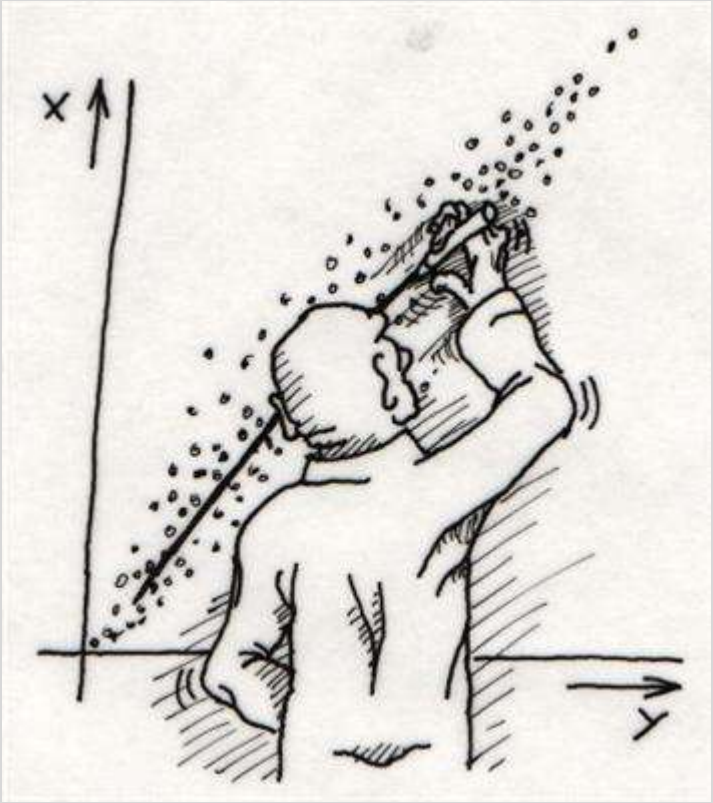


DigitalHouse > Coding School

DATA SCIENCE

MÓDULO 3

Intro a Regresión Lineal

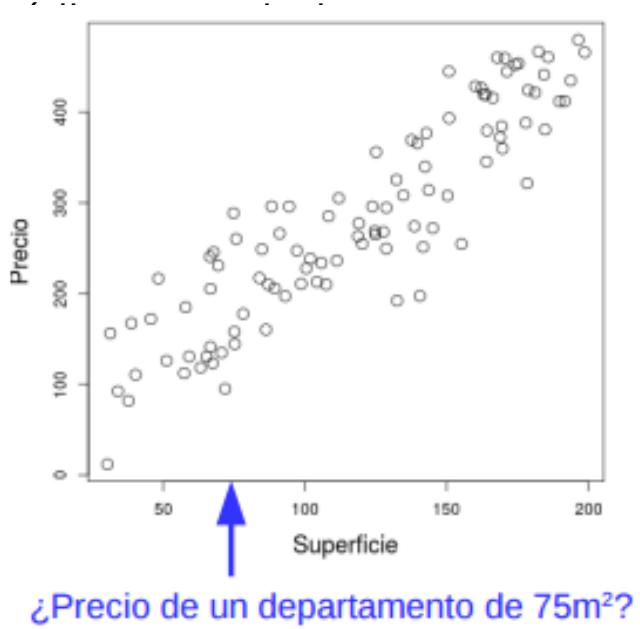
Septiembre de 2017

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

En particular, la regresión lineal es una herramient cuantitativa.

Predecir una cantidad:

- Tiempo de demora de un vuelo
- La probabilidad de que un mail sea SPAM
- El precio de una propiedad



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

Aunque parezca super 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.

OBJETIVOS DE ESTA CLASE

Veremos algunas de las ideas claves que soportan a los modelos de regresión lineal, así como la aproximación de mínimos cuadrados (least squares) que es la más frecuentemente usada para ajustar este modelo.

