

Estadística multivariada

Victor Gabriel Arredondo Dominguez

Table of contents

1	Análisis de producción de pescado	2
2	Introducción	2
3	1.- Descripción y carga de los datos	3
3.0.1	Variables Temporales y Geográficas	3
3.0.2	Variabes de Clasificación Biológica y Técnica	3
3.0.3	Variabes Cuantitativas (Numéricas)	3
4	2.- Regresión lineal	4
4.1	Carga de librerías	4
4.2	Matriz de correlación	7
4.3	Estimación y selección del modelo de regresión lineal múltiple	9
4.4	Validación de supuestos	11
4.4.1	Linealidad	11
4.4.2	Normalidad en los residuales	12
4.4.3	Homocedasticidad	12
4.4.4	Autocorrelación de los residuos	13
4.5	Predicciones	14
5	3.- Plots	15
5.1	Series de tiempo	15
5.2	Plots 3D	20
5.3	Otros plots	22
6	4.- PCA	23
7	5.- Clustering	25
8	Conclusiones	27
8.1	1. Modelado y Predicción del Valor Económico	27

8.2	2. Dinámica de la Producción y Valor	28
8.3	3. Eficiencia en la Reducción de Dimensionalidad (PCA)	28
8.4	4. Segmentación del Sector (Clustering)	28
8.5	Limitaciones y Trabajo Futuro	29
9	Bibliografia	29

1 Análisis de producción de pescado

2 Introducción

En el presente proyecto se realiza un análisis estadístico multivariante sobre la producción pesquera en México entre los años **2005** y **2018**, utilizando como base de datos el dataset *Fisheries Production in Mexico (2005 to 2018)* obtenido de la plataforma [kaggle](#) (1).

Este proyecto busca aplicar diversas técnicas de estadística multivariante para extraer conocimiento útil del dataset:

1. **Regresión Lineal Múltiple (RLM):** Se plantearán hipótesis relacionadas con la producción total o por grupo de especies en función de variables explicativas relevantes, justificando la elección del modelo y analizando sus resultados e interpretación.
2. **Métodos de visualización multivariante:** Se aplicará al menos una técnica de gráficas multivariadas para explorar relaciones entre múltiples variables simultáneamente y visualizar patrones o agrupamientos.
3. **Análisis de Componentes Principales (PCA):** Se usará PCA para reducir la dimensionalidad del conjunto de datos y evaluar cómo las distintas variables contribuyen a la variabilidad general, así como interpretar los componentes principales desde el punto de vista de la producción pesquera.
4. **Clustering:** Se implementará al menos una técnica de clasificación no supervisada (como *k-means* o *hierarchical clustering*) para identificar grupos homogéneos dentro de los datos.
5. **Resultados y conclusiones:** Se presentarán conclusiones generales basadas en los métodos anteriores, discutiendo hallazgos relevantes, limitaciones de los análisis y posibles implicaciones para futuras investigaciones en la temática de producción pesquera.

La selección de este dataset se basa en que contiene múltiples **variables numéricas y categóricas** (por ejemplo año, especie, volumen producido, tipo de pesca, etc.) necesarias para aplicar técnicas de estadística multivariante, permitiendo un análisis integral que abarque desde modelos de regresión hasta métodos exploratorios y de clusterización.

3 1.- Descripción y carga de los datos

3.0.1 Variables Temporales y Geográficas

- **ANÓ (Año):** Indica el año en que se realizó la captura o producción.
- **MES (Mes):** Mes de la captura.
- **LITORAL (Litoral):** Costa de procedencia del producto (Pacífico, Golfo de México/Caribe o Sin Litoral para estados interiores).
- **ENTIDAD (Entidad):** Estado de la República Mexicana donde se registró el desembarque del producto.

3.0.2 Variables de Clasificación Biológica y Técnica

- **GRUPO (Grupo):** Clasificación taxonómica o comercial del grupo de la especie (ANIMAL/VEGETAL).
- **NOMBRE_PRINCIPAL_ESPECIE (Especie):** Nombre común comercial de la especie en español (ej. Camarón, Atún, Mojarrá).
- **ORIGEN (Origen):** Indica si el producto proviene de “Captura” (pesca silvestre) o “Acuacultura” (cultivo).
- **ORIGEN DEL CAMARÓN (Origen del Camarón):** Campo específico para la especie camarón que detalla si proviene de alta mar, esteros o cultivo.
- **TIPO_AQUA (Tipo de Agua):** Clasificación del cuerpo de agua (Marina, Salobre o Dulce).
- **DESTINO (Destino):** Uso previsto para el producto (ej. Consumo Humano Directo, Industrialización).
- **ACUACULTURA (Acuacultura):** Información detallada sobre el sistema de cultivo utilizado, si aplica.

3.0.3 Variables Cuantitativas (Numéricas)

- **PESO_DESEMBARCADO_KG (Peso Desembarcado):** Peso total del producto al momento de llegar a puerto, expresado en kilogramos.
- **PESO_VIVO_KG (Peso Vivo):** Peso total del organismo al momento de la captura antes de cualquier proceso, expresado en kilogramos.

- **VALOR_MEXICAN_PESOS (Valor):** Valor comercial total de la producción en pesos mexicanos.

```
fishy <- read.csv("Fisheries_Production_Mexico_2005to2018.csv")
head(fishy)
```

	ANO	LITORAL	ENTIDAD	MES	GRUPO	NOMBRE_PRINCIPAL_ESPECIE	ORIGEN
1	2018	GOLFO Y CARIBE	CAMPECHE	ENERO	ANIMAL		ANCHOVETA CAPTURA
2	2018	GOLFO Y CARIBE	CAMPECHE	ENERO	ANIMAL		BAGRE CAPTURA
3	2018	GOLFO Y CARIBE	CAMPECHE	ENERO	ANIMAL		BAGRE CAPTURA
4	2018	GOLFO Y CARIBE	CAMPECHE	ENERO	ANIMAL		BANDERA CAPTURA
5	2018	GOLFO Y CARIBE	CAMPECHE	ENERO	ANIMAL		BERRUGATA CAPTURA
6	2018	GOLFO Y CARIBE	CAMPECHE	ENERO	ANIMAL		BESUGO CAPTURA
		ORIGEN.DEL.CAMARON		TIPO_AGUA		DESTINO	ACUACULTURA
1		NO APLICA	SALOBRE / MARINA	CONSUMO HUMANO	DIRECTO		NO APLICA
2		NO APLICA		DULCE CONSUMO	HUMANO DIRECTO		NO APLICA
3		NO APLICA	SALOBRE / MARINA	CONSUMO HUMANO	DIRECTO		NO APLICA
4		NO APLICA	SALOBRE / MARINA	CONSUMO HUMANO	DIRECTO		NO APLICA
5		NO APLICA	SALOBRE / MARINA	CONSUMO HUMANO	DIRECTO		NO APLICA
6		NO APLICA	SALOBRE / MARINA	CONSUMO HUMANO	DIRECTO		NO APLICA
		PESO_DESEMBARCADO_KG	PESO_VIVO_KG	VALOR_MEXICAN_PESOS			
1		292.50	308.50		2654.417		
2		334.40	334.40		1814.110		
3		6765.52	6765.52		52186.462		
4		152706.80	152706.80		1390232.571		
5		23780.40	23780.40		140230.047		
6		4430.80	4430.80		178721.342		

4 2.- Regresión lineal

El objetivo de esta sección es estimar un modelo de regresión lineal múltiple (RLM) que permita predecir el valor de la producción expresado en pesos mexicanos

