Análisis de Regresión Lineal Simple

Paula Sofía Torres
Mauricio Rodríguez
Pontificia Universidad Javeriana

1. Introducción

En esta práctica, vamos a explorar los fundamentos de la regresión lineal simple. El objetivo es entender cómo una variable independiente afecta a una variable dependiente y si hay una relación lineal significativa entre ellas.

La fuente de nuestros datos se encuentra en **Kaggle: Advertising Sales Dataset**.

Sin embargo, tuvimos que realizar algunos procesos de limpieza. Puedes encontrar el Excel ya modificado en este **enlace de Google Docs**.

2. Acerca de los datos

El conjunto de datos utilizado en este análisis proviene del archivo `Cleaned_Advertising_Budget_and_Sales.csv`, que contiene información sobre los presupuestos de publicidad en diferentes medios y las ventas correspondientes. Las variables incluidas son:

1. `Sales ($)`: las ventas totales en dólares.
2. `TV Ad Budget ($)`: el presupuesto de publicidad en televisión en dólares.
3. `Radio Ad Budget ($)`: el presupuesto de publicidad en radio en dólares.

3. Propósito del análisis

El propósito de este análisis es examinar la influencia de los presupuestos de publicidad en radio y televisión sobre las ventas totales. Se pretende comparar la eficacia de estos dos medios de publicidad para determinar cuál tiene una mayor relación lineal con el aumento de ventas.

4. Análisis de la variable Presupuesto para Radio

En este apartado nos enfocaremos en analizar el impacto del presupuesto de publicidad en radio.

Así, procederemos a visualizar la relación entre estas dos variables a través de un gráfico de dispersión, donde el eje X representa el presupuesto de publicidad en radio y el eje Y las ventas totales. En el gráfico, se observa cómo se distribuyen los puntos que representan las observaciones individuales, permitiendo visualizar de forma preliminar la existencia de una relación lineal entre ambas variables.

```
options(scipen = 999) #Penalizamos los valores científicos

df = Cleaned_Advertising_Budget_and_Sales
Y = df$Sales
```

```r
X = df$Radio_Budget

plot(X, Y)
```

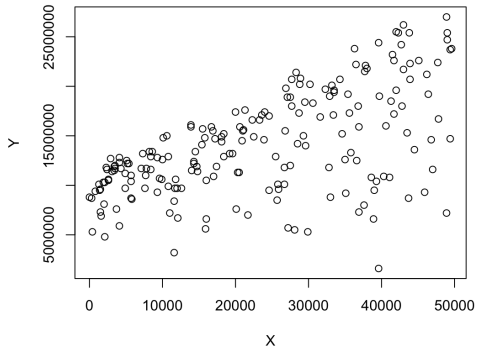

Figure 1: Presupuesto de Radio X Ventas

Gráficamente, se puede ver que es posibile que las variables sí esten linealmente relacionadas entre sí. Para comprobarlo, ajustamos un modelo de regresión lineal simple con lm(Y ~ X, data = df), utilizando el presupuesto de publicidad en radio como variable predictora de las ventas. Este modelo busca estimar cómo cambian las ventas en función de los cambios en el presupuesto de publicidad en radio.

```r
modelo_radio = lm(Y ~ X, data = df)
abline(modelo_radio, col='red')
```

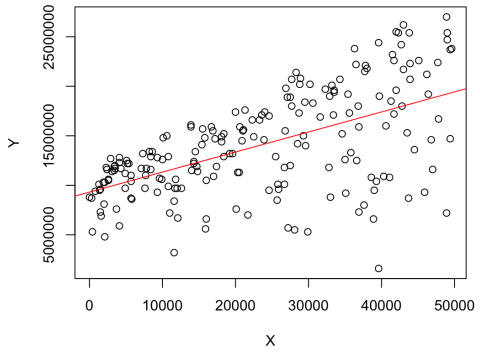

Figure 2: Presupuesto de Radio X Ventas con Modelo de Regresión Lineal

Con tal de validar nuevamente nuestros supuestos, haremos un análisis más detallado por medio de la salida de summary(modelo_radio)

```r
resumen_radio <- summary(modelo_radio)
```

