Multiple Linear Regression

Live Expectancy Between the States Model

April 24th, 2020

Nicole Del Villar

Abstract

This study wanted to demonstrate how to generate a multiple linear regression model that allowed predicting the average life expectancy of the habitants of the different states on different variables. The information was available on population, income, life expectancy, murder, HG. Grad, frost, area, and population density which was calculated to more accuracy in the results.

Introduction

The *State.x77* data set from R collected by the U.S. Bureau of the Census in the 1970s contained data of the 50 states. This data included 8 variables such as population, per capita income, illiteracy, life expectancy, murder rate, high school graduates, number of frost days (days with minimal temperature below freezing in the capital or large cities between 1931-1060), and land area in square per miles for each of the states. This dataset was used to fit a multiple linear regression model by analyzing the relationship between the variables, generating a model with all variables, selecting the best predictors, validating the conditions of a linear regression model, and identifying possible outliers.

Methods

The analysis started by establishing a relationship between the variables. This was to identify which may be the best predictors for the model, which variables had non-linear relationship, so they cannot be included, and to identify collinearity between the predictors. Also, the variable population had been calculated by 1000 habitants per area to create another variable called Density with the purpose of calculating the population density.

> library(dplyr)

> data(state)

> state.x77

Population Income Illiteracy Life Exp Murder HS Grad Frost Area

Alabama 3615 3624 2.1 69.05 15.1 41.3 20 50708

Alaska 365 6315 1.5 69.31 11.3 66.7 152 566432

Arizona 2212 4530 1.8 70.55 7.8 58.1 15 113417

Arkansas 2110 3378 1.9 70.66 10.1 39.9 65 51945

California 21198 5114 1.1 71.71 10.3 62.6 20 156361

Colorado 2541 4884 0.7 72.06 6.8 63.9 166 103766

Connecticut 3100 5348 1.1 72.48 3.1 56.0 139 4862

Delaware 579 4809 0.9 70.06 6.2 54.6 103 1982

Florida 8277 4815 1.3 70.66 10.7 52.6 11 54090

Georgia 4931 4091 2.0 68.54 13.9 40.6 60 58073

Hawaii 868 4963 1.9 73.60 6.2 61.9 0 6425

Idaho 813 4119 0.6 71.87 5.3 59.5 126 82677

Illinois 11197 5107 0.9 70.14 10.3 52.6 127 55748

Indiana 5313 4458 0.7 70.88 7.1 52.9 122 36097

Iowa 2861 4628 0.5 72.56 2.3 59.0 140 55941

Kansas 2280 4669 0.6 72.58 4.5 59.9 114 81787

Kentucky 3387 3712 1.6 70.10 10.6 38.5 95 39650

Louisiana 3806 3545 2.8 68.76 13.2 42.2 12 44930

Maine 1058 3694 0.7 70.39 2.7 54.7 161 30920

Maryland 4122 5299 0.9 70.22 8.5 52.3 101 9891

Massachusetts 5814 4755 1.1 71.83 3.3 58.5 103 7826

Michigan 9111 4751 0.9 70.63 11.1 52.8 125 56817

Minnesota 3921 4675 0.6 72.96 2.3 57.6 160 79289

Mississippi 2341 3098 2.4 68.09 12.5 41.0 50 47296

Missouri 4767 4254 0.8 70.69 9.3 48.8 108 68995

Montana 746 4347 0.6 70.56 5.0 59.2 155 145587

Nebraska 1544 4508 0.6 72.60 2.9 59.3 139 76483

Nevada 590 5149 0.5 69.03 11.5 65.2 188 109889

New Hampshire 812 4281 0.7 71.23 3.3 57.6 174 9027

New Jersey 7333 5237 1.1 70.93 5.2 52.5 115 7521

New Mexico 1144 3601 2.2 70.32 9.7 55.2 120 121412

New York 18076 4903 1.4 70.55 10.9 52.7 82 47831

North Carolina 5441 3875 1.8 69.21 11.1 38.5 80 48798

North Dakota 637 5087 0.8 72.78 1.4 50.3 186 69273

Ohio 10735 4561 0.8 70.82 7.4 53.2 124 40975

Oklahoma 2715 3983 1.1 71.42 6.4 51.6 82 68782

Oregon 2284 4660 0.6 72.13 4.2 60.0 44 96184

Pennsylvania 11860 4449 1.0 70.43 6.1 50.2 126 44966

Rhode Island 931 4558 1.3 71.90 2.4 46.4 127 1049

South Carolina 2816 3635 2.3 67.96 11.6 37.8 65 30225

South Dakota 681 4167 0.5 72.08 1.7 53.3 172 75955

Tennessee 4173 3821 1.7 70.11 11.0 41.8 70 41328

Texas 12237 4188 2.2 70.90 12.2 47.4 35 262134

