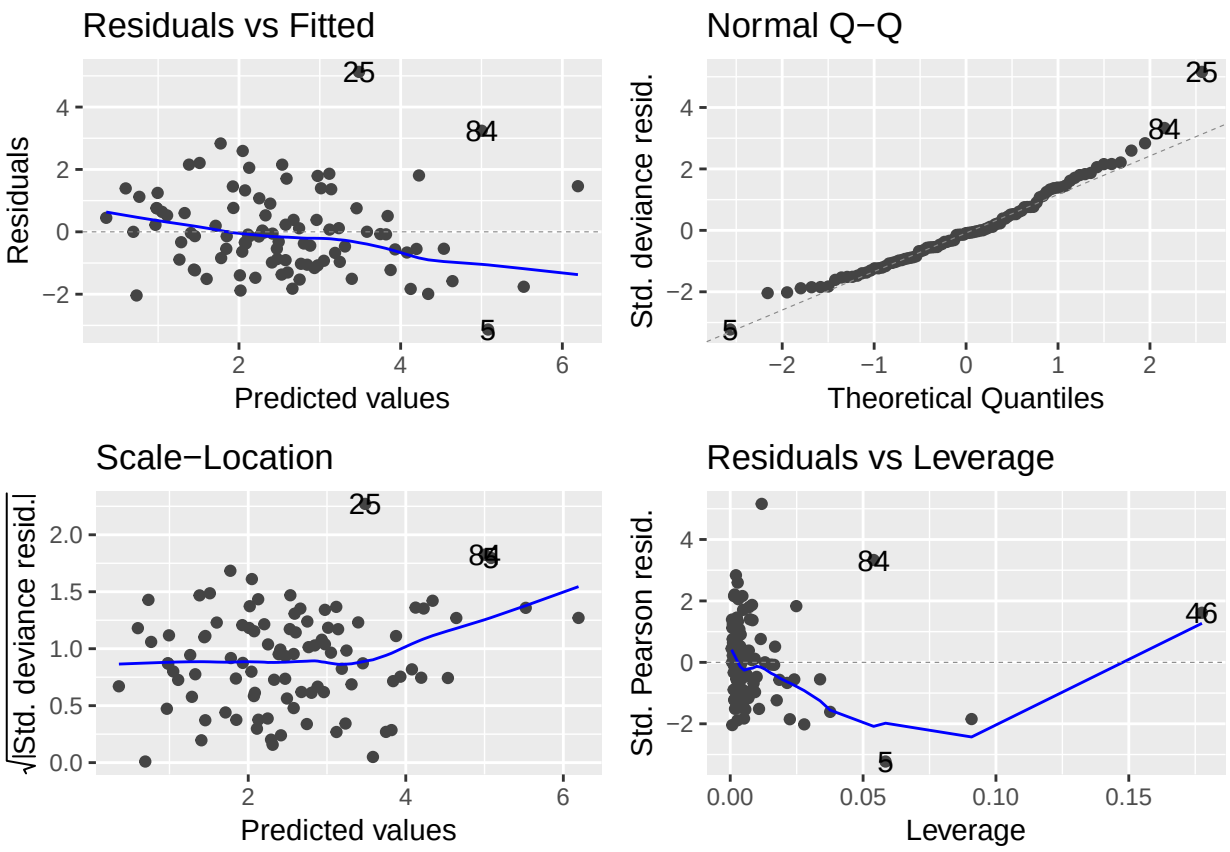


Figura 2: Diagnosis Modelo nulo

## 5.2. Modelo 1. Predictor basado en la ratio de plazas hoteleras por habitante

El nuevo modelo ya incluye, además del offset de la población, un primer predictor dado por la ratio de plazas hoteleras por habitante. El cuadro 1 muestra el resumen de dicho modelo.

Cuadro 1: Modelo 1. Predictor basado en la ratio de plazas hoteleras por habitante

| | BarResCaf |
|---|---|
| plazashotel.est | $2.2813^{***}$ |
| | (0.3749) |
| Constant | $-5.4944^{***}$ |
| | (0.0205) |
| Observations | 97 |
| Log Likelihood | −284.5392 |
| Akaike Inf. Crit. | 573.0784 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Se observa que el coeficiente del predictor añadido es significativo ya que el p-valor obtenido del contraste t-Student es menor que cualquier nivel de significación $\alpha$ fijado. También se debe apuntar que el signo positivo del coeficiente es coherente ya que se entiende que a medida que aumenta el número de plazas hoteleras, aumenta el número de bares, restaurantes y cafeterías.

El test de dispersión ofrece un p-valor de 0.0213. Por su parte, el p-valor del test $\chi^2$ de bondad del ajuste es de 0.001.

El supuesto de equidispersión debe rechazarse de nuevo, en favor de la hipótesis alternativa de que existe sobredispersión. El valor del p-valor se ha incrementado respecto al modelo nulo.

En cuanto a la bondad del ajuste, el resultado muestra que el ajuste obtenido sigue sin poder considerarse como adecuado ya que su p-valor sigue siendo pequeño pero ha aumentado considerablemente respecto al modelo anterior.

El coeficiente de determinación $R^2$ para valorar los efectos del modelo, del 97.85 %, apenas ha cambiado.

En cuanto a la diagnosis (Figura 3), observamos:

- El gráfico de residuos frente a valores del predictor lineal sigue mostrando un patrón que indica inadecuación del modelo.
- El gráfico de probabilidad normal de los residuos estudentizados también sigue mostrando un ajuste a la normal bastante bueno, salvo en algunas observaciones.
- El gráfico de la raíz cuadrada de la desviación de los residuos frente al valores del predictor lineal, tampoco muestra ningún patrón.
- Los residuos de Pearson frente al leverage y distancia de Cook, indica ahora la presencia de dos observaciones influyente Cazorla (25) que ya aparecía en el anterior modelo nulo y La Iruela(43) que aparece por primera vez.

Cuadro 2: Analysis of Deviance Table

| Resid. Df | Resid. Dev | Df | Deviance |
|---|---|---|---|
| 96 | 171.1 | | |
| 95 | 143.1 | 1 | 28.05 |

