Punto 1° Parcial: Ajuste de un modelo de R.L.S

Universidad Nacional de Colombia
Analisís de Regresión 2022-1S
Medellín, Colombia
2022

Daniel Villa 1005087556 Juan Pablo Vanegas 1000640165

Contents

1 Objetivos

Crear un modelo ajustado de R.L.S. por el cual predecimos el peso de los huevos (de gallina) por medio de la altura o el diámetro de este utilizando el software estadístico R

1.1 Objetivos específicos

- Interpretar los parámetros del modelo.
- Determinar si el efecto de la altura o diámetro sobre el peso de los huevos es significativo.
- Interpretar nuestro R^2.
- Calcular un I.C. al $100 * (1 - \alpha)\%$ para β_1.
- Calcular un intervalo de predicción al $100 * (1 - \alpha)\%$ para una altura o diámetro dado.
- Calcular un I.C. al $100 * (1 - \alpha)\%$ para una altura o diámetro dado.
- Validar los supuestos del modelo.
- Aplicar la prueba de falta de ajuste.
- Analizar los I.C's y I.P's para valores promedios y futuros de la respuesta.

2 Antecedentes Relevantes

Utilizando un *pie de rey o calibre* tomamos los datos de una cubeta de huevos tipo AA Marca kikes.

Gracias a que las empresas que distribuyen y producen los huevos, tienen muy controlados el peso, altura y diámetro por los controles de calidad, se esperan unos rangos de variabilidad angostos, esto se nota en este experimento pues al ser todos huevos de un mismo tipo y una misma marca se nota el control sobre ellos.

3 Variables de respuesta:

En nuestro caso será el **peso** para ajustar un modelo para predecir por medio de nuestra variable predictora el peso de un huevo.

4 Variable de Control:

Ya que haremos dos modelos tendremos dos variables de control; sin relacionarse entre ellas, es decir, serán dos R.L.S.

- 1er caso: Altura
- 2do Caso: Ancho

5 Creación de los modelos

La figura (1) nos muestra los 30 datos recolectados de los huevos medidos:

ID	Altura	Ancho	Peso
1	53.17	42.93	55.1
2	58.22	41.88	58
3	53.11	42.44	55
.	.	.	.
.	.	.	.
28	54.55	41.17	54.5
29	55.43	41.76	52.4
30	53.22	41.22	51.15

Definimos nuestras variables de forma aritmética

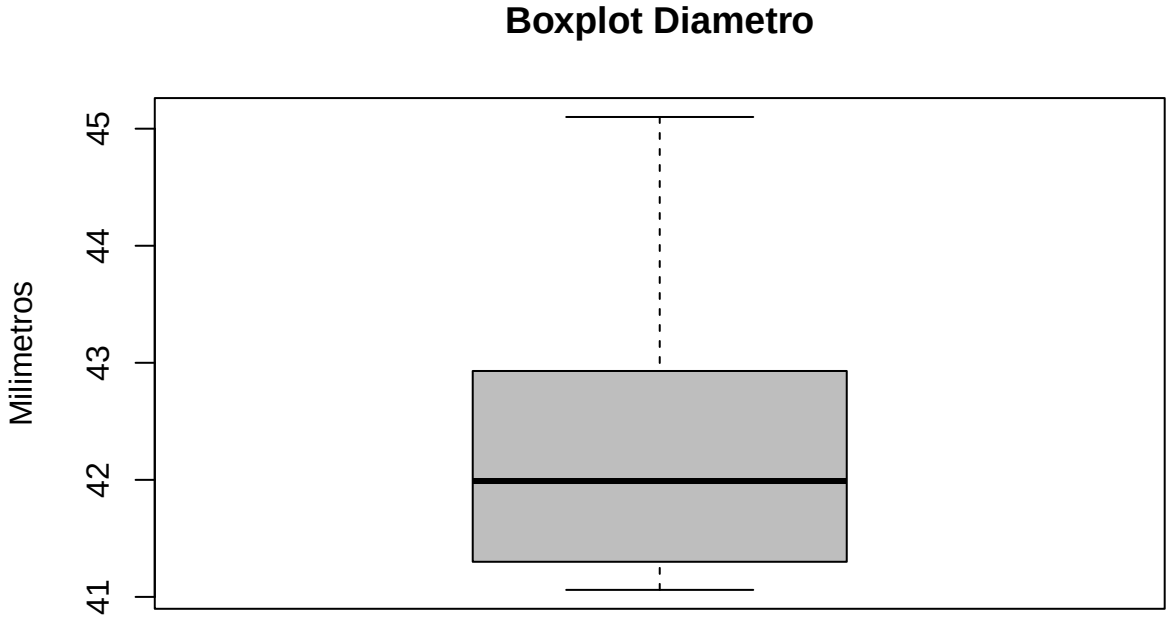
- Y = peso
- X = Altura ó Ancho
- Se midieron $n = 30$ huevos.

