Graduado en Informática

Prácticas de Estadística

Práctica 1: Estadística Descriptiva 1 variable.

Ejercicio 1:

El fichero 'datospr1.mat' contiene los valores medidos para la variable estadística X. Queremos detectar aquellos valores anómalos ('outliers') de X. Una de las técnicas empleadas es calcular Q_1, Me y Q_3 y considerar anómalos aquellos que sean menores de la mediana en $3(Me - Q_1)$, o mayores que esta en $3(Q_3 - Me)$.

1. *Encontrar los valores anómalos de X según la regla antes indicada.*
2. *Crear una nueva variable Y donde no se encuentren los valores anómalos.*
3. *Calcular media y varianza de los datos X y de los Y.*

Solución:

Debemos cargar la variable unidimensional X en Matlab (Scilab o similar), para ello debemos copiar el fichero en el directorio de trabajo de Matlab y ejecutar:

clear all, format long % Limpia de variables preexistentes y muestra todas las cifras.
load datospr1 % Carga los datos desde el fichero.
X % Comprobamos que ya existe la variable.
N=length(X) % Vemos su dimension.

Para calcular Q_1, Me y Q_3 hacemos:

XX=sort(X) % Ordenamos la variable, creando la variable ordenada XX.
n1=N*1/4, n2=N*1/2, n3=N*3/4 % Calculamos los lugares que deben ocupar los 3 cuartiles.
nn1=floor(n1), nn2=floor(n2), nn3=floor(n3) % Calculamos su parte entera.
% Si es entero saco la media, si no lo es tomo el siguiente.
% a==floor(a), es 'cierto' si a es entero y 'falso' si no lo es.
if n1==nn1, Q1=(XX(nn1)+XX(nn1+1))/2, else Q1=XX(nn1+1), end
if n2==nn2, Me=(XX(nn2)+XX(nn2+1))/2, else Me=XX(nn2+1), end
if n3==nn3, Q3=(XX(nn3)+XX(nn3+1))/2, else Q3=XX(nn3+1), end

Hallamos los límites:

Lmin=Me-3*(Me-Q1) % Limite inferior
Lmax=Me+3*(Q3-Me) % Limite superior

Hallamos los elementos fuera de [Lmin, Lmax]. Usamos los símbolos '|' y '&' para el 'or' y 'and' lógicos en Matlab.

La = **find**((**X** > **Lmax**)|(**X** < **Lmin**)) **% Lista elementos extraños**
L = **find**((**X** <= **Lmax**)&(**X** >= **Lmin**)) **% Lista elementos no extraños**
Y=X(L) % En Y están los no extraños

Calculamos media y varianza de X:

Sx=sum(X), mediaX=Sx/N
Sx2=sum(X.^2), m2X=Sx2/N
VarX=m2X-mediaX^2

Calculamos media y varianza de Y:

Ny=length(Y)
Sy=sum(Y), mediaY=Sy/Ny
S2Y=sum(Y.^2), m2Y=S2Y/Ny
VarY=m2Y-mediaY^2

Ejercicio 2:

El fichero 'TempMalaga.mat' obtenido a partir del suministrado por la Agencia Estatal de Meteorología (AEMET), muestra las temperaturas máximas, mínimas y media obtenidas diariamente en Málaga desde el año 2000, cada fila (registro del fichero consta de 6 columnas que indican:

Columnas 1, 2 y 3: Año, Mes y Día, respectivamente

Columnas 4, 5 y 6: Temperatura máxima, mínima y media en grados centígrados, respectivamente.

Calcular:

1. *La tabla de frecuencias absolutas de las temperaturas mínimas del mes de Febrero con las clases: 'Menor o igual a 3 grados', (3,6], (6,8], (8,10], (10,13], 'Más de 13 grados'.*

2. *A partir de la tabla, hallar:*

   (a) Q_1, Me, Q_3 *y Moda.*

   (b) *Media, varianza, sesgo y curtosis.*

   (c) *Desviación media y la media cuadrática de las desviaciones.*

3. *Repetir los cálculos del apartado 2 (excepto la moda) con los datos iniciales (sin agrupar).*

Solución:

Apartado 1: Cargamos el fichero en Matlab mediante:

clear all % Limpia de variables preexistentes.
load TempMalaga % Carga los datos desde el fichero.
whos % Muestra los nombres de las variables existentes.