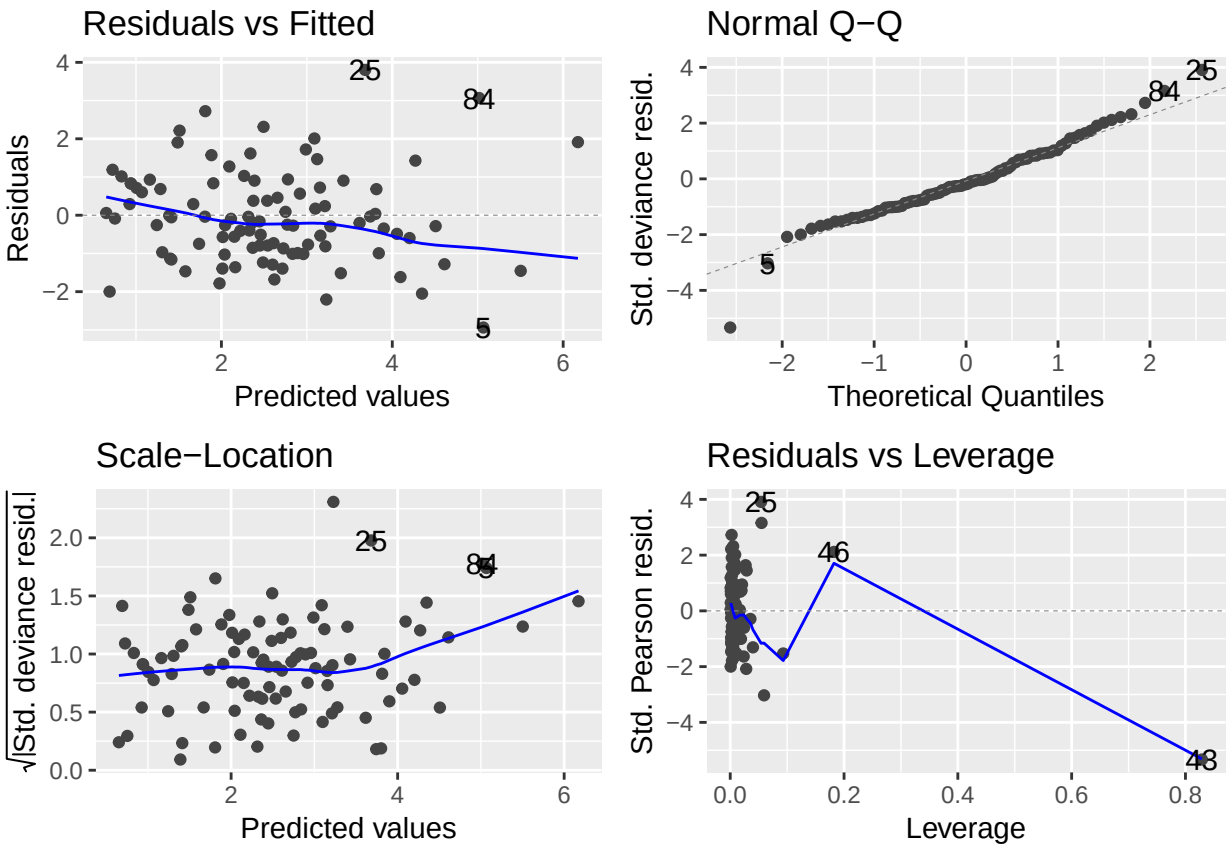


Figura 3: Diagnosis Modelo 1

| Chi.squared | p.value | df |
|---|---|---|
| 28.05 | 1.182e-07 | 1 |

Finalmente el Test de Razón de Verosimilitud (Cuadro 2), confirma la significación de la nueva variable del modelo 1.

### 5.3. Modelo 2. Predictor basado en la ratio de superficie (hectáreas) de espacios naturales protegidos por habitante

En vista de que el análisis de residuos ha mostrado de nuevo Cazorla y ahora también a la Iruela como valores influyentes y con residuos positivos y altos, hemos incluido en este nuevo modelo la variable superficie (hectáreas) de espacios naturales protegidos por habitante, que es una característica común de ambos municipios. Creemos que estos espacios naturales protegidos pueden ser un reclamo para el turimo en la zona, turismo que no necesariamente pernocte en hoteles pero que haga uso de bares, por lo que la nueva variable podría aportar información extra respecto de la que incluye el modelo 1. El cuadro 3 muestra el resumen de dicho modelo. Destacamos los siguientes aspectos:

Cuadro 4: Modelo 2. Predictor basado en la ratio de superficie (hectáreas) de espacios naturales protegidos por habitante

| | BarResCaf |
|---|---|
| plazashotel.est | $1.9998^{***}$ |
| | (0.4131) |
| Superficie.est | $0.0173^{**}$ |
| | (0.0084) |
| Constant | $-5.4991^{***}$ |
| | (0.0207) |
| Observations | 97 |
| Log Likelihood | −282.5915 |
| Akaike Inf. Crit. | 571.1831 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

1. El coeficiente del predictor añadido es significativo. El signo positivo del coeficiente también es coherente con nuestra hipótesis.
2. El valor del coeficiente $R^2$ es de 0.9772. El test de dispersión ofrece un p-valor de 0.031. Por su parte, el p-valor del test $\chi^2$ de bondad del ajuste es de 0.0017. El supuesto de equidispersión también debe rechazarse en este caso. El valor del p-valor se ha incrementado respecto al modelo 1 pero también en menor medida. El ajuste obtenido no se puede considerar como adecuado ya que su pvalor sigue siendo pequeño.
3. El test de Razón de Verosimilitud (Cuadro 4) muestra que el modelo es significativo con respecto al modelo 1.

En cuanto a la diagnosis (Figura 4), observamos:

- El gráfico de residuos frente a valores del predictor lineal sigue mostrando un patrón que indica inadecuación del modelo, pero ha mejorado.

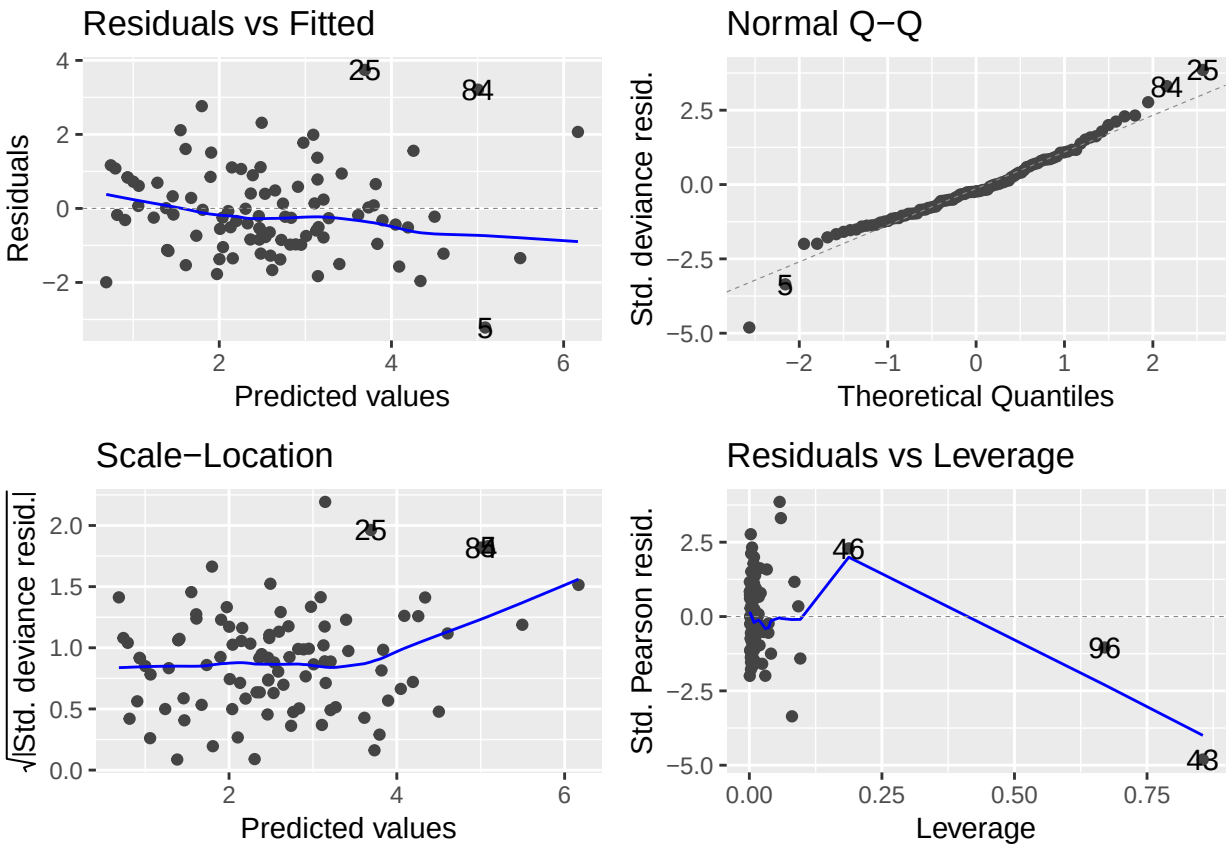


Figura 4: Diagnosis Modelo 2

- Los residuos de Pearson frente al leverage y distancia de Cook, muestra la presencia de dos observaciones influyentes, La Iruela(43) que apareció por primera vez en el modelo 1 y Santiago-Pontones(93) que aparece en el modelo 2.

