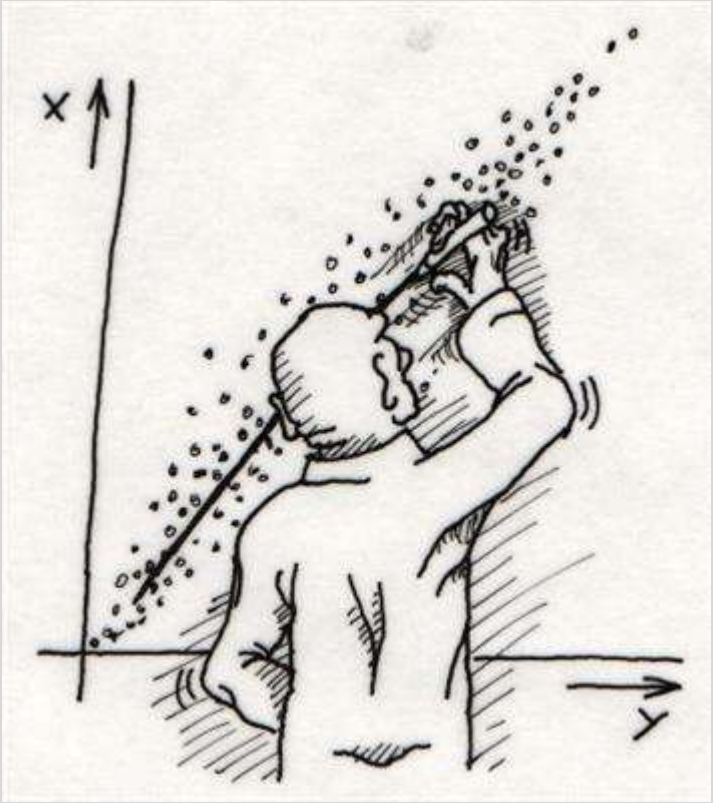


DigitalHouse > Coding School

DATA SCIENCE

MÓDULO 3

Regresión Lineal

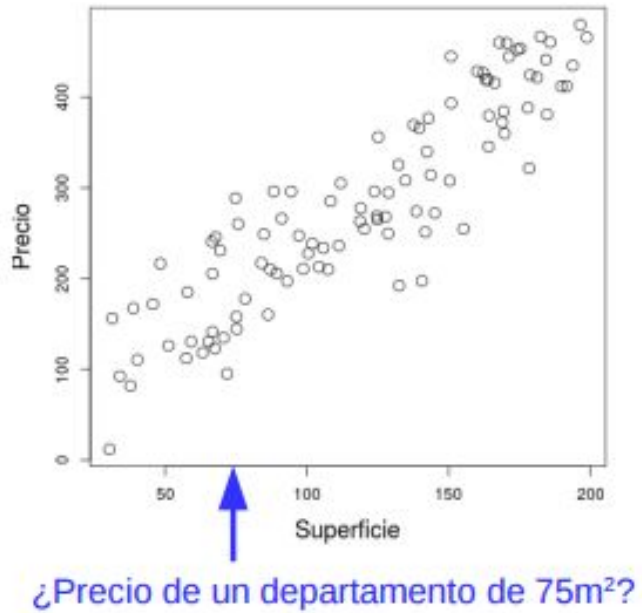
Introducimos la **regresión lineal**, una aproximación muy simple para **aprendizaje supervisado**.

En particular, la regresión lineal es una herramienta útil para predecir una **respuesta cuantitativa**.

Predecir una cantidad:

- Tiempo de demora de un vuelo
- El precio de una propiedad

¿Puede usarse para predecir variables cualitativas, quizá calcular la probabilidad de pertenercer a cierta clase?



Agenda de esta clase

Veremos algunas de las ideas claves que soportan a los modelos de regresión lineal, así como la estimación mediante mínimos cuadrados

1. **Regresión lineal simple**
2. **Estimación de coeficientes**
3. **Evaluación del Modelo**
4. **Regresión Múltiple**
5. **Variables Dummy**

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 (en miles de unidades), y el **presupuesto dedicado en publicidad** en 3 medios: TV, radio y diario (en miles de USD)

Si logramos identificar una **relación entre la inversión en publicidad y las ventas**, podremos recomendarle a nuestro cliente hacia dónde debe dirigir su inversión en publicidad.

