Entrega final - Determinación del peso de los peces

Fabricio Camacho,Matias Bajac

2024-06-25

Introducción

A lo largo de este trabajo vamos a interpretar y modelar datos referentes a las dimensiones de peces de la costa de Finlandia. Las variables con las que contamos son su peso, longitud, ancho y altura, además de la especie de cada pez.

El objetivo final es poder predecir el peso de los peces mediante las restantes variables, estas son:

Variable	Descripción
Especie	Bream, Parki, Perch, Pike, Roach, Smelt, Whitefish
Peso_gr	Peso del pez en gramos
Altura_cm	Altura en centímetros
Ancho_cm	Ancho en centímetros
Longitud1	Desde la nariz al comienzo de la cola
Longitud2	Desde la punta de la nariz hasta la muesca de la cola
Longitud3	Desde la nariz al final de la cola

La base de datos cuenta con 159 peces donde uno de ellos será quitado de la misma por tenér un 0 en la varialbe de peso, consiguiendo finalmente un total de 158 peces.

A efectos de tener un primer acercamiento con la estructura de los datos, se obtienen algunas estadisticas descripitvas, como la correlacion entre las variables cuantitativas y un diagrama de caja para visualizar el peso en relacion a cada especie

Metodologia

La idea es implementar las tecnicas de analisis estudiadas en el curso de Modelos Lineales, en partiuclar, el modelo de regresion multiple.

En una primera instancia, se procede a hacer un analisis explotatorio de los datos. Luego pasamos a una primera etapa de diagnostico, dado que la intención es poder inferir en una generalidad de peces, hay ciertos supuestos que tenemos que validar , estos son:

- Multicolinealidad: Donde nos va a interesar que ninguna variable sea combinación lineal del resto.
- Homoscedasticidad: Donde la varianza de los residuos para cada pez son iguales.
- Normalidad: Donde los residuos de los estimados tienen una distribución normal.

Para estudiar el supuesto de la multicolinealidad aproximada, en el cual nos sirve para quedarnos con las variables explicativas siguiendo el criterio de Vif<5. Para la Homoscedasticidad vamos a aplicar el test de Bresuch-Pagan, y para la normalidad el test de Kolmogorov-Smirnov.

Este analísis diagnóstico es aplicado para cada modelo candidato a responder nuestras inquietudes de investigación a efectos de encontrar el mejor, o en otras palabras el que pueda explicar en mayor medida la varianza.

En una siguiente etapa, se hizo un analisis ANOVA y ANCOVA. Bajo la hi potesis de que hay modelos mejores que otros y variables que puedan explicar de mejor forma el peso de los peces, es que tendremos particular interés en ver como interactuan distintas variables, haciendolas complementarse entre sí.

Para finalizar, se usaron técnicas de cross validation para evalular todos los modelos.

A efectos de tener un primer acercamiento con la estructura de los datos, se obtienen algunas estadisticas descripitvas, como la correlacion entre las variables cuantitativas y un diagrama de caja para visualizar el peso en relación a cada especie

Resultados

Análisis exploratorio de los datos

	Especie	Peso_gr	Longitud1	Longitud2	Longitud3	Altura_cm	Ancho_cm
X	Length:159	Length:159	Min. : 7.50	Min. : 8.40	Min. : 8.80	Min. :14.50	Min. : 8.70
X.1	Class :character	Class :character	1st Qu.:19.05	1st Qu.:21.00	1st Qu.:23.15	1st Qu.:24.25	1st Qu.:13.4
X.2	Mode :character	Mode :character	Median :25.20	Median :27.30	Median :29.40	Median :27.10	Median :14.
X.3			Mean :26.25	Mean :28.42	Mean :31.23	Mean :28.31	Mean :14.1
X.4			3rd Qu.:32.70	3rd Qu.:35.50	3rd Qu.:39.65	3rd Qu.:37.60	3rd Qu.:15.
X.5			Max. :59.00	Max. :63.40	Max. :68.00	Max. :44.50	Max. :20.90

```
##    Especie            Peso_gr           Longitud1       Longitud2
##  Length:159         Length:159         Min.   : 7.50   Min.   : 8.40
##  Class :character   Class :character   1st Qu.:19.05   1st Qu.:21.00
##  Mode  :character   Mode  :character   Median :25.20   Median :27.30
##                                        Mean   :26.25   Mean   :28.42
##                                        3rd Qu.:32.70   3rd Qu.:35.50
##                                        Max.   :59.00   Max.   :63.40
##    Longitud3       Altura_cm        Ancho_cm
##  Min.   : 8.80   Min.   :14.50   Min.   : 8.70
##  1st Qu.:23.15   1st Qu.:24.25   1st Qu.:13.40
##  Median :29.40   Median :27.10   Median :14.60
##  Mean   :31.23   Mean   :28.31   Mean   :14.12
##  3rd Qu.:39.65   3rd Qu.:37.60   3rd Qu.:15.30
##  Max.   :68.00   Max.   :44.50   Max.   :20.90
```

Podemos observar mediante la matriz de correlacion que existe correlacion fuerte entre las variables Longitud de los peces, lo que podria casuar problemas de multicolinealidad aproximada, como tambien problemas en la homoscedasticidad de la varianza de los residuos. Parece razonable a priori incluir solo una variable respecto a la longitud del pez para predecir su peso.

Análisis gráfico

Análisis de supuestos sobre modelos lineales

Multicolinealidad

```
##   Longitud1   Longitud2   Longitud3   Altura_cm    Ancho_cm
## 1817.971125 2391.862631  315.886137    4.724397    2.253745
```

```
##  Longitud1  Longitud3  Altura_cm   Ancho_cm
```

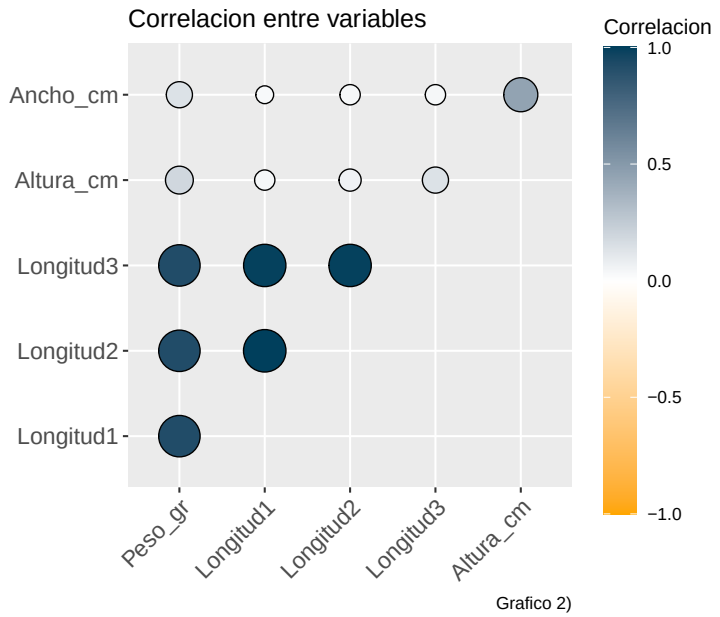


Figure 1: matriz de correlacion entre las variables cuantitativas.