4.1 Carga de librerías

Para ello cargamos las librerías correspondientes

```
library(leaps)
library(GGally)
```

```
Cargando paquete requerido: ggplot2
```

```
library(ggplot2)
library(olsrr)
```

```
Adjuntando el paquete: 'olsrr'
```

```
The following object is masked from 'package:datasets':
```

```
    rivers
```

```
library(MASS)
```

```
Adjuntando el paquete: 'MASS'
```

```
The following object is masked from 'package:olsrr':
```

```
    cement
```

```
library(lmtest)
```

```
Cargando paquete requerido: zoo
```

```
Adjuntando el paquete: 'zoo'
```

```
The following objects are masked from 'package:base':
```

```
    as.Date, as.Date.numeric
```

```
library(dplyr)
```

```
Adjuntando el paquete: 'dplyr'
```

```
The following object is masked from 'package:MASS':
```

```
select
```

```
The following objects are masked from 'package:stats':
```

```
filter, lag
```

```
The following objects are masked from 'package:base':
```

```
intersect, setdiff, setequal, union
```

```
library(plotly)
```

```
Adjuntando el paquete: 'plotly'
```

```
The following object is masked from 'package:MASS':
```

```
select
```

```
The following object is masked from 'package:ggplot2':
```

```
last_plot
```

```
The following object is masked from 'package:stats':
```

```
filter
```

```
The following object is masked from 'package:graphics':
```

```
layout
```

```
library(scales)
library(scatterplot3d)
library(tidyr)
library(cluster)
library(factoextra)
```

```
Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```

fishy$ANO <- as.numeric(fishy$ANO)
fishy$MES <- as.factor(fishy$MES)
fishy$LITORAL <- as.factor(fishy$LITORAL)
fishy$ENTIDAD <- as.factor(fishy$ENTIDAD)
fishy$GRUPO <- as.factor(fishy$GRUPO)
fishy$ORIGEN <- as.factor(fishy$ORIGEN)
fishy$TIPO_AQUA <- as.factor(fishy$TIPO_AQUA)

```

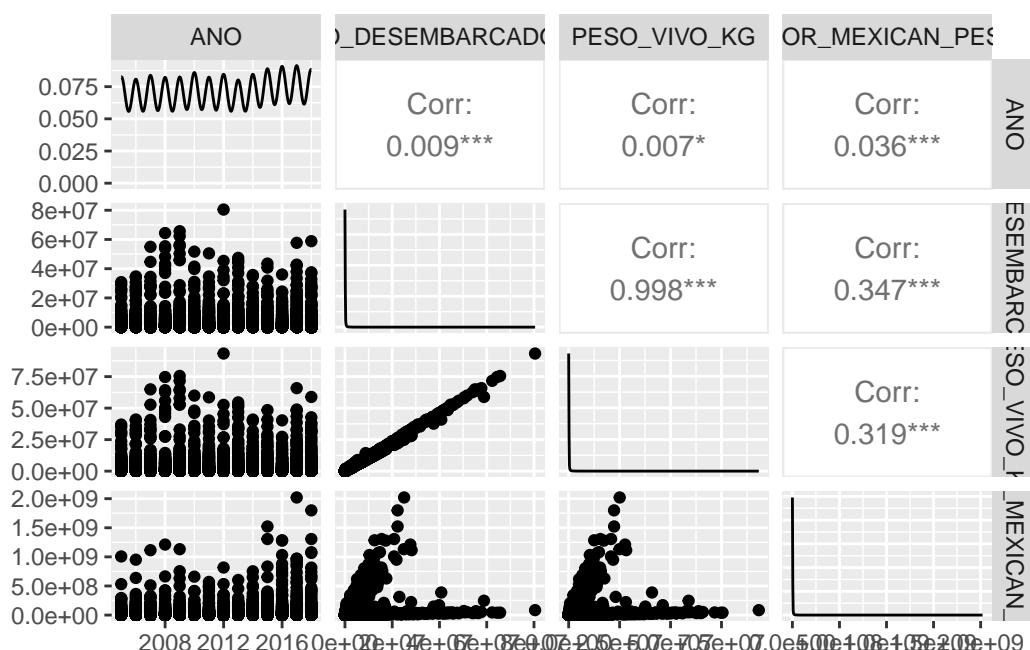
```

vars_numericas <- sapply(fishy, is.numeric)
fishy_num <- fishy[, vars_numericas]

```

4.2 Matriz de correlación

```
ggpairs(fishy_num)
```



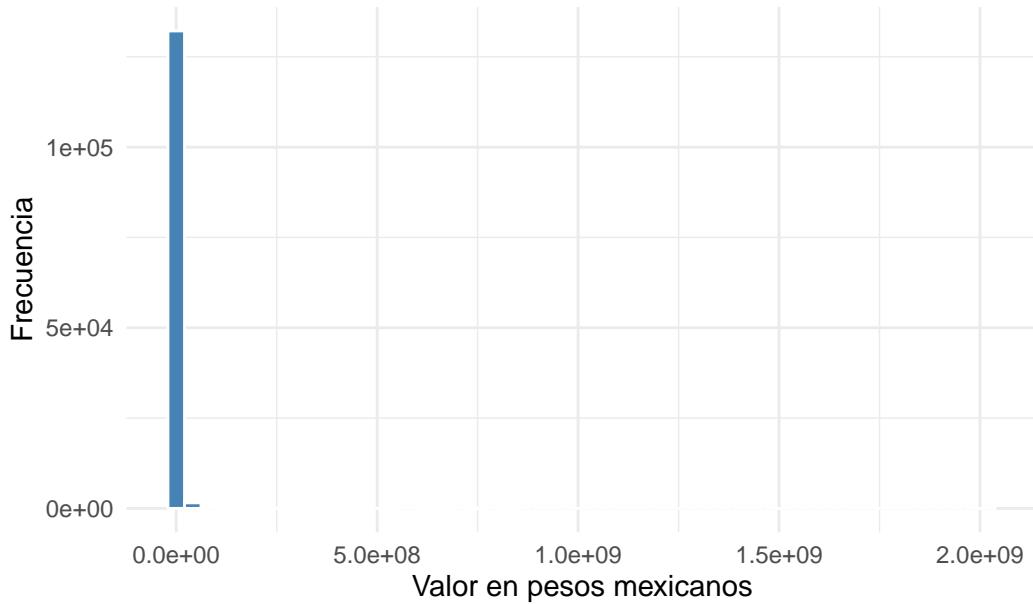
Se observa que las variables PESO_VIVO_KG y PESO_DESEMBARCADO_KG presentan una correlación extremadamente alta (0.998), lo que indica una relación lineal casi perfecta entre ambas. Esta alta correlación sugiere la presencia de multicolinealidad, por lo que se evitara incluir ambas variables simultáneamente en el modelo

```

ggplot(fishy, aes(x = VALOR_MEXICAN_PESOS)) +
  geom_histogram(
    bins = 50,
    fill = "steelblue",
    color = "white"
  ) +
  labs(
    title = "Distribución del valor de la producción",
    x = "Valor en pesos mexicanos",
    y = "Frecuencia"
  ) +
  theme_minimal()

