TAREA_4

Claret Rodríguez Jiménez- Mariana Méndez Pérez

2025-11-25

Colegio Universitario de Cartago

Ejercicios de Correlación y Regresión

Análisis de Correlación Lineal

Prof. David Martínez Salazar

Alumnas: Claret Rodríguez Jiménez, Mariana Méndez Pérez

Instrucciones Generales

Para cada ejercicio, determine si existe evidencia suficiente para concluir que existe una correlación lineal entre las variables presentadas. Utilice diagramas de dispersión y calcule el coeficiente de correlación cuando sea necesario.

En este análisis se estudia la relación lineal entre dos variables cuantitativas utilizando el **coeficiente de correlación de Pearson** (r) y la prueba de significancia basada en el **valor** p. El coeficiente r indica la fuerza y la dirección de la relación: valores cercanos a 1 o -1 representan relaciones fuertes, mientras que valores cercanos a 0 indican relaciones débiles o inexistentes. El signo muestra si la relación es directa (positiva) o inversa (negativa). Sin embargo, r por sí solo no demuestra si la relación es estadísticamente confiable, por lo que se utiliza la prueba `cor.test()` que genera un valor p. Si el valor p es menor o igual a 0.05, se rechaza la hipótesis nula y se concluye que existe evidencia suficiente de correlación lineal; si es mayor, **no existe evidencia estadística para afirmarlo**.

En cada gráfico de dispersión se incluye una **línea de regresión ajustada** mediante el método de mínimos cuadrados. Esta línea no implica necesariamente que la relación entre las variables sea lineal; más bien, funciona como una herramienta visual que permite evaluar si los puntos muestran una tendencia general ascendente, descendente o si están distribuidos de manera aleatoria. La finalidad de la línea es **facilitar la interpretación del patrón general** en los datos, independientemente de su fuerza o significancia estadística. Posteriormente, esta impresión visual se complementa con el cálculo del coeficiente de correlación de Pearson y la prueba de significancia asociada, lo que permite determinar formalmente si existe o no evidencia de una relación lineal en la población.

EJERCICIO 18

Tamaño de casinos e ingresos.

A continuación se presentan los tamaños (en miles de pies cuadrados) y los ingresos (en millones de dólares) de casinos de Atlantic City (según datos del New York Times). ¿Existe evidencia suficiente para concluir que existe una correlación lineal entre el tamaño de los casinos y sus ingresos?

