XS-2130 Modelos de Regresión Aplicados II Sem 2020 Grupo 2

Práctica examen parcial 3.

Se entrega el jueves 30 de octubre a más tardar 12 mediodía, al correo electrónico [Gilbert.brenes.camacho@gmail.com](mailto:Gilbert.brenes.camacho@gmail.com). Envían un solo documento de Word en el que aparezca cada pregunta que se efectúa, las respuestas respectivas pegando los comandos o resultados de R y una copia del script de R pegado al final del documento de Word. El nombre del archivo debe tener su nombre y la versión que le tocó. Por ejemplo, gilbertbrenesversion8.docx.

Total de puntos: 50 ptos.

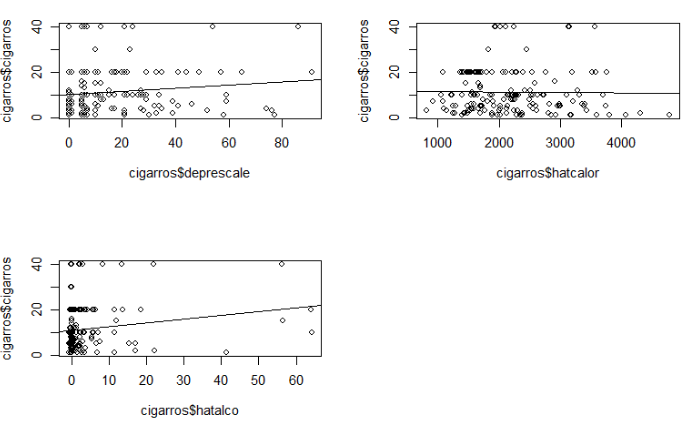
Pregunta 1.

Se le da un archivo cigarros.Rdata en formato R. Corresponde a una muestra de fumadores y se quiere investigar si la cantidad de alcohol consumido, la depresión, y la cantidad de calorías consumidas logran predecir la cantidad de cigarros que fuma una persona. La base de datos contiene entonces las siguientes variables:

|  |  |
| --- | --- |
| psalud | percepción de la salud |
| deprescale | Escala de depresión de 0 a 100, donde 100 es estar deprimido al máximo (medida en puntos porcentuales) |
| smoker | 1=Si la persona fuma |
| hatcalor | Calorías por día en kilocalorías diarias |
| hatalco | Consumo de alcohol en gramos por día |
| cigarros | Número de cigarros fumados en un día |

Conteste con R las siguientes preguntas.

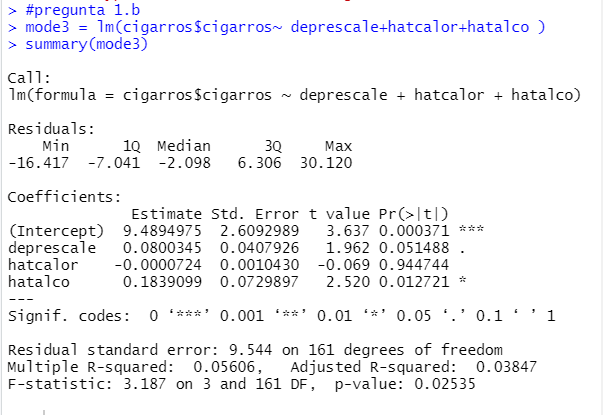
1. Haga una matriz de gráficos con el gráfico de dispersión de la variable dependiente (cigarros) vs. hatcalor, hatalco y deprescale (3 ptos.)



1. Según los gráficos de dispersión, ¿en cuáles asociaciones puede argumentar linealidad? (2 ptos.)

Se puede observar un poco de linealidad con las variables hatalco y deprescale

1. Estime una ecuación de regresión en la que prediga el número de cigarros en función de la escala de depresión, las calorías consumidas y la cantidad de alcohol consumido (3 ptos.)



1. Interprete los coeficientes de estas tres variables explicativas (3 ptos.)

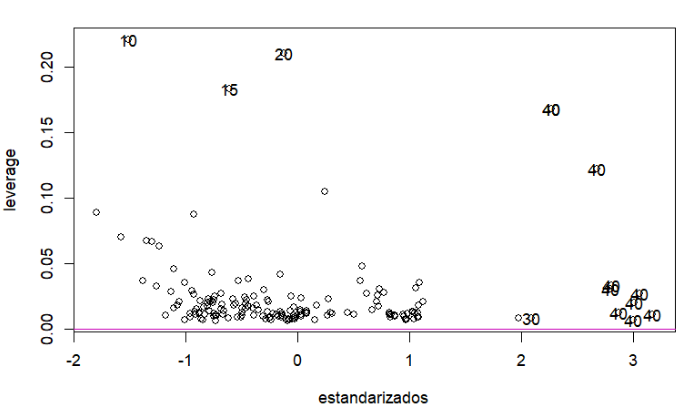
B0= la cantidad de cigarros consumidos en un día es de 9,489 cigarros cuando las demás variables son iguales a cero.

