**MÀSTER DE LOGÍSTICA, TRANSPORT I MOBILITAT (UPC).**

**CURS 11-12 Q1** – **TEST de LABORATORI 1**

**Mètodes de Captació, Anàlisi i Interpretació de Dades .**

**(Data: 24/11/2011 18:00-20:00 h Lloc: Aula H1.3)**

**Professor responsable**: Lídia Montero Mercadé

**Localització:** Edifici C5 D217

**Normativa de l’examen:** AMB APUNTS TEORIA, COMANDES DE R I TAULES ESTADÍSTIQUES SENSE ANOTACIONS MANUSCRITES

## ES POT DUR CALCULADORA

**Durada del test:** 2h 00 min

**Sortida de notes**: Abans 14 /12/11 al WEB de l’assignatura.

**Revisió**: El 14 /12/110 a les 14:30 hores (C5-217).

**Puntuació sobre 24 – 1 Punt per Apartat**

The data comes from a unpublished master's paper by Carl Hoffstedt. They relate the automobile accident rate, in accidents per million vehicle miles to several potential terms. The data include 39 sections of large highways in the state of Minnesota in 1973. The goal of this analysis was to understand the impact of design variables, acpt, slim, sigs1, and shld that are under the control of the highway department, on accidents. This data frame contains the following columns:

**rate** 1973 accident rate per million vehicle miles

**len**  length of the Highway1 segment in miles

**ADT** average daily traffic count in thousands

**trks** truck volume as a percent of the total volume

**sigs1** (number of signalized interchanges per mile times len + 1)/len, the number of signals per mile of roadway, adjusted to have no zero values.

**slim** speed limit in 1973

**shld** width in feet of outer shoulder on the roadway

**lane** total number of lanes of traffic

**acpt** number of access points per mile

**itg** number of freeway-type interchanges per mile

**lwid** lane width, in feet

**hwy** An indicator of the type of roadway or the source of funding for the road, either MC, FAI, PA, or MA

### References

Fox, J. and Weisberg, S. (2011) An R Companion to Applied Regression, Second Edition, Sage.

Weisberg, S. (2005) Applied Linear Regression, Third Edition. Wiley, Section 7.2.

> summary(high)

rate len ADT trks sigs1 slim

Min. :1.610 Min. : 2.960 Min. : 1.00 Min. : 6.000 Min. :0.04545 Min. :40

1st Qu.:2.630 1st Qu.: 7.995 1st Qu.: 5.00 1st Qu.: 8.000 1st Qu.:0.08738 1st Qu.:50

Median :3.050 Median :11.390 Median :13.00 Median : 9.000 Median :0.17666 Median :55

Mean :3.933 Mean :12.884 Mean :19.62 Mean : 9.333 Mean :0.51072 Mean :55

3rd Qu.:4.595 3rd Qu.:17.800 3rd Qu.:24.00 3rd Qu.:11.000 3rd Qu.:0.71515 3rd Qu.:60

