Universidad Técnica Federico Santa María
Departamento de Industrias
Econometría ICN312

“MODELO ECONOMÉTRICO PARA ESTABLECER LA RELACIÓN ENTRE LOS CARACTERES BIOLOGICOS Y GEOGRAFICOS DE LOS PINGÜINOS ADELIE

Presentado a: Prof. Mauricio Ahumada
Ayu. Amanda García

Desarrollado por: Daniel Inostroza
Rol: 20186016-2
Marco Subercaseaux
Rol: 201967061-1
Sebastián Zúñiga
Rol:201967069-7

Fecha: 31 de mayo de 2021

Contenido

1. RESUMEN EJECUTIVO

El trabajo evidencia, de forma explicativa y numérica la relación entre los caracteres biológicos y geográficos de los pingüinos Adelie. Se considere la masa corporal de la especie como variable explicativa, relacionándola con su morfología, sexo y habitad.

Para la conformación del análisis se utilizaron diversas técnicas econométrica:

1) Análisis exploratorio
2) Homocedasticidad o heterocedasticidad
3) Test de hipótesis
4) Análisis de datos atípicos e influyentes
5) Criterios de construcción de modelos

Se demuestra que según las herramientas señaladas, las islas no son significativas para el modelo. El modelo que explica de mejor manera la relación del peso de los pingüinos Adelie es Lm(Y~X1+X2+X3+D1).

2. ANTECEDENTES

El cambio climático trajo muchos efectos externos negativos y consecuencias catastróficas para la humanidad, pero sus efectos también existen en la flora y la fauna. Las aves del mundo son uno de los animales más gravemente afectados, especialmente los pingüinos. Según un estudio de SEO / Birdlife, su población pudiera reducirse hasta en un 50%, por lo que, debido a los nuevos desafíos que enfrenta esta especie, es muy importante comprender los aspectos filológicos que desarrolla cada una de sus subespecies y sus entorno geográfico en general. (BirdLife, 2018)

Para este proyecto, se examina la especie de pingüino “Adelie”, la cual es una de las dos únicas especies que viven en el continente antártico, distribuidas principalmente en las regiones circumpolar, océanos del sur e islas cercanas. Su población estimada es de 27 millones, lo que la convierte en la especie de pingüino más común en la península. Además, cabe mencionar que la Antártida es particularmente vulnerable al cambio climático, por lo que es importante estudiar en profundidad el peso corporal y la fisiología de la especie.

Se visualiza oportuno explicar la relación entre el peso, fisiología y habitad, a través de un **modelo econométrico** que explique la relación entre la **morfología de los pingüinos Adelie y los factores geográficos,** respecto a las condiciones de caracterización y hábitat.

Por otro lado, se considerará un análisis exploratorio para visualizar el comportamiento de los datos y determinar la existencia de correlación con las variables explicativas para completar la construcción del modelo.

La Dra. Kristen Gorman, miembro de la Red de Investigación Ecológica a Largo Plazo, y la Estación Palmer, Antártida LTER , Antártida, han recopilado y proporcionado estos datos. (Gorman, 2020)

3. VARIABLES

3.1. Variables consideradas en el modelo

Para abordar el modelo y de acuerdo con la investigación realizada, Adelie es la especie de pingüino más pequeña existente, con un peso de entre 4 y 5.5 kilogramos, según el primer censo realizado en 1974, se observo una reducción de 80% de su población. Eso hace considerar la existencia de factores relacionados entre la fisiología, geografía y peso que pudieran ser de incidencia al modelo. A partir de la evaluación de un modelo de regresión lineal se pretende definir la variable explicada o endógena para explicar el fenómeno, considerando una muestra de 147 pingüinos Adelie, sub detallado por sexo, geografía a la cual pertenecen, características físicas y fisiológicas.

3.1.1. Variable explicada

La variable explicada, es la que buscará ser explicada a través de distintas variables con ciertos grados de relación, que aportarán a hacer más precisa la regresión.

Y: Masa corporal de la especie [gramos]

Considerando como categorías base sexo femenino y la Isla Biscoe.

3.1.2. Variables explicativas

Se han definido las siguientes variables explicativas, las cuales permitirán explicar la relación entre el peso del pingüino y los datos de caracterización biológica obtenidos:

X_1: Longitud culmen[1] (mm)
X_2: Profundidad culmen[2] (mm)
X_3: Longitud de la aleta (mm)
D_1: Sexo (macho o hembra)
D_2: Isla (1 si es Torgersen y 0 si no lo es)
D_3: Isla (1 si es Dream y 0 si no lo es)

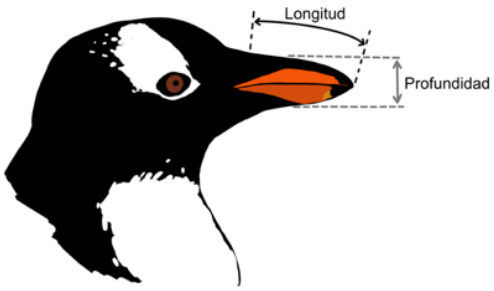


Ilustración 1: Morfología pingüinos

[1] Largo del vértice superior de la maxila (pico)
[2] Ancho del vértice superior de la maxila (pico)

4. METODOS

4.1. Herramientas de Software

1. RStudio: Software diseñado para hacer análisis estadístico y grafico.

2. Microsoft Excel: Herramienta de análisis, que permite encontrar estadísticas de regresión y análisis de varianza, entre otras disposiciones de los componentes.

4.2. Métodos econométricos

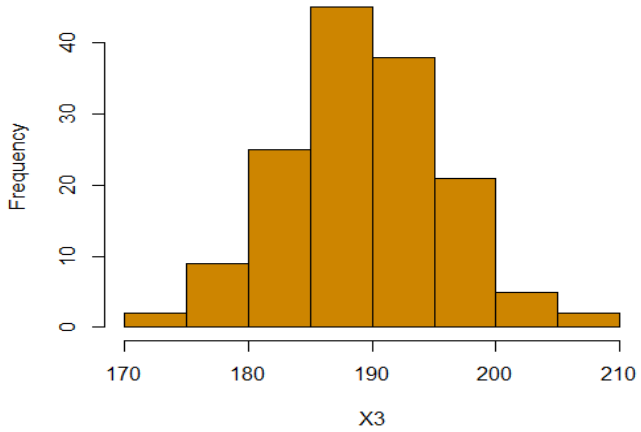
Para trabajar los datos se utiliza un modelo de regresión lineal múltiple en presencia de variables cuantitativas y cualitativas, las cuales permiten predecir el comportamiento de la masa de los pingüinos, respectivos a los datos morfológicos y geográficos.

