Health Insurence

Sudipta Das

2022-10-01

# Data Description

## Content

### Columns

age: age of primary beneficiary  
  
 sex: insurance contractor gender, female, male  
  
 bmi: Body mass index, providing an understanding of body, weights that are relatively high or low relative to height, objective index of body weight (kg / m ^ 2) using the ratio of height to weight, ideally 18.5 to 24.9  
  
 children: Number of children covered by health insurance / Number of dependents  
  
 smoker: Smoking  
  
 region: the beneficiary's residential area in the US, northeast, southeast, southwest, northwest.  
  
 charges: Individual medical costs billed by health insurance

# Calling some libraries  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

library(Hmisc)

## Loading required package: lattice

## Loading required package: survival

## Loading required package: Formula

##   
## Attaching package: 'Hmisc'

## The following objects are masked from 'package:dplyr':  
##   
## src, summarize

## The following objects are masked from 'package:base':  
##   
## format.pval, units

library(cowplot)  
library(WVPlots)

## Loading required package: wrapr

##   
## Attaching package: 'wrapr'

## The following object is masked from 'package:dplyr':  
##   
## coalesce

library(car)

## Loading required package: carData

##   
## Attaching package: 'car'

## The following object is masked from 'package:wrapr':  
##   
## bc

## The following object is masked from 'package:dplyr':  
##   
## recode

library(caTools)

# Calling my dataset  
  
data=read.csv("C:/Users/Sudipta/Downloads/insurance.csv")

head(data)

## age sex bmi children smoker region charges  
## 1 19 female 27.900 0 yes southwest 16884.924  
## 2 18 male 33.770 1 no southeast 1725.552  
## 3 28 male 33.000 3 no southeast 4449.462  
## 4 33 male 22.705 0 no northwest 21984.471  
## 5 32 male 28.880 0 no northwest 3866.855  
## 6 31 female 25.740 0 no southeast 3756.622

describe(data)

## data   
##   
## 7 Variables 1338 Observations  
## --------------------------------------------------------------------------------  
## age   
## n missing distinct Info Mean Gmd .05 .10   
## 1338 0 47 0.999 39.21 16.21 18 19   
## .25 .50 .75 .90 .95   
## 27 39 51 59 62   
##   
## lowest : 18 19 20 21 22, highest: 60 61 62 63 64  
## --------------------------------------------------------------------------------  
## sex   
## n missing distinct   
## 1338 0 2   
##   
## Value female male  
## Frequency 662 676  
## Proportion 0.495 0.505  
## --------------------------------------------------------------------------------  
## bmi   
## n missing distinct Info Mean Gmd .05 .10   
## 1338 0 548 1 30.66 6.893 21.26 22.99   
## .25 .50 .75 .90 .95   
## 26.30 30.40 34.69 38.62 41.11   
##   
## lowest : 15.960 16.815 17.195 17.290 17.385, highest: 48.070 49.060 50.380 52.580 53.130  
## --------------------------------------------------------------------------------  
## children   
## n missing distinct Info Mean Gmd   
## 1338 0 6 0.899 1.095 1.275   
##   
## lowest : 0 1 2 3 4, highest: 1 2 3 4 5  
##   
## Value 0 1 2 3 4 5  
## Frequency 574 324 240 157 25 18  
## Proportion 0.429 0.242 0.179 0.117 0.019 0.013  
## --------------------------------------------------------------------------------  
## smoker   
## n missing distinct   
## 1338 0 2   
##   
## Value no yes  
## Frequency 1064 274  
## Proportion 0.795 0.205  
## --------------------------------------------------------------------------------  
## region   
## n missing distinct   
## 1338 0 4   
##   
## Value northeast northwest southeast southwest  
## Frequency 324 325 364 325  
## Proportion 0.242 0.243 0.272 0.243  
## --------------------------------------------------------------------------------  
## charges   
## n missing distinct Info Mean Gmd .05 .10   
## 1338 0 1337 1 13270 12301 1758 2347   
## .25 .50 .75 .90 .95   
## 4740 9382 16640 34832 41182   
##   
## lowest : 1121.874 1131.507 1135.941 1136.399 1137.011  
## highest: 55135.402 58571.074 60021.399 62592.873 63770.428  
## --------------------------------------------------------------------------------

