GRAU INTERUNIVERSITARI D'ESTADÍSTICA I INVESTIGACIÓ OPERATIVA (UB- UPC)

CURS 2013-2014 Q1 – EXAMEN PARCIAL : MODEL LINEAL GENERALITZAT

(Data: 29 d'Octubre del 2013 a les 15:00h Aula 004-FME)

Professors:	Lídia Montero – Josep Anton Sànchez
Localització:	ETSEIB 6a Planta 6-67
Normativa de l'examen:	ÉS POT DUR APUNTS TEORIA *SENSE ANOTACIONS*, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	1h 00 min
Sortida de notes:	Abans del 14 de Novembre al Web Docent de MLGz
Revisió de l'examen:	14 de Novembre a 16:30 h a Sala Professors FME– Campus Sud

Problema 1 (2 punts): Forma Canònica

La forma canónica de una ley gamma se ha trabajado en los ejercicios del Tema 1.

La ley gamma $f_Y(y)=\left(\frac{y^{\beta-1}\,exp(-y/\alpha)}{\alpha^{\beta}\,\Gamma(\beta)}\right)\ con\ \alpha>0,\beta>0\quad y\geq 0\ real$ donde $\alpha\ i\ \beta$ son los parámetros de escala y forma respectivamente. Además, $\mathrm{E}[Y]=\alpha\beta\ \ i\ V[Y]=\alpha^2\beta$.

Expresión	$\exp\left(\ln\left(y^{\beta-1}\right)-\frac{y}{\alpha}-\ln\left(\alpha^{\beta}\right)-\ln\left(\Gamma(\beta)\right)\right)=\exp\left(-\frac{y}{\alpha}-\ln\left(\alpha^{\beta}\right)+\ln\left(y^{\beta-1}\right)-\ln\left(\Gamma(\beta)\right)\right)$ $=\exp\left(\frac{-\frac{y}{\alpha\beta}-\ln(\alpha)}{1/\beta}+\ln\left(y^{\beta-1}\right)-\ln\left(\Gamma(\beta)\right)\right)$
θ	$-\frac{1}{\alpha\beta}$
$b(\theta)$	$-\ln(-\theta)$ Ull $\ln(\alpha)=-\ln(-\theta)-\ln(\beta)$
$a(\phi)\ i\ \phi$	$a(\phi)=\phi\ \ i\ \phi=1/\beta$
$c(y,\phi)$	$\frac{1}{\phi}\ln\left(y^{\frac{1}{\phi}}\right)-\ln(y)-\ln\left(\Gamma\left(\frac{1}{\phi}\right)\right)\leftarrow\left\langle\ln\left(y^{\beta}\right)-\ln(y)-\ln\left(\Gamma(\beta)\right)+\ln(\beta)\right.$

1. Calcula la derivada de la función cumulante respecto al parámetro canónico.

$$b'(\theta)=-\frac{1}{\theta}\rightarrow \mathrm{E}[Y]=\mu$$

2. La derivada de la función cumulante es la esperanza matemática de la variable Y, ¿cuál es el enlace canónico para la ley gamma?

L'enllaç canònic expressa la relació entre el paràmetre canònic (natural) i l'esperança matemática de la variable per tant en aquest cas és directe d'afirmar a partir del resultat de l'apartat 1 que la funció inversa serà l'enllaç canònic per la distribució gamma.

