EXAMEN DE POSTGRADO 2022 - (MAESTRIA EN CIENCIA DE DATOS)

APELLIDOS Y NOMBRES: …SERDAN CARDENAS RAMON WILDER……… CEL:…… 73736998…

CORREO: wserdan@gmail.com

1. (Regresión múltiple)(Se pide dar solución en R) Un estudio quiere generar un modelo que permita predecir la esperanza de vida media de los habitantes de una ciudad en función de diferentes variables. Se dispone de información sobre: habitantes, analfabetismo, ingresos, esperanza de vida, asesinatos, universitarios, heladas, área y densidad poblacional. Con la siguiente información

library(dplyr)

datos <- as.data.frame(state.x77)

datos <- rename(habitantes = Population, analfabetismo = Illiteracy, ingresos = Income, esp\_vida = `Life Exp`, asesinatos = Murder, universitarios = `HS Grad`, heladas = Frost, area = Area,

.data = datos)

datos <- mutate(.data = datos, densidad\_pobl = habitantes \* 1000/area) datos

Descripción de variables:

Población : población estimada al 1 de julio de 1975.

Renta : renta per cápita (1974)

Analfabetismo : analfabetismo (1970, porcentaje de la población)

Life Exp : esperanza de vida en años (1969-71)

Asesinato : tasa de asesinato y homicidio no negligente por cada 100.000 habitantes (1976)

HS Grad : porcentaje de graduados de secundaria (1970)

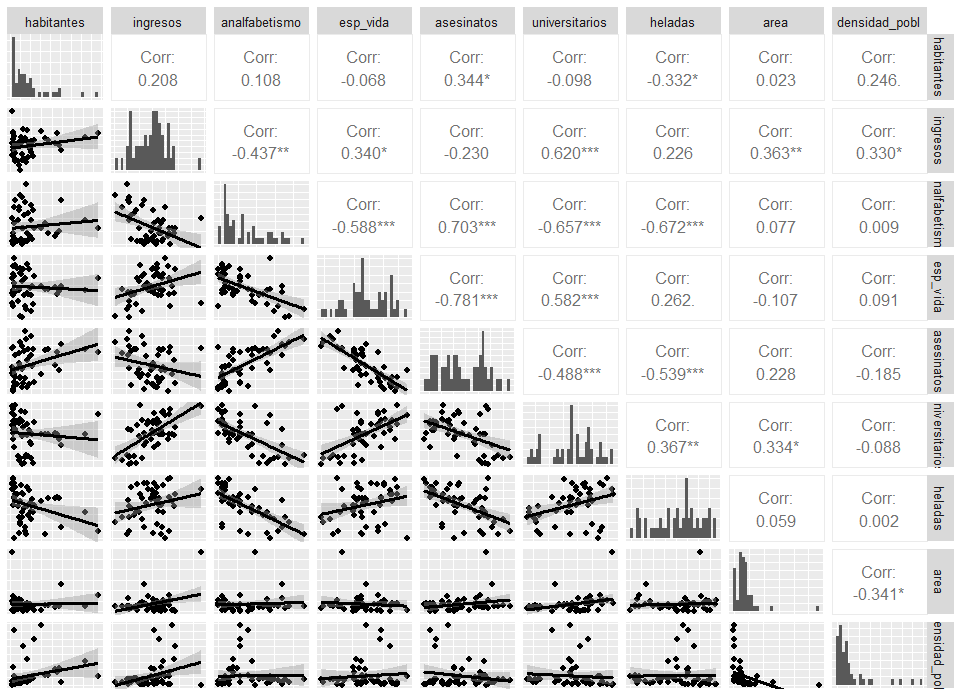
Helada : número medio de días con temperatura mínima bajo cero (1931-1960) en la capital o ciudad grande

