Regresión Lineal

Luis Mantilla

2023-10-06

Regresión Lineal

Se tiene la base de datos de nacimientos del año 2016 llamada "asriosgu_Base.csv" (La cual llamaremos a lo largo del trabajo como "datos"), tomada de la pagina del DANE, y se quiere crear un modelo lineal que me pronostique el peso del infante. Para esto primero debemos ver que variables tienen correlación con el peso del recién nacido:

Veamos que tipo de columnas tiene la base de datos:

```
names(datos)
```

```
##  [1] "Departamento.Nacimiento"             "Municipio.Nacimiento"
##  [3] "Área.Nacimiento"                     "Género"
##  [5] "Peso"                                "Talla"
##  [7] "Fecha.de.Nacimiento"                 "Hora.de.Nacimiento"
##  [9] "Parto.Atendido.por..."               "Tiempo.de.Gestación"
## [11] "Número.de.Consultas.Prenatales"      "Tipo.de.Parto"
## [13] "Multiplicidad.de.Embarazo"           "APGAR1"
## [15] "APGAR2"                              "Grupo.Sanguíneo"
## [17] "Factor.RH"                           "Pertenencia.Étnica"
## [19] "Tipo.de.Documento.de.la.Madre"       "Edad.de.la.Madre"
## [21] "Estado.Conyugal.Madre"               "Nivel.Educativo.de.la.Madre1"
## [23] "Nivel.Educativo.de.la.Madre"         "País.de.Residencia"
## [25] "Departamento.Residencia"             "Municipio.Residencia"
## [27] "Área.de.Residencia"                  "Barrio"
## [29] "Dirección"                           "Centro.Poblado"
## [31] "Rural.Disperso"                      "Número.de.Hijos.Nacidos.Vivos"
## [33] "Fecha.Anterior.del.Hijo.Nacido.Vivo" "Número.de.Embarazos"
## [35] "Régimen.Seguridad"                   "Tipo.de.Administración"
## [37] "Nombre.de.la.Administradora"         "Edad.del.Padre"
## [39] "Nivel.Educativo.del.Padre"           "Último.Año.Aprobado.del.Padre"
```

Veamos si la Talla, el tiempo de gestación, Edad de la madre o número de embarazos, afectan al peso del recién nacido. Para este propósito debemos realizar para cada una de las posibles variables un test de correlación con respecto al peso, en este caso implementaremos el método de Pearson.

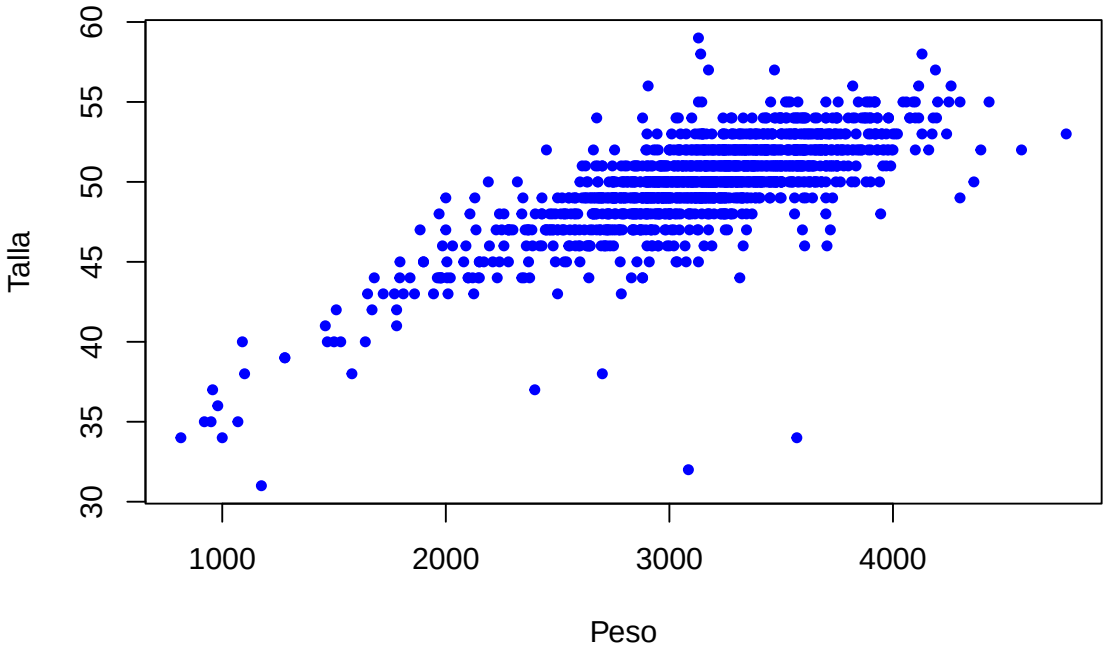
- Test de correlación del peso y la Talla

```
cor.test(datos$Peso, datos$Talla,method= "pearson" )
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos$Peso and datos$Talla
## t = 43.851, df = 1430, p-value < 2.2e-16
```

```
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.7343043 0.7785589
## sample estimates:
##       cor
## 0.7572996
```