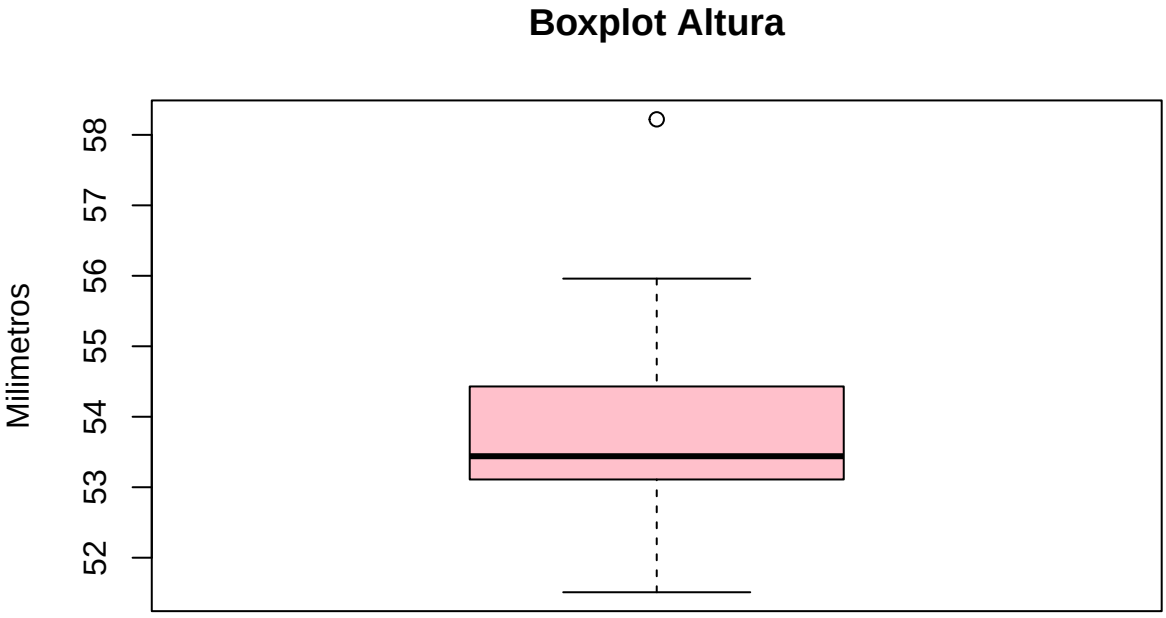
5.1 Resumen de los datos:

Hacemos un resumen de los datos para mirar de primera mano nuestros valores estadísticos comunes (μ, σ, mediana, máximo y mínimo valor)

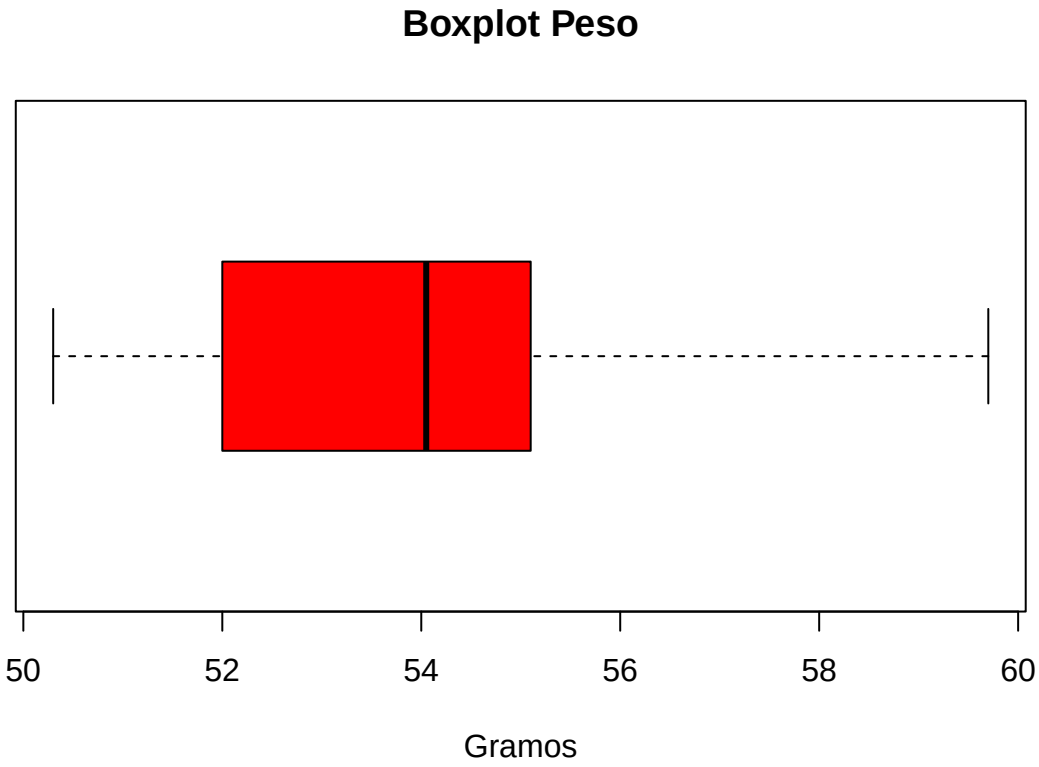
	largo	ancho	peso
	Min. :51.51	Min. :41.06	Min. :50.30
	1st Qu.:53.12	1st Qu.:41.32	1st Qu.:52.10
	Median :53.44	Median :41.99	Median :54.05
	Mean :53.75	Mean :42.26	Mean :54.05
	3rd Qu.:54.39	3rd Qu.:42.89	3rd Qu.:55.08
	Max. :58.22	Max. :45.10	Max. :59.70

Ahora para ser un poco más minuciosos haremos un *"boxplot"* de los datos para observar como se comportan nuestras variables de control y respuesta.