Utah 1203 4022 0.6 72.90 4.5 67.3 137 82096

Vermont 472 3907 0.6 71.64 5.5 57.1 168 9267

Virginia 4981 4701 1.4 70.08 9.5 47.8 85 39780

Washington 3559 4864 0.6 71.72 4.3 63.5 32 66570

West Virginia 1799 3617 1.4 69.48 6.7 41.6 100 24070

Wisconsin 4589 4468 0.7 72.48 3.0 54.5 149 54464

Wyoming 376 4566 0.6 70.29 6.9 62.9 173 97203

> dat <- as.data.frame(state.x77)

> head(dat)

Population Income Illiteracy Life Exp Murder HS Grad Frost Area

Alabama 3615 3624 2.1 69.05 15.1 41.3 20 50708

Alaska 365 6315 1.5 69.31 11.3 66.7 152 566432

Arizona 2212 4530 1.8 70.55 7.8 58.1 15 113417

Arkansas 2110 3378 1.9 70.66 10.1 39.9 65 51945

California 21198 5114 1.1 71.71 10.3 62.6 20 156361

Colorado 2541 4884 0.7 72.06 6.8 63.9 166 103766

> dat$Density <- dat$Population \* 1000 / dat$Area

> round(cor(x = dat, method = "pearson"), 3)

Population Income Illiteracy Life Exp Murder HS Grad Frost Area Density

Population 1.000 0.208 0.108 -0.068 0.344 -0.098 -0.332 0.023 0.246

Income 0.208 1.000 -0.437 0.340 -0.230 0.620 0.226 0.363 0.330

Illiteracy 0.108 -0.437 1.000 -0.588 0.703 -0.657 -0.672 0.077 0.009

Life Exp -0.068 0.340 -0.588 1.000 -0.781 0.582 0.262 -0.107 0.091

Murder 0.344 -0.230 0.703 -0.781 1.000 -0.488 -0.539 0.228 -0.185

HS Grad -0.098 0.620 -0.657 0.582 -0.488 1.000 0.367 0.334 -0.088

Frost -0.332 0.226 -0.672 0.262 -0.539 0.367 1.000 0.059 0.002

Area 0.023 0.363 0.077 -0.107 0.228 0.334 0.059 1.000 -0.341

Density 0.246 0.330 0.009 0.091 -0.185 -0.088 0.002 -0.341 1.000

The representation of the distribution of each variable was shown in the graphs below:

> library(psych)

> multi.hist(x = dat, dcol = c("blue", "red"), dlty = c("dotted", "solid"),

+ main = "")

A close up of a map

Description automatically generated

> library(GGally)

> ggpairs(dat, lower = list(continuous = "smooth"),

+ diag = list(continuous = "bar"), axisLabels = "none")

A close up of text on a black background

Description automatically generated

The variables that had the greater linear relationship with life expectancy were murder r = -0.78, illiteracy r = -0.59, and HS. Grad r = 0.58. Murder and illiteracy were moderate correlated with r = 0.7, so it was probably not useful to introduce both predictors in the model. The variables population, area, and density showed an exponential distribution.

To generate the model, a recreation of one model that included all the variables as predictors was generated before selecting the best ones. It was necessary to remove the space in the variable names Life expectancy and HS Grad to avoid error messages.

> colnames(dat)[4] <- "Life.Exp"

> colnames(dat)[6] <- "HS.Grad"

> model <- lm(Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad + Frost + Area + Density, data = dat)

> summary(model)

Call:

lm(formula = Life.Exp ~ Population + Income + Illiteracy + Murder +

HS.Grad + Frost + Area + Density, data = dat)

Residuals:

Min 1Q Median 3Q Max

-1.47514 -0.45887 -0.06352 0.59362 1.21823

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.995e+01 1.843e+00 37.956 < 2e-16 \*\*\*

Population 6.480e-05 3.001e-05 2.159 0.0367 \*

Income 2.701e-04 3.087e-04 0.875 0.3867

Illiteracy 3.029e-01 4.024e-01 0.753 0.4559

Murder -3.286e-01 4.941e-02 -6.652 5.12e-08 \*\*\*

HS.Grad 4.291e-02 2.332e-02 1.840 0.0730 .

Frost -4.580e-03 3.189e-03 -1.436 0.1585

Area -1.558e-06 1.914e-06 -0.814 0.4205

Density -1.105e-03 7.312e-04 -1.511 0.1385

---

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

Residual standard error: 0.7337 on 41 degrees of freedom

Multiple R-squared: 0.7501, Adjusted R-squared: 0.7013

F-statistic: 15.38 on 8 and 41 DF, p-value: 3.787e-10

To proceed to select the best predictors, the measurement *Akaike* (AIC) was used with the stepwise criteria. The AIC method was more restrictive and introduced less predictors than the adjusted R^2.