### In our dataset there are some catagorical variables present. We want to convert them into numerical

data$sex=factor(data$sex,levels = c("male","female"),labels = c(0,1))

data$region=factor(data$region,levels=c("northeast","northwest","southeast","southwest"),labels = c(0,1,2,3))

data$smoker=factor(data$smoker,levels = c("no","yes"),labels=c(0,1))

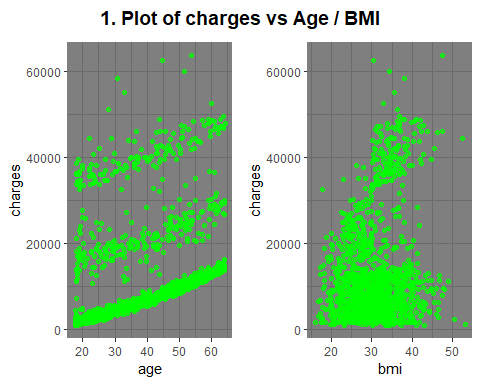
### Let’s see our modified dataset

head(data)

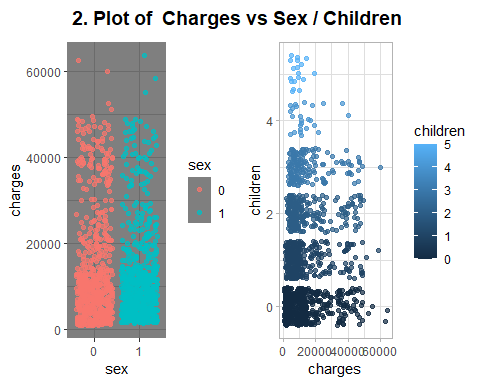
## age sex bmi children smoker region charges  
## 1 19 1 27.900 0 1 3 16884.924  
## 2 18 0 33.770 1 0 2 1725.552  
## 3 28 0 33.000 3 0 2 4449.462  
## 4 33 0 22.705 0 0 1 21984.471  
## 5 32 0 28.880 0 0 1 3866.855  
## 6 31 1 25.740 0 0 2 3756.622

### EDA

age<- ggplot(data, aes(age, charges)) +  
 geom\_jitter(color = "green", alpha = 0.7) +  
 theme\_dark()  
  
bmi <- ggplot(data, aes(bmi, charges)) +  
 geom\_jitter(color = "green", alpha = 0.7) +  
 theme\_dark()  
  
p <- plot\_grid(age, bmi)   
title <- ggdraw() + draw\_label("1. Plot of charges vs Age / BMI", fontface='bold')  
plot\_grid(title, p, ncol=1, rel\_heights=c(0.1, 1))

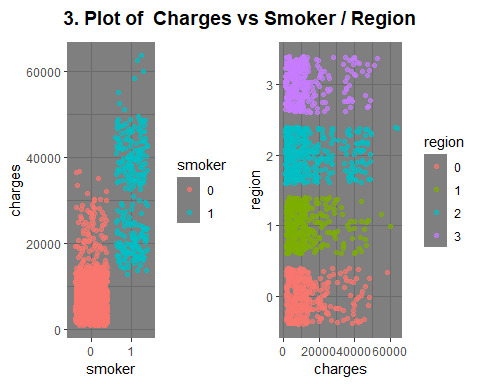


sex <- ggplot(data, aes(sex, charges)) +  
 geom\_jitter(aes(color = sex), alpha = 0.7) +  
 theme\_dark()  
  
children <- ggplot(data, aes(charges,children)) +  
 geom\_jitter(aes(color = children), alpha = 0.7) +  
 theme\_light()  
  
p <- plot\_grid(sex, children)   
title <- ggdraw() + draw\_label("2. Plot of Charges vs Sex / Children ", fontface='bold')  
plot\_grid(title, p, ncol=1, rel\_heights=c(0.1, 1))