Observemos que el p valor es menor al 5%, entonces no rechazamos el hecho de que existe una correlación del peso del infante con la talla, por otro lado la correlación de las dos variables es del 0.75, lo cual nos muestra una correlación sustancial entre variables. Además notemos que gráficamente si parece tener una tendencia los datos.

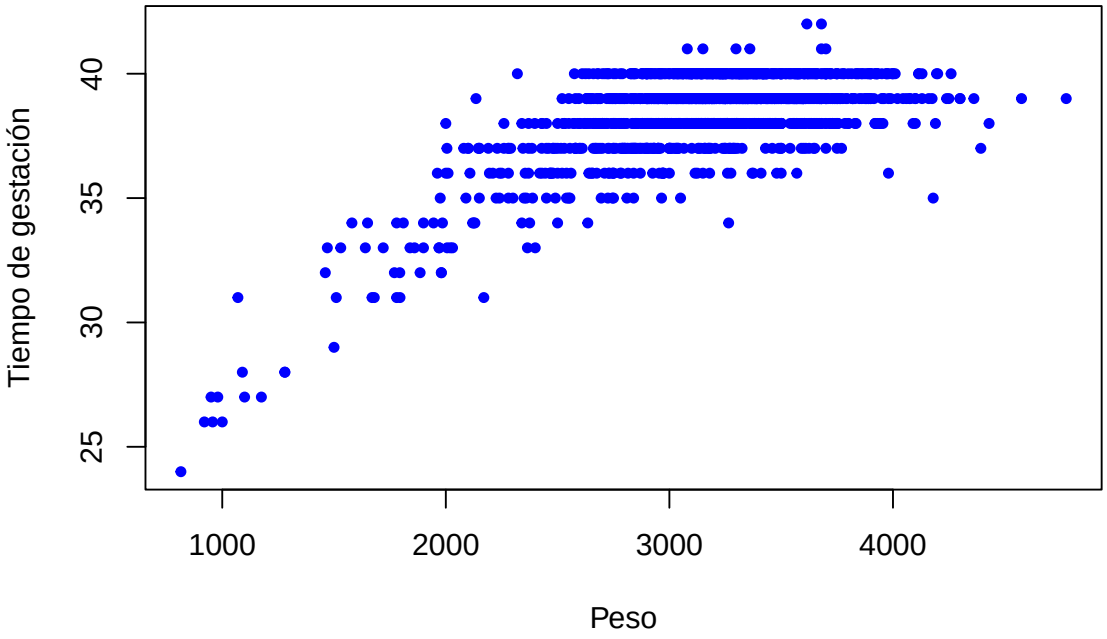


- Test de correlación del peso y el tiempo de gestación

```
cor.test(datos$Peso, datos$Tiempo.de.Gestación,method= "pearson" )
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos$Peso and datos$Tiempo.de.Gestación
## t = 33.322, df = 1430, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.6309308 0.6893190
## sample estimates:
##       cor
## 0.6611247
```

Como el valor p es menor al 5% entonces decimos tiene una buena correlación de manera similar al anterior test. Además veamos que la gráfica de los datos parecen seguir una tendencia.

