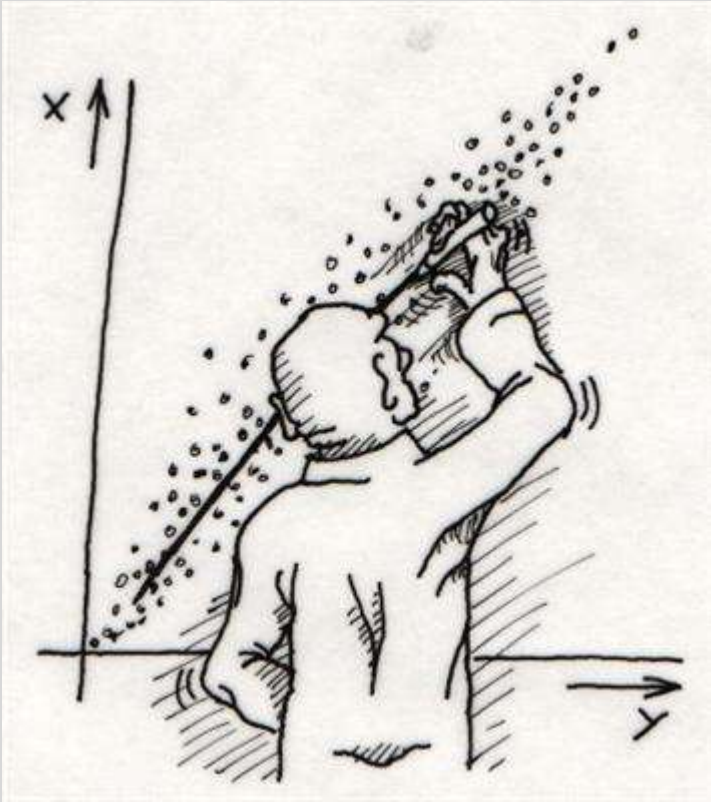


DigitalHouse > Coding School

DATA SCIENCE

MÓDULO 3

Regresión lineal Múltiple

Septiembre de 2017

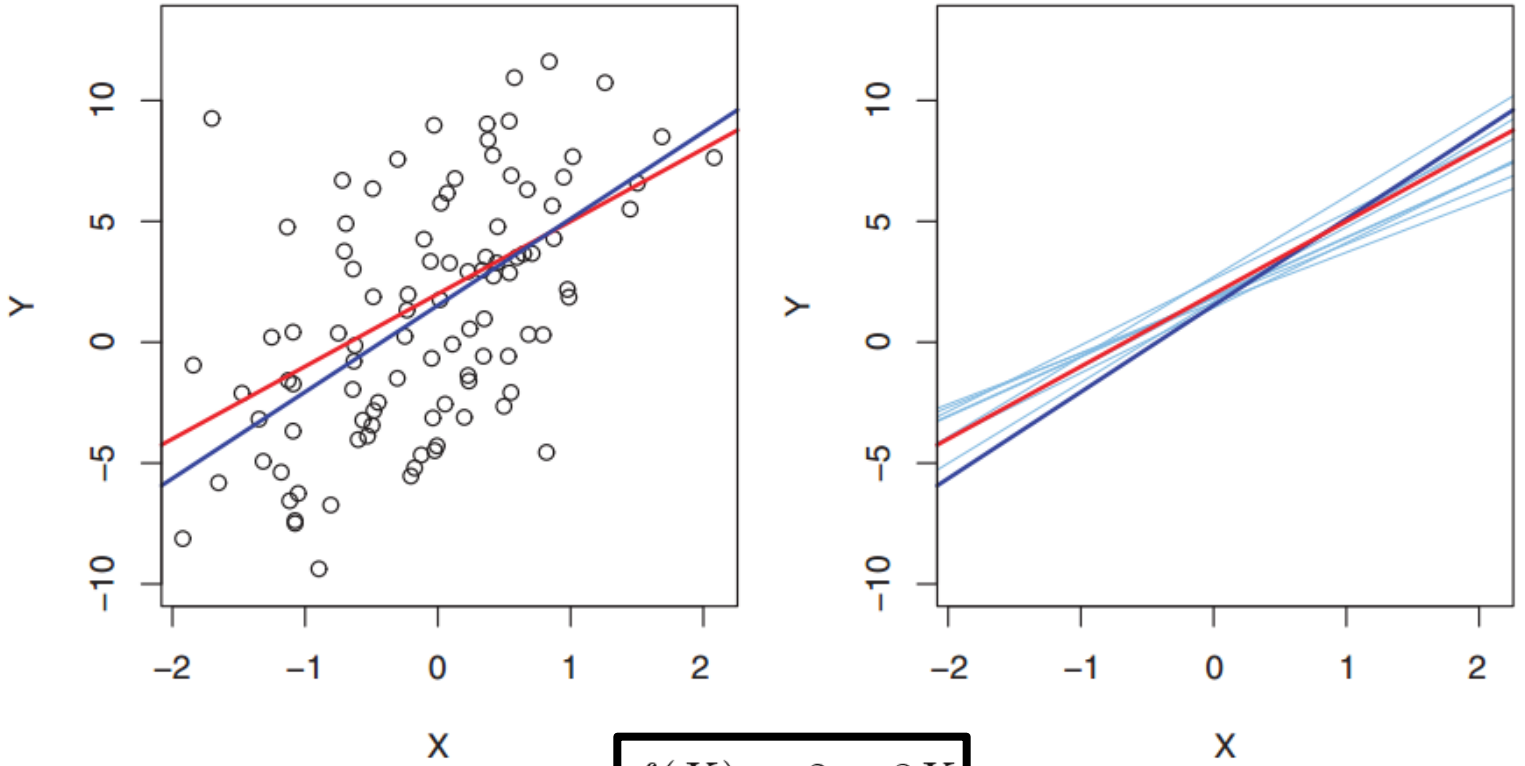
Objetivos de la clase

1. **Evaluar la precisión de los coeficientes estimados en Regresión Lineal**

2. **Profundizar algunos conceptos de Regresión Lineal Múltiple**

Precisión de los coeficientes estimados

Función de Regresión Poblacional y Muestral



$$f(X) = 2 + 3X$$

En el gráfico anterior se generaron datos a partir de una función conocida y ruido aleatorio con media cero.

En el gráfico de la Izquierda

- La línea roja representa la **verdadera relación** (f(x)=2+3X), que es llamada la "**función de regresión poblacional**").
- La línea azul representa la función de estimación por mínimos cuadrados de f(x) estimada en base a los datos.

En el gráfico de la Derecha

- La función de regresión poblacional está en rojo y la de mínimos cuadrados en azul oscuro.
- En azul claro, hay 10 funciones de mínimos cuadrados basadas en **submuestras independientes y aleatorias** de los datos. Cada línea de mínimos cuadrados es diferente pero **en promedio, las líneas están cerca de la función de regresión poblacional.**

¿Parece confuso?

Si tenemos un único dataset, qué significa que dos funciones de regresión diferentes describen la relación entre el predictor y la respuesta?

Fundamentalmente, este concepto es una extensión natural de la aproximación estadística estándar de usar información de una muestra para estimar las características de una población más grande.

Por ejemplo: se acuerdan cuando estimamos μ a partir de la media muestral.

- μ es una característica poblacional
- $\overline{X}$ es una característica de la muestra.

¿Cómo se relacionan μ y $\overline{X}$?

Estimador Insesgado

Podemos continuar con la analogía entre la regresión lineal y la estimación de la media de una variable aleatoria.

Si usamos la media muestral $\overline{X}$ para estimar μ este estimador es insesgado.

¿Qué significa?

Si promediamos todas las muestras posibles de tamaño parámetro obtenidas de un gran número de sets de observaciones todas del mismo tamaño n, entonces este promedio será *exactamente* igual a μ.

La propiedad de insesgadez se cumple para los estimadores de coeficientes de mínimos cuadrados bajo el supuesto de exogeneidad.

Error estándar asociado con los estimadores

¿Cuán lejos estará una única estimación $\overline{X}$ de del parámetro?