### From plot of sex vs chaarges we get that charges of health insurence does not depende upon whether a person male or female.

smoker<- ggplot(data, aes(smoker, charges)) +  
 geom\_jitter(aes(color = smoker), alpha = 0.7) +  
 theme\_dark()  
  
region<- ggplot(data, aes(charges,region)) +  
 geom\_jitter(aes(color = region), alpha = 0.7) +  
 theme\_dark()  
  
p <- plot\_grid(smoker, region)   
title <- ggdraw() + draw\_label("3. Plot of Charges vs Smoker / Region", fontface='bold')  
plot\_grid(title, p, ncol=1, rel\_heights=c(0.1, 1))



### As expected the charges of health insurence depends upon whether a person smoker or not.

### It seems that changes in region has not that much effect on charges.

## Let’s start Linear regression

### Splitting the dataset

set.seed(1338)  
n\_train <- round(0.8\* nrow(data))  
train\_indices <- sample(1:nrow(data), n\_train)  
data\_train <- data[train\_indices, ]  
data\_test <- data[-train\_indices, ]

### Let’s start model building

## My first formula  
formula\_1 <- as.formula("charges ~ age + sex + bmi + children + smoker + region")

## First model  
model\_1 <- lm(formula\_1, data = data\_train)  
summary(model\_1)

##   
## Call:  
## lm(formula = formula\_1, data = data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11005.5 -2823.9 -964.1 1548.7 25312.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11888.26 1088.17 -10.925 < 2e-16 \*\*\*  
## age 252.43 12.98 19.452 < 2e-16 \*\*\*  
## sex1 -25.08 362.46 -0.069 0.94485   
## bmi 345.99 31.36 11.032 < 2e-16 \*\*\*  
## children 478.91 151.58 3.159 0.00163 \*\*   
## smoker1 23687.62 446.10 53.100 < 2e-16 \*\*\*  
## region1 -743.75 519.34 -1.432 0.15241   
## region2 -1107.65 523.60 -2.115 0.03463 \*   
## region3 -1242.95 519.11 -2.394 0.01682 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5893 on 1061 degrees of freedom  
## Multiple R-squared: 0.7632, Adjusted R-squared: 0.7614   
## F-statistic: 427.4 on 8 and 1061 DF, p-value: < 2.2e-16

vif(model\_1)

## GVIF Df GVIF^(1/(2\*Df))  
## age 1.021712 1 1.010798  
## sex 1.011814 1 1.005889  
## bmi 1.111990 1 1.054510  
## children 1.004867 1 1.002431  
## smoker 1.018087 1 1.009003  
## region 1.104587 3 1.016717

### So there is no multicolinearity in model\_1

## Let's predict data on test dataset using first model  
prediction\_1 <- predict(model\_1, newdata = data\_test)  
  
#calculating the residuals  
residuals\_1 <- data\_test$charges - prediction\_1  
  
#calculating Root Mean Squared Error  
rmse\_1 <- sqrt(mean(residuals\_1^2))  
rmse\_1

## [1] 6705.281

## My second formula  
formula\_2 <- as.formula("charges ~ age + bmi + children + smoker+region")  
  
## Second model  
model\_2 <- lm(formula\_2, data = data\_train)  
summary(model\_2)

##   
## Call:  
## lm(formula = formula\_2, data = data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10994.4 -2833.6 -965.4 1541.9 25296.9   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -11903.34 1065.61 -11.170 < 2e-16 \*\*\*  
## age 252.41 12.97 19.465 < 2e-16 \*\*\*  
## bmi 346.06 31.33 11.047 < 2e-16 \*\*\*  
## children 479.04 151.50 3.162 0.00161 \*\*   
## smoker1 23690.59 443.83 53.378 < 2e-16 \*\*\*  
## region1 -743.74 519.09 -1.433 0.15222   
## region2 -1107.41 523.34 -2.116 0.03457 \*   
## region3 -1242.74 518.86 -2.395 0.01679 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5891 on 1062 degrees of freedom  
## Multiple R-squared: 0.7632, Adjusted R-squared: 0.7616   
## F-statistic: 488.9 on 7 and 1062 DF, p-value: < 2.2e-16

vif(model\_2)

