Exámen final Modelos Lineales

Problema 2

La concentració de lactat en sang serveix freqüentment per predir la resistència dels atletes. La identificació d'alguns predictors de la resistència pot ajudar als entrenadors i atletes a avaluar els canvis del seu rendiment.

Per identificar aquests predictors es van seleccionar vint-i-quatre dones ciclistes ben entrenades i se'ls va practicar una prova física en un cicloergonòmetre. La prova física va consistir en realitzar etapes de tres minuts fins l'esgotament. Als 30 segons d'haver finalitzat cada etapa de tres minuts, se'ls va

[1]Podeu fer servir les files 1,2,4,5,9.

treure sang capilar del dit índex per analitzar el lactat plasmàtic i obtenir els valors predictors següents: llindar de lactat (P-Tlac), DMax (DMax), LT-log-log (P-Tlac.II), P-4 mmol.L^{-1} (P-4mM) i taxa de treball corresponent a un increment de 1 mM del valor basal (Rise.1.PB). El rendiment de resistència (AV.Power) s'avaluà 7 dies més tard mitjançant una prova de bicicleta, d'una hora de durada, en la que les atletes havien d'aconseguir la potència de sortida més alta possible.

Els resultats es troben en el fitxer de dades `CyclingPower.xls`.

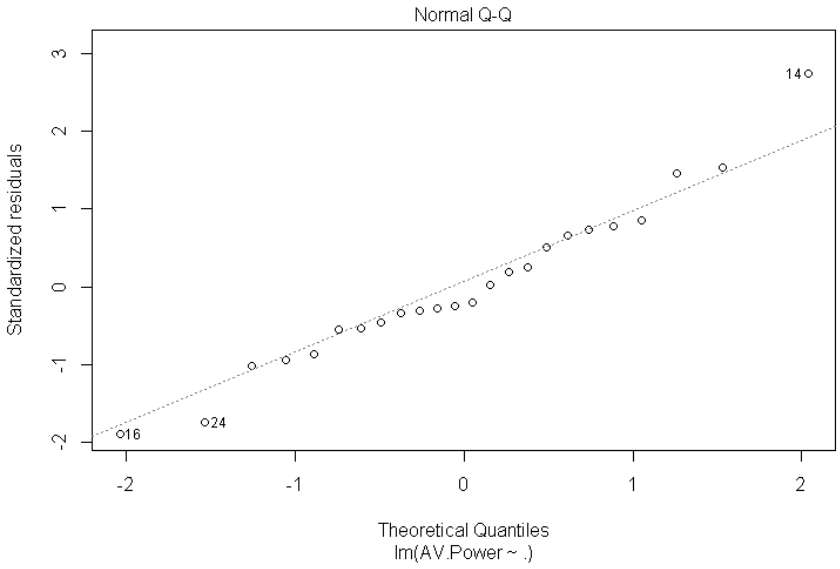
1. Calculeu l'hiperplà de regressió i el coeficient de correlació múltiple de AV.Power sobre les altres variables. Quina és la variància estimada de l'error?
2. És un model amb un bon ajust? Vol dir això que és significativa la regressió? Concreta què significa cada pregunta.
3. Amb els gràfics o els estadístics adients, investigueu la diagnosi d'aquest model en el següents punts:
   (i) Variància constant dels errors.
   (ii) Hipòtesi de normalitat.
   (iii) Punts amb influència potencial (leverage).
   (iv) Outliers.
   (v) Punts influents.
   (vi) Creieu que pot haver un problema de multicolinealitat? En què us baseu?

   Concreteu els punts problemàtics.
4. Què podem dir del punt 21? En què millora el model si eliminem aquest punt de les dades? I què podeu dir del 14?
5. Contrasteu si els coeficients de regressió de les variables P-Tlac i P-Tlac.II són iguals.
6. Si una corredora té uns valors de P-Tlac=175, DMax=158, P-Tlac.II=131, P-4mM=206 i Rise.1.PB=182, quina és l'estimació del seu rendiment de resistència (AV.Power) i l'error estàndard d'aquesta estimació.

   Amb una probabilitat del 0.95, entre quins valors es troba el rendiment de resistència (AV.Power) d'una corredora concreta que té els valors anteriors?
7. Calculeu la correlació parcial entre AV.Power i Rise.1.PB si eliminem la informació de P-Tlac, DMax, P-Tlac.II i P-4mM.