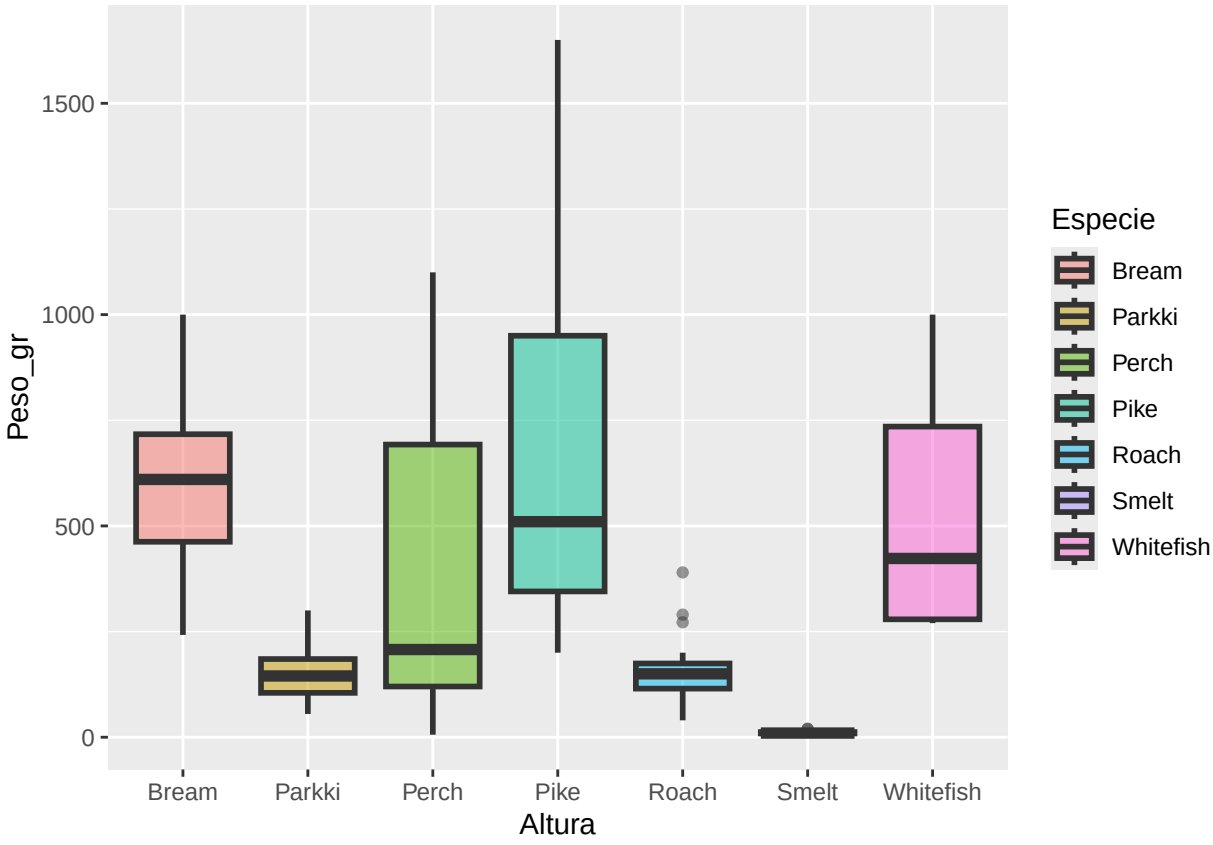


Figure 2: Diagrama de caja, relación entre cada especie y el peso

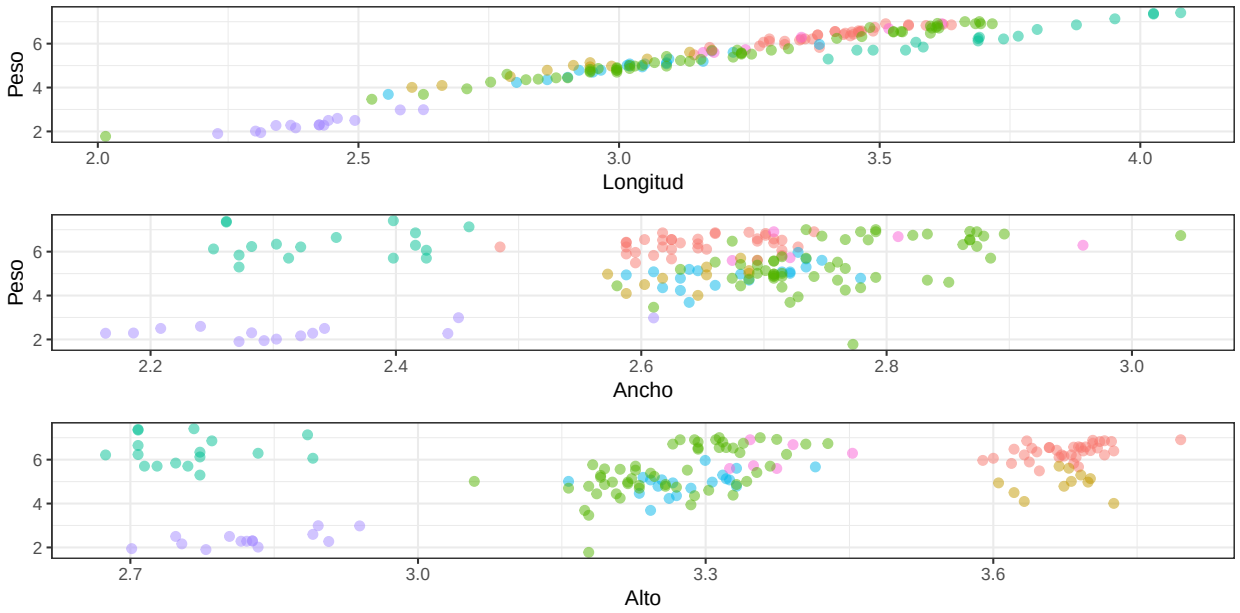


Figure 3: Diagrama de dispresion de los regresores respecto al peso

```
## 229.258810 233.347914   4.667798   1.812460
```

```
## Longitud1 Altura_cm  Ancho_cm
##  1.001651  1.258646  1.258136
```

En esta instancia analizamos la multicolinealidad, la idea es ver si hay variables que sean combinación lineal de otras, o sea, que en tengan la misma información. En caso de que el *vif* sea alto (mayor a 5), quitaremos la variable con el *vif* más alto. La presencia de multicolinealidad impide sobretodo la posibilidad de analizar el efecto de una variable predictora sobre lo que queremos predecir, en nuestro caso el peso.

En el primer paso de este análisis hallamos *vif* elevados en las distintas longitudes, aquí volvemos confirmar lo estudiado en el análisis de correlación previo, donde las longitudes están altamente correlacionadas, lo cual indica que contar todas las medidas es inviable.

$$VIF_j = \frac{1}{1 - R_j^2}$$

De esta forma las variables finales seran *Longitud1 , Altura_cm, Ancho_cm*

Modelos

Modelo 1 Como ya se menciono,el primer modelo estimado consiste en la regresión de la variable peso con las variables explicativas que fueron seleccionadas en el paso de multicolinealidad. El modelo queda esepecificado como:

$$peso_i = \beta_0 + \ \beta_1 Longitud1_i + \beta_2 Altura_i \ + \beta_3 Ancho_i \ + \ \epsilon_i$$

Diagnostico del modelo Homoscedasticidad

Aquí se opto por recurrir a un análisis visual de los residuos externamente estudientizados del modelo. A continuacion vemos el grafico de los residuos en el eje de las ordenadas. Con un $\alpha = 0.05$ rechazamos la

hipotesis nula, por lo que podemos afirmar que no hay homoscedasticidad con un $p - valor < 0.0001$.

$$H_0)\ E(\epsilon_i^2) = \sigma^2\ vs\ H_1)\ no\ H_0$$

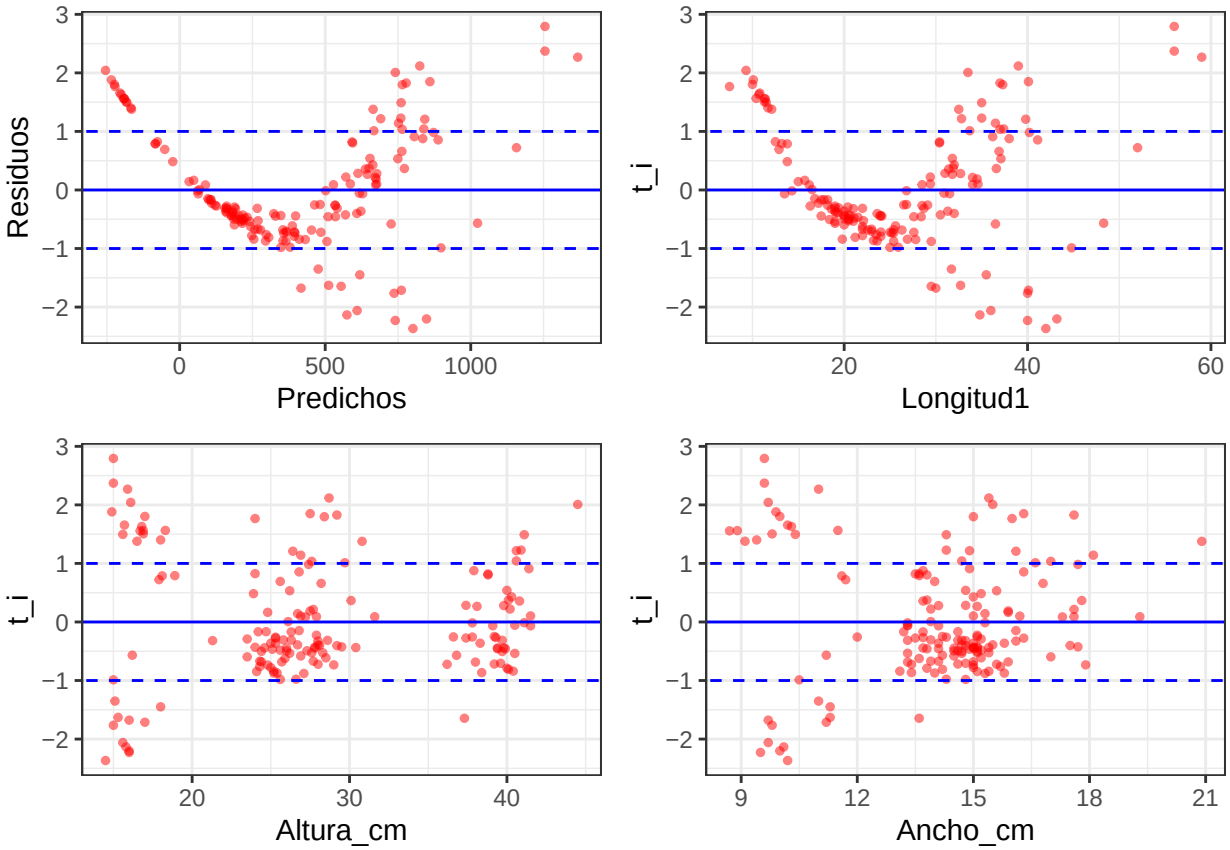


Figure 4: Análisis de los residuos externamente studientizados del modelo 1

```
## # A tibble: 1 x 5
##   statistic  p.value parameter method                 alternative
##       <dbl>    <dbl>     <dbl> <chr>                  <chr>
## 1      81.0 1.87e-17         3 Koenker (studentised) greater
```

Normalidad

El histograma de los residuos externamente estudentizados no se parece a una distribucion normal en los residuos. Ademas, el test de normalidad de Kolmogorov-Smirnov, segun el criterio del p_valor y para un $\alpha = 0.5$ se rechaza la hipotesis nula de normalidad de los residuos.

El modelo queda descartado al no superar el supuesto de homoscedasticidad.

```
##
##  Jarque Bera Test
##
## data:  datos$t_i
## X-squared = 3.7807, df = 2, p-value = 0.151
```