Tamaño 160 227 140 144 161 147 141

Ganancias 189 157 140 127 123 106 101

EJERCICIO 18 - Tamaño de casinos vs Ingresos

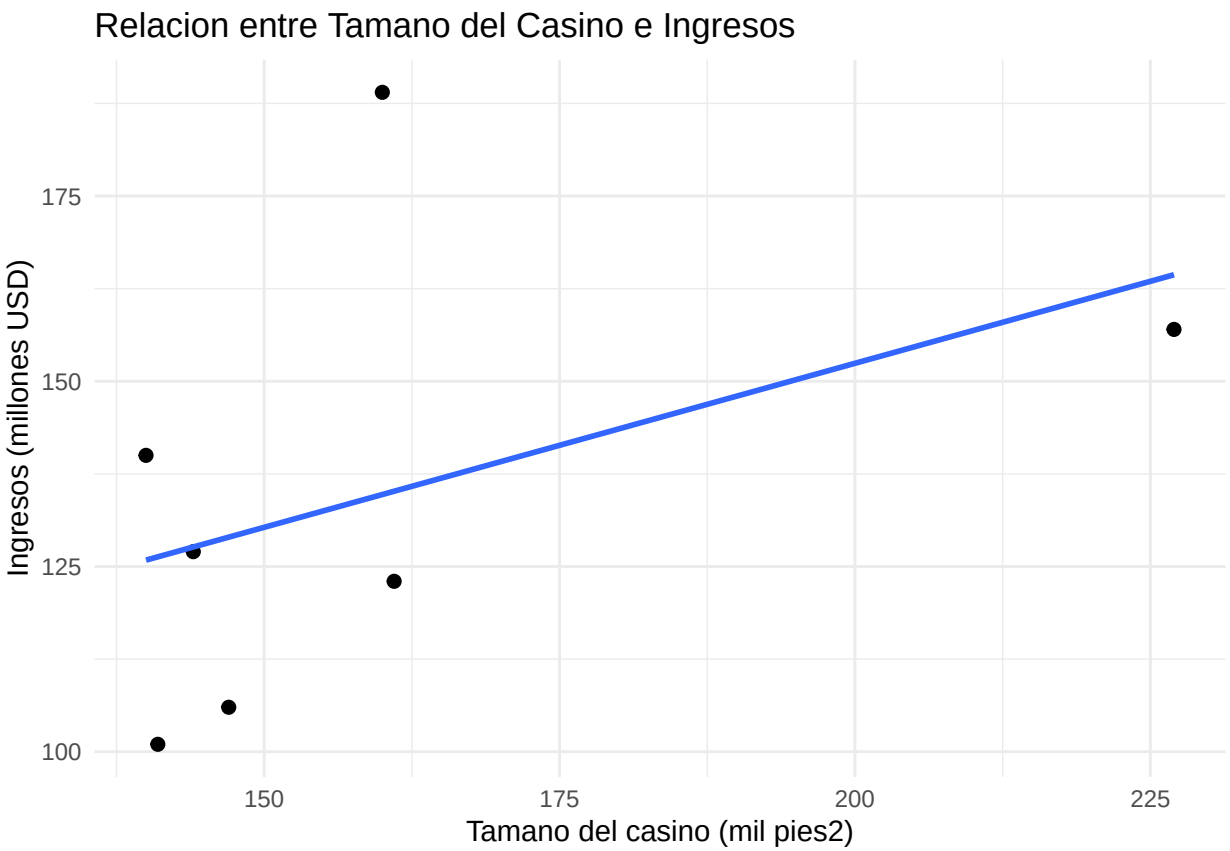
Datos

```
tamano <- c(160, 227, 140, 144, 161, 147, 141)
ingresos <- c(189, 157, 140, 127, 123, 106, 101)

df18 <- data.frame(tamano, ingresos)

library(ggplot2)
```

Diagramas de dispersión

```
p18 <- ggplot(df18, aes(x = tamano, y = ingresos)) +
geom_point(size=2) +
geom_smooth(method = "lm", se = FALSE) +
labs(title = "Relacion entre Tamano del Casino e Ingresos",
x = "Tamano del casino (mil pies2)",
y = "Ingresos (millones USD)") +
theme_minimal()

print(p18)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



INTERPRETACION:

El diagrama de dispersión titulado "Relación entre Tamaño del Casino e Ingresos'' presenta la asociación entre estas dos variables, donde el tamaño del casino (en miles de pies cuadrados) se representa en el eje horizontal y los ingresos (en millones de USD) en el eje vertical. Visualmente, se observa una **ligera tendencia ascendente** en la distribución de los puntos, sugiriendo que los casinos más grandes tienden a generar mayores ingresos. Sin embargo, esta tendencia **no es marcada**, ya que los puntos presentan una dispersión considerable alrededor de lo que sería una línea de tendencia imaginaria. Cabe destacar que existen casinos de tamaño similar que muestran ingresos notablemente diferentes, lo que indica que el tamaño no es el único factor determinante en el desempeño financiero. La nube de puntos se concentra principalmente entre los 150–225 mil pies^2 de tamaño y los 100–175 millones de USD en ingresos, mostrando una relación positiva pero con **variabilidad significativa**.

Cálculo del coeficiente de correlación (r)

```r
r18 <- cor(df18$tamano, df18$ingresos)
r18
```

```
## [1] 0.444569
```

INTERPRETACION:

El coeficiente de correlación de Pearson de $r = 0.445$ indica una **correlación positiva moderada** entre el tamaño del casino y los ingresos. Este valor sugiere que existe una tendencia de que a mayor tamaño del casino, mayores ingresos se generan, sin embargo, la fuerza de esta relación es solo **moderada**. El valor de $r^2 = 0.198$ indica que aproximadamente el 19.8% de la variabilidad en los ingresos puede explicarse por el tamaño del casino, dejando un 80.2% de la variación atribuible a otros factores no considerados en este análisis.

Prueba de hipótesis – cor.test()