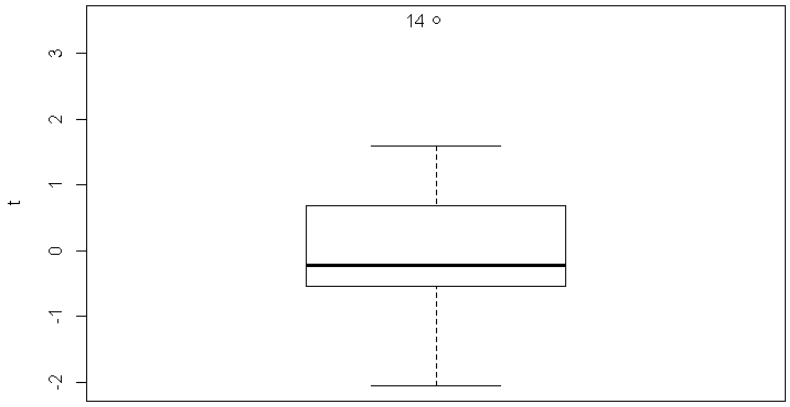
```
#Leemos los datos
datos<-read.table("CyclingPower.csv",sep=";",dec=",",header=TRUE)
datos

#1)
g <- lm(AV.Power ~ ., data = datos)
summary(g)
# Varianza estimada del error
summary(g)$sigma^2
#> summary(g)$sigma^2
#[1] 121.3527

#2)Una buena medida de ajuste es calcular la correlación:
summary(g)
#Multiple R-squared: 0.7334
#Podemos decir que tiene un buen ajuste, aunque no del todo exacto

#regresión significativa
summary(g)
#F-statistic: 9.905 on 5 and 18 DF,  p-value: 0.0001111
# La regresión es significativa.
```

```
#Que sea significativa quiere decir que es útil, que tiene sentido calcularla
#y que las regresoras afectan a la respuesta y nos ayudan a hacer predicciones.
#Si la regresión es significativa rechazamos la siguiente hipótesis:
#betaP.4mM=betaP.Tlac=betaP.Tlac.ll=betaDMax=betaRise.1.PB= 0


#3)
#a)
#varianza constante(Homocedasticidad)
# Recta de regressió entre valores absolutos y valores ajustados.
summary(lm(abs(residuals(g)) ~ fitted(g)))
#p-value: 0.2513
#La regresión no es significativa, porlo tanto el pendiente de la recta = 0 (Varianza constante)
# Método John Fox.
ncvTest(g)
#p = 0.3452906
#Por este método también vemos que la varianza es constante

#b)
#Hipótesis de normalidad
# Comprobamos gráficamente y analíticamente
#Grafico: QQ-plot
plot(g, which = 2)
```