```

Distribución del valor de la producción

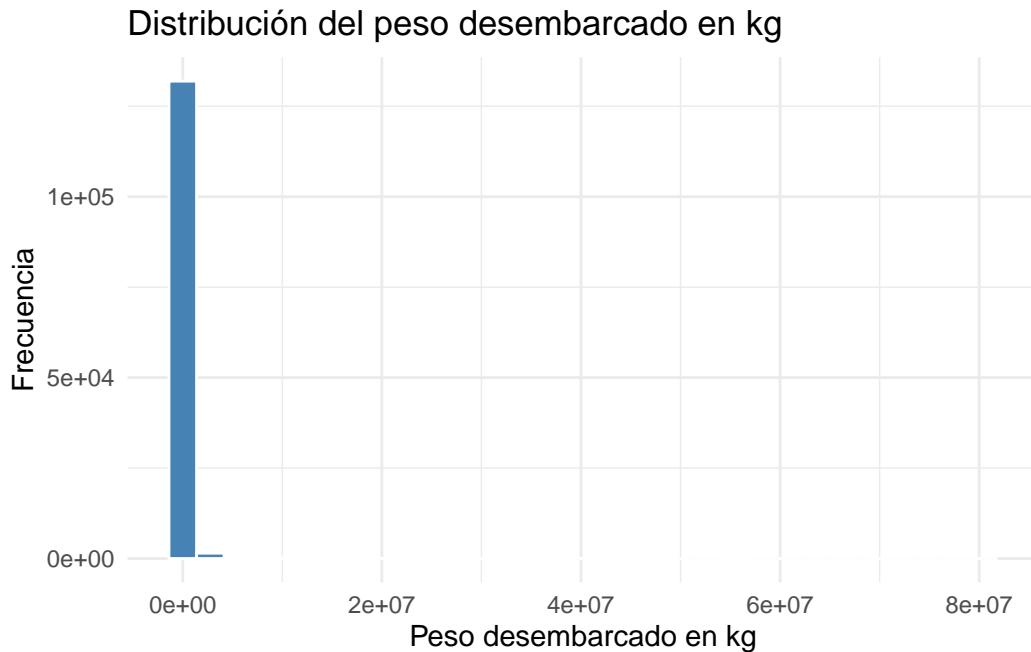


```

ggplot(fishy, aes(x = PESO_DESEMBARCADO_KG)) +
  geom_histogram(
    bins = 30,
    fill = "steelblue",
    color = "white"
  ) +
  labs(
    title = "Distribución del peso desembarcado en kg",
    x = "Peso desembarcado en kg",
    y = "Frecuencia"
  )

```

```
y = "Frecuencia"  
)+  
theme_minimal()
```



Al verificar la distribución de la variable objetivo notamos un sesgo extremo a la derecha

4.3 Estimación y selección del modelo de regresión lineal múltiple

```
modelo_rlm <- lm(  
  VALOR_MEXICAN_PESOS ~  
  .,  
  data = fishy  
)
```

Se estimó inicialmente un modelo de regresión lineal múltiple utilizando como variable **VALOR_MEXICAN_PESOS** el e incluyendo como variables explicativas todas las variables

```
paso_a_paso<-stepAIC(modelo_rlm, direction="both", trace=T)
```

El proceso de selección de variables se realizó utilizando el método paso a paso (*stepwise*) con base en el Criterio de Información de Akaike (AIC).

El procedimiento concluyó con un modelo final que presenta un AIC mínimo de 4,473,815, con las siguientes variables:**ANO, ENTIDAD, MES, NOMBRE_PRINCIPAL_ESPECIE, ORIGEN.DEL.CAMARON, DESTINO, ACUACULTURA, PESO_DESEMBARCADO_KG, PESO_VIVO_KG**

Se decidió excluir la variable **PESO_VIVO_KG** del modelo final. Aunque el procedimiento de selección basado en el AIC indicó que esta variable es estadísticamente relevante, se identificó que presenta una correlación extremadamente alta (0.998) con **PESO_DESEMBARCADO_KG**. La inclusión simultánea de ambas variables introduce problemas severos de multicolinealidad.

```
range(fishy$PESO_DESEMBARCADO_KG)
```

```
[1] 0 80452940
```

```
range(fishy$VALOR_MEXICAN_PESOS)
```

```
[1] 0 2020440294
```

```
fishy_log <- fishy %>%
  dplyr::filter(
    PESO_DESEMBARCADO_KG > 0,
    VALOR_MEXICAN_PESOS > 0
  ) # Se filtran los elemenots con valor 0
```

Se aplicó una transformación logarítmica a las variables **VALOR_MEXICAN_PESOS** y **PESO_DESEMBARCO_KG**. Esta decisión se fundamenta en la fuerte asimetría observada en sus distribuciones.

```
modelo_log <- lm(
  log(VALOR_MEXICAN_PESOS) ~
  log(PESO_DESEMBARCADO_KG) +
  ANO + MES + ENTIDAD +
  NOMBRE_PRINCIPAL_ESPECIE +
  ORIGEN.DEL.CAMARON +
  DESTINO + ACUACULTURA,
  data = fishy_log
)
```

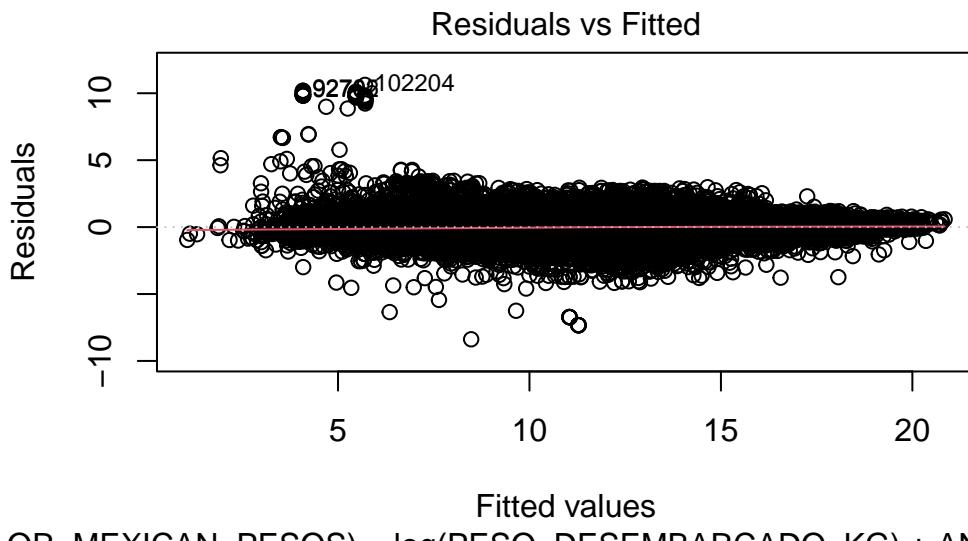
```
summary(modelo_log)
```

El modelo presenta un alto poder explicativo (R^2 ajustado = 0.9525) y es globalmente significativo ($p < 0.001$). El logaritmo del peso desembarcado tiene un efecto positivo y altamente significativo sobre el valor en pesos mexicanos, con un coeficiente cercano a uno, lo que indica una relación proporcional. Asimismo, el año y las variables estructurales por entidad, especie, origen, destino y acuacultura resultan estadísticamente relevantes.

4.4 Validación de supuestos

4.4.1 Linealidad

```
plot(modelo_log, which = 1)
```



```
.OR_MEXICAN_PESOS) ~ log(PESO_DESEMBARCADO_KG) + ANO + N
```

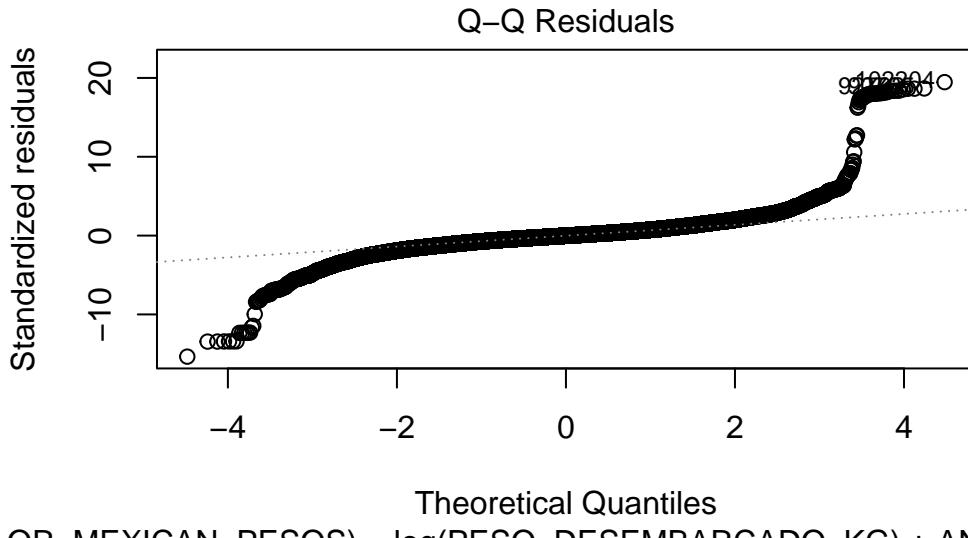
```
# Prueba de Rainbow  
raintest(modelo_log)
```

```
Rainbow test  
  
data: modelo_log  
Rain = 0.81135, df1 = 67171, df2 = 67064, p-value = 1
```

Tanto con la grafica y la prueba podemos comprobar que se cumple el supuesto de linealidad

4.4.2 Normalidad en los residuales

```
plot(modelo_log, which = 2)
```



En el gráfico observado, los puntos en los extremos izquierdo y derecho se alejan significativamente de la línea diagonal punteada, formando una estructura en forma de "S".