En general para calcular esta distancia, utilizamos la distribución de los estimadores del parámetro!

Recordemos que si las observaciones son independientes con igual varianza dada por σ^2

$$\mathrm{Var}(\hat{\mu}) = \mathrm{SE}(\hat{\mu})^2 = \frac{\sigma^2}{n},$$

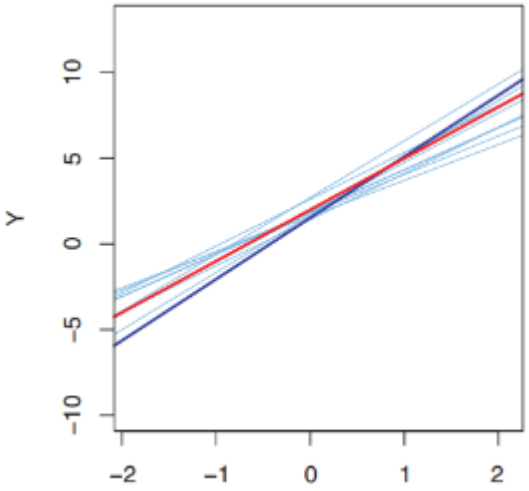
Los errores estándar pueden ser usados para calcular los intervalos de confianza de los estimadores de los coeficientes.

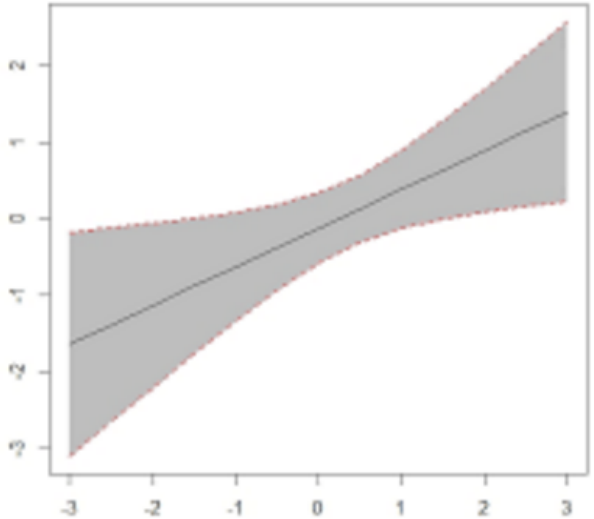
Un intervalo de confianza de 95% se define como un rango de valores tales que con una probabilidad de 95%, el rango contendrá el valor verdadero (poblacional y desconocido) del parámetro.

El rango se define en términos de límites inferior y superior, calculados a partir de la muestra de datos. Por lo tanto no pueden depender de parámetros desconocidos.

Con esta misma lógica nos preguntamos cuán cerca están $\hat{\beta}_0$ y $\hat{\beta}_1$ de los verdaderos valores β_0 y β_1

Nuestro objetivo en Regresión Lineal es obtener **intervalos de confianza** para los estimadores y las predicciones.





El **intervalo de confianza de la predicción** se hace más ancho a medida que se aleja del centro de los datos con los que calculamos el modelo

Intervalo de confianza para las predicciones

Se puede demostrar, que la fórmula para el **intervalo de confianza de las predicciones** es:

$$\hat{y} \pm t^{\star}_{n-2}\, s_y \sqrt{\frac{1}{n} + \frac{(x^{\star} - \bar{x})^2}{(n-1)s_x^2}}$$

Conclusiones a partir de esta fórmula:

- A medida que n crece, el estadístico t decrece, hasta un N aproximadamente igual a 20 donde es asintótico.
- A medida que n crece el intervalo de confianza se hace más estrecho.
- Cuando la varianza del regresor crece (hay más poder explicativo en el mismo/ modelamos todos los casos posibles), la confianza en la estimación crece, es decir, se achica el tamaño del intervalo de confianza
- Cuando el dato se encuentra lejos de la media (o centroide en n dimensiones), la calidad de la predicción decrece.

