UNAP
FINESI

IV Semestre

Modelos Discretos

Evaluación de Supuestos en Regresión Logística

Mtr. Alcides Ramos Calcina

Supuestos en Regresión Logística

Mtr. Alcides Ramos Calcina

1. Linealidad

- Uno de los supuestos más importantes en regresión logística es que la relación entre el logit o log-odds de la variable respuesta y cada variable predictora o variable independiente es **lineal**

- El supuesto de linealidad se verifica únicamente para las variables numéricas continuas que se tengan en el modelo.

- El logit es el logaritmo del odds ratio, en donde p = Probabilidad de éxito

$$logit(p) = log\left(\frac{p}{1-p}\right)$$

- En este caso utilizaremos el tes de Box Tidwell.

1. Linealidad

Tes de Box Tidwell

- Para verificar este supuesto, es posible usar el Test Box-Tidwell, la cual consiste en determinar la existencia de una relación lineal entre cada variable predictora y el logaritmo de la variable de respuesta.

- En R se realiza por medio de la función **`boxTidwell`** del paquete **`car`**. El test se plantea tomando como hipótesis nula el cumplimiento del supuesto de linealidad.

  H_0: La relación es lineal

  H_1: La relación no es lineal

1. Linealidad

- Según esto, realizaremos la prueba para las variables continuas edad, colesterol y triglicéridos, con los siguientes resultados:

- **Edad**

```
logodds <- modelo2$linear.predictors
boxTidwell(logodds ~ datos$edad)
 MLE of lambda Score Statistic (z) Pr(>|z|)
       0.81462             -0.4561   0.6483

iterations =  2
```

En la salida se muestra el valor *p* asociado al estadístico de prueba del test de Box Tidwell, a partir del cual, no se rechaza ($p = 0.6483 > 0.05$) el supuesto de linealidad entre el logit y la variable Edad.