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  datos$t_i
## D = 0.13593, p-value = 0.005824
## alternative hypothesis: two-sided
```

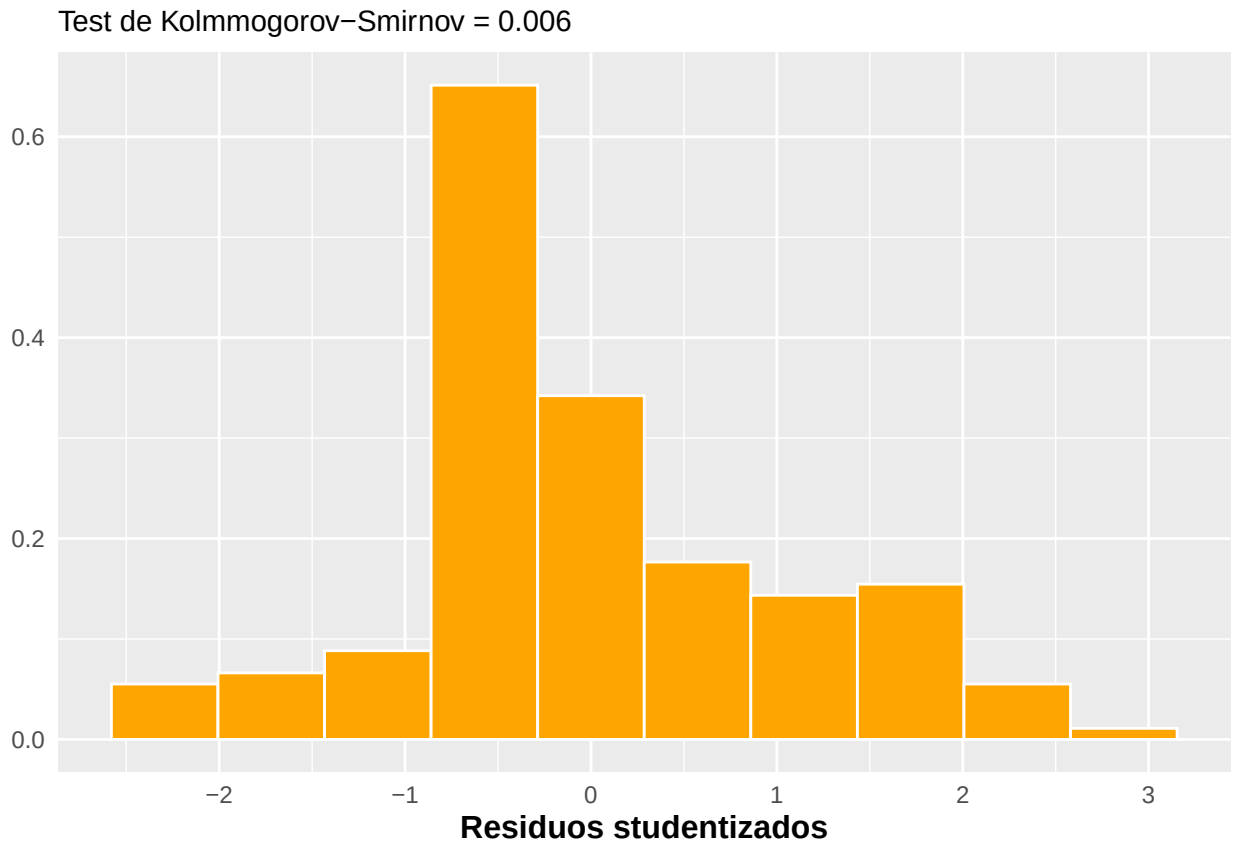


Figure 5: Histograma de los Residuos studientizados

Modelo 2 Como segundo modelo se estimo una regresion con transformacion logaritmica tanto en la variable dependiente como en la variables explicativas

$$Log(Peso_i) = \beta_0 \ + \beta_1 Log(Longitud_1) \ + \beta_2 Log(Altura_i) \ + \beta_3 Log(Ancho_i) + \epsilon_i$$

Diagnostico del modelo

Homoscedasticidad

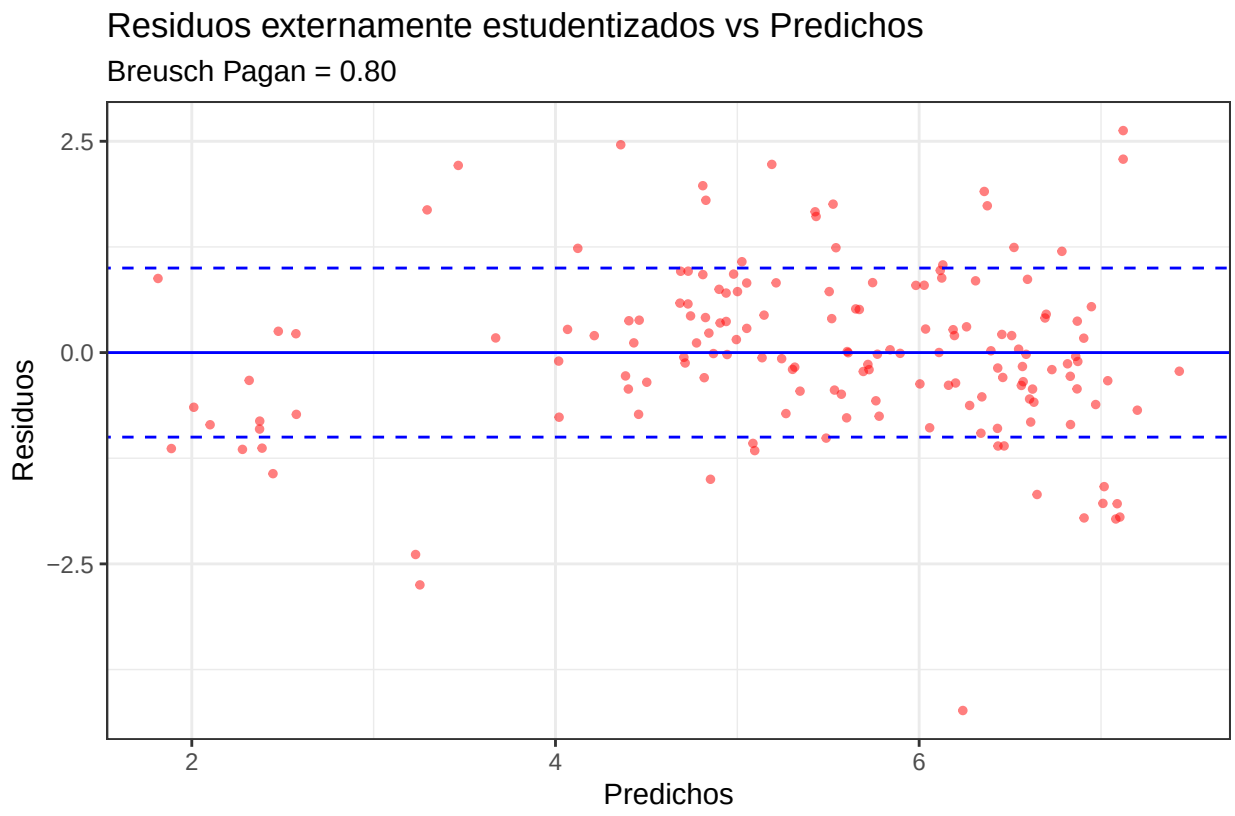


Gráfico 6)

Normalidad

El histograma de los residuos estandarizados se parecerse a una distribucion normal en los residuos. Ademas, el test de normalidad de Kolmogorov-Smirnov , segun el criterio del p_valor y para un $\alpha = 0.5$ no rechaza la hipotesis nula de normalidad de los residuos.

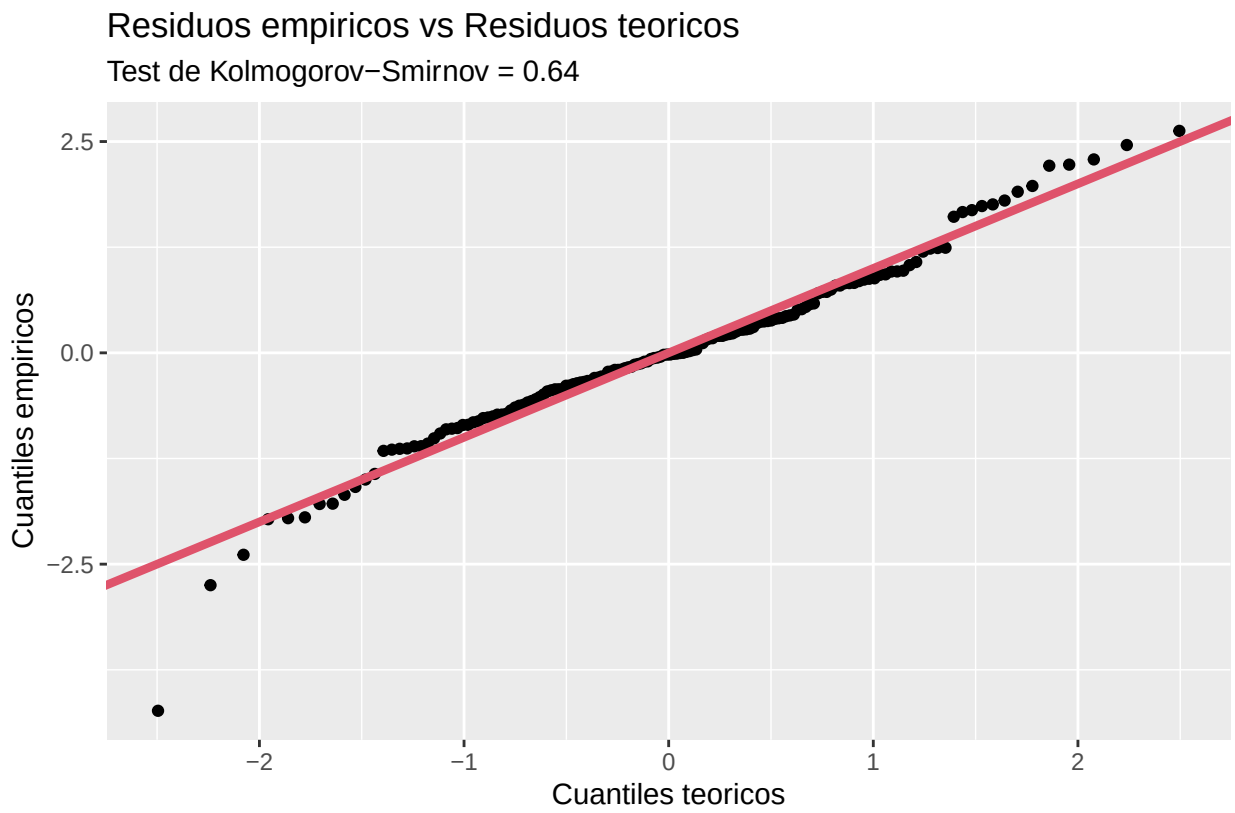


Gráfico 7)

