Estadística: Modelos Lineales

Examen Final

Junio de 2003

CUESTIONES

1. Se ha regresado la tasa de criminalidad (`y`) en una muestra de Estados de EE.UU. sobre variables como `M` (porcentaje de población formado por varones entre 14 y 24 años), `Ed`, `Po1` (gasto en Policía), `Ineq` (índice de desigualdad de la renta) y `Prob` (probabilidad estimada de que un delincuente sea capturado y no absuelto en juicio). Los resultados resumidos pueden verse en el estadillo a continuación:

```
> summary(a)

Call:
lm(formula = y ~ M + Ed + Po1 + Ineq + Prob,
                 data = UScrime)

Residuals:
     Min       1Q   Median       3Q      Max
-528.242  -74.000   -6.999  139.760  503.338

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -4064.574    816.279  -4.979 1.20e-05 ***
M               7.969      3.262   2.443 0.018964 *
Ed             16.015      4.342   3.688 0.000656 ***
Po1            12.123      1.406   8.621 9.47e-11 ***
Ineq            6.831      1.456   4.692 3.00e-05 ***
Prob        -3867.271   1596.552  -2.422 0.019930 *
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05

Residual standard error: 209.7 on 41 degrees of freedom
Multiple R-Squared: 0.7379,     Adjusted R-squared: 0.706
F-statistic: 23.09 on 5 and 41 DF,  p-value: 5.926e-11
```

Por su parte, la Figura 1 muestra gráficamente algunos resultados de interés. A la luz de todo lo anterior, responde a lo siguiente:

a) ¿Qué variables, si es que alguna, son significativas al nivel $\alpha = 0{,}05$?

b) ¿Cual es el tamaño de la muestra?

c) ¿Se rechazaría, al nivel de significación $\alpha = 10^{-5}$ la hipótesis nula “Las variables predictoras en su conjunto no explican en absoluto los valores de la variable respuesta”?

d) ¿Hay evidencia de multicolinealidad entre las variables predictoras?

e) ¿Cuál es el valor de SSE?

f) ¿Cuál es el valor de SST?

g) ¿Hay evidencia que apunte a alguna observación acusadamente distinta de las demás? Explica como llegas a tu conclusión.

h) ¿Hay evidencia de acusada no normalidad en las perturbaciones? ¿Como afecta la no normalidad (explica qué resultados y cuáles de las respuestas que hayas dado más arriba quedarían comprometidas por la falta de normalidad en las perturbaciones.

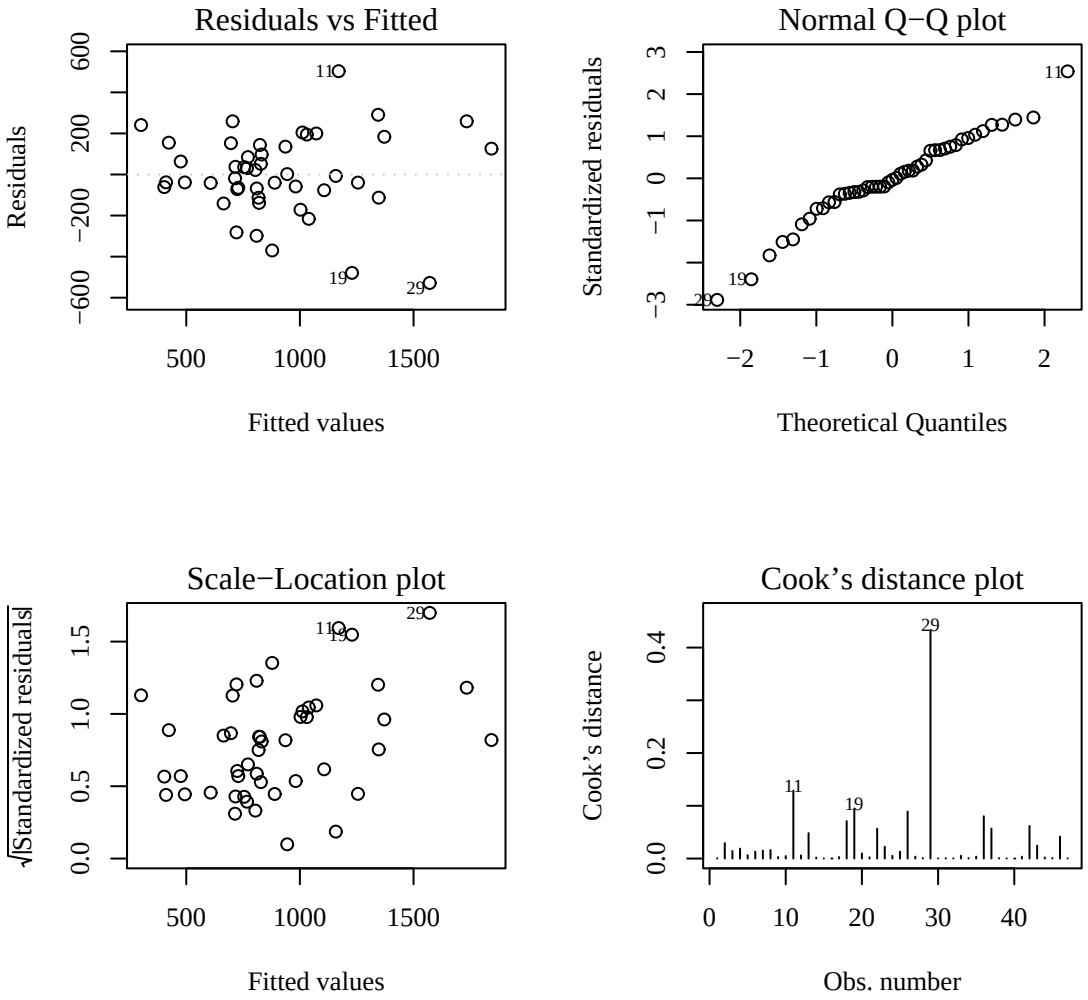
i) ¿Hay evidencia de acusada heteroscedasticidad en los residuos? ¿Qué efecto tendría, si fuera el caso, sobre: i) La insesgadez de las estimaciones de los parámetros, y ii) Sobre su eficiencia?