> step(object = model, direction = "both", trace = 1)

Start: AIC=-22.89

Life.Exp ~ Population + Income + Illiteracy + Murder + HS.Grad +

Frost + Area + Density

Df Sum of Sq RSS AIC

- Illiteracy 1 0.3050 22.373 -24.208

- Area 1 0.3564 22.425 -24.093

- Income 1 0.4120 22.480 -23.969

<none> 22.068 -22.894

- Frost 1 1.1102 23.178 -22.440

- Density 1 1.2288 23.297 -22.185

- HS.Grad 1 1.8225 23.891 -20.926

- Population 1 2.5095 24.578 -19.509

- Murder 1 23.8173 45.886 11.707

Step: AIC=-24.21

Life.Exp ~ Population + Income + Murder + HS.Grad + Frost + Area +

Density

Df Sum of Sq RSS AIC

- Area 1 0.1427 22.516 -25.890

- Income 1 0.2316 22.605 -25.693

<none> 22.373 -24.208

- Density 1 0.9286 23.302 -24.174

- HS.Grad 1 1.5218 23.895 -22.918

+ Illiteracy 1 0.3050 22.068 -22.894

- Population 1 2.2047 24.578 -21.509

- Frost 1 3.1324 25.506 -19.656

- Murder 1 26.7071 49.080 13.072

Step: AIC=-25.89

Life.Exp ~ Population + Income + Murder + HS.Grad + Frost + Density

Df Sum of Sq RSS AIC

- Income 1 0.132 22.648 -27.598

- Density 1 0.786 23.302 -26.174

<none> 22.516 -25.890

- HS.Grad 1 1.424 23.940 -24.824

+ Area 1 0.143 22.373 -24.208

+ Illiteracy 1 0.091 22.425 -24.093

- Population 1 2.332 24.848 -22.962

- Frost 1 3.304 25.820 -21.043

- Murder 1 32.779 55.295 17.033

Step: AIC=-27.6

Life.Exp ~ Population + Murder + HS.Grad + Frost + Density

Df Sum of Sq RSS AIC

- Density 1 0.660 23.308 -28.161

<none> 22.648 -27.598

+ Income 1 0.132 22.516 -25.890

+ Illiteracy 1 0.061 22.587 -25.732

+ Area 1 0.043 22.605 -25.693

- Population 1 2.659 25.307 -24.046

- Frost 1 3.179 25.827 -23.030

- HS.Grad 1 3.966 26.614 -21.529

- Murder 1 33.626 56.274 15.910

Step: AIC=-28.16

Life.Exp ~ Population + Murder + HS.Grad + Frost

Df Sum of Sq RSS AIC

<none> 23.308 -28.161

+ Density 1 0.660 22.648 -27.598

+ Income 1 0.006 23.302 -26.174

+ Illiteracy 1 0.004 23.304 -26.170

+ Area 1 0.001 23.307 -26.163

- Population 1 2.064 25.372 -25.920

- Frost 1 3.122 26.430 -23.877

- HS.Grad 1 5.112 28.420 -20.246

- Murder 1 34.816 58.124 15.528

Call:

lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,

data = dat)

Coefficients:

(Intercept) Population Murder HS.Grad Frost

7.103e+01 5.014e-05 -3.001e-01 4.658e-02 -5.943e-03

The best resulting model for the process had been the one below with the resulting AIC = -28.16. The model was called “modelo” to differentiated it from the original model that included all the variables.

> modelo <-lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,

+ data = dat)

> summary(modelo)

Call:

lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost,

data = dat)

Residuals:

Min 1Q Median 3Q Max

-1.47095 -0.53464 -0.03701 0.57621 1.50683

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 7.103e+01 9.529e-01 74.542 < 2e-16 \*\*\*

Population 5.014e-05 2.512e-05 1.996 0.05201 .

Murder -3.001e-01 3.661e-02 -8.199 1.77e-10 \*\*\*

HS.Grad 4.658e-02 1.483e-02 3.142 0.00297 \*\*

Frost -5.943e-03 2.421e-03 -2.455 0.01802 \*

---

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

Residual standard error: 0.7197 on 45 degrees of freedom

Multiple R-squared: 0.736, Adjusted R-squared: 0.7126

F-statistic: 31.37 on 4 and 45 DF, p-value: 1.696e-12

> confint(modelo)

2.5 % 97.5 %

(Intercept) 6.910798e+01 72.9462729104

Population -4.543308e-07 0.0001007343

Murder -3.738840e-01 -0.2264135705

HS.Grad 1.671901e-02 0.0764454870

Frost -1.081918e-02 -0.0010673977

For each unit that increased HS. Grad, the life expectancy increased on average 0.04658 units, keeping all the predictors constant.