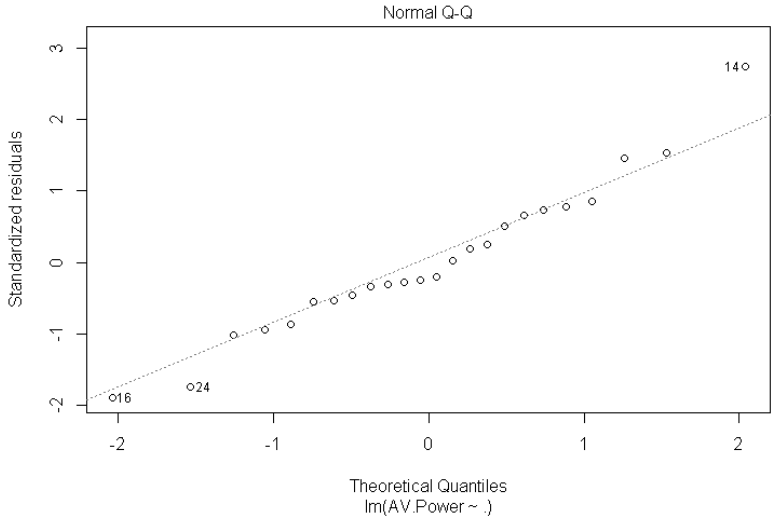



```
# Tenemos 3 puntos bastante alejados
# Test formals
shapiro.test(residuals(g))
#p-value = 0.5117   aceptamos normalidad

# c) Puntos con influencia potencial
leverage <- hatvalues(g)
# El criterio a seguir será: h > 2(k+1)/n
k <- 5 # k num vars
which(leverage > 2*(k+1)/24)
# Nos da solo el punto 21

# d) Outliers
# Criterio: |t| > 2
t <-rstudent(g)
which(abs(t)>2)
# Da los puntos 14 y 16
#Si probamos con Bonferroni vemos que solo nos da el punto 14
outlierTest(g)
```

```
# 14
#Gráficamente
Boxplot(t)
```




```
# e) Puntos influentes
#Utilizamos el método DFFITS
dffits(g)
# Calculamos el límite
k <- 2*sqrt((5 + 1)/24)
which(abs((dffits(g)))>k)
# Los puntos influyentes son 14 16 21 24

# f) Problemes de multicolinealitat
#Miramos la correlación entre las variables regresoras
round(cor(datos[,1:5]),3)
# Tenemos un 0.952 entre  P.Tlac y Dmax, que es el más alto
# Miramos los FIV's
vif(g)
# Son menores que 10, por lo tanto no podemos hablar de multicolinealidad


# 4) Quitamos el punto 21
# El punto 21 es un punto influente (influencia potencial)
g2 <- lm(AV.Power ~ ., data = datos[-21])
summary(g2)
summary(g2)$sigma^2
#[1] 121.3527 #Vemos que se mantiene
#Vemos que el ajuste aumenta un poco
plot(g2, which = 2)
```




```
# Miramos la normalidad
shapiro.test(residuals(g2))
# p-value = 0.5117   el hecho de quitar el punto 21 no afecta a la normalidad

# Ahroa mriamos la homocedasticidad
ncvTest(g2)
# p = 0.3452906 , seguimos considerando cierta la homocedasticidad, aunque ahora el p valor es más alto que con el punto 21.
#EL punto 14 vemos que es outlier y punto influyente

#5)

# 6) IC
predict(g, newdata = data.frame(P.4mM = 200,  P.Tlac = 175,          P.Tlac.ll = 131,       DMax = 152,
Rise.1.PB = 180), interval = 'prediction')
#fit     lwr     upr
#1 180.6768 153.6813 207.6722

#7)
g3 <- lm(AV.Power ~ Rise.1.PB, data = datos)
summary(g3)
round(cor(datos[,1:2]),3)
#Correlación = 0.803
```

Problema 3

Amb la mateixa base de dades del problema anterior i el mateix model de partida, sembla que tenim un problema de multicolinealitat.

1. Trobeu el "millor"model per dos mètodes diferents de selecció de variables com, per exemple, AIC i C_p de Mallows.
   (a) Quines són les variables seleccionades?
   (b) Quins són els coeficients de determinació ajustats d'aquests models? Compareu-los amb el del model complet. Llavors, què hem guanyat?
   (c) Calculeu l'interval de confiança al 95% per al coeficient de regressió de la variable Rise.1.PB en els models, el complet i els seleccionats.
2. Un altre possibilitat és fer servir la Ridge Regression. Quins són els coeficients obtinguts? Expliqueu breument les avantatges i inconvenients d'aquest mètode front a la selecció de variables.
3. Amb el model reduït de tres variables regressores (P.Tlac.ll, DMax i Rise.1.PB) el punt 21 encara fa nosa. Ajusteu un model per un mètode robust adient.

```
datos<-read.table("CyclingPower.csv",sep=";",dec=",",header=TRUE)
datos

#1)
g <- lm(AV.Power ~ ., data = datos)
summary(g)


# Selección por AIC
g_AIC <- step(g)
#Vemos dos opciones (diferentes AIC)
#Step:  AIC=118.44
#AV.Power ~ P.4mM + P.Tlac.ll + DMax + Rise.1.PB
#ó
#Step:  AIC=116.99
#AV.Power ~ P.Tlac.ll + DMax + Rise.1.PB


# Selección por Cp de Mallows
#No me sale

# Intento hacerlo en vez de Cp de Mallows, selección del modelo por Forward Stepwise
# Fórmula del modelo completo
gc.formula <- formula(g)
# Model simple
g0 <- lm(AV.Power ~ 1, data = datos)
# Forward stepwise
modelo <- g0
add1(modelo, scope = gc.formula, test="F")
#Añado la variable con la F mayor(Rise.1.PB)
model <- update(g0, ~ . + Rise.1.PB)
# Al modeol g0 le sumo Rise.1.PB
add1(modelo, scope = gc.formula, test="F")
#Single term additions
#Model:
#  AV.Power ~ Rise.1.PB
#Df Sum of Sq    RSS    AIC F value Pr(>F)
#<none>              2704.7 117.39
#P.4mM     1    29.086 2675.6 119.13  0.2283 0.6377
#P.Tlac    1    38.019 2666.7 119.05  0.2994 0.5900
#P.Tlac.ll  1   140.432 2564.3 118.11  1.1501 0.2957
#DMax      1   137.501 2567.2 118.14  1.1248 0.3009
#Añado la variable con la F mayor(P.Tlac.ll)
model <- update(modelo, ~ . + P.Tlac.ll)
# A el modelo ñe sumo P.Tlac.ll
```

```
add1(modelo, scope = gc.formula, test="F")
# Ya no tengo que añadir más

g_forward <- modelo

# Coeficients de determinación de los modelos

summary(g)$adj.r.squared
#[1] 0.6593934
summary(g_AIC)$adj.r.squared
#[1] 0.6840969
summary(g_forward)$adj.r.squared
#0 #está mal...

#c)
# Intervalos de confianza
confint(g, "income")
confint(g_AIC, "income")
confint(g_forward, "income")


#2)

# 3) Ajuste del modelo mediante el método LTS(Least trimmed squares)
library(MASS)
g_lts <- ltsreg(datos~ P.Tlac.ll + DMax + Rise.1.PB, data = datos, nsamp = "exact")
```