j) Decidimos investigar si el desempleo influye en la tasa de criminalidad. Introducimos para ello dos nuevas variables en nuestro modelo, `U1` (desempleo entre 14 y 24 años) y `U2` (desempleo entre 35 y 39 años), en ambos casos para varones. El nuevo modelo que obtenemos es:

```
Call:
lm(formula = y ~ M + Ed + Po1 + Ineq + Prob + U1 + U2,
             data = UScrime)

Residuals:
```

Figura 1: Criminalidad en relación con otras variables socio-económicas.
lm(formula = y ~ M + Ed + Po1 + Ineq + Prob, data = UScrime)



```
     Min       1Q   Median       3Q      Max
-520.756 -105.668    9.532  136.281  519.365

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -5095.552    896.895  -5.681 1.43e-06 ***
M              10.678      3.318   3.218  0.00260 **
Ed             21.845      4.833   4.520 5.62e-05 ***
Po1            10.596      1.572   6.738 4.91e-08 ***
Ineq            6.633      1.392   4.767 2.61e-05 ***
Prob        -3730.849   1522.207  -2.451  0.01883 *
U1             -3.542      3.022  -1.172  0.24822
U2             15.882      7.189   2.209  0.03310 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1

Residual standard error: 199.8 on 39 degrees of freedom
Multiple R-Squared: 0.7738,     Adjusted R-squared: 0.7332
F-statistic: 19.06 on 7 and 39 DF,  p-value: 8.805e-11
```

Contrasta al nivel de significación $\alpha = 0{,}05$ la hipótesis de que los parámetros que afectan a las dos variables `U1` y `U2` son simultáneamente cero.

k) Observa el estadillo siguiente. ¿Cuál sería, de acuerdo con el criterio AIC, el modelo que escogerías de entre los comparados?

```
Start:  AIC= 508.08
 y ~ M + Ed + Po1 + Ineq + Prob

       Df Sum of Sq     RSS     AIC
<none>              1803290     508
- Prob  1    258063 2061353     512
- M     1    262486 2065776     512
- Ed    1    598315 2401605     520
- Ineq  1    968199 2771489     526
- Po1   1   3268577 5071868     555

Call:
lm(formula = y ~ M + Ed + Po1 + Ineq + Prob, data = UScrime)

Coefficients:
(Intercept)            M           Ed          Po1         Ineq
  -4064.574        7.969       16.015       12.123        6.831
```

2. Obtén razonadamente la tabla de descomposición de la suma de cuadrados en un modelo ANOVA con dos tratamientos replicados sin interacción.

