Introducción a la analítica

Profesores César Augusto Gómez, Mauricio Alejandro Mazo y
Juan Carlos Salazar

Considere la analogía con $\bar{\mu}$ como un estimador de la media poblacional μ:

$$SE\left(\bar{\mu}\right) = \frac{\sigma^2}{n}$$

Este error estándar dice la cantidad promedio que esta estimación $\bar{\mu}$ difiere del verdadero valor poblacional μ y además, dice cómo este error estándar se encoje (shrink en inglés) con n: a mayor n, menor error estándar para $\bar{\mu}$. De manera similar, se puede pensar qué tan cerca están $\widehat{\beta}_0$ y $\widehat{\beta}_1$ de β_0 y de β_1.

Los errores estándar[1] de $\widehat{\beta}_0$ y $\widehat{\beta}_1$ están dados por:

$$SE\left(\widehat{\beta}_0\right) = \sigma^2 \left[\frac{1}{n} + \frac{\bar{x}^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}\right]$$

y

$$SE\left(\widehat{\beta}_1\right) = \frac{\sigma^2}{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2}$$

Donde $\sigma^2 = Var(\varepsilon)$. Se asume que los errores ε_i para cada observación son incorrelacionados con varianza común σ^2.

[1] Se debería escribir $\widehat{SE}\left(\widehat{\beta}_0\right)$ pero se suprime el primer 'gorro' por simplicidad en la notación

Note en la fórmula para $SE\left(\widehat{\beta}_1\right)$ que cuando los x_i están más dispersos el error estándar es más pequeño, intuitivamente se tiene más *leverage*[2] para estimar la pendiente β_1 cuando este es el caso. También, $SE\left(\widehat{\beta}_0\right)$ debe ser igual a $SE\left(\widehat{\mu}\right)$ cuando $\bar{x}$ es cero (en cuyo caso $\widehat{\beta}_0$ es igual a $\bar{y}$).

[2] Término que se define más adelante; intuitivamente se refiere a una medida de qué tan lejos están los valores de una variable independiente de los de las otras observaciones

Puesto que en general, σ^2 no se conoce, se debe estimar con base en los datos disponibles. Esta estimación se conoce como *Error Estándar Residual* (RSE en inglés):

$$RSE = \sqrt{\frac{RSS}{(n-2)}}$$

Los errores estándar se pueden usar para calcular intervalos de confianza (IC). Un IC del 95 % se define como el rango de valores tales que, con 95 % de probabilidad, ese rango contendrá el verdadero valor desconocido del parámetro. Los IC se calculan con base en los datos. Un IC del 95 % para β_1 es:

$$\widehat{\beta}_1 \pm 2 \times SE\left(\widehat{\beta}_1\right)$$

y un IC del 95 % para β_0 es:

$$\widehat{\beta}_0 \pm 2 \times SE\left(\widehat{\beta}_0\right)$$

Los errores estándar también se pueden usar para llevar a cabo pruebas de hipótesis acerca de los coeficientes. La prueba de hipótesis más común tiene que ver con probar la hipótesis nula:

$$H_0: \text{ No hay relación entre } X \text{ y } Y$$

versus la hipótesis alternativa

$$H_1: \text{ Hay relación entre } X \text{ y } Y$$

Que son equivalentes a probar:

$$H_0: \ \beta_1 = 0$$

versus

$$H_1: \ \beta_1 \neq 0$$

En la práctica, para probar estas hipótesis, se calcula un test estadístico t-student:

$$t = \frac{\widehat{\beta}_1 - 0}{SE\left(\widehat{\beta}_1\right)} \sim T_{(n-2)} \quad \text{, Bajo } H_0$$

el cual mide el número de desviaciones estándar que $\widehat{\beta}_1$ se aleja de 0. Generalmente, usando $|t|$ se calcula un valor-p que es la probabilidad de observar cualquier valor igual o mayor a $|t|$. Se rechaza H_0 si el valor-p< 0.05. Un valor-p pequeño indica que es improbable observar tal asociación tan sustancial entre el predictor y la respuesta por el mero azar, en ausencia de cualquier asociación real entre el predictor y la respuesta.