Aunque el gráfico indica que los residuos no siguen perfectamente una distribución normal, para los fines de este estudio el modelo se considera válido. Dado que el conjunto de datos cuenta con 134,354 observaciones, se puede invocar el Teorema del Límite Central, el cual garantiza que las inferencias estadísticas son confiables en muestras grandes.

4.4.3 Homocedasticidad

```
library(lmtest)
```

bptest(modelo_log)

studentized Breusch-Pagan test

```
data: modelo_log
BP = 11791, df = 105, p-value < 2.2e-16
```

La prueba de Breusch-Pagan estudiada sobre el modelo logarítmico (studentized BP = 11,791, df = 105, p < 2.2e-16) indica que no se cumple la homocedasticidad, es decir, los residuos presentan varianza no constante. Para corregir este problema, se utilizaron errores estándar robustos de tipo HC1 mediante `coeftest`, lo que permite obtener estimaciones de los coeficientes más confiables frente a la heterocedasticidad.

```
library(sandwich)
library(lmtest)

coeftest(modelo_log, vcov = vcovHC(modelo_log, type = "HC1"))
```

Debido a que la prueba de Breusch-Pagan indicó heterocedasticidad en el modelo (BP = 11,791, df = 105, p < 2.2e-16), se recurrió a errores estándar robustos para obtener estimaciones confiables de los coeficientes. La tabla resultante muestra los coeficientes ajustados, sus errores estándar robustos, los valores t y los niveles de significancia.

- Variables como `log(PESO_DESEMBARCADO_KG)` y `ANO` resultan altamente significativas (p < 2.2e-16), indicando un fuerte efecto sobre la variable dependiente.
- Algunas categorías de meses (`MESAGOSTO`, `MESJULIO`, `MESOCTUBRE`) muestran significancia marginal.
- La mayoría de los efectos de entidad (`ENTIDAD*`) y especie principal (`NOMBRE_PRINCIPAL_ESPECIE*`) son estadísticamente significativos, confirmado diferencias importantes entre regiones y especies en el modelo.
- Variables de `origen` y `destino`, así como `sistemas de acuacultura`, también muestran efectos significativos con errores robustos.

4.4.4 Autocorrelación de los residuos

```
dwtest(modelo_log)
```

Durbin-Watson test

```
data: modelo_log
DW = 1.2857, p-value = 0.7603
alternative hypothesis: true autocorrelation is greater than 0
```

El estadístico obtenido fue **DW = 1.286** con un valor p de **0.7603**, lo que indica que no hay evidencia significativa de autocorrelación positiva en los residuos. Por lo tanto, se asume que los errores del modelo son independientes.

Debido a la heterocedasticidad detectada y a la significancia marginal de algunas variables de mes, se decidió agrupar los meses en trimestres para simplificar el modelo y reducir la dispersión de los coeficientes, mejorando así la estabilidad de las estimaciones.

```
fishy_log$TRIMESTRE <- with(fishy_log,
                                ifelse(MES %in% c("ENERO", "FEBRERO", "MARZO"), "T1",
                                       ifelse(MES %in% c("ABRIL", "MAYO", "JUNIO"), "T2",
                                             ifelse(MES %in% c("JULIO", "AGOSTO", "SEPTIEMBRE"), "T3",
                                                   "T4"))))

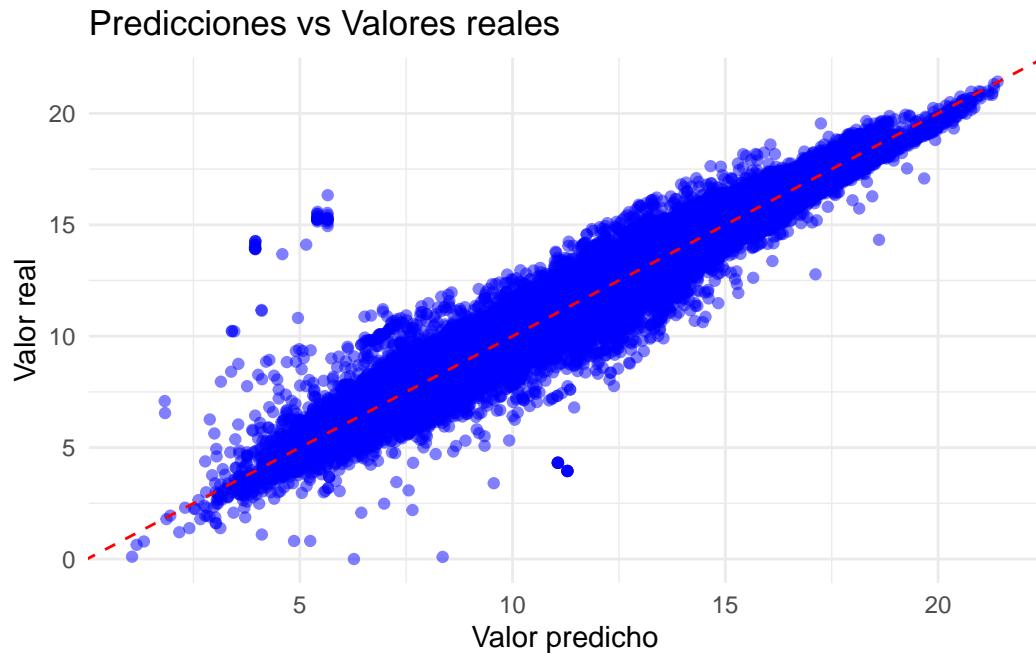
fishy_log$TRIMESTRE <- factor(fishy_log$TRIMESTRE)

modelo_log_trim <- lm(log(VALOR_MEXICAN_PESOS) ~ log(PESO_DESEMBARCADO_KG) + ANO + TRIMESTRE - 1)
```

4.5 Predicciones

```
fishy_log$pred <- predict(modelo_log_trim, newdata = fishy_log)

ggplot(fishy_log, aes(x = pred, y = log(VALOR_MEXICAN_PESOS))) +
  geom_point(alpha = 0.5, color = "blue") +
  geom_abline(intercept = 0, slope = 1, color = "red", linetype = "dashed") +
  labs(
    x = "Valor predicho",
    y = "Valor real",
    title = "Predicciones vs Valores reales"
  ) +
  theme_minimal()
```



El modelo de regresión lineal múltiple presenta un ajuste robusto y de alta fidelidad, sin embargo, la presencia de ciertos grupos de puntos fuera de la diagonal sugiere que todavía hay margen para mejorar el ajuste

5 3.- Plots

En esta sección se generarán diversos gráficos con el fin de realizar un análisis exploratorio de los datos.

