ESCUELA DE INGENIERÍA
FACULTAD DE INGENIERÍA

EDUCACIÓN PROFESIONAL

Programación en R para Ciencia de Datos

Miguel Jorquera

DBDC-202010

Educación Profesional

Escuela de Ingeniería

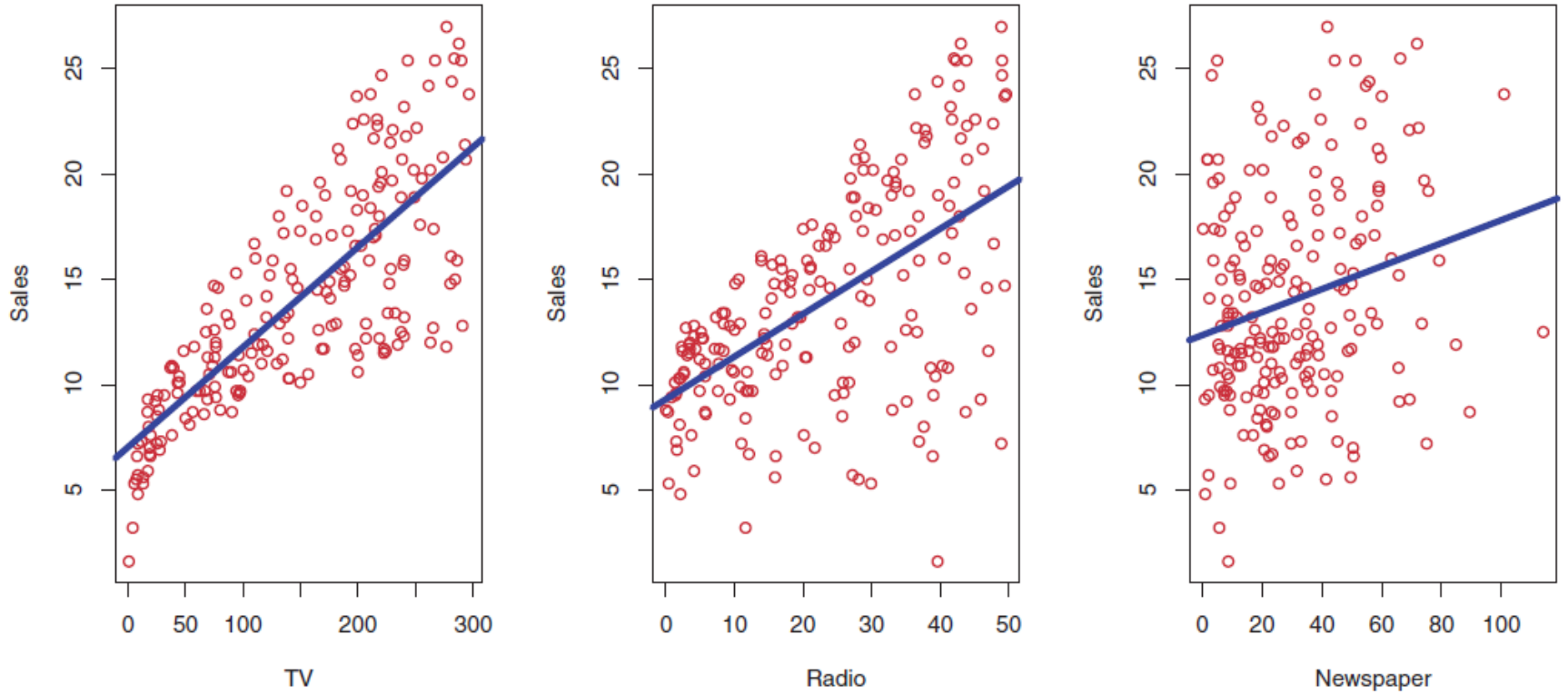
TEMAS PARA HOY

TEMAS PARA HOY

- Testeo de normalidad: Caso aplicado
- Modelo de regresión lineal en R
- Modelo de regresión logística en R

MODELOS DE REGRESIÓN

- Supongamos que interesa predecir el nivel de **ventas** de un determinado producto **en función** de los montos invertidos en distintos medios de publicidad (**tv, radio, periódico**).



MODELOS DE REGRESIÓN

- En general, el problema podría expresarse matemáticamente de la siguiente manera:

$$Y = f(X) + \epsilon \qquad (1)$$

- Donde X contiene a las variables explicativas (monto en tv, radio y periódico en el ejemplo), y ϵ es un error aleatorio no observable.
- En esta especificación, f es una función desconocida y que buscamos estimar.
- En términos simples, diremos que un modelo es de regresión, cuando en la expresión *(1)*, la variable de interés a predecir, Y, es una variable numérica.