Ejemplo. Refiérase a los datos de Advertising. Un IC del 95 % para β_0 es $(6.130,\ 7.935)$ y para β_1 es $(0.042,\ 0.053)$. Se concluye que en ausencia de cualquier gasto en publicidad en TV, las ventas estarán, en promedio entre 6.130 y 7.940 unidades. Además, por cada 1000 dolares de incremento en publicidad por TV, habrá un incremento promedio en ventas entre 42 y 53 unidades. Esto se puede corrobar usando la opción *confint* en conjunto con la función *lm* del R:

```r
library(MASS)
Advertising<-read.csv(file="F:/INTRODUCCIÓN A LA ANALÍTICA/IA Virtual 02_2020/Advertising.csv",
                      header=T,sep=',',dec='.')
Sales=Advertising$sales
Tv=Advertising$TV
Radio=Advertising$radio
Newspaper=Advertising$newspaper
fit.lm<-lm(Sales~Tv)
list(summary=summary(fit.lm),IC=confint(fit.lm))
```

```
## $summary
##
## Call:
## lm(formula = Sales ~ Tv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## Tv          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 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: < 2.2e-16
##
##
## $IC
##                  2.5 %     97.5 %
## (Intercept) 6.12971927 7.93546783
## Tv          0.04223072 0.05284256
```

Evaluando la precisión del modelo. Una vez se rechaza H_0, se quiere cuantificar el grado al cual el modelo se ajusta a los datos. En regresión lineal esto se hace usualmente usando dos cantidades relacionadas: El error estándar residual (RSE) y el estadístico R^2. La tabla anterior muestra estas dos cantidades en relación a los datos de Advertising donde se ajustó un modelo de regresión lineal de Sales y Tv Budget.

RSE. En el modelo

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad i = 1, 2, \ldots, n$$

cada observación tiene su propio término de error ε_i. Debido a la presencia de este término de error, aún si se conociera la verdadera línea de regresión, no se podría predecir perfectamente a Y usando X. El RSE es una estimación de la desviación estándar de ε. Es decir, es la cantidad promedio que la respuesta se desviará de la verdadera recta de regresión. Se calcula así:

$$\begin{aligned} RSE &= \sqrt{\frac{1}{n-2} RSS} \\ &= \sqrt{\frac{1}{n-2} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2} \end{aligned}$$

```
##
## Call:
## lm(formula = Sales ~ Tv)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.3860 -1.9545 -0.1913  2.0671  7.2124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 7.032594   0.457843   15.36   <2e-16 ***
## Tv          0.047537   0.002691   17.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.259 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: < 2.2e-16

## $Mean_of_Sales
## [1] 14.0225
```

De la salida enterior, se observa que el $RSE = 3.26$. Es decir, las ventas reales en cada mercado se desvían, en promedio, de la recta de regresión por aproximadamente 3260 unidades. Otra forma de interpretar el RSE sería: aún si el modelo fuera correcto y se conocieran de manera exacta a β_0 y a β_1, cualquier predicción de ventas con base en publicidad por TV, estaría alejada por aproximadamente 3260 unidades en promedio. El RSE se considera una medida de la falta de ajuste del modelo a los datos.

Si 3260 unidades es o no un error de predicción aceptable depende del contexto del problema. En este ejemplo particular de los datos de Advertising, el valor promedio de ventas es de 14022.5 unidades, y por lo tanto el porcentaje de error es $3260/14022.5 = 0.2324835 \approx 23\,\%$.

El RSE se considera una medida de la falta de ajuste del modelo a los datos. Esto quiere decir que si las predicciones obtenidas usando el modelo están muy cercanas a los verdaderos valores observados (es decir, si $\hat{y}_i \approx y_i \quad \text{para } i = 1, 2, \ldots, n$) entonces el RSE será pequeño y se podría concluir que el modelo se ajusta muy bien a los datos. Por el contrario, si $\hat{y}_i$ está muy alejado de y_i para una o más observaciones, entonces el RSE puede ser muy grande e indicar que el modelo no se ajusta bien a los datos.