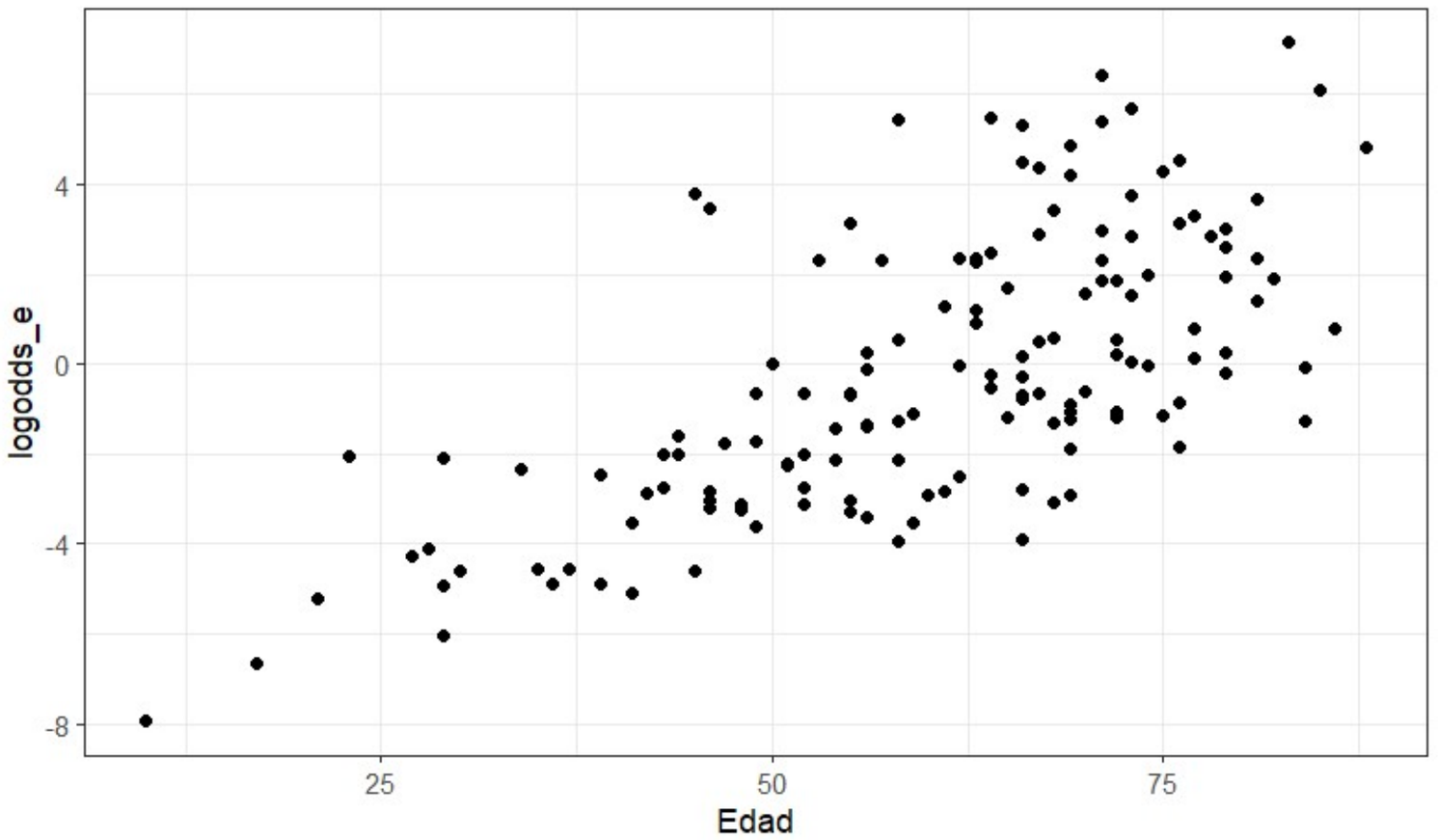
1. Linealidad

- Así mismo, es posible apreciar el gráfico de la edad vs. el logaritmo de las probabilidades estimadas:

```
logEdad <- data.frame(logodds, Edad = datos$edad)
ggplot(data = logEdad, aes(x = Edad, y = logodds)) +
  geom_point() +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

Se muestra la siguiente figura:

1. Linealidad



El diagrama de dispersión de la Figura, parece tener un patrón lineal creciente fuerte.

1. Linealidad

- **Colesterol**

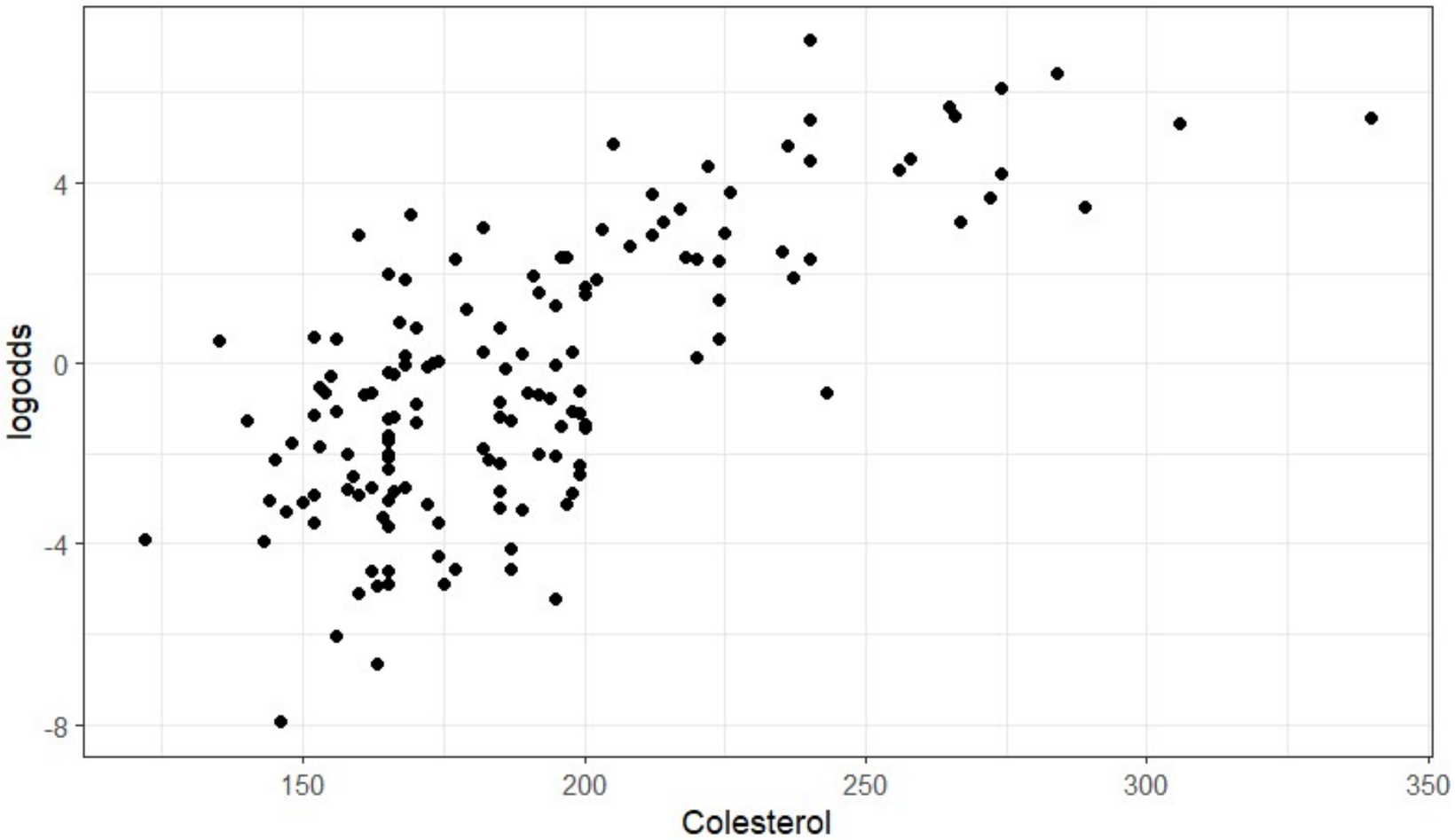
```
logodds <- modelo2$linear.predictors
boxTidwell(logodds ~ datos$coles)
 MLE of lambda Score Statistic (z) Pr(>|z|)
       0.78061             -0.4453   0.6561