Área : superficie terrestre en millas cuadradas

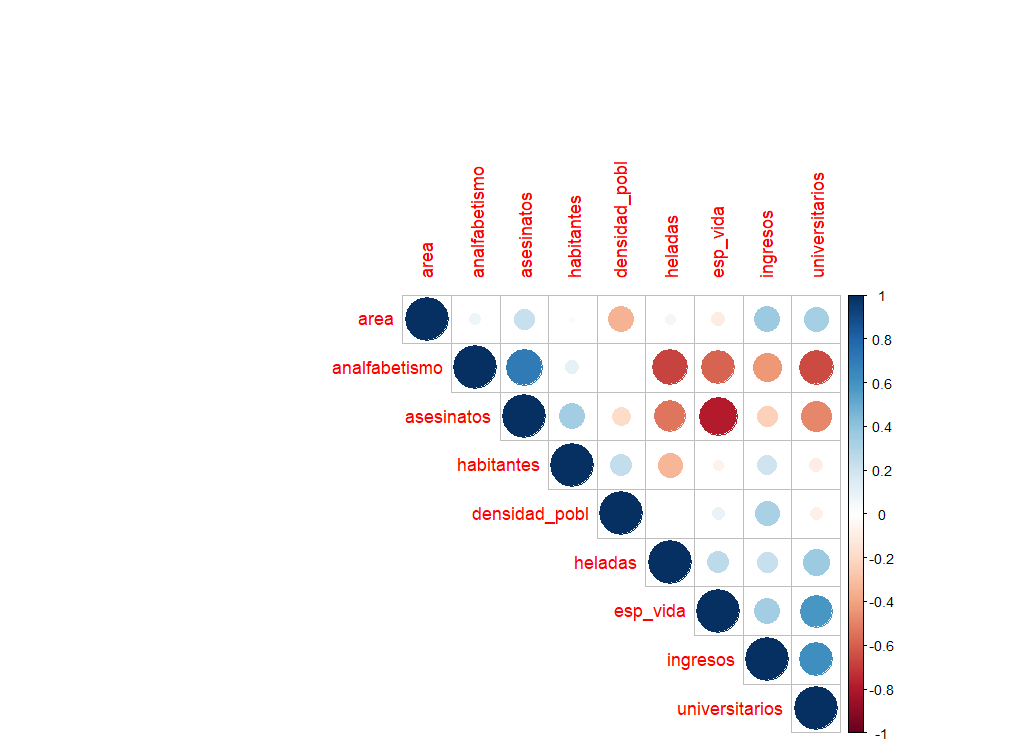
se pide: (todo con R)

1. Analizar la relación entre variables

Las correlaciones de todas las variables cuantitativas son las siguientes:



Utilizando la librería “corrplot” se determina que las correlaciones más importantes son :

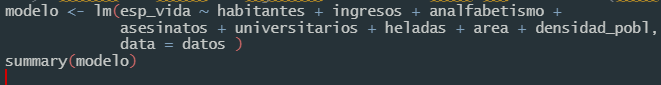


ANALISIS

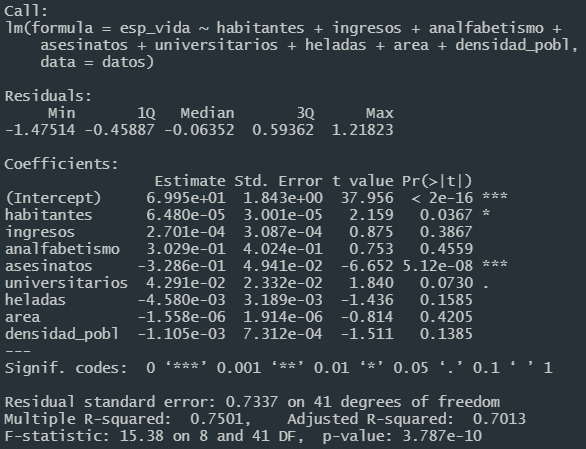
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Correlaciones positivas | |  | Correlaciones negativas | |
| Porcentaje de **Analfabetismo** | Tasa de **asesinatos** |  | Número de días con **helada** | Porcentaje de **Analfabetismo** |
| Porcentaje de graduados (**universitarios**) | Esperanza de vida |  | Esperanza de vida | Tasa de **asesinatos** |
| Porcentaje de graduados (**universitarios**) | **Ingresos** |  | Porcentaje de graduados (**universitarios**) | Porcentaje de **Analfabetismo** |

1. Generar un modelo de regresión lineal con todas las variables (hacer un summary y analizar)

Aplicamos el modelo a las 9 variables:



Con el siguiente resultado:



ANALISIS

De acuerdo al Summary del modelo se tiene un R2 importante de 0,75 que muestra una alta capacidad de explicar la variabilidad observada en la esperanza de vida.

El p-value del modelo es prácticamente cero (3,787e-10) por lo que se puede aceptar que el modelo no es por el azar.