El estadístico R^2. El RSE proporciona una medida absoluta de falta de ajuste del modelo $Y = \beta_0 + \beta_1 X + \varepsilon$, pero en vista de que se mide en las unidades de Y, no siempre es claro que se entiende por un buen RSE. El estadístico R^2 proporciona una medida alternativa de ajuste. El R^2 toma la forma de una proporción (la proporción de varianza explicada) y por lo tanto siempre es un valor entre 0 y 1. Se calcula como:

$$\begin{aligned} R^2 &= \frac{TSS - RSS}{TSS} \\ &= 1 - \frac{RSS}{TSS} \end{aligned}$$

donde

$$TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2 \quad \text{Suma Total de Cuadrados (TSS)}$$

y

$$RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \quad \text{Suma de cuadrados de residuales (RSS)}$$

El TSS mide la varianza total en la respuesta Y y se puede pensar como la cantidad de variabilidad inherente en la respuesta antes de que se lleve a cabo la regresión. En contraste, el RSS mide la cantidad de variabilidad que queda sin explicar después de llevar a cabo la regresión. De esta manera TSS-RSS mide la cantidad de variabilidad en la respuesta que es explicada (o removida) al llevar a cabo la regresión, y el R^2 mide la proporción de variabilidad en Y que es explicada por X. Un R^2 cercano a 1 indica una gran proporción de variabilidad explicada por la regresión.

Un R^2 cercano a 0 indica una poca proporción de variabilidad explicada por la regresión y generalmente se asocia a modelos mal especificados. Refiérase al ejemplo de Advertising:

```
library(MASS)
 Advertising<-read.csv(file="F:/INTRODUCCIÓN A LA ANALÍTICA/IA Virtual 02_2020/Advertising.csv",
                      header=T,sep=',',dec='.')
#Advertising<-read.csv(file="Advertising.csv",header=T,sep=',',dec='.')
Sales=Advertising$sales
Tv=Advertising$TV
Radio=Advertising$radio
Newspaper=Advertising$newspaper
fit.lm<-lm(Sales~Tv)
#summary(fit.lm)
#names(summary(fit.lm))
summary(fit.lm)$adj.r.squared
```

```
## [1] 0.6099148
```

Este valor implica que aproximadamente dos tercios de la variabilidad en las ventas se explica con una regresión lineal sobre TV advertising.

Puede ser retador determinar que es un buen R^2 y en general su interpretación depende del problema. El R^2 es una medida de la relación lineal entre X y Y. La correlación muestral se define como:

$$\widehat{\rho} = \widehat{Cor(X, Y)} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2}}$$

La cual también es una medida de la relación lineal entre X y Y. $\widehat{\rho} \in [-1,\ 1]$

Refiérase al ejemplo de Advertising:

```
library(MASS)
 Advertising<-read.csv(file="F:/INTRODUCCIÓN A LA ANALÍTICA/IA Virtual 02_2020/Advertising.csv",
                       header=T, sep=',',dec='.')
Sales=Advertising$sales
Tv=Advertising$TV
Radio=Advertising$radio
Newspaper=Advertising$newspaper
cor(Tv, Sales,  method = "pearson", use = "complete.obs")
```

```
## [1] 0.7822244
```

Este valor indica una asociación lineal positiva, cercana a 0.8, entre Sales y Tv Advertising.

La regresión lineal simple es una aproximación útil para predecir una respuesta con base en un solo predictor. Sin embargo, en la práctica, frecuentemente se tiene más de un predictor. Por ejemplo, en los datos de advertising, solo se ha examinado la relación entre Ventas (Sales) y publicidad por TV (TV advertising) a pesar de que se tiene información del gasto en publicidad en Radio y en Periódico y también se quiere saber si estos dos medios publicitarios se asocian con las ventas ¿Cómo se pueden incorporar en el modelo?

Una opción, es correr tres modelos separados cada uno con un medio publicitario distinto como predictor:

Modelo 1

$$Sales = \beta_0 + \beta_1 TV + \varepsilon$$

Modelo 2

$$Sales = \beta_0 + \beta_1 Radio + \varepsilon$$

Modelo 3

$$Sales = \beta_0 + \beta_1 Newspaper + \varepsilon$$

Refiérase al ejemplo de Advertising:

```
library(MASS)
 Advertising<-read.csv(file="F:/INTRODUCCIÓN A LA ANALÍTICA/IA Virtual 02_2020/Advertising.csv",
                       header=T,sep=',',dec='.')
#Advertising<-read.csv(file="Advertising.csv",header=T,sep=',',dec='.')
Sales=Advertising$sales
Tv=Advertising$TV
Radio=Advertising$radio
Newspaper=Advertising$newspaper
Modelo1<-summary(lm(Sales~Tv))$coefficients
Modelo2<-summary(lm(Sales~Radio))$coefficients
Modelo3<-summary(lm(Sales~Newspaper))$coefficients
list(Modelo1=Modelo1,Modelo2=Modelo2,Modelo3=Modelo3)
```

```
## $Modelo1
##               Estimate  Std. Error  t value     Pr(>|t|)
## (Intercept) 7.03259355 0.457842940 15.36028 1.40630e-35
## Tv          0.04753664 0.002690607 17.66763 1.46739e-42
##
## $Modelo2
##              Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 9.3116381 0.56290050 16.542245 3.561071e-39
## Radio       0.2024958 0.02041131  9.920765 4.354966e-19
##
## $Modelo3
##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 12.3514071 0.62142019 19.876096 4.713507e-49
## Newspaper    0.0546931 0.01657572  3.299591 1.148196e-03
```

Del Modelo 2 (Sales sobre Radio) se observa que un incremento de 1000 dolares en publicidad por radio se asocia con un incremento en ventas de cerca de 203 unidades (aprox 0.2024958×1000). Del Modelo 3 (Sales sobre Newspaper) se observa que un incremento de 1000 dolares en publicidad por prensa se asocia con un incremento en ventas de cerca de 55 unidades (aprox 0.0546931×1000). Sin embargo, esta solución de ajustar modelos separados no es del todo satisfactoria y puede conducir a reportar asociaciones inexistentes.

Primero que todo, no es claro cómo hacer una sola predicción de las ventas dados niveles de valores de los tres medios publicitarios ya que cada presupuesto está asociado con una regresión diferente. Segundo, cada una de las tres ecuaciones de regresión ignora los otros dos medios en la generación de las estimaciones de los coeficientes de regresión (es decir, se ignora la correlación entre estos medios). Por estas razones, es mejor usar regresión lineal múltiple.

Esto se puede hacer asignándole a cada predictor una pendiente separada, pero un intercepto común. Es decir

$$Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper + \varepsilon$$

En general, si se tiene p predictores distintos, el modelo de regresión lineal múltiple toma la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \varepsilon$$

Donde X_j representa el j-ésimo predictor y β_j cuantifica la asociación entre esa variable X_j y la respuesta Y. β_j se interpreta como el efecto promedio en Y por un incremento de una unidad en X_j, manteniendo fijos todos los otros predictores.

Estimación de los coeficientes de regresión en el modelo lineal múltiple. El vector de coeficientes $\theta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ no se conoce y se debe estimar, usando los datos, para obtener $\hat{\theta} = \left(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p\right)$. Con $\hat{\theta}$ se pueden hacer predicciones usando la fórmula

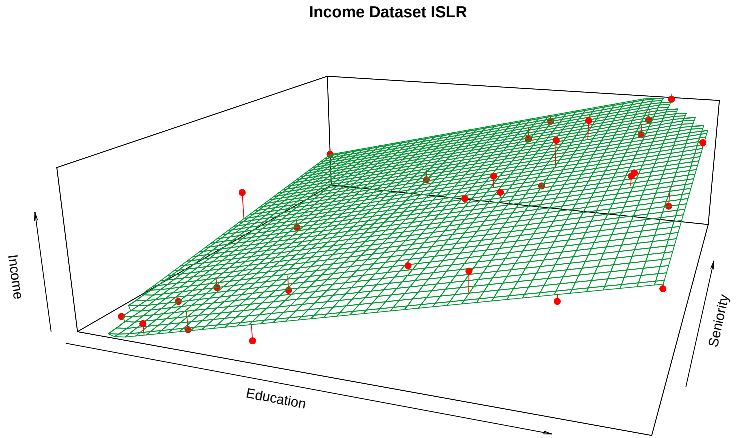
$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Estos parámetros se estiman usando OLS. En OLS se obtiene un $\hat{\theta}$ (coeficientes de OLS estimados) que minimiza al RSS:

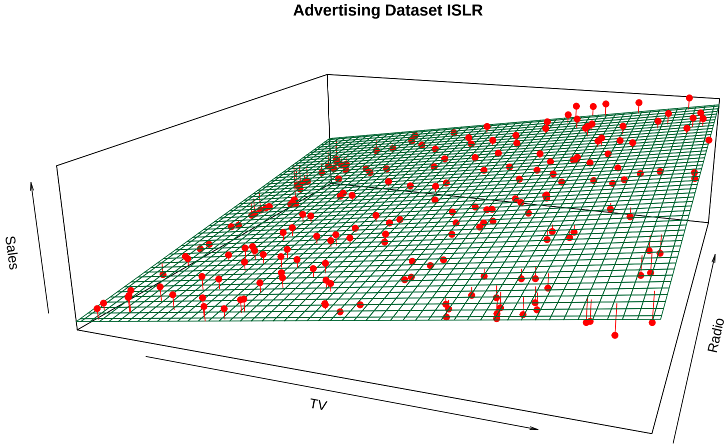
$$\begin{aligned} RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 \\ &= \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \hat{\beta}_2 x_{i2} + \cdots + \hat{\beta}_p x_{ip}\right)^2 \end{aligned}$$

Para ilustrar, en el caso de tres dimensiones (dos predictores y una respuesta) la figura que se obtiene, con los OLS, es un plano. Este plano minimiza la suma de cuadrados de las distancias verticales entre cada observación y el plano. A continuación se presentan algunos ejemplos de estos planos OLS.

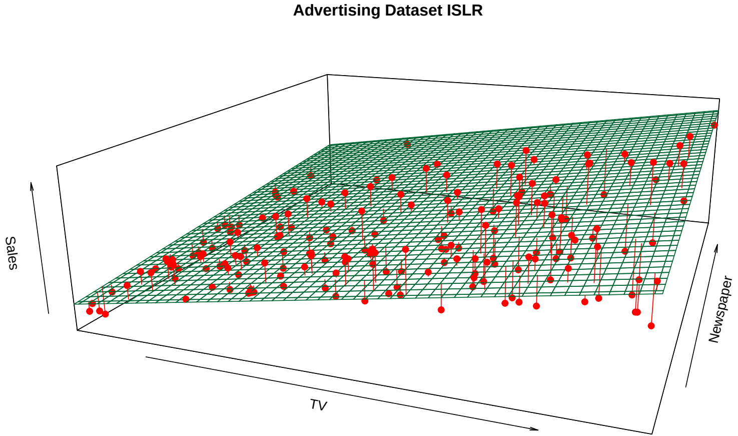
Refiérase al ejemplo de Income:



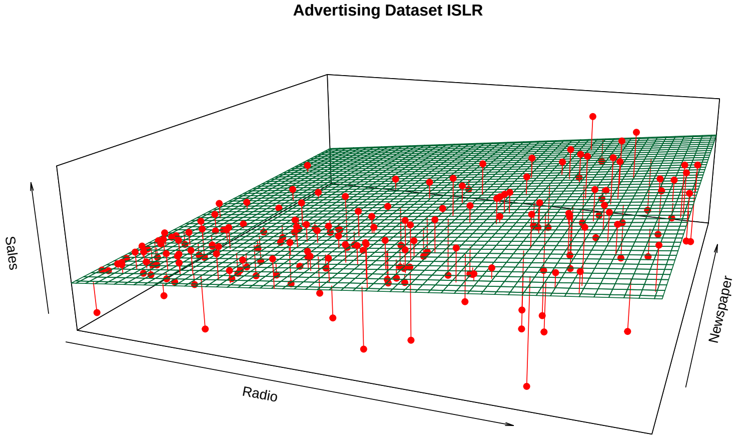
Refiérase al ejemplo de Advertising (Sales$\sim$Tv+Radio):



Refiérase al ejemplo de Advertising (Sales~Tv+Newspaper):



Refiérase al ejemplo de Advertising (Sales~Radio+Newspaper):



De los últimos tres gráficos, se observa que el modelo estimado por OLS $Sales = \hat{\beta}_0 + \hat{\beta}_1 Radio + \hat{\beta}_2 Newspaper$ es el que muestra el ajuste más pobre de los tres, ya que los residuales asociados se perciben mayores a los de los otros dos modelos, es decir la suma de los cuadrados de las distancias verticales entre cada observación (que es un punto en $\mathbb{R}^2$) y el plano estimado OLS no es la mínima. Sin embargo, y similar al caso de modelos lineales simples individuales, considerar modelos de a dos predictores, podría no ser la mejor idea. Una aproximación más adecuada consiste en considerar los tres predictores: TV, Radio y Newspaper.

