Regresión Lineal con R

Clasica y Bayesiana

Jorge Mario Estrada A. MSc.

Comfamiliar Risaralda

¿ Qué revisaremos ?

- Objetivos e importancia (Conceptos teóricos)
- Procedimiento en R
- Ejemplo de aplicación

Objetivo e importancia

- Estudia la relación entre variables: Describe, modela y predice
- Base para crear modelos mas avanzados
- Central en Data Science y Machine Learning

Condiciones iniciales para su implementación

1. **Variable respuesta (Y):** numericá, continua

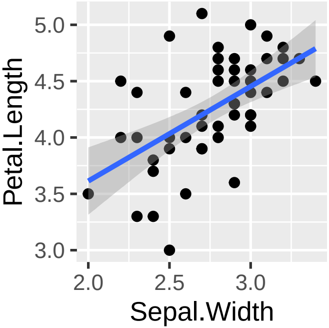
Variables explicativas (X´s): continua, categórica.

2. Supuestos: La relacion a modelar entre las variables es lineal y se desea describir de una forma mas detallada su relacion e incluso llegar a predecir la variables de respuesta en funcion de la(s) explicativa(s).

El modelo lineal

Se supone en un principio que se tiene evidencia de una relación lineal (Y ~ x) con correlación entre ellas.

Paso 1: cuantificar y observar dicha relación



Correlación $= 0.5605221$

t $=48$

p-valor $= 2.302168 \times 10^{-5}$

Modelo lineal

Asumiendo que la relación entre las variables podria modelarse mediante una relación lineal, la distribución de los valores de Y se daria asi.

$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_k X_k + \epsilon$$

con $k = 1, 2, 3, ...k - variables$ y $i = 1, 2, 3, .., n$

Supuesto estadistico: Y asume distribución Normal $Y \sim N(\mu, \sigma)$.

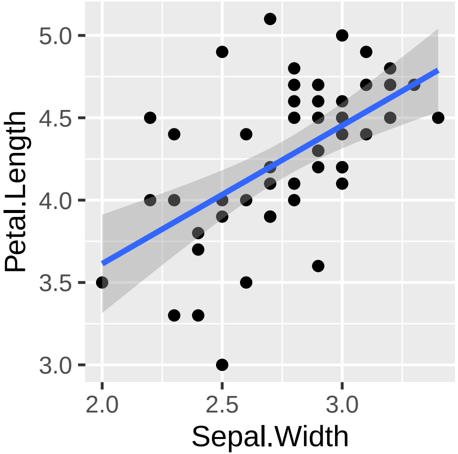
Problema a resolver: Conocer los valores que asume los β_k

Metodos:

- Minimos cuadrados
- Maxima verosimilitud
- Enfoque bayesiano

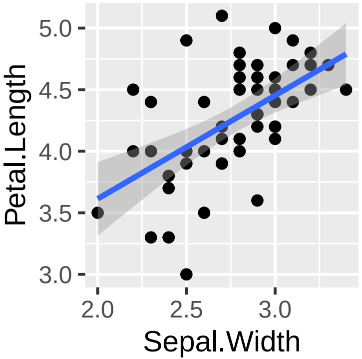
Interpretación de coeficientes β_k

- Intercepto β_0: el valor esperado para la variable respuesta Y cuando la variable explicativa X_k toma el valor de cero. refleja la media de la variable respuesta.



Interpretación de coeficientes β_k

- Pendiente β_1: el aumento promedio en la variable respuesta asociado a una unidad de aumento en la variable explicativa.



$$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_k X_k + \epsilon$$

Estimación por maximaverosimilitud con R

R para regresion lineal clasica

- Función base `lm(Y~x1+x2+x3, data = datos)`

```
rls <- lm(Petal.Length ~ Sepal.Width, data = iris)
```