B1= en la escala de depresión la depresión aumenta 0,08 puntos porcentuales, por cada aumento de un cigarro fumado en el dia

B2= la cantidad de calorías consumidas disminuye 0, 00007 calorías diarias, por cada aumento de un cigarro fumado en el dia, manteniendo las demás variables constantes

B3= el consumo de alcohol aumenta 0,18 gramos, por cada aumento de un cigarro fumado en un dia.

1. Haga un gráfico de residuos estandarizados vs. leverage y, si hay valores extremos y/o influyentes, identifique los casos con la variable cigarros (4 ptos.).

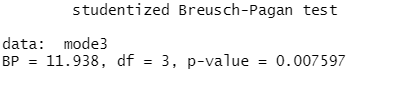


1. ¿Hay valores extremos en el consumo de cigarros que tengan influencia en los resultados de la regresión? ¿Diga por qué? (2 ptos.)
2. Analice el supuesto de homoscedasticidad con una prueba de Breusch-Pagan, con un =0.05. ¿Se puede mantener el supuesto de homoscedasticidad? (3 ptos.)

H0: Sigma\_i = Sigma

H1: Sigma\_i <> Sigma

Hay suficiente evidencia estadística para rechazar la hipótesis nula de varianzas iguales

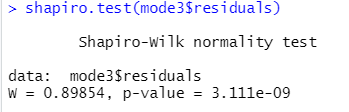


1. Analice la normalidad de los residuos con una prueba de Shapiro-Wilks, con un =0.05. ¿Qué concluye? (3 ptos.)

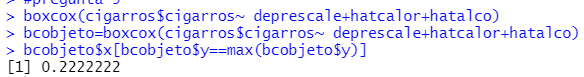
H0: miu=0

H1: miu <> 0

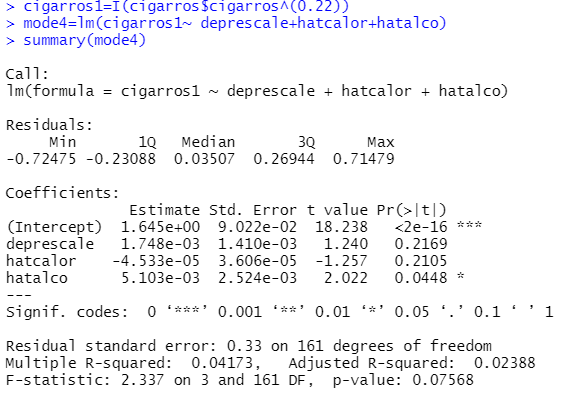
Hay suficiente evidencia estadística para rechazar la hipótesis nula de normalidad de los residuos



1. Haga una transformación de BoxCox a la variable dependiente escogiendo la transformación que más se ajuste a los datos (3 ptos.)

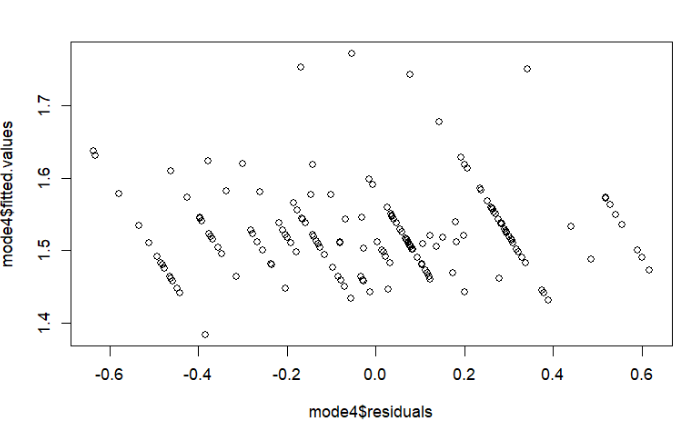


1. Vuelva a estimar el modelo en el que la variable transformada de cigarros es predicha por las calorías consumidas, el alcohol consumido y la escala de depresión. Usando como base las pruebas de hipótesis para los coeficientes de regresión, cómo cambia la intepretación de los factores que determinan la cantidad de cigarros al usar la variable transformada en comparación con la variable sin transformar? (3 ptos.)



