Carlos García Sancho

15 de abril de 2020

Car Sales

```
carsalesRAW=read.table("car_sales.txt",header=T,dec = ".")
carsales=na.omit(carsalesRAW)
carsales2=carsales[,-c(1,2,5)]
attach(carsales2)
```

Primer tratamiento de los datos. Leemos y eliminamos las muestras en las que falten datos.

A).- Indicar las dos variables explicativas que tienen mayor relación lineal con la variable respuesta. Justificación. Importancia.

```
c = cor(carsales2)
det(c)
heatmap(c, scale="none")
```

Deducimos a partir del mapa de calor y la matriz de covarianzas que las variables explicativas con mayor relación lineal para la variable respuesta (reventa) son caballos y precio, por tanto deducimos que a partir de ambas podremos obtener un buen modelo lineal para predecir la variable reventa. (Antes debemos ver la colinealidad entre estas dos variables, si es alta puede existir el problema de multicolinealidad, con lo cual el modelo obtenido no sería bueno).

B).- Indicar las dos parejas de variables explicativas que tienen mayor relación lineal entre ellas. Justificación. ¿Qué consecuencias puede tener estas relaciones en este estudio?

En la tabla de correlaciones observamos que las parejas de variables explicativas con mayor relación lineal son Longitud-Batalla (0.853) y Caballos-motor (0.86).
La existencia variables explicativas tan correlacionadas implica que pueda existir el problema de multicolinealidad. También podemos ver que el determinante de la matriz de correlaciones es muy próximo a cero, lo que de igual forma nos muestra que puede existir el problemas de multicolinealidad y que nuestro modelo no sea bueno.

C).- Relacionar a través de un modelo lineal la variable respuesta en función del tipo de vehículo. Interpretar el estimador del coeficiente asociado en el modelo al tipo de vehículo.

```
carsales3 = carsales[,-c(1,2)]
attach(carsales3)
tipo
class(tipo)
levels(tipo)
contrasts(tipo)

r1 = lm(reventa~tipo, data = carsales)
plot(r1)
summary(r1)
shapiro.test(residuals(r1))
```

La variable tipo es de clase factor, esta codificada por cero y uno (automóvil y camión, respectivamente) por lo que podemos hacer la regresión lineal de la reventa respecto el tipo.
Planteamos un modelo Y = B0 + B1 · tipo.
Planteamos el Contraste de Hipótesis
H0 : B1 = 0
H1: B1 /= 0
Obtenemos un p-valor de 0.326, mayor que el nivel de significación, luego no podemos rechazar la hipótesis nula. B1 es 0, por tanto no influye el valor del tipo en nuestro modelo. Hemos obtenido un modelo que no sirve para explicar la variable reventa.

D).- Obtener el modelo que permita explicar la variable respuesta en función de todas las variables explicativas (sin incluir marcas ni modelo).

```
todas.car.lm = lm(reventa ~ ventas+tipo+precio+motor_s+caballos+batalla+anchura+longitud+peso_revestimiento+tapón_combustible+kpl, data = carsales)
todas.car.lm
summary(todas.car.lm)
plot(todas.car.lm)
shapiro.test(residuals(todas.car.lm))
# Hay normalidad en los residuos (aunque el p valor = 0.28 pequeño)
```

Obtenemos mediante el comando lm el modelo ***todas.car.lm*** que explica la variable reventa en función del conjunto de variables
{ventas, tipo, precio, motor_s, caballos, batalla, anchura, longitud, peso_revestimiento, tapón_combustible, kpl}.
En el contraste fundamental obtenemos un p-valor < 2.2e-16, por tanto rechazamos la hipótesis nula H0, es decir, el vector Beta formado por los por coeficientes de cada variable explicativa en cada modelo es distinto del vector nulo. Esto significa que estas variables explican la variable reventa, con un r-cuadrado de 0.9472. Esto significa que la variabilidad de la variable respuesta queda explicada en un 94.72 % por este hiperplano de regresión o modelo.
Mediante el comando plot vemos que parecen cumplirse las hipótesis del modelo lineal.

E).- Aplicar los métodos de selección paso a paso y:
- Comparar, brevemente, los diferentes modelos obtenidos, en los apartados D y E.

1. Aplicamos los métodos de regresión paso a paso.

```
step(car.lm,direction = "both")
#reventa ~ precio + motor_s + longitud + peso_revestimiento + tapón_combustible + kpl
# Toma como variables explicativas precio, motor_s, longitud, peso_revestimiento, tapón_-
combustible, kpl
car.stepbnf.lm = lm(reventa ~ precio + motor_s + longitud + peso_revestimiento + tapón_-
combustible + kpl, data = carsales)
summary(car.stepbnf.lm)
plot(car.stepbnf.lm)
shapiro.test(residuals(car.stepbnf.lm))
# p valor = 0,12, no podemos rechazar normalidad en los residuos
```

Regresión hacia atrás y hacia delante. Una mezcla de los procesos descritos a continuación. Se realiza regresión hacia atrás y después vemos si se mejora el resultado añadiendo alguna variable. Nos quedamos con las variables siguientes:
{precio, motor_s, longitud, peso_revestimiento, tapón_combustible, kpl}
Llamamos a este modelo ***car.stepbnf.lm***.

```
step(car.lm,direction = "backward")
```

Regresión hacia atrás. Partimos de todas las variables y vamos eliminando aquellas con menos importancia (mayores p-valores). En cada paso actualizamos el modelo con las variables que queden. Con este método obtenemos el modelo anterior.

```
nulo=lm(reventa~1)

step(nulo,direction = “forward", scope=list(lower=nulo, upper=car.lm))

car.stepforward.lm = lm(reventa ~ precio + longitud + peso_revestimiento + tapón_combusti-
ble, data = carsales)
summary(car.stepforward.lm)
```

Regresión hacia delante. Partimos de un modelo nulo, sin variables, y vamos añadiendo variables hasta encontrar el mejor resultado. Hacemos el modelo ***car.stepforward.lm*** con las variables obtenidas
{precio, longitud, peso_revestimiento, tapón_combustible}.

