Modelo de Regresión Lineal Robusta

2024-05-16

Por medio de este análisis se busca estimar la expectativa de vida en distintos países considerando el conjunto de variables proporcionadas por la Organización mundial de salud.

Fuente: https://www.kaggle.com/datasets/kumarajarshi/life-expectancy-who

Variables del set de datos

```
lifeExp <- read.csv("~/Downloads/lifeExp.csv")
lifeExp <- na.omit(lifeExp)
str(lifeExp)
```

```
## 'data.frame':    1649 obs. of  22 variables:
##  $ Country                        : chr  "Afghanistan" "Afghanistan" "Afghanistan" "Afghanistan" ...
##  $ Year                           : int  2015 2014 2013 2012 2011 2010 2009 2008 2007 2006 ...
##  $ Status                         : chr  "Developing" "Developing" "Developing" "Developing" ...
##  $ Life.expectancy                : num  65 59.9 59.9 59.5 59.2 58.8 58.6 58.1 57.5 57.3 ...
##  $ Adult.Mortality                : int  263 271 268 272 275 279 281 287 295 295 ...
##  $ infant.deaths                  : int  62 64 66 69 71 74 77 80 82 84 ...
##  $ Alcohol                        : num  0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.03 0.02 0.03 ...
##  $ percentage.expenditure         : num  71.3 73.5 73.2 78.2 7.1 ...
##  $ Hepatitis.B                    : int  65 62 64 67 68 66 63 64 63 64 ...
##  $ Measles                        : int  1154 492 430 2787 3013 1989 2861 1599 1141 1990 ...
##  $ BMI                            : num  19.1 18.6 18.1 17.6 17.2 16.7 16.2 15.7 15.2 14.7 ...
##  $ under.five.deaths              : int  83 86 89 93 97 102 106 110 113 116 ...
##  $ Polio                          : int  6 58 62 67 68 66 63 64 63 58 ...
##  $ Total.expenditure              : num  8.16 8.18 8.13 8.52 7.87 9.2 9.42 8.33 6.73 7.43 ...
##  $ Diphtheria                     : int  65 62 64 67 68 66 63 64 63 58 ...
##  $ HIV.AIDS                       : num  0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 0.1 ...
##  $ GDP                            : num  584.3 612.7 631.7 670 63.5 ...
##  $ Population                     : num  33736494 327582 31731688 3696958 2978599 ...
##  $ thinness..1.19.years           : num  17.2 17.5 17.7 17.9 18.2 18.4 18.6 18.8 19 19.2 ...
##  $ thinness.5.9.years             : num  17.3 17.5 17.7 18 18.2 18.4 18.7 18.9 19.1 19.3 ...
##  $ Income.composition.of.resources: num  0.479 0.476 0.47 0.463 0.454 0.448 0.434 0.433 0.415 0.405
##  $ Schooling                      : num  10.1 10 9.9 9.8 9.5 9.2 8.9 8.7 8.4 8.1 ...
##  - attr(*, "na.action")= 'omit' Named int [1:1289] 33 45 46 47 48 49 58 59 60 61 ...
##   ..- attr(*, "names")= chr [1:1289] "33" "45" "46" "47" ...
```