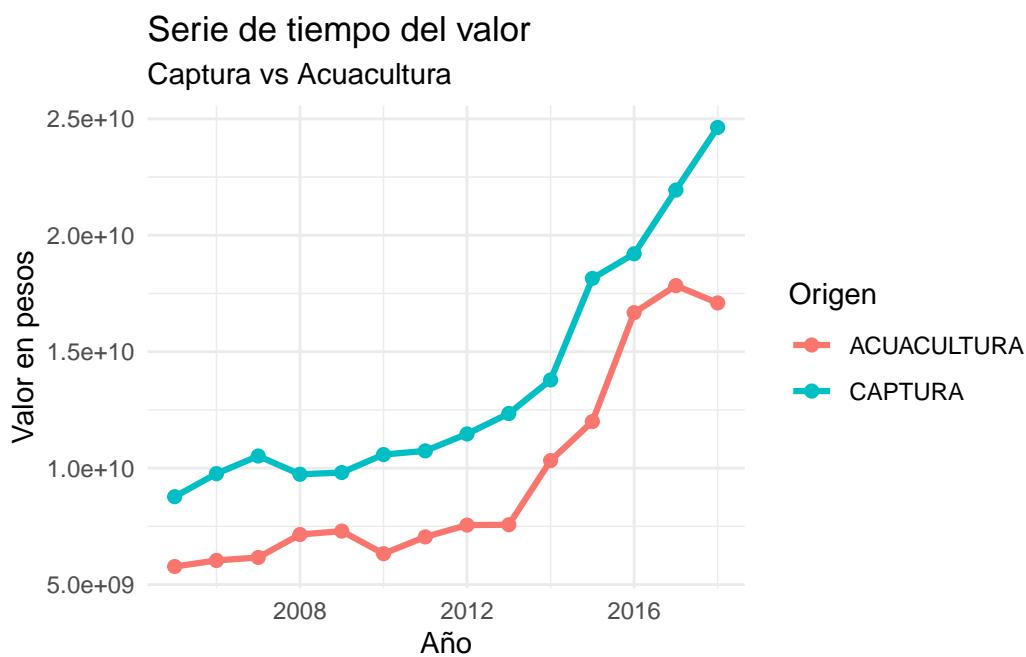
5.1 Series de tiempo

```
fishy_ts <- fishy %>%
  group_by(ANO, ORIGEN) %>%
  summarise(
    valor_total = sum(VALOR_MEXICAN_PESOS, na.rm = TRUE),
    .groups = "drop"
  )
```

```

ggplot(fishy_ts, aes(x = ANO, y = valor_total, color = ORIGEN)) +
  geom_line(linewidth = 1.1) +
  geom_point(size = 2) +
  labs(
    title = "Serie de tiempo del valor",
    subtitle = "Captura vs Acuacultura",
    x = "Año",
    y = "Valor en pesos",
    color = "Origen"
  ) +
  theme_minimal()

```



```

fishy_ts <- fishy %>%
  group_by(ANO, ORIGEN) %>%
  summarise(
    produccion_total = sum(PESO_DESEMBARCADO_KG, na.rm = TRUE),
    .groups = "drop"
  )

```

```

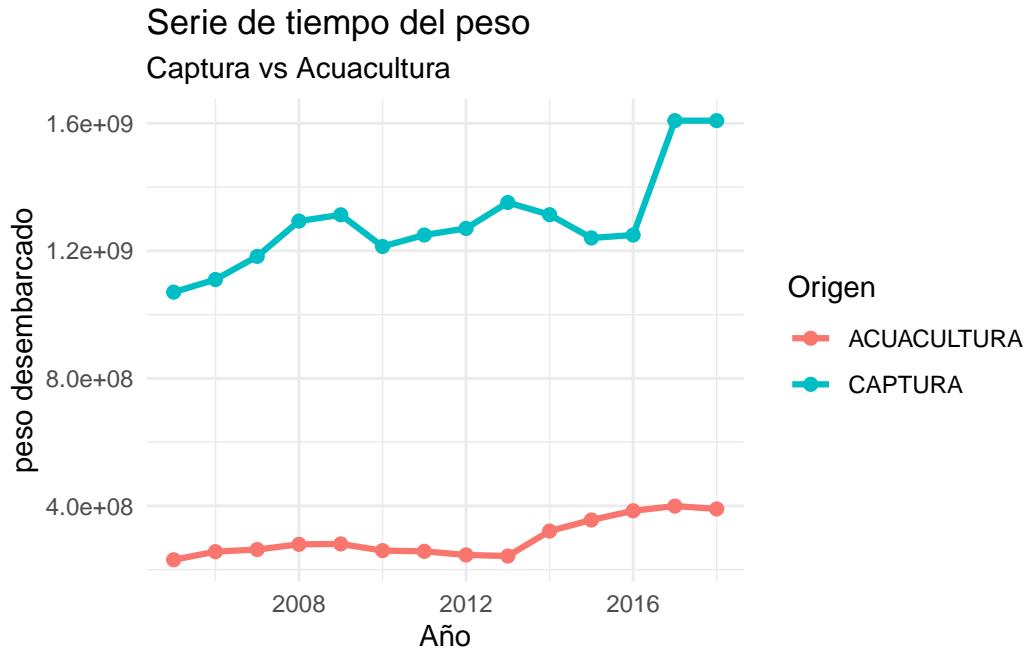
ggplot(fishy_ts, aes(x = ANO, y = produccion_total, color = ORIGEN)) +
  geom_line(linewidth = 1.1) +
  geom_point(size = 2) +

```

```

  labs(
    title = "Serie de tiempo del peso",
    subtitle = "Captura vs Acuacultura",
    x = "Año",
    y = "peso desembarcado",
    color = "Origen"
  ) +
  theme_minimal()

```



Se puede observar que tanto el valor de la captura como el de la acuacultura han experimentado un crecimiento anual constante. Sin embargo, en 2018, se produjo una disminución en el valor de la acuacultura. Por otro lado, la producción por acuacultura ha mostrado un aumento muy leve en comparación con la captura durante los mismos períodos.

```

top_especies <- fishy %>%
  group_by(NOMBRE_PRINCIPAL_ESPECIE) %>%
  summarise(
    peso_total = sum(PESO_DESEMBARCADO_KG, na.rm = TRUE)
  ) %>%
  arrange(desc(peso_total)) %>%
  slice_head(n = 5) %>%
  pull(NOMBRE_PRINCIPAL_ESPECIE)

```

```

top_especies_valor <- fishy %>%
  group_by(NOMBRE_PRINCIPAL_ESPECIE) %>%
  summarise(
    peso_total = sum(VALOR_MEXICAN_PESOS, na.rm = TRUE)
  ) %>%
  arrange(desc(peso_total)) %>%
  slice_head(n = 5) %>%
  pull(NOMBRE_PRINCIPAL_ESPECIE)

```

```

fishy_ts_especie <- fishy %>%
  filter(NOMBRE_PRINCIPAL_ESPECIE %in% top_especies) %>%
  group_by(ANO, NOMBRE_PRINCIPAL_ESPECIE) %>%
  summarise(
    peso_anual = sum(PESO_DESEMBARCADO_KG, na.rm = TRUE),
    .groups = "drop"
  )

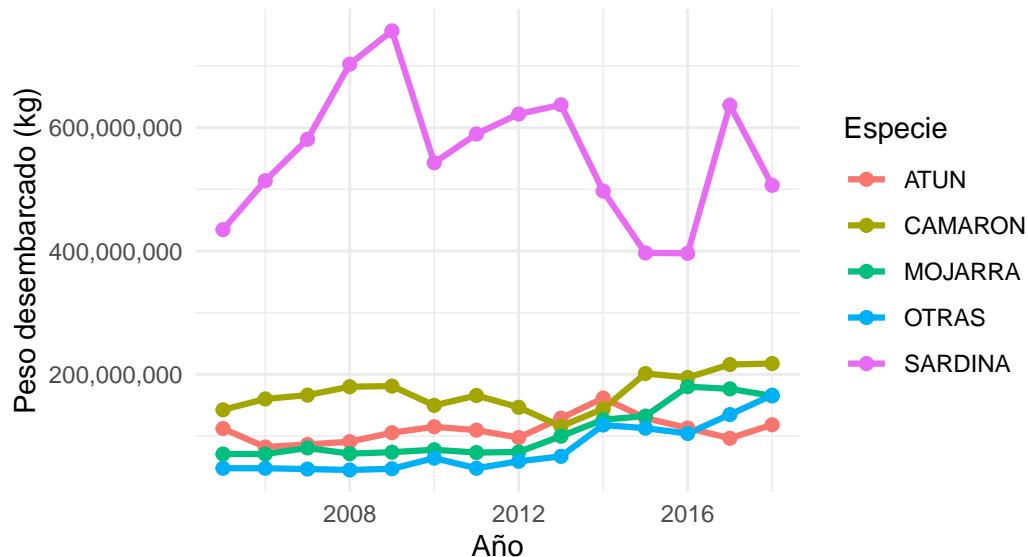
```

```

ggplot(fishy_ts_especie,
       aes(x = ANO, y = peso_anual, color = NOMBRE_PRINCIPAL_ESPECIE)) +
  geom_line(linewidth = 1.1) +
  geom_point(size = 2) +
  scale_y_continuous(labels = comma) +
  labs(
    title = "Serie de tiempo anual por especie",
    subtitle = "Top 5 especies por peso desembarcado",
    x = "Año",
    y = "Peso desembarcado (kg)",
    color = "Especie"
  ) +
  theme_minimal()