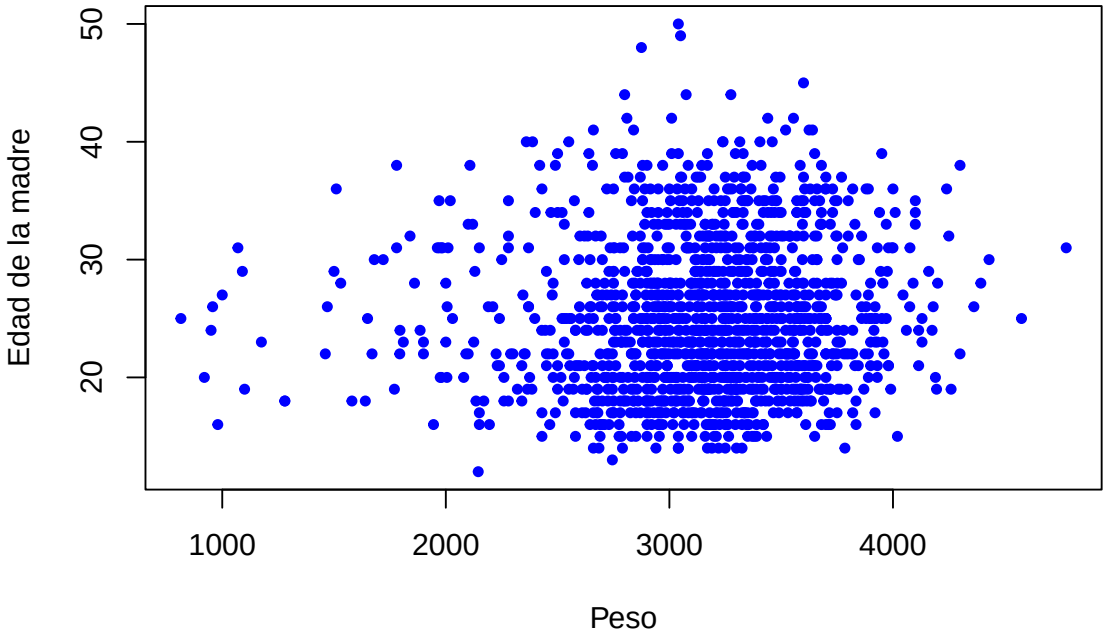


- Test de correlación del peso y la edad de la madre

```
cor.test(datos$Peso, datos$Edad.de.la.Madre,method= "pearson" )
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos$Peso and datos$Edad.de.la.Madre
## t = 2.3725, df = 1430, p-value = 0.0178
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.01085042 0.11404846
## sample estimates:
##        cor
## 0.06261681
```

Veamos que el valor p es menor al 5%, entonces no rechazamos la hipótesis de que existe correlación entre las variables, sin embargo veamos que la correlación entre variables es insignificante y si vemos la gráfica de los datos, no vemos alguna tendencia.

