Coeficiente de correlación de Pearson

Debe satisfacer las siguientes condiciones

La relación que se quiere estudiar entre ambas variables es lineal
Las dos variables deben de ser cuantitativas
Normalidad: ambas variables se tienen que distribuir de forma normal (test K-S)
Homocedasticidad: La varianza de Y debe ser constante a lo largo de la variable X (test de Bartlett)

Características

Toma valores entre [-1, +1], siendo +1 una correlación lineal positiva perfecta y -1 una correlación lineal negativa perfecta. Si es 0 significa que NO existe relación lineal entre las variables consideradas.
No varía si se aplican transformaciones a las variables
no equivale a la pendiente de la recta de regresión
es necesario calcular su significatividad (p-valor) si no es significativo, se ha de interpretar que la correlación de ambas variables es 0

1.1. Evidencias gráficas: Scatterplot

```
library(ggplot2)
ggplot(data = Cars93,
  aes(x = Weight, y = Horsepower)) +
  geom_point(colour = "red4") +
  ggtitle("Diagrama de dispersión") +
  theme_bw() +
  theme(plot.title = element_text(hjust =
0.5))
```

1.2. Evidencias gráficas: Análisis de normalidad

```
par(mfrow = c(1, 2))
hist(Cars93$Weight, breaks = 10, main =
"",
  xlab = "Weight", border = "darkred")
hist(Cars93$Horsepower, breaks = 10,
main = "",
  xlab = "Horsepower", border = "blue")
par(mfrow = c(1, 1))
```

Coeficiente de correlación de Pearson (cont)

1.3. Evidencias gráficas: gráfico quantil-quantil (Q-Q plot)

```
qqnorm(Cars93$Weight, main = "Wei-
ght",
  col = "darkred")
qqline(Cars93$Weight)
qqnorm(Cars93$Horsepower, main = "-
Horsepower",
  col = "blue")
qqline(Cars93$Horsepower)
```

```
par(mfrow = c(1, 2))
qqnorm(Cars93$Weight, main = "Wei-
ght",
  col = "darkred")
qqline(Cars93$Weight)
qqnorm(Cars93$Horsepower, main = "-
Horsepower",
  col = "blue")
qqline(Cars93$Horsepower)
par(mfrow = c(1, 1))
```

2.1 Evid. contr.: test K-S-L (H0= dist. normal)

```
require(nortest)
lillie.test(Cars93$Weight)
lillie.test(Cars93$Horsepower)
```

2.2 Si no es dist. normal: tipificación

a cada observación se resta la media y se divide por la desviación típica.
xt = scale(x)

2.3 Si no es dist. normal: transformación no lineal

Asimetría negativa: $x2$
Asimetría positiva: $\sqrt{x}$ (poco) $ln(x)$
(medio) $1/x$ (alto)
Ahora se repiten todos los pasos a partir de evidencias gráficas: normalidad

Coeficiente de correlación de Pearson (cont)

3.1. Análisis de la homocedasticidad: Gráficas

```
ggplot(data = Cars93, aes(x = Weight,
  y = Horsepower)) + geom_point(colour = "re
  geom_segment(aes(x = 1690, y = 70, xend
  yend = 300), linetype = "dashed") +
  geom_segment(aes(x = 1690, y = 45, xend
  yend = 100),linetype = "dashed") +
  ggtitle("Diagrama de dispersion") + theme_
  theme(plot.title = element_text(hjust = 0.5))
```

3.2. Análisis de la homocedasticidad: ev. cont. Bartlett

El p-valor menor que 5% permite rechazar la hipótesis nula $H0$
bartlett.test(list(Cars93$Weight,Cars93$Hors
Test de Bartlett en las variables transformad
bartlett.test(list(log(Cars93$Weight),log(Cars
rsepower)))

4.1. Coeficiente de Pearson (ev. num. solo)

```
cor(x = Cars93$Weight, y = log(Cars93$Hors
method = "pearson")
```

4.2. Coeficiente de Pearson (ev. contrastada)

Comrpobamos que p-valor sea menor que o 0,05

```
cor.test(x = Cars93$Weight, y = log(Cars93$
wer),alternative = "two.sided",conf.level = 0.9
method = "pearson")
```

Coeficiente de correlación de Pearson (cont)

5. Coeficiente de determinación $R2$

cantidad de varianza de la variable Y explicada por X y que se obtiene elevando al cuadrado el coeficiente de correlación r

El $R2$ presenta valores entre 0 y 1. Lo más próximo a 1 significará que con nuestra variable X seremos capaces de explicar en gran medida el comportamiento de nuestra variable Y. En cambio, si se aproxima a 0 querrá decir que si nuestro objetivo es explicar la variable Y en función de X, tendremos que cambiar de variable explicativa (X) ya que prácticamente no nos aporta información sobre la variable Y.

Si no se cumplen condiciones del coef. Pearson

Coeficiente Rho de Spearman

Cuando estamos ante variables categóricas (no numéricas) (género y educación)

Cuando los valores no pertenecen a una distribución normal

```
require(MASS)
cor.test(birthwt$bwt, birthwt$lwt, alternative="two.sided", method="spearman")
cor.test(birthwt$bwt, birthwt$lwt, alternative="less", method="spearman")
cor.test(birthwt$bwt, birthwt$lwt, alternative="greater", method="spearman")
```

Coeficiente Tau de Kendall

Cuando estamos ante variables categóricas (no numéricas) (género y educación)

Cuando los valores no pertenecen a una distribución normal

```
require(MASS)
cor.test(birthwt$bwt, birthwt$lwt, alternative="two.sided", method="kendall")
etc
```

La matriz de correlaciones

```
cor(x = datos, method = "pearson")
pairs(x = datos, lower.panel = NULL)
require(corrplot)
corrplot(corr = cor(x = datos, method = "-pearson"), method = "number")
require(psych)
pairs.panels(x = datos, ellipses = FALSE, lm = TRUE, method = "pearson")
```

El coeficiente de correlación parcial

Se quiere estudiar la relación entre las variables precio y peso de los automoviles. Se sospecha que esta relación podría estar influenciada por la variable potencia del motor, ya que a mayor peso del vehículo se requiere mayor potencia y, a su vez, motores más potentes son más caros.

```
ggplot(data = Cars93, aes(x = Weight, y = log(Price))) +
geom_point(colour = "red4") +
ggtitle("Diagrama de dispersion") +
theme_bw() +
theme(plot.title = element_text(hjust = 0.5))
cor.test(x = Cars93$Weight, y = log(Cars93$Price), method = "pearson")
require(ppcor)
pcor.test(x = Cars93$Weight, y = log(Cars93$Price),
  z = Cars93$Horsepower, method = "pearson")
```

El coeficiente de correlación parcial (cont)

La correlación entre el peso y el logaritmo del precio es alta (r=0.764) y significativa (p-value < 2.2e-16). Sin embargo, cuando se estudia su relación bloqueando la variable potencia de motor, a pesar de que la relación sigue siendo significativa (p-value = 6.288649e-05) pasa a ser baja (r=0.4047).

Luego, podemos concluir que la relación lineal existente entre el peso y el logaritmo del precio está influenciada por el efecto de la variable potencia de motor. Si se controla el efecto de la potencia, la relación lineal existente es baja (r=0.4047).