```

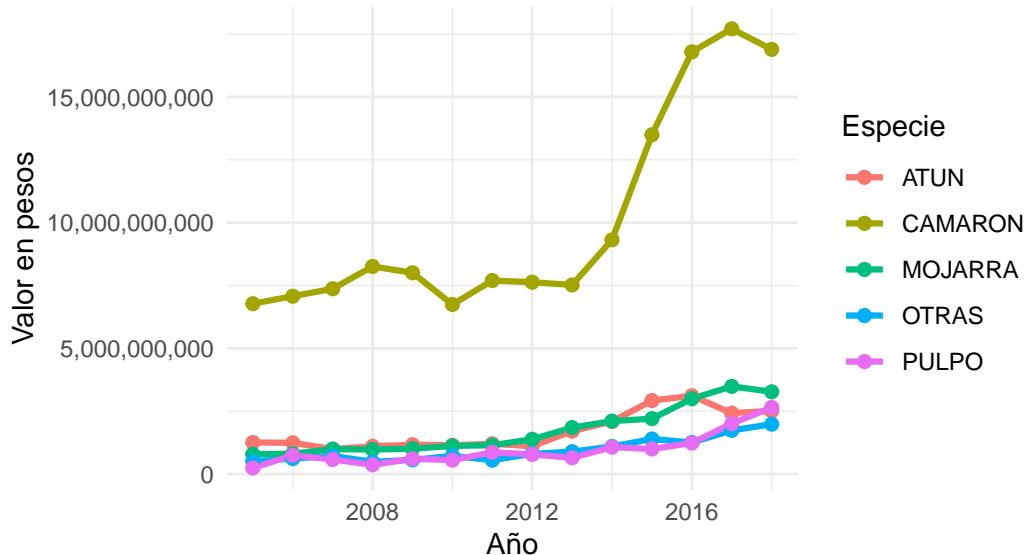
Serie de tiempo anual por especie Top 5 especies por peso desembarcado



```
fishy_ts_especie <- fishy %>%
  filter(NOMBRE_PRINCIPAL_ESPECIE %in% top_especies_valor) %>%
  group_by(ANO, NOMBRE_PRINCIPAL_ESPECIE) %>%
  summarise(
    valor_anual = sum(VALOR_MEXICAN_PESOS, na.rm = TRUE),
    .groups = "drop"
  )
```

```
ggplot(fishy_ts_especie,
       aes(x = ANO, y = valor_anual, color = NOMBRE_PRINCIPAL_ESPECIE)) +
  geom_line(linewidth = 1.1) +
  geom_point(size = 2) +
  scale_y_continuous(labels = comma) +
  labs(
    title = "Serie de tiempo anual por especie",
    subtitle = "Top 5 especies por valor",
    x = "Año",
    y = "Valor en pesos",
    color = "Especie"
  ) +
  theme_minimal()
```

Serie de tiempo anual por especie Top 5 especies por valor



Se observa que, a pesar de ser la especie con la mayor producción por un margen considerable, la sardina no figura entre las cinco especies con mayor valor. En contraste, especies como el atún, la mojarra y el camarón se mantienen en ambas gráficas, ocupando las mismas posiciones. De estas, el camarón destaca como la especie más valiosa, con una diferencia significativa respecto a las demás.

5.2 Plots 3D

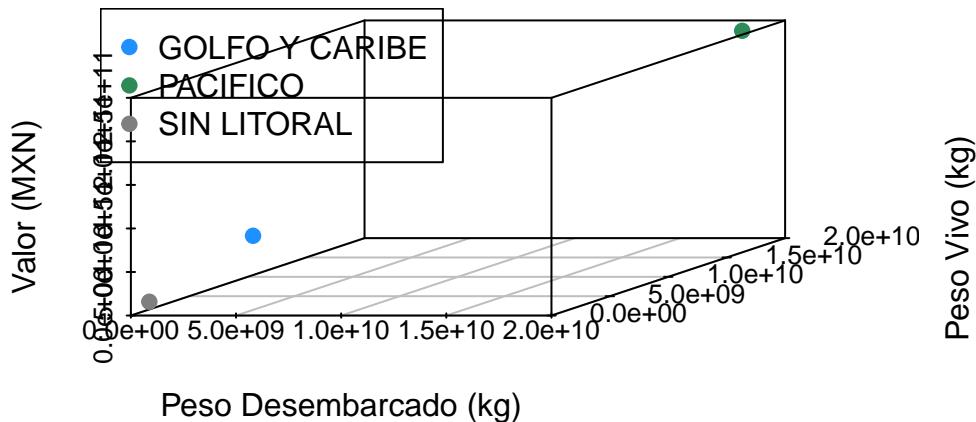
```
fishy_3d <- fishy %>%
  group_by(LITORAL) %>%
  summarise(
    peso_desembarcado = sum(PESO_DESEMBARCADO_KG, na.rm = TRUE),
    peso_vivo = sum(PESO_VIVO_KG, na.rm = TRUE),
    valor = sum(VALOR_MEXICAN_PESOS, na.rm = TRUE),
    .groups = "drop"
  )
colores <- c("dodgerblue", "seagreen", "gray50")
col_litoral <- colores[as.factor(fishy_3d$LITORAL)]
scatterplot3d(
  x = fishy_3d$peso_desembarcado,
  y = fishy_3d$peso_vivo,
```

```

z = fishy_3d$valor,
color = col_litoral,
pch = 19,
main = "Relación 3D entre Peso, Peso Vivo y Valor por Litoral",
xlab = "Peso Desembarcado (kg)",
ylab = "Peso Vivo (kg)",
zlab = "Valor (MXN)"
)
legend("topleft", legend = fishy_3d$LITORAL, col = col_litoral, pch = 19)

```

Relación 3D entre Peso, Peso Vivo y Valor por Litoral



La gráfica ilustra claramente la hegemonía económica y productiva del litoral del Pacífico en el sector pesquero mexicano, seguida por el Golfo y Caribe, dejando a las zonas sin litoral con una participación mínima.