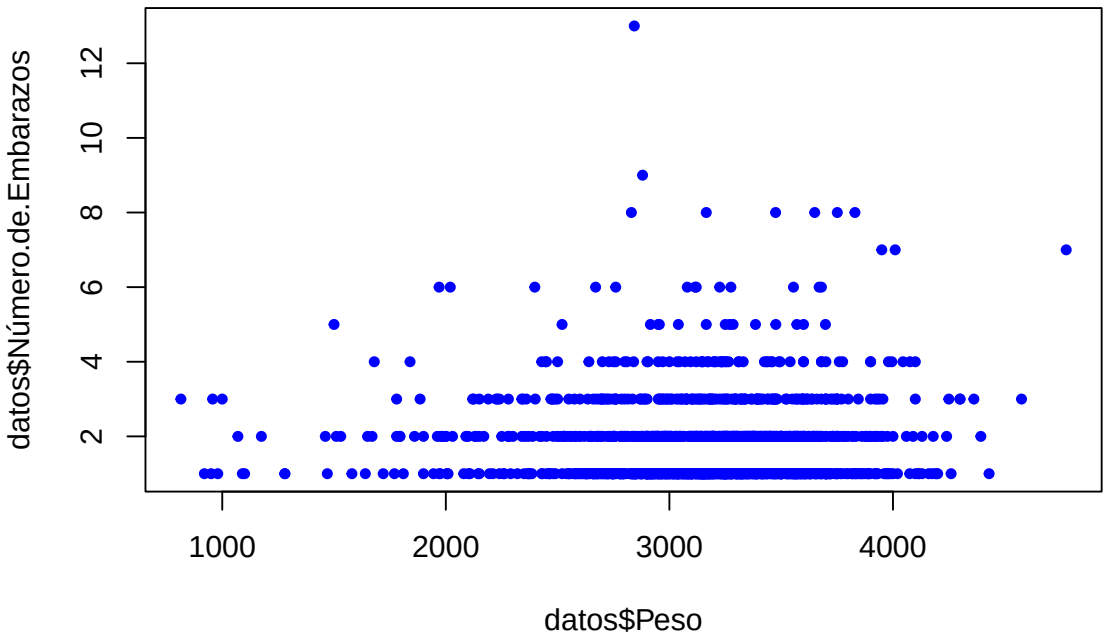


- Test de correlación del peso y el número de embarazos

```
cor.test(datos$Peso, datos$Número.de.Embarazos,method= "pearson" )
```

```
##
##  Pearson's product-moment correlation
##
## data:  datos$Peso and datos$Número.de.Embarazos
## t = 0.98942, df = 1430, p-value = 0.3226
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  -0.02568070  0.07785179
## sample estimates:
##        cor
## 0.02615569
```

Veamos que le valor p es mayor al 5%, entonces rechazamos el hecho de que existe correlación entre variables, más aun su coeficiente de correlación es muy cercano a cero, lo cual implica que definitivamente no deberíamos agregar esta variable a la regresión lineal. Además de manera similar al anterior indice no vemos ninguna relación entre los datos de manera gráfica.



Dados los anteriores tests de pearson entre variables, entonces deberíamos agregar las variables Talla, tiempo de gestación y Edad de la madre (Agregamos esta ultima variable, ya que la prueba de Pearson nos mostró correlación) al modelo lineal, es decir, quedaría de la forma

$$Peso = \beta_0 + \beta_T \text{ Talla} + \beta_G \text{ Tiempo de Gestación} + \beta_M \text{ Edad de la madre} + \varepsilon$$

Donde $\beta_0, \beta_T, \beta_G$ y β_M son los pesos de la regresión lineal y ε el error. El siguiente código es la regresión lineal múltiple en R.

```
regresion1= lm((datos$Peso ~ datos$Talla + datos$Tiempo.de.Gestación+ datos$Edad.de.la.Madre))
summary(regresion1)
```

```
##
## Call:
## lm(formula = (datos$Peso ~ datos$Talla + datos$Tiempo.de.Gestación +
##     datos$Edad.de.la.Madre))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1019.34  -204.72    -8.28   180.06  1861.91
##
## Coefficients:
##                              Estimate Std. Error t value Pr(>|t|)
## (Intercept)                -4907.840    173.193 -28.337  < 2e-16 ***
## datos$Talla                   96.323      3.563  27.035  < 2e-16 ***
## datos$Tiempo.de.Gestación     81.448      5.548  14.680  < 2e-16 ***
## datos$Edad.de.la.Madre         4.884      1.263   3.867 0.000115 ***
## ---
```