1. A lo anterior se incluye el análisis con prueba de hipótesis, dócimas individuales, especificación, estimación, validación y análisis de corrección del modelo. Obtención de la matriz de correlaciones.

2. Análisis exploratorio de cada variable exógena.
    a) Correlaciones con la variable explicada y con las explicativas.
    b) Indicadores de tendencia central y de dispersión.
    c) Valores extremos.

5. Resultados: Análisis exploratorio

En el Histograma 1: "Longitud de la aleta", podemos ver la distribución de la longitud de la aleta en la muestra observada, en donde se logra visualizar una tendencia central que varía 185 mm y 195 mm.



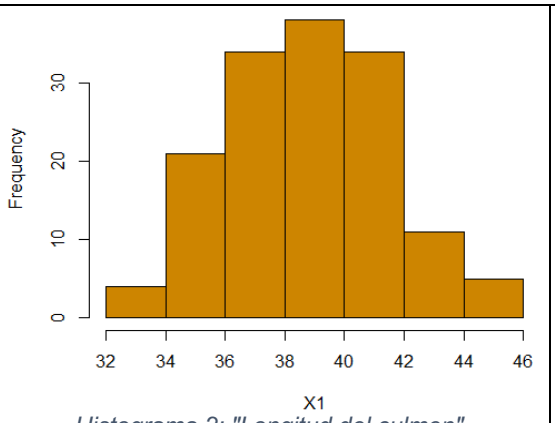
Histograma 1: "Longitud de la aleta"

A continuación, dada la Tabla 1: "Pingüinos por isla" correspondiente al área geográfica de la muestra (isla), se observa que la colonia de pingüinos se concentra mayoritariamente en Isla Dream, seguida de Isla Torgersen e Isla Biscoe.

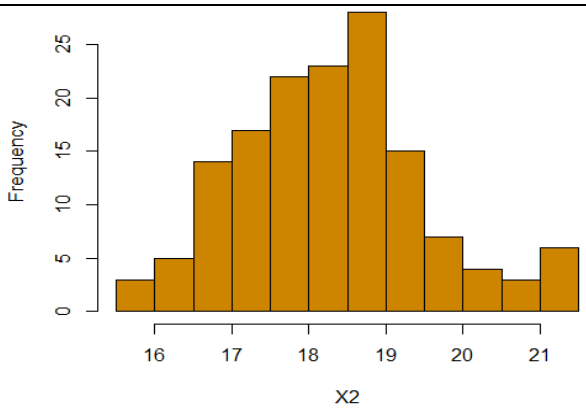
Especies por ubicación Isla	Total
Biscoe	44
Dream	56
Torgersen	51
Total general	151

Tabla 1: "Pingüinos por isla"

Por otro lado, el histograma relacionado con el culmen del pingüino (la placa central superior de la mandíbula superior) muestra que, en la mayoría de los casos, la longitud se concentra en los valores de 34 y 42 mm, alcanzando un máximo de 46 mm y un mínimo de 32 mm. La profundidad es concentrada en el valor entre 15 y 19 mm, alcanzando un máximo de 21,5 mm y un mínimo de 15,5 mm, todos los cuales se muestran en el Histograma 2: "Longitud del culmen" y el Histograma 3: "Profundidad del culmen".

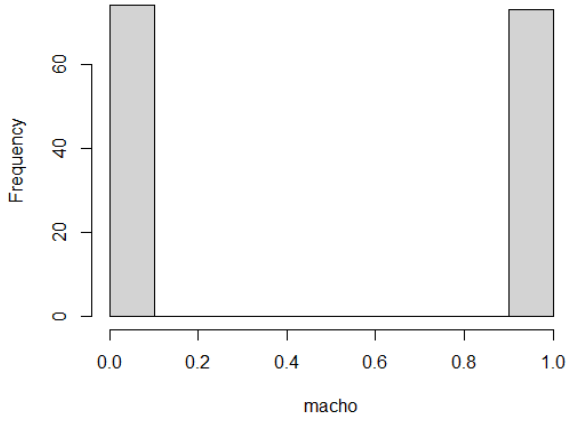


Histograma 2: "Longitud del culmen"



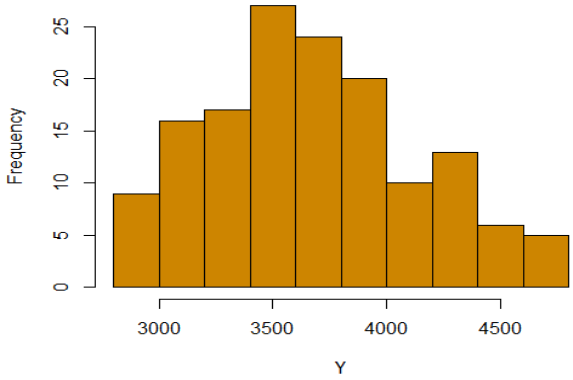
Histograma 3: "Profundidad del culmen"

Al revisar el Histograma 4: "Sexo de los pingüinos" se aprecia una distribución igualitaria de la población entre machos y hembras, 50% para cada uno.**¡Error! No se encuentra el origen de la referencia.**



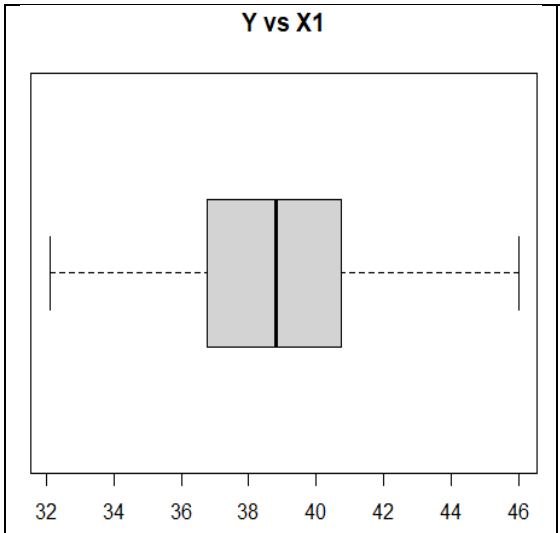
Histograma 4: "Sexo de los pingüinos"

En el Histograma 5: "Peso [gramos]" (variable explicativa) correspondiente al peso del pingüino, se puede observar que los datos se concentran entre 3000 y 4000 gramos, alcanzando un límite superior de 6300 gr y un límite inferior es de 2700 gr.