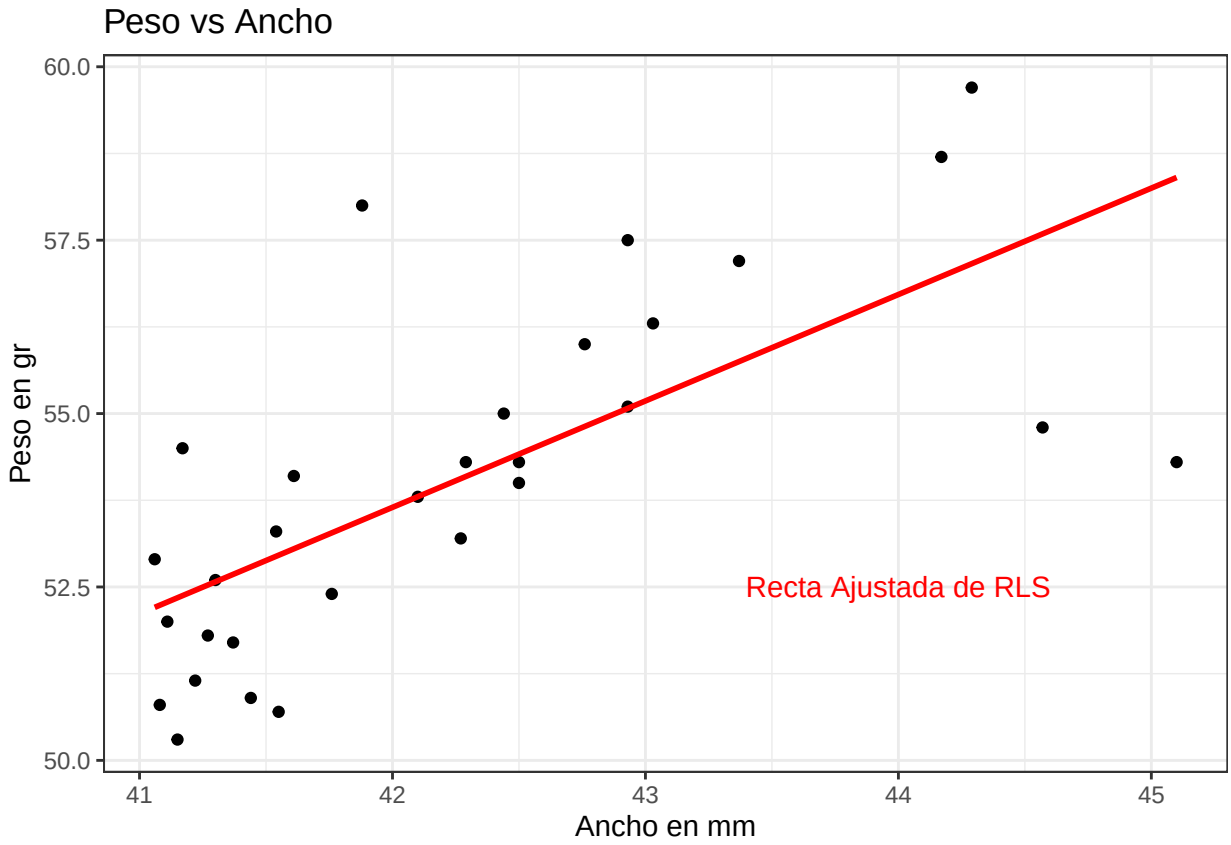


observamos un dato atípico en nuestra variable del Altura (en mm)



Ahora viendo el comportamiento individual de los datos, observaremos el comportamiento en grafico de disperción para comparar variables con otras.

5.2 Grafico de Disperción Ancho ~ Peso



La relación entre el *Ancho* y el *Peso* se puede aproximar utilizando un modelo de regresión lineal.

Dado que el modelo de RLS puede aproximar a la relación entre el *Ancho* y el *Peso*, se puede plantear el modelo:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{con } \varepsilon_1 \overset{\text{i.i.d}}{\sim} N(0, \sigma^2), i = 1, ..., 30$

5.3 Ajuste del modelo, estadísticos de resumen y Tabla ANOVA

Model Coefficients - peso

Predictor	Estimate	SE	t	p
Intercept	−10.78	12.551	−0.859	0.398
ancho	1.53	0.297	5.167	< .001

Figure 1: figura_6

Del resultado anterior vemos que el modelo ajustado está dado por:

$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i = -10.78066 + 1.53402 x_i$

6 Interpretación de los parámetros estimados del modelo.

- Interpretación de $\hat{\beta_0}$: Es el valor promedio de la respuesta cuando la predictora toma el valor de cero. Esto sólo si $x = 0, \in [x_{min}, x_{max}]$
- Interpretación de $\hat{\beta_1}$: que por cada unidad de aumento en la predictora el promedio de la respuesta cambia en $\hat{\beta_1}$ u.

6.1 Prueba de significancia de la regresión

se quiere probar que:

$H_0 : \beta_1 = 0$ vs. $\beta_1 \neq 0$

- Hay una tabla que se conoce como tabla de parámetros estimados, identificada en R como coefficients (la tabla de la figura 6).

Model Coefficients - peso

Predictor	Estimate	SE	t	p
Intercept	−10.78	12.551	−0.859	0.398
ancho	1.53	0.297	5.167	< .001

Figure 2: figura_6

Cuyas columnas son:

- **Parámetro:** con valores (intercept) ó β_0 y Ancho ó β_1.
- **Estimación:** con valores $\hat{\beta}_0$ y $\hat{\beta}_1$.
- **Error estándar:** con valores $se(\hat{\beta}_0)$ y $se(\hat{\beta}_1)$.
- **Valor t:** Valores de estadísticos de prueba para la significancia de β_0 y β_1, respectivamente.
- **Valor P:** Valores p para la prueba de significancia de β_0 y β_1, respectivamente.

