GRAU INFORMÀTICA (FIB- UPC)

CURS 2013-2014 Q1 – QUIZ : STATISTICAL MODELING

(Data: 20 de Desembre del 2013 a les 13:00h Aula C6S306)

Nom de l´alumne:	
Professor:	Lídia Montero
Localització:	C5 – 217 – Campus Nord UPC
Normativa de l'examen:	ÉS POT DUR APUNTS TEORIA, CALCULADORA I TAULES ESTADÍSTIQUES
Durada de l'examen:	1h 00 min
Sortida de notes:	Abans del 17 de Gener al Web Docent de MLGz
Revisió de l'examen:	17 de Gener 2014 a 16:00 h a C5-217 Campus Nord

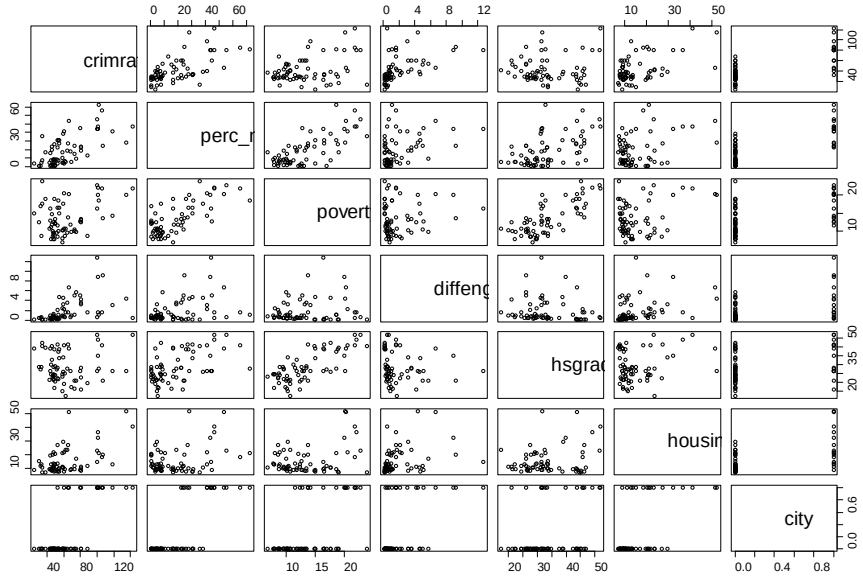
Problema 1 (10 punts): Tasa de criminalidad en USA

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:

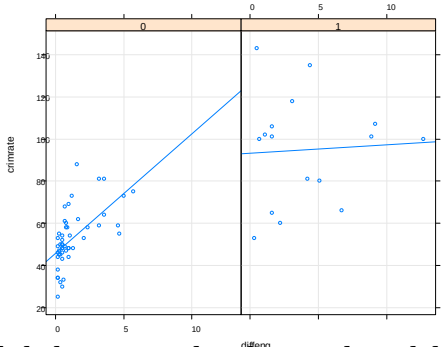
```
* area     = "Nombre del Área (áreas marcadas _R son resto del estado)" ;
* crimrate = "Tasa de crímenes mayores por 1000 habitantes" ;
* perc_min = "Porcentaje de minorías (población hispana o de color)" ;
* poverty  = "Porcentaje de pobreza" ;
* diffeng  = "Porcentaje que tienen problemas con el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

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



Pregunta 1 (5 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. Summary del modelo:




```
             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
```

- Determinar qué porcentaje de la variabilidad en la tasa de la criminalidad viene explicado por el modelo.

Pide el coeficiente de determinación del modelo que es del 61.13%, ésa es la variabilidad de la respuesta explicada por las 2 variables incluídas en el modelo.

- Escribid la/las ecuación/nes del modelo lineal estimado.

Si estado (ref 0): crimrate= 45.685 + 5.692 diffeng

Si ciudad (1): crimrate= (45.685 + 47.652)+ (5.692-5.307) diffeng = 93.337 + 0.385 diffemg

- Interpretad el efecto de la variable dificultades con el inglés en la respuesta.

Si estado (ref 0) el incremento de un 1% en las dificultades con el inglés supone un incremento de la tasa de criminalidad por mil hab de 5.692 unidades.

Si ciudad (1) el incremento de un 1% en las dificultades con el inglés supone un incremento de la tasa de criminalidad por mil hab de 0.385 unidades.

- Suponiendo validado el modelo, ¿la relación entre el porcentaje de población con dificultades con el inglés y la tasa de criminalidad, depende de que el área sea ciudad o estados?

Pide si la interacción es estadísticamente significativa y por lo que se puede ver en el listado la HO de que no lo es, se rechaza con un p valor de 0.0084, por tanto la relación entre la variable dificultades con el inglés y la respuesta crimrate SI depende de si el área considerada es estado o ciudad.

- Determinar cuál sería la tasa de criminalidad por 1000 habitantes en la ciudad de Miami si el porcentaje de población con dificultades con el inglés es del 10%.

Si ciudad - Miami (1): crimrate= (45.685 + 47.652)+ (5.692-5.307) 10 = 97.187

Pregunta 2 (2 puntos)

Se ajusta un modelo con algunas variables, obteniendo la tabla de los coeficientes (método summary).