Cuadro 5: Analysis of Deviance Table

| Resid. Df | Resid. Dev | Df | Deviance |
|---|---|---|---|
| 95 | 143.1 | | |
| 94 | 139.2 | 1 | 3.895 |

| Chi.squared | p.value | df |
|---|---|---|
| 3.895 | 0.04842 | 1 |

## 5.4. Modelo 3. Predictor basado en la existencia de centro histórico en el municipio

Cuadro 7: Modelo 3. Predictor basado en la existencia de centro histórico en el municipio

| | BarResCaf |
|---|---|
| plazashotel.est | $1.8801^{***}$ |
| | (0.4180) |
| Superficie.est | $0.0189^{**}$ |
| | (0.0085) |
| centroSí | $0.0912^{**}$ |
| | (0.0386) |
| Constant | $-5.5464^{***}$ |
| | (0.0291) |
| Observations | 97 |
| Log Likelihood | −279.7908 |
| Akaike Inf. Crit. | 567.5817 |
| *Note:* | $^{*}$p<0.1; $^{**}$p<0.05; $^{***}$p<0.01 |

Como tercera covariable hemos introducido una variable dicotómica que indica si el municipio cuenta con centro histórico, ya que puede ser otro factor relacionado con el turismo.

Destacamos los siguientes resultados del cuadro 5:

1. El coeficiente del predictor añadido es significativo. El signo positivo del coeficiente también es coherente.
2. El valor del coeficiente $R^2$ es de 0.9819. El test de dispersión ofrece un p-valor de 0.0513. Por su parte, el p-valor del test $\chi^2$ de bondad del ajuste es de 0.0037. Por primera vez no se rechaza la hipótesis de equidispersión, aunque el test $\chi^2$ de bondad de ajuste sigue rechazando el modelo.
3. De nuevo la variable cualitativa añadida hace que el modelo 3 sea significativo respecto al modelo 2.

En cuanto a la diagnosis (Figura 5), observamos:

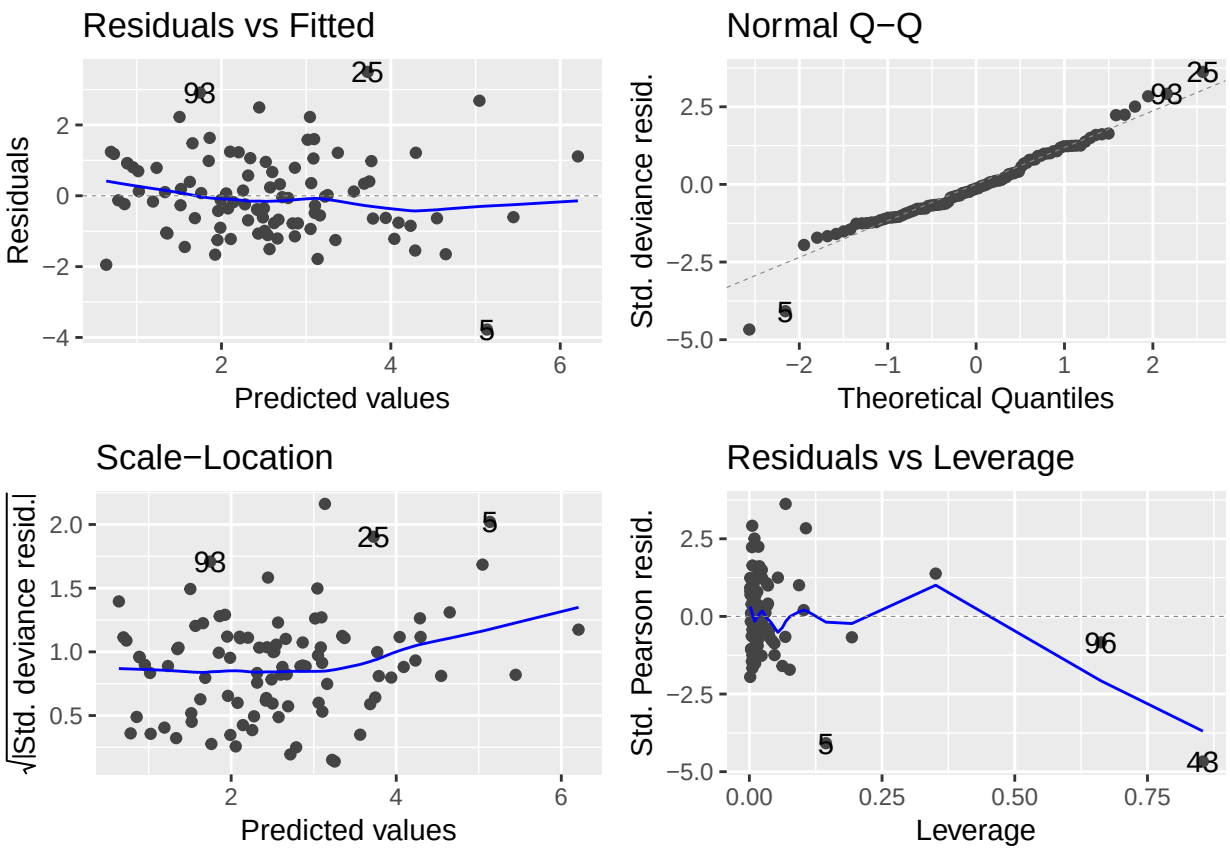


Figura 5: Diagnosis Modelo 3

- El gráfico de residuos frente a valores del predictor lineal muestra por primera vez adecuación del modelo en cuanto a linealidad.
- El gráfico de probabilidad normal de los residuos estudentizados mejora para observaciones con residuos positivos pero sigue sin ser del todo buena para residuos negativos, incluso podría decirse que empeora respecto al modelo 2.

Cuadro 8: Analysis of Deviance Table

| Resid. Df | Resid. Dev | Df | Deviance |
|---|---|---|---|
| 94 | 139.2 | | |
| 93 | 133.6 | 1 | 5.601 |

| Chi.squared | p.value | df |
|---|---|---|
| 5.601 | 0.01795 | 1 |

A continuación vamos a analizar con detalle este modelo en el que, por primera vez, se acepta la hipótesis de equidispersión.

A partir de los valores predichos por el modelo y los valores observados, obtenemos el siguiente gráfico (Figura 6), que consiste en un diagrama de puntos en el que el eje X cuenta con los valores esperados según el modelo y en el eje Y, los valores observados. La recta representa $y = x$, es decir, cuando el número de bares esperados es igual a los observados. De esta forma, se interpretan aquellos municipios que se sitúan por encima de la recta como aquellos que cuentan con más bares observados de lo esperado (residuos positivos), mientras que los que se sitúan por debajo son aquellos que tienen menos bares observados de lo esperado según el modelo (residuos negativos).

Entre aquellos municipios que tienen más de 100 bares destacan, aunque no en exceso, Jaén y Úbeda por tener más bares que los predichos por el modelo. Sin embargo, Linares y Andújar, destacan por todo lo contrario, aunque con valores que no pueden considerarse anómalos, más bien están en torno a la recta $y = x$. En cualquier caso, el autor del trabajo reconoce su sorpresa ante el hecho de que Linares muestre un residuo negativo en el modelo.

Si nos centramos en aquellos municipios que cuentan con más de 30 bares pero no llegan a los 100 (Figura 7), la mayoría de ellos cuentan con residuos de aproximadamente 0 ya que están en torno a la recta $y = x$, excepto Cazorla y Baeza, que tienen residuos positivos, y Bailén, con residuos negativos.

De igual forma tenemos a los municipios con menos de 30 bares (Figura 8).

Ahora vamos a tratar de arrojar luz sobre el análisis de los residuos mediante mapas. En primer lugar vamos a visualizar los municipios de la provincia de Jaén según el area territorial a la que pertenecen (Figura 9).

Podemos visualizar también la predicción que realiza el modelo 3 en cuanto al número de bares esperados en cada municipio (Figura 10).