3. Michelson realizó una serie de mediciones de la velocidad de la luz en cinco días consecutivos durante 1879. Los resultados, después de restar 299000 Km/h, están recogidos en la variable `Speed` y la variable `Expt` codifica el día en que se verificaron las mediciones (20 en cada jornada). La Figura 2 recoge gráficamente el resultado de todas las mediciones, en que parecen apreciarse diferencias de unos días a otros. Como ayuda para dilucidar si, realmente, hubo diferencias de unos días a otros (por el calibrado de los instrumentos, por alteraciones atmosféricas que hicieran variar la velocidad de la luz...), se ha realizado el análisis de varianza a continuación. Obtén las conclusiones que quepa obtener de los resultados que se te proporcionan.

```
Call:
   aov(formula = Speed ~ Expt, data = michelson)

Terms:
                   Expt Residuals
Sum of Squares    94514    523510
Deg. of Freedom       4        95

Residual standard error: 74.23363
```

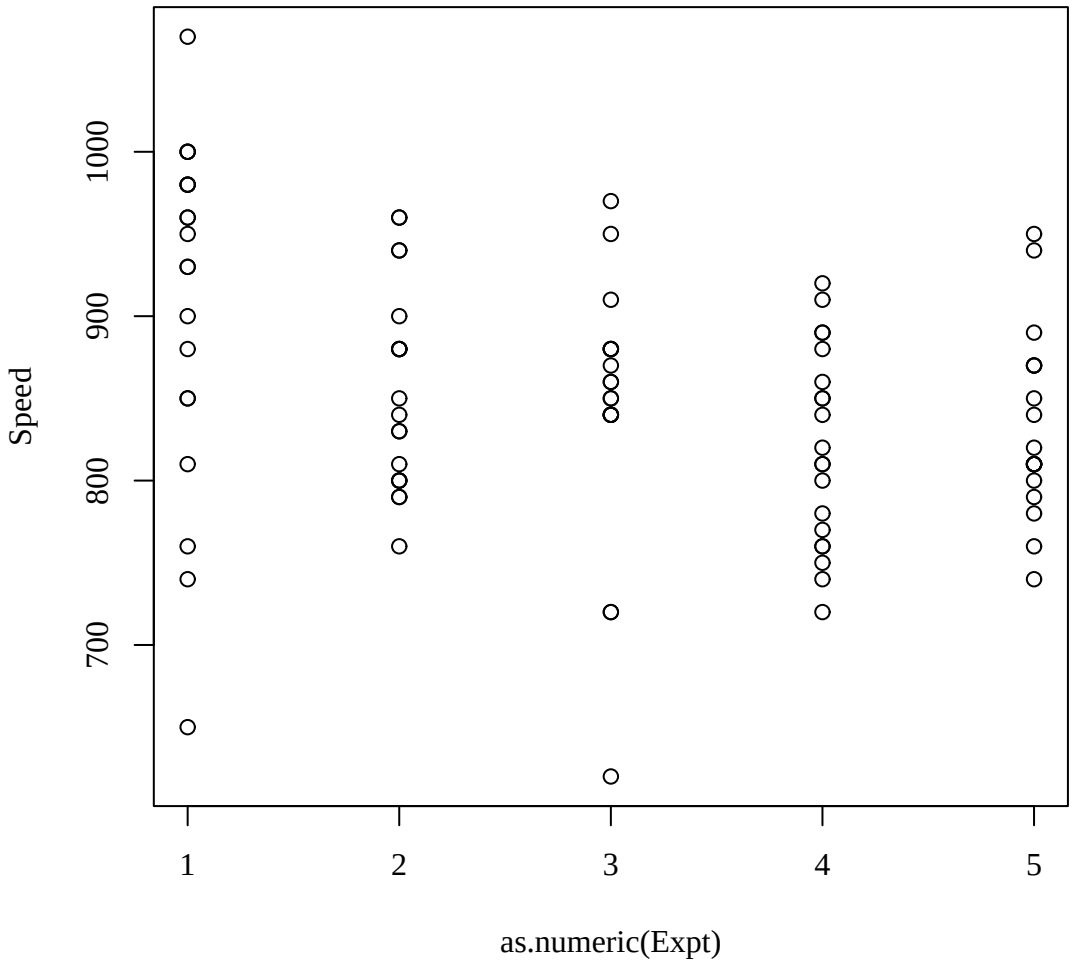


Figura 2: Mediciones de Michelson de la velocidad de la luz