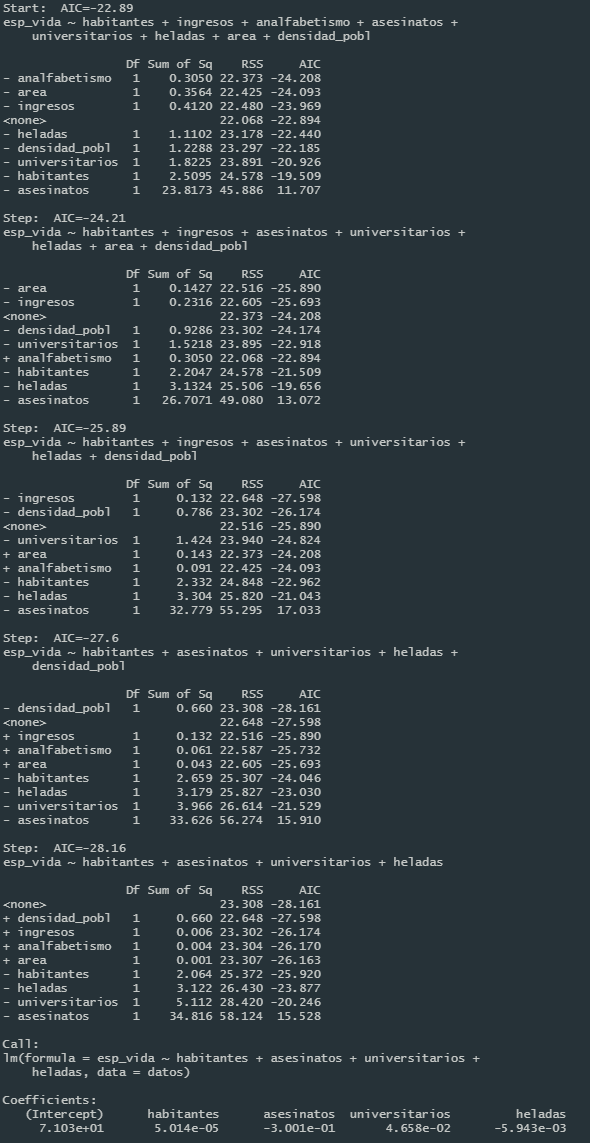
Una buena parte de los coeficientes no son significativos con un aporte casi nulo al modelo.

1. Selección de los mejores predictores (hacer un summary y analizar)

Aplicando



Se obtienen los siguientes resultados:

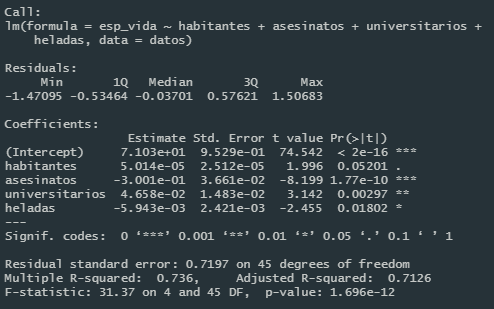


ANALISIS

Entonces el mejor modelo estaría dado por:



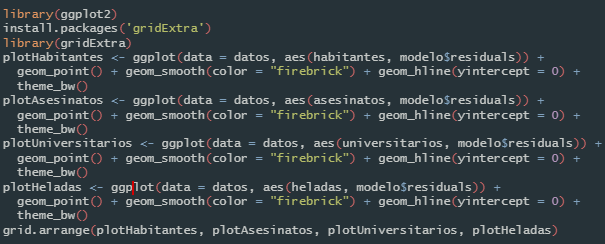
Resultando:



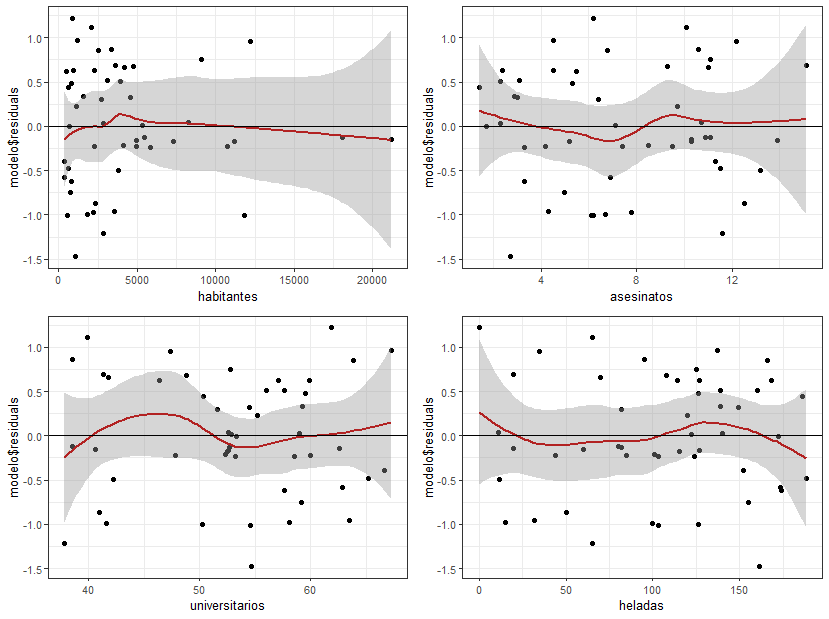
Se pueden interpretar cada uno de los coeficientes como elasticidades que explican la variable dependiente manteniendo las demás constantes. Según los resultados una reducción de los días con heladas en una unidad porcentual, la esperanza de vida aumenta en promedio 0,0059 puntos porcentuales manteniendo constantes los demás predictores.