```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  datos$t_i
## D = 0.059245, p-value = 0.6361
## alternative hypothesis: two-sided
```

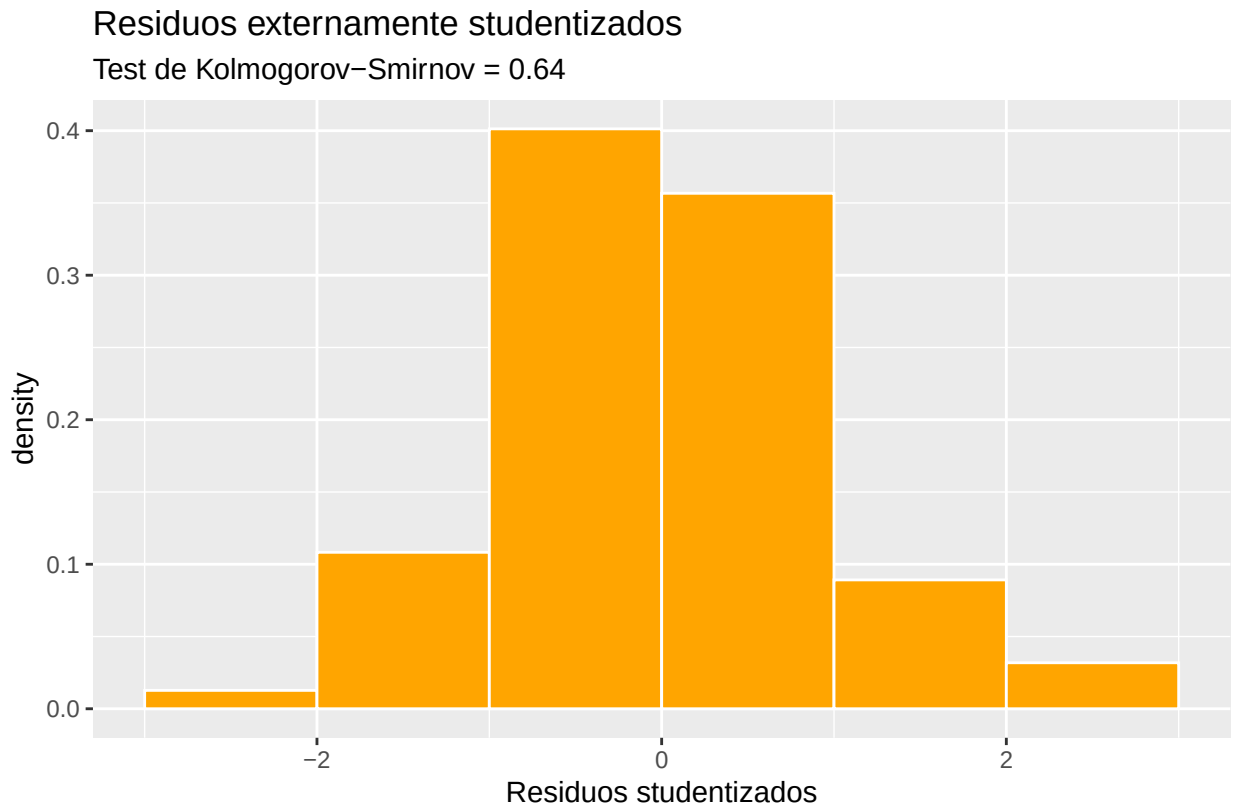


Gráfico 8)

Modelo 3 El modelo 3 cumple con todos los supuestos y es óptimo para realizar el análisis de inferencia y responder las preguntas de investigación. De todas formas, podemos llegar a la conclusión de que el aporte de las variables *Ancho_cm* y *Altura_cm* es marginal, conecntrando en *Longitud1* la mayor explicación de la varianza de los pesos. Siendo este un modelo mas parsimonioso para poder explicar el peso de los peses.

De esta manera, de aquí en más vamos a trabajar con el Modelo 3.

$$Log(Peso_i) = \beta_0 \ + \beta_1 Log(Longitud1_i) \ \ + \ \epsilon_i$$

```
mod3 = lm(log(Peso_gr) ~ log(Longitud1)  , data = datos)



summary(mod3)
```

```
##
## Call:
## lm(formula = log(Peso_gr) ~ log(Longitud1), data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.90870 -0.07280  0.07773  0.26639  0.50636
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -4.62769    0.23481  -19.71   <2e-16 ***
```

```
## log(Longitud1)  3.14394    0.07296   43.09   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## s: 0.3704 on 156 degrees of freedom
## Multiple R-squared: 0.9225,
## Adjusted R-squared: 0.922
## F-statistic:  1857 on 1 and 156 DF,  p-value: < 2.2e-16
```

Significacion individual

Para cada uno de las variables explicativas se realiza la siguiente prueba de hipotesis:

$$H_0)B_i = 0 \; vs \; H_1)B_i \neq 0$$

Con región critica $RC = \left\{ \left(Xy \right) \Big/ |t| \geq t_{n-k-1}(1 - {}^{\alpha}/_{2}) \right\}$

Se usa el estadístico: $t = \frac{\hat{\beta}_i}{\hat{V}(\hat{\beta}_i)} \sim t_{n-k-1}\$$

Siguiendo el criterio del p_valor, la evidencia empirica sugiere que las variables Longitud1 en centimetros es individualmente significativa para explicar el peos del pez a un nivel de confianza del 5%.

Signficacion global del modelo

Siguiendo el criterio del p_valor, a un nivel del 5%, la evidencia empirica sugiere que el modelo es globalmente significativo. Esto implica que, dada la evidencia empirica con la que se cuenta, no es posible rechazar la hipotesis de que las variables explicativas usadas no contribuyen a explicar el peso del pez.

##Homoscedasticidad