Max. :9.230 Max. :40.090 Max. :73.00 Max. :15.000 Max. :2.78933 Max. :70

shld lane acpt itg lwid hwy

Min. : 1.000 Min. :2.000 Min. : 2.20 Min. :0.0000 Min. :10.00 FAI: 5

1st Qu.: 4.000 1st Qu.:2.000 1st Qu.: 6.95 1st Qu.:0.0000 1st Qu.:12.00 MA :13

Median : 8.000 Median :2.000 Median :10.30 Median :0.1300 Median :12.00 MC : 2

Mean : 6.872 Mean :3.128 Mean :12.16 Mean :0.2964 Mean :11.95 PA :19

3rd Qu.:10.000 3rd Qu.:4.000 3rd Qu.:14.60 3rd Qu.:0.3600 3rd Qu.:12.00

Max. :10.000 Max. :8.000 Max. :53.00 Max. :1.5400 Max. :13.00

>

**La variable de respuesta es *rate*.**

1. Determinar si la variable de respuesta (accidentes) tiene una distribución aceptablemente normal.
2. Valorad la estadística descriptiva de las variables explicativas numéricas presente en el espacio de trabajo.
3. Indicad por exploración de los datos cuales son aparentemente las variables más asociadas con la variable de respuesta (emplead únicamente las variables indicadas).
4. Indicad a nivel marginal, una a una, cuáles de las covariables disponibles tienen una correlación lineal significativamente distinta de cero.
5. Calculad e interpretar la explicabilidad del modelo anova de una vía que explica la TASA DE ACCIDENTES POR MILLA y ADT según el tipo de carretera (HWY).
6. Pensáis que la variabilidad de la respuesta POR MILLA y ADT depende del tipo de carretera?
7. Pensáis que la tasa media de accidentes POR MILLA y ADT depende del tipo de carretera?
8. Calculad el modelo de regresión lineal que explica la respuesta a partir de la longitud en millas del segmento, el aforo diario medio, el porcentaje de camiones y la tasa de señalización por milla. Interpretad el modelo de regresión y valorad la aportación de cada una de sus variables.
9. Cuál es el porcentaje de la variabilidad del número de accidentes que viene explicada por la longitud del segmento una vez que el aforo medio diario, el porcentaje de camiones y la tasa de señalización por milla ya están incluídos en el modelo?
10. Valorad a grandes rasgos la calidad del modelo en función de la primera impresión de la diagnosis de sus residuos.
11. Aplicad alguna transformación a la variable de respuesta que mejore las propiedades del modelo de regresión lineal.
12. Estudiad alguna transformación de las variables longitud del segmento, intensidad media diaria y porcentaje de camiones que mejore la calidad del modelo de regresión lineal para explicar la respuesta a partir de las variables explicativas len, ADT, trks y sigs1.
13. Después de controlar por la longitud del trayecto, indicad si el efecto aditivo del tipo de carretera es estadísticamente significativo.
14. Contruir el modelo que relaciona el logaritmo de la respuesta con los términos logarítmicos de la longitud, la intensidad media diaria, el porcentaje de camiones y la tasa de señalización por milla.Valorar el modelo respecto las propuestas anteriormente estudiadas.
15. En el modelo que relaciona el logaritmo de la respuesta con los términos logarítmicos de la longitud, la intensidad media diaria, el porcentaje de camiones y la tasa de señalización por milla son significativamente necesarias en algúno de los casos la inclusión simultánea de efectos lineales y cuadráticos para mejorar la explicabilidad del modelo?.
16. Argumentar si el efecto neto del tipo de carretera es significativo en el modelo que relaciona el logaritmo de la respuesta con los términos logarítmicos de la longitud, la intensidad media diaria, el porcentaje de camiones y la tasa de señalización por milla.

**Se construye un nuevo modelo en escala logarítmica de la respuesta con los términos logarítmicos de la longitud, la intensidad media diaria, el porcentaje de camiones y la tasa de señalización por milla y el resto de las variables numéricas disponibles (sin transformarlas).**

1. Determinar qué variables son estadísticamente significativas en el modelo descrito. Valorad la calidad del modelo.
2. Determinar si en el modelo resultante del punto anterior, el efecto aditivo del tipo de carretera merece la pena incluirlo en el modelo.
3. Determinar si en el modelo resultante del **punto 17**, la relación entre la tasa de accidentes (respuesta) y el límite de velocidad (slim) depende del tipo de carretera.
4. Valorad la presencia de outliers en los residuos estudentizados a un nivel de confianza del 99%. Indicad cuáles son esas observaciones **en el modelo resultante del Punto 19**.
5. Estudiad la presencia de valores influyentes a priori, indicando su número según el criterio estudiado en clase **en el modelo resultante del Punto 19**.
6. Estudiad la presencia de valores influyentes a posteriori, indicando el criterio estudiado en clase y las observaciones realmente atípicas **en el modelo resultante del Punto 19**.
7. Dado un segmento con valores de las variables explicativas en la mediana, cuál sería la tasa de accidentes esperada con un intervalo de confianza del 95% para cada tipo de carretera ?
8. Valorad lo que habéis aprendido trabajando con este interesante juego de datos reales.