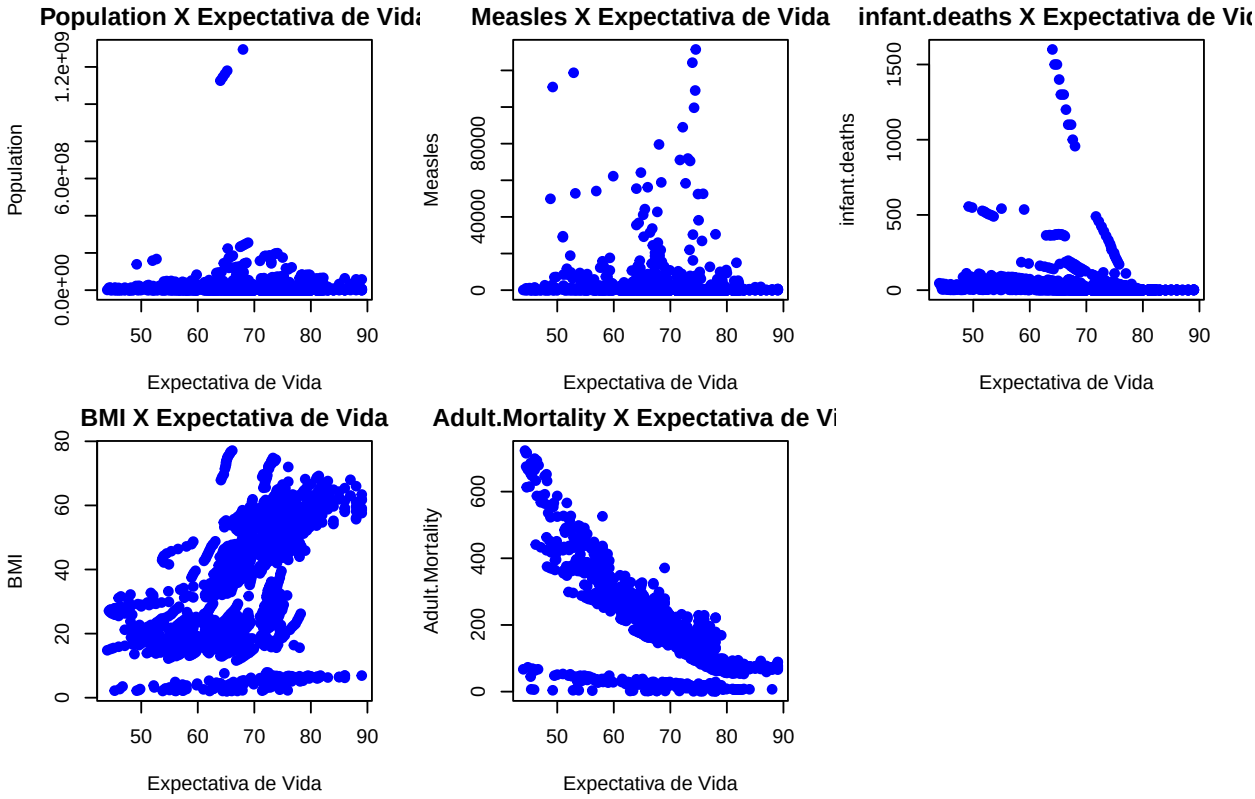
Gráficas de las variables selccionadas para el análisis

Para este análisis, se utilizarán variables que contienen valores atípicos (outliers), considerando el analisis expolotario previo a seleccionamos las variables "Population", "Measles", "infant.deaths" , "Adult.Mortality" y "BMI" para evaluar su comportamiento.

```r
par(mfrow = c(2, 3), mar = c(4, 4, 2, 2))

columnas <- c("Population", "Measles" , "infant.deaths" , "BMI", "Adult.Mortality")

for (col in columnas) {
  plot(lifeExp$Life.expectancy, lifeExp[[col]],
       main = paste(col, "X Expectativa de Vida"),
       xlab = "Expectativa de Vida",
       ylab = col,
       pch = 19,
       col = "blue")
}
```



Así ajustaremos un modelo inicial (el cual se ajustó el el trabajo previo a este) y se contrastaran con los resultados que nos proporcione el modelo de regresión robusta.