3. Mejor selección de subconjuntos.

Observamos el subconjunto de variables cuyo modelo tiene mayor R cuadrado y obtenemos el modelo de regresión lineal a partir de las variables {precio, motor_s, caballos, batalla, longitud, peso_revestimiento, tapón_combustible, kpl}, ***car.subset.rsq***.

También tomamos el subconjunto de variables cuyo modelo tiene menor BIC, {precio, longitud, peso_revestimiento, tapón_combustible} y observamos que es el mismo que obtuvimos con el método de selección de variables por p-valor ***car.pvalor.lm***.

```
## Comparamos brevemente los modelos
# Por Bondad de ajuste
summary(car.lm)            # Multiple R-squared:  0.9472,     Adjusted R-squared:  0.9417
summary(car.pvalor.lm)     # Multiple R-squared:  0.9441,   Adjusted R-squared:  0.9421
summary(car.stepbnf.lm)    # Multiple R-squared:  0.946,    Adjusted R-squared:  0.9431
summary(car.stepforward.lm) # Multiple R-squared:  0.9441,Adjusted R-squared:  0.9421
summary(car.subset.rsq)    # Multiple R-squared:  0.9468, Adjusted R-squared:  0.9429

# Por BIC
BIC(car.lm)               # 622.3848
BIC(car.pvalor.lm)        # 599.2804
BIC(car.stepbnf.lm)       # 601.2307
BIC(car.stepforward.lm)   # 595.8155
BIC(car.subset.rsq)       # 608.8155
```

Hemos obtenido los modelos ***todas.car.lm***, ***car.pvalor.lm***, ***car.stepbnf.lm***, ***car.stepforward.lm*** y ***car.subset.rsq***.
Los comparamos por bondad de ajuste y vemos que todas son parecidos.
Los comparamos por el indicador BIC y vemos que el menor lo tiene el modelo ***car.stepforward.lm***
BIC = 595.8155.
El modelo con el más bajo valor de BIC es considerado el mejor en explicar los datos con el mínimo número de parámetros.

F).- ¿Qué modelo propondría? Justificar la respuesta e intenta dar una interpretación del mismo, desde el punto de vista del vendedor de vehículos.

```
summary(car.stepforward.lm)
```

Por lo anterior propondría el modelo ***car.stepforward.lm***, que hace uso de las variables {precio, longitud, peso_revestimiento, tapón_combustible}.
Tiene un menor BIC, un indicador que premia el ajuste o la capacidad explicativa pero penaliza su complejidad.

Un modelo menos complejo puede suponer una mejor predicción para nuevos datos.

Vamos a analizar este modelo desde un punto de vista matemático.
En el contraste fundamental obtenemos un p-valor < 2.2e-16. Por tanto, rechazamos hipótesis nula, el vector Beta es no nulo. Concluimos que estas variables explican conjuntamente la variable reventa. Obtenemos un r-cuadrado de 0.9441, lo que significa que nuestro modelo explica la variabilidad de la reventa en un 94.41%.
Se cumple la normalidad de los residuos.

El coeficiente de la variable precio en nuestro modelo es 0.84, teniendo en cuenta que la variable respuesta reventa y la variable precio están en distintas escalas (reventa en euros y precio en miles de euros) podemos ver que está es la variable más explicativa, lo cuál tiene lógica.
El coeficientes de longitud y peso_revestimiento son negativos, lo cual muestra que a mayor reventa, menor longitud e igual con peso_revestimiento. Esto puede significar que en el mercado de segunda mano tienen mayor valor los coches pequeños que los coches grandes.

G).- Realizar una predicción puntual del precio de reventa de un vehículo con las características medias de la base de datos. Cálculo e interpretación de los intervalos asociados a dicha predicción a un nivel de confianza del 95%.

```
predict(car.stepforward.lm,newdata=data.frame(precio=(mean(precio)),longitud=mean(longitud),peso_revestimiento=mean(peso_revestimiento),tapón_combustible=mean(tapón_combustible)),interval="confidence")
mean(reventa)
```

Realizamos dicha predicción y obtenemos un valor de 18.03154, que coincide con la media de la reventa de nuestros datos.
Obtenemos el siguiente intervalo de confianza del 95%: [17.51998, 18.5431]. Esto quiere decir que la probabilidad de que el precio de reventa de un vehículo con tales características estará en ese intervalo es del 95%. Parece un intervalo de confianza suficientemente pequeño.

H).- Obtener el mejor modelo, desde el punto de vista matemático, con solamente tres variables explicativas. Breve estudio del mismo.

```
summary(car.subsets)

tres.subset.lm=lm(reventa~precio,peso_revestimiento,tapón_combustible,data=carsales)

summary(tres.subset.lm)
plot(tres.subset.lm)
```

El mejor modelo con tres variables explicativas se obtiene a partir de {precio, peso_revestimiento, tapón_combustible} según el comando regsubsets de la librería leaps,
pero el modelo que obtenemos a partir de estas tres variables
reventa = B0 + B1 · precio + B2 · peso_revestimiento + B3 · tapón_combustible

tiene B2=B3=0, luego es en realidad
reventa = B0 + B1 · precio.
Con el comando plot observamos que no se cumplen muchas de las hipótesis previas. Es un mal modelo desde el punto de vista matemático.