Se puede demostrar, que la fórmula para el **intervalo de confianza de los coeficientes** es:

$$\hat{\beta}_j \pm t_{n-m-1,\alpha/2}\,\widehat{SE}$$

Vamos a desarrollar en detalle la fórmula de la estimación del desvío estándar de los coeficientes $\widehat{SE}$

Veremos que la magnitud de este desvío está íntimamente relacionada con la capacidad de explicar linealmente los coeficientes en función de los demás. Cuando la explicación de un coeficiente por los otros es muy buena (tiene un R2 elevado), entonces el intervalo de confianza para el estimador se hace muy alto.

Este problema se conoce como **multicolinealidad**.

$$\widehat{\mathrm{SE}}(\hat{\beta}_1)$$

Estrictamente hablando, cuando σ^2 es estimado a partir de los datos deberíamos escribir el error estándar de esta forma, para indicar que se ha hecho una estimación.

Pero para simplificar la notación, **no utilizaremos el sombrero extra en nuestras presentaciones.**

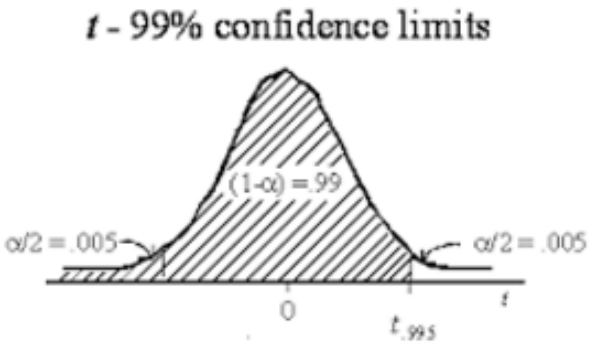
Significatividad de los coeficientes estimados

Estimando los desvíos estándar podremos evaluar la significatividad de cada uno de los coeficientes con una prueba de hipótesis.

¿Qué probabilidad existe de observar lo que observamos si el verdadero valor del coeficiente es cero?

p-valor bajo ($p<0.05$ o $p<0.01$): Es improbable observar los datos si el verdadero valor es 0.

p-valor alto : Es probable que la asociación observada sea producto del azar.

