Regresión Lineal

Felipe José Bravo Márquez

15 de noviembre de 2013

Introducción

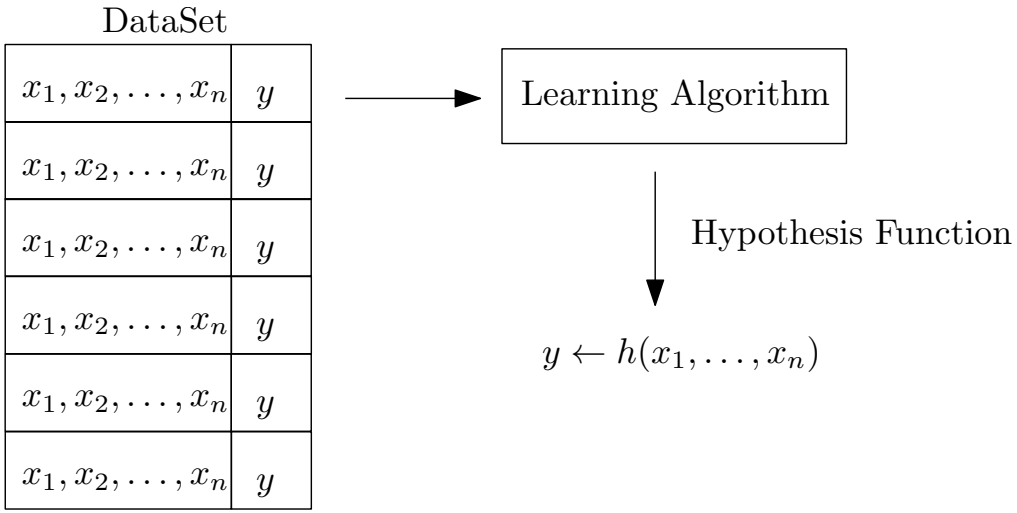
- Un modelo de regresión se usa para modelar la relación de una variable dependiente **y** numérica con n variables independientes $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$.
- A grandes rasgos queremos conocer el valor esperado de **y** a partir los valores de **x**:

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

- 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 β_0 representa el intercepto de la recta (el valor de y cuando x vale cero).
- El parámetro β_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 ϵ_i corresponden a los errores asociados al modelo.
- Tenemos que encontrar una función lineal o recta h_β 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 **y** y de **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 β_0 y β_1. Luego igualamos las derivadas a cero y resolvemos la ecuación para despejar los parámetros.

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

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

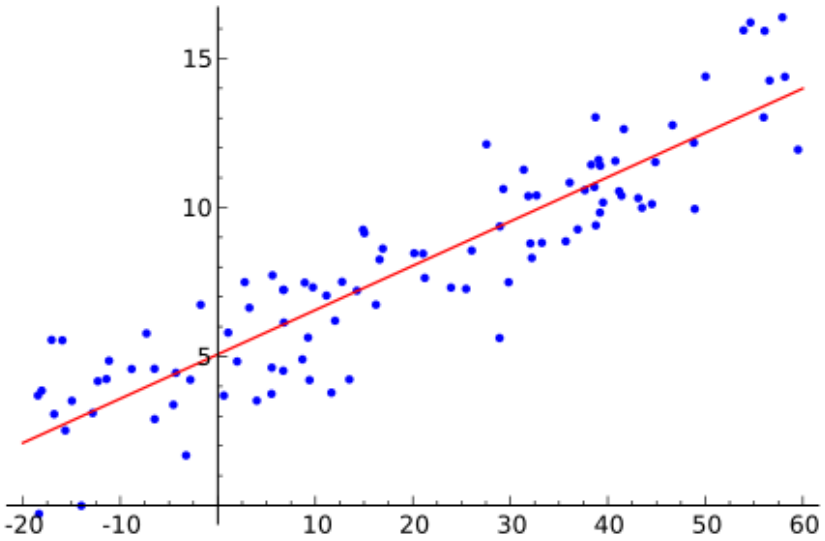
Mínimos Cuadrados (2)

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

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

$$\hat{\beta}_0 = \overline{y} - \beta_1 \overline{x} \tag{5}$$

- 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 $\overline{y}$ para predecir la variable de respuesta y (es muy similar a la varianza de la variable):

$$\text{SST} = \sum_{i}^{m} (y_i - \overline{y})^2$$

- 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:

$$\text{SSM} = \sum_{i}^{m} (\hat{y}_i - \overline{y})^2$$

Coeficiente de Determinación R^2 (2)