The graphs below demonstrated the validation of the conditions for a normal distribution.

> qqnorm(model$residuals)

> qqline(modelo$residuals)

A close up of a map

Description automatically generated

The graph above showed that residuals were normally distributed.

> ggplot(data = dat, aes(model$fitted.values, model$residuals)) +

+ geom\_point() +

+ geom\_smooth(color = "firebrick", se = FALSE) +

+ geom\_hline(yintercept = 0) +

+ theme\_bw()

`geom\_smooth()` using method = 'loess' and formula 'y ~ x'

A picture containing text, outdoor, photo, old

Description automatically generated

In the graph above, the residuals were equally spread around the horizontal line without a very different pattern, there was not non-linear relationship.

To identify possible outliers, the Breusch-Pagan test was necessary to fit the linear regression model to the residuals of the model.

> library(lmtest)

> bptest(modelo)

studentized Breusch-Pagan test

data: modelo

BP = 6.2721, df = 4, p-value = 0.1797

There was no evidence of lack of homoscedasticity, so the variance around the regression line was the same for all the values of the predictor variable. This was showed in the graph below.

> datos$studentized\_residual <- rstudent(modelo)

> dat$studentized\_residual <- rstudent(modelo)

> ggplot(data = dat, aes(x = predict(modelo), y = abs(studentized\_residual))) +

+ geom\_hline(yintercept = 3, color = "grey", linetype = "dashed") +

+ geom\_point(aes(color = ifelse(abs(studentized\_residual) > 3, 'red', 'black'))) +

+ scale\_color\_identity() +

+ labs(title = "Distribution of Studentized Residuals",

+ x = "modelo prdiction") +

+ theme\_bw() + theme(plot.title = element\_text(hjust = 0.5))

A picture containing photo, bird, white, side

Description automatically generated

> which(abs(datos$studentized\_residual) > 3)

integer(0)

No outliers were identified.

The table below contains observations that were significantly influencing at least one predictor.

> summary(influence.measures(modelo))

Potentially influential observations of

lm(formula = Life.Exp ~ Population + Murder + HS.Grad + Frost, data = dat) :

dfb.1\_ dfb.Pplt dfb.Mrdr dfb.HS.G dfb.Frst dffit cov.r cook.d hat

Alaska 0.41 0.18 -0.40 -0.35 -0.16 -0.50 1.36\_\* 0.05 0.25

California 0.04 -0.09 0.00 -0.04 0.03 -0.12 1.81\_\* 0.00 0.38\_\*

Hawaii -0.03 -0.57 -0.28 0.66 -1.24\_\* 1.43\_\* 0.74 0.36 0.24

Nevada 0.40 0.14 -0.42 -0.29 -0.28 -0.52 1.46\_\* 0.05 0.29

New York 0.01 -0.06 0.00 0.00 -0.01 -0.07 1.44\_\* 0.00 0.23

The result was showed in the graph below:

> library(car)

> influencePlot(modelo)

A close up of a map

Description automatically generated

StudRes Hat CookD

California -0.1500614 0.38475924 0.002879053

Hawaii 2.5430162 0.23979244 0.363778638

Maine -2.2012995 0.06424817 0.061301962

Nevada -0.8120831 0.28860921 0.053917754

Washington -1.4895722 0.17168830 0.089555784

The analysis showed that several influential observations such as California and Hawaii, exceed the limit of concern for the leverage or cook distance value. More exhaustive studies would consist of redoing the model without the observation and seeing the impact.

Conclusion

The multiple linear regression model Life Exp = Population + Murder + HS. Grad + Frost was capable to explain 73% of the variability observed in life expectancy (R^2:0.736, R^2 Adjusted: 0.7126). The F test demonstrated that was significant (p-value:1,696e-12). All the conditions for this multiple linear regression were satisfied. Two observations California and Hawaii could be notably influencing the model*.*

Literature Cited

Queensborough Lab.

<http://www.simonqueenborough.info/R/statistics/multiple-regression>

R Documentation.

<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/step>

R Documentation.

<http://finzi.psych.upenn.edu/R/library/lmtest/html/bptest.html>

R Documentation

<https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/influence.measures>

U.S. Department of Commerce, Bureau of the Census (1977) Statistical Abstract of the United States, and U.S. Department of Commerce, Bureau of the Census (1977) County and City Data Book.