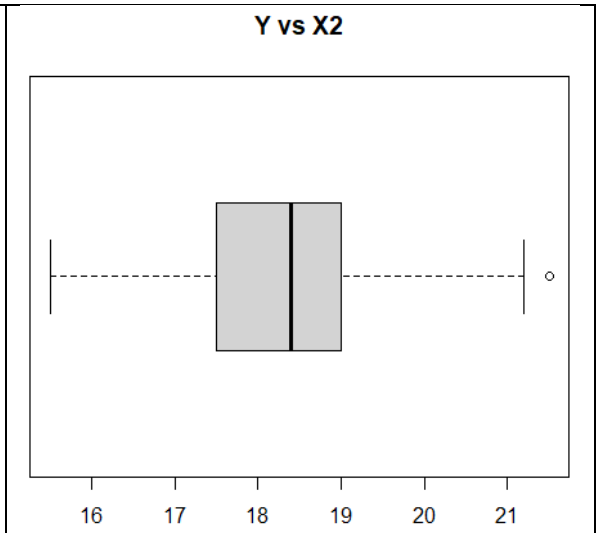


Histograma 5: "Peso [gramos]"

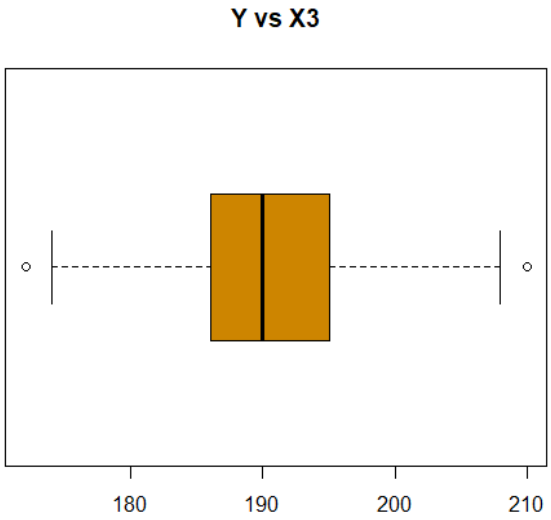
El método Box-Plot considera un método gráfico con las mismas variables, en donde se identifica la existencia de datos atípicos de forma descriptiva, estos datos muestran el comportamiento de las variables.



Box-plot 1: "Masa corporal vs longitud culmen"



Box-plot 2: "Masa corporal vs profundidad culmen"



Box-plot 3: "Masa corporal vs longitud de la aleta"

El Gráfico 1: "Longitud y profundidad del culmen", muestra una relación proporcional entre las variables X1 y X2. Es de suma importancia práctica, dado que a mayor longitud del culmen mayor es mayor la profundidad de este.

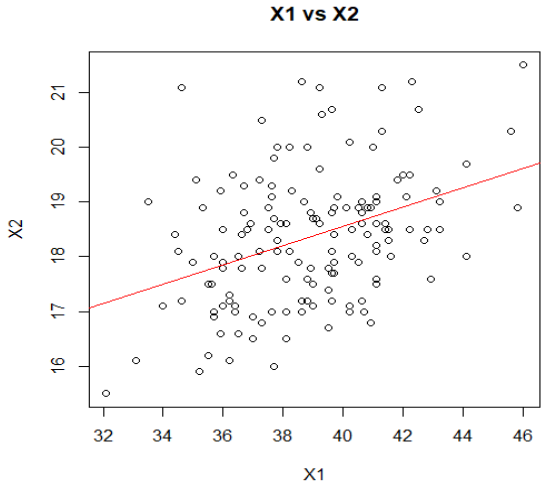


Gráfico 1: "Longitud y profundidad del culmen"

```
> m4=lm(X2~X1)
> plot(X1,X2, main = "X1 vs X2")
> abline(m4, col="red")
> cor.test(X1,X2)

        Pearson's product-moment correlation

data:  X1 and X2
t = 5.0029, df = 145, p-value = 1.611e-06
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.2364691 0.5136619
sample estimates:
      cor
0.3836743
```

Cálculo 1: "Correlación X1 vs X2"

El Gráfico 2: "Longitud del culmen y longitud de la aleta" muestra una relación proporcional del tipo lineal positiva entre X1 y X3, Se entiende que a mayor longitud del culmen, mayor es la longitud de aleta.

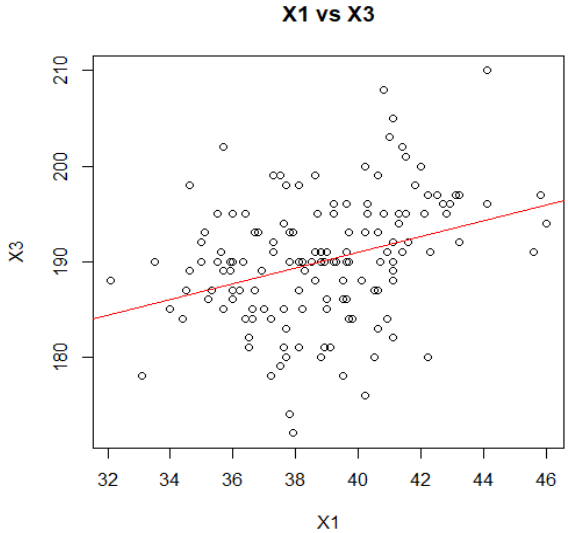


Gráfico 2: "Longitud del culmen y longitud de la aleta"

```
> m5=lm(X3~X1)
> plot(X1,X3, main = "X1 vs X3")
> abline(m5, col="red")
> cor.test(X1,X3)

        Pearson's product-moment correlation

data:  X1 and X3
t = 4.2739, df = 145, p-value = 3.463e-05
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1824703 0.4708777
sample estimates:
      cor
0.3344827
```

Cálculo 2: "Correlación X1 vs X3"

Al igual que los casos anteriores, se muestra que las variables explicativas X2 y X3 tienen una relación lineal directamente proporcional positiva, donde a mayor profundidad del culmen, mayor longitud de la aleta, lo anterior se puede apreciar en la siguiente gráfica:

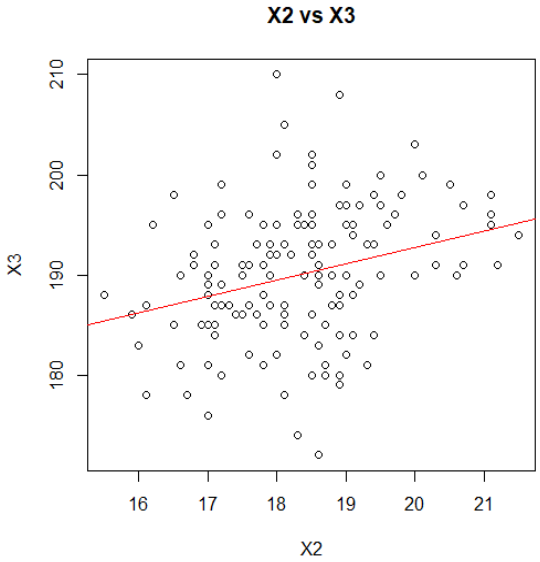


Gráfico 3: "Profundidad del culmen y longitud de la aleta"

```
> m6=lm(X3~X2)
> plot(X2,X3, main = "X2 vs X3")
> abline(m6, col="red")
> cor.test(X2,X3)

        Pearson's product-moment correlation

data:  X2 and X3
t = 3.8204, df = 145, p-value = 0.000197
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.1477500 0.4426325
sample estimates:
      cor
0.3024097
```

Cálculo 3: "Correlación X2 vs X3"

Se observa una relación proporcional directa lineal positiva entre el peso del pingüino (Y) y la longitud del culmen (X1), de esto se interpretar que, a mayor longitud del culmen, el pingüino tiende a tener un mayor masa corporal.

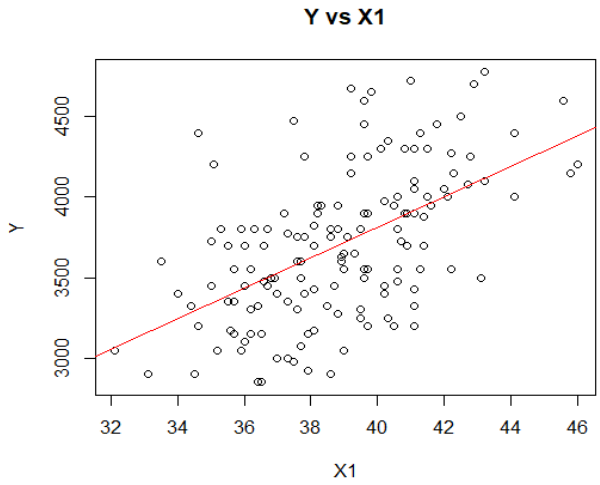


Gráfico 4: "Masa y longitud del culmen"

En el Gráfico 5: "Masa y profundidad del culmen" relaciona la masa del pingüino, respecto a la profundidad del culmen, en la que se muestra una línea de tendencia central positiva ascendente, lo que afirma que, a mayor tamaño, mayor profundidad del culmen.

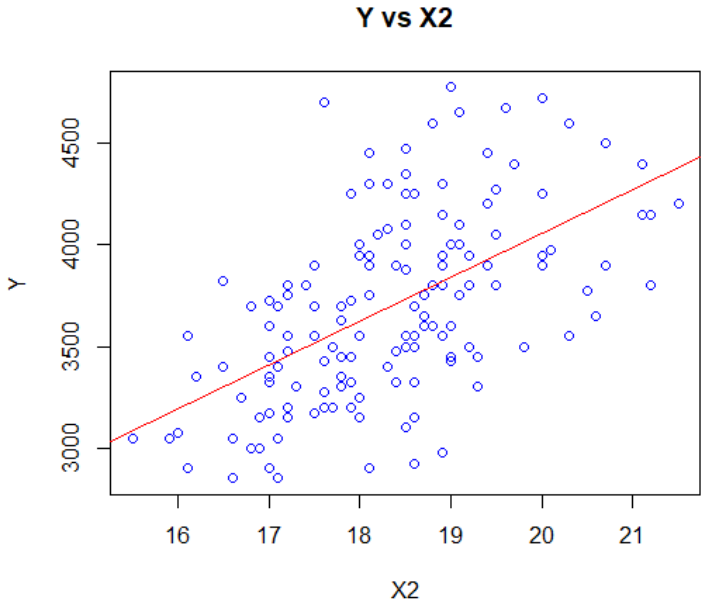


Gráfico 5: "Masa y profundidad del culmen"

Para el caso Y vs X3, la comparación entre la masa y la longitud de la aleta, el análisis arroja proporcionalidad positiva, por lo que se traza una lógica natural de que cuanto mayor es la longitud de la aleta, mayor es la masa del pingüino.

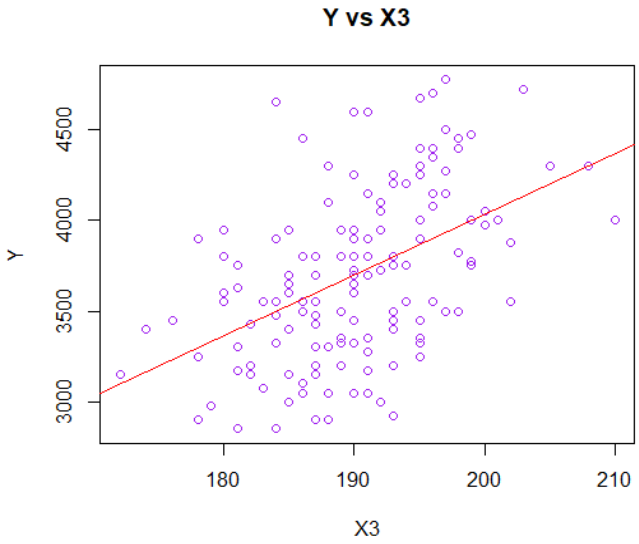


Gráfico 6: "Masa y longitud de la aleta"

En conclusión, al aumentar las condiciones morfológicas del pingüino se puede deducir preliminarmente que se relacionan de manera lineal con el peso (masa) en cierta proporción.

5.1. Matriz de correlación

A continuación, se muestra la correlación que existe entre todas las variables. Se destacan D2 vs D3, ya que presentan una alta correlación.

