

# Linear Regression

Felipe José Bravo Márquez

May 11, 2021

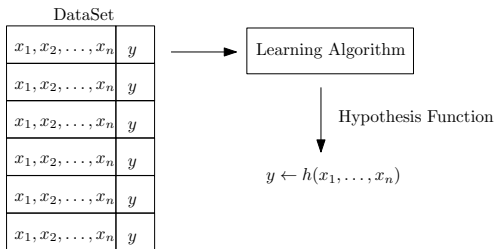
- Un modelo de regresión se usa para modelar la relación de una variable dependiente  $y$  numérica con  $n$  variables independientes  $x_1, x_2, \dots, x_n$ .
- A grandes rasgos queremos conocer el valor esperado de  $y$  a partir los valores de  $x$ :

$$\mathbb{E}(y|x_1, x_2, \dots, x_n)$$

- Usamos estos modelos cuando creemos que la variable de respuesta  $y$  puede ser modelada por otras variables independientes también conocidas como covariables o atributos.
- Para realizar este tipo de análisis necesitamos un dataset formado por  $m$  observaciones que incluyan tanto a la variable de respuesta como a cada uno de los atributos.
- Nos referimos al proceso de **ajustar** una función de regresión al proceso en que a partir de los datos inferimos una función de hipótesis  $h$  que nos permite predecir valores de  $y$  desconocidos usando los valores de los atributos.

# Introducción (2)

- A este proceso de ajustar una función a partir de los datos se le llama en las áreas de minería de datos y aprendizaje de máquinas como **entrenamiento**.
- En esas disciplinas se dice que las funciones **aprenden** a partir de los datos.
- Como necesitamos observaciones donde el valor de **y** sea conocido para aprender la función, se le llama a este tipo de técnicas como técnicas de **aprendizaje supervisado**.
- Cuando **y** es una variable categórica hablamos de un problema de **clasificación**.



# Regresión Lineal Simple

- En la regresión lineal simple se tiene una única variable independiente  $x$  para modelar la variable dependiente  $y$ .
- Se asume la siguiente relación lineal entre la variables:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad \forall i$$

- El parámetro  $\beta_0$  representa el intercepto de la recta (el valor de  $y$  cuando  $x$  vale cero).
- El parámetro  $\beta_1$  es la pendiente y representa el cambio de  $y$  cuando variamos el valor de  $x$ . Entre mayor sea la magnitud de este parámetro mayor será la relación lineal entre las variables.
- Los valores  $\epsilon_i$  corresponden a los errores asociados al modelo.
- Tenemos que encontrar una función lineal o recta  $h_\beta$  que nos permita encontrar una estimación de  $y$ ,  $\hat{y}$  para cualquier valor de  $x$  con el mínimo error esperado.

$$h(x) = \beta_0 + \beta_1 x$$

# Mínimos de Cuadrados

- El método de mínimos cuadrados ordinarios se usa para estimar  $\hat{\beta}_0$  y  $\hat{\beta}_1$  minimizando la suma de los errores cuadráticos (SSE) de los datos observados.
- Supongamos que tenemos  $m$  observaciones de  $\mathbf{y}$  y de  $\mathbf{x}$ , calculamos la suma de los errores cuadráticos (SSE) o  $E$  de error de la siguiente forma:

$$E = \sum_{i=1}^m (y_i - h(x_i))^2 = \sum_{i=1}^m (y_i - \beta_0 - \beta_1 x_i)^2 \quad (1)$$

- Para encontrar los parámetros que minimizan el error calculamos las derivadas parciales de SSE respecto a  $\beta_0$  y  $\beta_1$ . Luego igualamos las derivadas a cero y resolvemos la ecuación para despejar los parámetros.

$$\frac{\partial E}{\partial \beta_0} = -2 \sum_{i=1}^m (y_i - \beta_0 - \beta_1 x_i) = 0 \quad (2)$$

$$\frac{\partial E}{\partial \beta_1} = -2 \sum_{i=1}^m (y_i - \beta_0 - \beta_1 x_i) x_i = 0 \quad (3)$$