```r
test18 <- cor.test(df18$tamano, df18$ingresos)
test18
```

```
##
##  Pearson's product-moment correlation
##
## data:  df18$tamano and df18$ingresos
## t = 1.1098, df = 5, p-value = 0.3176
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.4637444  0.8972426
## sample estimates:
##      cor
## 0.444569
```

INTERPRETACION:

La prueba de significancia estadística revela un valor $p = 0.318$, muy por encima del nivel de significancia convencional de $\alpha = 0.05$. Esto indica que la correlación observada de 0.445 **no es estadísticamente significativa** y podría deberse al azar. El intervalo de confianza del 95% $[-0.464, 0.897]$ que incluye el cero confirma esta falta de significancia estadística. Por lo tanto, aunque visualmente existe una tendencia positiva y el coeficiente de correlación es moderado, **no tenemos evidencia suficiente para afirmar que exista una relación lineal genuina** entre el tamaño del casino y los ingresos en la población.

CONCLUSION

A pesar de que se observa una **correlación positiva moderada** ($r = 0.445$), la prueba estadística ($p = 0.318$) indica que **no existe evidencia suficiente** para concluir que la correlación lineal entre el tamaño del casino y los ingresos sea **estadísticamente significativa** al nivel de significancia del 5%.

**Ejercicio 18 (r = 0.445, p = 0.318)*

-Fuerza: correlación débil

-Dirección: positiva

-Prueba estadística: p > 0.05 → NO existe correlación lineal

EJERCICIO 19

Tarifas aéreas.

A continuación se presentan los precios (en dólares) de tarifas aéreas de diferentes aerolíneas desde la ciudad de Nueva York (JFK) a San Francisco. Los precios se basan en boletos comprados con 30 días de anticipación y un día de anticipación; las aerolíneas son US Air, Continental, Delta, United, American, Alaska y Northwest. ¿Hay evidencia suficiente para concluir que existe una correlación lineal entre los precios de los boletos comprados con 30 días de anticipación y los boletos comprados con un día de anticipación?

30 días 244 260 264 264 278 318 280

Un día 456 614 567 943 628 1088 536

** EJERCICIO 19 - Tarifas aéreas**

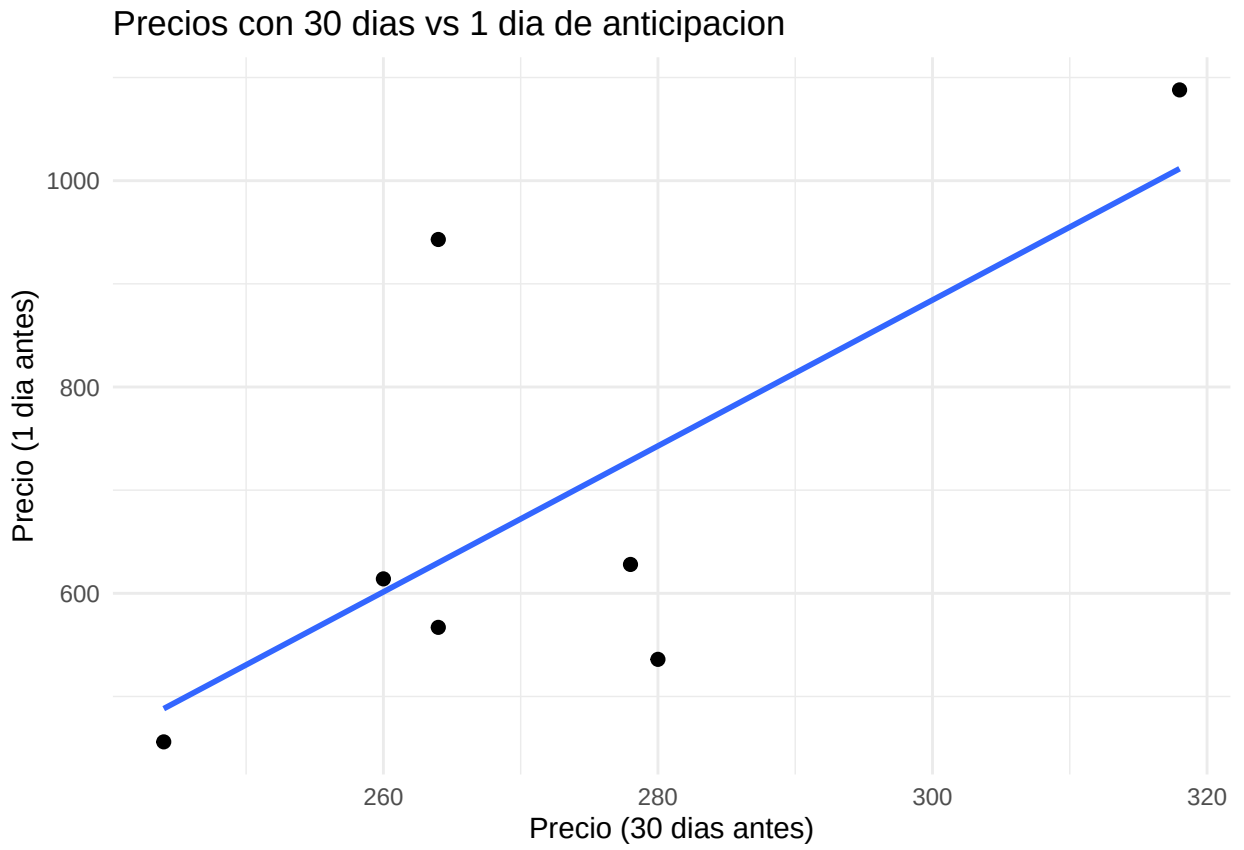
Datos

“‘ r precio_30 <- c(244, 260, 264, 264, 278, 318, 280) precio_1 <- c(456, 614, 567, 943, 628, 1088, 536)
df19 <- data.frame(precio_30, precio_1)
library(ggplot2) “‘

Diagramas de dispersión

“‘ r p19 <- ggplot(df19, aes(x = precio_30, y = precio_1)) + geom_point(size=2) +
geom_smooth(method = “lm”, se = FALSE) + labs(title = “Precios con 30 dias vs 1 dia de anticipacion”,
x = “Precio (30 dias antes)”, y = “Precio (1 dia antes)”) + theme_minimal()
print(p19) “‘