1. Realizar la validación de condiciones para la regresión múltiple lineal

Aplicando:



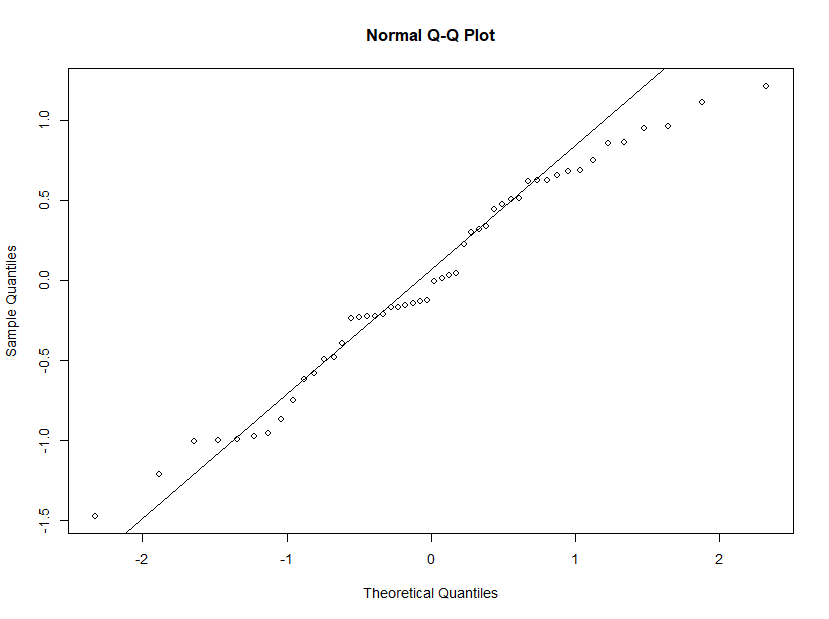
Gráficamente se verifica que la relacion es lineal ya que los residuos se distribuyen aleatoriamente en torno al origen con una variabilidad constante en el x.



1. Determinar si cumple con la distribución normal de los residuos

Vemos gráficamente aplicando el siguiente comando.



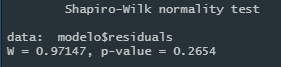


1. Realizar el test de normalidad

Aplicamos:



Con los siguientes resultados:

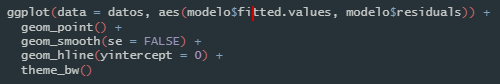


ANALISIS

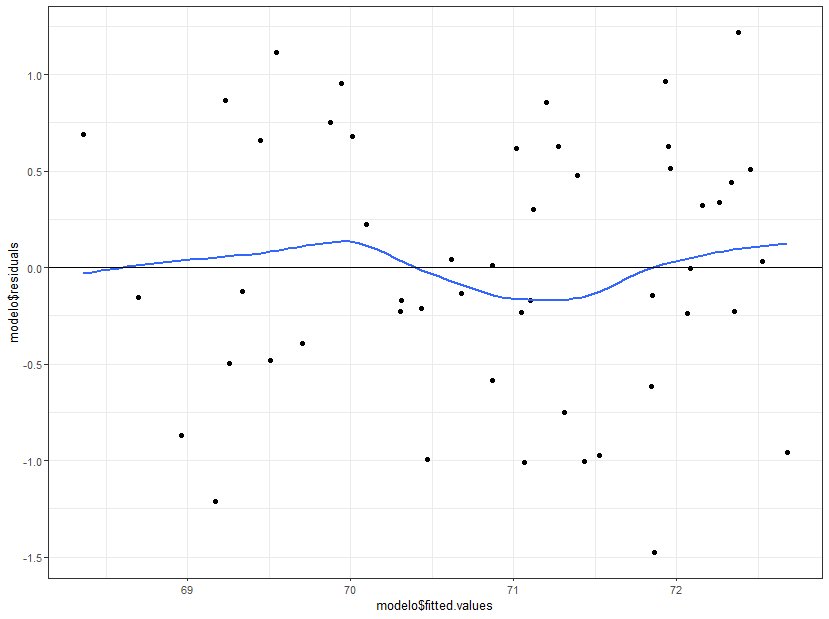
Debido a que el p-value es mayor a un nivel de significancia de 5%, no se rechaza la hipótesis nula de distribución normal de los residuos.