Observamos que solo existe una variable llamada 'TMA' con 4383 filas y 6 columnas a la que le vamos a filtrar los elementos correspondientes al mes de febrero (segunda columna con un '2':

L=find(TMA(:,2)==2) % Nos da las filas (posiciones) de los registros de febrero.
X=TMA(L,5) % Crea la variable X con las temperaturas mínimas de los registros de febrero.
N=length(X) % Número de elementos tras el filtro.

Calculamos las frecuencias absolutas de cada clase:

```
n=zeros(1,6); % Pues hay 6 modalidades
for k=1:N,
   if X(k)<= 3, n(1)=n(1)+1;
   elseif X(k)<= 6, n(2)=n(2)+1;
   elseif X(k)<= 8, n(3)=n(3)+1;
   elseif X(k)<= 10, n(4)=n(4)+1;
   elseif X(k)<= 13, n(5)=n(5)+1;
   else n(6)=n(6)+1;
   end
end, n % Mostramos al final las frecuencias absolutas.
```

Apartado 2-a: Tomo las marcas de clase 'x' y las alturas del histograma 'h':

```
Li=[0,3,6,8,10,13,16], x=[1.5 4.5 7 9 11.5 14.5]
a=diff(Li), h=n./a % Amplitud y altura.
```

Calculamos la moda:

```
[val,ind]=max(h) % ind nos indica el intervalo modal.
h1=h(ind)-h(ind-1), h2=h(ind)-h(ind+1) % NOTA: val=h(ind)
Mo=Li(ind)+a(ind)*h1/(h1+h2)
```

NOTA: El método no funcionaría si ind=1 ó ind=6(última clase) pues no existe h(0) ni h(7).

Calculamos Q_1, Mediana y Q_3:

```
n1=N/4, n2=N/2, n3=3*N/4
nac=cumsum(n) % frecuencias acumuladas
L=find(n1<nac),ind=L(1);
Q1=Li(ind)+a(ind)*(n1-nac(ind-1))/n(ind)
L=find(n2<nac),ind=L(1);
Me=Li(ind)+a(ind)*(n2-nac(ind-1))/n(ind)
L=find(n3<nac),ind=L(1);
Q3=Li(ind)+a(ind)*(n3-nac(ind-1))/n(ind)
```

Apartado 2-b: Para hallar la media, varianza, sesgo y curtosis:

```
m1=sum(n.*x)/N, m2=sum(n.*x.^2)/N, m3=sum(n.*x.^3)/N, m4=sum(n.*x.^4)/N
mu3=m3-3*m2*m1+2*m1^3, mu4=m4-4*m3*m1+6*m2*m1^2-3*m1^4
```

media=m1, var=m2-m1^2, s=sqrt(var), sesgo=mu3/s^3, curtosis=mu4/s^4-3

Apartado 2-c: Para hallar la desviación media y la media cuadrática de las desviaciones:
d=x-m1, DM=sum(n.*abs(d))/N, MCdesv=sqrt(sum(n.*d.^2)/N)

Apartado 3: Repetimos para los datos sin agrupar:
Calculamos Q_1, Mediana y Q_3:
nn1=floor(n1), nn2=floor(n2), nn3=floor(n3) % n1, n2, n3 fueron calculados antes.
XX=sort(X); % Ordenamos los datos
if n1==nn1, Q1X=(XX(nn1)+XX(nn1+1))/2, else Q1X=XX(nn1+1),end
if n2==nn2, MeX=(XX(nn2)+XX(nn2+1))/2, else MeX=XX(nn2+1),end
if n3==nn3, Q3X=(XX(nn3)+XX(nn3+1))/2, else Q3X=XX(nn3+1),end
Para hallar la media, varianza, sesgo y curtosis:
m1X=sum(X)/N, m2X=sum(X.^2)/N, m3X=sum(X.^3)/N, m4X=sum(X.^4)/N
mu3X=m3X-3*m2X*m1X+2*m1X^3
mu4X=m4X-4*m3X*m1X+6*m2X*m1X^2-3*m1X^4
mediaX=m1X, varX=m2X-m1X^2
sX=sqrt(varX), sesgo=mu3X/sX^3, curtosis=mu4X/sX^4-3
Para hallar la desviación media y la media cuadrática de las desviaciones:
dX=X-m1X, DMX=sum(abs(dX))/N, MCdesvX=sqrt(sum(dX.^2)/N)