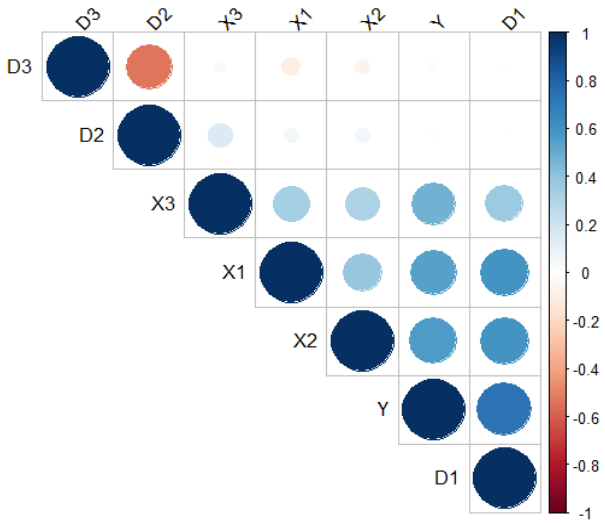


Tabla 2: "Matriz de correlación"

5.2. Criterios de multicolinealidad

Para corroborar la existencia de colinealidad múltiple entre las variables se utiliza el VIF, este valor debe ser menor a 5 lo que indica que no existe multicolinealidad, para el caso

de las variables utilizadas se concluye que todas cumplen con el requisito, como se muestra a continuación para el modelo full lm Y~X1+X2+X3+D1+D2+D3:

```
> library(carData)
> library(car)
> vif(modelo)
      X1       X2       X3       D1       D2       D3
1.599968 1.570882 1.234321 2.091083 1.454732 1.431725
```

Cálculo 4: "VIF"

5.3. Criterios de homocedasticidad o heterocedasticidad

Para corroborar si los datos del modelo full (lm Y~X1+X2+X3+D1+D2+D3) y ajustado preliminarmente (lm Y~X1+X2+X3+D1) son homocedasticos o heterocedasticos se aplican los siguientes criterios:

5.3.1. Prueba Goldfeld-Quandt (Solo modelo full)

```
#Goldfeld-Quandt
#ordenar segun X1
datos <- Datos_Penguins[with(Datos_Penguins, order(Datos_Penguins$X1)), ]
#c=49 c<1/3*n
cjo1 <- datos[-c(49:147), ]
cjo2 <- datos[-c(1:98), ]

model1 = lm(Y~X1+X2+X3+D1+D2+D3, data=cjo1)
model2 = lm(Y~X1+X2+X3+D1+D2+D3, data=cjo2)
anova(model1)
summary(model1)

anova1 =anova(model1)
anova2 =anova(model2)

CME1 = anova1$`Mean Sq`[7]
CME2 = anova2$`Mean Sq`[7]

model1$df.residual
#F(n1-p-1)=F(167-3-1) ##GRADOS DE LIBERTAD## EN ESTE CASO = 41
E = CME2/CME1
qf(0.95,model1$df.residual,model2$df.residual)#valor de tabla
pf(E, model2$df.residual,model1$df.residual,lower.tail=FALSE) #valores p
# E>V.tabla se rechaza ho
## dado que E=1.3979<1.674758 se ACEPTA h0, osea hay homocedasticidad segun la prueba de Goldfeld - Quandt
```

Cálculo 5: "Prueba de Goldfeld-Quandt"

```
> E
[1] 1.39797
> qf(0.95,model1$df.residual,model2$df.residual)#valor de tabla
[1] 1.674758
```

Se concluye que las variables son homocedasticas para el modelo full dados la siguiente dócima:

h0: hay homocedasticidad / no hay heterocedasticidad
ha: no hay homocedasticidad / hay heterocedasticidad
IEI > Valor tabla se rechaza ho

Ya que, 1.39797(E) < 1.674758 (valor tabla) se acepta h0 y la dócima no es significativa.

5.3.2. Prueba de White (Modelo full y ajustado)

```
#white
Err<-resid(m1)
Aux<-lm(I(Err^2)~X1+X2+X3+D1+D2+D3+I(X1^2)+I(X2^2)+I(X3^2)+I(D1^2)+I(D2^2)+I(D3^2)+
          I(X1*X2)+I(X1*X3)+I(X1*D1)+I(X1*D2)+I(X1*D3)+
          I(X2*X3)+I(X2*D1)+I(X2*D2)+I(X2*D3)+
          I(X3*D1)+I(X3*D2)+I(X3*D3)+
          I(D1*D2)+I(D1*D3)+
          I(D2*D3)  ) #27
summary(Aux)
Aux1<-lm(I(Err^2)~X1+X2+X3+D1+I(X1^2)+I(X2^2)+I(X3^2)+
           I(X1*X2)+I(X1*X3)+I(X1*D1)+
           I(X2*X3)+I(X2*D1)+
           I(X3*D1)) #13
summary(Aux1)


Raux <- summary(Aux)$r.squared
E<-147*Raux
E
Raux1 <- summary(Aux1)$r.squared
E1<-147*Raux1

#vatb= chi(0,95,p*(p+3)/2)
Vtab1<-qchisq(0.95,5*(3+5)/2)#malo, YA NO NOS SIRVE POR TENER VARIABLES CUALITATIVAS
Vtab<-qchisq(0.95,27)#correcto
Vtab
E
## el estadistico es menor que el valor de tabla (28.57<40.1132)
##se acepta h0 y hay homocedasticidad

E1
Vtab1<-qchisq(0.95,13)#correcto
Vtab1
## el estadistico es menor que el valor de tabla (11.49<22.3620)
##se acepta h0 y hay homocedasticidad
```

Cálculo 6: "Prueba de White"

Para modelo full

```
> Vtab
[1] 40.11327
> E
[1] 28.57526
```

Se concluye que las variables son homocedasticas para el modelo full dadas la siguiente dócimas:

h0: hay homocedasticidad/ no hay heterocedasticidad
ha: no hay homocedasticidad/ hay heterocedasticidad
IEI > Valor tabla se rechaza ho

Ya que, 28.57 (E) < 40.11 (valor tabla) se acepta h0 y la dócima no es significativa.

Para modelo ajustado

```
> Vtab1
[1] 22.36203
> E1
[1] 11.4966
```

Se concluye que las variables son homocedasticas para el modelo ajustado dadas la siguientes dócimas:

ho: hay homocedasticidad / no hay heterocedasticidad
ha: no hay homocedasticidad / hay heterocedasticidad
IEI > Valor tabla se rechaza ho

Ya que, 11.49 (E) < 22.36 (Valor tabla) se acepta h0 y la docima no es significativa.