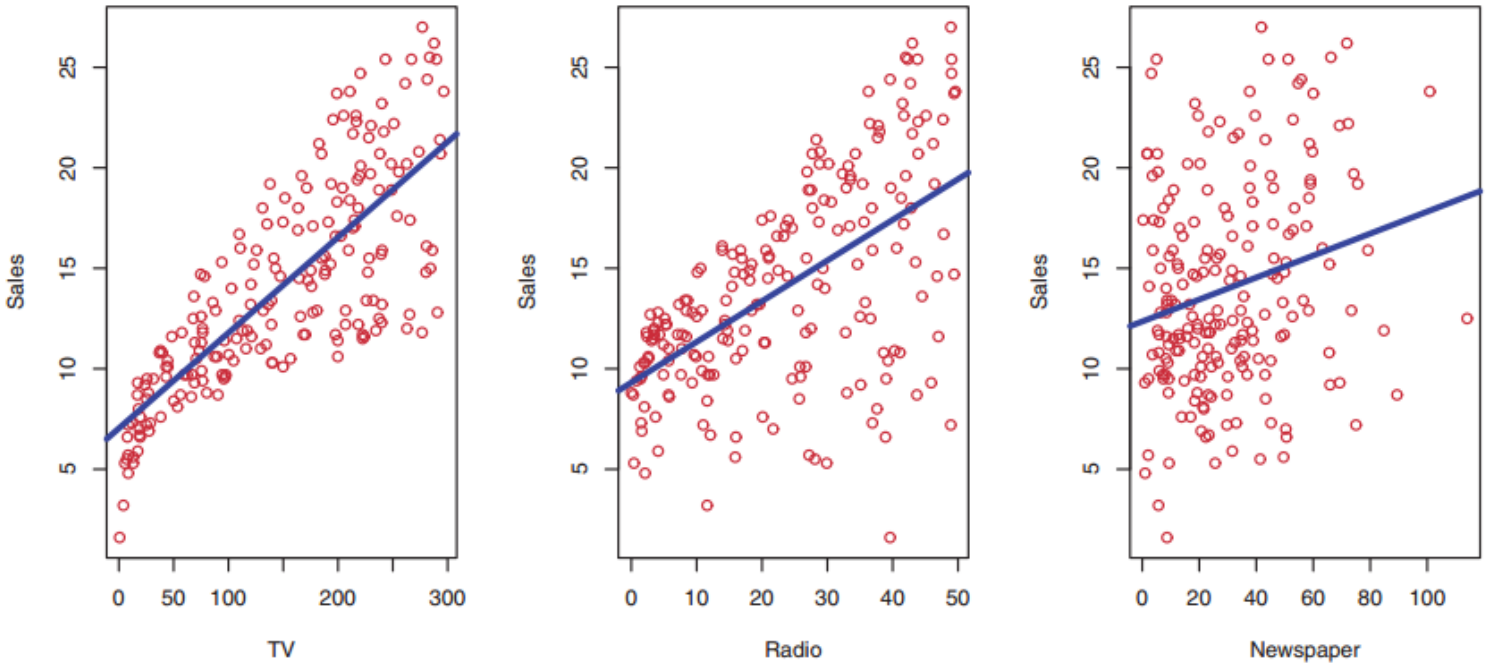


Recordemos el dataset Advertising

Supongamos que que somos consultores estadísticos, y nos contratan con el objetivo de aumentar las ventas de un determinado producto.
El dataset Advertising consiste en las ventas del producto en 200 mercados, y el presupuesto dedicado en publicidad en 3 medio: TV, radio y diario.
Si logramos identificar una relación entre la inversión en publicidad y las ventas, podremos recomendarle a nuestro cliente donde hacía donde debe dirigir su inversión en publicidad.

La variables predictoras serán los presupuestos para cada canal y la label serán las ventas.

Así se ve una primera visualización del dataset.



Intervalos de Confianza - Interpretación

En el caso del dataset **advertising,** el intervalo de confianza de 95% para β_0 es [6.130, 7.935] y para β_1 es [0.042, 0.053]. Por lo tanto:

- Podemos concluir que en ausencia de cualquier publicidad, las ventas (**sales**) caerán en algún valor entre 6130 y 7940 unidades (con un 95% de confianza)..

- Y además, que por cada incremento de $1000 en **TV**, habrá un incremento promedio en **sales** de entre 42 y 53 unidades.

¿Existe evidencias para afirmar que hay relación entre X e Y?

Los errores estándar de los estimadores de los coeficientes también pueden ser usados para realizar tests de hipótesis.

El test de significación individual tiene las siguientes hipótesis

H_0: No hay relación entre X e Y

$$H_0 : \beta_1 = 0$$

versus la **hipótesis alternativa:**

H_a: Hay alguna relación entre X e Y

$$H_a : \beta_1 \neq 0,$$

¿Existe evidencias para afirmar que hay relación entre X e Y?

Si $\beta_1 = 0$, entonces el modelo se reduce a $Y = \beta_0 + e$

Por lo tanto X no estaría asociado a Y

Necesitamos determinar $\hat{\beta}_1$ (nuestro estimador para β_1) está lo **suficientemente lejos de cero**, para que podamos estar seguros de que β_1 no es cero.

¿Cuánto creen que es "**suficientemente lejos de cero**"?

Esto, por supuesto, depende de la precisión de $\hat{\beta}_1$, es decir depende del error estándar de nuestro estimador del coeficiente (que a su vez depende del desvío estándar del estimador y del tamaño muestral n).

Si el error estándar de $\hat{\beta}_1$ es pequeño, entonces incluso valores pequeños de nuestro estimador pueden darnos evidencia fuerte de que β_1 != 0 y por lo tanto que hay relación entre X e Y.

En contraste, si el error estándar es grande, entonces $\hat{\beta}_1$ debe ser grande en valores absolutos para que podamos rechazar la hipótesis nula.