$$\eta=g(\mu)=\theta\rightarrow g(\mu)=-\frac{1}{\mu}=\theta$$

Problema 2 (8 punts): Tasa de criminalidad en USA

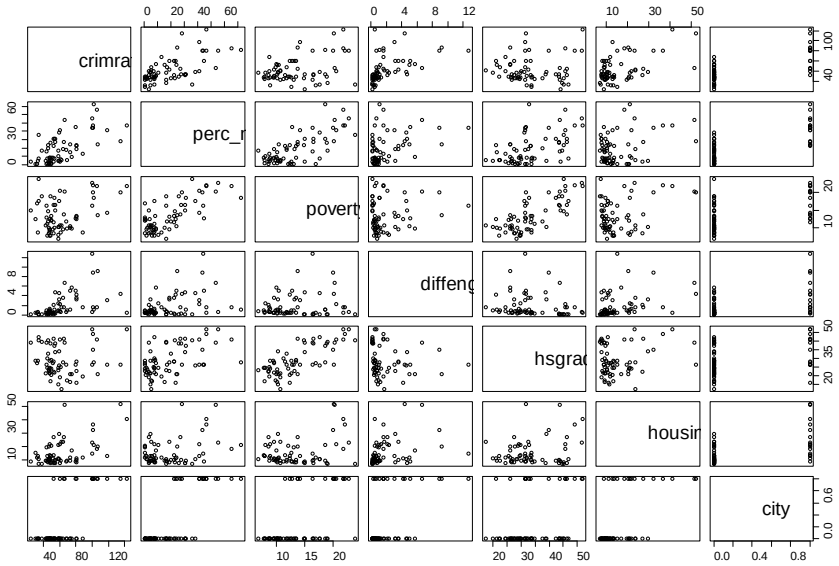
El conjunto de datos "`eriksen`" contiene información de áreas del censo de USA del 1980. Los primeros 50 datos corresponden a los 50 estados y las 16 últimas observaciones son las principales ciudades. Los datos de los estados con alguna de estas ciudades se refieren al resto del estado sin incluir la ciudad. Si

bien el objetivo inicial era ajustar el censo de hogares, en este caso analizaremos los factores relacionados con la tasa de criminalidad.

Los campos son los siguientes:

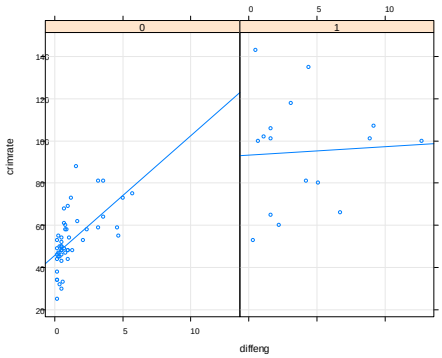
```
  * area     = "Nombre del Área (áreas marcadas _R son resto del estado)"
  * crimrate = "Tasa de crímenes mayores por 1000 habitantes"
  * perc_min = "Porcentaje de minorías (población hispana o de color)”
  * poverty  = "Porcentaje de pobreza"
  * diffeng  = "Porcentaje que tienen problemas con el inglés hablado y/o escrito "
  * hsgrad   = "Porcentaje con edad mayor de 25 que no tienen estudios secundarios "
  * housing  = "Porcentaje de hogares en pisos pequeños en grandes edificios "
  * city     = "Ciudad 1=si, 0=no (Estado)" ;
```

El matrix-plot de los datos (método pairs) es el siguiente:



Pregunta 2.1 (2 puntos): ANCOVA

Exploramos con detalle la relación entre el porcentaje de habitantes con dificultades con el inglés y la tasa de criminalidad, según sea una ciudad principal o un área mayor. El siguiente plot representa los diagramas de puntos con las rectas ajustadas en cada grupo por mínimos cuadrados. Se incluye el modelo con ambas variables y la interacción entre ellas, calculado con dos contrastes activos diferentes



Contraste Activo tipo baseline con referencia: primera categoría (contr.treatment en R)

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)       45.685      3.020  15.127  < 2e-16 ***
diffeng            5.692      1.596   3.566 0.000705 ***
city1             47.652      6.696   7.117 1.35e-09 ***
diffeng:city1     -5.307      1.949  -2.723 0.008383 **

Residual standard error: 15.89 on 62 degrees of freedom
Multiple R-squared:  0.6113,Adjusted R-squared:  0.5925
F-statistic:  32.5 on 3 and 62 DF,  p-value: 9.532e-13
```

Contraste Activo tipo baseline con referencia: última categoría (contr.SAS en R)

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      93.3371     5.9763  15.618  < 2e-16 ***
diffeng           0.3851     1.1180   0.344  0.73169
city0           -47.6523     6.6960  -7.117 1.35e-09 ***
diffeng:city0     5.3073     1.9488   2.723  0.00838 **
```

1. Suponiendo validado el modelo, ¿hay relación entre el porcentaje de población con dificultades con el inglés y la tasa de criminalidad, según sea ciudad o un área mayor (estados)? Justifica la respuesta indicando el modelo ajustado, la prueba estadística que sirve para determinar la relación y el p-valor asociado en cada uno de los dos estratos (Ciudad/Estado).