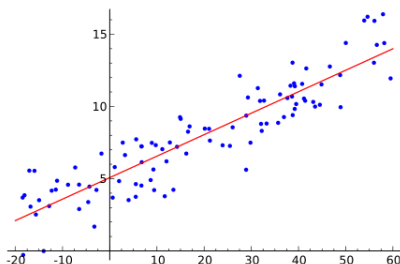
# Mínimos Cuadrados (2)

- Del sistema de ecuaciones anterior se obtienen las soluciones normales:

$$\hat{\beta}_1 = \frac{\sum_i^m (x_i - \bar{x})(y_i - \bar{y})}{\sum_i^m (x_i - \bar{x})^2} \quad (4)$$

$$\hat{\beta}_0 = \bar{y} - \beta_1 \bar{x} \quad (5)$$

- El modelo ajustado representa la recta de mínimo error cuadrático.



# Coeficiente de Determinación $R^2$

- Una vez ajustado nuestro modelo lineal debemos evaluar la calidad del modelo.
- Una medida muy común es el coeficiente de determinación  $R^2$ .
- Para calcularlo debo calcular otros errores distintos a los errores cuadráticos SSE.
- Se define a la suma cuadrática total (SST) como el error predictivo cuando usamos la media  $\bar{y}$  para predecir la variable de respuesta  $y$  (es muy similar a la varianza de la variable):

$$SST = \sum_i^m (y_i - \bar{y})^2$$

- Luego tenemos a la suma de los cuadrados explicada por el modelo (SSM) que nos indica la variabilidad de los valores predichos por el modelo respecto a la media:

$$SSM = \sum_i^m (\hat{y}_i - \bar{y})^2$$

# Coeficiente de Determinación $R^2$ (2)

- Se define el coeficiente de determinación para un modelo lineal  $R^2$  como:

$$R^2 = \frac{SSM}{SST} = \frac{\sum_i^m (\hat{y}_i - \bar{y})^2}{\sum_i^m (y_i - \bar{y})^2} \quad (6)$$

- El coeficiente adquiere valores entre 0 a 1 y mientras más cercano a 1 sea su valor mayor será la calidad del modelo.
- El valor de  $R^2$  es equivalente a la correlación lineal (Pearsons) entre  $y$  e  $\hat{y}$  al cuadrado.

$$R^2 = \text{cor}(y, \hat{y})^2$$



# Regresión Lineal Múltiple

- Supongamos que tenemos  $n$  variables independientes:  $x_1, x_2, \dots, x_n$ .
- Intuitivamente, estas variables en conjunto podrían explicar de mejor manera la variabilidad de la variable de respuesta  $y$  que un modelo simple.
- Se define un modelo lineal multi-variado de la siguiente manera:

$$y_i = \beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \dots + \beta_n x_{i,n} + \epsilon_i \quad \forall i \in \{1, m\}$$

- En el modelo multi-variado se extienden todas las propiedades del modelo lineal simple.
- Se puede representar el problema de manera matricial:

$$Y = X\beta + \epsilon$$

- Donde  $Y$  es un vector de  $m \times 1$  de variables de respuesta:

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_m \end{pmatrix}$$

# Regresión Lineal Múltiple (2)

- $X$  es una matriz de  $m \times (n + 1)$  con las variables explicativas. Tenemos  $m$  observaciones de las  $n$  variables. La primera columna es constante igual a 1 ( $x_{i,0} = 1 \quad \forall i$ ) para modelar la variable de intercepto  $\beta_0$ .

$$X = \begin{pmatrix} x_{1,0} & x_{1,1} & x_{1,2} & \cdots & x_{1,n} \\ x_{2,0} & x_{2,1} & x_{2,2} & \cdots & x_{2,n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{m,0} & x_{m,1} & x_{m,2} & \cdots & x_{m,n} \end{pmatrix}$$

- Luego,  $\beta$  es un vector de parámetros de  $(n + 1) \times 1$

$$\beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{pmatrix}$$

# Regresión Lineal Múltiple (2)

- Finalmente,  $\epsilon$  es un vector con los errores del modelo de  $m \times 1$ .