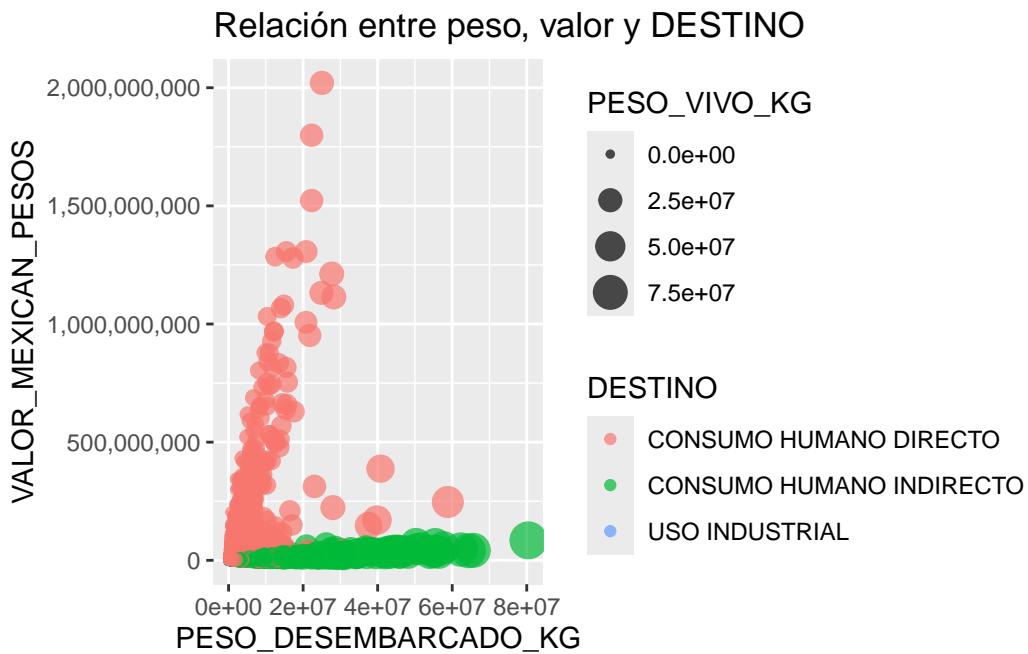
```

fishy_3d_especie <- fishy %>%
  filter(NOMBRE_PRINCIPAL_ESPECIE %in% top_especies) %>%
  group_by(NOMBRE_PRINCIPAL_ESPECIE) %>%
  summarise(
    peso_desembarcado = sum(PESO_DESEMBARCADO_KG, na.rm = TRUE),
    peso_vivo = sum(PESO_VIVO_KG, na.rm = TRUE),
    valor = sum(VALOR_MEXICAN_PESOS, na.rm = TRUE),
    .groups = "drop"
  )

```

5.3 Otros plots

```
ggplot(fishy, aes(x = PESO_DESEMBARCADO_KG,  
                   y = VALOR_MEXICAN_PESOS,  
                   color = DESTINO,  
                   size = PESO_VIVO_KG)) +  
  geom_point(alpha = 0.7) +  
  scale_y_continuous(labels = scales::comma) +  
  labs(title = "Relación entre peso, valor y DESTINO")
```



Esta gráfica revela que las pesquerías con mayor producción son, en su mayoría, aquellas destinadas al consumo humano indirecto. Sin embargo, estas no son las especies de mayor valor en el mercado, ya que las más valiosas corresponden a aquellas destinadas al consumo humano directo.

```
fishy_state <- fishy %>%  
  group_by(ENTIDAD) %>%  
  summarise(  
    peso_total = sum(PESO_DESEMBARCADO_KG, na.rm = TRUE),  
    valor_total = sum(VALOR_MEXICAN_PESOS, na.rm = TRUE),  
    .groups = "drop"  
)
```

```

ggplot(fishy_state, aes(x = 1, y = ENTIDAD, fill = valor_total)) +
  geom_tile(color = "white") +
  scale_fill_gradient(low = "white", high = "dodgerblue") +
  labs(title = "Valor total por Estado",
       x = "",
       y = "Estado",
       fill = "Valor (MXN)") +
  theme_minimal() +
  theme(axis.text.x = element_blank(),
        axis.ticks.x = element_blank())

```



En esta gráfica se puede observar que Sonora y Sinaloa son los estados que concentran la mayor producción en términos de valor, seguidos de las Baja Californias. Esto destaca que, a pesar de que el Pacífico en general domina la producción, es la zona del Mar de Cortés la que concentra la mayor parte del valor generado.

6 4.- PCA

```

pca_result <- prcomp(fishy_num, center = TRUE, scale. = TRUE)
summary(pca_result)

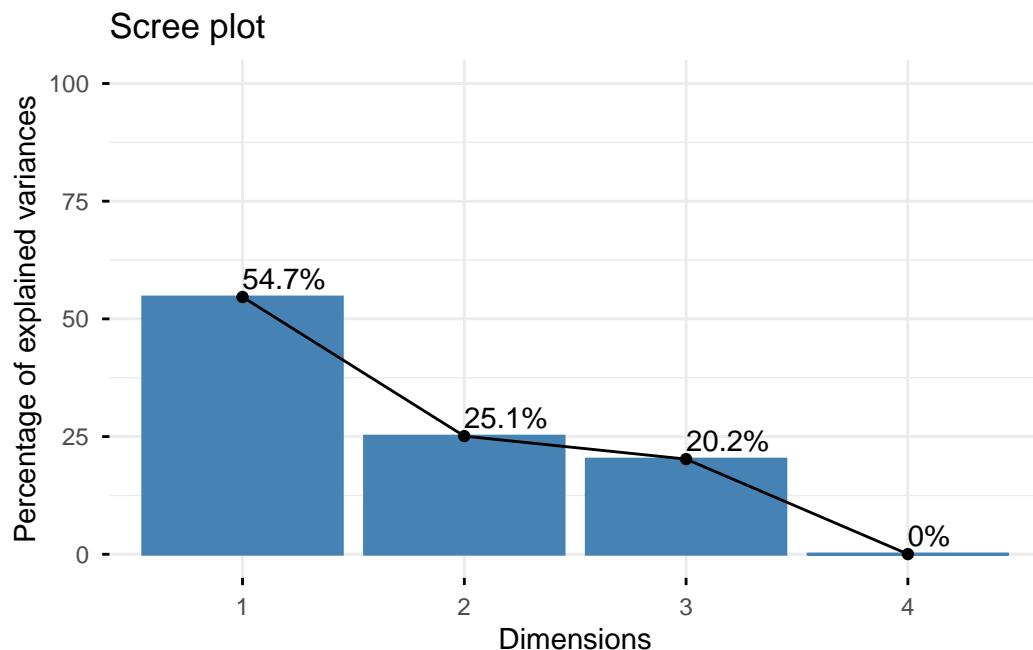
```

Importance of components:

	PC1	PC2	PC3	PC4
Standard deviation	1.4785	1.0019	0.8993	0.03689
Proportion of Variance	0.5465	0.2510	0.2022	0.00034
Cumulative Proportion	0.5465	0.7975	0.9997	1.00000

```
fh <- prcomp(fishy_num,scale. = TRUE)
fviz_eig(fh, addlabels = TRUE, ylim = c(0, 100))
```

Warning in geom_bar(stat = "identity", fill = barfill, color = barcolor, :
Ignoring empty aesthetic: `width`.



El análisis de componentes principales revela lo siguiente:

- **PC1:** es la componente que captura la mayor parte de la varianza, con una desviación estándar total de 1.4785, explica el 54.65% de la varianza de los datos. Esta componente es la más relevante en términos de representar la variabilidad en el conjunto de datos.
- **PC2:** con una desviación estándar de 1.0019, explica el 25.10% de la varianza. Junto con la primera componente, captura un 79.75% de la varianza acumulada. Esta componente proporciona información adicional importante, aunque en menor medida que la primera.

- **PC3:** con una desviación estándar de 0.8993, explica el 20.22% de la varianza, lo que aumenta la proporción acumulada de varianza explicada al 99.97%. Aunque su contribución es significativa, es mucho menor que las dos primeras componentes.
- **PC4:** con una desviación estándar muy baja de 0.03689, tiene una contribución mínima a la varianza, lo que hace que su impacto sea casi insignificante en la reducción de la dimensionalidad.