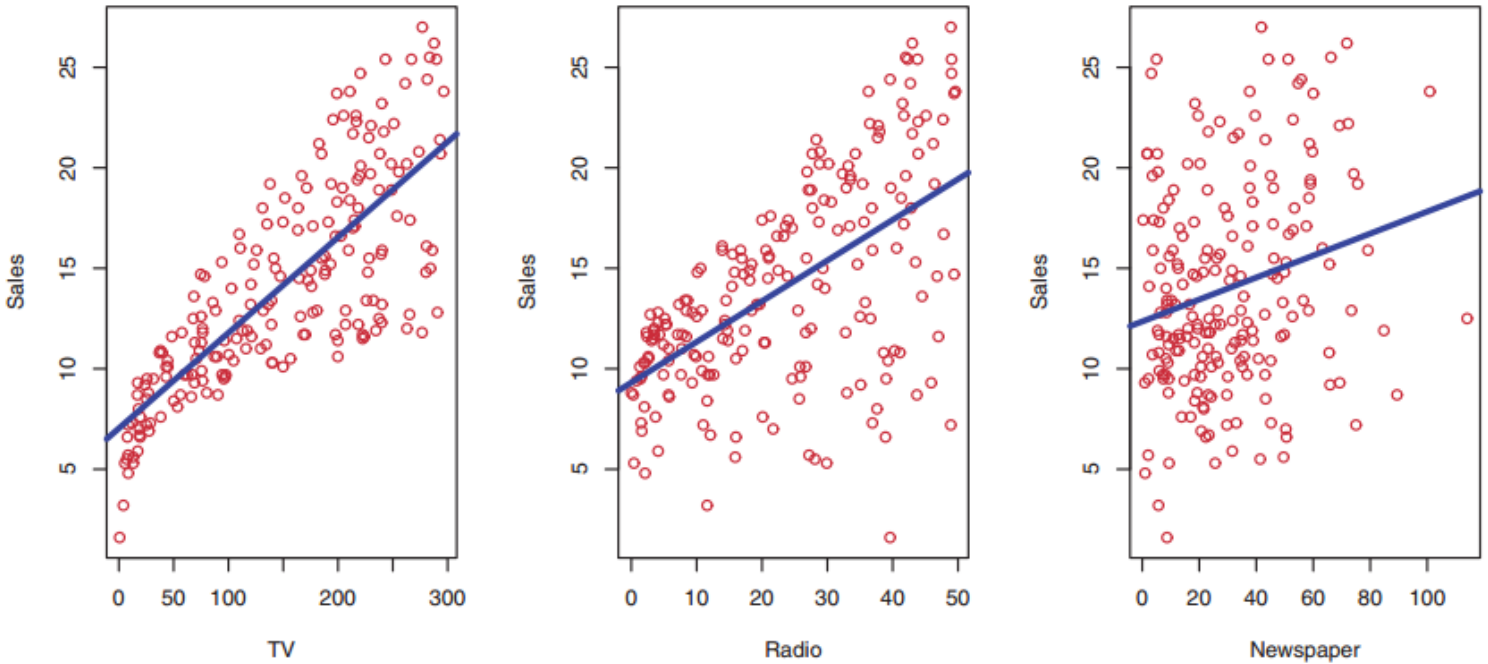
1. **Regresión lineal simple**
2. **Estimando coeficientes**
3. **Evaluación del Modelo**
4. **Introducción a Regresión Múltiple**

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 medios: 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 variable de respuesta será las ventas.

Así se ve una primera visualización del dataset.



Pensemos en estos datos. Algunas preguntas que podrían surgir:

- ¿Hay alguna **relación** entre el presupuesto en publicidad y las ventas?
- ¿**Qué tan fuerte** es esa relación?
- ¿**Cuáles** de los medios mencionados contribuyen a las ventas?
- ¿Con cuánta **precisión** podemos predecir las ventas futuras?
- ¿Es esta **relación lineal**?

Resulta que la regresión lineal puede ser usada para responder cada una de estas preguntas y algunas más.
Veamos algunos conceptos y luego intentaremos responderlas.

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** o coeficientes del modelo.