1. Realizar el test homocedasticidad

Realizamos un análisis gráfico con los siguientes comandos:



Obteniendo el siguiente resultado:



1. Realizar el Análisis de Inflación de Varianza (VIF)

Utilizando la librería:



Obtenemos el siguiente resultado:

**

1. Realizar ANOVA

https://rpubs.com/daniballari/Celec\_metodos\_estadisticos\_r

1. Considerando el estudio de rendimiento antes y después cuya información es:

grupoAntes: 2, 4, 6, 1, 3

grupoDespues: 5, 2, 7, 1, 6

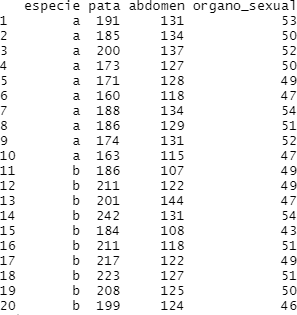
Se pide:

1. Realizar la prueba de rangos de signos de Wilcoxon en sus tres tipos de pruebas) (Se pide dar solución en R)
2. Verifique el resultado de R con el cálculo a mano. (en ambos incisos realizar la interpretación de todo el proceso)

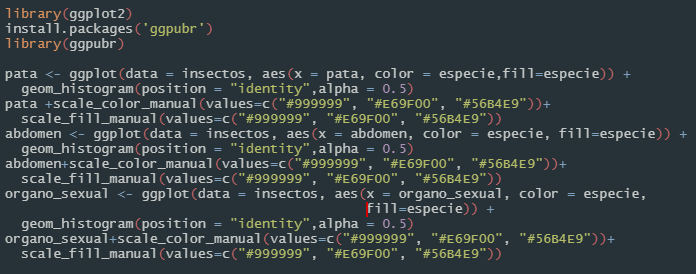
<https://gauss.inf.um.es/feir/50/>

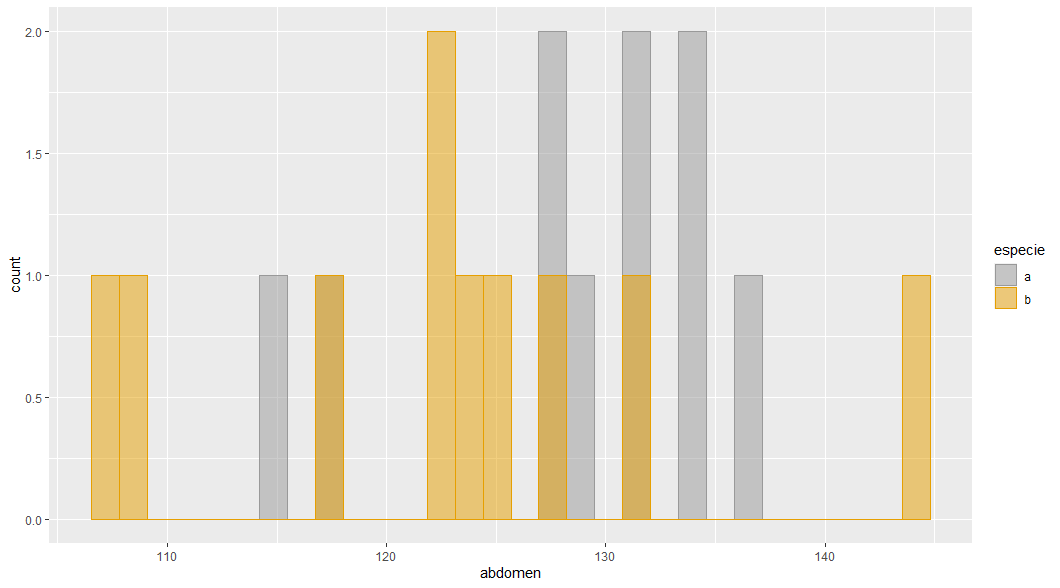
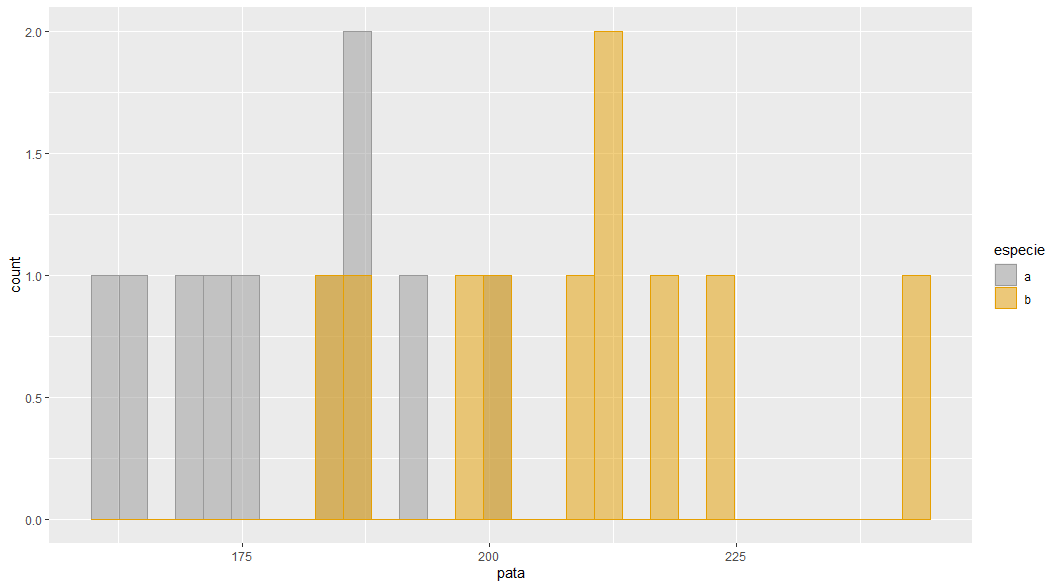
<https://www.cienciadedatos.net/documentos/18_prueba_de_los_rangos_con_signo_de_wilcoxon>