```
## # A tibble: 1 x 5
##   statistic p.value parameter method                alternative
##       <dbl>   <dbl>     <dbl> <chr>                 <chr>
## 1     0.885   0.347         1 Koenker (studentised) greater
```

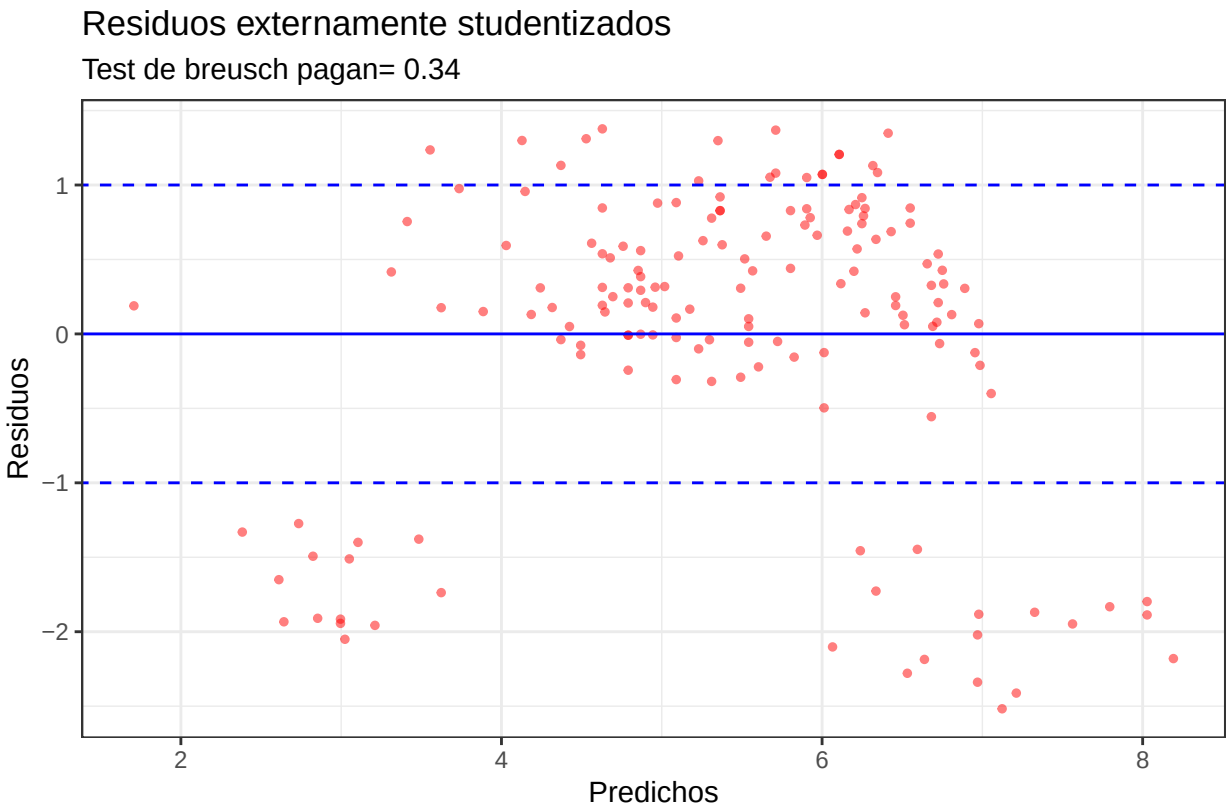
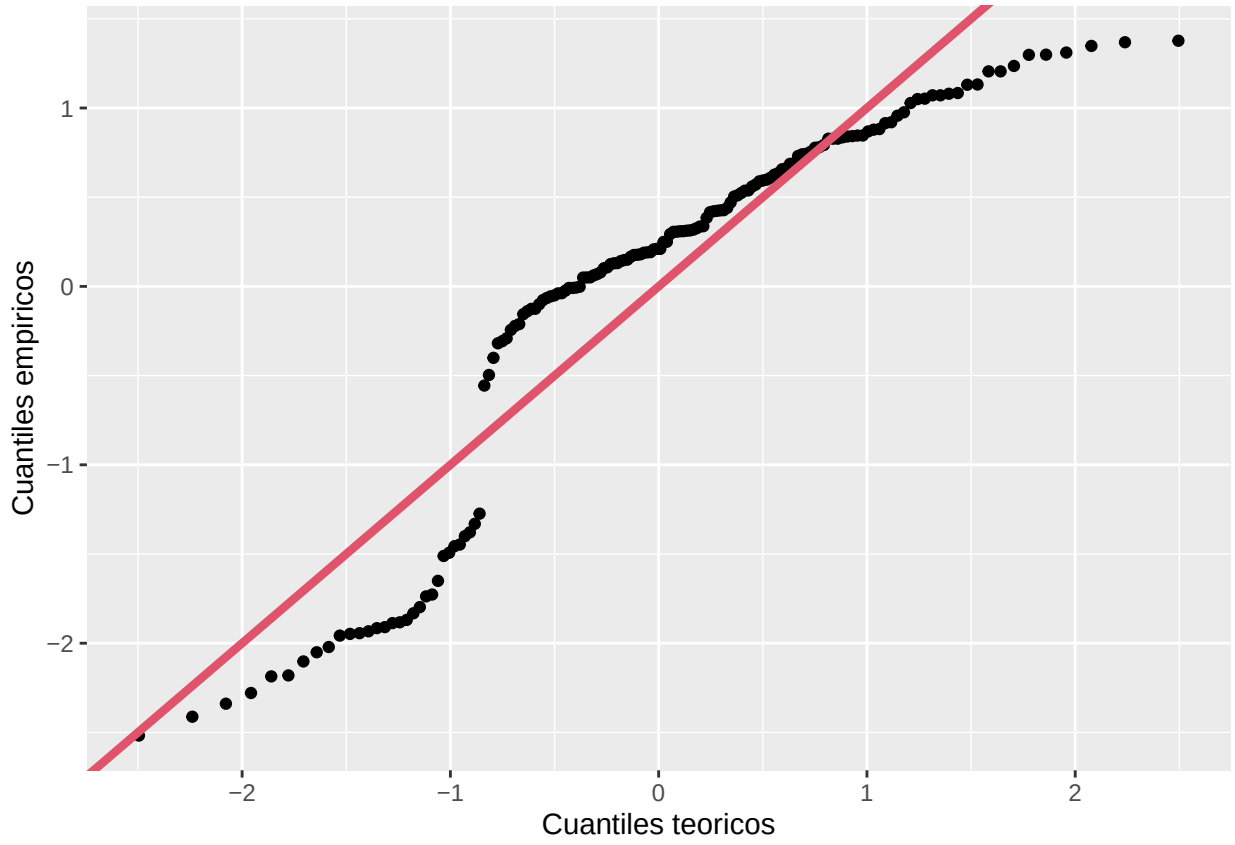

Residuos externamente studentizados
Test de breusch pagan= 0.34
Residuos
1
0
−1
−2
2
4
6
8
Predichos

Gráfico 9)

Normalidad



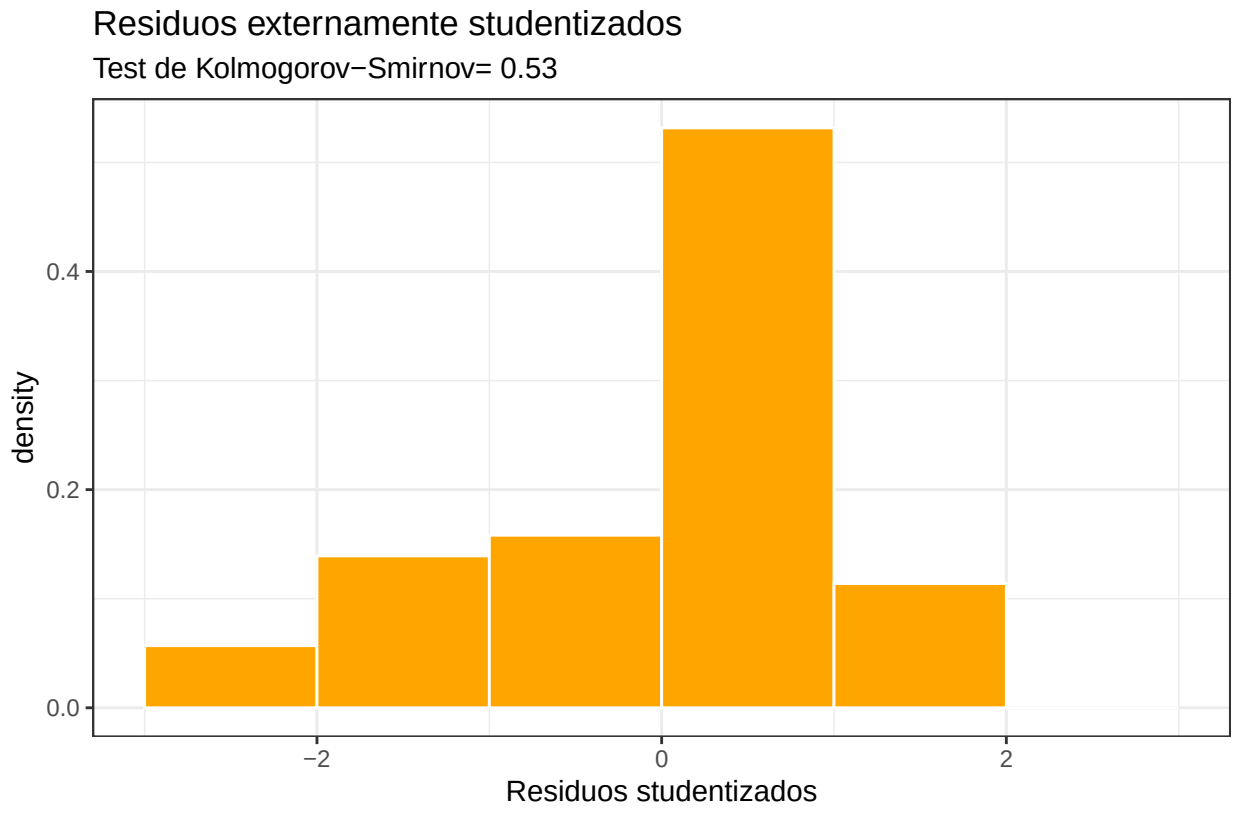


Gráfico 10)

ANOVA a 1 vía

El objetivo es estudiar si existe igualdad de medias entre las categorias de la variable Especie. Haciendo el analisis de varianza a una via, nos plantemos 2 modelos, uno solo con la constante y otro especificando la especie.

El modelo con la constante queda especificado de la siguiente manera:

$$Peso_{ij} = \mu + \epsilon_{ij}$$

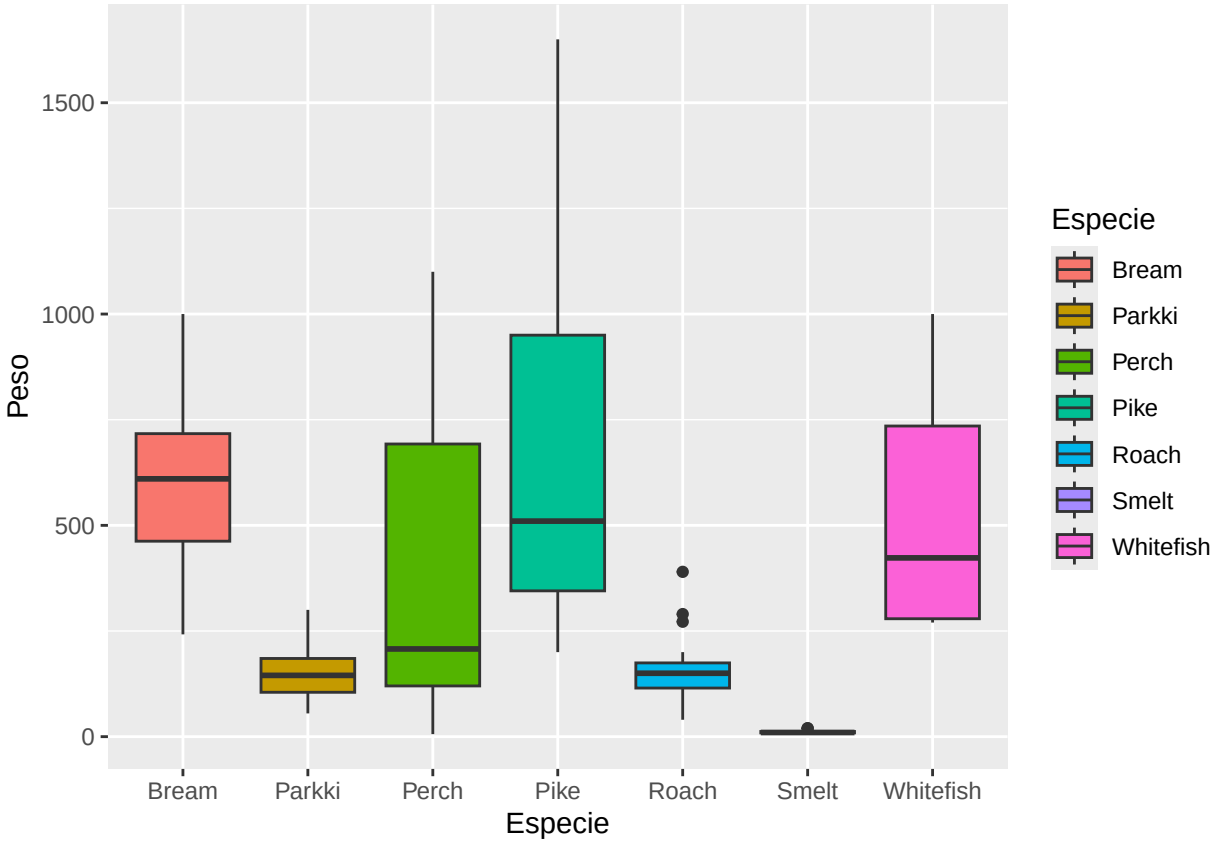
mieintras que si le agregamos el efecto especie queda:

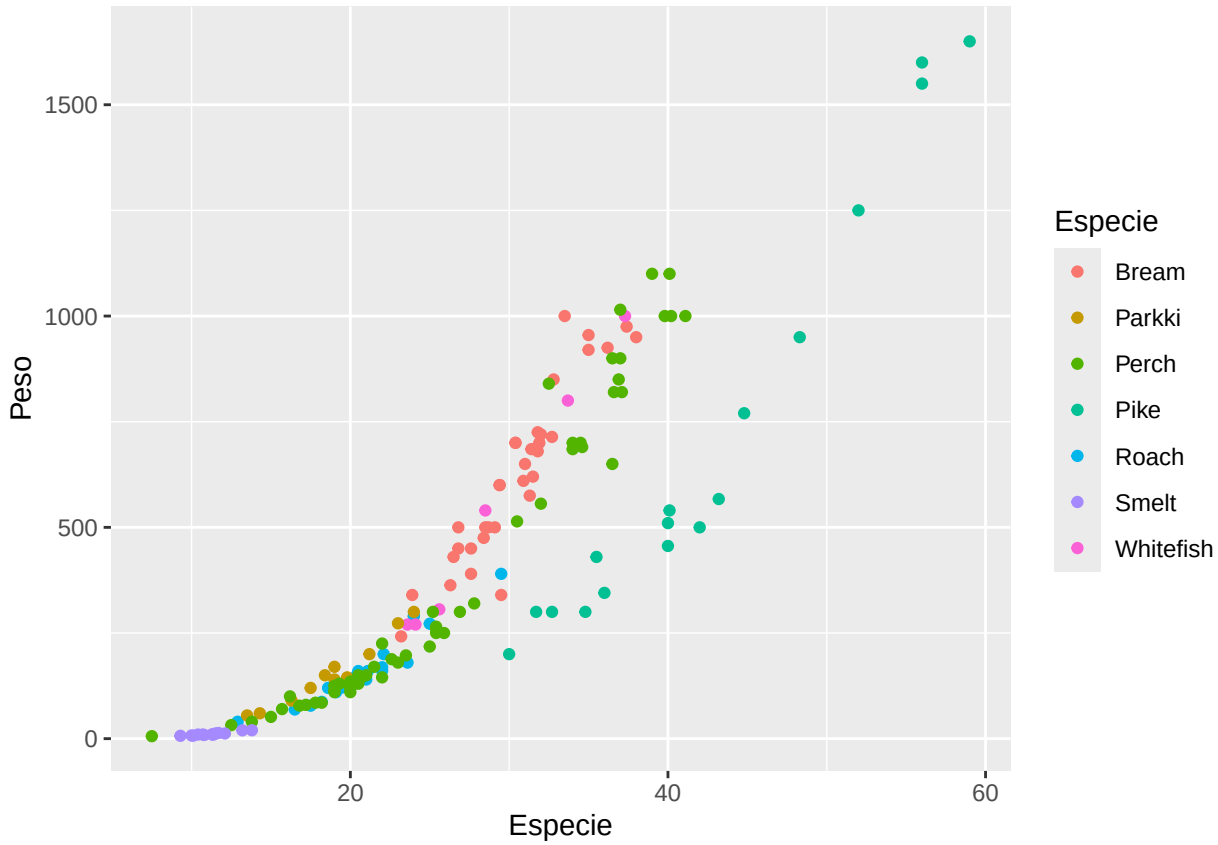
$$Peso_{ij} = \mu + Especie_{ij} + \epsilon_{ij}$$

A un nivel de significación del 5%, podemos afirmar que tenemos evidencia suficiente para rechaza la hipotesis nula de igualdad de medias.

Especie	media(peso)	desvio(peso)	min(peso)	max(peso)
Bream	617.83	209.21	242.0	1000.0
Parkki	154.82	78.76	55.0	300.0
Perch	382.24	347.62	5.9	1100.0
Pike	718.71	494.14	200.0	1650.0
Roach	160.05	83.53	40.0	390.0
Smelt	11.18	4.13	6.7	19.9
Whitefish	531.00	309.60	270.0	1000.0