iterations =  6
```

En este caso el valor *p* asociado al estadístico de prueba del test de Box Tidwell, es mayor a 0.05, por tanto, no se rechaza el supuesto de linealidad entre el *logit* y la variable Colesterol.

Verificamos gráficamente:

1. Linealidad



El diagrama de dispersión de la Figura, presenta también un patrón lineal creciente.

1. Linealidad

- **Trigliceridos**

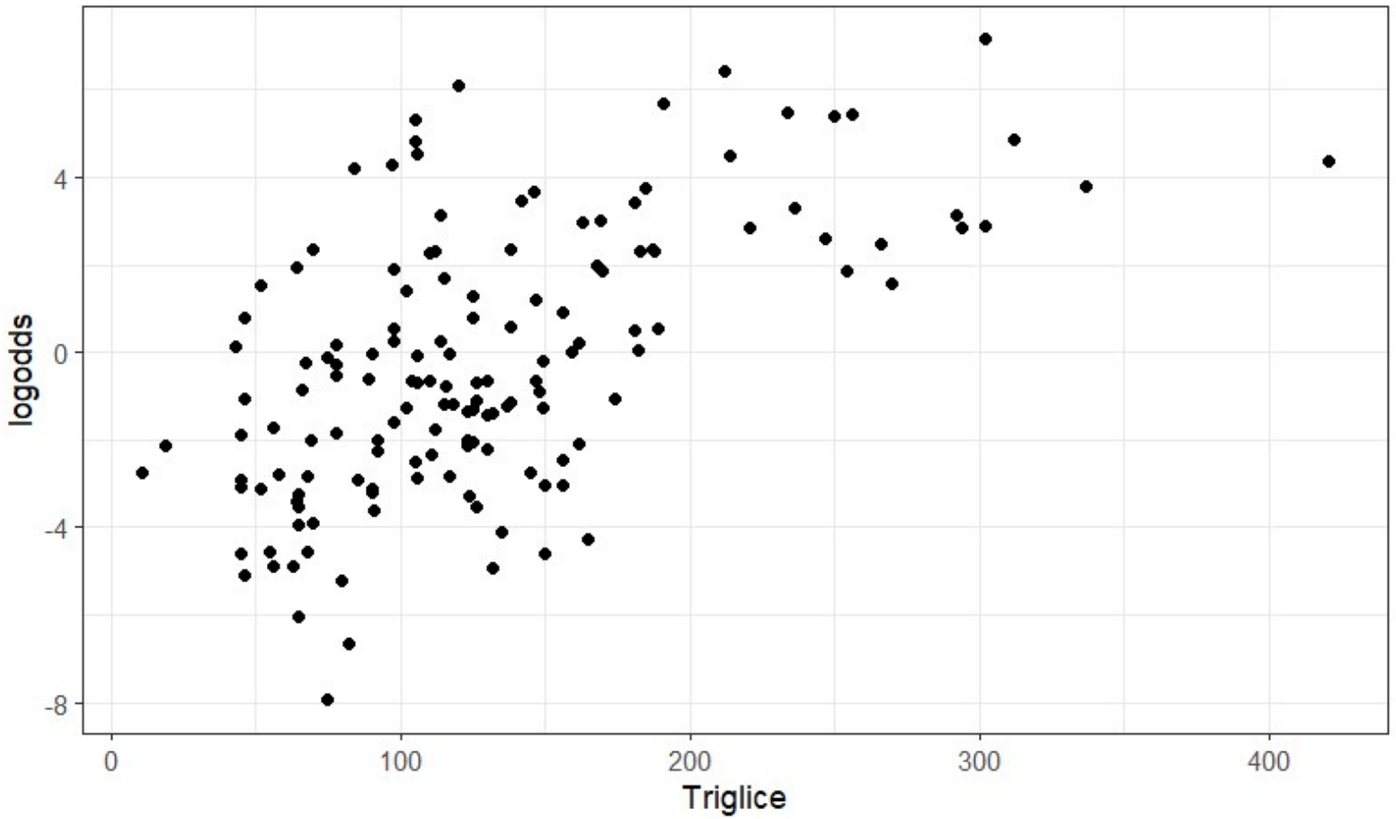
```
logodds <- modelo2$linear.predictors
boxTidwell(logodds ~ datos$trigl)
 MLE of lambda Score Statistic (z) Pr(>|z|)
       0.87172              -0.4897   0.6244