5.4. Datos atípicos e influyentes

En el análisis realizado a los datos sobre el modelo full (lm Y~X1+X2+X3+D1+D2+D3) fueron encontrados 11 datos atípicos mediante el análisis de los residuales estudentizados, los cuales se muestran a continuación:

```
#datos atipicos
round(head(rstandard(modelo),n=147),d=4)  #residuales estudentizados y 4 es que sean 4 decimales
which(abs(rstandard(modelo))>2) #cuales son los atipicos (MAYORES A 2)
sum(abs(rstandard(modelo))>2) #cantidad total de atipicos
###de 147 datos existen 11 atipicos que corresponen a los numeros  7  35  41  77  81  89 100 105 115 117 127
> which(abs(rstandard(modelo))>2) #cuales son los atipicos (MAYORES A 2)
  7  35  41  77  81  89 100 105 115 117 127
  7  35  41  77  81  89 100 105 115 117 127
> sum(abs(rstandard(modelo))>2) #cantidad total de atipicos
[1] 11
```

Cálculo 7: "Datos atípicos"

En el análisis del apalancamiento del modelo full (lm Y~X1+X2+X3+D1+D2+D3) fueron detectados 4 datos influyentes que corresponden a los números 10, 15, 110 y 125, según el criterio de leverage:

```
#datos influyentes
#leverage
round(head(hatvalues(modelo),n=147),d=4)
hat=abs(hatvalues(modelo))
which(hat>(2*((ncol(Datos_Penguins)-1)+1)/nrow(Datos_Penguins))) #cuales son influyentes
sum(hat>(2*((ncol(Datos_Penguins)-1)+1)/nrow(Datos_Penguins)))
###de 147 datos existen 4 influyentes que corresponen a los numeros  10  15 110 125

> which(hat>(2*((ncol(Datos_Penguins)-1)+1)/nrow(Datos_Penguins))) #cuales son influyentes
 10  15 110 125
 10  15 110 125
> sum(hat>(2*((ncol(Datos_Penguins)-1)+1)/nrow(Datos_Penguins)))
[1] 4
```

Cálculo 8: "Datos influyentes"

5.5. Distancia de Cook's

En el Gráfico 7: "Distancia de Cook", muestra 6 valores influyentes, estos datos son las observaciones 35, 77, 100, 105, 117 y 127, los cuales deben ser considerados para decidir su eliminación dentro del modelo.

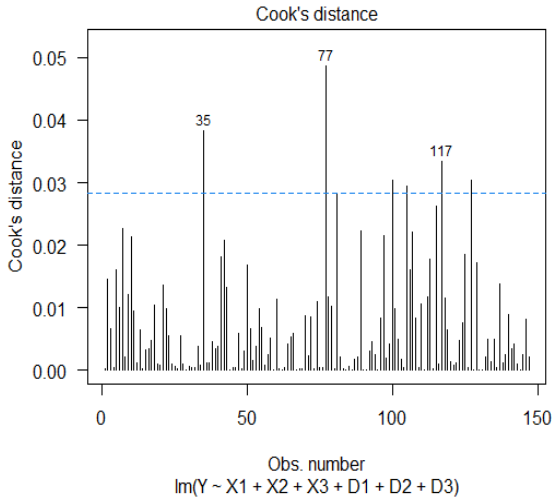


Gráfico 7: "Distancia de Cook"

Al realizar la eliminación de los datos influyentes y evaluar las observaciones mediante la distancia de Cook, los resultados obtenidos son los mostrados en Gráfico 8: "Distancia de Cook corregida", la que cuenta con 4 datos influyentes, ubicados en la posición 35, 77,1 17 y 127.

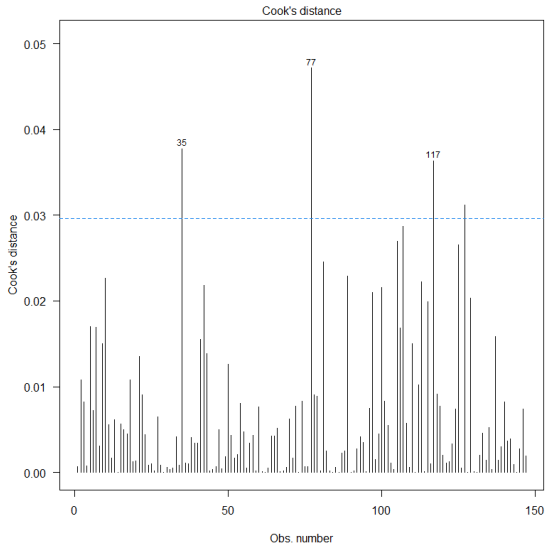


Gráfico 8: "Distancia de Cook corregida"

Se concluye que para el modelo full con datos corregidos siguen existiendo datos influyentes, esta situación se pudiera repetir al eliminar los 4 datos señalados y calcular nuevamente.

6. Criterios de construcción de los modelos

6.1. ANDEVA

En la construcción de la tabla ANDEVA, se pudo determinar que las variables D2 y D3 no son significativas.

```
Response: Y
           Df   Sum Sq Mean Sq  F value    Pr(>F)
X1          1  9200279 9200279 111.0577 < 2.2e-16
X2          1  4740274 4740274  57.2205 4.698e-12
X3          1  1742178 1742178  21.0301 9.936e-06
D1          1  3706229 3706229  44.7384 4.976e-10
D2          1    34596   34596   0.4176    0.5192
D3          1     7688    7688   0.0928    0.7611
Residuals 140 11597921   82842
```

Tabla 3: "ANDEVA"

6.2. Significancia de las variables

A continuación, se presenta la probabilidad de significancia estadística de la relación de las variables, con el fin de comprender cómo influyen en los resultados.

Hipótesis	
h0: bi=0	
ha: bi=!0	
Se rechaza ho cuando |E|>v.tabla o p<0,05	

Tabla 4: "Test de hipótesis"

- X1: Se rechaza H0, X1 si es significativo.
- X2: Se rechaza H0, X2 si es significativo.
- X3: Se rechaza H0, X3 si es significativo.
- D1: Se rechaza H0, D1 si es significativo.
- D2: Se acepta H0, D2 no es significativo.
- D3: Se acepta H0, D3 no es significativo.

6.3. Modelo paso a paso, ascendente

```
Call:
lm(formula = Y ~ D1 + X3 + X2 + X1)

Coefficients:
(Intercept)           D1           X3           X2           X1
   -1225.45       455.94        14.35        64.23        20.45
```

Cálculo 9: "Modelo paso a paso, ascendente"

6.4. Modelo paso a paso, descendente

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + D1, data = Datos_Penguins)

Coefficients:
(Intercept)           X1           X2           X3           D1
   -1225.45        20.45        64.23        14.35       455.94
```

Cálculo 10: "Modelo paso a paso, descendente"

6.5. Método en ambas direcciones

```
Call:
lm(formula = Y ~ X1 + X2 + X3 + D1, data = Datos_Penguins)

Coefficients:
(Intercept)           X1           X2           X3           D1
   -1225.45        20.45        64.23        14.35       455.94
```

Cálculo 11: "Método en ambas direcciones"

6.6. Método de todas las regresiones posibles

A continuación, es presentada la tabla que contiene todos los modelos posibles, considerando las mejores variables que explican el modelo, en base a esto se aplica los criterios R^2 ajustado, CP Mallows y el Criterio de residuales.