Una vez que hemos usado nuestro set de entrenamiento para producir los estimaciones $\hat{\beta}_0$ y $\hat{\beta}_1$ 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 = 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

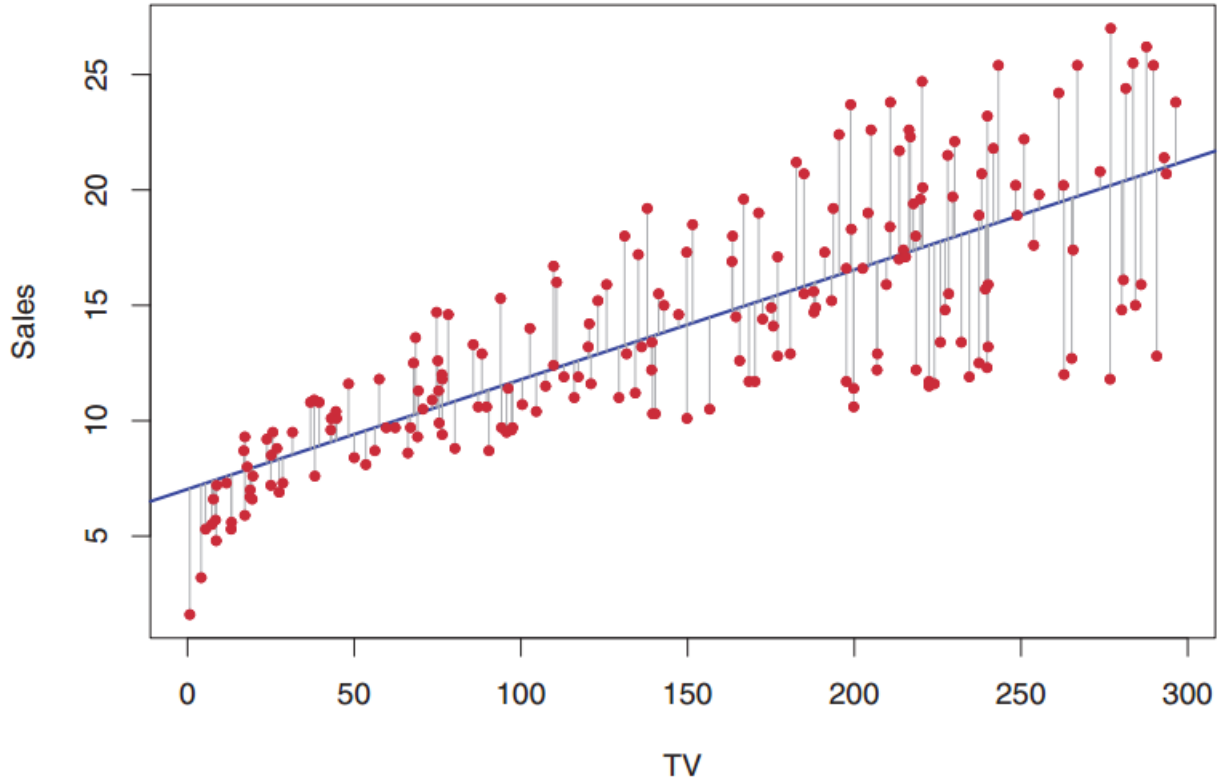
Estimación de β_0

Estimación de β_1

Nueva instancia

Estimando los coeficientes

Ejemplo de ajuste por mínimos cuadrados, para el dataset Advertising



Para hacer predicciones, debemos usar los datos para **estimar los coeficientes β_0 y β_1** para que la línea resultante esté tan *cerca* como sea posible a los puntos de entrenamiento.

Hay varias formas de medir ***cercanía***, pero la aproximación más común se relaciona con el **criterio de mínimos cuadrados**

En su forma más simple, intenta **minimizar la suma de cuadrados** de las diferencias en las ordenadas (llamadas **residuos**) entre los puntos generados por la función elegida y los correspondientes valores en los datos.

Ejemplo de ajuste por mínimos cuadrados, para el dataset Advertising

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.

Explicación de la figura

Para el dataset **Advertising,** la figura anterior muestra el ajuste por mínimos cuadrados para la regresión de **sales** contra **TV.**

El ajuste se encuentra minimizando la suma de los cuadrados de los errores.

Cada segmento de línea gris representa un error y el ajuste toma una decisión de compromiso promediando sus cuadrados.

En este caso el ajuste lineal captura la esencia de la relación pero ¿qué pasa con los residuos a la izquierda y a la derecha del plot?