   iterations =  4
```

Al igual que las variables anteriores el valor p asociado al estadístico de prueba del test de Box Tidwell, es mayor a 0.05, por tanto, no se rechaza el supuesto de linealidad entre el logit y la variable Trigliceridos.

Verificamos gráficamente:

1. Linealidad



El diagrama de dispersión de la Figura, presenta también un patrón lineal creciente.

Finalmente, se puede concluir de las tres variables cuantitativas analizadas que, no se observa una violación al **supuesto de linealidad** para el modelo.

2. Independencia

- La independencia entre observaciones es uno de los supuestos más importantes.
- Suele ser el resultado del proceso de recolección de los datos, por lo que usualmente no es fácil detectarlo mediante la evaluación de residuales del modelo.
- Si los datos no fueron recopilados a lo largo del tiempo, este supuesto puede ser verificado por medio del gráfico de los residuos vs. los residuos rezagados un período de tiempo.
- En un escenario ideal, no se debe observar ningún tipo de patrón aparente.

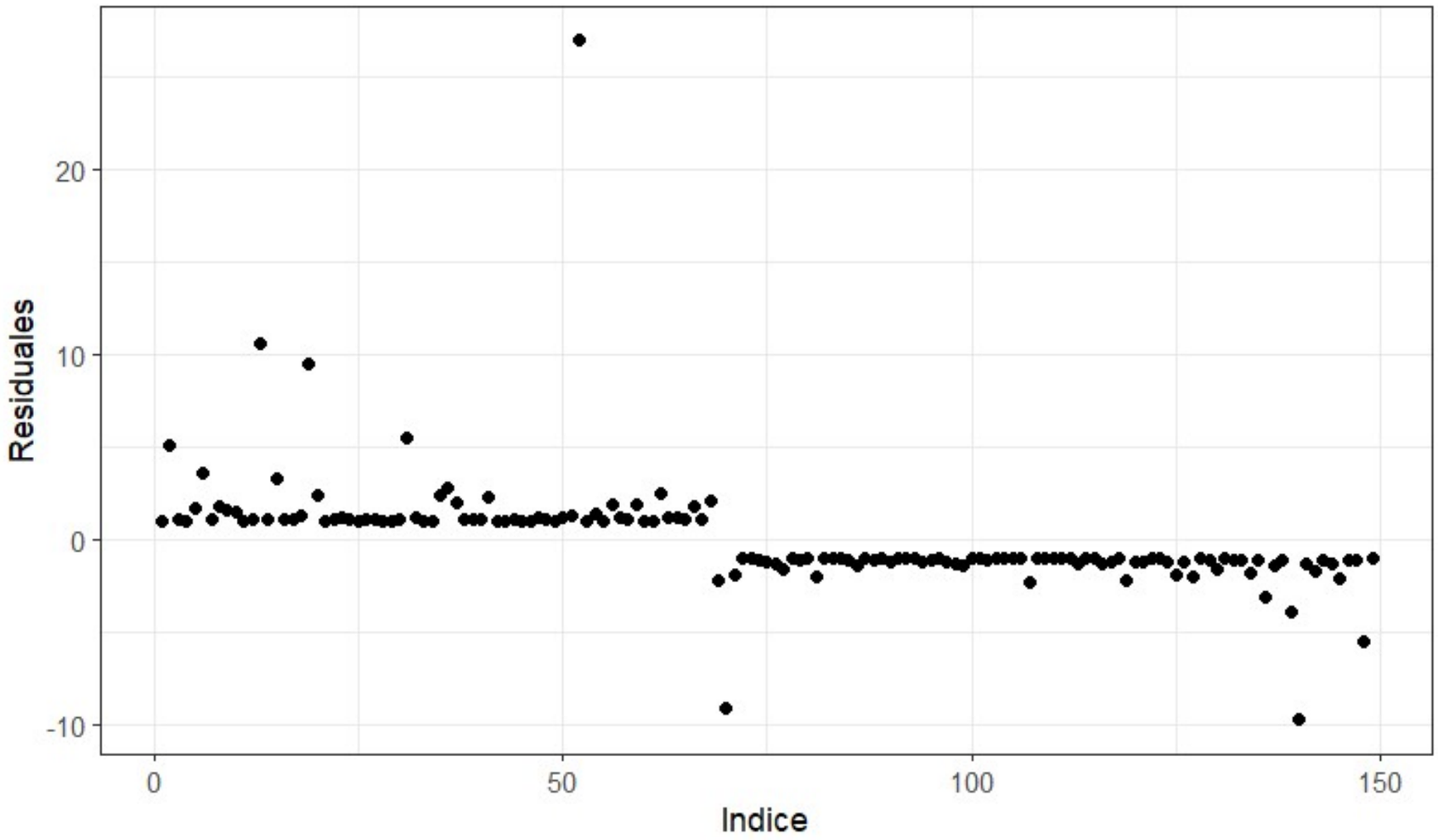
2. Independencia

- En particular, podemos crear la gráfica de los residuales de desviación del modelo logit contra los números índice de las observaciones.