Aplicación del modelo robusto vs lineal

```r
library(robust)
```

```
## Loading required package: fit.models
```

```r
modelo1 <- lm(Life.expectancy ~ Population + Measles + infant.deaths + BMI + Adult.Mortality , data = li
summary(modelo1)
```

```
##
## Call:
## lm(formula = Life.expectancy ~ Population + Measles + infant.deaths +
##     BMI + Adult.Mortality, data = lifeExp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -29.0758  -2.4333   0.5662   3.0972  20.2374
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.100e+01  4.372e-01 162.397  < 2e-16 ***
## Population       9.726e-09  2.646e-09   3.676 0.000245 ***
## Measles          2.114e-05  1.611e-05   1.312 0.189705
## infant.deaths   -9.879e-03  1.754e-03  -5.633 2.08e-08 ***
## BMI              1.401e-01  7.665e-03  18.282  < 2e-16 ***
## Adult.Mortality -4.106e-02  1.171e-03 -35.052  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.565 on 1643 degrees of freedom
## Multiple R-squared:  0.601,  Adjusted R-squared:  0.5998
## F-statistic: 494.9 on 5 and 1643 DF,  p-value: < 2.2e-16
```

```
model_rob <- lmRob(Life.expectancy ~ Population + Measles + infant.deaths + BMI  + Adult.Mortality, data
```

```
## Warning in lmRob.fit.compute(x, y, x1.idx = x1.idx, nrep = nrep, robust.control
## = robust.control, : Max iteration for refinement reached.
```

```
summary(model_rob)
```

```
##
## Call:
## lmRob(formula = Life.expectancy ~ Population + Measles + infant.deaths +
##     BMI + Adult.Mortality, data = lifeExp)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -33.4100  -2.7738  -0.0217   1.8753  15.5382
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      7.732e+01  3.646e-01 212.028  < 2e-16 ***
## Population       8.145e-09  1.944e-09   4.191 2.92e-05 ***
## Measles          1.711e-05  1.080e-05   1.584    0.113
## infant.deaths   -1.181e-02  1.369e-03  -8.621  < 2e-16 ***
## BMI              7.551e-02  5.552e-03  13.602  < 2e-16 ***
## Adult.Mortality -5.609e-02  1.060e-03 -52.933  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.274 on 1643 degrees of freedom
## Multiple R-Squared: 0.501
```

```
##
## Test for Bias:
##             statistic  p-value
## M-estimate      47.91 1.23e-08
## LS-estimate    182.92 0.00e+00
```

```r
calculate_aic <- function(model) {
  n <- length(model$residuals)
  k <- length(coef(model))
  rss <- sum(residuals(model)^2)
  aic <- n * log(rss / n) + 2 * k
  return(aic)
}

calculate_aic(modelo1)
```

```
## [1] 5667.078
```

```r
calculate_aic(model_rob)
```

```
## [1] 5913.201
```

Claramente, en este punto, el modelo lineal muestra el menor AIC. Ahora, investiguemos si el mejor modelo seleccionado mediante los criterios de best subset, backward y forward es capaz de superar o reducir considerablemente el AIC para el caso robusto. A continuación se crea manualmente una función para realizar la selección del mejor subconjunto de características, considerando que el método robusto no es aplicable bajo funciones de otros modelos como lo son los lineales.

Forward

```r
# Función para ajustar cuatro modelos distintos y mostrar sus AIC
compare_models <- function(data) {
  # Definir las combinaciones de variables para los cuatro modelos
  formulas <- list(
    Life.expectancy ~ Measles,
    Life.expectancy ~ BMI + Population,
    Life.expectancy ~ Adult.Mortality + infant.deaths,
    Life.expectancy ~ Measles + Adult.Mortality + BMI,
    Life.expectancy ~ Population + Measles + infant.deaths
  )

  # Ajustar los modelos y calcular los AIC
  for (i in 1:length(formulas)) {
    model <- lmRob(formulas[[i]], data = data)
    aic <- calculate_aic(model)
    cat("Model", i, "formula:", deparse(formulas[[i]]), "\n")
    cat("AIC:", aic, "\n\n")
  }
}
```

```r
# Aplicar la función a los datos
compare_models(lifeExp)
```

```
## Model 1 formula: Life.expectancy ~ Measles
## AIC: 7186.502
##
## Model 2 formula: Life.expectancy ~ BMI + Population
## AIC: 6609.141
##
## Model 3 formula: Life.expectancy ~ Adult.Mortality + infant.deaths
## AIC: 6148.942
##
## Model 4 formula: Life.expectancy ~ Measles + Adult.Mortality + BMI
## AIC: 5939.501
##
## Model 5 formula: Life.expectancy ~ Population + Measles + infant.deaths
## AIC: 10923.05
```

Selección hacia Adelante (Forward Selection)

Descripción

La selección hacia adelante es un método de construcción de modelos que comienza con un modelo vacío y agrega variables predictor una por una. En cada paso, se elige la variable que más reduce el criterio de información de Akaike (AIC). Este proceso continúa hasta que no se puede mejorar significativamente el AIC agregando más variables.