R para regresion lineal clasica

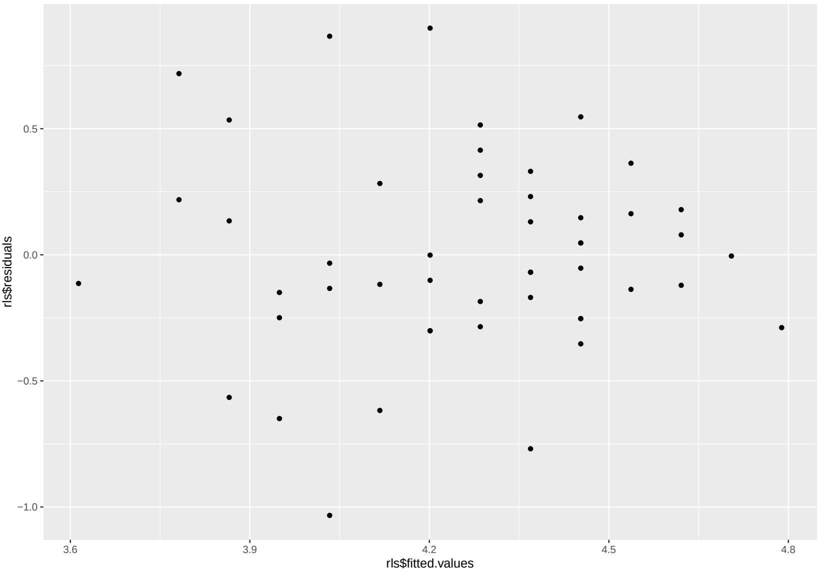
```
Call:
lm(formula = Petal.Length ~ Sepal.Width, data = iris)

Residuals:
     Min       1Q   Median       3Q      Max
-1.03337 -0.23337 -0.04321  0.21754  0.89876

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9349     0.4989   3.878  0.00032 ***
Sepal.Width   0.8394     0.1790   4.689  2.3e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 '

Residual standard error: 0.3932 on 48 degrees of freedom
Multiple R-squared:  0.3142,    Adjusted R-squared:  0.299
F-statistic: 21.99 on 1 and 48 DF,  p-value: 2.302e-05
```

Diagnostico del modelo



Estimación clasica vs Estimación bayesiana

Frecuentista: Los datos muestreados de la población se consideran aleatorios y los valores de los parámetros de la población, son fijos (pero desconocidos). Para estimar buscamos los parámetros muestrales que maximicen la probabilidad de los datos.

x = Datos

θ_i = parametros a estimar

$$modeloprobabilistico + datos$$

$$p(Datos, \theta) = f(x/\theta)$$

Estimación clasica vs Estimación bayesiana

El enfoque bayesiano: Este enfoque se basa en el teorema de Bayes, por ejemplo, si tenemos un parámetro θ de una población y tenemos algunos datos muestreados D al azar de esta población, se podría estimar la distribución de valores de θ dado los datos muestreados D que se tienen.

$$p(\theta/D) = f(\theta/x) = f(x/\theta) \times f(\theta)$$

$f(\theta/x)$: Función de distribución posterior

$f(x/\theta)$: Función de verosimilitud

$f(\theta)$: Función a priori

Resumiendo: Estimación bayesiana en RLS

$Y_i = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + .. + \beta_k X_k + \epsilon$

$p(Y_i) = f(x/\beta, \sigma)$

entonces los coeficientes son parametros a estimar por tanto su comportamiento depende de conocimiento previo, que se ve reflejado en una distribución de probabilidad conocida.

$$p(\beta_k) = f(\beta_k)$$

$$p(\sigma) = f(\sigma)$$

La distribución posterior estaria representada por:

$$f(\beta, \sigma/x) = f(x/\beta, \sigma) \times f(\beta, \sigma)$$

R para regresión lineal bayesiana

```
stan_glm(formula = , # modelo
         family = gaussian(), # regesion lineal es
                              # normal por defecto
         data = , # dataset a usar
         prior = NULL, # define el prior para
                       # los coeficientes beta
         seed = , # semilla para la simulacion
         iter = 4000, # numero de iteraciones
         prior_intercept = , # define un prior
         #diferente para el intercepto
         algorithm = "sampling", # metodo de muestrear
                                 # el posterior
         chains = 4, # numero de cadenas para MCMC
         warmup = 1000)
```

R para regresion lineal bayesiana

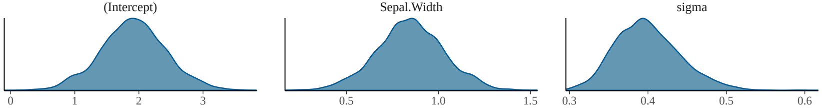
```
rlsbayes <- stan_glm(formula = Petal.Length ~ Sepal.Width,
         family = gaussian(),
         data = iris,
         prior = NULL,
         seed = 111,
         algorithm = "sampling")
```

R para regresion lineal bayesiana

```
                      mean         sd        50%
(Intercept)      1.9137691 0.51349544   1.9114014
Sepal.Width      0.8468717 0.18444708   0.8468133
sigma            0.3999444 0.04153150   0.3966277
mean_PPD         4.2603286 0.08183443   4.2594691
log-posterior  -26.9060720 1.27836442 -26.5678172
```

R para regresion lineal bayesiana

```r
mcmc_dens(rlsbayes,
          pars = c("(Intercept)","Sepal.Width","sigma"))
```




```r
rlsbayes$coefficients
```

```
(Intercept) Sepal.Width
  1.9114014   0.8468133
```

```r
hdi(rlsbayes)
```

```
Highest Density Interval

Parameter   |      95% HDI
--------------------------
(Intercept) | [0.85, 2.90]
Sepal.Width | [0.48, 1.22]
```

Practiquemos