1. Un equipo de biólogos quiere generar un modelo estadístico que permita identificar a que especie (a o b) pertenece un determinado insecto. Para ello se han medido tres variables (longitud de las patas, diámetro del abdomen y diámetro del órgano sexual) en 10 individuos de cada una de las dos especies. Los datos son los siguientes (aplique análisis discriminante con R). Con sus correspondientes interpretaciones.



Primero, hacemos un análisis de las variables:

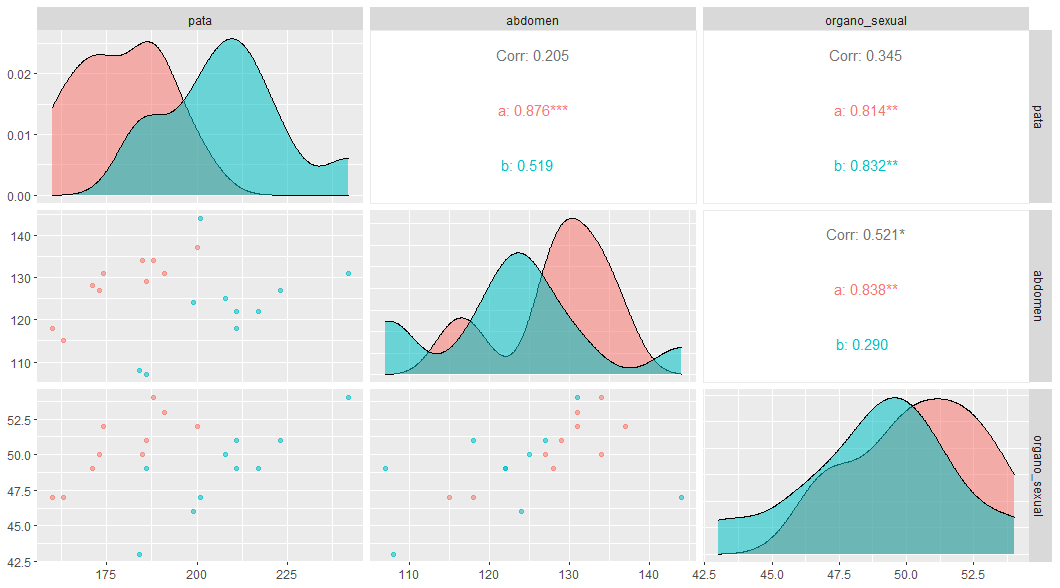




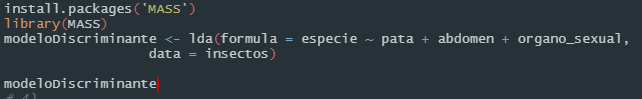
ANALISIS

El largo de la pata de los insectos muestra una clara separación en las dos espacies de insectos a y b.

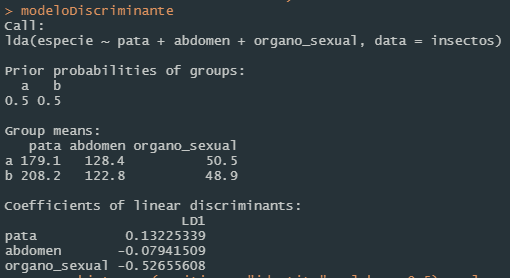
sdf



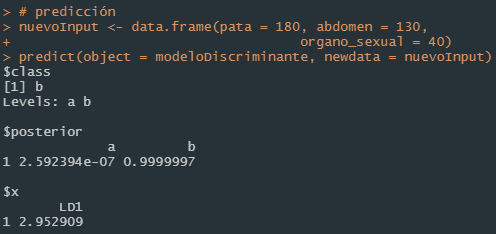
Análisis Discriminante Lineal



Con el siguiente resultado:



Podemos hacer una predicción en base al modelo:



ANALISIS

Con base a la función discriminante, la probabilidad posterior de que la nueva información del insecto pertenezca a la especie b es del 99,9% en contraposición al 0,1% de que pertenezca a la especie a.