REGRESIÓN LINEAL

- Cuando el modelo f a estimar, se asume como una función lineal, diremos que *(1)* es un modelo de regresión lineal. En tal caso, el modelo matemático queda expresado de la siguiente manera:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p \ + \epsilon \qquad (1)$$

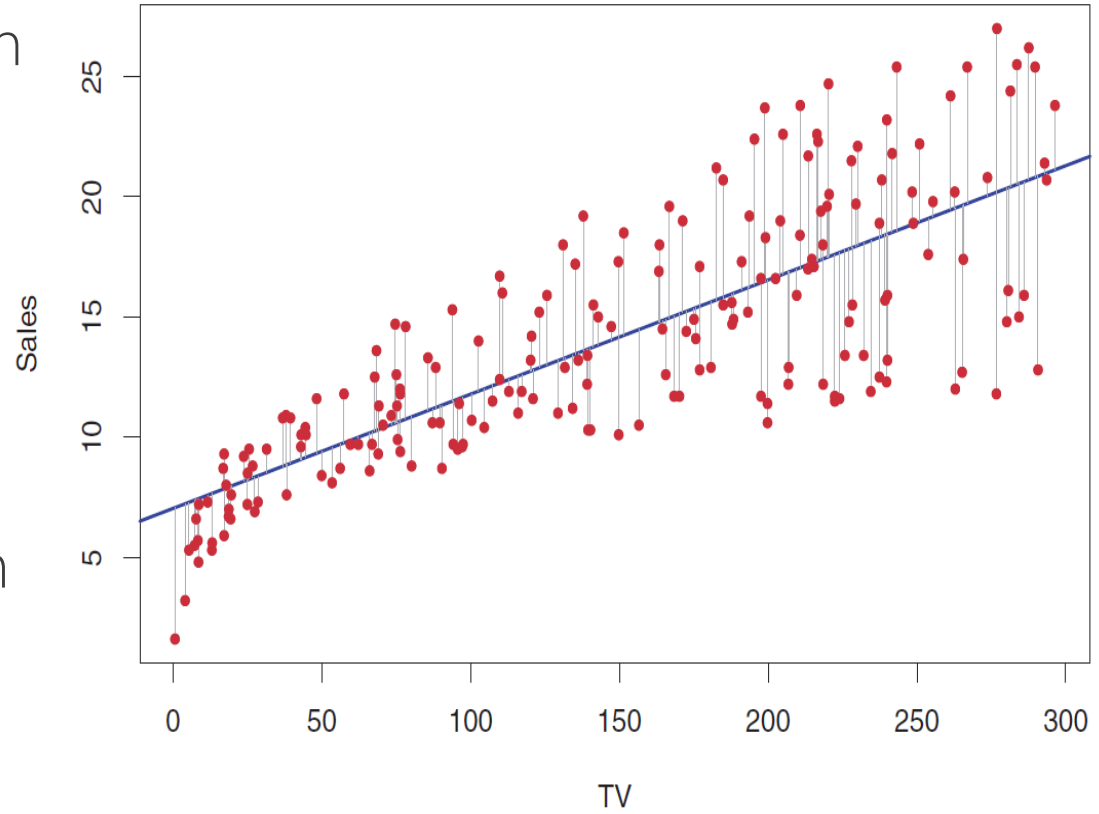
- Donde $\beta_i, i = 0, \ldots, p$ son los parámetros a estimar (estos parámetros definen al modelo), y ϵ es un error aleatorio no observable, típicamente siguiendo una distribución aleatoria normal $N(0, \sigma^2)$.

REGRESIÓN LINEAL

- ¿Cómo se estiman los parámetros en una regression lineal?
- Tanto en R como en la mayoría de los softwares, la manera estándar de estimar los coeficientes en un modelo de regresión, es mediante la estimación vía mínimos cuadrados, donde se busca minimizar la suma residual:

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

- No entraremos en detalle, respecto de las bondades de esta estimación y que coincide con otros estimadores en el caso de la regresión lineal con errores normales.