En la práctica, se computa el **estadístico t** que mide la cantidad de desviaciones estándar a las que el estimador $\hat{\beta}_1$ se encuentra de cero.

$$t = \frac{\hat{\beta}_1 - 0}{\mathrm{SE}(\hat{\beta}_1)},$$

Básicamente, es cuestión de calcular la probabilidad de observar cualquier valor igual a $|t|$ o mayor, asumiendo que $\beta_1 = 0$ (para una variable t con distribución T-Student con n-1 grados de libertad).

Recordemos que llamamos a esta probabilidad ***p-value***.

Simplificando, interpretamos el *p-value* de la siguiente forma:

Un *p-value* pequeño indica que es poco probable observar un valor del estadístico como el observado o más extremo asumiendo que H0 es verdadera. Por lo tanto, si el p-value es chico podemos rechazar H0 con baja probabilidad de equivocarnos.

Los coeficientes del modelo de mínimos cuadrados

	Coefficient	Std. error	t-statistic	p-value
Intercept	7.0325	0.4578	15.36	< 0.0001
TV	0.0475	0.0027	17.67	< 0.0001

Un incremento de $1000 en el presupuesto de publicidad en TV está asociado con un incremento de alrededor de 50 unidades en las ventas (la variable dependiente está en miles de unidades y la de presupuesto en miles de U$S).

Los coeficientes del modelo de mínimos cuadrados

- Notar que los coeficientes para $\hat{\beta}_0$ y $\hat{\beta}_1$ son muy grandes en relación a sus errores estándar. Por ende, los estadísticos t también son grandes.

- Las probabilidades de "observar" tales valores si H_0 fuera verdadera serían casi cero. Por ende, podemos rechazar tanto la H_0 de que el intercepto como la H_0 de que el coeficiente de pendiente son cero en la población.

- Un p-value pequeño para el intercepto implica que podemos rechazar la hipótesis nula que afirma que $\beta_0 = 0$. Lo mismo sucede con un p-value pequeño para el coeficiente asociado a TV. Podemos concluir, entonces, que hay relación entre TV y **sales**.

- Al correr una regresión lineal, es común reportar el error estándar de cada estimador:

$$\mathrm{SE}(\hat{\beta}_0) \text{ y } \mathrm{SE}(\hat{\beta}_1)$$

- Esto es útil para construir intervalos de confianza de los estimadores de los coeficientes.
- Evaluar la significatividad de cada estimador, mediante un test estadístico.
  - **p-valor bajo (típicamente, p<0.05 o p<0.01) → es improbable observar al azar una asociación semejante entre X e Y.**
  - **p-valor alto → es probable que la asociación observada sea sólo consecuencia del azar.**

Regresión Lineal Múltiple

- En la práctica tenemos más de un único predictor
- Por ejemplo en el dataset Advertising, tenemos datos sobre **radio** y **newspaper** y podemos querer saber si estos medios están asociados con **sales**

¿Cómo podemos extender nuestra análisis para agregar estos dos predictores adicionales?

Una opción es calcular tres regresiones lineales simples por separado, cada una usando un medio como predictor.
Sin embargo esto no es del todo satisfactorio:

1. No es claro cómo hacer una única predicción de ventas a partir de los 3 predictores ya que cada uno tiene una ecuación de regresión separada
2. Cada una de las regresiones simples **ignora** a los otros dos medios al estimar los coeficientes de regresión

Hoy veremos que si los presupuestos de publicidad están correlacionados entre ellos en nuestro dataset, esto puede llevar a estimaciones erróneas de los efectos individuales de cada medio en las ventas. Vamos a tenerlo en cuenta

En lugar de ajustar un modelo distinto de regresión simple para cada predictor, una mejor aproximación es extender el modelo de regresión simple para que puede incluir múltiples predictores.

Podemos darle a cada predictor un coeficiente separado en un único modelo.

En general, supongamos que tenemos p predictores distintos, entonces el modelo de regresión lineal múltiple toma la siguiente forma:

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

X_j representa el predictor j y β_j cuantifica la asociación entre esa variable y la respuesta.