```
Indice <- seq(1,149,1)
Residuales <- modelo2$residuals
residual <- data.frame(Indice, Residuales)
ggplot(data = residual, aes(x = Indice, y = Residuales)) +
  geom_point() +
  theme_bw() +
  theme(plot.title = element_text(hjust = 0.5))
```

- En la figura no se observa ningún patrón aparente que pudiera indicar un grado de dependencia entre las observaciones, por lo que se puede concluir que se cumple el supuesto de independencia

2. Independencia



3. Multicolinealidad

- La multicolinealidad es un problema como en la regresión lineal.

- Las variables predictoras no deben estar altamente correlacionadas. Cuando esto sucede, debe eliminarse aquella covariable que genera un grado de dependencia con las demás.

- Usualmente es posible trabajar con un grado de correlación moderado, ya que cuando la correlación entre variables es muy alta, se genera un incremento en los errores estándar, lo cual hace que los coeficientes estimados no sean confiables, y en consecuencia, las estimaciones sean poco creíbles.

3. Multicolinealidad

La metodología a seguir para probar este supuesto consiste en:

1) Evaluar los coeficientes de correlación entre las variables explicativas.

2) Verificar colinealidad por medio de los valores de tolerancia o los factores infladores de varianza (VIF).

3) Si a partir de las medidas anteriores hay sospecha de multicolinealidad, esto puede ser confirmado por medio del índice de condición, los valores propios y las proporciones de varianza.

3. Multicolinealidad

a) Correlación

Inicialmente calculamos la matriz de correlaciones entre las variables cuantitativas, para ello extraemos las variables cuantitativas del objeto datos y generamos una matriz de correlaciones:

```
datos.cuanti <- datos[, c(3, 5, 6)]
colnames(datos.cuanti) <- c('Edad', 'Colesterol', 'Triglice')
M <- round(cor(datos.cuanti), digits=2); M
            Edad Colesterol Triglice
Edad        1.00        0.25     0.19
Colesterol  0.25        1.00     0.36
Triglice    0.19        0.36     1.00
```

En esta se observa que la mayor correlación es entre las variables colesterol y trigliceridos

3. Multicolinealidad

Se presenta un mapa de calor con la correlación existente entre las dos variables numéricas que fueron consideradas en el ajuste del modelo:

```
# install.packages("corrplot")
# library(corrplot)
corrplot.mixed(M)
```

3. Multicolinealidad

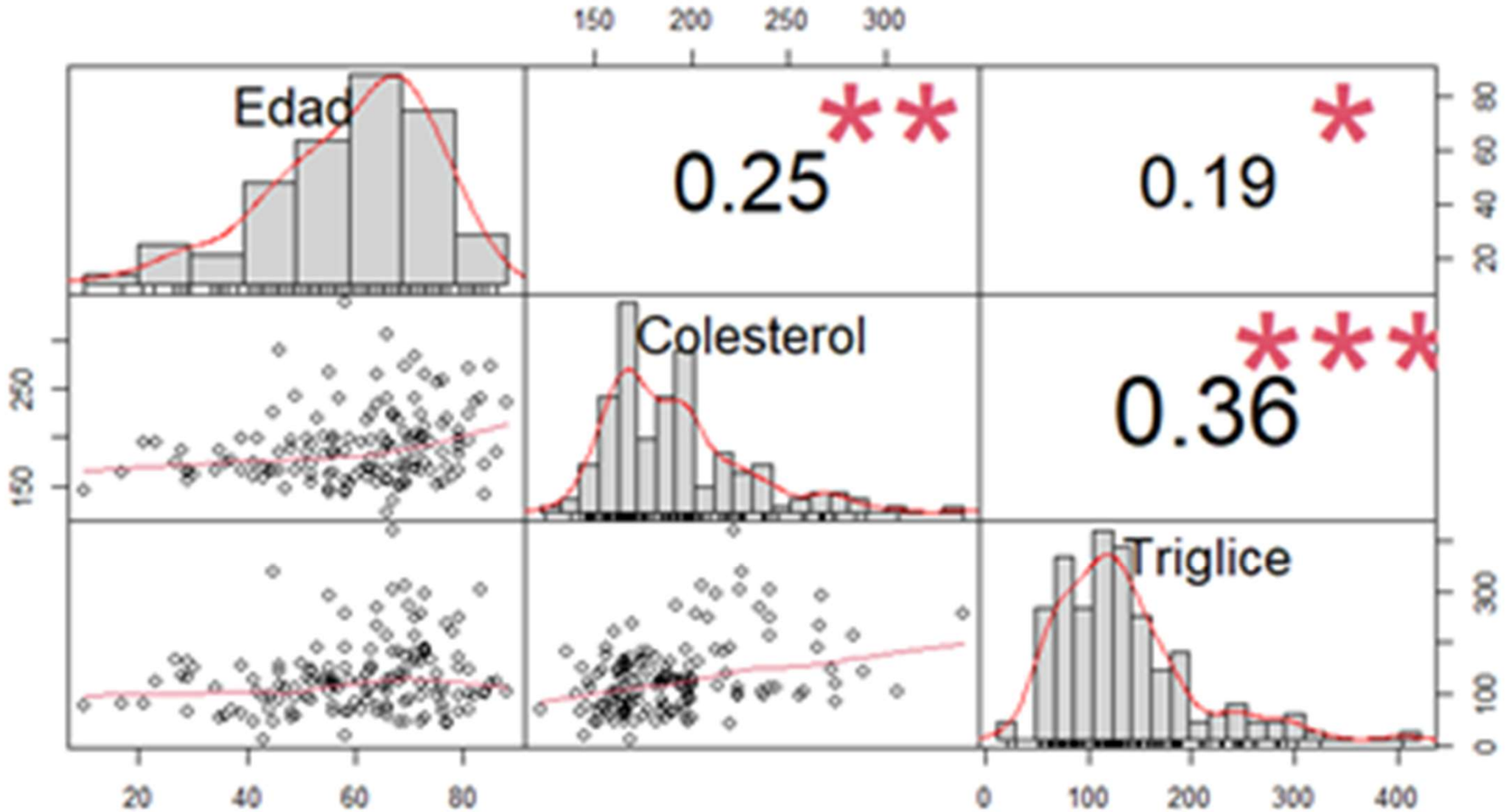
También se puede realiza el siguiente gráfico, donde se presenta el histograma para cada variable, valor de correlación y gráfico de dispersión.