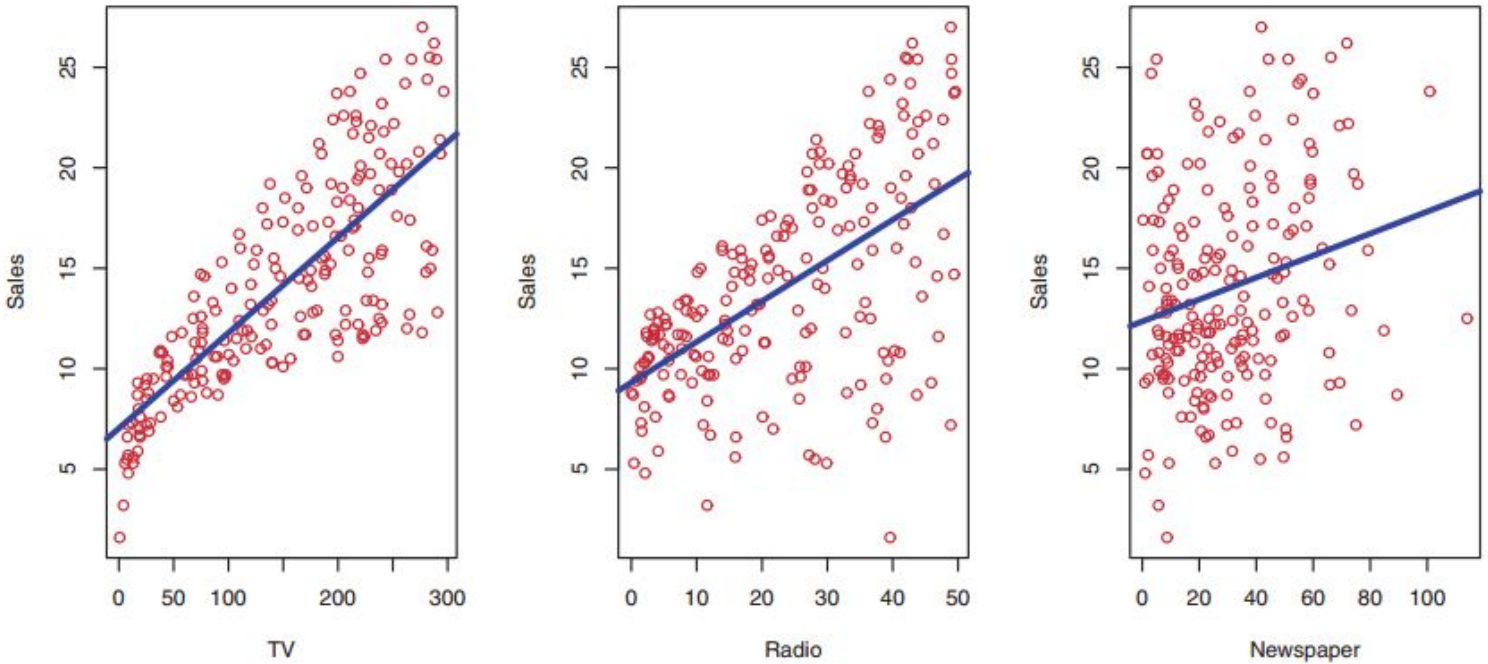
```
1 advertising.head()
```

	TV	radio	newspaper	sales
0	230.1	37.8	69.2	22.1
1	44.5	39.3	45.1	10.4
2	17.2	45.9	69.3	9.3
3	151.5	41.3	58.5	18.5
4	180.8	10.8	58.4	12.9

```
1 advertising.shape
```

```
(200, 4)
```

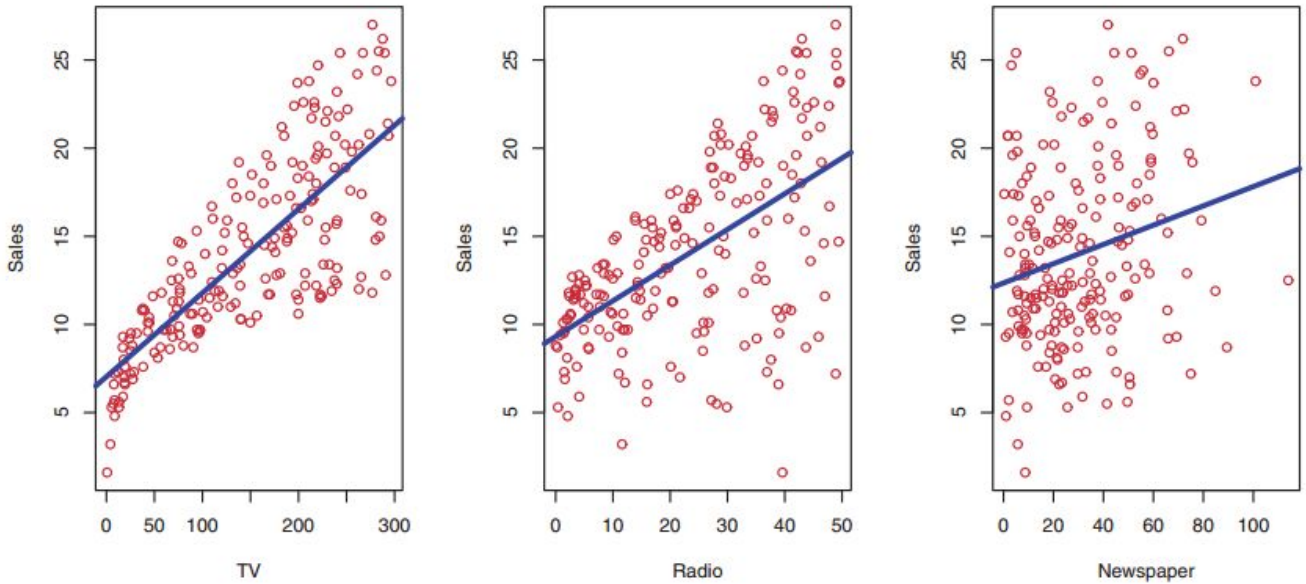
Así se ve una primera visualización del dataset.



Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

Pensemos en estos datos. Algunas preguntas que podrían surgir:

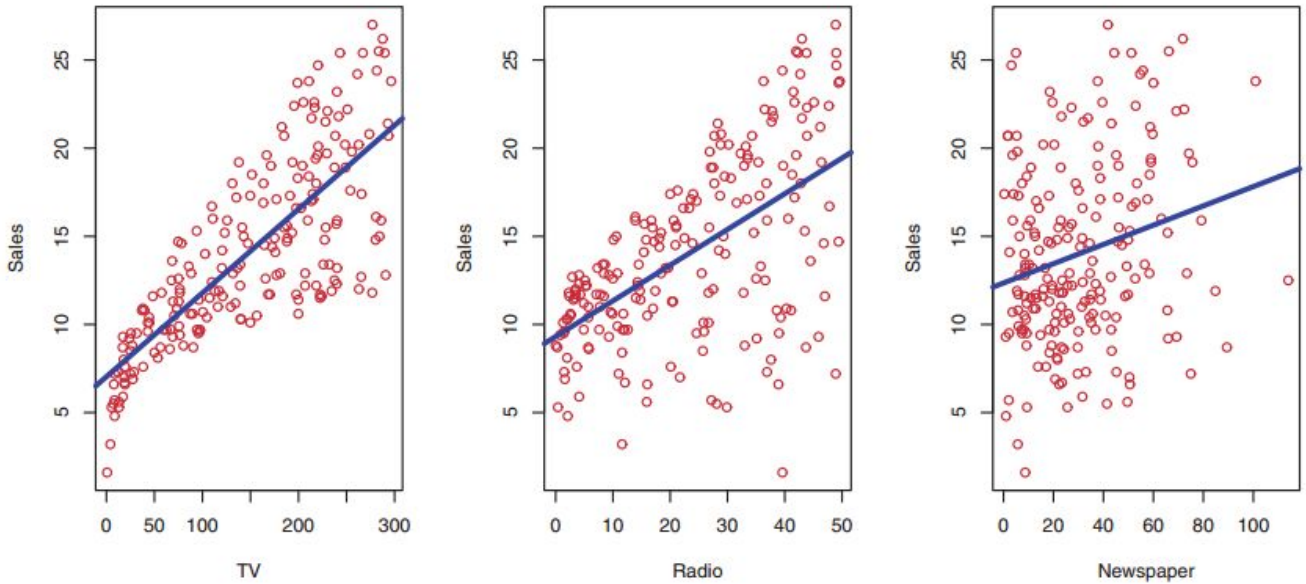
- ¿Hay alguna **relación** entre el presupuesto en publicidad y las ventas?



Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

Pensemos en estos datos. Algunas preguntas que podrían surgir:

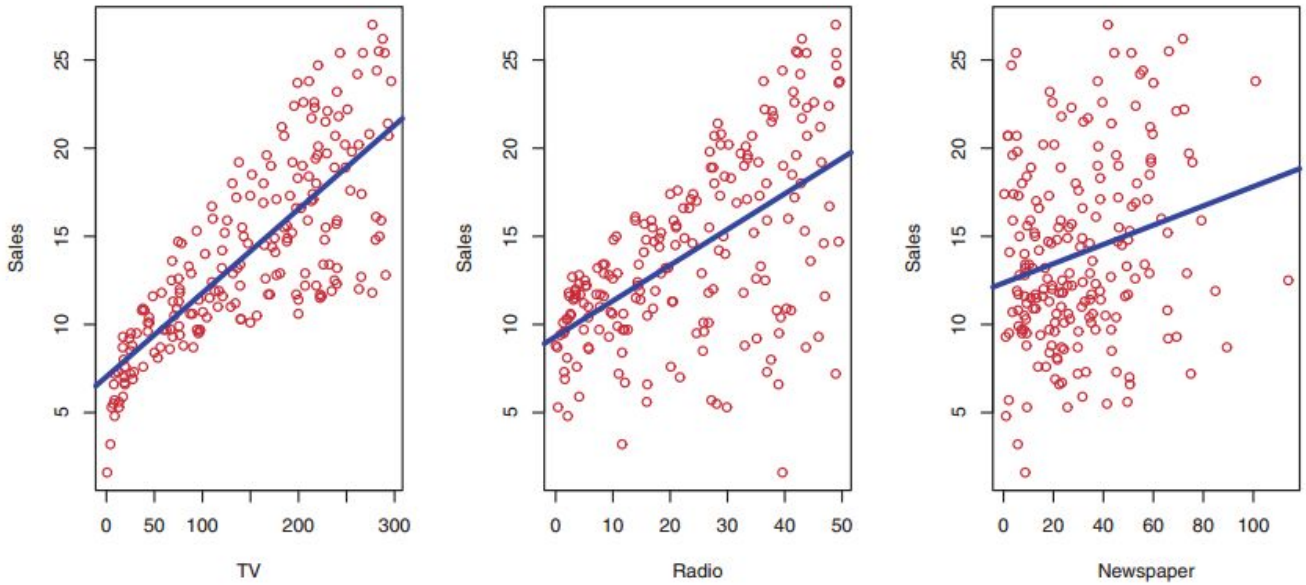
- ¿**Qué tan fuerte** es esa relación?



Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

Pensemos en estos datos. Algunas preguntas que podrían surgir:

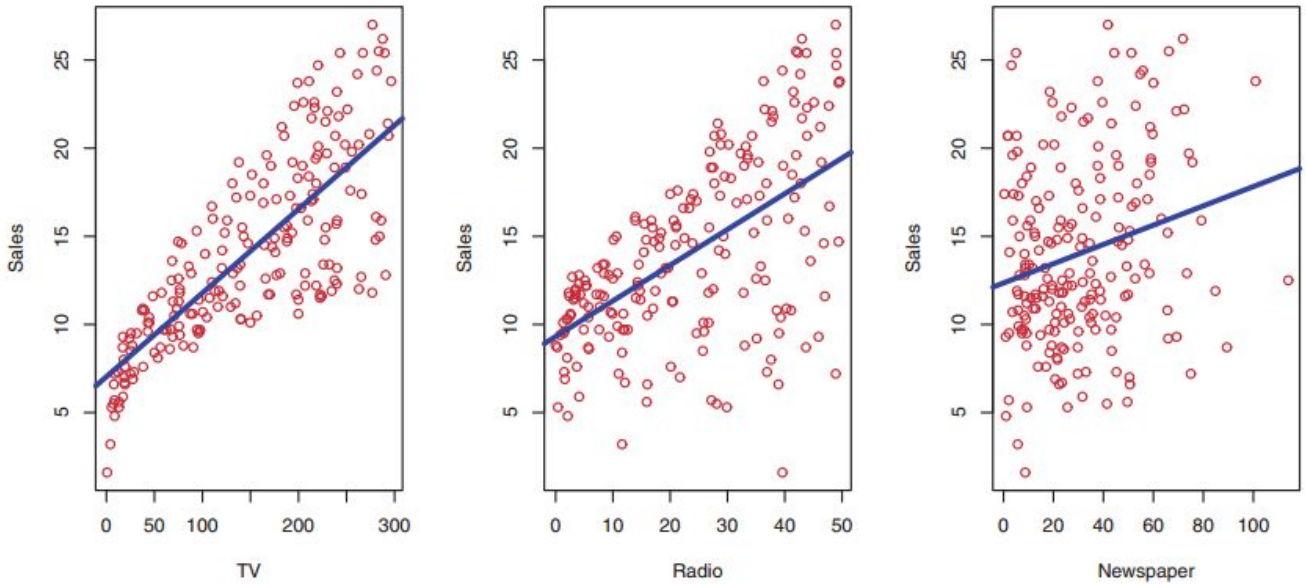
- ¿**Cuáles** de los medios mencionados contribuyen a las ventas?



Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

Pensemos en estos datos. Algunas preguntas que podrían surgir:

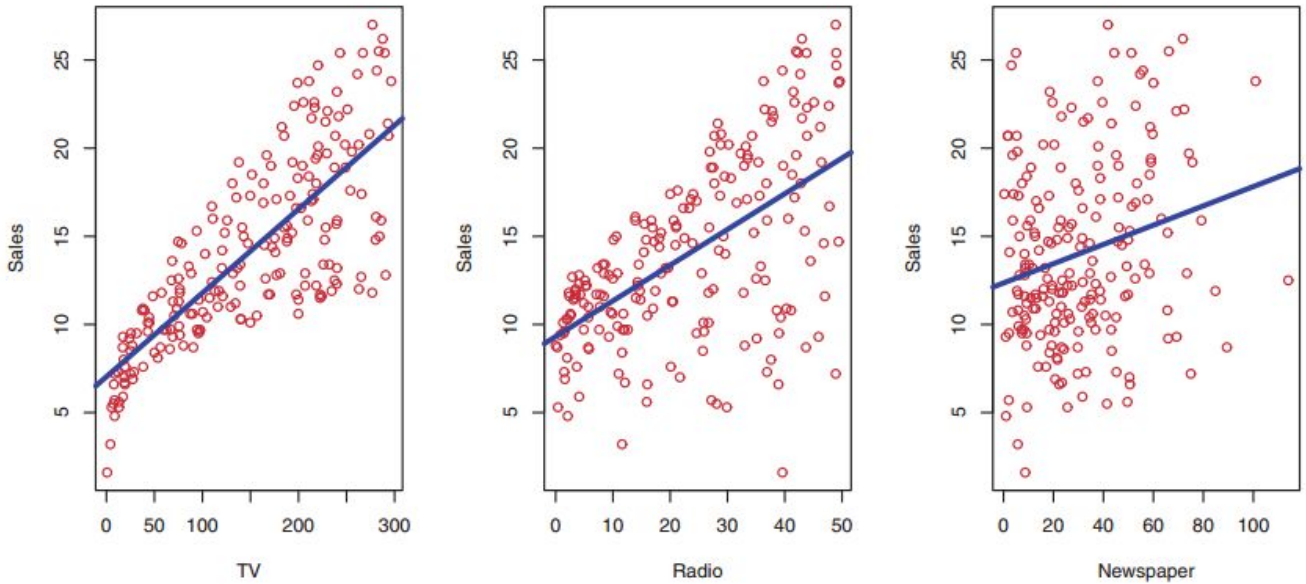
- ¿Con cuánta **precisión** podemos predecir las ventas futuras?



Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

Pensemos en estos datos. Algunas preguntas que podrían surgir:

- ¿Es esta **relación lineal**?



Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

Regresión Lineal Simple

La **regresión lineal simple** intenta predecir una respuesta cuantitativa **Y** en base a una única variable predictora **X.**

Asume que hay aproximadamente una relación lineal entre X e Y. Matemáticamente:

$$Y \approx \beta_0 + \beta_1 X.$$

Podemos leer esta expresión como "se modela aproximadamente como".

Por ejemplo, X puede representar el presupuesto en publicidad en **TV** e Y las ventas (**sales**)

$$\texttt{sales} \approx \beta_0 + \beta_1 \times \texttt{TV}.$$

β_0 y β_1 son dos constantes que representan el intercepto y la pendiente en el modelo lineal.

Juntos, β_0 y β_1 son conocidos como los **parámetros** del modelo.

Una vez que hemos usado nuestro set de entrenamiento para producir los estimadores $\hat{\beta}_1$ y $\hat{\beta}_0$ para los coeficientes del modelo, podemos predecir futuras ventas en base a un valor particular de **TV.**

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x,$$

donde $\hat{y}$ indica una predicción de Y en base a X.

Aquí usamos un símbolo ^ para denotar el valor estimado para un parámetro o coeficiente desconocido, o para denotar el valor predicho de la respuesta.

Entonces:

- Consiste en predecir una respuesta cuantitativa Y en base a una única variable predictora X.

$$Y \approx \beta_0 + \beta_1 \cdot X$$

Ejemplo:

$$Precio \approx \beta_0 + \beta_1 \cdot Superficie$$

Ordenada al origen (*intercept*)

Pendiente (*slope*)

- β_0 y β_1 son los coeficientes desconocidos que vamos a estimar, o ajustar en base a los datos de entrenamiento. Una vez estimados, los podemos usar para predecir:

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 \cdot x$$

Valor predicho para Y cuando X=x

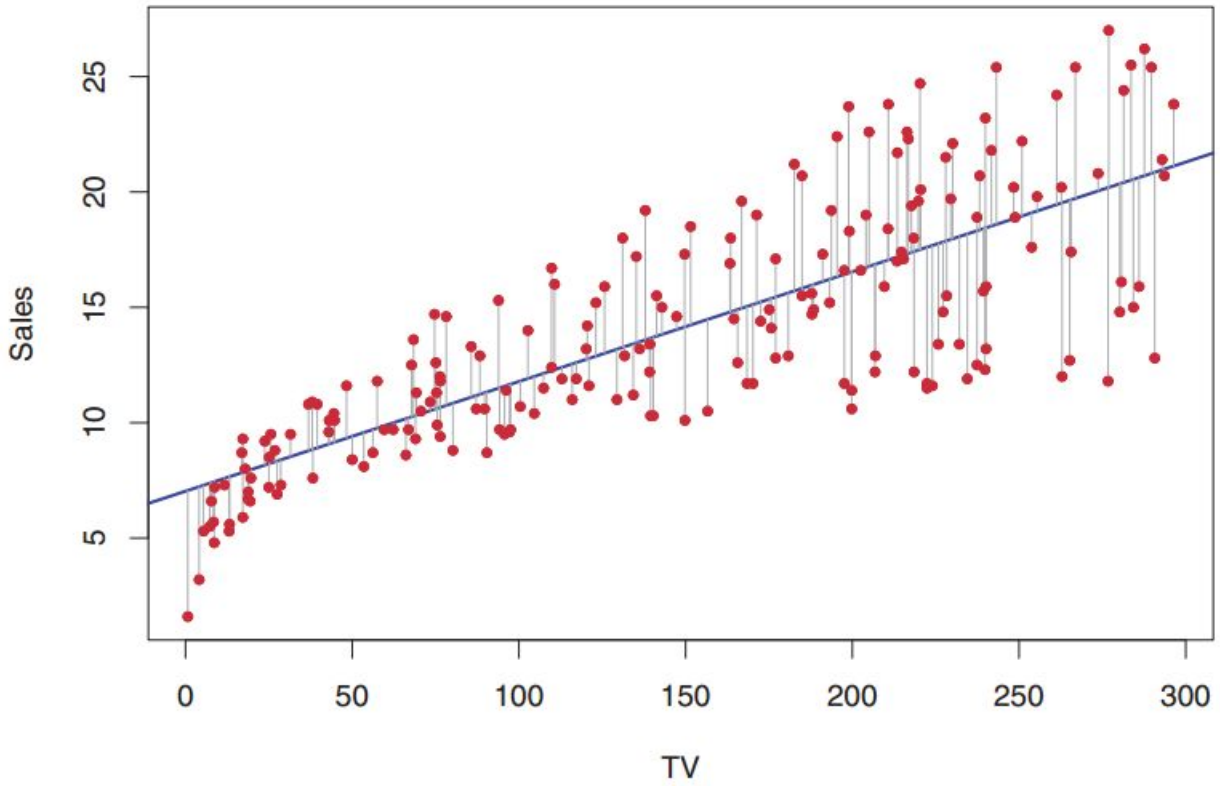
Estimación de β_0

Estimación de β_1

Nueva instancia

Estimación de los coeficientes

Ejemplo de ajuste por mínimos cuadrados, para el dataset Advertising



Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

- Definición: Residuo o error de predicción

$$e_i = y_i - \hat{y}_i$$

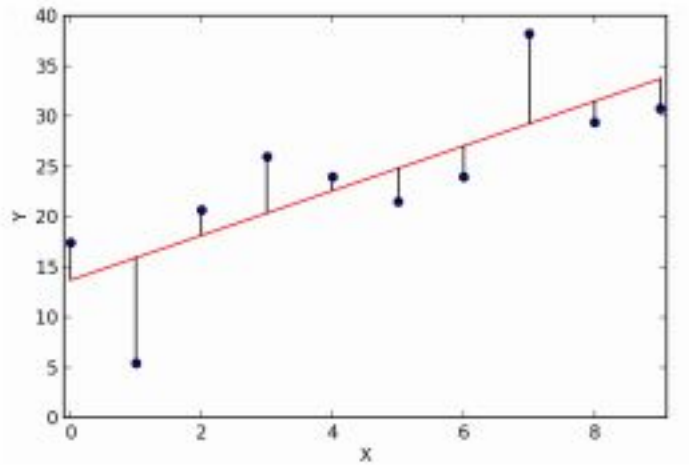
- Residual sum of squares:

$$\mathrm{RSS} = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 \cdot x_i\right)^2$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x},$$

Donde $\bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$ y $\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i$



La figura muestra el ajuste de una regresión lineal simple al dataset **Advertising,** donde:

$$\hat{\beta}_0 = 7.03 \text{ y } \hat{\beta}_1 = 0.0475$$

En otras palabras, de acuerdo a esta aproximación, un adicional de $1000 gastados en publicidad en TV está asociado con vender aproximadamente 47.5 unidades adicionales del producto. ¿ En qué unidades están expresados los coeficientes del modelo?

Evaluación del Modelo

- TSS (Total Sum of Squares): Variabilidad total de los datos

$$\text{TSS} = \sum_{i=1}^{n} \left(y_i - \overline{y}\right)^2$$

- RSS: Variabilidad no explicada por el modelo

$$\text{RSS} = \sum_{i=1}^{n} \left(y_i - \hat{y}_i\right)^2$$

- R^2: Proporción de la variabilidad explicada por el modelo

$$R^2 = \frac{\text{TSS} - \text{RSS}}{\text{TSS}}$$

$R^2 \rightarrow 0$ cuando el modelo explica poco de la variabilidad de los datos.

$R^2 \rightarrow 1$ cuando el modelo explica mucho de la variabilidad de los datos.

- Un estadístico **R^2** cercano a 1 indica que una gran proporción de la variabilidad en la respuesta ha sido explicada por la regresión.

- Un **R^2** cercano a 0 indica que la regresión no explicó mucha de la variabilidad en la respuesta; esto podría ocurrir porque el modelo lineal está mal, o porque el error inherente σ^2 es alto, or ambas.

- Por ejemplo, en la regresión del **ejemplo anterior, el R^2 fue 0.61,** y por lo tanto menos de dos tercios de la variabilidad en **sales** está explicada por una regresión lineal sobre TV.

Precisión de los coeficientes estimados

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

$$SE(\beta_0) \quad SE(\beta_1)$$

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

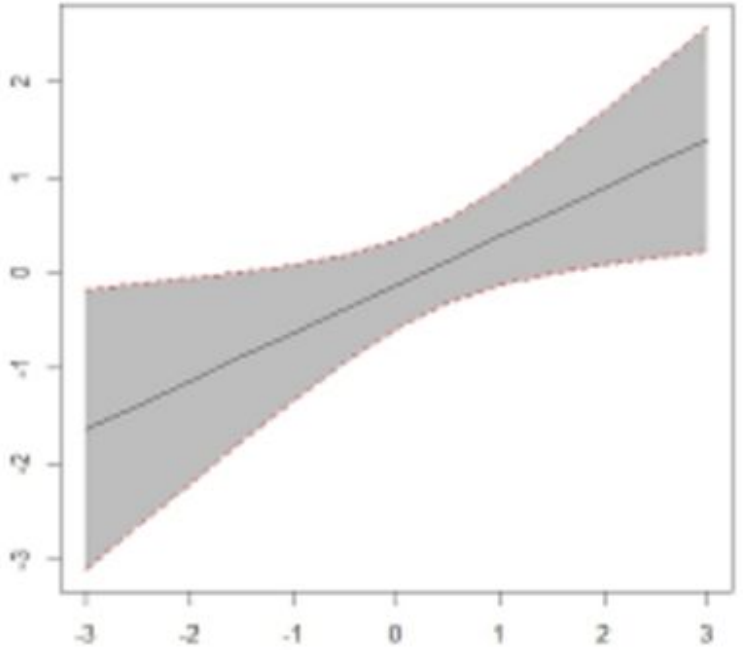
Significatividad de los coeficientes estimados

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.