$$\epsilon = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_m \end{pmatrix}$$

- Usando la notación matricial, podemos ver que la suma de los errores cuadráticos (SSE) se puede expresar como:

$$\text{SSE} = (Y - X\beta)^T(Y - X\beta)$$

- Minimizando esta expresión derivando el error en función de  $\beta$  e igualando a cero se llega a las ecuaciones normales:

$$\hat{\beta} = (X^T X)^{-1} X^T Y$$

# Supuestos del Modelo Lineal

Cada vez que ajustamos un modelo lineal estamos asumiendo implícitamente ciertos supuestos sobre los datos.

## Supuestos

- 1 Linealidad: la variable de respuesta se relaciona linealmente con los atributos.
- 2 Normalidad: los errores tienen distribución normal de media cero:  $\epsilon_i \sim N(0, \sigma^2)$
- 3 Homocedasticidad: los errores tienen varianza constante (mismo valor  $\sigma^2$ ).
- 4 Independencia: los errores son independientes entre sí.

# Interpretación Probabilística

- Considerando los supuestos anteriores podemos ver que la densidad de probabilidad (PDF) de los errores  $\epsilon$  esta definida por una normal de media cero y varianza constante:

$$\text{PDF}(\epsilon_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\epsilon_i^2}{2\sigma^2}\right)$$

- Esto implica que:

$$\text{PDF}(y_i|x_i; \beta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - h_\beta(x_i))^2}{2\sigma^2}\right)$$

- Lo que implica que la distribución de  $\mathbf{y}$  dada los valores de  $\mathbf{x}$  y parametrizada por  $\beta$  sigue una distribución normal.
- Luego si uno estima los parámetros de  $\beta$  usando una técnica de estimación llamada máxima verosimilitud llega a los mismos resultados que haciendo estimación por mínimos cuadrados.
- Esto nos dice que cuando estimamos los parámetros del modelo usando mínimos cuadrados estamos realizando las mismas hipótesis probabilísticas mencionados anteriormente.

# Regresiones en R

- En R los modelos lineales se crean con el comando `lm` que recibe como parámetro una fórmula de la forma  $y \sim x$  ( $y = f(x)$ ).
- Vamos a trabajar con el dataset `USArrests` que tiene información sobre los arrestos ocurridos en Estados Unidos el año 1973.
- Cada observación corresponde a un estado.
- Tiene las siguientes variables:
  - 1 Murder**: arrestos por homicidio (por 100.000 habitantes).
  - 2 Assault**: arrestos por asalto (por 100.000 habitantes).
  - 3 UrbanPop**: porcentaje de la población total del estado.
  - 4 Rape**: arrestos por violación (por 100.000 habitantes).
- Para ver si vale la pena hacer un análisis de regresión lineal vemos las correlaciones lineales entre las variables:

```
> data(USArrests)
> attach(USArrests)
> cor(USArrests)
```

	Murder	Assault	UrbanPop	Rape
Murder	1.00000000	0.8018733	0.06957262	0.5635788
Assault	0.80187331	1.0000000	0.25887170	0.6652412
UrbanPop	0.06957262	0.2588717	1.0000000	0.4113412
Rape	0.56357883	0.6652412	0.41134124	1.0000000

# Regresiones en R (2)

- Podemos ver que hay una correlación positiva importante entre `Murder` y `Assault`.
- Vamos a modelar los asesinatos en función de los asaltos usando una regresión lineal simple:

$$\text{Murder}(\text{Assault}) = \beta_0 + \beta_1 * \text{Assault}$$

```
> reg1<-lm(Murder~Assault,USArrests)
> reg1
```

Call:

```
lm(formula = Murder ~ Assault, data = USArrests)
```

Coefficients:

(Intercept)	Assault
0.63168	0.04191

- Podemos ver que los coeficientes del modelo son  $\beta_0 = 0.632$  y  $\beta_1 = 0.042$ .