Alternativa al apartado 3:
Matlab cuenta con las funciones estadísticas: **mean, median, std, min, max, moment, skewness, kurtosis, mode y cov** que nos permiten simplificar los cálculos:
MeX=median(X) % Sin necesidad de ordenar.
m1X=moment(X,1), m2X=moment(X,2), m3X=moment(X,4), m4X=moment(X,4)
mediaX=mean(X), varX=var(X), sesgoX=skewness(X), curtosisX=kurtosis(X)-3

Prácticas de Estadística

Práctica 2: Estadística Descriptiva 2 variables.

Ejercicio 3:

El fichero 'Hamb.m' contiene las medidas obtenidas en un estudio sobre hamburguesas. Cada análisis (registro) tiene 3 campos:

- Tipo: Con los valores { Ternera, Resto de carnes, Ave} codificadas respectivamente como 1, 2 y 3.
- Calorías de la hamburguesa.
- Sodio: Miligramos de sodio en la hamburguesa.

Hallar:

1. Las calorías medias y varianzas para cada tipo de hamburguesa.
2. Lo mismo pero para el contenido en sodio.
3. Representar los resultados medios gráficamente.
4. Ajustar la recta de regresión de Y=calorías respecto a X=sodio.
5. Hallar varianza residual y coeficiente de correlación lineal del ajuste Y/X.
6. Ajustar la recta de regresión de X=sodio respecto a Y=calorías.
7. Hallar varianza residual y coeficiente de correlación lineal del ajuste X/Y.
8. Representar conjuntamente la nube de puntos (con 'x') y ambas rectas.
9. Representar conjuntamente la nube de puntos (con 'x') y las rectas, pero ahora los puntos correspondientes a hamburguesas de ternera en rojo (r), las de carne en general en azul (b) y las de ave en magenta (m)'.
10. Ajustar un polinomio de tercer grado Y/X. Hallar la varianza residual, suma de los cuadrados de los errores y coeficiente de determinación.
11. Representar conjuntamente los puntos (con 'x'), la recta de Y/X y el polinomio anterior.