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

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

- El coeficiente adquiere valores entre 0 a 1 y mientras más cercano a 1 sea su valor mayor sera 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: $x_1, x_2, \ldots, x_n$.
- 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:

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

- 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 β_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, β 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, ϵ 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 β 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 σ^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 ϵ 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 **y** dada los valores de **x** y parametrizada por β sigue una distribución normal.
- Luego si uno estima los parámetros de β 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 lo modelos lineales se crean con el comando `lm` que recibe como parámetro una fórmula de la forma `y~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 1.00000000 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:

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

```
> 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
```

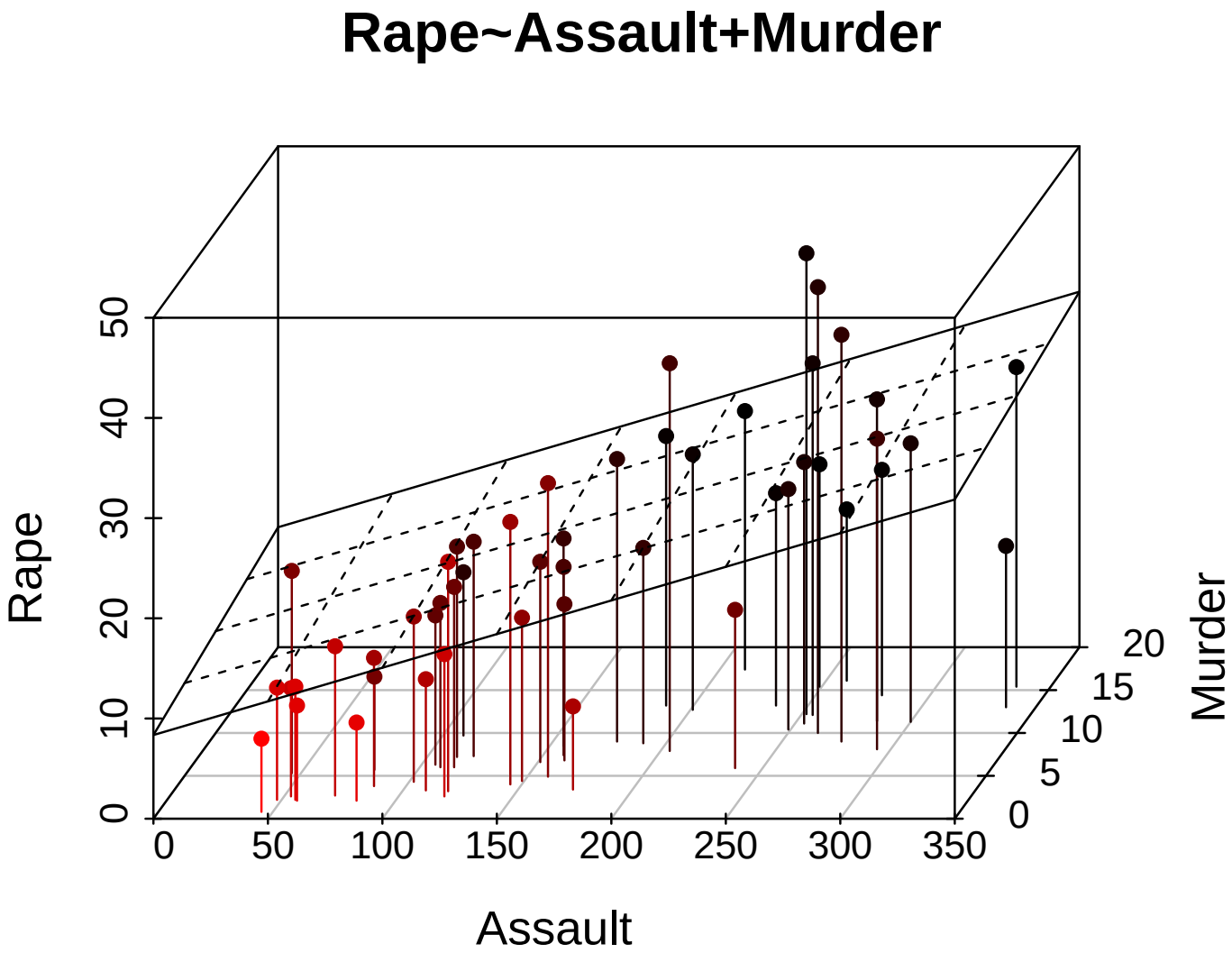
- En este caso el coeficiente de determinación es bajo. Por lo que tendremos baja confianza en la calidad del modelo.

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")
```

Regresiones Lineales en R (9)



Bilbiografía I


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