De esta tabla solo nos interesa los valores de las dos ultimas columnas y especificamente en la segunda fila (β_1), de esta se sacan el valor del estadístico $t = 5.166734$ y nuestro $P_{valor} = 1.758177e - 05$.

como $P_{valor} < \alpha = 0.05$ entonces rechazo 0 y concluyo que en efecto el Ancho de los huevos sobre el promedio de sus pesos es significativo.

6.2 Coeficiente de determinación para el modelo de regresion

se sabe que:

$R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SSR+SSE}$

De la tabla ANOVA se obtiene:

El $48,8\%$ **de la variabilidad total del Peso es explicado por el Ancho.**

Model Fit Measures

Model	R^2
1	0.488

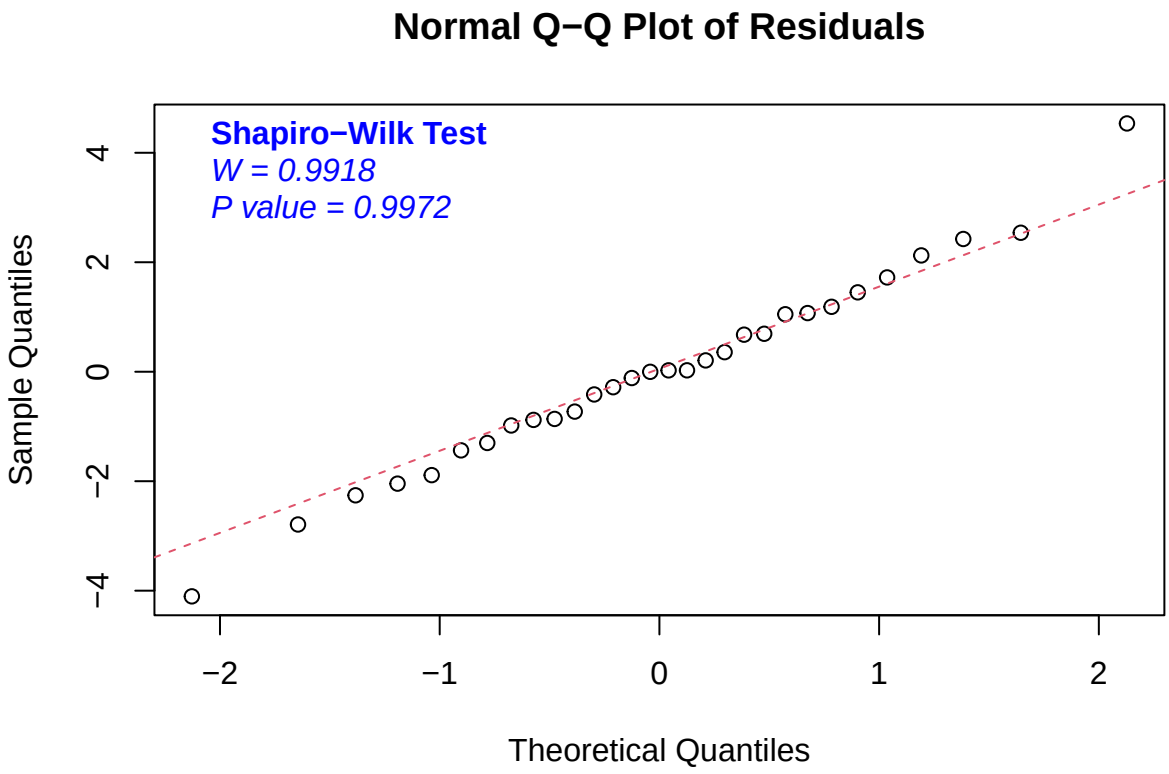
Figure 3: figura_8

7 Validación del modelo de regresión

Se deben validar los siguentes supuestos:

- Normalidad
- Varianza Constante
- Linealidad

7.1 Supuesto de normalidad - Gráfica de normalidad y prueba de Shapiro-Wilk

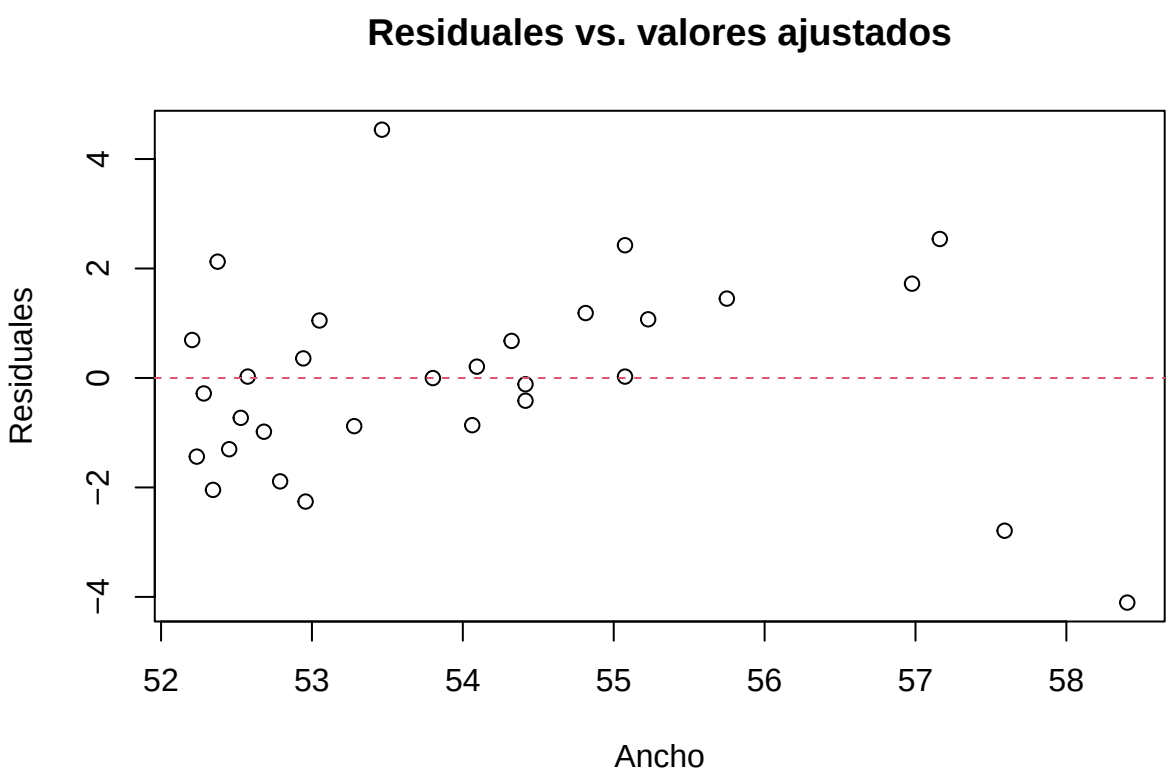


Se quiere probar:

$H_0 : \varepsilon_i \sim Normal$ vs $H_1 : \varepsilon_i \nsim Normal$

En la gráfica se observa que los datos se alinean mejor en el centro, sin despreciar los extremos, por lo tanto se concluye que el supuesto se cumple. Esto es ratificado por la prueba de $Shapiro - Wilk \ \ test$, ya que $p - valor = 0.9972 > 0.05 = \alpha$ y, nos lleva a no rechazar H_0

8 Validación del supuesto de varianza constante




```
## integer(0)
```

se quierprobar que:

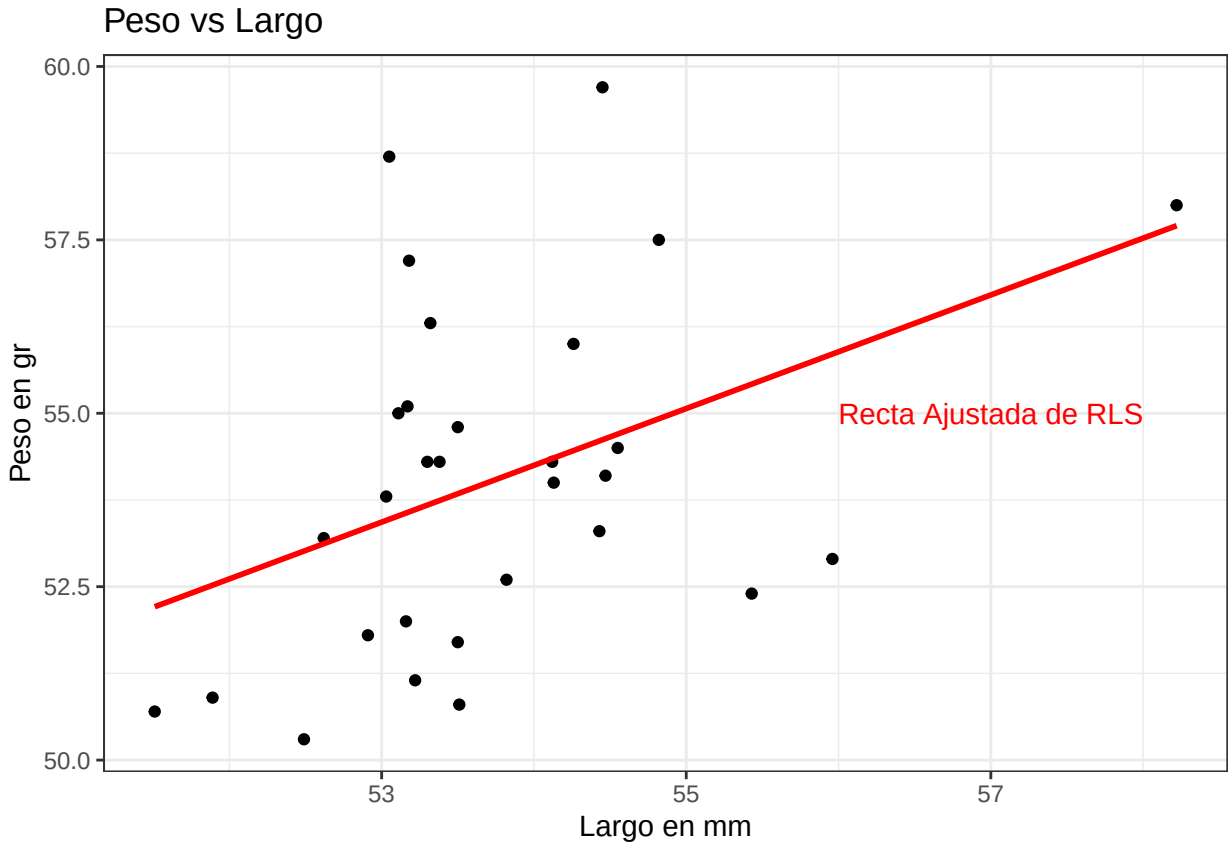
$H_0 : V[\varepsilon_i] = \sigma^2$ vs $H_1 : V[\varepsilon_i] \neq \sigma^2$