Proceso

1. **Inicio:** Comienza con un modelo que solo incluye el intercepto.
2. **Agregar Variables:** Iterativamente agrega la variable que más reduce el AIC.
3. **Criterio de Parada:** El proceso se detiene cuando agregar nuevas variables no reduce significativamente el AIC.

Ejemplo

Para los modelos considerados, un posible resultado de la selección hacia adelante podría ser:

1. **Model 1:** `Life.expectancy ~ Measles` (AIC: 7186.502)
2. **Model 4:** `Life.expectancy ~ Measles + Adult.Mortality + BMI` (AIC: 5939.501)

Selección hacia Atrás (Backward Selection)

Descripción

La selección hacia atrás es un método de reducción de modelos que comienza con un modelo completo, incluyendo todas las variables predictoras, y elimina variables una por una. En cada paso, se elimina la variable cuya eliminación cause la menor pérdida de ajuste, medida por el AIC. Este proceso continúa hasta que la eliminación de más variables no mejora el AIC significativamente.

Proceso

1. **Inicio:** Comienza con un modelo que incluye todas las variables.
2. **Eliminar Variables:** Iterativamente elimina la variable cuya eliminación menos aumenta el AIC.
3. **Criterio de Parada:** El proceso se detiene cuando eliminar más variables aumenta significativamente el AIC.

Ejemplo

Para los modelos considerados, un posible resultado de la selección hacia atrás podría ser:

1. **Model 5:** `Life.expectancy ~ Population + Measles + infant.deaths` (AIC: 10923.05)
2. **Model 2:** `Life.expectancy ~ BMI + Population` (AIC: 6609.141)

Mejor Subconjunto (Best Subset Selection)

Descripción

La selección del mejor subconjunto es un método exhaustivo que evalúa todos los posibles modelos de diferentes tamaños y selecciona el modelo con el AIC más bajo. Este método garantiza encontrar el mejor modelo en términos de AIC, pero puede ser computacionalmente intensivo, especialmente con muchas variables.

Proceso

1. **Generar Subconjuntos:** Considera todos los posibles modelos, desde aquellos con una sola variable hasta aquellos con todas las variables.
2. **Evaluar Modelos:** Calcula el AIC para cada modelo.
3. **Seleccionar el Mejor:** Elige el modelo con el AIC más bajo.

Ejemplo

Para los modelos considerados, los AIC calculados son:

1. **Model 1:** `Life.expectancy ~ Measles` (AIC: 7186.502)
2. **Model 2:** `Life.expectancy ~ BMI + Population` (AIC: 6609.141)
3. **Model 3:** `Life.expectancy ~ Adult.Mortality + infant.deaths` (AIC: 6148.942)
4. **Model 4:** `Life.expectancy ~ Measles + Adult.Mortality + BMI` (AIC: 5939.501)
5. **Model 5:** `Life.expectancy ~ Population + Measles + infant.deaths` (AIC: 10923.05)