La salida que obtenemos es

```
Residuals:
      Min        1Q    Median        3Q       Max
-15730471  -2132427    770692   2777527   8181043

Coefficients:
             Estimate Std. Error t value            Pr(>|t|)
(Intercept) 9311638.10  562900.50  16.542 <0.0000000000000002 ***
X               202.50      20.41   9.921 <0.0000000000000002 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4275000 on 198 degrees of freedom
Multiple R-squared:  0.332, Adjusted R-squared:  0.3287
F-statistic: 98.42 on 1 and 198 DF,  p-value: < 0.00000000000000022
```

Al observar el p-valor asociado al estadístico (tanto de la F como de la T), que es significativamente pequeño (< 0.05), concluimos que **hay evidencia suficiente para afirmar que existe una relación entre el presupuesto de publicidad en radio y las ventas.** Esto indica que el presupuesto de publicidad en radio tiene un impacto estadísticamente significativo en las ventas o, en otras palabras, que la pendiente de la linea de regresion es distinta de cero.

Considerando la interpretación del coeficiente para el presupuesto de publicidad en radio, el cual es de 202.50, nos dice que por cada unidad adicional invertida en publicidad de radio (suponiendo que esta unidad está en los mismos términos que los datos de entrada, por ejemplo, miles de dólares), podemos esperar un incremento promedio de 202.50 en las ventas, asumiendo que todos los demás factores se mantienen constantes.

Este análisis revela que aunque el presupuesto de publicidad en radio no es el único factor que influye en las ventas, sí constituye un componente significativo de la variabilidad de las mismas.

5. Análisis de la variable de Presupuesto para Televisión

En este apartado nos enfocaremos en analizar el impacto del presupuesto de publicidad en televisión.

Así, procederemos a visualizar la relación entre estas dos variables a través de un gráfico de dispersión, donde el eje X2 representa el presupuesto de publicidad en televisión y el eje Y las ventas totales. En el gráfico, se observa cómo se distribuyen los puntos que representan las observaciones individuales, permitiendo visualizar de forma preliminar la existencia de una relación lineal entre ambas variables.

```
y = df$Sales
x2 = df$TV_Budget
plot(x2, y)
```

En la figura 3, observamos que la relación podría ser logaritmica. Por lo cuál podría ser beneficioso aplicar una transformación logarítmica a los datos para mejorar la estimación del modelo. Esto se debe a que una transformación logarítmica puede ayudar a estabilizar la varianza y hacer que la relación entre las variables sea más lineal.

Analizaremos ambos a continución ajustando un modelo de regresión lineal simple con lm(Y ~ X, data = df), para cada caso.

1. Modelo Regular

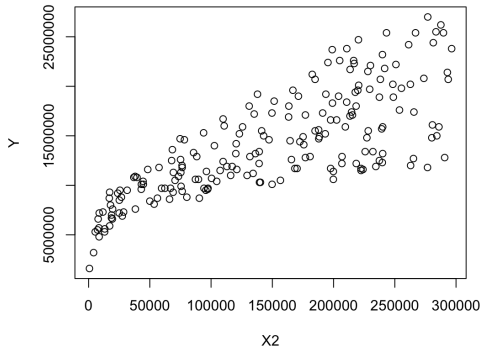

Figure 3: Presupuesto de Televisión X Ventas

```
modelo_tv = lm(y ~ x2, data = df)
abline(modelo_tv,col='red')
```

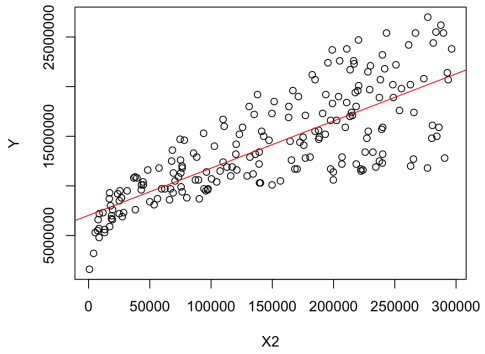

Figure 4: Presupuesto de Televisión X Ventas con Modelo de Regresión Lineal

```
resumen_tv <- summary(modelo_tv)
```

Con tal de validar nuevamente nuestros supuestos, haremos un análisis más detallado por medio de la salida de summary(modelo_tv)

La salida que obtenemos del resumen es

```
Residuals:
     Min       1Q   Median       3Q      Max
-8385982 -1954522  -191265  2067109  7212369

Coefficients:
              Estimate  Std. Error t value            Pr(>|t|)
(Intercept) 7032593.549  457842.940   15.36 <0.0000000000000002 ***
x2               47.537       2.691   17.67 <0.0000000000000002 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3259000 on 198 degrees of freedom
Multiple R-squared:  0.6119,    Adjusted R-squared:  0.6099
F-statistic: 312.1 on 1 and 198 DF,  p-value: < 0.00000000000000022
```