Si $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ es la predicción de Y basada en el i-ésimo valor de X, entonces, $e_i = y_i - \hat{y}_i$ representa el i-ésimo residuo (es decir, la diferencia entre el i-ésimo valor observado y la i-ésima predicción basada en el modelo lineal). Así, podemos definir la "suma de los residuos al cuadrado" (RSS, por sus siglas en inglés) de la siguiente forma:

$$\text{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$

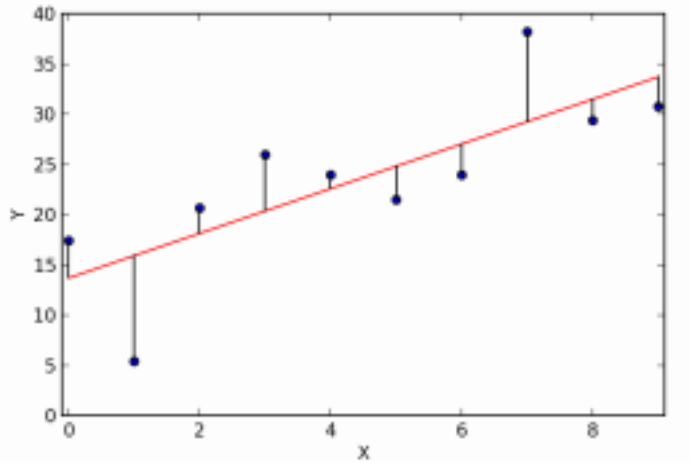
- Definición: Residuo o error de predicción

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

- Residual sum of squ

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

$$\mathrm{RSS} = e_1^2 + e_2^2 + \cdots + e_n^2,$$



- Los residuos se elevan al cuadrado para sacar el signo y para que RSS sea diferenciable.
- Hay que tener **cuidado con los outliers** en los datos, porque RSS penaliza los residuos grandes

- Estimados los coeficientes, los podemos usar para predecir:

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

- En nuestro ejemplo, podríamos predecir el precio de un departamento de 75m^2

$$\widehat{Precio} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 75$$

Precisión de los coeficientes estimados

- 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.

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

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.**

La próxima clase entraremos más en detalle en la estadística de los estimadores de coeficientes

Evaluación del Modelo

- Queremos cuantificar **hasta qué punto el modelo se ajusta a los datos**

- Típicamente, la calidad de un ajuste lineal se mide usando dos magnitudes relacionadas: el error estándar de los residuos (RSE, **Residual Standard Error**) y el estadístico R^2.

- El RSE se define como

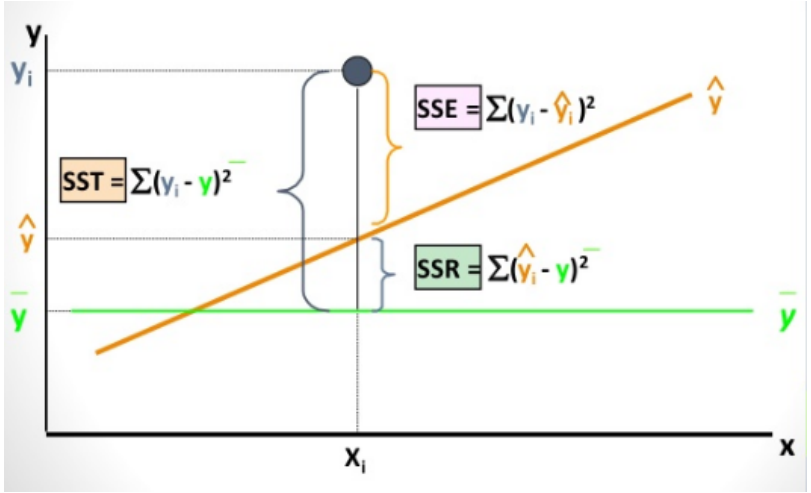
$$\mathrm{RSE} = \sqrt{\frac{1}{n-2}\mathrm{RSS}} = \sqrt{\frac{1}{n-2}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

- Hoy veremos **R^2**

- El **coeficiente de determinación**, denominado **R²** y pronunciado **R cuadrado** provee una medida de ajuste para la regresión.