Problema 4 (Test de Breusch-Pagan)

Sota les condicions de Gauss-Markov, que inclou l'homocedasticitat, els mínims quadrats ordinaris proporcionen el millor estimador lineal no esbiaixat (BLUE), és a dir, no esbiaixat i eficient. Malgrat això, l'eficiència es perd en presència d'heterocedasticitat. Per estudiar la presència d'heterocedasticitat podem fer servir el test de Breusch-Pagan.
Donat el model $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$ assumim que l'heterocedasticitat pren la forma:

$$E(\epsilon_i) = 0 \qquad \sigma_i^2 = E(\epsilon_i^2) = h(\mathbf{z}_i'\boldsymbol{\alpha}) \qquad \text{per a tota } i = 1, \ldots, n$$

on $\mathbf{z}_i' = (1, z_{1i}, z_{2i}, \ldots, z_{pi})$ i $\boldsymbol{\alpha} = (\alpha_0, \alpha_1, \cdots, \alpha_p)$ és un vector de coeficients desconeguts i $h(\cdot)$ és una funció no especificada que només pot prendre valors positius.
La hipòtesi nul·la d'homocedasticitat és:

$$H_0 : \alpha_1 = \alpha_2 = \cdots = \alpha_p = 0$$

i sota aquesta hipòtesi tenim que $\sigma_i^2 = h(\alpha_0)$ constant.
Considereu el model a estimar per mínims quadrats ordinaris i assumiu la normalitat dels errors.

El procediment a seguir és el següent:

1. Apliqueu el mètode dels mínims quadrats al model considerat i calculeu els seus residus e_i. Estimeu la variància dels errors $\tilde{\sigma}^2 = \sum e_i^2/n$.
2. Feu una regressió auxiliar de la variable $e_i^2/\tilde{\sigma}^2$ com a resposta i les variables $\mathbf{z}$ com regressores i calculeu la suma de quadrats explicada ESS o suma de quadrats de la regressió.
3. L'estadístic del test és BP = ESS/2 amb distribució asimptòtica χ_p^2 i l'homocedasticitat es rebutja si l'estadístic supera el valor crític de la taula.

El procediment explicat requereix el coneixement de las variables regressores $\mathbf{z}$ i no és necessari conèixer la funció h. Sovint les regressores $\mathbf{z}$ són algunes de les regressores del model original.
Amb la base de dades `Ornstein` del paquet `car` considerem el model lineal amb la variable `interlocks` com a resposta i les variables `assets`, `sector` i `nation` com a regressores.

(a) Contrasteu l'homocedasticitat d'aquest model amb el test de Breusch-Pagan segons el procediment explicat i amb les mateixes regressores per al model auxiliar que pel primer model.

Calculeu l'estadístic i el seu p-valor.

(b) Compareu el resultat amb la funció `ncvTest` del paquet `car`. Configureu correctament el paràmetre `var.formula`.

(c) Calculeu el test amb la funció `bptest` del paquet `lmtest`. Configureu correctament el paràmetre `studentize`.

Nota 1: Observeu que algunes variables regressores són factors, de forma que en el model de regressió es transformen en vàries dicotòmiques i això modifica els graus de llibertat.
Nota 2: Una versió simplificada del test[2] demana la regressió de e_i^2 sobre $\mathbf{z}$, llavors l'estadístic és nR^2 d'aquesta regressió auxiliar. Aquí no farem servir aquesta aproximació.