<https://www.cienciadedatos.net/documentos/28_linear_discriminant_analysis_lda_y_quadratic_discriminant_analysis_qda>

<https://gc.scalahed.com/recursos/files/r161r/w24932w/S2_M1CDN107_C.pdf>

<https://rstudio-pubs-static.s3.amazonaws.com/575414_ee44bb2a707345dbb2d0bf91fee9fb75.html>

<https://rpubs.com/Joaquin_AR/233932>

1. Considerando la siguiente información. La cuantificación del contenido en grasa de la carne pude hacerse mediante técnicas de analítica química, sin embargo, este proceso es costoso en tiempo y recursos. Una posible alternativa para reducir costes y optimizar tiempo es emplear un espectrofotómetro (instrumento capaz de detectar la absorbancia que tiene un material a diferentes tipos de luz en función de sus características). Para comprobar su efectividad se mide el espectro de absorbancia de 100 longitudes de onda en 215 muestras de carne, cuyo contenido en grasa se obtiene también por análisis químico para poder comparar los resultados. El set de datos meatspec del paquete faraway contiene toda la información.

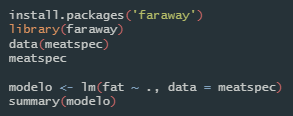
library(faraway)

data(meatspec)

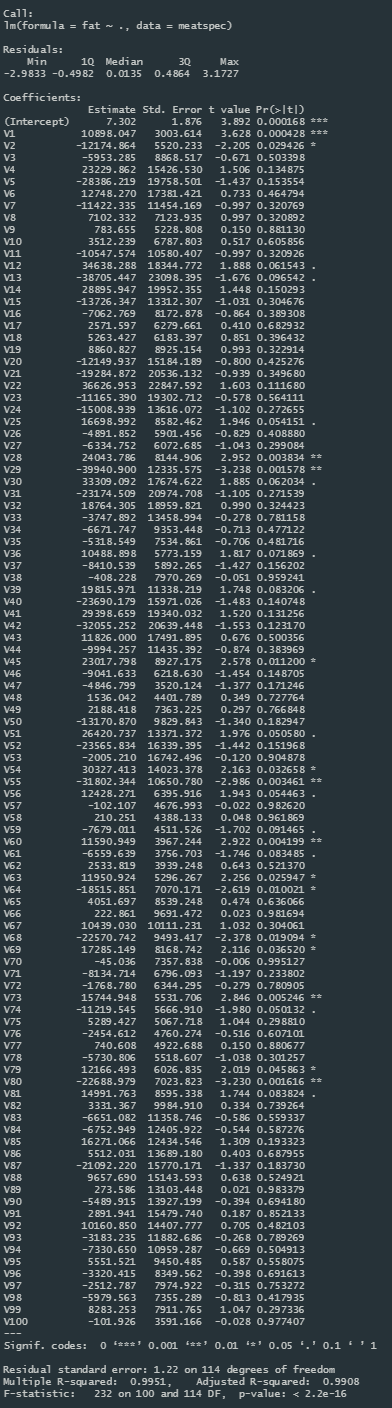
Se pide determinar:

* + - 1. el modelo inicial

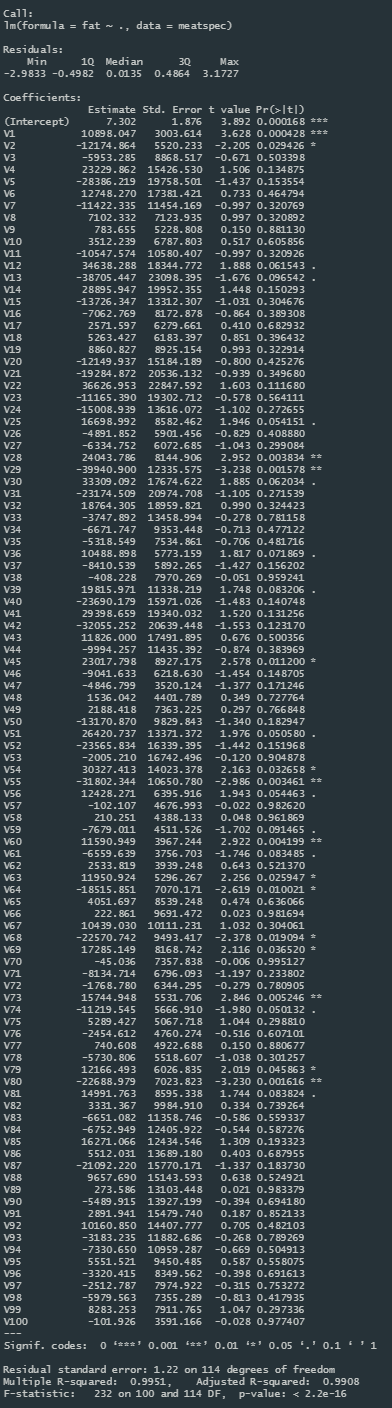
Instalamos el paquete “faraway” y aplicamos el modelo lineal a la totalidad de los datos.