SOLUCIÓN:
El programa MATLAB puede quedar como:

```
clear all, format compact,clf          %  1 Borro variables y graficas.
Hamb,whos                              %  2 Cargo los datos y miro que variables hay
CC                                     %  3 Visualizo la unica variable existente
Y=CC(:,2);                             %  4 La segunda columna es la Y (sodio)
X=CC(:,3);N=length(X)                  %  5 La tercera es la X y N (numero de registros).
L1=find(CC(:,1)==1);N1=length(L1)      %  6 Lugares clase 1 y cuantos hay
L2=find(CC(:,1)==2);N2=length(L2)      %  7 Lo mismo para clase 2
L3=find(CC(:,1)==3);N3=length(L3)      %  8 Lo mismo para clase 3
Y1=Y(L1),Y2=Y(L2),Y3=Y(L3)             %  9 Variable Y para cada clase
X1=X(L1),X2=X(L2),X3=X(L3)             % 10 Variable X para cada clase

mY1=sum(Y1)/N1,%mY1b=mean(Y1)          % 11 media de 1 (ternera) de dos formas. (sodio)
mY2=sum(Y2)/N2,%mY2b=mean(Y3)          % 12 media de 2 (otra carne).
mY3=sum(Y3)/N3,%mY3b=mean(Y3)          % 13 media de 3 (ave).
mX1=sum(X1)/N1,%mX1b=mean(X1)          % 14 media de 1 (ternera) de dos formas. (calorias)
mX2=sum(X2)/N2,%mX2b=mean(X2)          % 15 media de 2 (otra carne) (calorias)
mX3=sum(X3)/N3,%mX3b=mean(X3)          % 16 media de 3 (ave) (calorias)
vY1=sum(Y1.^2)/N1-mY1^2,%vY1b=var(Y1)*(N1-1)/N1 % 17 VARIANZAS sodio de 2 formas
vY2=sum(Y2.^2)/N2-mY2^2,%vY2b=var(Y2)*(N2-1)/N2 % 18    "
vY3=sum(Y3.^2)/N3-mY3^2,%vY3b=var(Y3)*(N3-1)/N3 % 19    "
vX1=sum(X1.^2)/N1-mX1^2,%vX1b=var(X1)*(N1-1)/N1 % 20    "   ahora para calorias
vX2=sum(X2.^2)/N2-mX2^2,%vX2b=var(X2)*(N2-1)/N2 % 21    "
```

```
vX3=sum(X3.^2)/N3-mX3^2,%vX3b=var(X3)*(N3-1)/N3 % 22    "

disp('Apartado 3')                                % 23 Saca el mensaje
line([0.6,3.4],[0,0])                             % 24 Dibuja una recta entre (0.6,0) y (3.4,0)
line([1,1],[0,mY1],'Linewidth',8,'Color','r')     % 25 Linea en rojo y ancho 8 de (1,0) a (1,mY1)
line([2,2],[0,mY2],'Linewidth',8,'Color','g')     % 26 Linea en verde, ancho 8 de (2,0) a (2,mY2)
line([3,3],[0,mY3],'Linewidth',8,'Color','b')     % 27 Linea en azul y ancho 8 de (3,0) a (3,mY3)
pause,clf                                         % 28 Pausa para ver el grafico y lo borra
line([0.6,3.4],[0,0])                             % 29 Lo mismo para el sodio
line([1,1],[0,mX1],'Linewidth',8,'Color','r')     % 30
line([2,2],[0,mX2],'Linewidth',8,'Color','g')     % 31
line([3,3],[0,mX3],'Linewidth',8,'Color','b')     % 32
pause,clf                                         % 33 Para salir del pause pulsar tecla

disp('Apartados 4 y 5')                           % 34
p=polyfit(X,Y,1)                                  % 35 Ajuste lineal por la via rapida
a=p(2), b=p(1)                                    % 36 Coeficientes de la recta Y/X
mX=mean(X),mY=mean(Y),                            % 37 Medias
vX=sum(X.^2)/N-mX^2,vY=sum(Y.^2)/N-mY^2           % 38 Varianzas
covar=sum(X.*Y)/N-mX*mY,                          % 39 Covarianza
r=covar/sqrt(vX*vY)                               % 40 Coeficiente correlación lineal
Vr=(1-r^2)*vY                                     % 41 Varianza residual caso lineal

disp('Apartados 6 y 7')                           % 42
p1=polyfit(Y,X,1),b1=p1(1),a1=p1(2)               % 43 Ajuste de la recta X/Y: X=a1+b1*Y
Vr1=(1-r^2)*vX                                    % 44 Varianza residual ajuste X/Y

disp('Apartado 8')                                % 45
xx=100:700;yy1=polyval(p,xx);                     % 46 Auxiliar para dibujar recta Y/X
yy=80:200;xx1=polyval(p1,yy);                     % 47    "       "     "      "   X/Y
plot(X,Y,'x',xx,yy1,xx1,yy),grid                  % 48 Dibuja todo. Grid pone la cuadrícula
pause,clf                                         % 49

disp('Apartado 9')                                % 50
plot(X1,Y1,'xr',X2,Y2,'xb',X3,Y3,'xm',xx,yy1,xx1,yy),grid % 51 Dibuja lo pedido
pause,clf                                         % 52

disp('Apartado 10')                               % 53
pp=polyfit(X,Y,3)                                 % 54 Ajusta el polinomio y=a+bx+cx^2+dx^3
ye=polyval(pp,X)                                  % 55 Calcula los y estimados
re=Y-ye                                           % 56 Calcula vector residuo o error.
Vre=var(re)*(N-1)/N                               % 57 Varianza residual.
SSE=sum(re.^2)                                    % 58 Suma de errores al cuadrado
R2=1-Vre/vY                                       % 59 Coeficiente de determinación

disp('Apartado 11')                               % 60
yy2=polyval(pp,xx);                               % 61 Auxiliar para dibujar polinomio
plot(X,Y,'X',xx,yy1,xx,yy2),grid                  % 62 Dibuja todo conjuntamente
```

Resultados obtenidos: Comentaremos los principales: Número de elementos totales N =54, del tipo 1 (ternera) hay N1=20, del tipo2 hay N2=17 y de ave hay N3=17.