Con la expresión del modelo bajo un contraste ya se podrían deducir los modelos para cada estrato:

Si el dato corresponde a un estado (City=0):

$$crimrate = 45.685 + 5.692\ diffeng + \varepsilon$$

Si el dato corresponde a una ciudad (City=1):

$$crimrate = (45.685 + 47.652) + (5.692 - 5.307)\ diffeng + \varepsilon$$
$$crimrate = 93.337 + 0.385\ diffeng + \varepsilon$$

La expresión de ambos sub-modelos también se puede extraer de la ordenada en el origen y el coeficiente de la variable diffeng de cada modelo, porque, puesto que la variable es binaria, ambas categorías aparecen como referencia baseline en alguno de los dos modelos. Para establecer la relación entre la variable numérica (diffeng) y la respuesta (crimrate) en cada estrato es necesario realizar un test de significación del parámetro del modelo correspondiente al coeficiente de la variable diffeng:

Relación crimrate-diffeng en estados (city=0)

Prueba estadística de significación del coeficiente (Test de Wald):

$$H_0: \beta_{diffeng} = 0$$
$$H_1: \beta_{diffeng} \neq 0$$

Estadístico (t-ratio):

$$\hat{t} = \frac{\hat{\beta}_{diffeng} - 0}{S_{\hat{\beta}_{diffeng}}} = \frac{5.692}{1.596} = 3.566$$

Distribución de referencia bajo H_0: t-Student con N-p grados de libertad: t_{62}

P-valor: 0.000705<0.05 ➔ Rechazamos la hipótesis nula y por lo tanto existe relación entre ambas variables.

Relación crimrate-diffeng en ciudades (city=1)

Prueba estadística de significación del coeficiente (Test de Wald):

$$H_0: \beta_{diffeng} = 0$$
$$H_1: \beta_{diffeng} \neq 0$$

Estadístico (t-ratio):

$$\hat{t} = \frac{\hat{\beta}_{diffeng} - 0}{S_{\hat{\beta}_{diffeng}}} = \frac{0.385}{1.118} = 0.344$$

Distribución de referencia bajo H_0: t-Student con N-p grados de libertad: t_{62}

P-valor: 0.73169>0.05 ➔ No hay evidencias estadísticas significativas que permitan establecer relación entre ambas variables.

El modelo ANCOVA en este caso establece que la interacción es significativa y por lo tanto la relación entre la covariable y la respuesta es diferente en ambos estratos (no es un modelo de rectas paralelas, sino de rectas con diferente pendiente). El resultado es consecuente con la representación gráfica donde se aprecia que la recta en el caso de las ciudades es prácticamente horizontal.

Pregunta 2.2 (2 puntos): Inferencia

Se ajusta un modelo con algunas variables, obteniendo la tabla de los coeficientes (método `summary`) y la tabla de análisis de la varianza del modelo (método `anova`).

```
> summary(m3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  86.9211    11.3242   7.676 1.59e-10 ***
perc_min      0.7816     0.1829   4.274 6.86e-05 ***
hsgrad       -1.1093     0.2577  -4.306 6.14e-05 ***
housing       0.6726     0.2270   2.963  0.00434 **
city0       -16.1846     7.4870  -2.162  0.03457 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 14.52 on 61 degrees of freedom
Multiple R-squared:  0.6807,Adjusted R-squared:  0.6598
F-statistic: 32.52 on 4 and 61 DF,  p-value: 1.638e-14

> anova(m3)
Analysis of Variance Table

Response: crimrate
          Df  Sum Sq Mean Sq F value    Pr(>F)
perc_min   1 17272.1 17272.1 81.9462 6.968e-13 ***
hsgrad     1  4549.0  4549.0 21.5824 1.856e-05 ***
housing    1  4608.4  4608.4 21.8643 1.666e-05 ***
city       1   984.9   984.9  4.6729   0.03457 *
Residuals 61 12857.2   210.8
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

2. En esta caso ambos métodos establecen la significación de las variables introducidas, pero no todos los p-valores coinciden. Indica que tipo de test corresponde a ambos métodos y a qué es debido las diferencias en los p-valores. ¿Por qué el único p-valor que coincide es el asociado a la variable binaria “`city`”?

En la tabla summary aparecen los resultados de la prueba de significación del cada coeficiente (test de Wald) en donde se compara la contribución neta de la variable al modelo (comparamos el modelo con todas las variables, con el modelo en el que se suprime la variable correspondiente para determinar si es significativa).

En el método anova estándar la comparación de las sumas de cuadrados por defecto se hace de forma incremental. En el caso de respuesta normal corresponde a un test de Wald o a uno de razón de verosimilitudes ya que son equivalentes. Puesto que la comparación se hace de forma secuencial, los modelos que se comparan son: el modelo con las variables indicadas anteriormente con el modelo que incluye esas variables y la de la línea correspondiente, pero no tiene en cuenta las variables que se incluyen en líneas posteriores. Por ejemplo, el primer p-valor (6.97e-13) resuelve el test que compara el modelo con el intercept solamente con el modelo que posee el intercept y la variable perc_min y por lo tanto establece la significación del coeficiente asociado a esta variable cuando en el modelo no hay más predictores. Esto explica que no coincidan los p-valores de ambas tablas, ya que comparan modelos diferentes, excepto en la ultima variable, donde los modelos que se comparan son los mismos en ambas tablas. Por esta razón,

el p-valor de la variable city es la misma en ambas tablas por ser la última variable introducida en el modelo.

Pregunta 2.3 (2 puntos): Interpretación

Aplicando el mecanismo stepwise usando el criterio BIC e incluyendo todas las variables y las interacciones entre las variables númericas y la binaria (city) se obtiene el siguiente modelo:

```
Residuals:
    Min      1Q  Median      3Q     Max
-40.723  -4.866  -0.819   4.191  44.517

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     58.9060    15.7469   3.741 0.000417 ***
perc_min        -0.1483     0.3019  -0.491 0.624994
poverty          5.0582     1.2324   4.104 0.000126 ***
hsgrad          -1.2310     0.3038  -4.052 0.000151 ***
city0           33.4987    17.1336   1.955 0.055308 .
perc_min:city0   1.1437     0.3822   2.993 0.004033 **
poverty:city0   -6.0391     1.2841  -4.703 1.59e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 13.45 on 59 degrees of freedom
Multiple R-squared:  0.7351,Adjusted R-squared:  0.7082
F-statistic: 27.29 on 6 and 59 DF,  p-value: 2.545e-15
```

3. Suponiendo validado el modelo, especifica los modelos que se obtienen para predecir la tasa de criminalidad en cada estrato (ciudad/estado).

Hay que tener en cuenta que el contraste activo es de tipo baseline con la última categoría como nivel de referencia. Por ello, los coeficientes de las variables corresponden a la categoría city=1, es decir, cuando la información hace referencia a una ciudad.

Si el dato corresponde a un estado (City=0):

$$crimrate = (58.91 + 33.50) + (1.14 - 0.15)percmin + (5.06 - 6.04)\ poverty - 1.23\ hsgrad + \varepsilon$$
$$crimrate = 92.41 + 0.99\ percmin - 0.98\ poverty - 1.23\ hsgrad + \varepsilon$$

Si el dato corresponde a una ciudad (City=1):

$$crimrate = 58.91 - 0.15\ permin + 5.06\ poverty - 1.23\ hsgrad + \varepsilon$$

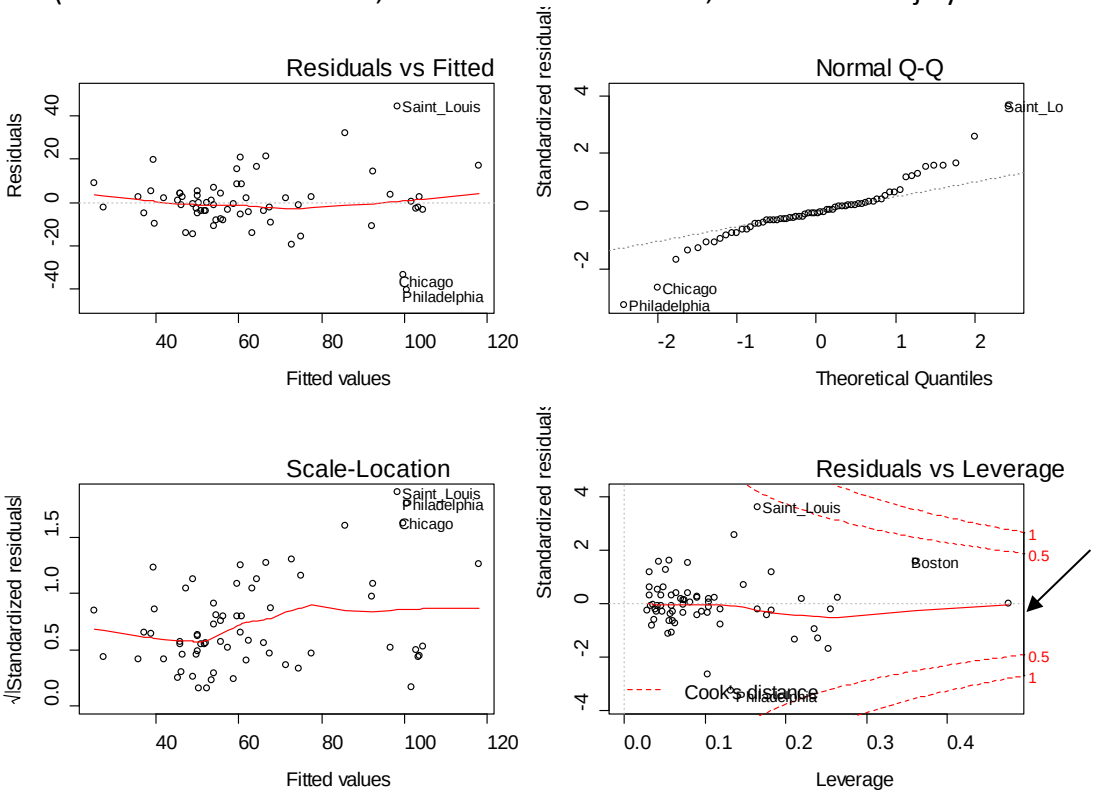
4. Según el modelo, ¿qué factores están relacionados con la tasa de criminalidad en los datos de las ciudades y en qué sentido lo están? ¿Y en los estados?

En el caso de los estados, el porcentaje de minorías está relacionado con la respuesta de forma directa (mayor presencia de minorías está relacionada con una mayor tasa de criminalidad). En cambio, el porcentaje de pobreza posee signo negativo indicando relación lineal inversa, pero no disponemos del p-valor para este coeficiente por lo que podría ser no significativo. Finalmente, el porcentaje de habitantes sin estudios secundarios tiene relación inversa, debido a su signo negativo. La interpretación sería que a mayor porcentaje de gente sin estudios secundarios, menor criminalidad. Hay que tener en cuenta que no se ha analizado la correlación entre los predictores seleccionados, y por tanto, las interpretaciones podrían estar influidas por una posible multicolinealidad del modelo.
En el caso de las ciudades, el porcentaje de pobreza es claramente significativo y con signo positivo, indicando una relación directa con la respuesta. El porcentaje de minorías no es significativo y por lo tanto no se establece relación entre esta covariable y la tasa de criminalidad. Respecto al porcentaje de

habitantes sin estudios secundarios, el comentario es el mismo que en el caso de los estados, ya que la interacción de esta variable con la binaria no aparece en el modelo como significativa.

Pregunta 2.4 (2 punts): Validació