REGRESIÓN LINEAL

- Sin entrar en más detalles técnicos, veamos como podemos ajustar una regresión lineal en R.
- Para ello podemos utilizar la función **`lm()`**, del paquete **`base`**. Esta recibe como argumento principal una formula y un datasets, del siguiente modo.

```
lm(formula = y ~ x1+x2+…+xp, data = dataset)
```

- Generemos nuestra primera regresión lineal con el dataset "Advertising". El cual contiene las ventas totales de un producto y los montos invertidos en tres tipos de publicidad (tv, radio, periódico). En esta primera iteración consideremos solamente la variable *newspapper*

```
lm(formula = sales ~ newspapper, data = Advertising)
```

- Hablaremos sobres los coeficientes estimados y la salida que genera R en el notebook *Clase8_2_Regresion_lineal_en_r.ipynb*

REGRESIÓN LINEAL

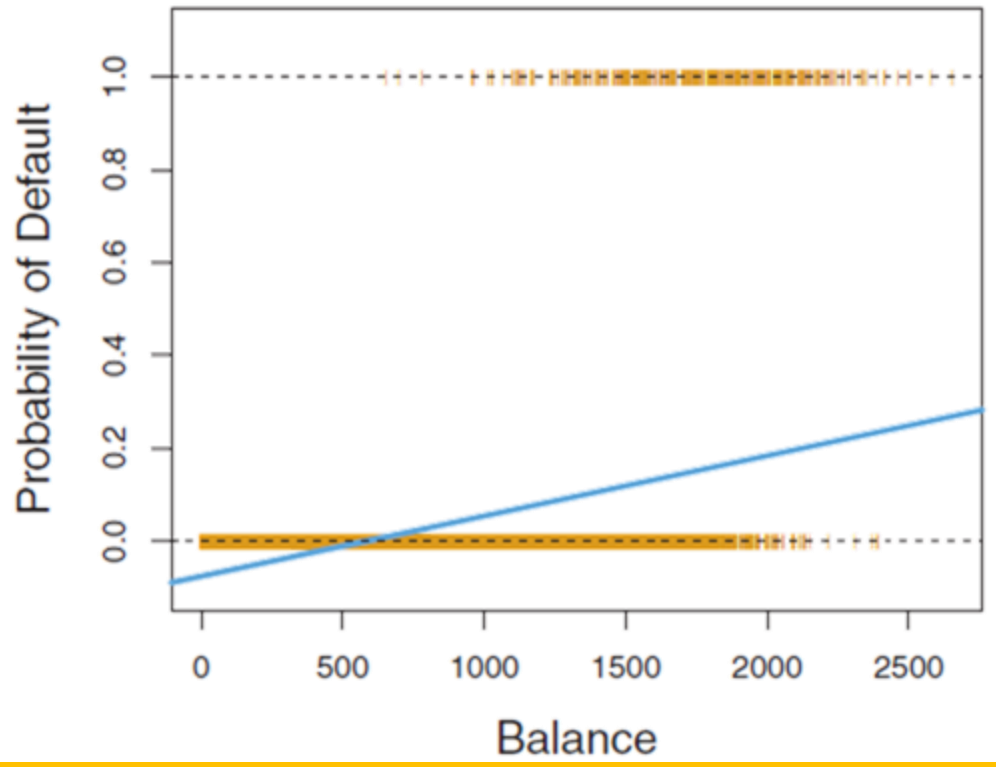
- Interpretación de coeficientes

REGRESIÓN LOGÍSTICA

- En el caso del problema de las ventas, el interés se centraba en modelar el total de ventas, variable numérica posiblemente continua si pensamos en USD.
- Si nuestro interés no es una variable cuantitativa, si no más bien **categórica**, una regresión lineal no es lo más adecuado.
- Por ejemplo, supongamos que nos interesa saber si una persona quedará en "default" (no terminará de pagar su tarjeta de crédito), en función de atributos como su ingreso, nivel de deuda y si es estudiante o no.
- Para tal efecto nos interesa modelar entonces la probabilidad de caer en "default".