Refiérase al ejemplo de Advertising, se ajustará el modelo $Sales = \beta_0 + \beta_1 TV + \beta_2 Radio + \beta_3 Newspaper + \varepsilon$:

```
##                 Estimate  Std. Error    t value     Pr(>|t|)
## (Intercept)  2.938889369 0.311908236  9.4222884 1.267295e-17
## TV           0.045764645 0.001394897 32.8086244 1.509960e-81
## Radio        0.188530017 0.008611234 21.8934961 1.505339e-54
## Newspaper   -0.001037493 0.005871010 -0.1767146 8.599151e-01
```

El modelo ajustado es

$$\widehat{Sales} = 2.94 + 0.046\,TV + 0.189\,Radio - 0.001\,Newspaper$$

A pesar de que individualmente cada medio se asocia de manera importante con las ventas (en los modelos lineales simples independientes), cuando se consideran juntos en el modelo de regresión lineal múltiple esto ya no es cierto, puesto que el efecto del predictor Newspaper no es importante (Valor-p=0.86). En conclusión, los coeficientes estimados con regresión simple y múltiple pueden ser completamente diferentes.

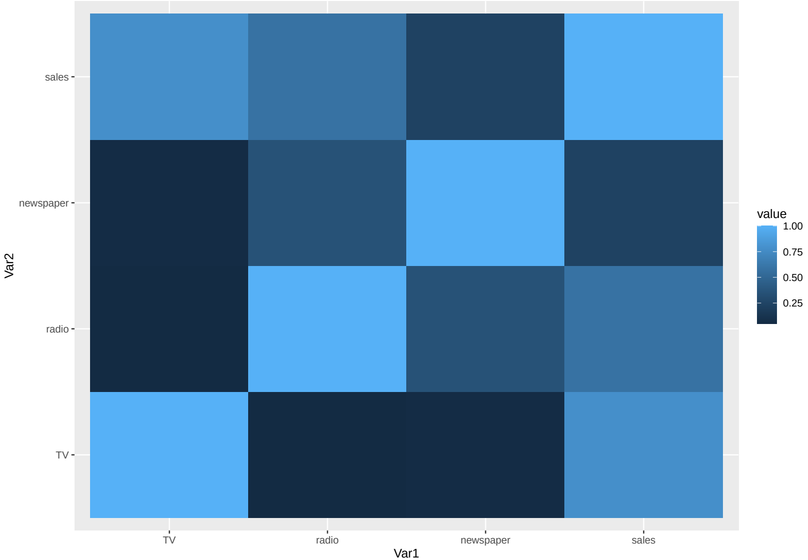
Esta diferencia se explica por el hecho de que en el caso de regresión lineal simple, las estimaciones de las pendientes representan el respectivo efecto en las ventas al incrementar en una unidad ese predictor ignorando los otros predictores. Por el contrario, en el caso de regresión lineal múltiple, los coeficientes de las pendientes representan el efecto promedio de incrementar en una unidad ese predictor sobre las ventas, mientras los otros se mantienen fijos (no se ignoran). Pero ¿tiene sentido que el modelo de regresión lineal múltiple sugiera ninguna relación entre Sales y Newspaper mientras que el modelo de regresión lineal simple implica lo contrario? Pues sí tiene sentido y se explica la razón.

Considere la matriz de correlación de los predictores y la respuesta:

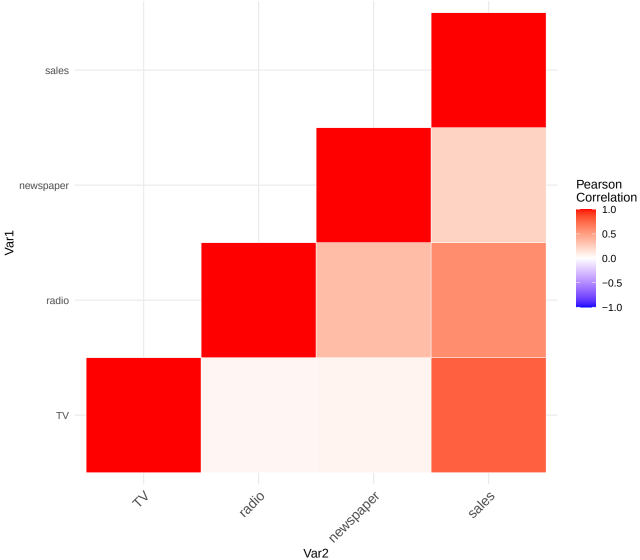
```
library(ISLR)
library(ggplot2)
library(MASS)
library("plot3D")
 Advertising<-read.csv(file="F:/INTRODUCCIÓN A LA ANALÍTICA/IA Virtual 02_2020/Advertising.csv",
                       header=T,sep=',',dec='.')
#Advertising<-read.csv(file="Advertising.csv",
#                      header=T,sep=',',dec='.')
cor(Advertising[,2:5])
```

```
##                   TV      radio  newspaper     sales
## TV        1.00000000 0.05480866 0.05664787 0.7822244
## radio     0.05480866 1.00000000 0.35410375 0.5762226
## newspaper 0.05664787 0.35410375 1.00000000 0.2282990
## sales     0.78222442 0.57622257 0.22829903 1.0000000
```