```
> 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
```

- Indicad la significación individual y conjunta de las variables introducidas.

El test global de regresión indica que el modelo actual es significativamente distinto del modelo nulo (p valor test global 10^{-14}) y cada uno de los efectos principales incluídos en el modelo son estadísticamente significativos a juzgar por los p valores de los test individuales de coeficiente igual a 0 que se incluye en la salida estándar de R.

- Escribid las ecuaciones del modelo.

Si estado (0): crimrate= (86.92 -16.185) + 0.78 perc_min -1.11 hsgrad + 0.67 housing

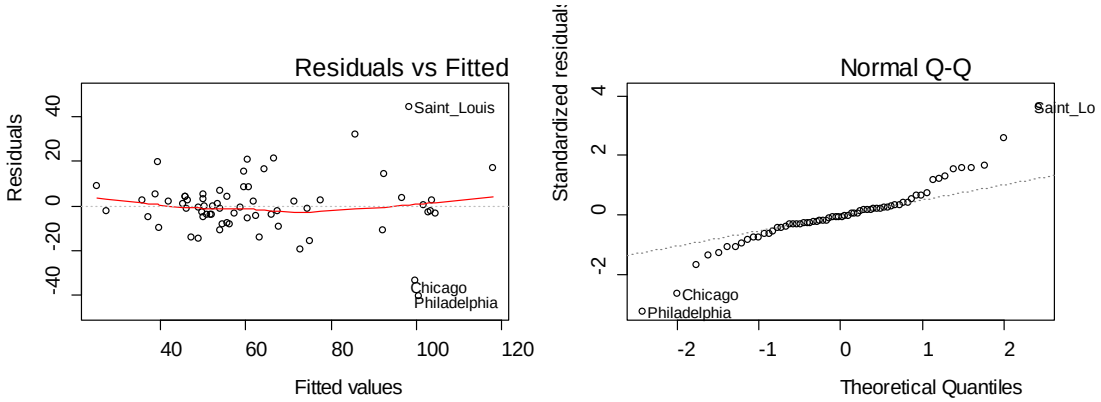
Si ciudad (ref 1): crimrate = 86.92 + 0.78 perc_min -1.11 hsgrad + 0.67 housing

Pregunta 3 (3 punts): Validació

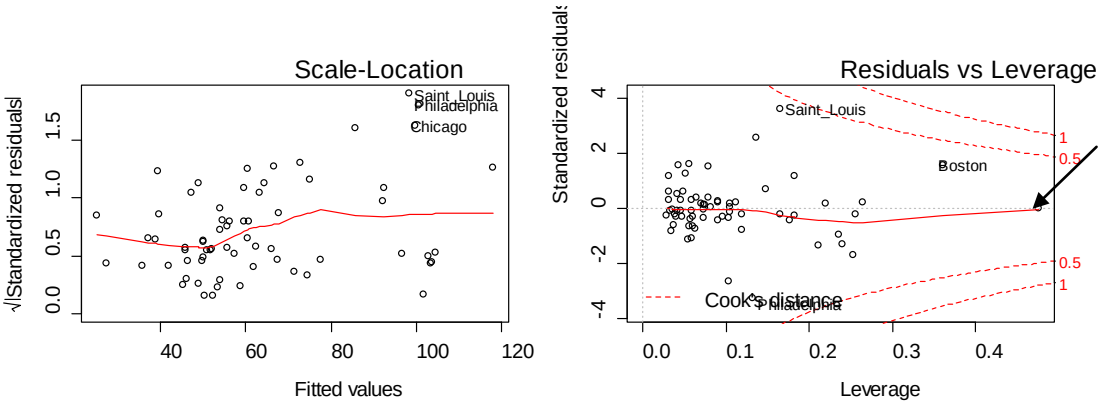
Aplicando el mecanismo stepwise usando el criterio BIC e incluyendo todas las variables y las interacciones entre las variables númericas y la binaria (city) se obtiene el siguiente modelo:

```
Call:
lm(formula = crimrate ~ perc_min + poverty + hsgrad + city +
    perc_min:city + poverty:city, data = dades2)

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
```



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

- 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 son 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 elevado el 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. La cota de Cook de 4/(n-p)=4/59=0.068 indicaría que son influyentes las observaciones de Mississippi, NewMéxico, Boston, Chicago, Philadelphia y Sant Louis.

- Indicad las observaciones que presentan un gran desajuste en este modelo y el criterio estadístico considerado?

El tamaño de la muestra sugiere que un valor absoluto de residuo estudentizado superior a 2 indica desajuste entre la observación y la predicción facilitada por el modelo. Esto sucede en las ciudades de Chicago, Philadelphia, Dallas y Sant Louis.

- 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 (0.0011) 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.