Aptdo. 1: Las calorías medias para 'ternera' es $\mathbf{mY1 = 156.85}$ (significado 1.5685e+002), para 'otra carne' $\mathbf{mY2 = 158.7059}$ y para 'ave' vale $\mathbf{mY3 = 122.4706}$.

Las varianza para 'ternera' es $vY1 = 487.0275$, para 'otra carne' $vY2 = 599.3841$ y para 'ave' de $vY3 = 611.1903$

Aptdo. 2: El contenido medio de sodio es para 'ternera' de $\mathbf{mX1 = 401.15}$, para 'carne general' de $\mathbf{mX2 = 418.5294}$ y para 'ave' de $\mathbf{mX3 = 459}$.

Las varianzas son respectivamente $\mathbf{vX1 = 9968.2275}$, $\mathbf{vX2 = 8293.6609}$ y $\mathbf{vX3 = 6758.3529}$.

Aptdo. 4: La recta obtenida es $\mathbf{Y = 0.1565X + 80.1073}$.

Aptdo. 5 La media de X (sodio) es $\mathbf{mX = 424.8333}$, la de Y (calorías) es $\mathbf{mY = 146.6111}$, la varianza de X vale $\mathbf{vX = 9018.2870}$ y la de Y $\mathbf{vY = 8298.3025}$, la covarianza es $\mathbf{cov = 1411.7315}$ y por último el coeficiente de correlación lineal $\mathbf{r = 0.5161}$ y la varianza residual $\mathbf{Vr = 608.8364}$.

Aptdo. 6: La recta obtenida es: $\mathbf{X = 1.7012Y + 175.4142}$. La varianza residual del ajuste X/Y es $Vr1 = 6616.6082$.

Notese que el coeficiente de correlación lineal es el mismo en el ajuste Y/X y X/Y, pero la varianza residual vale $(1 - r^2)V_Y$ en un caso y $(1 - r^2)V_X$ en el otro.

Aptdo. 10: El polinomio obtenido es (obtenido en pp con 'format long'):

$$\mathbf{Y = -0.0000011327908X^3 + 0.0013465122356X^2 - 0.3429711494881X + 136.9271160836836}$$

La varianza residual es: $\mathbf{Vre = 603.3198}$, la suma del cuadrado de los errores $\mathbf{SSE = 32579.2684}$ y el coeficiente de determinación vale $\mathbf{R2 = 0.2730}$.

Ejercicio 4:

El fichero 'car.m' contiene características de diversos modelos de coche. una de las variables contiene la marca, mientras la otra contiene una matriz con 5 columnas donde:

- Columna 1: **VOL:** Volumen interior en pies cúbicos. (ft^3)
- Columna 2: **HP:** Potencia del motor en CV. (Hp)
- Columna 3: **MPG:** Consumo medio en millas por galón. (mpg)
- Columna 4: **VM:** Velocidad máxima en millas por hora. (mph)
- Columna 5: **WT:** Peso del vehículo en libras dividido por 100. (100 lb)

Hallar:

1. Ajustar una recta que nos de el consumo en función del peso. Hallar r y Vr.
2. Ajustar un hiperplano que nos de el consumo en función de las restantes. Hallar el coeficiente de determinación.
3. ¿Existe alguna variable poco significativa? Es decir, que variables influyen mucho en el consumo y cuales no.
4. ¿Qué modelo presenta el consumo más bajo para sus características? ¿Cuál lo tiene más alto?
5. Imputación de datos: Supongamos que del modelo de coche XXXX conocemos los valores de $VOL = 70$, $HP = 84$, $VM = 100$ y $WT = 30$, pero el dato de MPG es desconocido (missing) y necesitamos usar un valor para él. Una primera aproximación es darle el de una medida de tendencia central de los valores MPG conocidos. Otra posibilidad es imputar su valor a partir del ajuste.
   (a) Imputar el valor medio.
   (b) Imputar el valor mediano.
   (c) Imputar el valor ajustado por el hiperplano.