Conclusión

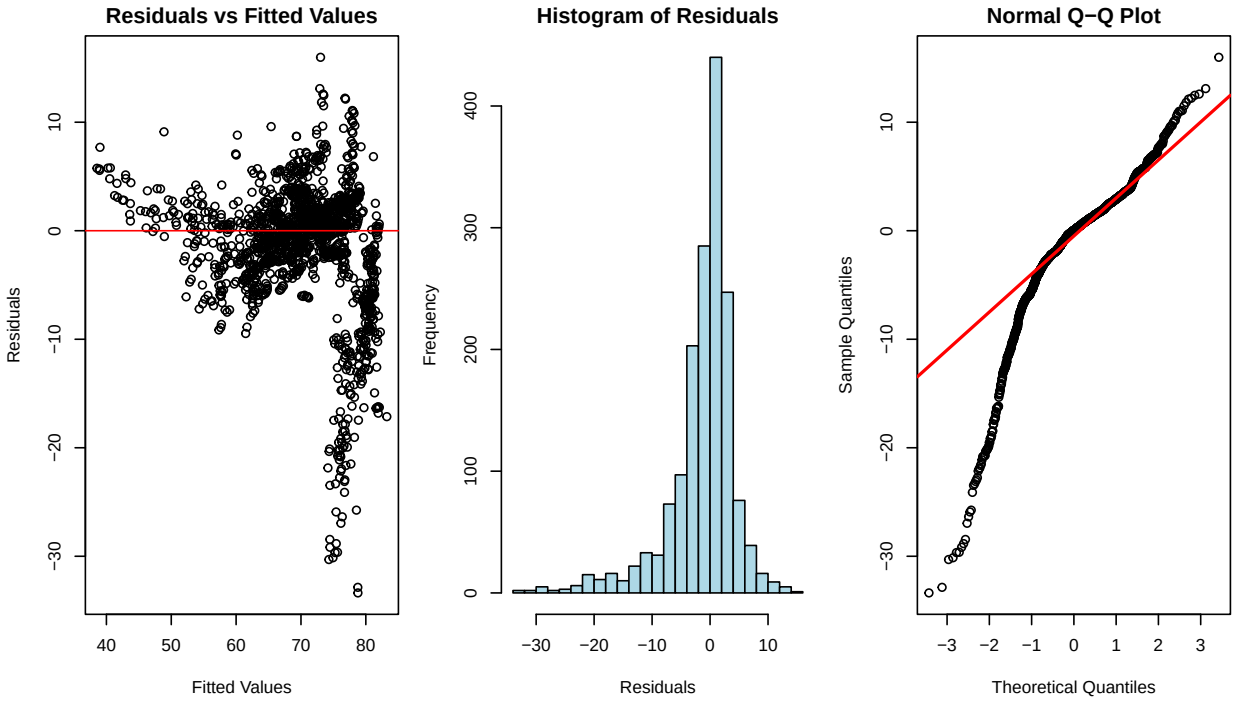
El modelo que presentó el mejor desempeño según el criterio AIC es el **Model 4: Life.expectancy ~ Measles + Adult.Mortality + BMI con un AIC de 5939.501.** Este modelo superó significativamente en términos de AIC a otros modelos evaluados, incluidos aquellos ajustados con métodos menos robustos.

La menor influencia de los outliers en este modelo robusto permitió capturar mejor las tendencias subyacentes en los datos, reflejando de manera más precisa las relaciones entre las variables predictoras y la esperanza de vida.

```r
model_rob <- lmRob(Life.expectancy ~ Measles + Adult.Mortality + BMI, data = lifeExp)
par(mfrow = c(1, 3), mar = c(4, 4, 2, 1))

residuals_rob <- residuals(model_rob)
fitted_values <- fitted(model_rob)
plot(fitted_values, residuals_rob,
     main = "Residuals vs Fitted Values",
     xlab = "Fitted Values",
     ylab = "Residuals")
abline(h = 0, col = "red")

hist(residuals_rob, breaks = 20,
     main = "Histogram of Residuals",
     xlab = "Residuals",
     col = "lightblue")
qqnorm(residuals_rob)
qqline(residuals_rob, col = "red", lwd = 2)
```



El análisis de residuales del modelo robusto revela algunas características importantes. En el gráfico de **Residuales vs Valores Ajustados**, se observa *heterocedasticidad*, ya que la variabilidad de los residuales no es constante a lo largo de los valores ajustados, y hay patrones estructurados que sugieren que el modelo puede no estar capturando todas las características de los datos adecuadamente. El Histograma de Residuales muestra **una distribución aproximadamente normal, aunque con colas pesadas que indican la presencia de outliers**. Finalmente, el Gráfico Q-Q de Residuales evidencia desviaciones de la normalidad en las colas, confirmando la presencia de outliers. **A pesar de que el modelo robusto mejora el ajuste en presencia de estos outliers, la heterocedasticidad y la variabilidad no capturada sugieren que podrían ser necesarias transformaciones adicionales de las variables o una investigación más profunda de los outliers para mejorar la capacidad predictiva del modelo.**