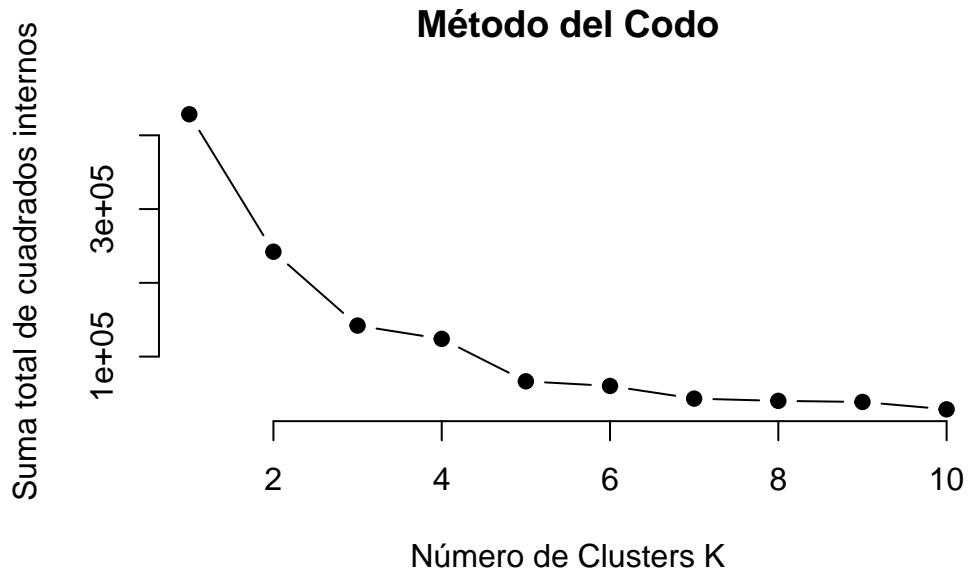
Podemos concluir que las primeras dos componentes principales (PC1 y PC2) explican en conjunto el 79.75% de la varianza, lo que sugiere que al utilizar solo estas dos componentes se puede capturar la mayor parte de la información de los datos.

7 5.- Clustering

```
pca_scores_2D <- pca_result$x[, 1:2]
set.seed(123)
kmeans_result <- kmeans(pca_scores_2D, centers = 2)

wss <- numeric(10)
for (k in 1:10) {
  modelo <- kmeans(pca_scores_2D, centers = k, nstart = 25)
  wss[k] <- modelo$tot.withinss
}

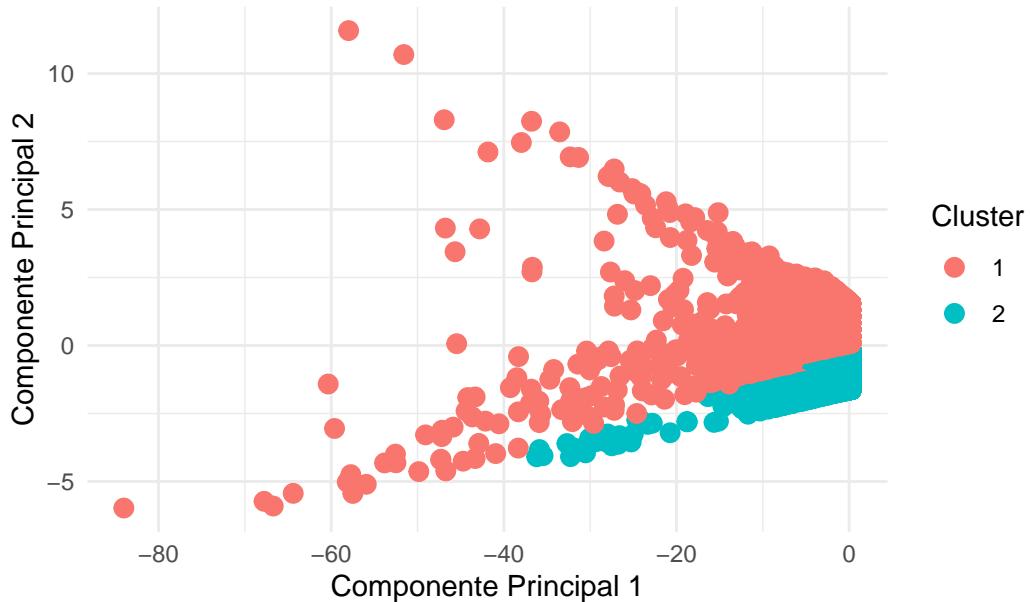
# Dibujar la gráfica
plot(1:10, wss, type = "b", pch = 19, frame = FALSE,
      xlab = "Número de Clusters K",
      ylab = "Suma total de cuadrados internos",
      main = "Método del Codo")
```



El metodo del codo nos indica que el numero de clusters optimo esta entre 2 o 3, procederemos a usar un k=2

```
fishy$Cluster <- factor(kmeans_result$cluster)
ggplot(fishy, aes(x = pca_scores_2D[,1], y = pca_scores_2D[,2], color = Cluster)) +
  geom_point(size = 3) +
  labs(title = "KMEANS en el Espacio PCA", x = "Componente Principal 1",
       y = "Componente Principal 2") +
  theme_minimal()
```

KMEANS en el Espacio PCA



El modelo ha identificado dos clusters claramente diferenciados por color:

- **Cluster 1 (Salmón):** Es el grupo más numeroso y disperso. Incluye tanto la masa crítica de datos cerca del origen como los valores atípicos que se alejan hacia la izquierda y hacia arriba.
- **Cluster 2 (Cian):** Es un grupo mucho más compacto y lineal situado en la parte inferior derecha.

8 Conclusiones

El análisis multivariante realizado sobre la producción pesquera en México (2005-2018) permite extraer las siguientes determinaciones clave:

8.1 1. Modelado y Predicción del Valor Económico

El modelo de **Regresión Lineal Múltiple** alcanzó un coeficiente de determinación ajustado (R^2) de **0.9525**, lo que indica que las variables seleccionadas explican más del 95% de la variabilidad en el valor de la producción.

- **Elasticidad Peso-Valor:** Se identificó una relación proporcional casi unitaria entre el logaritmo del peso desembarcado y el valor comercial.

- **Robustez:** A pesar de la presencia de heterocedasticidad (corregida mediante errores estándar robustos HC1) y una distribución de residuos con colas pesadas, el volumen de datos permitió obtener inferencias estadísticamente significativas que validan la estructura del modelo.

8.2 2. Dinámica de la Producción y Valor

A través de la visualización multivariante y el análisis de series de tiempo, se revelaron contrastes críticos en el sector:

- **Especies “Volumen vs. Valor”:** Existe una disparidad notable representada por la Sardina (alta producción, bajo valor) frente al Camarón (líder indiscutible en generación de riqueza).
- **Dominio Regional:** El litoral del Pacífico, específicamente la zona del **Mar de Cortés** (Sonora y Sinaloa), se consolida como el motor económico de la pesca nacional, concentrando los mayores niveles de peso desembarcado y valor comercial.

8.3 3. Eficiencia en la Reducción de Dimensionalidad (PCA)

El Análisis de Componentes Principales demostró que es posible simplificar la estructura del dataset de 4 variables numéricas a solo 2 componentes principales, reteniendo el 79.75% de la varianza total.

8.4 4. Segmentación del Sector (Clustering)

La aplicación de K-means sobre el espacio transformado del PCA permitió identificar dos perfiles de producción claramente diferenciados:

- **Cluster 1:** Representa la mayor parte de las operaciones, con una alta dispersión que sugiere una gran variedad de especies y escalas.
- **Cluster 2:** Un grupo compacto que representa nichos de producción con comportamientos lineales y predecibles, posiblemente asociados a pesquerías industriales de alta eficiencia.

8.5 Limitaciones y Trabajo Futuro

Si bien el modelo es altamente explicativo, el sesgo en la normalidad de los residuos y la presencia de *outliers* indican que el valor de la pesca no solo depende de variables físicas y geográficas, sino también de factores externos no incluidos en este dataset, como fluctuaciones de precios internacionales, vedas estacionales y fenómenos climáticos. Para futuras investigaciones, se recomienda integrar variables macroeconómicas.

9 Bibliografia

- 1.- <https://www.kaggle.com/datasets/viclopez0306/fisheries-production-in-mexico-2005-to-2018>