```
# install.packages("PerformanceAnalytics")
# library(PerformanceAnalytics)
chart.Correlation(datos.cuanti, histogram=TRUE, pch=15))
```

En la figura siguiente se observa que, las variables edad, colesterol y trigliceridos no posee una distribución simétrica, ya que están sesgadas.

Además, los gráficos de dispersión no muestran patrones claros de correlación, sin embargo, la correlación más alta entre las variables colesterol y triglicéridos es de 0.36, lo cual no es preocupante en términos de violación del supuesto.

3. Multicolinealidad



3. Multicolinealidad

b) Factores infladores de varianza o VIF

- Los factores infladores de varianza o VIF determinan el grado de relación entre variables independientes. Además, vienen determinados por el ajuste de una regresión entre ambas variables.

- El Factor de Inflación de la Varianza se define así: $$VIF = \frac{1}{1 - R_k^2}$$

  Siendo R_k^2 el coeficiente de determinación de la regresión auxiliar de la variable X_k sobre el resto de las variables explicativas.

- El VIF toma valores entre 1 e ∞.

3. Multicolinealidad

- Si hay uno o más VIF grandes, hay multicolinealidad.
- La experiencia indica que, si cualquiera de los VIF es mayor que 5 o 10, es indicio de que los coeficientes asociados de regresión están mal estimados debido a la multicolinealidad.
- A continuación, se presentan los valores para los estadísticos *VIF* para las variables numéricas que se tienen en el modelo.

```
# install.packages("knitr")
# library(knitr)
kable(vif(modelo2), digits = 2)
```

```
|      |    x|
|:-----|----:|
|edad  | 1.12|
|coles | 1.09|
|trigl | 1.01|
|hiper | 1.21|
```

3. Multicolinealidad

- También podemos obtener a través de la librería car.

```
# library(car)
vif(modelo2)
    edad    coles    trigl    hiper
1.121075 1.086154 1.007599 1.210846
```

Para el conjunto de datos considerado, los valores en la tabla parecen indicar que **no hay un problema evidente de multicolinealidad**.

FINESI

Modelos Discretos

IV Semestre

https://aulavirtual2.unap.edu.pe/

GRACIAS