Al observar el p-valor asociado al estadístico tanto de la F como de la T, al igual que el caso anterior, es extremadamente bajo, concluimos que también **hay evidencia suficiente para afirmar que existe una relación significativa entre el presupuesto de publicidad en televisión y las ventas.** Esto indica que el presupuesto de publicidad en televisión tiene un impacto estadísticamente significativo en las ventas. En otras palabras, la pendiente de la línea de regresión es significativamente distinta de cero, lo que sugiere una relación positiva entre estas variables.

Para la interpretación del coeficiente para el presupuesto de publicidad en televisión, que es de 47.537, nos indica que por cada unidad adicional invertida en publicidad de televisión, podemos esperar un incremento promedio de 47.537 en las ventas.

2. Modelo transformado

```
Y <- round(log(y),2) # Se realiza la transformación de las variables considerando la linealización Loga
X2 <- round(log(x2),2)
modelo_tv_ajustado = lm(Y ~ X2, data = df) #Aplicamos el modelo a las variables transformado
modelo_tv_ajustado$coefficients
beta0_est <- exp( 12.2696539 )
beta1_est <- 0.3549444
points(x2,beta0_est*(x2^{beta1_est}),col='blue', pch=20) #revisamos visualmente el modelo
```

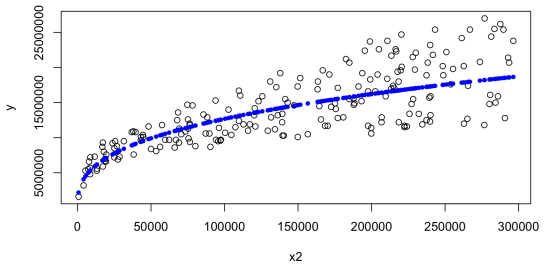

Figure 5: Presupuesto de Televisión X Ventas con Modelo de Regresión Lineal Transformado

```
summary(modelo_tv_ajustado)
```

```
Residuals:
    Min      1Q  Median      3Q     Max
-0.4371 -0.1580  0.0146  0.1716  0.3929

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept) 12.26965    0.17252   71.12 <0.0000000000000002 ***
X2           0.35494    0.01484   23.91 <0.0000000000000002 ***
```

```
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.2105 on 198 degrees of freedom
Multiple R-squared:  0.7428,    Adjusted R-squared:  0.7415
F-statistic: 571.8 on 1 and 198 DF,  p-value: < 0.00000000000000022
```

Se puede denotar que para este modelo ajustado el R-cuadrado aumento significativamente si lo contrastamos con el modelo no ajustaso.

6. Conclusiones

La **comparación directa de los coeficientes de los modelos de regresión para la publicidad en televisión y radio nos ofrece una perspectiva interesante sobre cómo cada medio contribuye a las ventas.** A primera vista, podría parecer que la publicidad en radio es más eficiente, dado que su coeficiente es significativamente más alto (202.50) en comparación con el coeficiente de la publicidad en televisión (47.537) en el modelo no ajustado. Esto sugeriría que, por cada unidad adicional invertida, la publicidad en radio aporta un incremento mayor en las ventas que la televisión.

Sin embargo, es crucial considerar el contexto más amplio, incluyendo el R-cuadrado de ambos modelos. El modelo de televisión tiene un R-cuadrado de 0.7428 (en el modelo que se seleccionaría, es decir, el ajustado), lo que indica que aproximadamente el 74.28% de la variabilidad en las ventas puede explicarse por el presupuesto de publicidad en televisión. Por otro lado, el modelo de radio tiene un R-cuadrado de 0.332, sugiriendo que solo el 33.2% de la variabilidad en las ventas se puede explicar por la publicidad en radio. Esto significa que, aunque la publicidad en radio parece tener un impacto más directo por unidad de inversión, la publicidad en televisión, en general, explica una mayor proporción de la variabilidad en las ventas.

Así, si trabajaramos con una empresa que busca seleccionar si invertir en publicidad de televisión versus radio debería basarse en los objetivos específicos de la campaña, el presupuesto disponible, y el público objetivo. **Mientras que la televisión ofrece un impacto más generalizado en las ventas, la radio ofrece una eficiencia notable por cada dólar gastado.**