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

where s_y is the standard deviation of the residuals, calculated as

$$s_y = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}.$$



$$Y = \beta_0 + \beta_1 X + \epsilon.$$

$$\mathrm{SE}(\hat{\beta}_1)^2 = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2},$$

$$\text{where } \sigma^2 = \mathrm{Var}(\epsilon)$$

Significatividad de los coeficientes estimados

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

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

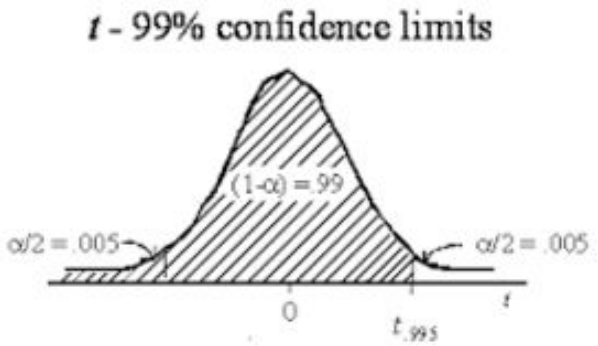
Pero para simplificar la notación, **no utilizaremos el sombrero extra en nuestras presentaciones.**

Significatividad de los coeficientes estimados

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

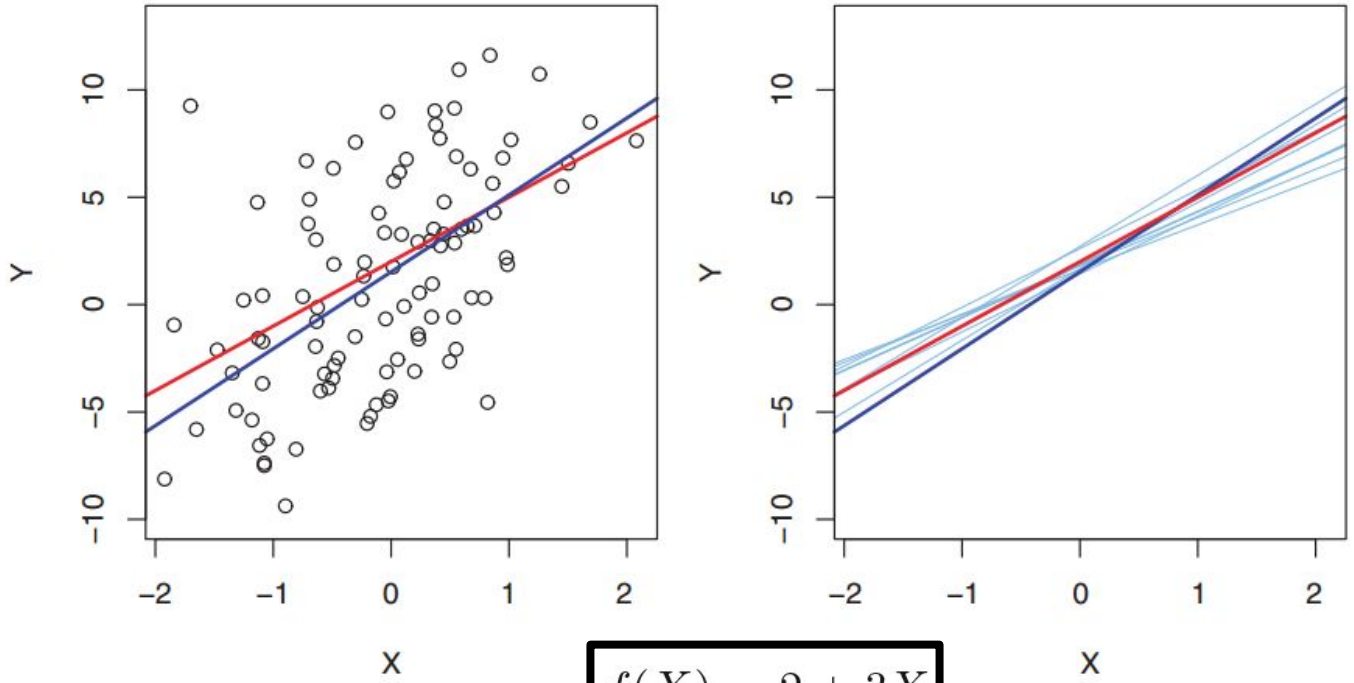
¿Qué probabilidad existe de observar lo que observamos si el verdadero valor del coeficiente es cero?

- *p-valor bajo* (p<0.05 o p<0.01): Es improbable observar un coeficiente de la magnitud del estimado en la muestra si el verdadero valor del coeficiente es 0 en la población.
- *p-valor alto* : Es probable que la asociación observada sea producto del azar.



Función de Regresión Poblacional y Muestral

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



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

Gareth James • Daniela Witten • Trevor Hastie
Robert Tibshirani (2017)

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.**

Expresión para un intervalo de confianza de aproximadamente 95% para β_1 :

$$\left[\hat{\beta}_1 - 2 \cdot \mathrm{SE}(\hat{\beta}_1),\ \hat{\beta}_1 + 2 \cdot \mathrm{SE}(\hat{\beta}_1)\right]$$

En el caso del dataset **advertising,** el intervalo de confianza de aproximadamente 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_o: 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 evidencia 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 si $\hat{\beta}_1$ (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 entre otras cosas del tamaño muestral n.

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 del cero.

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

con *n-2* grados de libertad. En general, cuando tengamos más variables, los grados de libertad van a ser *n-p-1*, es decir la cantidad de observaciones menos los parámetros que estamos estimando).

Básicamente, es cuestión de calcular la probabilidad de observar cualquier valor igual a $|t|$ o mayor asumiendo que $\beta_1 = 0$, por lo que bajo H0 el estadístico de prueba es una variable con distribución T-Student con n-p-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

Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

Un incremento de $1000 en el presupuesto de publicidad en TV está asociado con un incremento de 47,5 unidades en las ventas (la variable dependiente está en miles de unidades y la de presupuesto en miles de U$S).

- 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 que lidiar con más de un predictor.
- Por ejemplo, en el caso de los datos de Advertising, examinamos la relación entre ventas y publicidad en **TV**.
- Pero también hay datos sobre otras variables: publicidad en **Diarios** y en **Radio**.

- En la práctica tenemos que lidiar con más de un predictor.
- Por ejemplo, en el caso de los datos de Advertising, examinamos la relación entre ventas y publicidad en **TV**.
- Pero también hay datos sobre otras variables: publicidad en **Diarios** y en **Radio**.