Solución:

El programa MATLAB puede ser:

```
clear all,clf,clc,format compact                                  %  1 Preparacion
car                                                               %  2 Carga de variables
VOL=CC(:,1);HP=CC(:,2);MPG=CC(:,3);VM=CC(:,4);WT=CC(:,5);         %  3 Cambio nombres a variables
p=polyfit(WT,MPG,1)                                               %  4 Ajustamos recta MPG=a+b(WT)
N=length(VOL)                                                     %  5 Numero de registros
mWT=mean(WT),vWT=var(WT)*(N-1)/N                                  %  6 Media y varianza de WT
mMPG=mean(MPG),vMPG=var(MPG)*(N-1)/N                              %  7 Media y varianza de MPG
cov1=sum(MPG.*WT)/N-mMPG*mWT                                      %  8 Covarianza entre MPG y WT
%cov1b=cov(MPG,WT)*(N-1)/N                                        %  9 Otra forma de hallar covar.
r=cov1/sqrt(vMPG*vWT) % rb=corrcoef(WT,MPG)                       % 10 Coef. correlacion (2 formas)
Vr=(1-r^2)*vMPG                                                   % 11 Varianza residual caso recta
SSE=Vr*N                                                          % 12 Suma de errores al cuadrado
CD_lineal=r^2                                                     % 13 Coef. determinacion caso lineal
disp('Aptdo 2:')                                                  % 14
A=[N sum(VOL) sum(HP) sum(VM) sum(WT);                            % 15 Matriz para ajustar:
   sum(VOL) sum(VOL.^2) sum(HP.*VOL) sum(VM.*VOL) sum(WT.*VOL); % 16 VMG=A+B*VOL+C*HP+D*VM+E*WT
   sum(HP) sum(VOL.*HP) sum(HP.^2) sum(VM.*HP) sum(WT.*HP);   % 17 Ec1: <E,1>=0, Ec2: <E,VOL>=0,
   sum(VM) sum(VOL.*VM) sum(HP.*VM) sum(VM.^2) sum(WT.*VM);   % 18 Ec3: <E,HP>=0, Ec4: <E,VM>=0,
   sum(WT) sum(VOL.*WT) sum(HP.*WT) sum(VM.*WT) sum(WT.^2)]   % 19 Ec5: <E,WT>=0
B=[sum(MPG);sum(MPG.*VOL);sum(MPG.*HP);sum(MPG.*VM);sum(MPG.*WT)] % 20 Vector para el ajuste
sol=A\B                                                           % 21 Resolvemos el sistema lineal
MPGe=sol(1)+sol(2)*VOL+sol(3)*HP+sol(4)*VM+sol(5)*WT              % 22 MPG estimado por el ajuste
res=MPG-MPGe                                                      % 23 Vector residuo o error
Vres=var(res)*(N-1)/N                                             % 24 Varianza residual del ajuste
R2=1-Vres/vMPG                                                    % 25 Coeficiente de determinación
disp('Apartado 3:')                                               % 26
CORR=corrcoef([MPG,VOL,HP,VM,WT])                                 % 27 Matriz de coef. de corr. lineal
```

```
disp('Apartado 4:')                                                  % 28
[MAX,modmax]=max(res),modelMAX=MODEL(modmax,:)                       % 29 Modelo con maximo residuo pos.
[MIN,modmin]=min(res),modelMIN=MODEL(modmin,:)                       % 30 Modelo con máximo residuo neg.
disp('Apartado 5')                                                    % 31
ImpMedia=mMPG                                                         % 32 La media ya estaba calculada
ImpMediana=median(MPG)                                                % 33 Imputado por la mediana
ImpHiperpl=sol(1)+sol(2)*70+sol(3)*84+sol(4)*100+sol(5)*30 % 34 Imputado por el hiperplano
```

RESULTADOS:

Aptdo. 1: La recta es: $\mathbf{MPG} = -\mathbf{1.1122(WT)} + \mathbf{68.1655}$.

Hay $N = 82$ registros, la media de $mWT = 30.9146$ y la varianza es $vWT = 65.4744$. La media de MPG es $mMPG = 33.7817$ y la varianza es $vMPG = 98.8715$.

La covarianza es $cov1 = -72.8217$, elcoeficiente de correlación lineal vale $r = -0.9051$ y la varianza residual $Vr = 17.8781$.