Interpretamos β_j como el efecto promedio en Y de un incremento unitario en X_j, manteniendo todos los otros predictores constantes.

En el ejemplo del dataset Advertising, esto se representa así:

$$\texttt{sales} = \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times \texttt{newspaper} + \epsilon.$$

	Coefficient	Std. error	t-statistic	p-value
Intercept	2.939	0.3119	9.42	< 0.0001
TV	0.046	0.0014	32.81	< 0.0001
radio	0.189	0.0086	21.89	< 0.0001
newspaper	−0.001	0.0059	−0.18	0.8599

TABLE 3.4. *For the* Advertising *data, least squares coefficient estimates of the multiple linear regression of number of units sold on radio, TV, and newspaper advertising budgets.*

La tabla anterior muestra los coeficientes estimados de la regresión múltiple cuando **TV**, **radio** y **newspaper** son usadas para predecir las ventas de productos (**sales**) usando el **Advertising**.

Interpretamos los resultados de esta forma: para cantidades fijas de TV y newspaper, gastar un adicional de $1000 dataset en radio, produce un incremento en ventas de aproximadamente 189 unidades.

Revisemos los resultados para la regresión Simple de **sales** contra **newspaper**

Simple regression of `sales` on `newspaper`

	Coefficient	Std. error	t-statistic	p-value
`Intercept`	12.351	0.621	19.88	< 0.0001
`newspaper`	0.055	0.017	3.30	< 0.0001

Comparemos los estimadores de los coeficientes de la regresión múltiple contra los de la regresión simple para **newspaper** contra **sales**

En la regresión lineal el estimador del coeficiente para newspaper era estadísticamente distinto de cero.

En la regresión múltiple es cercano a cero y el p-value correspondiente no es significativo, con un valor cercano a 0.86

Tiene sentido que la regresión lineal indique que no hay relación entre sales y newspaper mientras que la regresión simple implica lo contrario?

De hecho, si lo tiene. Consideremos la matriz de correlación para las tres variables predictoras y la variable respuesta.

Notemos que la correlación entre radio y newspaper es 0.35. Esto revela una tendencia a gastar más en publicidad en newspaper en mercados donde más se gasta en publicidad en radio.

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0567	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

TABLE 3.5. *Correlation matrix for* TV, radio, newspaper, *and* sales *for the* Advertising *data.*

Ahora supongamos que la regresión que la regresión múltiple es correcta y la publicidad en newspaper no tiene impacto directo en **sales** pero la publicidad en radio incrementa **sales**.

Entonces en mercados donde gastamos más en **radio**, nuestras **sales** tenderán a ser más altas, como muestra la matriz de correlación, también tendemos a gastar más en publicidad en newspaper en esos mismos mercados

Por lo tanto, en una regresión lineal simple que sólo examina **sales vs newspaper,** sólo observamos que valores más grandes de **newspaper** tienden a estar asociados con valores más altos de **sales,** incluso aunque la publicidad en newspaper no afecta a las ventas en **sales**.

Por lo tanto la variable **newspaper** puede esconder el el efecto de la variable **radio;** es decir **newspaper** recibe el “crédito” por los efectos de **radio** en **sales**

Este resultado, un poco contra-intuitivo, es muy común en muchas situaciones de la vida real. Consideremos un ejemplo absurdo para ilustrar el punto. Si computamos una regresión de ataques de tiburones contra ventas de helado para datos recolectados en una playa específica sobre un período de tiempo nos mostraría una relación positiva, similar a la que hemos visto entre sales y newspaper.

Por supuesto, nadie ha sugerido (hasta ahora) que los helados deberían ser prohibidos en las playas para reducir ataques de tiburones.

En realidad, las altas temperaturas causan que más gente visite la playa, lo cual a su vez resulta en más ventas de helados y más ataques de tiburones.

Una regresión múltiple de ataques versus ventas de helados y temperatura, revelará que, como nos dice la intuición, el primer predictor ya no es significativo luego del ajuste que incluye la temperatura.