# Regresiones en R (3)

- Podemos acceder directamente a los coeficientes y guardarlos en una variable:

```
> reg1.coef<-reg1$coefficients
> reg1.coef
(Intercept)      Assault
  0.63168266   0.04190863
```

- Podemos ver diversos indicadores sobre el modelo lineal con el comando **summary**:

```
> summary(reg1)
Residuals:
    Min       1Q   Median       3Q      Max
-4.8528 -1.7456 -0.3979  1.3044  7.9256

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.631683   0.854776   0.739    0.464
Assault      0.041909   0.004507   9.298 2.6e-12 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.629 on 48 degrees of freedom
Multiple R-squared:  0.643, Adjusted R-squared:  0.6356
F-statistic: 86.45 on 1 and 48 DF,  p-value: 2.596e-12
```



# Regresiones en R (4)

- Vemos que el coeficiente de determinación  $R^2$  tiene un valor de 0.643 lo cual no es tan bueno pero aceptable.
- Podemos concluir que el nivel de asaltos si bien provee información útil para modelar una parte de la variabilidad del nivel de homicidios no es suficiente para construir un modelo altamente confiable.
- Puedo guardar los resultados del comando `summary` en una variable y así acceder directamente al coeficiente de determinación:

```
> sum.reg1<-summary(reg1)
> sum.reg1$r.squared
[1] 0.6430008
```

- También puedo acceder a los valores ajustados que son los valores predichos por mi modelo para los datos usados:

```
> reg1$fitted.values
```

Alabama	Alaska	Arizona	Arkansas
10.522119	11.653652	12.952819	8.594322

# Regresiones en R (5)

- Podemos ver que la correlación lineal al cuadrado entre mis valores ajustados y los observados para la variable de respuesta es equivalente al coeficiente de determinación:

```
> cor(Murder, reg1$fitted.values) ^2  
[1] 0.6430008
```

- Supongamos ahora que conozco el nivel de asalto de dos estados en otro período para dos lugares pero no conozco el nivel de homicidios.
- Podría usar mi modelo lineal para predecir el nivel de de homicidios.
- Para hacerlo en R debo usar el comando `predict.lm` que recibe el modelo lineal y un `data.frame` con los datos nuevos:

```
> nuevos.arrestos<-data.frame(Assault=c(500,12))  
> predict.lm(object=reg1,newdata=nuevos.arrestos)  
          1          2  
21.585997  1.134586  
> # Esto es equivalente a:  
> reg1.coef[1]+reg1.coef[2]*nuevos.arrestos  
      Assault  
1 21.585997  
2  1.134586
```

# Regresiones en R (6)

- Ahora estudiaremos una regresión lineal múltiple.
- Podemos ver que la variable **Rape** que representa el nivel de violaciones tiene una correlación menor con el número de asaltos y con el número de homicidios que la correlación que presentan estas dos variables entre sí.
- Vamos a ajustar el siguiente modelo lineal multi-variado:

$$\text{Rape} = \beta_0 + \beta_1 * \text{Assault} + \beta_2 * \text{Murder}$$

- En R para agregar más variables al modelo lineal las agregamos con el signo **+** :  
`reg2<-lm(Rape~Assault+Murder,USArrests)`

# Regresiones en R (7)

```
> summary(reg2)
```

Residuals:

Min	1Q	Median	3Q	Max
-17.243	-3.171	-1.171	3.281	18.511

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	8.35011	2.32912	3.585	0.000799	***
Assault	0.06716	0.02044	3.286	0.001927	**
Murder	0.18155	0.39108	0.464	0.644619	

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 7.124 on 47 degrees of freedom