```
## `geom_smooth()` using formula = 'y ~ x'
```



INTERPRETACION:

El gráfico de dispersión titulado "Precios con 30 días vs 1 día de anticipación' ' muestra la relación entre el precio de los boletos comprados con 30 días de anticipación (eje horizontal) y aquellos comprados con 1 día de anticipación (eje vertical). Visualmente, se observa una **clara tendencia lineal ascendente** en la distribución de los puntos, indicando que existe una **relación directa** entre ambos precios. Los puntos se agrupan de manera relativamente consistente alrededor de lo que sería una línea de tendencia diagonal, sugiriendo que, en general, cuando el precio con 30 días de anticipación es mayor, el precio con 1 día de anticipación también tiende a ser mayor, y viceversa.

Cálculo del coeficiente de correlación (r)

```
r r19 <- cor(df19$precio_30, df19$precio_1) r19
## [1] 0.7087568
```

INTERPRETACION:

El coeficiente de correlación de Pearson de $r = 0.709$ indica una **correlación positiva fuerte** entre los precios con 30 días de anticipación y los precios con 1 día de anticipación. Este valor sugiere que existe una relación lineal considerablemente consistente entre ambas variables, donde aproximadamente el 50.3% ($r^2 = 0.503$) de la variabilidad en los precios de 1 día de anticipación puede explicarse por los precios de 30 días de anticipación.

Prueba de hipótesis – cor.test()