Para validar el modelo anterior, se realizan los plots para el análisis de residuos y las medidas de influencia para cada dato (residuos estandarizados, residuos estudentizados, factor de anclaje y distancia de Cook).



	rstandard	rstudent	hatvalue	cooks.d
Alabama	0.21073	0.20901	0.09052	0.00063
Alaska	-0.31746	-0.31502	0.10393	0.00167
Arizona	1.27566	1.28262	0.05157	0.01264
Arkansas	0.17208	0.17066	0.10542	0.00050
California_R	-0.10888	-0.10796	0.10460	0.00020
Colorado	0.13257	0.13146	0.07380	0.00020
Connecticut	-0.05733	-0.05685	0.05936	0.00003
Delaware	0.63495	0.63171	0.03130	0.00186
Florida	1.56135	1.58107	0.04345	0.01582
Georgia	0.39324	0.39040	0.07836	0.00188
Hawaii	1.17105	1.17482	0.03116	0.00630
Idaho	-0.29542	-0.29312	0.05987	0.00079
Illinois_R	-0.63132	-0.62807	0.05981	0.00362
Indiana_R	-0.06785	-0.06728	0.04633	0.00003
Iowa	-0.20507	-0.20340	0.03802	0.00024
Kansas	-0.26372	-0.26163	0.02906	0.00030
Kentucky	0.71896	0.71598	0.14740	0.01277
Louisiana	0.05151	0.05108	0.10423	0.00004
Maine	0.16806	0.16667	0.07019	0.00030
Maryland_R	-0.76435	-0.76162	0.11962	0.01134
Massachusetts_R	-0.08513	-0.08441	0.04093	0.00004
Michigan_R	0.52913	0.52587	0.04109	0.00171
Minnesota	-0.30838	-0.30601	0.03971	0.00056
Mississippi	-1.27620	-1.28318	0.24051	0.07368
Missouri_R	-0.09240	-0.09162	0.03292	0.00004
Montana	-0.02473	-0.02452	0.07277	0.00001
Nebraska	-0.82832	-0.82609	0.03472	0.00353
Nevada	1.63336	1.65736	0.05525	0.02229
New_Hampshire	-0.30212	-0.29978	0.04754	0.00065
New_Jersey	0.16475	0.16338	0.07419	0.00031
New_Mexico	-1.34624	-1.35577	0.21104	0.06926
New_York_R	-0.57380	-0.57051	0.03733	0.00182
North_Carolina	0.05891	0.05841	0.08109	0.00004

	rstandard	rstudent	hatvalue	cooks.d
North_Dakota	-0.74380	-0.74095	0.06320	0.00533
Ohio_R	-0.02522	-0.02500	0.03619	0.00000
Oklahoma	0.30468	0.30232	0.03152	0.00043
Oregon	0.32325	0.32078	0.04651	0.00073
Pennsylvania_R	-1.08761	-1.08933	0.05856	0.01051
Rhode_Island	1.52667	1.54449	0.07859	0.02840
South_Carolina	0.23797	0.23605	0.11173	0.00102
South_Dakota	-0.42700	-0.42402	0.17675	0.00559
Tennessee	0.40554	0.40265	0.06456	0.00162
Texas_R	-0.42963	-0.42664	0.09010	0.00261
Utah	-0.33674	-0.33420	0.07268	0.00127
Vermont	0.32705	0.32456	0.05898	0.00096
Virginia	-0.64693	-0.64372	0.05581	0.00353
Washington	0.64111	0.63788	0.04909	0.00303
West_Virginia	-0.19106	-0.18949	0.11933	0.00071
Wisconsin_R	-0.38553	-0.38273	0.05696	0.00128
Wyoming	-1.09516	-1.09705	0.05391	0.00976
Baltimore	-0.24572	-0.24375	0.18250	0.00193
Boston	1.58350	1.60449	0.36032	0.20177
Chicago	-2.64704	-2.79577	0.10336	0.11538
Cleveland	-0.18912	-0.18757	0.16505	0.00101
Dallas	2.57807	2.71353	0.13649	0.15008
Detroit	0.20209	0.20044	0.22023	0.00165
Houston	0.21180	0.21008	0.26367	0.00229
Indianapolis	-1.69922	-1.72756	0.25316	0.13982
Los_Angeles	0.26351	0.26142	0.09084	0.00099
Milwaukee	-0.22049	-0.21870	0.25609	0.00239
New_York_City	-0.27766	-0.27548	0.09595	0.00117
Philadelphia	-3.25087	-3.55752	0.13217	0.22992
Saint_Louis	3.62433	4.07572	0.16563	0.37250
San_Diego	-0.94865	-0.94783	0.23556	0.03962
San_Francisco	1.19083	1.19515	0.18219	0.04513
Washington_DC	0.02878	0.02853	0.47523	0.00011