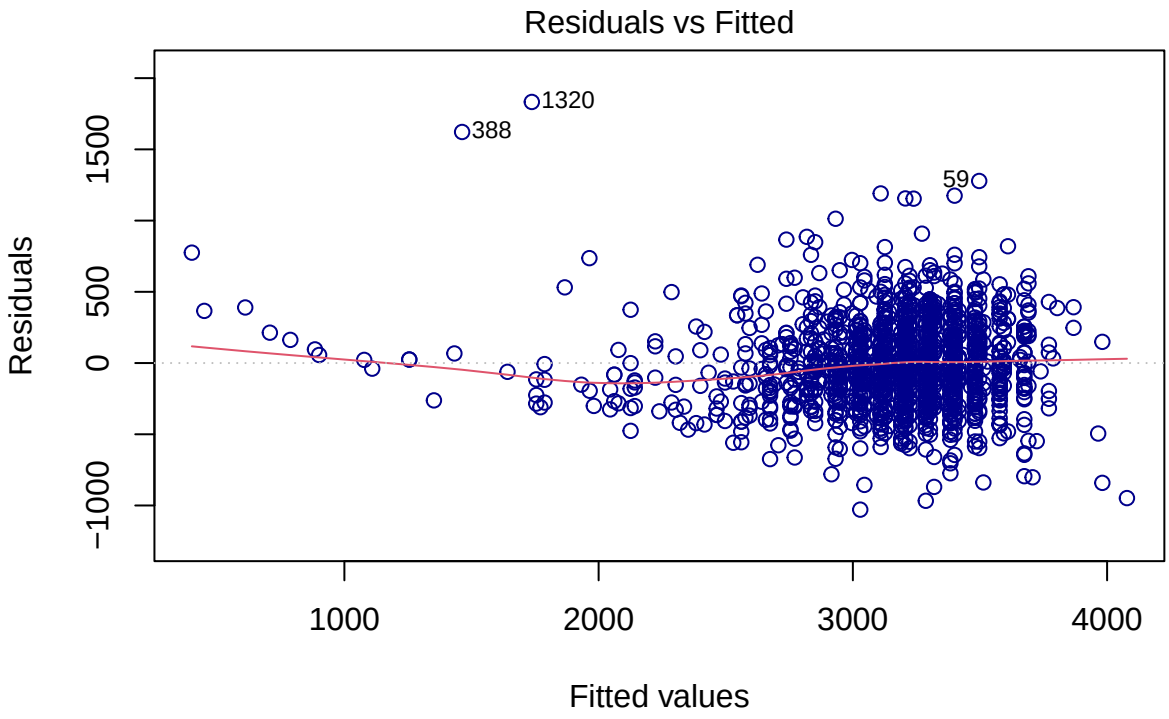
```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 301.6 on 1428 degrees of freedom
## Multiple R-squared:  0.6318, Adjusted R-squared:  0.631
## F-statistic: 816.7 on 3 and 1428 DF,  p-value: < 2.2e-16
```

Veamos que tenemos los pesos son $\beta_0 = -4907.84$, $\beta_T = 96.32, \beta_G = 81.44$ y $\beta_M = 4.88$, más aun veamos que R-squared el cual es 0.6318 los cual es un indicador de que el modelo predice de una buena manera. Sin embargo tratemos de hacer la regresión sin la variable "Edad de la madre".