Multiple R-squared: 0.4451, Adjusted R-squared: 0.4215

F-statistic: 18.85 on 2 and 47 DF, p-value: 9.755e-07

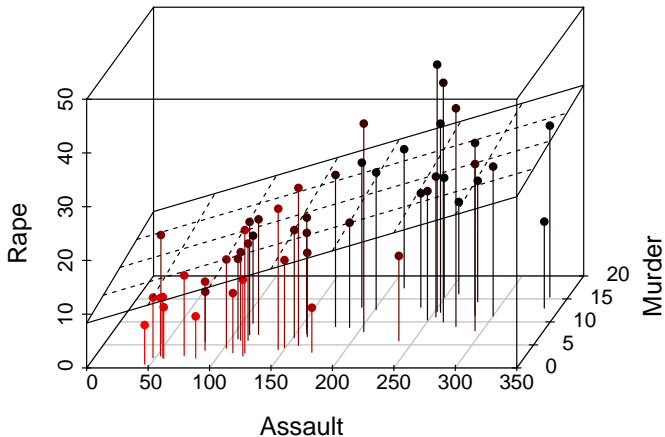
- En este caso el coeficiente de determinación es bajo. Por lo que tendremos baja confianza en la calidad del modelo.

# Regresiones en R (8)

- Cuando teníamos una regresión simple podíamos ver el modelo ajustado como una recta.
- Ahora que tenemos dos variables independientes podemos ver el modelo ajustado como un plano.
- Si tuviésemos más variables independientes nuestro modelo sería un hiper-plano.
- Podemos graficar en R el plano de nuestro modelo lineal de dos variables independientes y una dependiente de la siguiente manera:

```
library("scatterplot3d")
s3d <- scatterplot3d(USArrests[,c("Assault", "Murder", "Rape")],
                     type="h", highlight.3d=TRUE,
                     angle=55, scale.y=0.7, pch=16,
                     main="Rape~Murder+Rape")
s3d$plane3d(reg2, lty.box = "solid")
```

## Rape~Assault+Murder



# FOUR CARDINAL RULES OF STATISTICS by Daniela Witten

- ONE: CORRELATION DOES NOT IMPLY CAUSATION. Yes, I know you know this, but it's so easy to forget! Yeah, YOU OVER THERE, you with the p-value of 0.0000001 — yes, YOU!! That's not causation.
- No matter how small the p-value for a regression of IQ onto shoe size is, that doesn't mean that big feet cause smarts!! It just means that grown-ups tend to have bigger feet and higher IQs than kids.
- So, unless you can design your study to uncover causation (very hard to do in most practical settings — the field of causal inference is devoted to understanding the settings in which it is possible), the best you can do is to discover correlations. Sad but true.
- TWO: A P-VALUE IS JUST A TEST OF SAMPLE SIZE. Read that again — I mean what I said! If your null hypothesis doesn't hold (and null hypotheses never hold IRL) then the larger your sample size, the smaller your p-value will tend to be.
- If you're testing whether  $\text{mean}=0$  and actually the truth is that  $\text{mean}=0.000000001$ , and if you have a large enough sample size, then YOU WILL GET A TINY P-VALUE.
- Why does this matter? In many contemporary settings (think: the internet), sample sizes are so huge that we can get TINY p-values even when the deviation from the null hypothesis is negligible. In other words, we can have STATISTICAL significance w/o PRACTICAL significance.

# FOUR CARDINAL RULES OF STATISTICS by Daniela Witten

- Often, people focus on that tiny p-value, and the fact that the effect is of **\*\*literally no practical relevance\*\*** is totally lost.
- This also means that with a large enough sample size we can reject basically ANY null hypothesis (since the null hypothesis never exactly holds IRL, but it might be “close enough” that the violation of the null hypothesis is not important).
- Want to write a paper saying Lucky Charms consumption is correlated w/blood type? W/a large enough sample size, you can get a small p-value. (Provided there's some super convoluted mechanism with some teeny effect size... which there probably is, b/c IRL null never holds)
- **THREE: SEEK AND YOU SHALL FIND.** If you look at your data for long enough, you will find something interesting, even if only by chance! In principle, we know that we need to perform a correction for multiple testing if we conduct a bunch of tests.
- But in practice, what if we decide what test(s) to conduct **AFTER** we look at data? Our p-value will be misleadingly small because we peeked at the data. Pre-specifying our analysis plan in advance keeps us honest... but in reality, it's hard to do!!!
- Everyone is asking me about the mysterious and much-anticipated fourth rule of statistics. The answer is simple: we haven't figured it out yet.... that's the reason we need to do research in statistics





L. Wasserman *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics, 2005.