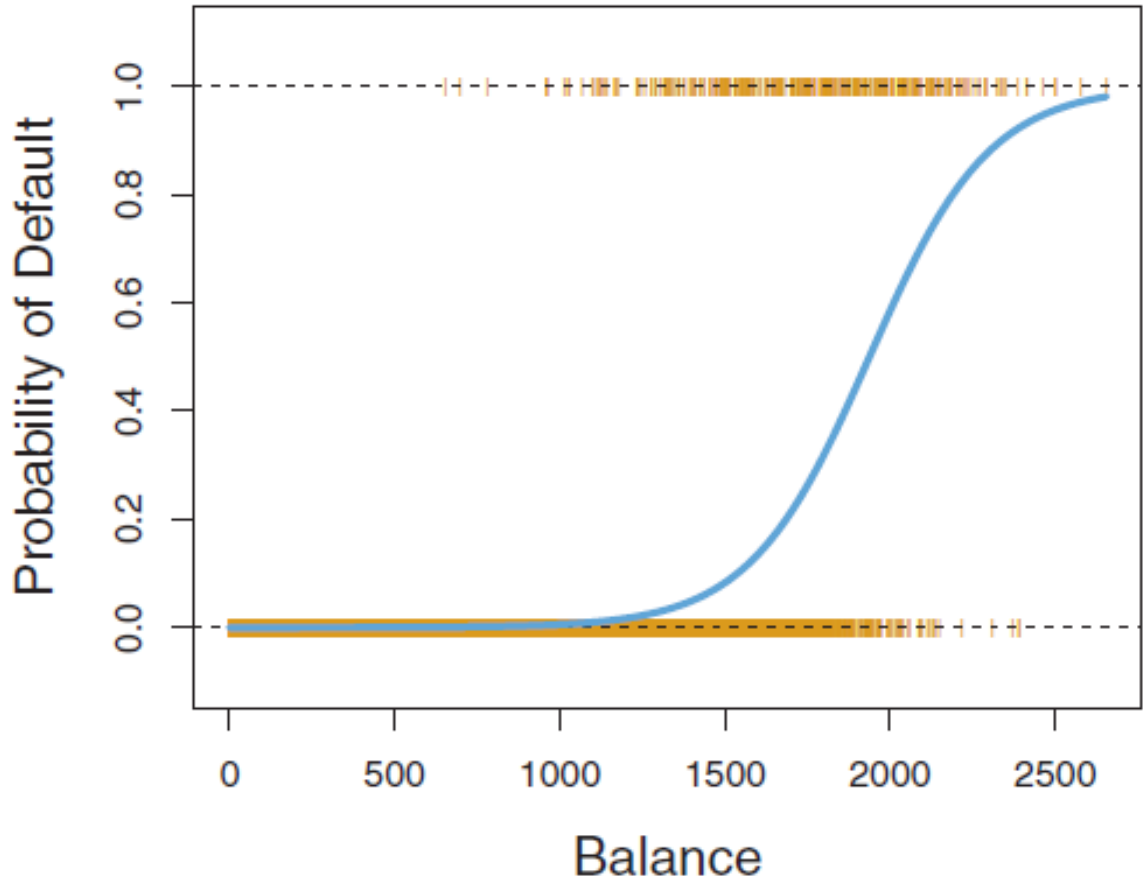
REGRESIÓN LOGÍSTICA

- Supongamos que en una primera aproximación sólo modelamos "default" en función de los ingresos.
- ¿Que problema podríamos tener si modelamos la probabilidad de "default" mediante una regresión lineal?



REGRESIÓN LOGÍSTICA

- Dado lo anterior, nos gustaría poder establecer un modelo que nos permita modelar la probabilidad de "default", por ejemplo, de la siguiente manera:



REGRESIÓN LOGÍSTICA

- Un posible modelo, corresponde a la regresión logística, el cual se especifica de la siguiente manera:
  - Supongamos que Y representa el estado del cliente, el cual se codifica como "Yes" si el cliente cayó en "default" y "No" , en caso contrario.

$$\log\left(\frac{P(Y = \text{Yes}|income)}{1 - P(Y = \text{Yes}|income)}\right) = \beta_0 + \beta_1 income$$

- Dada la ecuación anterior, es posible estimar los parámetros $\beta_0 y\ \beta_1$, mediante un procedimiento "similar" al de una regresión lineal.
- De manera más general, cuando incluímos más de una variable explicativa, el modelo se especifica de la siguiente manera.

$$\log\left(\frac{P(Y = 1|\mathbf{X} = \mathbf{x})}{1 - P(Y = 1|\mathbf{X} = \mathbf{x})}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

REGRESIÓN LOGÍSTICA

- En R, es posible ajustar un modelo logístico mediante la función glm la cual recibe como principales argumentos:
  - `formula:` Fórmula correspondiente al modelo planteado.
  - `family:` Argumento que especifica la distribución de los errores así como la función de enlace al modelo lineal generalizado que se busca ajustar. Para ajustar una regresión logística este parámetro por defecto está seteado a modo de ajustar una reg. logística cuando la variable respuesta es binaria.

```
family = binomial(link = “logit”)
```

  - `data:` Dataset desde donde se mapen las variables indicadas en `formula.`

REGRESIÓN LOGÍSTICA

- Veamos un caso en R…

REGRESIÓN LOGÍSTICA

- Interpretación de coeficientes

INGENIERÍA UC

EXPANDIENDO
CONOCIMIENTO
Y EXPERIENCIA