Para finalizar, dada la dificultad de extraer conclusiones claras del mapa anterior, se proporciona a continuación el mapa de Jaén según el signo de los residuos (Figura 11). Es el más interesante, ya que, como se puede observar, la zona Este de la provincia está claramente definida por municipios que tienen más bares de lo esperado y viceversa.

Por tanto, este análisis indica que parece haber una correlación espacial, ya que los residuos parecen estar relacionados con la situación geográfica.

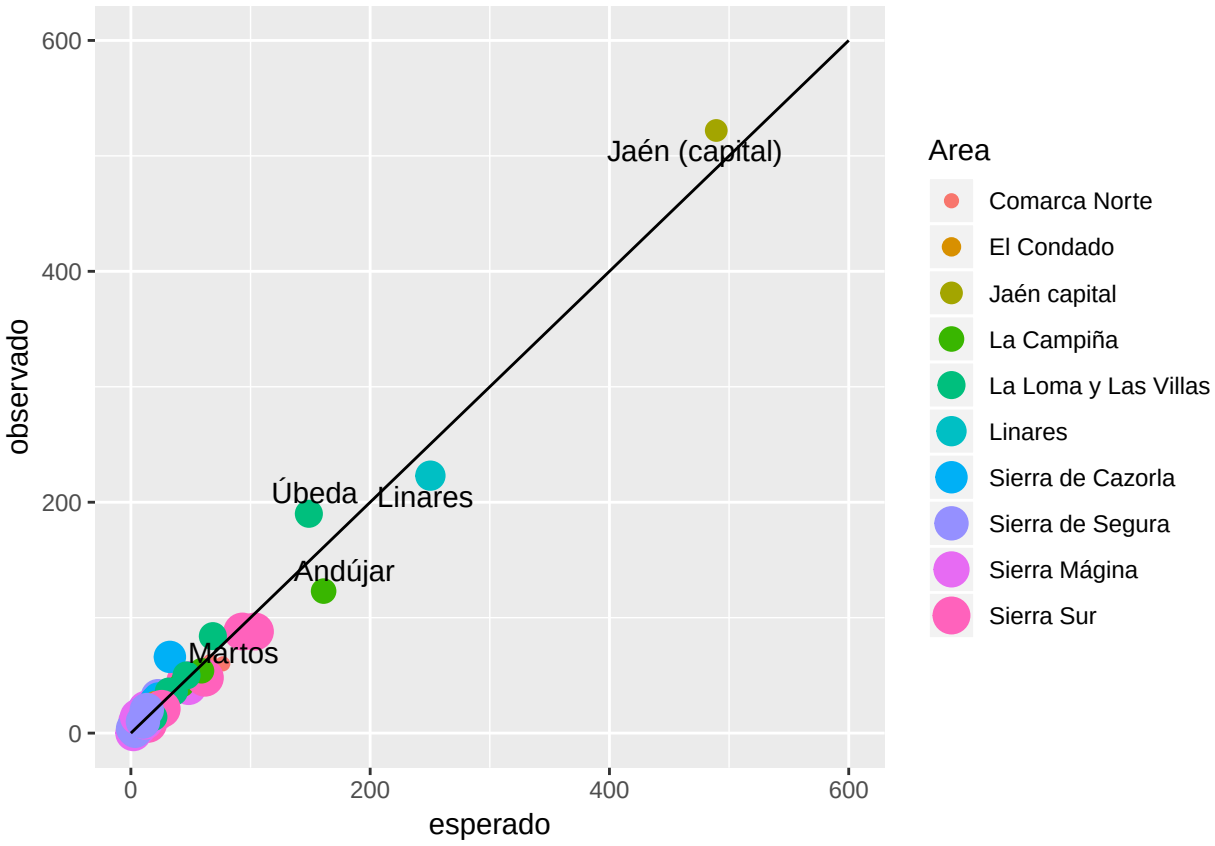


Figura 6: Municipios con más de 100 bares

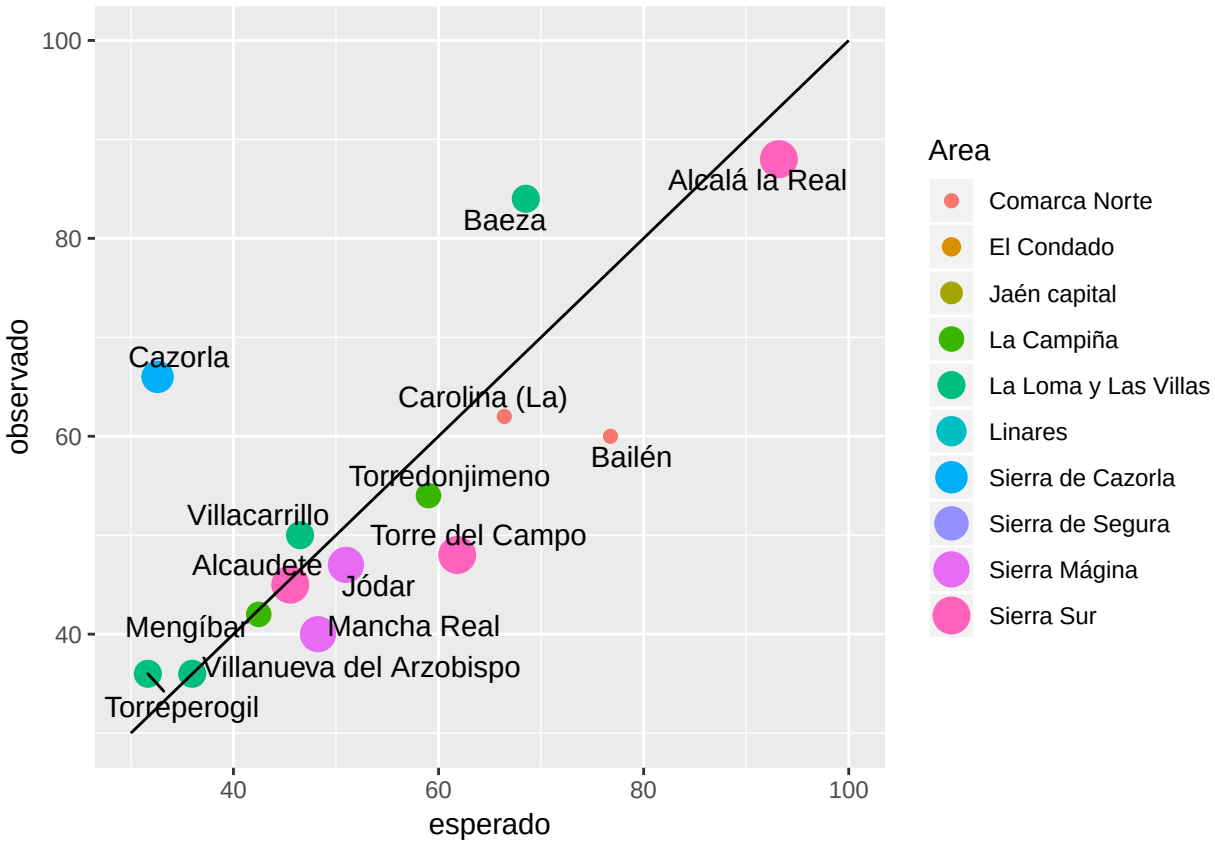


Figura 7: Municipios que tienen entre 30 y 100 bares

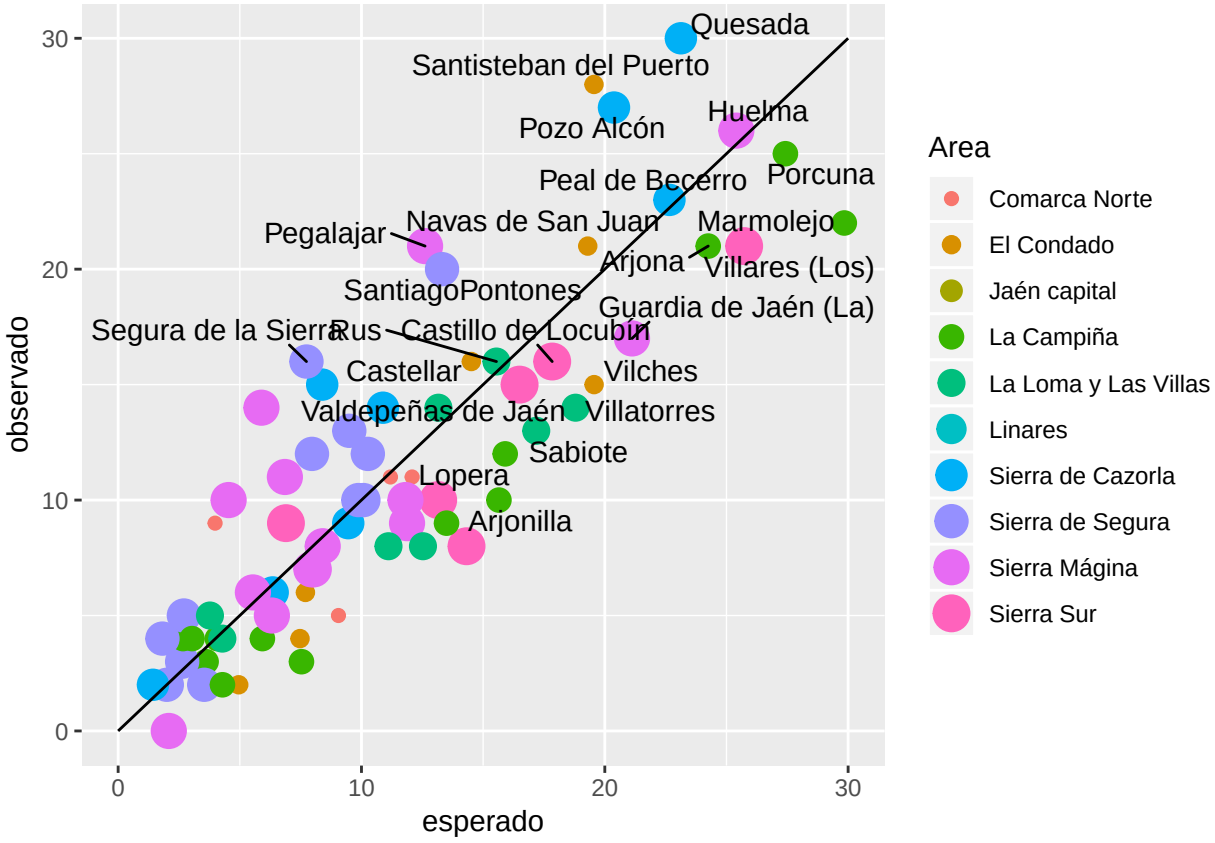


Figura 8: Municipios con menos de 30 bares

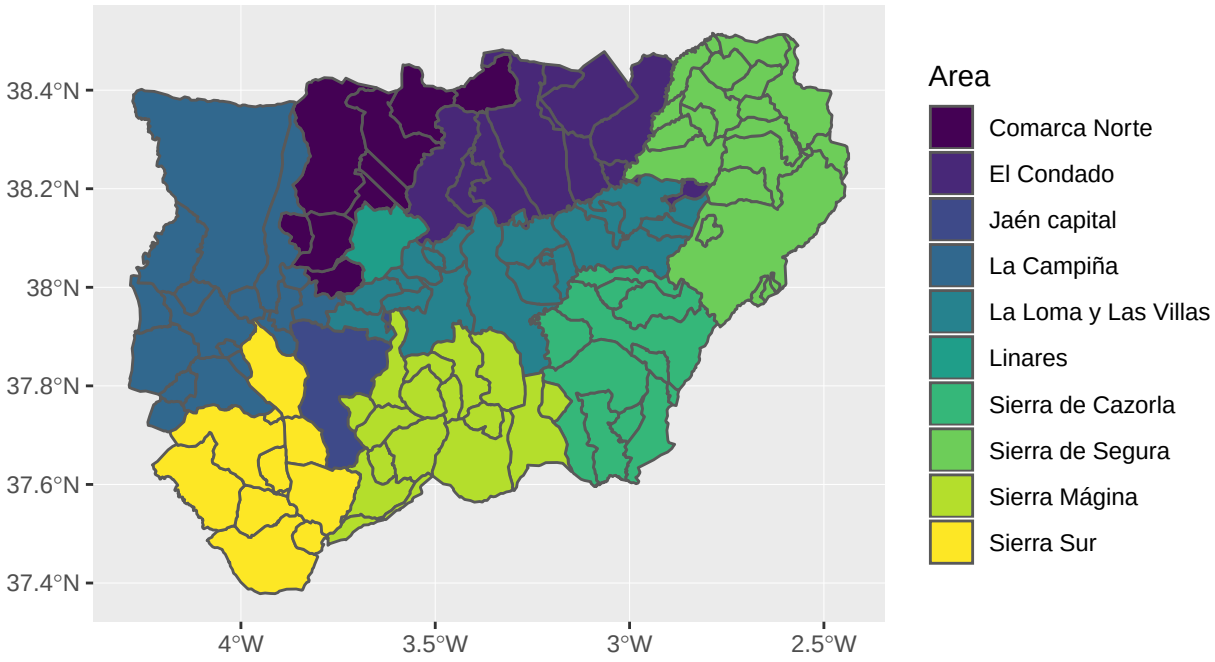


Figura 9: Mapa de Jaén según área territorial

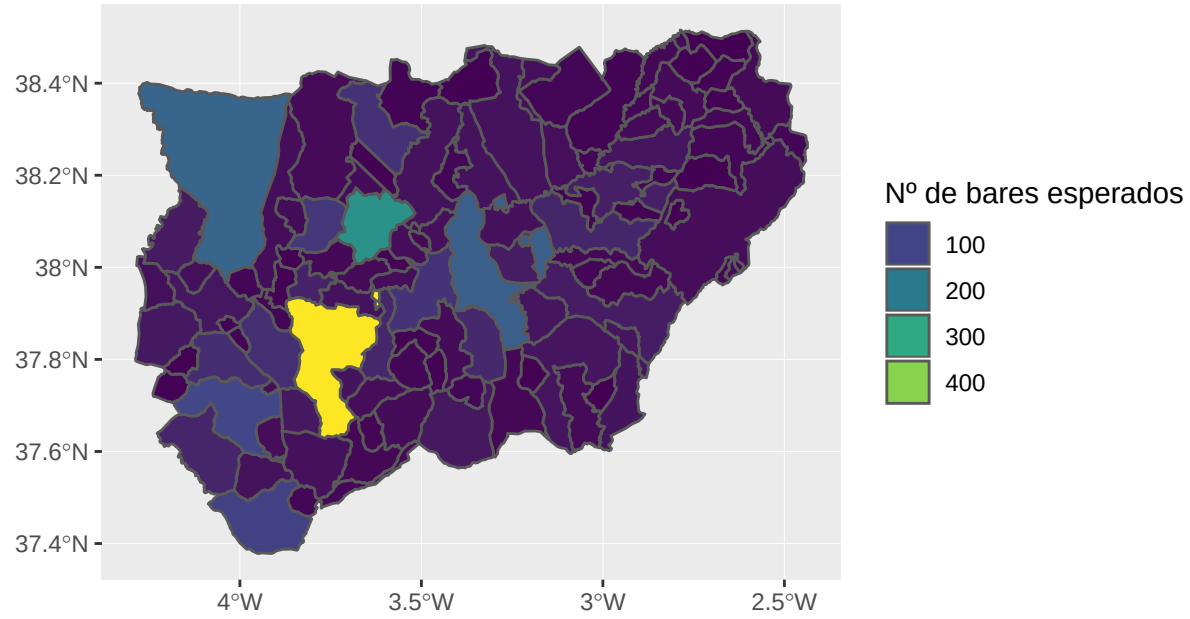


Figura 10: Mapa de Jaén según número de bares esperados por el modelo 3

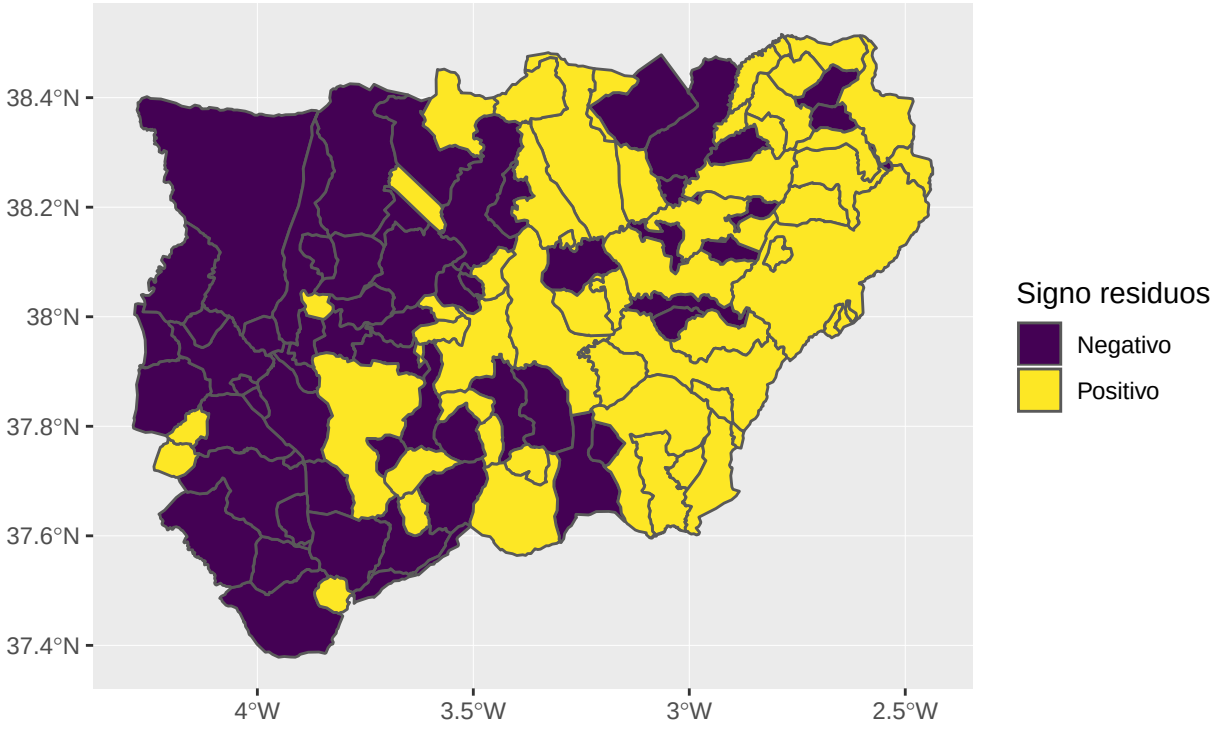


Figura 11: Mapa de Jaén según signo de los residuos

### 5.5. Otros modelos

Además de estos modelos se han probado otros en el que se incluían las siguientes covariables:

1. Número de oficinas de entidades de crédito por habitante. (Entidades de depósito)
2. Parque de turismos con mas de 2.000 cc por habitante
3. Renta neta declarada por habitante
4. Número de pensionistas
5. Pensión media
6. Número de hoteles por habitante

Dichos modelos se han omitido ya que estas covariables no aportaban información significativa a la hora de explicar el comportamiento de los bares.

A través de la siguiente aplicación web [2] realizada en Shiny, el usuario puede interactuar con el modelo, incluyendo covariables al mismo y obteniendo tanto resultados numéricos como gráficos.

## 6. Conclusiones y vías futuras de estudio futuras

El objetivo principal que planteábamos en este trabajo era analizar el comportamiento, desde el punto de vista aleatorio, de bares, restaurantes y cafeterías en el contexto de los municipios de la provincia de Jaén, considerando como variables explicativas diferentes indicadores de aspectos sociales o económicos. En este sentido, nuestra principal conclusión es que los resultados mostrados demuestran que el modelo de regresión de Poisson ajustado supone una descripción adecuada, por lo que podemos afirmar que, dadas las covariables del modelo, la variable bares, restaurantes y cafeterías queda explicada por el puro azar, carecterística fundamental de la distribución de Poisson.