¿Cómo podemos extender el análisis para incorporar estos dos predictores nuevos?

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**.

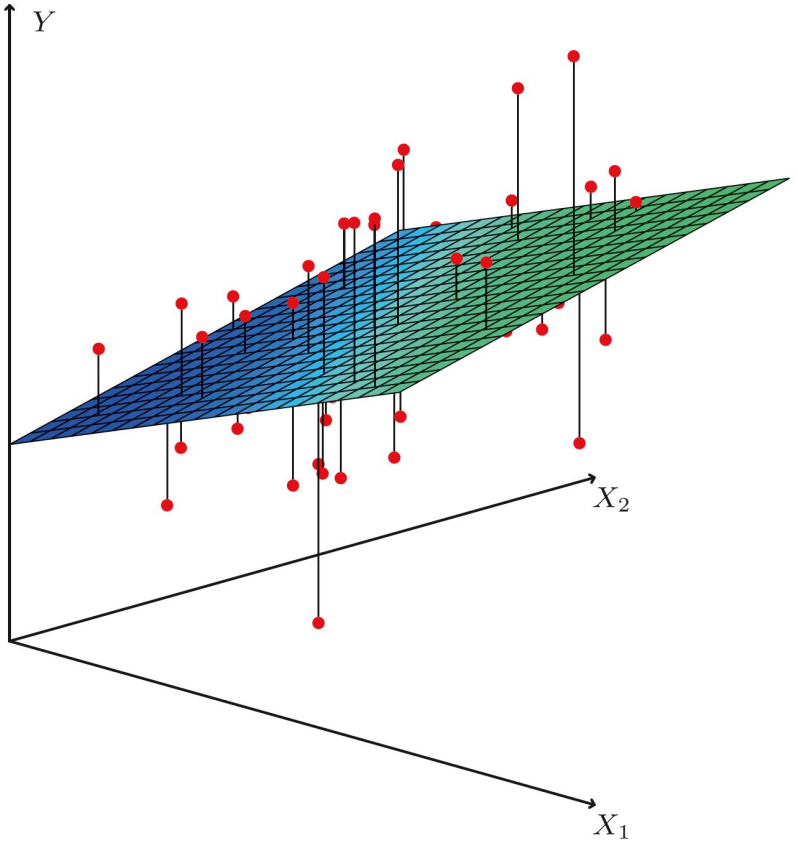
Podemos darle a cada predictor un coeficiente separado en un único modelo.

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

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

Ejemplo de regresión lineal con dos predictores y una respuesta.

La “línea” de regresión mínimo cuadrática se vuelve un plano. El plano es “buscado” para minimizar la suma de los cuadrados de las distancias entre cada observación y el plano.



Dados estimadores de los coeficientes de pendiente podemos pronosticar la variable de respuesta para una observación con valores dados de los predictores como:

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

Elegimos los valores para los estimadores de los coeficientes que minimizan la suma de residuos al cuadrado

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

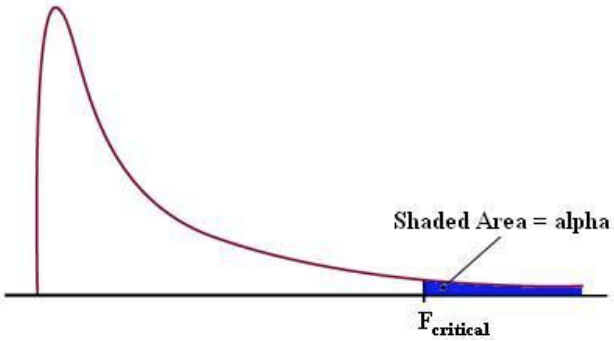
¿Hay alguna relación entre la variable objetivo y todas las variables explicativas?

$$H_0 : \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a : \text{ at least one } \beta_j \text{ is non-zero.}$$

Usamos el estadístico *F*.

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



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.*

Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

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

Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

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

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

En la regresión múltiple el estimador del coeficiente es cercano a cero y no es significativo, con un p-value cercano a 0.86.

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

De hecho, sí 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.*

Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

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 efecto de la variable **radio;** es decir **newspaper** recibe el "crédito" por los efectos de **radio** en **sales.**

El **escenario ideal** es cuando los **predictores no están correlacionados**

Las correlaciones entre los predictores causan problemas

- La varianza de los estimadores de los coeficientes tiende a aumentar, a veces dramáticamente.
- Entonces los estimadores de los coeficientes son menos precisos.

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

- **aditiva:** el efecto del cambio de X_j sobre Y es aditivo a los efectos de los otros predictores.
- **lineal:** el efecto del cambio en una unidad de X_j sobre Y es constante.

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 tienen influencia sobre las ventas.

Hasta aquí asumimos que el efecto de pautar en TV sobre las ventas no dependía del gasto en publicidad en otros medios.

¿Y si esto no fuera así? Si el gasto en TV incrementa, también el efecto del gasto en radio quizás tendría más sentido repartir el gasto entre ambos medios en lugar de asignar el presupuesto entero a uno solo.

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 el efecto de X_1 está afectado 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 (o al revés).