```
r test19 <- cor.test(df19$precio_30, df19$precio_1) test19
## ##  Pearson's product-moment correlation ## ## data:  df19$precio_30 and df19$precio_1
## t = 2.2465, df = 5, p-value = 0.0746 ## alternative hypothesis: true correlation is
not equal to 0 ## 95 percent confidence interval: ##  -0.09501321  0.95310784 ## sample
estimates: ##       cor ## 0.7087568
```

INTERPRETACION:

El valor p obtenido fue 0.0746, mayor que el nivel de significancia de 0.05, por lo que **no se rechaza la hipótesis nula**. Aunque la correlación observada es relativamente fuerte ($r = 0.709$), **no es estadísticamente significativa**. Además, el intervalo de confianza del 95% $[-0.095, 0.953]$ incluye el cero, lo que confirma que no hay evidencia suficiente para afirmar la existencia de una correlación lineal en la población.

CONCLUSION

En conclusión, aunque el diagrama de dispersión y el coeficiente de correlación de Pearson ($r = 0.709$) muestran una **correlación lineal positiva moderadamente fuerte** entre los precios de boletos comprados con 30 días y 1 día de anticipación, la prueba de significancia estadística indica lo contrario. El valor $p = 0.0746$, mayor que $\alpha = 0.05$, y el intervalo de confianza del 95% que incluye el cero, revelan que **no existe evidencia estadística suficiente** para afirmar que esta relación lineal se mantenga en la población. Por lo tanto, aunque la tendencia observada sugiere una relación positiva, **no se puede concluir que la correlación sea significativa**.

Ejercicio 19 (r = 0.709, p = 0.0746)

-Fuerza: correlación fuerte

-Dirección: positiva

-Prueba estadística: p > 0.05 → NO existe correlación lineal

EJERCICIO 20

####**Pasajeros y espacios de estacionamiento.**

A continuación se presentan los números de pasajeros y los números de espacios de estacionamiento en diferentes estaciones del tren Metro-North (según datos de Metro- North). ¿Existe una correlación lineal entre los números de pasajeros y los números de espacios de estacionamiento?

Pasajeros 3453 1350 1126 3120 2641 277 579 2532

Espacios de estacionamiento 1653 676 294 950 1216 179 466 1454

** EJERCICIO 20 - Pasajeros y espacios de estacionamiento.**

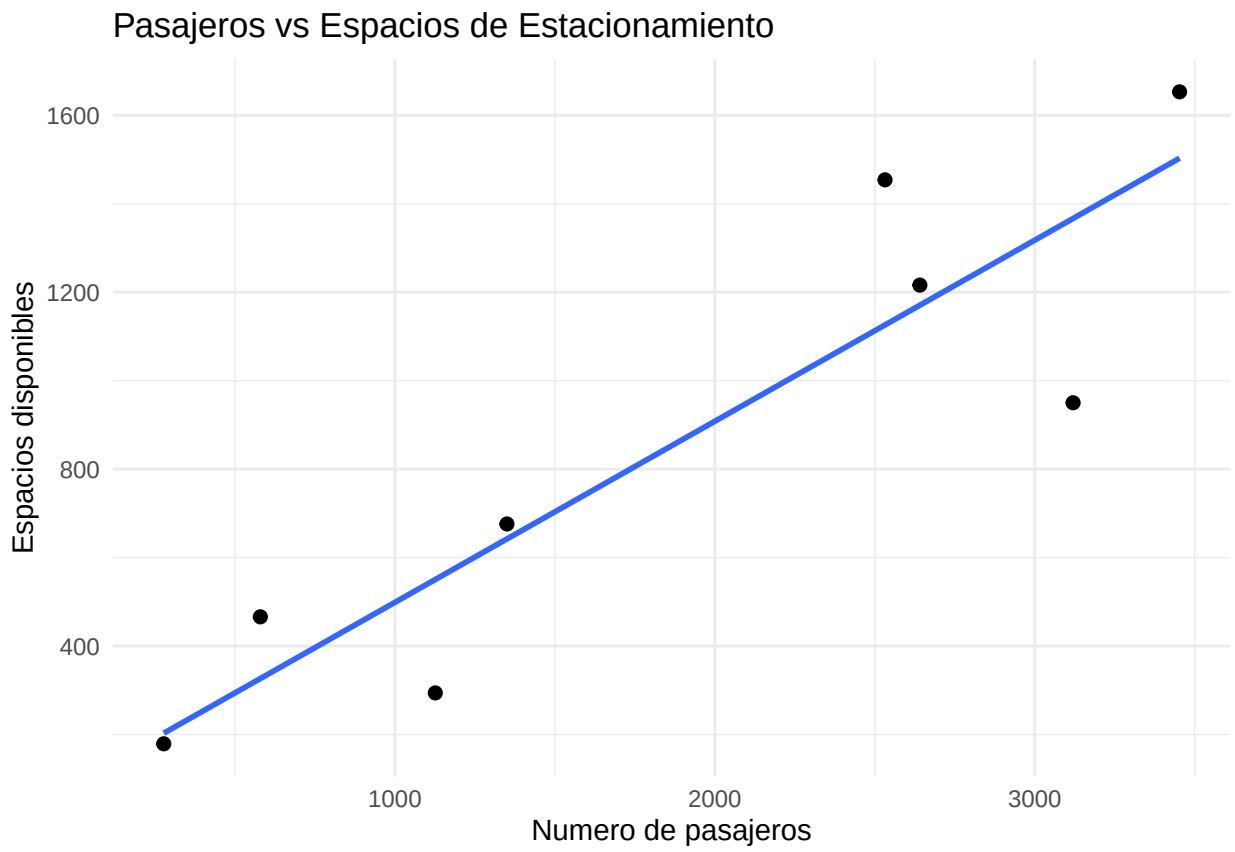
Datos

“‘ r pasajeros <- c(3453, 1350, 1126, 3120, 2641, 277, 579, 2532) espacios <- c(1653, 676, 294, 950, 1216, 179, 466, 1454)
df20 <- data.frame(pasajeros, espacios) “‘

Diagramas de dispersión

“‘ r library(ggplot2)
p20 <- ggplot(df20, aes(x = pasajeros, y = espacios)) + geom_point(size=2) + geom_smooth(method = “lm”, se = FALSE) + labs(title = “Pasajeros vs Espacios de Estacionamiento”, x = “Numero de pasajeros”, y = “Espacios disponibles”) + theme_minimal()
print(p20) “‘