5. Realiza la validación del modelo, indicando en cada gráfico las premisas que permite analizar.

Las premisas del modelo son : linealidad , homocedasticidad , normalidad e independencia .

El primer plot es el de los residuos frente las predicciones , permite ver si la disposición de los residuos es aleatoria alrededor del cero , sin que se observe ningún patrón que indicas desviaciones de la relación lineal. El ajuste local (línea roja) es prácticamente horizontal , confirmando en este caso no parece haber patrones de no linealidad . En este plot también se puede verificar descriptivamente si la varianza puede considerarse constante, frente a las predicciones. En este caso, no se observa incremento de la variabilidad de los residuos a medida que aumenta la predicción, indicando que se puede asumir homocedasticidad. También en este plot , aparecen etiquetadas las observaciones con residuos estandarizados superior a 2 (aprox) en valor absoluto (valores atípicos).

El segundo plot es el plot de normalidad, que permite determinar si podemos considerar que la distribución Normal es adecuada para los residuos. Si los puntos están alineados podemos asumir Normalidad de los residuos. Este plot permitiría ver patrones de asimetría o colas pesadas en los residuos que irían en contra de la hipótesis de normalidad. También se etiquetan los atípicos. En este caso, la disposición de los puntos no está del todo alineada y existen valores que se separan en las colas, sugiriendo que la normalidad de los residuos es dudosa.

El tercer plot representa la raíz cuadrada de los valores absolutos de los residuos frente a las predicciones. Es un plot que permite determinar de forma más clara la presencia de heteroscedasticidad . El ajuste local mediante la recta no indica un claro incremento de los valores que constituyen una estimación de la varianza de los residuos. No es concluyente para confirmar la presencia de varianza no constante y además se ve influido por la presencia de atípicos que están relacionados con valores altos de las predicciones. La presencia de éstos últimos podría explicar el ligero incremento del ajuste local.

El cuarto modelo permite identificar y caracterizar los datos influyentes. Representa los residuos estandarizados frente al factor de anclaje/apalancamiento (leverage). Además incluye curvas de nivel para indicar la distancia de Cook de las observaciones. Valores con una distancia de Cook alta pueden ser valores influyentes y se debe analizar su efecto en el ajuste del modelo. La distancia de Cook es una función creciente de los residuos al cuadrado y del leverage. Las observaciones que tienen un valor alto de la distancia de Cook aparecen etiquetadas (pueden ser por tener muy leverage , o tener un residuos alto en valor absoluto o una combinación de ambas situaciones no tan extremas). Las observaciones etiquetadas como influyentes parece que tienen un leverage alto ya la vez tienen un residuo de magnitud elevada. Habría que analizar qué efecto tienen en la estimación del modelo.