Resultando:

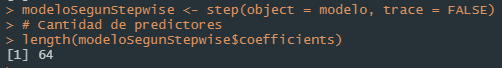


* + - 1. la suma de cuadrados del error

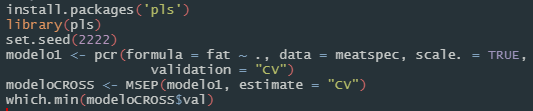




* + - 1. los predictores mediante stepwise



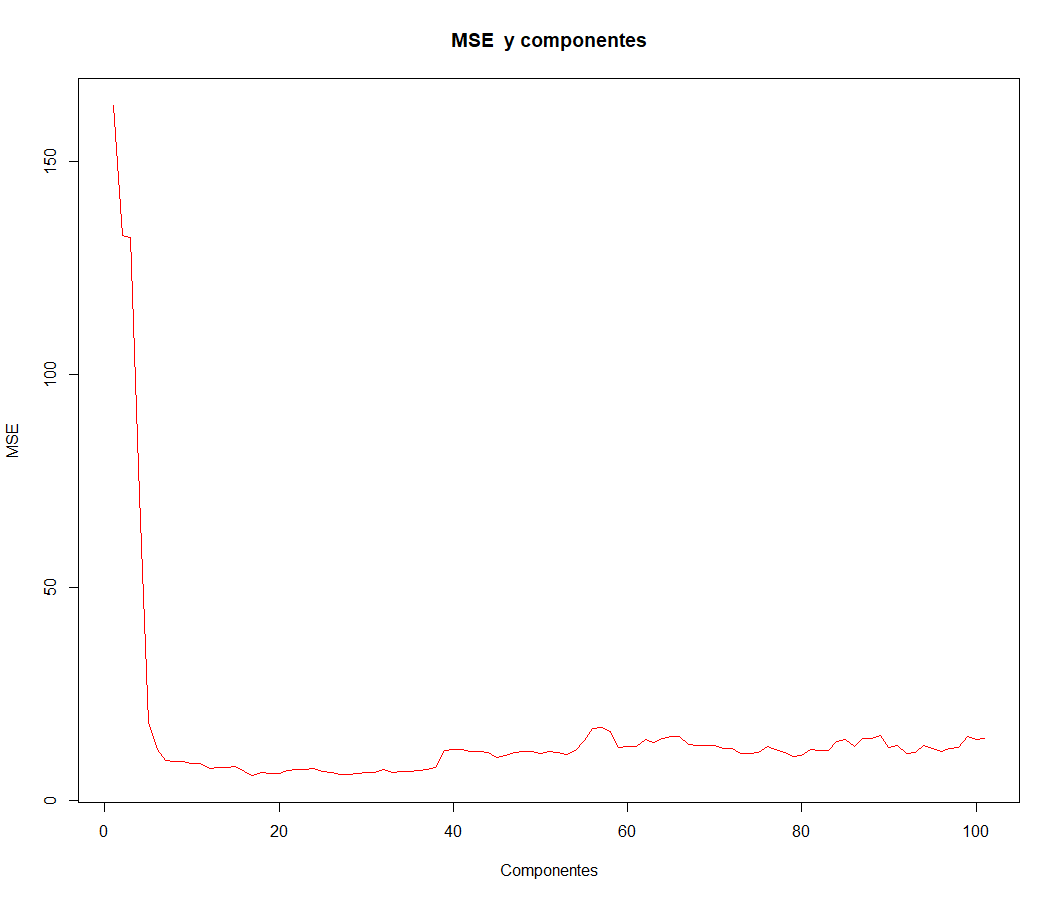
* + - 1. el número óptimo de componentes principales identificado por cross validation



* + - 1. el test-MSE



* + - 1. los gráficos correspondientes.



(puede realizar otros aspectos no solicitados en esta pregunta que será tomado en cuenta en la evaluación)

https://www.cienciadedatos.net/documentos/35\_principal\_component\_analysis