```
## `geom_smooth()` using formula = 'y ~ x'
```



INTERPRETACION:

El gráfico de dispersión, titulado “Pasajeros vs Espacios de Estacionamiento’ ’, muestra una **clara relación lineal positiva** entre el número de pasajeros (eje horizontal) y los espacios de estacionamiento disponibles (eje vertical). La nube de puntos se distribuye formando un patrón ascendente de izquierda a derecha, donde a medida que aumenta el número de pasajeros, también aumenta consistentemente la cantidad de espacios de estacionamiento. La línea de tendencia azul, que atraviesa centralmente la nube de puntos, confirma visualmente esta relación positiva y directa entre ambas variables.

Cálculo del coeficiente de correlación (r)

```
r r20 <- cor(df20$pasajeros, df20$espacios) r20
## [1] 0.9011516
```

INTERPRETACION:

El coeficiente de correlación de Pearson de r = 0.901 indica una correlación positiva muy fuerte entre las variables. Este valor, cercano a 1, significa que existe una relación lineal casi perfecta entre el número de pasajeros y los espacios de estacionamiento: cuando una variable aumenta, la otra tiende a aumentar proporcionalmente. La fuerza de esta correlación sugiere que el número de pasajeros explica una gran proporción de la variabilidad observada en la cantidad de espacios de estacionamiento disponibles.

Prueba de hipótesis – cor.test()