	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

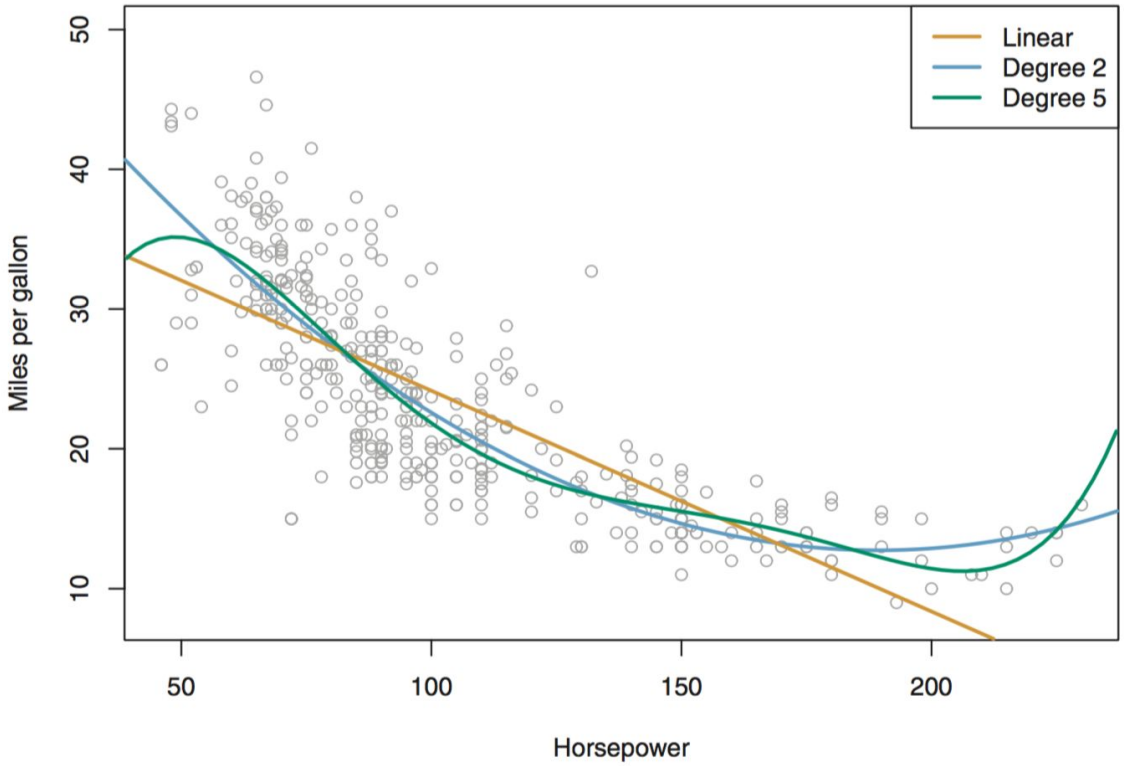
- Al ver los resultados del modelo con efecto de interacción aparece que el modelo con interacción es “superior” al que solo contiene los efectos principales.
- El p-value para el término de interacción es muy bajo, lo cual sugiere que hay evidencia para rechazar que β3 es igual a cero en la población..
- A su vez, el R2 es 96.8 %, comparado con el 89.7% del modelo solo con efectos principales.

Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

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

El **principio jerárquico** plantea que si incluimos interacciones en un modelo también debemos incluir los efectos principales, aún si los coeficientes de estos últimos son estadísticamente no significativos.

Extendiendo el modelo lineal: Relaciones no lineales



Podemos generar una variable nueva, elevando a *horsepower* al cuadrado:

$$\texttt{mpg} = \beta_0 + \beta_1 \times \texttt{horsepower} + \beta_2 \times \texttt{horsepower}^2 + \epsilon$$

Notar que el modelo sigue siendo lineal, porque es lineal en los parámetros.

	Coefficient	Std. error	t-statistic	p-value
Intercept	56.9001	1.8004	31.6	< 0.0001
horsepower	-0.4662	0.0311	-15.0	< 0.0001
horsepower^2	0.0012	0.0001	10.1	< 0.0001

Gareth James • Daniela Witten • Trevor Hastie Robert Tibshirani (2017)

Variables dummy

A veces nos encontraremos con problemas de regresión lineal que presentan predictores cualitativos (variables categóricas nominales u ordinales).

Por ejemplo, el dataset **Credit Cards,** donde la variable dependiente es el saldo de deuda, presenta algunos predictores con estas características:

- gender
- student (condición de estudiante)
- status (estado civil)
- ethnicity (caucásico, afroamericano, etc.)

Problema: estimar diferencias entre el saldo de la tarjeta de crédito entre hombres y mujeres (ignorando el resto de las variables).

Regresión simple con predictor cualitativo

Variable dummy (xi) tal que $$x_i = \begin{cases} 1 & \text{if } i\text{th person is female} \\ 0 & \text{if } i\text{th person is male} \end{cases}$$

Así, el modelo de regresión toma la siguiente forma

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i = \begin{cases} \beta_0 + \beta_1 + \epsilon_i & \text{if } i\text{th person is female} \\ \beta_0 + \epsilon_i & \text{if } i\text{th person is male.} \end{cases}$$

Resultados del modelo anterior...

	Coefficient	Std. Error	t-statistic	p-value
`Intercept`	509.80	33.13	15.389	< 0.0001
`gender[Female]`	19.73	46.05	0.429	0.6690

¿Cómo interpretamos esto?

β_0 puede ser interpretado como el saldo promedio de los hombres
$\beta_0 + \beta_1$ expresa el saldo promedio de las mujeres
β_1 expresa la diferencia media en el saldo entre ambos grupos

En la tabla anterior, el crédito medio entre hombres se estimó en $509,80; el de las mujeres, en cambio, se estimó en $509,80 + $19,73 = $529,53. Es decir, que hay una diferencia de $19,73.

Sin embargo, notar que el p-value del coeficiente estimado de la variable dummy es demasiado elevado. Esto indica que no parece haber evidencia de una diferencia significativa en el saldo promedio de crédito entre sexos.

La decisión de codificación es arbitraria, en caso de haberlo codificado al revés. solamente hubiesen cambiado los valores de los coeficientes β_0 hubiese sido $529,53 y β_1 = - $19,73 (negativo)

Práctica Guiada

Es un método que tiene muchos años y está presente en toda la bibliografía.

Aunque parezca súper simple comparado con las técnicas modernas de machine learning, la regresión lineal aún es un método útil y ampliamente usado.

Principalmente, sirve como un **buen punto de partida para aproximaciones más nuevas**: muchas de las técnicas fancy pueden interpretarse como generalizaciones o extensiones de la regresión lineal.

Por lo tanto es super importante tener una buena compresión de la regresión lineal antes de estudiar los algoritmos más complejos de machine learning.

- TSS mide toda la varianza en la respuesta **Y**; puede pensarse como la cantidad de variabilidad inherente en la respuesta antes de que se realice la regresión.

- En contraste, RSS mide la cantidad de variabilidad que permanece no explicada luego de realizar la regresión.

- Por lo tanto, TSS-RSS mide la cantidad de variabilidad en la respuesta que es explicada (o removida) al realizar la regresión.

- Finalmente, el estadístico **R^2 mide la proporción de variabilidad en Y que puede explicarse usando X.**