```
## Analysis of Variance Table
##
## Response: log(Peso_gr)
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Especie     6 187.941  31.323   53.64 < 2.2e-16 ***
## Residuals 151  88.177   0.584
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


1500
1000
500
0
Peso
Bream
Parkki
Perch
Pike
Roach
Smelt
Whitefish
Especie
Especie
Bream
Parkki
Perch
Pike
Roach
Smelt
Whitefish


1500
1000
500
0
Peso
20
40
60
Especie
Especie
Bream
Parkki
Perch
Pike
Roach
Smelt
Whitefish

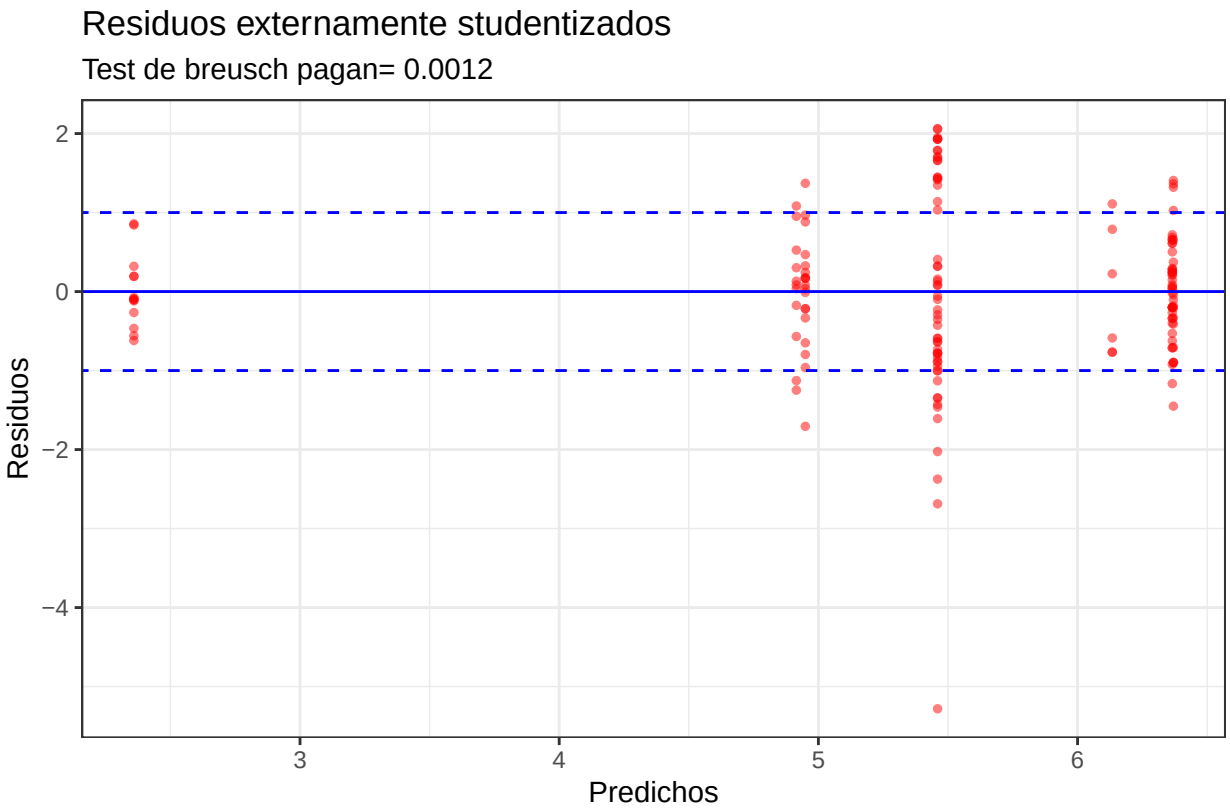
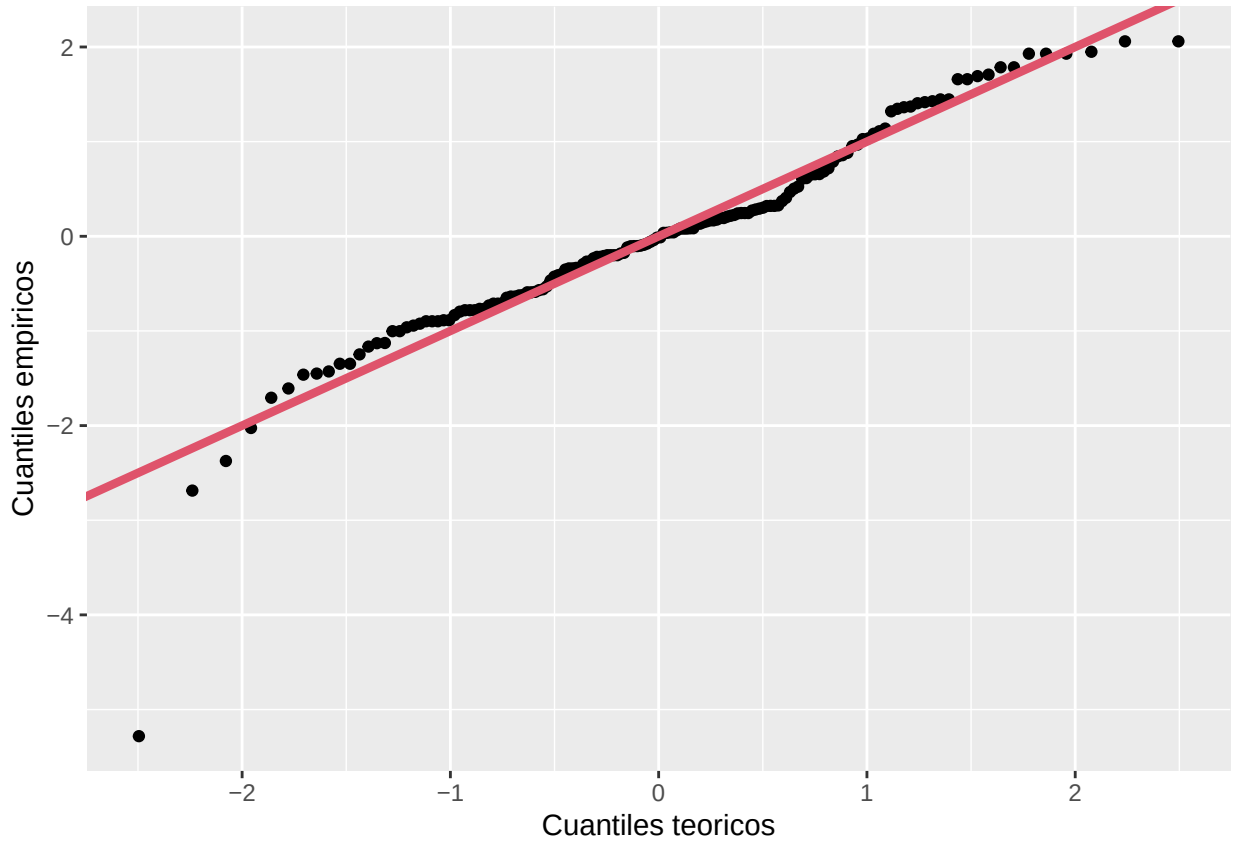
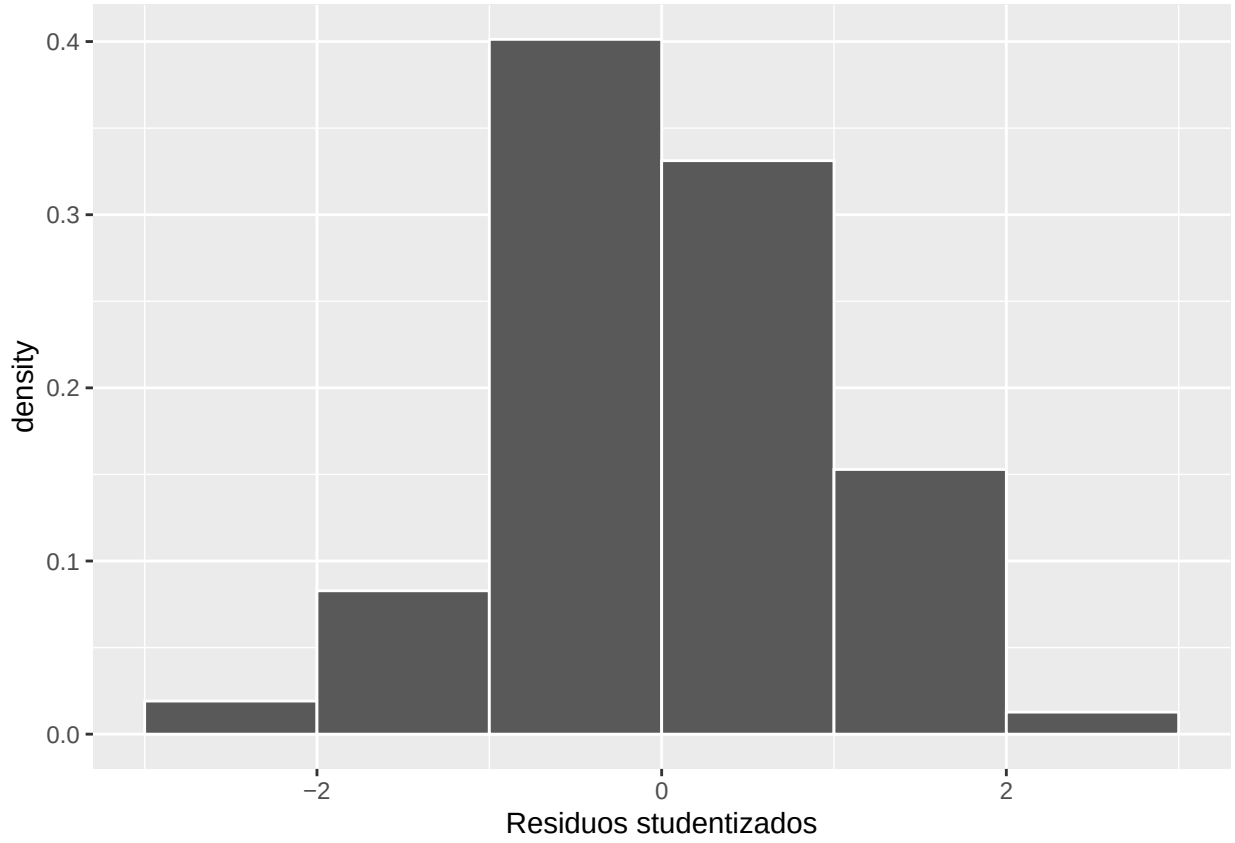