```
r test20 <- cor.test(df20$pasajeros, df20$espacios) test20
## ##  Pearson's product-moment correlation ## ## data:  df20$pasajeros and df20$espacios
## t = 5.0919, df = 6, p-value = 0.002239 ## alternative hypothesis: true correlation is
not equal to 0 ## 95 percent confidence interval: ##  0.5383229 0.9821454 ## sample
estimates: ##       cor ## 0.9011516
```

INTERPRETACION:

La prueba de significancia estadística para la correlación revela un valor $p = 0.002239$, muy por debajo del nivel de significancia convencional de $\alpha = 0.05$. Esto indica que la correlación observada de 0.901 es **estadísticamente significativa** y no producto del azar. El intervalo de confianza del 95% $[0.538, 0.982]$ confirma que existe una correlación positiva significativa en la población, descartando con un 95% de confianza que la verdadera correlación sea cero. Por lo tanto, podemos concluir que existe evidencia sólida de una **relación lineal positiva genuina** entre el número de pasajeros y los espacios de estacionamiento.

CONCLUSION

En conclusión, el análisis del diagrama de dispersión, junto con el coeficiente de correlación de Pearson ($r = 0.901$) y la prueba de significancia ($p = 0.002$), demuestra que existe una **correlación lineal positiva muy fuerte** y **estadísticamente significativa** entre el número de pasajeros y los espacios de estacionamiento. Esto significa que, a mayor cantidad de pasajeros en una estación, generalmente se dispone de un mayor número de espacios de estacionamiento. Dado que el intervalo de confianza del 95% **no incluye el cero**, existe evidencia suficiente para afirmar que esta relación lineal positiva se mantiene en la población.}

Ejercicio 20 (r = 0.901, p = 0.0022)

-Fuerza: correlación muy fuerte

-Dirección: positiva

-Prueba estadística: p < 0.05 → SÍ existe correlación lineal

EJERCICIO 21

####**Costos de reparación de automóviles.**

A continuación se presentan los costos de reparación (en dólares) para automóviles que participaron en pruebas de choques frontales a una velocidad de 6 mi/h y en pruebas de choques traseros a una velocidad de 6 mi/h (según datos del Insurance Institute for Highway Safety). Los automóviles son Toyota Camry, Mazda 6, Volvo S40, Saturn Aura, Subaru Legacy, Hyundai Sonata y Honda Accord. ¿Hay evidencia suficiente para concluir que existe una correlación lineal entre los costos de reparación de los choques frontales y de los choques traseros?

Choques frontales 936 978 2252 1032 3911 4312 3469

Choques traseros 1480 1202 802 3191 1122 739 2767

** EJERCICIO 21 - Costos de reparación de automóviles.**

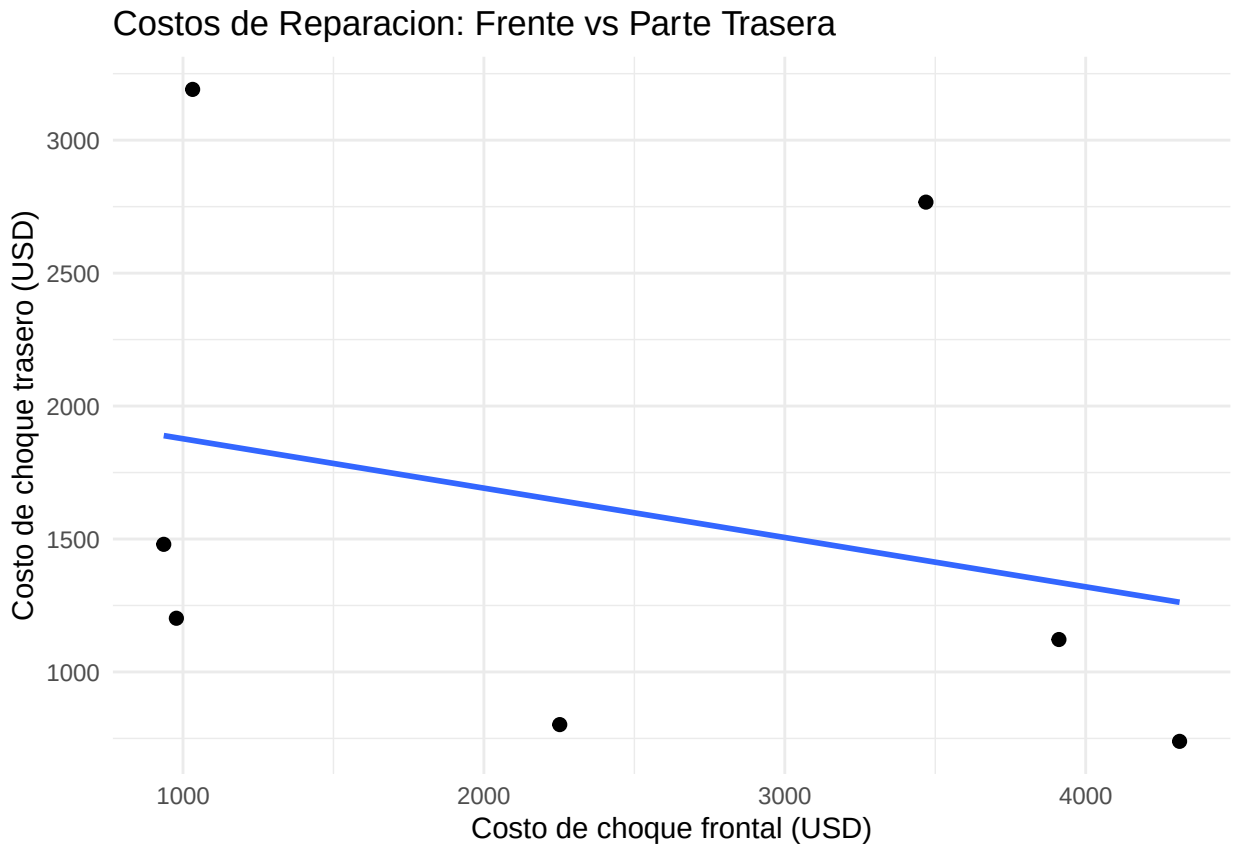
Datos

“‘ r frontales <- c(936, 978, 2252, 1032, 3911, 4312, 3469) traseros <- c(1480, 1202, 802, 3191, 1122, 739, 2767)
df21 <- data.frame(frontales, traseros) “‘

Diagramas de dispersión

“‘ r library(ggplot2)
p21 <- ggplot(df21, aes(x = frontales, y = traseros)) + geom_point(size=2) + geom_smooth(method = “lm”, se = FALSE) + labs(title = “Costos de Reparacion: Frente vs Parte Trasera”, x = “Costo de choque frontal (USD)”, y = “Costo de choque trasero (USD)”) + theme_minimal()
print(p21) “‘