Adicionalmente, el proceso de ajuste realizado mediante un análisis de los residuos partiendo del modelo nulo también nos conduce a destacar los siguientes hallazgos:

1. El modelo nulo que sólo incluye la población como offset presenta un muy elevado coeficiente $R^2$, lo que confirma que el número de establecimientos de este tipo queda principalmente explicado por la población, en términos multiplicativos.
2. A la luz de los municipios que han destacado, bien atípicos por su elevado residuo, bien influyentes en el modelo correspondiente, las covariables que han resultado ser regresores significativos han sido la ratio plazas hoteleras por habitante, la ratio de hectáreas de espacios naturales por habitante y una variable que indica la existencia en el municipio de centro histórico.
3. Ninguna del resto de covariables que fueron obtenidas como posibles regresores ha alcanzado la significación estadística.
4. En el modelo final ajustado se puede aceptar la hipótesis de equidispersión. Sin embargo, aún se rechaza el contraste de bondad de ajuste del modelo. Ello parece indicar todavía la presencia de cierta sobredispersión que no ha podido ser explicada por el modelo.
5. En este sentido, el análisis de los residuos desde un punto de vista espacial mediante mapas (Figura 11) ha permitido establecer como hipótesis para un futuro estudio la existencia de una componente espacial que puede formar parte de un modelo ampliado.

Existen, no obstante, otras vías futuras para la propuesta de un modelo alternativo más adecuado para los datos. Entre ellas, quizá la más inmediata sea el planteamiento de un ajuste mediante un modelo de regresión binomial negativo que, probablemente, se adapte mejor que el de Poisson precisamente en los casos en los que aún no se ha podido corregir la sobredispersión.

[2] http://estio.ujaen.es:3838/sample-apps/AppTFG/

# 7. Bibliografía


- Cameron, A.C. and Trivedi, P.K. (1990). Regression-based Tests for Overdispersion in the Poisson Model. Journal of Econometrics, 46, 347–364.