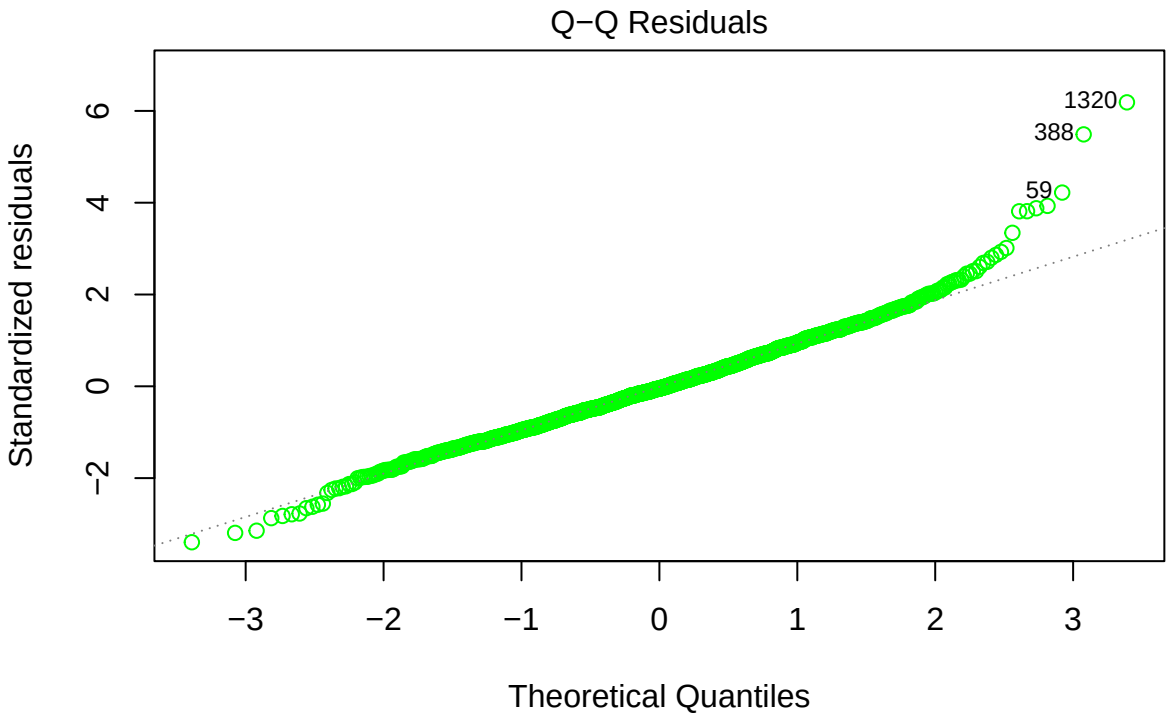
```
regresion2= lm(Peso ~ Talla + Tiempo.de.Gestación, data=datos)
summary(regresion2)
```

```
##
## Call:
## lm(formula = Peso ~ Talla + Tiempo.de.Gestación, data = datos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1029.05  -196.26   -15.52   189.45  1832.71
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -4776.415    170.652  -27.99   <2e-16 ***
## Talla                  96.853      3.578   27.07   <2e-16 ***
## Tiempo.de.Gestación    80.518      5.570   14.46   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 303 on 1429 degrees of freedom
## Multiple R-squared:  0.6279, Adjusted R-squared:  0.6274
## F-statistic:  1206 on 2 and 1429 DF,  p-value: < 2.2e-16
```

Notemos que $R^2 = 0.62$ lo cual comparándolo con el 0.63 de la anterior regresión, podríamos considerar quitar la variable "Edad de la madre", pues no cambia mucho el ajuste de la regresión lineal. La siguiente gráfica nos muestra con la linea roja que la media de los errores es cercana a cero.



De la siguiente gráfica vemos que la mayoría de los Residuales (errores) se acerca a la recta normal, a excepción de algunos (aunque esto ultimo es común en la practica) entonces podemos asumir normalidad en los errores (esto es un fundamento teórico, pues teóricamente $\varepsilon \sim N(0, \sigma^2)$).



Ahora teniendo el modelo, podemos predecir el peso del infante según su talla y el tiempo de gestación, supongamos que la talla sea 40 cm y 38 semanas de gestación

```
nuevos=data.frame(c(40),c(38))
colnames(nuevos)= c("Talla", "Tiempo.de.Gestación" )

predict(regresion2,newdata = nuevos)
```

```
##        1
## 2157.375
```

Según el modelo el infante debería pesar 2157 kilos aproximadamente. Sin embargo este tipo de predicciones es poco concluyente, es por esto que se recomienda realizar un intervalo de confianza para el valor dado, de la siguiente manera se muestra el intervalo de confianza para nuestro caso con un nivel de confianza del 95%.

```
predict(regresion2,newdata = nuevos, interval = "confidence")
```

```
##        fit      lwr      upr
## 1 2157.375 2088.188 2226.563
```

De lo anterior concluimos que el infante con talla 40 cm y 38 semanas de gestación, tendrá un peso entre 2088 Kg y 2226 Kg con una confianza del 95%.