Residuos externamente studentizados
Test de breusch pagan= 0.0012
Residuos
2
0
−2
−4
3
4
5
6
Predichos

Gráfico 11)




```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  datos$t_i
## D = 0.093715, p-value = 0.1246
## alternative hypothesis: two-sided
```



Ancova

Sabiendo que la variable especie es significativo para predecir el peso de cada pez, vamos a plantearnos un modelo en el cual tenga una variable cuantitativa (longitud1) y la variable categorica Especie.

$$Peso_i = \beta_0 \ + \ \beta_1 log(Longitud_i) \ + \beta_2 Especie_i \ + \ \epsilon_i$$

Luego, para ver si existe un efecto Especie sobre la pendiente, debemos plantearnos un modelo con interaccion. El modelo con interaccion queda definido como:

$$Peso_i = \beta_0 \ + \ \beta_1 log(Longitud_i) \ + \beta_2 Especie_i \ + \ \beta_3 log(Longitud1_i) : Especie_i \ + \ \epsilon_i$$

Ahora pasamos a estudiar si efectivamente existe igualdad de pendientes entre las especies. No hay evidencia suficiente para decir que la pendiente sean distinta entre las especies.

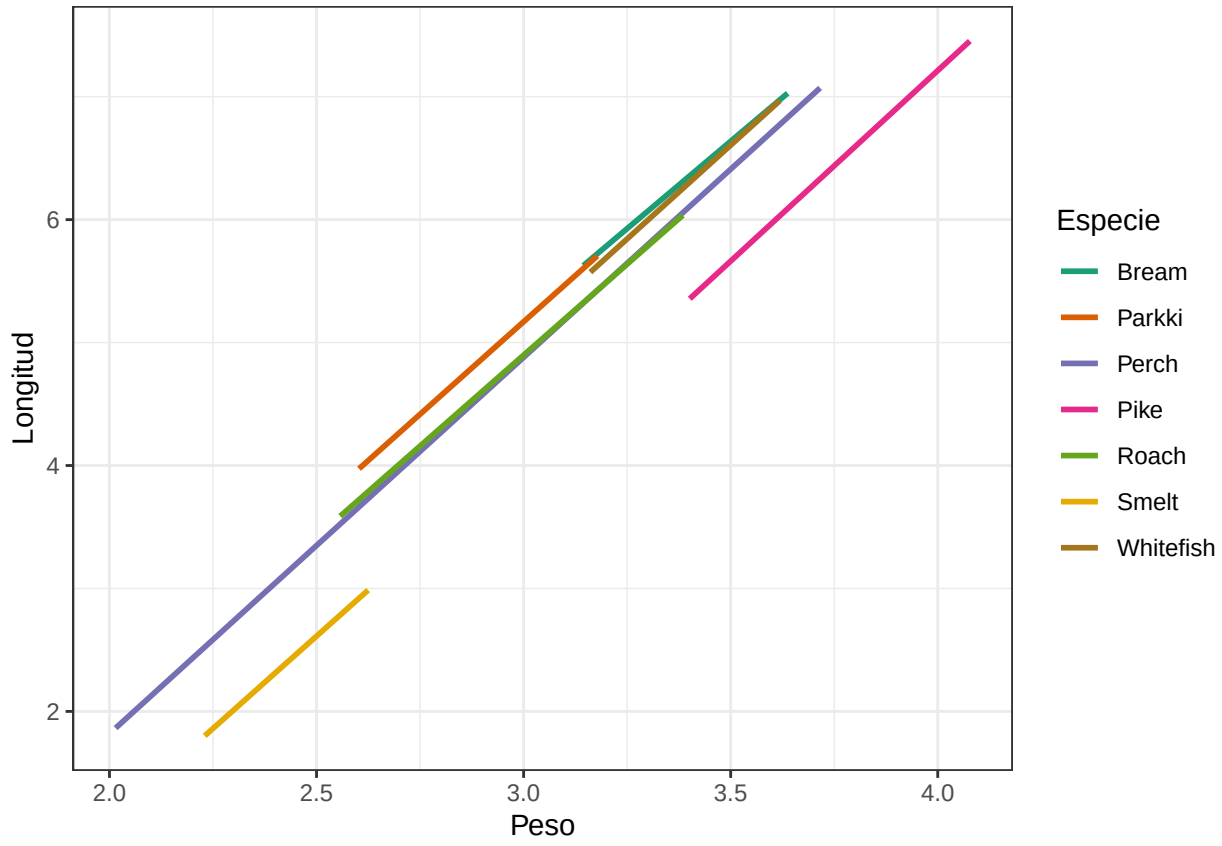
$$H_0) B_i = 0 \ vs \ H_1) B_i \neq 0$$

Se puede observar tambien mediante un grafico de puntos que existe una relacion lineal entre el peso y la longitud. Mas aun haciendo una transformacion logaritmica a ambas variables.

Al plantearnos el modelo con la covariable longitud1 y especie , podemos ver que es suficiente para predecir el peso. Por lo que existe un efecto de la longitud y tambien un efecto de la especie.

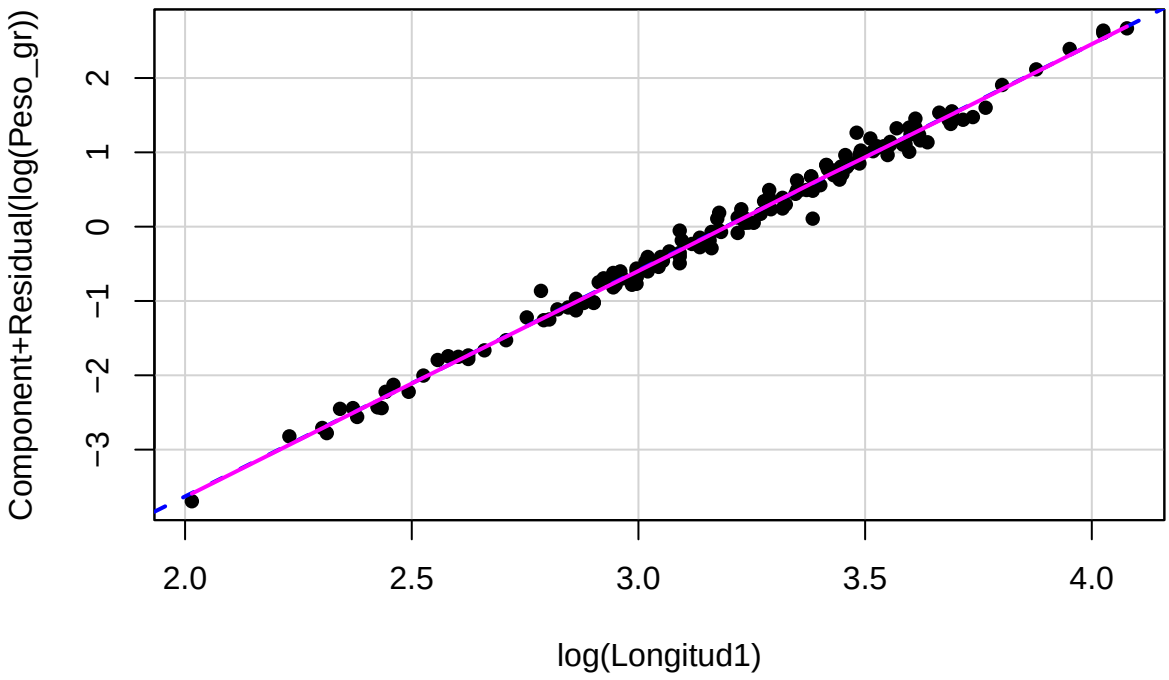
```
## Analysis of Variance Table
##
```

```
## Model 1: log(Peso_gr) ~ log(Longitud1) + Especie
## Model 2: log(Peso_gr) ~ log(Longitud1) + Especie + log(Longitud1):Especie
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    150 2.0118
## 2    144 1.9830  6  0.028848 0.3491 0.9095
```