```
> best.subset <- regsubsets(Y~X1+X2+X3+D1+D2+D3,data=BDP,nbest = 4, nvmax=NULL)
> summary.out <-summary(best.subset)
> as.data.frame(summary.out$outmat)
         X1 X2 X3 D1 D2 D3
1  ( 1 )            *
1  ( 2 )     *
1  ( 3 )  *
1  ( 4 )        *
2  ( 1 )        *  *
2  ( 2 )     *     *
2  ( 3 )  *        *
2  ( 4 )           *     *
3  ( 1 )     *  *  *
3  ( 2 )  *     *  *
3  ( 3 )        *  *  *
3  ( 4 )        *  *     *
4  ( 1 )  *  *  *  *
4  ( 2 )     *  *  *  *
4  ( 3 )     *  *  *     *
4  ( 4 )  *     *  *  *
5  ( 1 )  *  *  *  *  *
5  ( 2 )  *  *  *  *     *
5  ( 3 )     *  *  *  *  *
5  ( 4 )  *     *  *  *  *
6  ( 1 )  *  *  *  *  *  *
> ##
```

Cálculo 12: "Método de todas las regresiones posibles"

Para el criterio de selección, se determina que el modelo con las 4 mejores variables corresponde al modelo ajustado (lm Y~X1+X2+X3+D1).

6.6.1. Criterio R^2 ajustado

El modelo escogido al utilizar el análisis R^2 ajustado es Y= X1+X2+X3+D1 con una explicación del 61%, según se muestra a continuación:

```
> #r2ajustado
> summary.out$adjr2
 [1] 0.5444743 0.3200860 0.2916525 0.5911907 0.5685749 0.5591635 0.6080227 0.5982216 0.5887577 0.6142952 0.6061022
[12] 0.6052983 0.6127142 0.6115816 0.6039675 0.6102062
> which.max(summary.out$adjr2)
[1] 10
```

Cálculo 13: "Criterio R^2 ajustado"

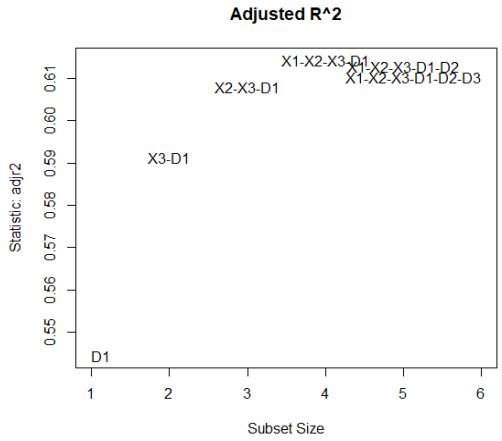


Gráfico 9: "Criterio R^2 ajustado"

6.6.2. Criterio CP Mallows

De acuerdo con el criterio del CP de Mallows, el modelo seleccionado es Y=X1+X2+X3+D1 al igual que en el criterio del R^2 ajustado.

```
> #cpmallows
> summary.out$cp
 [1]  26.451734 109.922292 120.499307  10.024834  18.379707  21.856514   4.801054   8.396690  11.868647   3.510419
[11]   6.495101   6.787947   5.092801   5.502498   8.256744   7.000000
> which.min(summary.out$cp)
[1] 10
```

Cálculo 14: "Criterio CP Mallow"

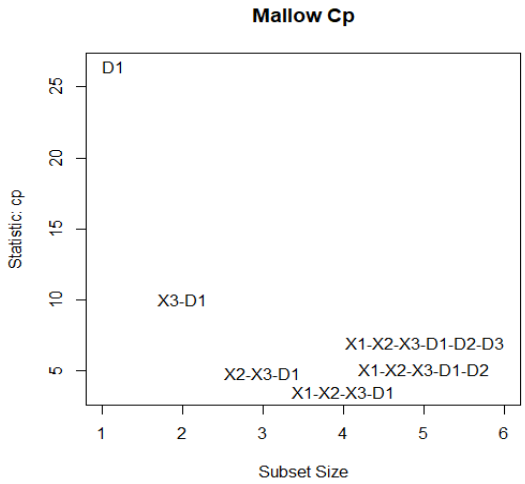


Gráfico 10: "Criterio CP Mallow"

6.6.3. Criterio de residuales

Al aplicar la suma cuadrado de los residuales el modelo seleccionado es el que contempla todas variables Y=X1+X2+X3+D1+D2+D3, según se muestra en la tabla de todas las regresiones posibles.

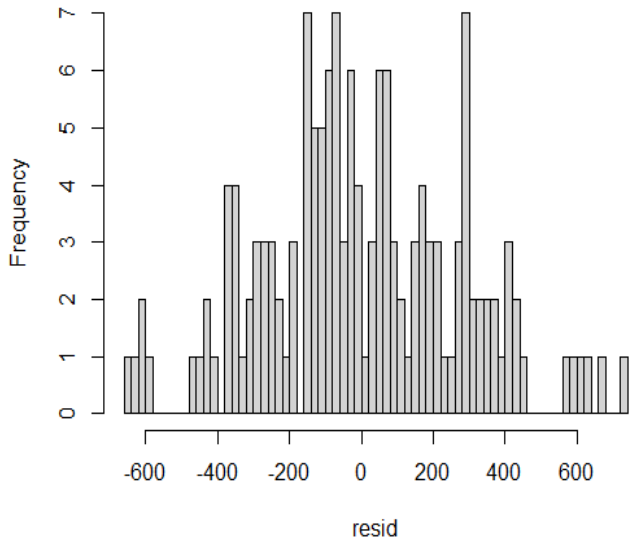
```
> #suma cuadrada de residuos
> summary.out$rss
 [1] 14037771 20952663 21828888 12511244 13203381 13491408 11912809 12210680 12498305 11640206 11887464 11911724
[13] 11605609 11639550 11867718 11597921
> which.min(summary.out$rss)
[1] 16
```

Cálculo 15: "Suma cuadrado de los residuos"

Según los datos anteriormente expuesto de la construcción de modelos, se define como el mejor modelo Y=X1+X2+X3+D1 ya que amplia mayoría de criterios (todas las regresiones posibles, Mallows, R^2 ajustado, modelos ascendente, descendente y mixto) este modelo fue seleccionado.

6.7. Normalidad del error

A continuación, se realiza un análisis de los residuales, donde se aprecia una concentración entre -450 y 450, se percibe que el centro es equidistante a cero, a primera vista son datos simétricos, similares a una campana gaussiana excepto por algunos puntos exclusivos.



Histograma 6: "Residuales"

6.8. Test de Jarque-Bera

Dado lo que se observa en el Histograma 6: "Residuales" se aplica el test de Jarque-Bera, en donde se obtiene un valor de 0.438, que es mejor a los 5.991 correspondientes al Chi-cuadrado,por tanto, se afirma que el error se comporta en forma normal.