- Toma la forma de una proporción (la proporción de varianza explicada) y por lo tanto siempre se encontrará entre los valores 0 y 1. (siempre y cuando el modelo incluya intercepto). Si no se incluye intercepto el **R² podría ser negativo. También puede ser negativo al ser usado para evaluar otros modelos (no regresión lineal).**

- Es independiente de la escala de Y, esto es bueno.



- RSS: Variabilidad no explicada por el modelo

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

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

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

- 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.

- 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.**

- 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.

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.
- Una pregunta relevante podría ser, ¿está alguno de estos dos medios asociado a las ventas?
- ¿Cómo podemos extender el análisis para incorporar estos dos predictores nuevos?

Un enfoque es extender el modelo de regresión simple para que pueda incorporar múltiples predictores.

Así, tenemos una pendiente separada para cada coeficiente en un solo modelo. En general, si tenemos p predictores, el modelo de regresión múltiple toma la forma:

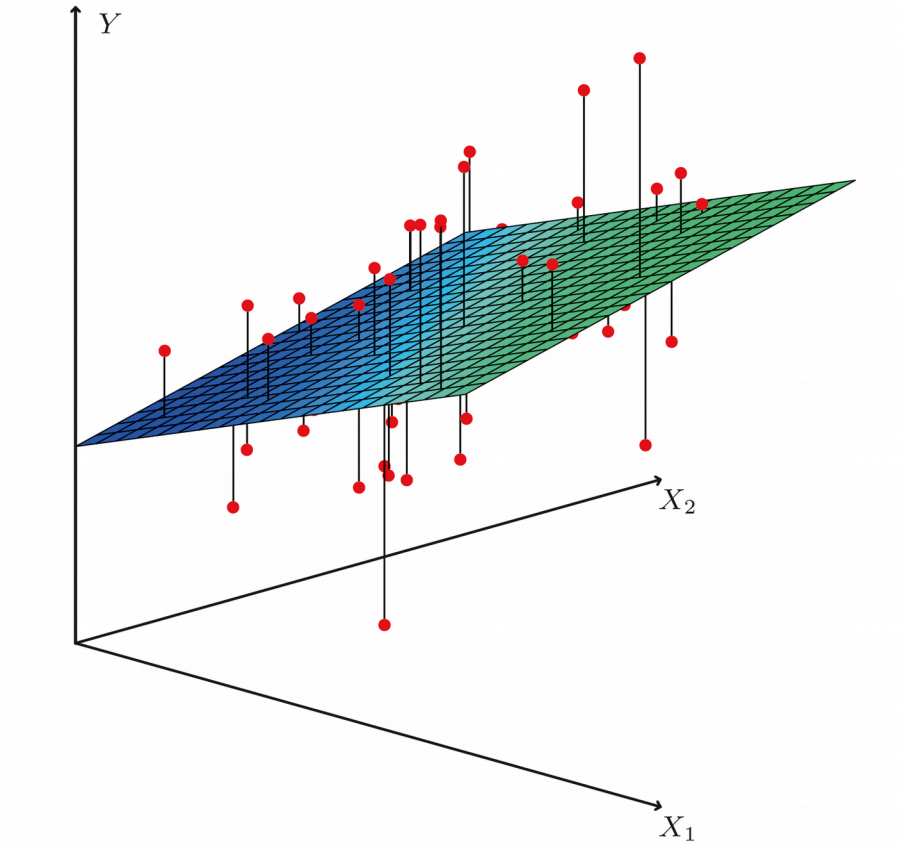
$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon,$$

donde X_j representa el j-ésimo predictor

β_j cuantifica la asociación entre la variable y la respuesta

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.



En nuestro ejemplo, el modelo toma la forma

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

Interpretamos los β_j como el “efecto promedio” del incremento de una unidad de X_j sobre Y, **manteniendo todos los demás factores constantes**.

- Consiste en predecir una respuesta cuantitativa Y en base a una de múltiples variable predictoras X_1, X_2, ... X_p

$$Y \approx \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

- RSS se define igual que para la regresión lineal simple

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

- Los coeficientes se estiman en forma análoga
- TSS, R^2 también se definen en forma similar.
- ¿Qué pasa si hay variables categóricas?

Vamos a ver las respuestas a estas preguntas y mucho más background técnico de regresiones múltiples en la próxima clase.

Por ahora, vamos a practicar un poco...

Práctica Guiada