- Bartlett, J (2014). Deviance goodness of fit test for Poisson regression. *The Stat Geek* https://thestatsgeek.com/2014/04/26/deviance-goodness-of-fit-test-for-poisson-regression/.
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
- Yuan Tang, Masaaki Horikoshi, and Wenxuan Li. "ggfortify: Unified Interface to Visualize Statistical Result of Popular R Packages." The R Journal 8.2 (2016): 478-489.
- H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.
- Dabao Zhang (2018). rsq: R-Squared and Related Measures. R package version 1.1. https://CRAN.R-project.org/package=rsq.
- Christian Kleiber and Achim Zeileis (2008). Applied Econometrics with R. New York: Springer-Verlag. ISBN 978-0-387-77316-2. URL https://CRAN.R-project.org/package=AER.
- Hadley Wickham and Jennifer Bryan (2019). readxl: Read Excel Files. R package version 1.3.1. https://CRAN.R-project.org/package=readxl.
- Hadley Wickham (2017). tidyverse: Easily Install and Load the 'Tidyverse'. R package version 1.2.1. https://CRAN.R-project.org/package=tidyverse.
- Kamil Slowikowski (2018). ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.8.0. https://CRAN.R-project.org/package=ggrepel.
- Erich Neuwirth (2014). RColorBrewer: ColorBrewer Palettes. R package version 1.1-2. https://CRAN.R-project.org/package=RColorBrewer.
- Pebesma, E., 2018. Simple Features for R: Standardized Support for Spatial Vector Data. The R Journal, https://journal.r-project.org/archive/2018/RJ-2018-009/.
- Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables. R package version 5.2.1. https://CRAN.R-project.org/package=stargazer.
- Winston Chang, Joe Cheng, JJ Allaire, Yihui Xie and Jonathan McPherson (2018). shiny: Web Application Framework for R. R package version 1.2.0. https://CRAN.R-project.org/package=shiny.
- Ato, Manuel y López-García, Juan. (1996). Análisis estadístico para datos categóricos.
- Judd C.M, McClelland G. H and Ryan C. S (2009), Data analysis: a model comparison approach, Routledge.
- Ato, Manuel & Vallejo, Guillermo. (2005). Diseños experimentales en psicología.
- Instituto de Estadística y Cartografía de Andalucía. https://www.juntadeandalucia.es/institutodeestadisticaycartografi
- Clasificación Nacional de Actividades Económicas. https://www.cnae.com.es/
- Diario Expansión. http://www.expansion.com/
- R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