```
Title:
 Jarque - Bera Normalality Test

Test Results:
  STATISTIC:
    X-squared: 0.4386
  P VALUE:
    Asymptotic p Value: 0.8031
```

Cálculo 16: "Test de Jarque-Bera"

```
> #valor del chi-cuadrado
> qchisq(0.95,2)
[1] 5.991465
```

Cálculo 17: "Chi cuadrado Jarque-Bera"

ho: error se comporta de forma normal.
ha: error no se comporta de forma normal.
IEI > Valor tabla se rechaza ho

Se acepta ya que el valor del estadístico es menor 0.438 < 5.99

6.9. Modelo ajustado

Modelo Full
Lm(Y~X1+X2+X3+D1)

Bajo los criterios de todas las regresiones posibles, modelo paso a paso ascendente, descendente y en ambas direcciones, considerando además el análisis de los criterios de R^2 ajustado y CP de Mallows antes mencionados, se considera Y=X1+X2+X3+D1 como el mejor modelo que explica el peso en los pingüinos Adelie.

```
> summary(Majustado)

Call:
lm(formula = Y ~ X1 + X2 + X3 + D1)

Residuals:
    Min      1Q  Median      3Q     Max
-672.96 -160.42  -17.62  191.86  737.56

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1225.447    854.458  -1.434 0.153719
X1             20.450     11.214   1.824 0.070315 .
X2             64.230     24.348   2.638 0.009269 **
X3             14.355      3.948   3.636 0.000386 ***
D1            455.937     67.807   6.724 3.99e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 286.3 on 142 degrees of freedom
Multiple R-squared:  0.6249,    Adjusted R-squared:  0.6143
F-statistic: 59.13 on 4 and 142 DF,  p-value: < 2.2e-16
```

Cálculo 18: "Modelo ajustado"

7. CONCLUSIONES

En síntesis se logra concluir que el modelo ajustado obtenido por el modelo paso a paso ascendentes, descendente, en ambas direcciones y todas las regresiones posibles, corresponderá Y= X1 +X2+X3+D1, esto gracias a las medida de bondad encontradas.

Dicho modelo considera tres variable explicativas, longitud y profundidad del culmen, longitud de la aleta, y una cualitativa (sexo). Demostrando que dos de las variables explicadas D3 y D4 (islas) no eran influyentes sobre la variable Y (masa). Esto es aceptable, ya que, se percibe una proximidad geográfica entre las islas, lo cual no genera un cambio en las condiciones de habitabilidad y ecosistemicas en los pingüinos.

Al tratarse de un base de datos de corte transversal la variabilidad de la información es espacial y representa un momento determinado en el tiempo, representa un estudio observacional, descriptivo y analítico.

Alguna de las limitantes, es que puede presentar sesgos por omisión del observador, se pierde trazabilidad temporal de su morfología y depende del momento en cuando se tomaron los datos.

El modelo ajustado no presenta problemas de multicolinealidad, se desplaza la existencia de relaciones lineales entre 2 o más variables, existiendo una adecuada variabilidad en las observaciones independientes.

Se concluye que las variables expuestas son las que mejor explican el comportamiento del peso de los pingüinos.

8. BIBLIOGRAFÍA

BirdLife. (2018). Pingünos, testigos del cambio global. *Ave y naturaleza*, 58.
Gorman, D. K. (1 de Julio de 2020). *Kaggle*. Obtenido de Kaggle: https://www.kaggle.com/parulpandey/palmer-archipelago-antarctica-penguin-data

9. ANEXOS

Ver archivos adjuntos:

- Base de datos “Datos Penguins.xlsx”
- Análisis de datos en R “R Final”

Análisis de Datos Excel:

Macho	Categoria base	D1
Torgersen		D2
Biscoe	Categoria base	
Dream		D3
D1= 1 si es Macho y 0 si es mujer		
D2= 1 si es de la isla de Torgersen y 0 si no es		
D3= 1 si es de la isla de Dream		
Las variables son:		
Y=	Masa corporal	
	Explicadas:	
X1=	Longitud Culmen (mm)	
X2=	Profundidad Culmen (mm)	
X3=	Longitud de la aleta (mm)	

Estadísticas de la regresión	
Coeficiente de correlación múltiple	0.791343901
Coeficiente de determinación R^2	0.62622517
R^2 ajustado	0.610206249
Error típico	287.823377
Observaciones	147

	Grados de libertad	*Suma de cuadrados*	*Promedio de los cuadrados*	*F*	*Valor crítico de F*
Regresión	6	19431245.18	3238540.863	39.09284273	1.23318E-27
Residuos	140	11597921.49	82842.29636		
Total	146	31029166.67			

	Coeficientes	*Error típico o estandar*	*Estadístico t o IEI --- Docima individual*	*Probabilidad o Valor P*	*Inferior 95%*	*Superior 95%*
Intercepción	-1309.615656	872.1602594	-1.501576852	0.135458522	-3033.923341	414.6920285
X1	20.47466253	11.3455332	1.804645244	0.073279608	-1.956066082	42.90539114
X2	64.72310795	24.55185456	2.63618	0.009330816	16.18277256	113.2634433
X3	14.86298355	4.031982228	3.686272089	0.000324561	6.89153809	22.83442901
D1	452.4362537	68.65829742	6.589680647	8.29185E-10	316.6951114	588.177396
D2	-43.52055211	61.39418507	-0.708870914	0.479583334	-164.9001541	77.85904986
D3	-17.81869836	58.49244956	-0.304632453	0.761098475	-133.4614133	97.8240166

Hipótesis	
h0: bi=0	
ha: bi=!0	
Se rechaza ho cuando |E|>v.tabla o p<0,05	
v.tabla	1.97705372