Como el patrón de la nube de puntos en la gráfica ε_i vs. $\hat{y}$ no se asemeja a un rectángulo ni a una u, entonces se concluye que el modelo de regresión no tiene varianza constante

8.1 Conclusión:

Como el supuesto para los residuales no se cumplen (para el caso de la varianza constante), decimos que el modelo de regresión lineal del peso respecto al ancho no sirve para realizar predicciones/estimaciones.

9 Modelo N°2 (Peso~Altura)



La relación entre la *Largo* y el *Peso* se puede aproximar utilizando un modelo de regresión lineal.

Dado que el modelo de RLS puede aproximar a la relación entre la *Largo* y el *Peso*, se puede plantear el modelo:

$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{con } \varepsilon_1 \overset{\text{i.i.d}}{\sim} N(0, \sigma^2), i = 1, ..., 30$

9.1 Ajuste del modelo, estadísticos de resumen y Tabla ANOVA

Model Coefficients - peso

Predictor	Estimate	SE	t	p
Intercept	10.031	18.026	0.556	0.582
largo	0.819	0.335	2.442	0.021

Figure 4: figura_9

Del resultado anterior vemos que el modelo ajustado está dado por:

$\hat{y_i} = \hat{\beta_0} + \hat{\beta_1} x_i = 10.0310 + 0.8189 x_i$

9.1.1 Interpretación de los parámetros estimados del modelo

- Interpretación de $\hat{\beta_0}$: Es el valor promedio de la respuesta cuando la predictora toma el valor de cero. Esto sólo si $x = 0, \in [x_{min}, x_{max}]$

- Interpretación de $\hat{\beta}_1$: que por cada unidad de aumento en la predictora el promedio de la respuesta cambia en $\hat{\beta}_1$ *u*.

9.1.1.1 Prueba de significancia de la regresión:

se quiere probar que: $H_0 : \beta_1 = 0$ vs. $\beta_1 \neq 0$

Con la tabla de parámetros estimados, identificada en R como coefficients.

Model Coefficients - peso

Predictor	Estimate	SE	t	p
Intercept	10.031	18.026	0.556	0.582
largo	0.819	0.335	2.442	0.021

Figure 5: figura_9

En particular, nos enfocaremos en tal prueba para β_1 (segunda fila).De ahí que el valor del estadístico es 2.442 , y el valor de $P_{valor} = 0.021$

como $P_{valor} < \alpha = 0.05$ entonces rechazo H_0 y concluyo que en efecto el largo de los huevos sobre el promedio de sus pesos es significativo.

9.2 Coeficiente de determinación para el modelo de regresion

Model Fit Measures

Model	R^2
1	0.176

Figure 6: figura_10

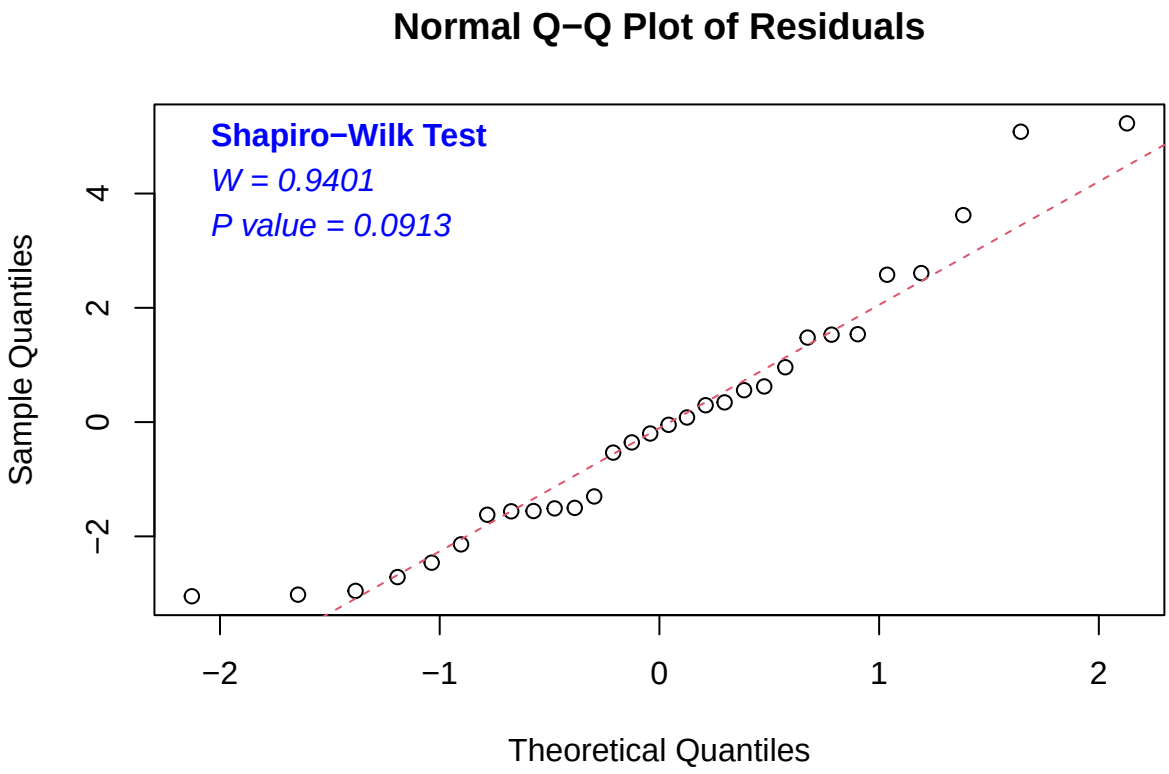
\

El 17.6% de la variabilidad total del Peso es explicado por el largo.