# 8. Anexos: Scripts de RStudio

## 8.1. Anexo I: Depurado de los datos

```
Cargamos los ficheros de datos

Directorio de trabajo
setwd("~/Universidad/Beca Colaboración/Segunda publicación")

Hostelería
library(readxl)
hosteleria <- read_excel("~/Universidad/Beca Colaboración/Segunda publicación/hosteleria.xlsx",
                         col_types = c("text", "text", "numeric",
                                       "numeric", "numeric", "numeric",
                                       "numeric", "numeric", "numeric",
                                       "numeric", "numeric"))
hosteleria<-hosteleria[,c(1,2,5,7)]
library(stringr)

hosteleria<-hosteleria[str_detect(hosteleria$CODIGO_INE4, "23"),]
hosteleria<-hosteleria[7:103,]
hosteleria[98,]<-c("Provincia","23",sum(hosteleria$Comercio, na.rm = TRUE),
                   sum(hosteleria$Hostelería, na.rm = TRUE))
hosteleria$Comercio<-as.numeric(hosteleria$Comercio)
hosteleria$Hostelería<-as.numeric(hosteleria$Hostelería)
names(hosteleria)[2]<-"codigo"
Hoteles
library(readr)
hoteles <- read_delim("~/Universidad/Beca Colaboración/Segunda publicación/hoteles.txt",
                      ";", escape_double = FALSE, col_types = cols(Valor = col_character()),
                      trim_ws = TRUE)
hoteles<-hoteles[,c(1,3,4,6)]
hoteles<-hoteles[str_detect(hoteles$CODIGO_INE3,"23"),]
hoteles<-hoteles[str_detect(hoteles$`Tipo de establecimiento hotelero`,"Total"),]
hoteles<-hoteles[42:138,c(2,4)]
names(hoteles)<-c("codigo","hoteles")
hoteles$hoteles<-as.numeric(hoteles$hoteles)
hoteles[98,]<-c("23",sum(hoteles$hoteles,na.rm = TRUE))
hoteles$hoteles<-as.numeric(hoteles$hoteles)

library(dplyr)
bares<-full_join(hosteleria,hoteles,by="codigo")
bares$bares<-bares$Hostelería-bares$hoteles


Cilindrada
cilindrada <- read_delim("~/Universidad/Beca Colaboración/Segunda publicación/cilindrada.txt",
                         ";", escape_double = FALSE, col_types = cols(Valor = col_character()),
                         trim_ws = TRUE)
cilindrada<-cilindrada[,c(1,3,4,6)]
cilindrada<-cilindrada[str_detect(cilindrada$CODIGO_INE3,"23"),]
cilindrada<-cilindrada[str_detect(cilindrada$`Tipo de cilindrada`,"2.000"),]
cilindrada<-cilindrada[42:138,c(2,4)]
```

```
names(cilindrada)<-c("codigo","mas2000cc")
cilindrada$mas2000cc<-as.numeric(gsub("[[:punct:]]","",cilindrada$mas2000cc))
cilindrada[98,]<-c("23",sum(cilindrada$mas2000cc,na.rm = TRUE))
cilindrada$mas2000cc<-as.numeric(cilindrada$mas2000cc)

bares<-full_join(bares,cilindrada,by="codigo")
bares<-data.frame(bares)

Entidades de depósito

oficinas <- read_excel("~/Universidad/Beca Colaboración/Segunda publicación/oficinas.xlsx",
                       col_types = c("text", "text", "numeric"))
names(oficinas)<-c("Area","Territorio","oficinas")

a<-anti_join(oficinas,bares,by="Territorio")
a<-data.frame(a)
oficinas<-data.frame(oficinas)

v<-numeric()
for (i in 1:length(a$Territorio)) {
  b<-rownames(oficinas[oficinas$Territorio==a$Territorio[i],])
  grep(a$Territorio[1],oficinas$Territorio)
  v<-c(v,b)
}

v<-as.numeric(v)
oficinas[v,2]<-c("Carolina (La)","Espelúy","Jaén (capital)","Iruela (La)",
                 "Bedmar y Garcíez","Bélmez de la Moraleda","Cambil","Guardia de Jaén (La)",
                 "Huelma","Noalejo","Pegalajar","Hornos","Puerta de Segura (La)",
                 "SantiagoPontones","Torres de Albánchez","Torre del Campo","Villares (Los)")

bares<-full_join(bares,oficinas,by="Territorio")
bares[98,8:9]<-c("Provincia",sum(bares$oficinas,na.rm = TRUE))
bares$oficinas<-as.numeric(bares$oficinas)

Renta neta declarada

renta <- read_delim("~/Universidad/Beca Colaboración/Segunda publicación/renta.txt",
                    ";", escape_double = FALSE, col_names = FALSE,
                    col_types = cols(X4 = col_character()),
                    trim_ws = TRUE)

renta<-renta[,c(2,4)]
names(renta)<-c("codigo","renta")
renta<-renta[str_detect(renta$codigo,"23"),]
renta<-renta[7:103,]
renta$renta<-gsub(".","",renta$renta, fixed = TRUE)
renta$renta<-as.numeric(gsub(",",".",renta$renta,fixed = TRUE))

bares<-full_join(bares,renta,by="codigo")
bares<-data.frame(bares)
bares[98,10]<-sum(bares$renta, na.rm = TRUE)
```

```
load("/Users/antonio/Universidad/Cuarto/TFG/script/datos.rda")
bares<-bares[,-c(3,4,6)]

Importamos bares
library(readxl)
BarResCaf <- read_excel("Universidad/Cuarto/TFG/script/Datos_definitivos/BarResCaf.xlsx",
                        col_types = c("text", "numeric"))
library(tidyverse)
bares<-full_join(bares,BarResCaf,by="codigo")

library(readxl)
espacios <- read_excel("Universidad/Cuarto/TFG/script/Primera solución/espacios.xltx")
espacios<-espacios[,c(2,11)]
names(espacios)[1]<-"codigo"
espacios$codigo<-as.character(espacios$codigo)
bares<-full_join(bares,espacios,by="codigo")


library(GLDEX)
bares$Superficie[which.na(bares$Superficie)]<-0
bares$centro<-bares$codigo
bares$centro[c(2,5,9,11,13,17,21,25,34,35,39,40,43,44,46,49,56,70,71,73,75,80,84)]<-"Sí"
bares$centro[-c(2,5,9,11,13,17,21,25,34,35,39,40,43,44,46,49,56,70,71,73,75,80,84)]<-"No"


library(readr)
pensiones <- read_delim("Universidad/Cuarto/TFG/script/Primera solución/pensiones.txt",";",
                        escape_double = FALSE, col_names = FALSE,trim_ws = TRUE)
pensiones<-pensiones[,c(2,5,6,7,8,10)]
pensiones<-na.omit(pensiones)
pensiones<-pensiones[8521:10266,]
pensiones<-pensiones[pensiones$X6=="Ambos sexos",]
pensiones<-pensiones[pensiones$X7=="TOTAL",]
pensiones<-pensiones[,c(1,2,5,6)]
pensiones<-spread(pensiones,X8,X10)
names(pensiones)[1:2]<-c("Territorio","codigo")
bares<-full_join(bares,pensiones[,2:4],by="codigo")
bares$`Número de pensionistas`<-as.numeric(bares$`Número de pensionistas`)
bares$`Pensión media`<-as.numeric(bares$`Pensión media`)
save(bares,file = "datos.rda")
```