La suma de los errores al cuadrado vale $SSE = 1466.0016$ y el coeficiente de determinación del caso lineal es $CD_l ineal = 0.8192$

Aptdo 2: Vamos a ajustar un hiperplano del tipo $MPG = A + B * VOL + C * HP + D * VM + E * WT$, es decir, hallar el vector generado por $B = \{\vec{1}, V\vec{O}L, H\vec{P}, V\vec{M}, W\vec{T}\}$ más próximo a $M\vec{P}G$. Plantando las condiciones: $< \vec{e}, \vec{1} >= 0$, $< \vec{e}, V\vec{O}L >= 0$, $< \vec{e}, H\vec{P} >= 0$, $< \vec{e}, V\vec{M} >= 0$, $< \vec{e}, W\vec{T} >= 0$, obtenemos el sistema a resolver:

$$\begin{pmatrix} N & \sum_i VOL_i & \sum_i HP_i & \sum_i VM_i & \sum_i WT_i \\ \sum_i VOL_i & \sum_i VOL_i^2 & \sum_i HP_i * VOL_i & \sum_i VM_i * VOL_i & \sum_i WT_i * VOL_i \\ \sum_i HP_i & \sum_i VOL_i * HP_i & \sum_i HP_i^2 & \sum_i VM_i * HP_i & \sum_i WT_i * HP_i \\ \sum_i VM_i & \sum_i VOL_i * VM_i & \sum_i HP_i * VM_i & \sum_i VM_i^2 & \sum_i WT_i * VM_i \\ \sum_i WT_i & \sum_i VOL_i * WT_i & \sum_i {*HP_i WT_i} & \sum_i VM_i * WT_i & \sum_i WT_i^2 \end{pmatrix} \begin{pmatrix} A \\ B \\ C \\ D \\ E \end{pmatrix} =$$

$$= \begin{pmatrix} \sum_i MPG_i \\ \sum_i MPG_i * VOL_i \\ \sum_i MPG_i * HP_i \\ \sum_i MPG_i * VM_i \\ \sum_i MPG_i * WT_i \end{pmatrix}$$

que al resolverlo nos proporciona el modelo:

$$VMG = 192.4378 - 0.01564 * VOL + 0.3922 * HP - 1.2948 * VM - 1.8598 * WT$$

La varianza residual vale $Vres = 12.5290$ y la razón de determinación: $R2 = 0.8733$

Aptdo. 3: Nos resulta la matriz de correlaciones lineales:

```
CORR =
    1.0000   -0.3686   -0.7899   -0.6884   -0.9051
   -0.3686    1.0000    0.0765   -0.0431    0.3850
   -0.7899    0.0765    1.0000    0.9665    0.8322
   -0.6884   -0.0431    0.9665    1.0000    0.6785
   -0.9051    0.3850    0.8322    0.6785    1.0000
```

que podemos interpretar como que la variable que más afecta a la primera variable (MPG) es la que tiene el módulo mayor en la primera fila, es decir la 5ª variable (WT) y la que menos la 2ª , es decir (VOL).

Además todas tienen correlación negativa, indicando que si aumentan, disminuye MPG (millas recorridas por galón de combustible).

Es destacable que el modelo lineal con solo la variable WT explica el 81.92% de la varianza de MPG, mientras que el hiperplano (mucho más complejo) explica el 87.33% de la varianza. Evidentemente el hiperplano ajusta mejor, pero un buen modelo de ajuste debe ser además lo más simple posible, no quedando en este caso claro si compensa usar 4 variables, en lugar de 1, para explicar un 6.41% más de las variaciones observadas en MPG.

Aptdo 4: Podemos buscar que modelo se aparta más del hiperplano de regresión de forma positiva, es decir, tiene un valor MPG superior al estimado por la regresión: $MAX = 11.9854$ (lo que se aparta), $modmax = 1$ (registro que lo hace) y lo hace el modelo $modelMAX = GM/GeoMetroXF1$.

También podemos conseguir cual se aparta más del valor estimado por la regresión pero en sentido negativo: $MIN = -9.0108$, $modmin = 29$ y lo hace el $modelMIN = SubaruLoyale$.

Aptdo. 5: Los valores imputados son:

ImpMedia = 33.7817, **ImpMediana = 32.45**, **ImpHiperpl = 39.0125**.