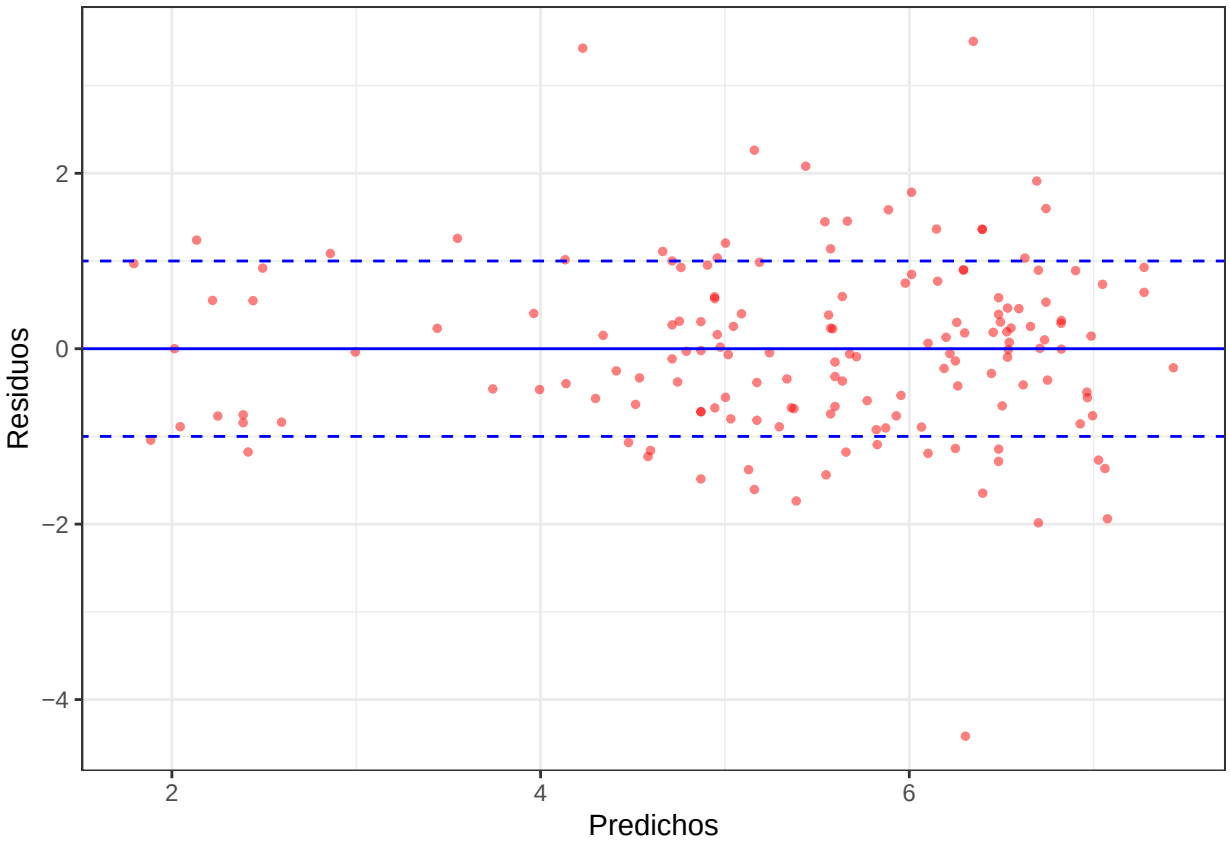


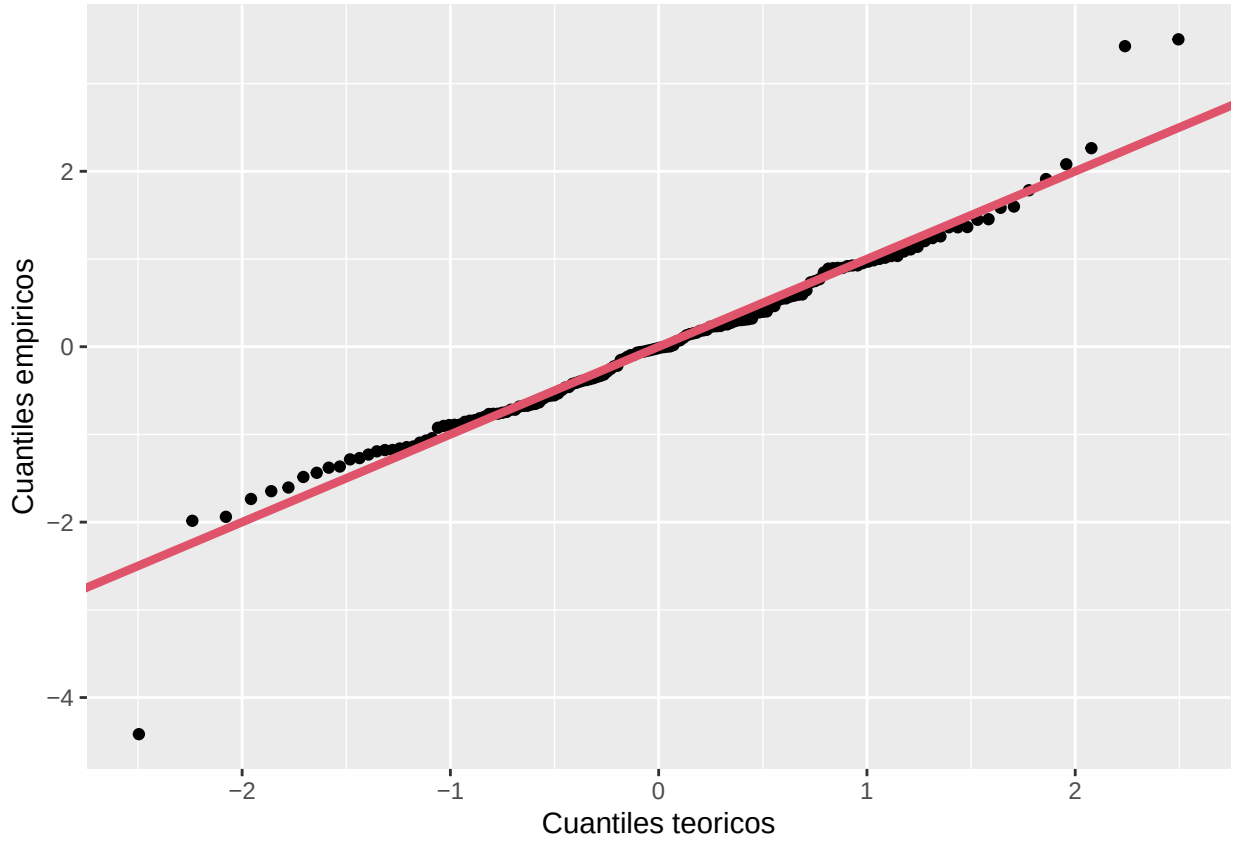
No rechazamos H_0) por lo tanto vemos que las pendientes son iguales entre las especies con un nivel de significación del 5%.

```
## [1] 0.9157195
```

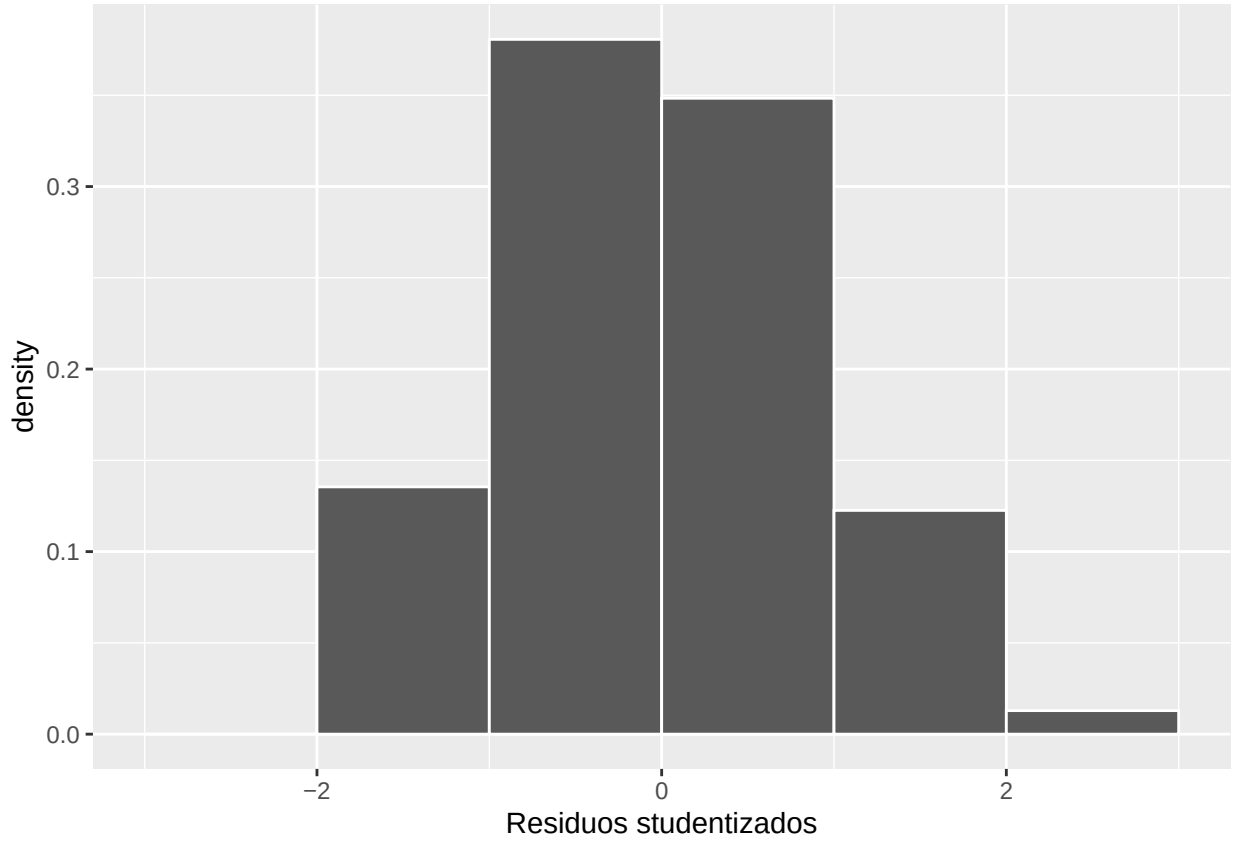



```
## # A tibble: 1 x 5
##   statistic p.value parameter method                alternative
##       <dbl>   <dbl>     <dbl> <chr>                 <chr>
## 1      2.60   0.919         7 Koenker (studentised) greater
```


Residuos
2
0
−2
−4
2
4
6
Predichos




```
##
##  Asymptotic one-sample Kolmogorov-Smirnov test
##
## data:  datos$t_i
## D = 0.05098, p-value = 0.8061
## alternative hypothesis: two-sided
```



Cross-Validation

El objetivo de la validacion cruzada es separar primero la base en 2, una de testeo y otra de entrenamiento. Luego se usa esta ultima para estimar el modelo. Luego se utiliza el modelo para predecir la variable peso en la base de testeo. Por ultimo se calcula el ECM de dichas predicciones

$$CV_{[n]} = \frac{1}{k}\sum_{i=1}^{k} ECM_i$$

modelo	k_folds_cv
modelo 1	1.76
modelo 2	0.60
modelo 3	0.01

Conclusiones

Mediante ensayo y error llegamos a la conclusion de que el modelo con una variable explicativa alcanza para poder predecir el peso de los peces, lo cual lo hace un modelo mas eficiente que la otra alternativa, por mas que el indicador del R^2 ajustado sea mas chico que otro modelo con mas covariables, no es determinante para no quedarnos con este modelo.

Como era de esperarse, al hacer el analisis de varianza a una via, podemos afirmar que existe diferencia de medias entre los grupos, siendo especie una buena variable para distinguir el peso de los peces.

En tanto al analisis de covarianza (ANCOVA), vimos que las covariables longitud y especie, son buenas predictoras para predecir el peso, y no importa la interaccion entre ellas.