## 8.2. Anexo II: Modelado

Modelo nulo

```
load("datos.rda")
modelo <- glm(BarResCaf ~ offset(log(Poblacion)), data = bares[-c(98),], family = poisson)
```

Modelo 1

```
library(readxl)
plazahotel <- read_excel("~/Universidad/Cuarto/TFG/script/Datos_definitivos/plazahotel.xlsx",
    col_types = c("text", "numeric"))
library(dplyr)
```

```r
bares<-full_join(bares,plazahotel,by="codigo")

bares$plazashotel.est<-bares$plazashotel/bares$Poblacion

modelo1<-glm(BarResCaf~offset(log(Poblacion))+plazashotel.est,
             data = bares[-c(98),], family = poisson)

anov<-data.frame("Chi-squared"=modelo$deviance-modelo1$deviance,
                 "p-value"=1-pchisq(modelo$deviance-modelo1$deviance,
                                    modelo$df.residual-modelo1$df.residual),
                 "df"=modelo$df.residual-modelo1$df.residual)
```

Modelo 2

```r
bares$Superficie.est<-bares$Superficie/bares$Poblacion
modelo2<-glm(BarResCaf~offset(log(Poblacion))+plazashotel.est+Superficie.est,
             data = bares[-c(98),], family = poisson)

anov<-data.frame("Chi-squared"=modelo$deviance1-modelo2$deviance,
                 "p-value"=1-pchisq(modelo1$deviance-modelo2$deviance,
                                    modelo1$df.residual-modelo2$df.residual),
                 "df"=modelo1$df.residual-modelo2$df.residual)
```

Modelo 3

```r
modelo3<-glm(BarResCaf~offset(log(Poblacion))+plazashotel.est+Superficie.est+centro,
             data = bares[-c(98),], family = poisson)
anov<-data.frame("Chi-squared"=modelo$deviance2-modelo3$deviance,
                 "p-value"=1-pchisq(modelo2$deviance-modelo3$deviance,
                                    modelo2$df.residual-modelo3$df.residual),
                 "df"=modelo2$df.residual-modelo3$df.residual)
```

## 8.3. Anexo III: Representaciones gráficas

Figura 1

```r
library(ggplot2)
ggplot() + geom_point(aes(x = log(bares$Poblacion[1:97]),
                          y = bares$BarResCaf[1:97])) +
  geom_line(aes(x = log(bares$Poblacion[1:97]),
                y = modelo$fitted.values))+
  labs(x="Logaritmo Población",y="Bares, restaurantes y cafeterías")
```

Figura 6

```r
#Gráfico observado vs esperado
df<-data.frame(bares$Territorio[1:97],bares$BarResCaf[1:97],
               modelo$fitted.values,bares$Area[1:97])
names(df)<-c("municipio","observado","esperado","Area")
library(ggplot2)
library(ggrepel)
linea<-data.frame(y=seq(0,600),x=seq(0,600))
ggplot()+geom_point(data = df, aes(esperado,observado,color=Area,size=Area))+
  geom_line(data=linea, aes(x=x,y=y))+
  geom_text_repel(data = df[df$observado>100|df$esperado>100,],
                  aes(esperado,observado,label=municipio))
```

Figura 7

```
#Gráfico observado vs esperado haciendo "zoom" en los municipios con menos bares
ggplot()+geom_point(data = df, aes(esperado,observado,color=Area,size=Area))+
  geom_line(data=linea, aes(x=x,y=y))+
  geom_text_repel(data = df[df$observado>30|df$esperado>30,],
                  aes(esperado,observado,label=municipio))+
  lims(x=c(30,100),y=c(30,100))
```

Figura 8

```
#Gráfico observado vs esperado haciendo "zoom" en los municipios con menos bares
ggplot()+geom_point(data = df, aes(esperado,observado,color=Area,size=Area))+
  geom_line(data=linea, aes(x=x,y=y))+
  geom_text_repel(data = df[(df$observado<=30|df$esperado<=30)&
                              (df$observado>15|df$esperado>15),],
                  aes(esperado,observado,label=municipio))+
  lims(x=c(0,30),y=c(0,30))
```

Coordenadas mapa de Jaén

```
library(sf)
library(tidyverse)
nc = st_read("~/Universidad/espana-municipios.geojson")
indice<-as.numeric(rownames(nc[nc$municipio=="Los Villares"|nc$municipio=="La Iruela"|
                                 nc$municipio=="La Guardia de Jaén"|nc$municipio=="Jaén"|
                                 nc$municipio=="La Carolina"|
                                 nc$municipio=="La Puerta de Segura"|
                                 nc$municipio=="Santiago-Pontones",]))
nc$municipio<-as.character(nc$municipio)
nc$municipio[indice]<-c("Villares (Los)","Iruela (La)","Carolina (La)",
                        "Puerta de Segura (La)","Guardia de Jaén (La)",
                        "SantiagoPontones","Jaén (capital)")
a<-full_join(nc,df,by="municipio")
a<-a[-c(91),]
df2 <- data.frame(matrix(unlist(a$geo_point_2d), nrow=length(a$geo_point_2d)
                         , byrow=T),a$municipio,a$observado,a$esperado)
library(RColorBrewer)
library(ggrepel)
```

Figura 9

```
ggplot() + geom_sf(data = a, aes(fill=a$Area))+
  scale_fill_viridis_d()+
  guides(fill=guide_legend(title="Area"))
```

Figura 10

```
#Más bares observados de lo esperado
ggplot() + geom_sf(data = a, aes(fill=a$esperado))
+scale_fill_viridis_c()
+guides(fill=guide_legend(title="Nº de bares esperados"))
+labs(x="",y="")
```

Figura 11

```r
for (i in 1:nrow(a)) {
  if(a$observado[i]>a$esperado[i]){
    a$dic[i]<-"Positivo"
  } else a$dic[i]<-"Negativo"

}
ggplot() + geom_sf(data = a, aes(fill=a$dic))+
  scale_fill_viridis_d()+
  guides(fill=guide_legend(title="Signo residuos"))
```

## 8.4. Anexo IV: Aplicación en Shiny

```r
library(shiny)
library(ggplot2)
load("~/Universidad/Cuarto/TFG/datos.rda")

ch<-names(datos)[c(4,6:7,11:13,15:17)]

shinyUI(fluidPage(

  headerPanel("Modelo de Regresión de Poisson"),
  sidebarPanel(
    #Probar con checkbox
    selectInput("var1", "Covariable", choices = ch),
    selectInput("var2", "Covariable", choices = ch),
    selectInput("var3", "Covariable", choices = ch),
    selectInput("var4", "Covariable", choices = ch),
    selectInput("var5", "Covariable", choices = ch),
    selectInput("var6", "Covariable", choices = ch)
  ),

  mainPanel(
    h4('Regresión:'),
    verbatimTextOutput("modelo")

  )
))

shinyServer(function(input, output) {
  load("~/Universidad/Cuarto/TFG/datos.rda")
  fit<-reactive({

    glm(as.formula(paste("BarResCaf","~",input$var1,"+",input$var2,
                         "+",input$var3,"+",input$var4,"+",
                         input$var5,"+",input$var6)),
        data = datos[-c(98),], family = poisson)
  })


  output$modelo<-renderPrint({
    summary(fit())
  })
```

```
  }
)
```