9.3 Validación del modelo de regresión

Se deben validar los siguentes supuestos:

- Normalidad
- Varianza Constante
- Linealidad

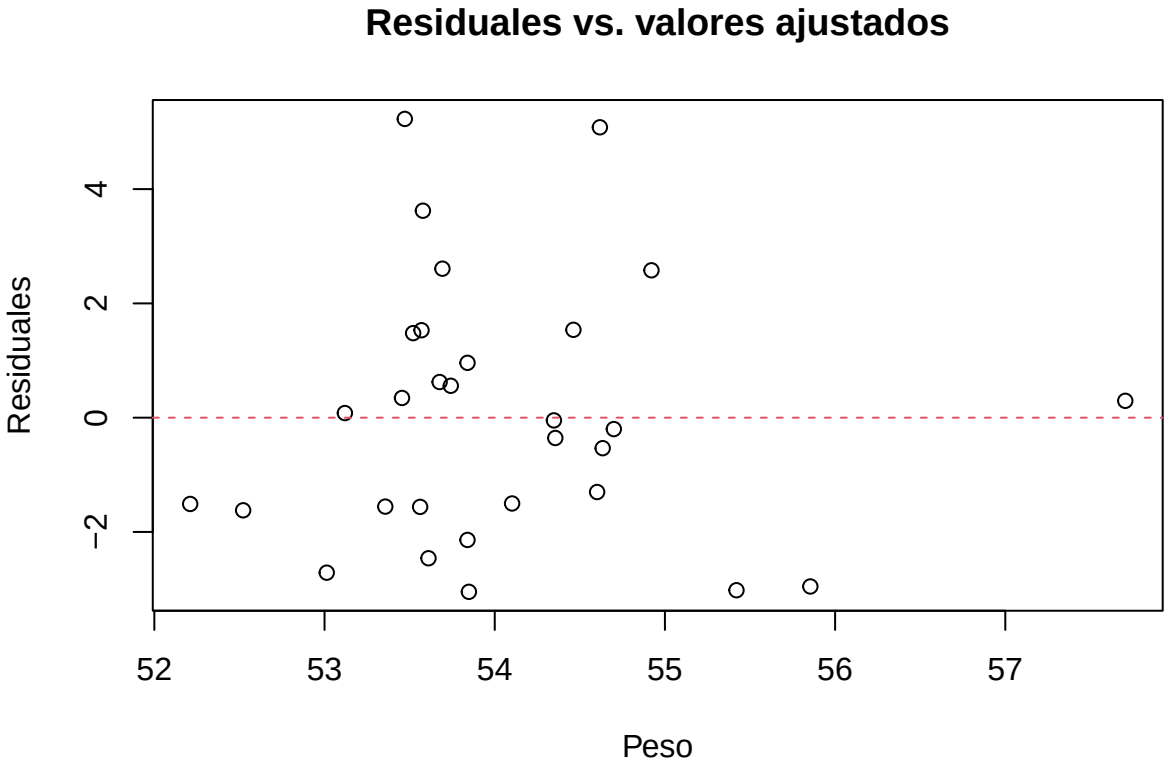


Se quiere probar:

$H_0 : \varepsilon_i \sim Normal$ vs $H_1 : \varepsilon_i \nsim Nomal$

De la grafica anterior podemos ver que el modelo N°1 se ajustaba mejor a este test de prueba de normalidad, es decir, que los datos estan un poco dispersos apriori, pero gracias al P_{valor} podemos ver que si cumple ya que este es mayor al $\alpha = 0.05$; entonces no se rechaza H_0 y se concluye que los errores son normales.

9.4 Validación del supuesto de varianza constante




```
## integer(0)
```

donde se quiere probar:

$H_0 : V[\varepsilon_i] = \sigma^2$ vs $H_1 : V[\varepsilon_i] \neq \sigma^2$

De la gráfica se observa que el patrón formado por la nube de puntos no se aleja mucho de un patrón rectangular, de manera que se puede concluir que el supuesto de varianza constante se cumple.

9.5 Conclusión

Como los *supuestos de normalidad* y *varianza constante* de los residuales se cumplen podemos decir que el modelo de regresion lineal sirve para realizar estimaciones/predicciones.

para terminar se dice que la altura o el largo de los huevos sirve para prededir su altura.