6. ¿Qué diferencia hay entre residuos estandarizados y estudentizados? ¿A qué es debido que las observaciones con distancia de Cook mayor sean los que presenten mayores diferencias entre los residuos estandarizados y estudentizados (St. Louis, Philadelphia)?

Los residuos estandarizados se calculan dividiendo los residuos crudos por la estimación del modelo para la varianza residual. En cambio, los residuos estudentizados utilizan para cada residuo la estimación del modelo obtenido con todas las observaciones excluyendo la de ese residuo. Puesto que los residuos extremos suelen implicar un aumento en la estimación de la varianza residual, el residuo estandarizado correspondiente a un atípico podría estar enmascarado por haber utilizado una estimación inflada de la varianza residual. En este sentido, los residuos estudentizados presentan una mayor sensibilidad a la hora de identificar residuos atípicos o extremos, ya que la desviación estándar que utiliza no está afectada por ese atípico.

Una distancia de Cook elevada puede estar relacionada con un residuo de gran magnitud, así como del factor de apalancamiento. La distancia de Cook es una medida global de influencia de la observación en el modelo.

Aquellas observaciones con distancia de Cook elevadas corresponden a observaciones que, en caso de excluirse de modelo dan lugar a cambios en los estimadores de los parámetros. En particular, la estimación de la varianza residual suele verse afectada si el dato es influyente. Esto explica que habitualmente, las observaciones con distancia de Cook elevada son las que presentan mayores diferencias entre residuos estandarizados y estudentizados.

7. El punto situado más a la derecha en el cuarto gráfico (señalado con una flecha) es Washington DC. Caracteriza este dato en términos de su influencia en el modelo. ¿Es necesario eliminarlo del ajuste?

El factor de apalancamiento (leverage) de este dato es 0.475 (el mayor de todos). Por ello aparece situado en el extremo del gráfico, indicando que es un punto alejado de la muestra respecto al espacio de las X's (predictoras). Quiere decir que los valores de las variables predictoras para este caso son muy "diferentes" y alejadas de las que se dan en este conjunto de datos. Así pues, es un dato influyente a priori, ya que su presencia en el conjunto de datos podría inducir cambios importantes en los estimadores de los coeficientes. Sin embargo, su residuo estandarizado es muy pequeño (0.02878), lo que implica que esta observación está bien explicada por el modelo. Como resultado, su distancia de Cook es muy pequeña (0.00011) por lo que se concluye que no es una observación influyente a posteriori y que no introduce cambios importantes en el proceso de estimación. Pese a ser un dato potencialmente influyente (leverage alto) finalmente no se puede considerar que sea realmente influyente (distancia de Cook baja debido a su pequeño residuo). No es necesario eliminarla del conjunto de datos, ya que el modelo estimado con y sin esta observación es prácticamente el mismo y posee las mismas propiedades.