**Vargas Carrillo Raul Paulino**

**Regresión Lineal Múltiple Dummys**

Obtener la información utilizando la librería lm de R deberá adjuntar tanto el código utilizado como los resultados obtenidos.

Utilizando el conjunto de datos simulado que contiene las ventas de sillas de coche para niños en 400 tiendas diferentes (Carseats.xlsx), obtenga las estimaciones de los siguientes modelos y proporcione una interpretación de los resultados.

Salesi = β0 + β1Advertisingi + β2USi + εi (1)

Salesi = β0 + β1Income + β2ShelveLoci + εi (2)

Calcule el coeficiente de determinación para los dos modelos anteriores y para el modelo que incluye todas las variables del estudio en el conjunto de datos(lm(Sales∼., data =Carseats)). Utilice este valor para determinar cuál es el mejor modelo.

**# Sales = B0 + B1\*Advertising + B2\*US + e**

> Carseats$US <- factor(Carseats$US)

> modSales <- lm(Sales ~ Advertising + US, Carseats)

> modSales

Call:

lm(formula = Sales ~ Advertising + US, data = Carseats)

Coefficients:

(Intercept) Advertising USYes

6.76295 0.11848 -0.08177

**# Sales = 6.76 + 0.12\*Advertising + -0.082\*US + e 🡨--- MODELO**

> summary(modSales)

Call:

lm(formula = Sales ~ Advertising + US, data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-7.3938 -1.9653 -0.1202 1.7542 8.3564

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.76295 0.22923 29.503 < 2e-16 \*\*\*

Advertising 0.11848 0.02815 4.209 3.18e-05 \*\*\*

USYes -0.08177 0.39075 -0.209 0.834

---

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

Residual standard error: 2.726 on 397 degrees of freedom

Multiple R-squared: 0.07274, Adjusted R-squared: 0.06806

F-statistic: 15.57 on 2 and 397 DF, p-value: 3.089e-07

Para el primer modelo se puede observar que por cada unidad aumentada de Abversiting, las ventas aumentan en un 0.11 unidades, y para la ubicación, si la tienda está en US respecto a cuando no lo está, este tiende a bajar -0.08177 en las ventas. Y el coeficiente de determinación es de 0.068 para este modelo.

**# Sales = B0 + B1\*Income + B2\*ShelveLoc + e**

> modSales2 <- lm(Sales ~ Income + ShelveLoc, Carseats)

> modSales2

Call:

lm(formula = Sales ~ Income + ShelveLoc, data = Carseats)

Coefficients:

(Intercept) Income ShelveLocGood ShelveLocMedium

4.22771 0.01793 4.76752 1.87130

**# Sales = B0 + 0.01793\*Income + 4.76752\* ShelveLocGood + 1.87130\* ShelveLocMedium + e**

> summary(modSales2)

Call:

lm(formula = Sales ~ Income + ShelveLoc, data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-7.3695 -1.5868 -0.1491 1.6550 6.1990

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 4.227713 0.377355 11.204 < 2e-16 \*\*\*

Income 0.017929 0.004103 4.369 1.59e-05 \*\*\*

ShelveLocGood 4.767518 0.341192 13.973 < 2e-16 \*\*\*

ShelveLocMedium 1.871297 0.280764 6.665 8.90e-11 \*\*\*

---

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

Residual standard error: 2.288 on 396 degrees of freedom

Multiple R-squared: 0.3486, Adjusted R-squared: 0.3437

F-statistic: 70.65 on 3 and 396 DF, p-value: < 2.2e-16

Para este otro modelo, si tomamos en cuenta la variable income las ventas aumentan un 0.018 unidades, además, el factor de calidad de ubicación en el estante para los asientos, en los niveles buena y medio aumentan un 4.76 y 1.87 unidades de ventas respectivamente respecto a un nivel de calidad malo. Además, el coeficiente de determinación es de 0.3437, más alto que el modelo anterior.

**# Todas las variables**

> modSales3 <- lm(Sales ~ ., Carseats)

> summary(modSales3)

Call:

lm(formula = Sales ~ ., data = Carseats)

Residuals:

Min 1Q Median 3Q Max

-2.8692 -0.6908 0.0211 0.6636 3.4115

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 5.6606231 0.6034487 9.380 < 2e-16 \*\*\*

CompPrice 0.0928153 0.0041477 22.378 < 2e-16 \*\*\*

Income 0.0158028 0.0018451 8.565 2.58e-16 \*\*\*

Advertising 0.1230951 0.0111237 11.066 < 2e-16 \*\*\*

Population 0.0002079 0.0003705 0.561 0.575

Price -0.0953579 0.0026711 -35.700 < 2e-16 \*\*\*

ShelveLocGood 4.8501827 0.1531100 31.678 < 2e-16 \*\*\*

ShelveLocMedium 1.9567148 0.1261056 15.516 < 2e-16 \*\*\*

Age -0.0460452 0.0031817 -14.472 < 2e-16 \*\*\*

Education -0.0211018 0.0197205 -1.070 0.285

UrbanYes 0.1228864 0.1129761 1.088 0.277

USYes -0.1840928 0.1498423 -1.229 0.220

---

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

Residual standard error: 1.019 on 388 degrees of freedom

Multiple R-squared: 0.8734, Adjusted R-squared: 0.8698

F-statistic: 243.4 on 11 and 388 DF, p-value: < 2.2e-16

Para un modelo que incluye todas las variables, el coeficiente de determinación es de 0.8698, más alto que los dos modelos anteriores, por lo tanto, incluir todas las variables realiza un mejor ajuste que incluir algunas de estas.

**> anova(modSales3)**

Analysis of Variance Table

Response: Sales

Df Sum Sq Mean Sq F value Pr(>F)

CompPrice 1 13.07 13.07 12.5855 0.0004363 \*\*\*

Income 1 79.07 79.07 76.1616 < 2.2e-16 \*\*\*

Advertising 1 219.35 219.35 211.2741 < 2.2e-16 \*\*\*

Population 1 0.38 0.38 0.3683 0.5442756

Price 1 1198.87 1198.87 1154.7211 < 2.2e-16 \*\*\*

ShelveLoc 2 1047.47 523.74 504.4519 < 2.2e-16 \*\*\*

Age 1 217.39 217.39 209.3831 < 2.2e-16 \*\*\*

Education 1 1.05 1.05 1.0117 0.3151346

Urban 1 1.22 1.22 1.1753 0.2789892

US 1 1.57 1.57 1.5094 0.2199750

Residuals 388 402.83 1.04

---

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

Con la función anova se observa que la variable ShelveLoc si es significativa para el modelo, en cambio, la variable US o la ubicación no lo es.