10 Apéndice

10.1 Lista de figuras

Figura:

1. Tabla de los 30 datos de la base de datos
2. tabla de un resumen de las variables numericas de la base de datos
3. Grafico (Boxplot) del diametro
4. Grafico (Boxplot) de la altura
5. Grafico (Boxplot) del peso
6. Grafico (Scatter plot) Ancho vs Peso

7. tabla de coeficientes del modelo N°1
8. Tabla del R^2 del modelo ajustado N°1
9. Grafico (QQplot) de los residuales
10. Grafico (Scatter plot) Valores ajustados vs Residuales
11. Grafico (Scatter plot) Largo vs Peso
12. tabla de coeficientes del modelo N°2
13. Tabla del R^2 del modelo ajustado N°2
14. Grafico (QQplot) de los residuales
15. Grafico (Scatter plot) Valores ajustados vs Residuales

10.2 Codigo:

```
## ----message=FALSE, warning=FALSE, include=FALSE----
library(magrittr)
library(tidyverse)
library(knitr)
library(kableExtra)
library(janitor)

myQQnorm <- function(modelo, student = F, ...){
  if(student){
    res <- rstandard(modelo)
    lab.plot <- "Normal Q-Q Plot of Studentized Residuals"
  } else {
    res <- residuals(modelo)
    lab.plot <- "Normal Q-Q Plot of Residuals"
  }
  shapiro <- shapiro.test(res)
  shapvalue <- ifelse(shapiro$p.value < 0.001, "P value < 0.001",
  paste("P value = ", round(shapiro$p.value, 4), sep = ""))
  shapstat <- paste("W = ", round(shapiro$statistic, 4), sep = "")
  q <- qqnorm(res, plot.it = FALSE)
  qqnorm(res, main = lab.plot, ...)
  qqline(res, lty = 2, col = 2)
  text(min(q$x, na.rm = TRUE), max(q$y, na.rm = TRUE)*0.95, pos = 4,
  'Shapiro-Wilk Test', col = "blue", font = 2)
  text(min(q$x, na.rm = TRUE), max(q$y, na.rm = TRUE)*0.80, pos = 4,
  shapstat, col = "blue", font = 3)
  text(min(q$x, na.rm = TRUE), max(q$y, na.rm = TRUE)*0.65,
  pos = 4, shapvalue, col = "blue", font = 3)
}


## ----message=TRUE, warning=TRUE, include=FALSE----
# Lectura de los datos:
datos <- read_delim("eggs.csv", delim = ";") %>% clean_names()


## ----echo=FALSE-----------------------
## figura 1
```

```
kableExtra::kable(rbind(head(datos,n=3),rep(".",ncol(datos)),
rep(".",ncol(datos))
,tail(datos,n=3)),col.names = c("ID","Altura","Ancho","Peso"),
digits = 3,align = 'c')


## ----echo=FALSE-------------------------
## figura 2
datos %>% select(-"huevos") %>% summary() %>% kable(align = 'c')


## ----echo=FALSE-------------------------
#figura 3 diametro
boxplot(datos$ancho, main = "Boxplot Diametro", col = "Grey",
ylab = "Milimetros")

#figura 4 altura
boxplot(datos$largo,main="Boxplot Altura",col="pink", ylab = "Milimetros")


## ----echo=FALSE, message=FALSE, warning=FALSE----
#figura 5
boxplot(datos$peso,main="Boxplot Peso",col="red", xlab = "Gramos",
horizontal = T)


## ----echo=FALSE, message=FALSE, warning=FALSE----
attach(datos)
ggplot(datos, aes(x=ancho, y=peso))+
  geom_point()+geom_smooth(method = "lm", color="red",se=F)+
  theme_bw()+
  labs(x="Ancho en mm", y="Peso en gr", title = "Peso vs Ancho")+
  annotate(geom = "text", y=52.5, x=44, label="Recta Ajustada de RLS",
  color="red")


## ----echo=FALSE, message=FALSE, warning=FALSE----
lm(peso~ancho,data = datos) %>% myQQnorm()


## ----echo=FALSE, message=FALSE, warning=FALSE----
modelo1 <- lm(peso~ancho, data=datos)
plot(fitted(modelo1), residuals(modelo1), xlab = "Ancho",
ylab = "Residuales", main = "Residuales vs. valores ajustados")+
abline(h = 0, lty = 2, col = 2)


## ----echo=FALSE, message=FALSE, warning=FALSE----
ggplot(datos, aes(x=largo, y=peso))+
  geom_point()+geom_smooth(method = "lm", color="red",se=F)+
  theme_bw()+
  labs(x="Largo en mm", y="Peso en gr", title = "Peso vs Largo")+
  annotate(geom = "text", y=55, x=57, label="Recta Ajustada de RLS", color="red")
```

```
## ----echo=FALSE, message=FALSE, warning=FALSE----
modelo2 <- lm(peso~largo,data=datos)
modelo2 %>% myQQnorm()


## ----echo=FALSE, message=FALSE, warning=FALSE----
plot(fitted(modelo2), residuals(modelo2), xlab = "Peso",
     ylab = "Residuales", main = "Residuales vs. valores ajustados")+
abline(h=0,lty = 2, col = 2)
```

11 Referencias

[**1**] Coleman, D. E., & Montgomery, D. C. (1993). A Systematic Approach to Planning for a Designed Industrial Experiment. Technometrics, 35(1), 1–12. https://doi.org/10.2307/1269280

[**2**] The jamovi project (2021). jamovi. (Version 2.2) [Computer Software]. Retrieved from https://www.jamovi.org.

[**3**] R Core Team (2021). R: A Language and environment for statistical computing. (Version 4.0) [Computer software]. Retrieved from https://cran.r-project.org. (R packages retrieved from MRAN snapshot 2021-04-01).

[**4**] Fox, J., & Weisberg, S. (2020). car: Companion to Applied Regression. [R package]. Retrieved from https://cran.r-project.org/package=car.