El escenario ideal es cuando los predictores no están correlacionados

- Cada coeficiente puede ser estimado y testeado por separado.
- Son posibles las interpretaciones como “un cambio en una unidad en Xj está asociado con un cambio de Bj en Y , mientras todas las demás variables se mantienen constantes

Como ya mencionamos, las correlaciones entre los predictores (multicolinealidad) causan problemas

- La varianza de todos los coeficientes tiende a aumentar, a veces dramáticamente.
- Las interpretaciones se vuelven riesgosas, cuando cambia Xj todo lo demás también cambia.

Las afirmaciones de causalidad deben ser evitadas para datos observacionales.

Dos de los supuestos básicos determinados por la forma del modelo lineal es que la relación entre predictor y target es:

- **aditiva:** el efecto del cambio de X_j sobre Y es independiente ***de los valores del resto de los predictores***
- **lineal:** el efecto del cambio en una unidad de X_j sobre Y es independiente ***de los valores de*** X_j

Veremos muchos modelos a lo largo del curso que relajan estos supuestos. Veamos ahora dos formas de “relajarlos” dentro del marco de una regresión lineal.

En el caso anterior, concluímos que TV y radio influian sobre las ventas.

El modelo lineal asumía que el efecto de pautar en TV sobre las ventas era independiente del efecto del gasto en otros medios.

¿Y si esto no fuera así? Si el gasto en TV se incrementa, también el efecto del gasto en radio... en ese caso, quizás tendría más sentido repartir el gasto entre ambos que alocarlo todo a un solo medio...

En marketing se llama a esto “efecto sinergia”... en estadística... **efecto interacción**,

Una forma de modelar el efecto interacción es incluir un tercer predictor en el modelo: el producto de los predictores. Por ejemplo, $X_1 * X_2$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \epsilon.$$

¿Cómo afecta esto al modelo? Relajando el supuesto de aditividad. Podemos reescribir el modelo de la siguiente forma:

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \epsilon \\ &= \beta_0 + \tilde{\beta}_1 X_1 + \beta_2 X_2 + \epsilon \end{aligned}$$

Ahora X_1 está afectada por los valores de X_2 . De esta forma, el efecto de X_2 sobre Y no es constante respecto de X_2. “Mover” X_2 hace que el efecto de X_1 sobre Y cambie.

Volvamos al caso de estudio. Incluyendo un término de interacción el modelo quedaría

$$\begin{aligned} \texttt{sales} &= \beta_0 + \beta_1 \times \texttt{TV} + \beta_2 \times \texttt{radio} + \beta_3 \times (\texttt{radio} \times \texttt{TV}) + \epsilon \\ &= \beta_0 + (\beta_1 + \beta_3 \times \texttt{radio}) \times \texttt{TV} + \beta_2 \times \texttt{radio} + \epsilon. \end{aligned}$$

Podemos interpretar $\beta 3$ como el incremento en la efectividad de la pauta en TV por cada unidad de incremento de la pauta en radio.

	Coefficient	Std. error	t-statistic	p-value
Intercept	6.7502	0.248	27.23	< 0.0001
TV	0.0191	0.002	12.70	< 0.0001
radio	0.0289	0.009	3.24	0.0014
TV×radio	0.0011	0.000	20.73	< 0.0001

- Al ver los resultados del modelo con efecto de interacción aparece que el modelo con interacción es “superior” al que solo contiene los efectos principales (el simple).
- El p-value para el término de interacción es bastante bajo, lo cual sugiere que hay evidencia para asumir que β3 es diferente a cero.
- A su vez, el R2 es 96.8 %, comparado con el 89.7% del modelo solo con efectos principales.

En el ejemplo anterior todos los p-valores eran significativos.

Puede suceder que los efectos principales (los que no tienen interacción) no sean significativos, mientras que los términos de interacción sí lo sean.

El principio jerárquico plantea que si incluimos efectos de interacción en un modelo de regresión lineal, también debemos incluir los efectos principales… aún si estos últimos no son significativos….

Dos citas famosas...

- **“Essentially, all models are wrong, but some are useful” George Box**

- **“The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively” Fred Mosteller and John Tukey, paraphrasing George Box**

Práctica Guiada