Considere la matriz de correlación de los predictores y la respuesta:



Considere la matriz de correlación de los predictores y la respuesta:



Observe que la correlación entre Radio y Newspaper es 0.35. Esto revela una tendencia a gastar más en publicidad por periódico en mercados donde se gasta más en publicidad por Radio. Ahora suponga que la regresión múltiple es correcta y que la publicidad por periódico no tiene un impacto directo en las ventas, mientras que publicidad por Radio sí incrementa las ventas. Entonces, en mercados donde se gasta más en Radio, las ventas tenderán a ser mayores y, tal como lo muestra la matriz de correlación, tambien se tiende a gastar más en publicidad por periódico en esos mismos mercados.

Por lo tanto, en un modelo lineal simple que solo examina Sales versus Newspaper, se observará que valores grandes de gasto en publicidad por periódico tienden a asociarse con valores grandes en las ventas, aún cuando en el modelo múltiple se observa que Newspaper no afecta de manera real las ventas. Esto implica que Newspaper es un sustituto (surrogate en inglés) para publicidad por Radio; es decir Newspaper se lleva algo de crédito por el efecto de Radio en las ventas[3]. Este resultado parece ser contraintuitivo.

[3] Es un efecto confusor: en presencia de Radio, Newspaper modifica su efecto. En ausencia de Radio, Newspaper parece explicar una asociación importante

A pesar de que este resultado parecer ser contraintuitivo, es muy común en la práctica. Considere un ejemplo absurdo para ilustrar este punto. Suponga que se implementa una regresión entre ataques de tiburones y ventas de helado con base en datos recolectados en una playa en un período de tiempo (verano, altas temperaturas) y se observa una correlación positiva (similar a Sales y Newspaper). Por supuesto nadie a recomendado prohibir (aún) la venta de helados a fin de reducir los ataques de tiburones.

En realidad, las altas temperaturas hacen que más gente vaya a las playas, lo cual incrementa las ventas de helados y los ataques de tiburones. Un modelo de regresión múltiple de ataques de tiburón versus ventas de helado y temperatura revela que la variable ventas de helado ya no es estadísticamente significativa después de ajustar por temperatura.

Algunas preguntas importantes. Cuando se implementa regresión lineal múltiple, usualmente interesa responder algunas preguntas importantes:

- ¿Hay al menos un predictor útil (del conjunto $X_1, \ldots, X_p$) para predecir la respuesta?
- ¿Todos los predictores ayudan a explicar Y o solamente un subconjunto de los predictores es útil?
- ¿Qué tan bien se ajusta el modelo a los datos?
- Dado un conjunto de predictores ¿qué valor de Y se debería predecir y qué tan precisa es esta predicción?

Se responderá cada una de estas preguntas.

- **¿Hay al menos un predictor útil (del conjunto $X_1, \ldots, X_p$) para predecir la respuesta?** Recuerde que en RLS a fin de determinar si hay una relación entre la respuesta y el predictor se chequea si $\beta_1 = 0$. En el caso de regresión múltiple con p predictores, se necesita preguntar si todos los coeficientes son cero $\beta_1 = \beta_2 = \cdots = \beta_p = 0$. Usualmente se recurre a pruebas de hipótesis para responder esta preguntas:

$$H_o: \ \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

versus

$$H_1: \ \text{Al menos un } \beta_j \neq 0 \text{ para } j = 1, 2, \ldots, p$$

Estadístico de prueba (EP):

$$F = \frac{(TSS - RSS)/p}{RSS/(n-p-1)} \sim F(p,\ n-p-1) \quad \text{, Bajo } H_0$$

donde $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ y $RSS = \sum_{i=1}^{n} (y_i - \hat{y})^2$