```
## `geom_smooth()` using formula = 'y ~ x'
```



INTERPRETACION:

El gráfico de dispersión titulado “Costos de Reparación: Frente vs Parte Trasera’ ’ muestra la relación entre los costos de reparación por choques frontales (eje horizontal) y los costos por choques traseros (eje vertical), ambos medidos en USD. Visualmente, no se aprecia ningún patrón claro o tendencia definida en la distribución de los puntos. Los datos aparecen dispersos de manera aleatoria en el plano, sin mostrar una dirección consistente. Los costos frontales se distribuyen entre aproximadamente 1000 y 4000 USD, mientras que los costos traseros oscilan entre 1000 y 3000 USD, sin evidencia visual de una relación sistemática entre ambas variables.

Cálculo del coeficiente de correlación (r)

```
r r21 <- cor(df21$frontales, df21$traseros) r21
## [1] -0.2825546
```

INTERPRETACION:

El coeficiente de correlación de Pearson de $r = -0.283$ indica una correlación **negativa débil** entre los costos de reparación frontales y traseros. Este valor sugiere que existe una tendencia muy leve a que cuando los costos frontales aumentan, los costos traseros tienden a disminuir ligeramente, pero esta relación es mínima. El valor de $r^2 = 0.08$ indica que solo el **8%** de la variabilidad en los costos traseros puede explicarse por los costos frontales, lo que refuerza la debilidad de esta asociación.

Prueba de hipótesis – cor.test()

```
r test21 <- cor.test(df21$frontales, df21$traseros) test21
## ##  Pearson's product-moment correlation ## ## data:  df21$frontales and df21$traseros
## t = -0.65865, df = 5, p-value = 0.5392 ## alternative hypothesis: true correlation is
not equal to 0 ## 95 percent confidence interval: ##  -0.8539163  0.5976773 ## sample
estimates: ##        cor ## -0.2825546
```

INTERPRETACION:

La prueba de significancia estadística revela un valor $p = 0.5392$, muy por encima del nivel de significancia convencional de $\alpha = 0.05$. Esto indica que la correlación observada de -0.283 no es estadísticamente significativa y es altamente probable que sea producto del azar. El intervalo de confianza del 95% $[-0.854, 0.598]$, que incluye ampliamente el cero, confirma la falta de significancia estadística. Por lo tanto, podemos concluir que no existe evidencia de una relación lineal genuina entre los costos de reparación frontales y traseros en la población estudiada.

CONSCLUSION

En conclusión, aunque el coeficiente de correlación de Pearson muestra una **relación negativa débil** ($r = -0.283$), el valor $p = 0.5392$ y un intervalo de confianza del 95% que incluye ampliamente el cero indican que **no existe evidencia estadística suficiente** para afirmar que haya una correlación lineal entre los costos de reparación por choques frontales y traseros. Los datos no presentan un patrón visible en el diagrama de dispersión, y solo un 8% de la variabilidad en los costos traseros es explicada por los costos frontales. Por lo tanto, se concluye que **no hay relación lineal significativa** entre ambas variables en la población estudiada.

Ejercicio 21 (r = –0.283, p = 0.5392)
-Fuerza: correlación muy débil

-Dirección: negativa

-Prueba estadística: $p > 0.05 \rightarrow$ NO existe correlación lineal