1. Haga un gráfico de residuos vs. predichos y argumente si la forma de dicho gráfico sugiere o no homoscedasticidad (3 ptos.)

Debido a que los datos siguen un patrón, diría que no sugiere homocedasticidad

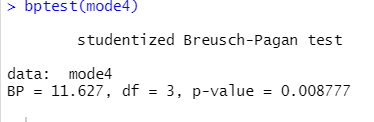


1. Estime una prueba de Breusch-Pagan para el modelo con la variable transformada, con un =0.05 y concluya apropiadamente (3 ptos.).

Ho: sigma\_i = sigma

H1= sigma\_i<> sigma

Hay suficiente evidencia estadística para rechazar la hipótesis nula de varianzas iguales



1. Conteste: Si la transformación de Box-Cox fue creada para corregir el supuesto de normalidad de los residuos, cómo una transformación de BoxCox puede afectar los resultados de una prueba de homoscedasticidad (ya sea White y Breusch-Pagan)? **Nota: No estoy buscando que digan que la prueba de Breusch-Pagan supone normalidad y la de White es más robusta ante tal violación**. (3 ptos)

La prueba de box cox trata de formar una distribución simétrica con los datos, entonces al manipular los datos de esa manera también puede alterar otros supuestos como la homocedasticidad.

Pregunta 2.

Un investigador en deportes está investigando si la presión arterial predice la tasa de sudoración (en g/mm). Para ello, toma una muestra de 10 corredores aficionados, les pide que corran durante media hora, y les coloca un medidor de presión así como unos parches para medir la tasa de sudoración. Logra medir dos variables independientes: presión sistólica y presión diastólica (en mm/Hg) máxima durante el recorrido. Al final de la carrera, les toma la tasa de sudoración. Hizo varios análisis en R y obtiene los resultados que se describen más abajo. Con base en ello, conteste las siguientes preguntas:

1. El hecho de que los coeficientes de diastólica y sistólica en las regresiones lineales simples son significativamente distintos de cero, pero en la regresión lineal múltiple no lo son, señala que hay un problema de multicolinealidad. ¿Qué otra información en las salidas de R sugiere que hay problemas de multicolinealidad? (3 ptos.)

Que el error estándar sea muy similares

1. ¿De qué manera específica la multicolinealidad afecta los resultados (estimación e inferencia) de un modelo de regresión lineal múltiple? (3 ptos.)

Al presentar multicolinealidad los intervalos para predecir se vuelven mas amplios y por emde menos precisas las estimaciones.

1. Calcule el VIF para los coeficientes de las variables diastólica y sistólica en el modelo de regresión lineal múltiple (3 ptos.)

1/(1-0,9419) = 17,21

Es decir según los vifs, se presentan problemas de multicolinealidad

> mod1=(lm(sudoracion~sistolica+diastolica))

>

> summary(mod1)

Call:

lm(formula = sudoracion ~ sistolica + diastolica)

Residuals:

Min 1Q Median 3Q Max

-2.0324 -0.5082 0.1852 0.6082 1.7950

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.3460 0.4234 3.179 0.0155 \*

sistolica 2.6760 1.4292 1.872 0.1033

diastolica 2.1385 1.4292 1.496 0.1782

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.339 on 7 degrees of freedom

Multiple R-squared: 0.9419, Adjusted R-squared: 0.9253

F-statistic: 56.76 on 2 and 7 DF, p-value: 4.722e-05

>

> cor(diastolica,sistolica)

[1] 0.95

>

> mod1a=(lm(sudoracion~sistolica))

> mod1b=(lm(sudoracion~diastolica))

>

> summary(mod1a)

Call:

lm(formula = sudoracion ~ sistolica)

Residuals:

Min 1Q Median 3Q Max

-1.9814 -1.2373 0.2859 0.9908 1.8126

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.3460 0.4550 2.959 0.0182 \*

sistolica 4.7076 0.4796 9.816 9.75e-06 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.439 on 8 degrees of freedom

Multiple R-squared: 0.9233, Adjusted R-squared: 0.9138

F-statistic: 96.36 on 1 and 8 DF, p-value: 9.746e-06

> summary(mod1b)

Call:

lm(formula = sudoracion ~ diastolica)

Residuals:

Min 1Q Median 3Q Max

-2.88766 -0.37647 -0.02034 0.81210 2.54408

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.3460 0.4852 2.774 0.0241 \*

diastolica 4.6807 0.5114 9.153 1.64e-05 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.534 on 8 degrees of freedom

Multiple R-squared: 0.9128, Adjusted R-squared: 0.9019

F-statistic: 83.77 on 1 and 8 DF, p-value: 1.637e-05