## GVIF Df GVIF^(1/(2\*Df))  
## age 1.021186 1 1.010538  
## bmi 1.110581 1 1.053841  
## children 1.004713 1 1.002354  
## smoker 1.008696 1 1.004338  
## region 1.104505 3 1.016704

### So there is no multicolinearity in model\_2

## Let's predict data on test dataset using second model  
prediction\_2 <- predict(model\_2, newdata = data\_test)  
  
## Calculating residuals  
residuals\_2 <- data\_test$charges - prediction\_2  
rmse\_2 <- sqrt(mean(residuals\_2^2))  
rmse\_2

## [1] 6704.563

## Here is my formula three  
formula\_3 <- as.formula("charges ~ age + bmi + children + smoker")  
  
## Third model  
model\_3 <- lm(formula\_3, data = data\_train)  
summary(model\_3)

##   
## Call:  
## lm(formula = formula\_3, data = data\_train)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11630.6 -2915.7 -957.5 1528.7 26183.3   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -12258.58 1025.58 -11.953 < 2e-16 \*\*\*  
## age 252.83 12.97 19.490 < 2e-16 \*\*\*  
## bmi 331.31 30.00 11.044 < 2e-16 \*\*\*  
## children 479.27 151.69 3.159 0.00163 \*\*   
## smoker1 23708.16 443.01 53.516 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 5901 on 1065 degrees of freedom  
## Multiple R-squared: 0.7616, Adjusted R-squared: 0.7608   
## F-statistic: 850.8 on 4 and 1065 DF, p-value: < 2.2e-16

vif(model\_3)

## age bmi children smoker   
## 1.018330 1.014894 1.003790 1.001460

### Here also no multicolinearity in model\_3

## Let's predict the test dataset using third model  
prediction\_3 <- predict(model\_3, newdata = data\_test)  
  
residuals\_3 <- data\_test$charges - prediction\_3  
rmse\_3 <- sqrt(mean(residuals\_3^2))  
rmse\_3

## [1] 6695.038

r\_sq\_1 <- summary(model\_1)$r.squared  
r\_sq\_2 <- summary(model\_2)$r.squared  
r\_sq\_3 <- summary(model\_3)$r.squared  
  
print(paste0("R-squared for first model:", round(r\_sq\_1, 4)))

## [1] "R-squared for first model:0.7632"

print(paste0("R-squared for second model:", round(r\_sq\_2, 4)))

## [1] "R-squared for second model:0.7632"

print(paste0("R-squared for third model:", round(r\_sq\_3, 4)))

## [1] "R-squared for third model:0.7616"

print(paste0("RMSE for first model: ", round(rmse\_1, 2)))

## [1] "RMSE for first model: 6705.28"

print(paste0("RMSE for second model: ", round(rmse\_2, 2)))

## [1] "RMSE for second model: 6704.56"

print(paste0("RMSE for third model: ", round(rmse\_3, 2)))

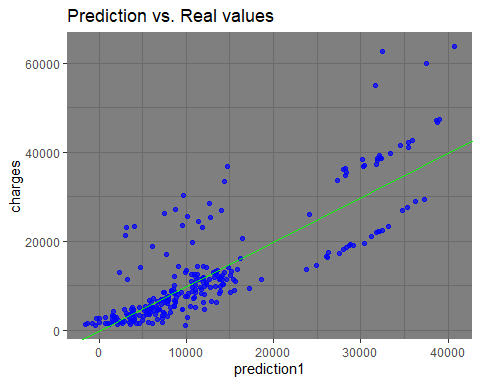
## [1] "RMSE for third model: 6695.04"

### As we can see, performance is quite similar between those models so we can drop the first model since rest two models are little bit simpler.

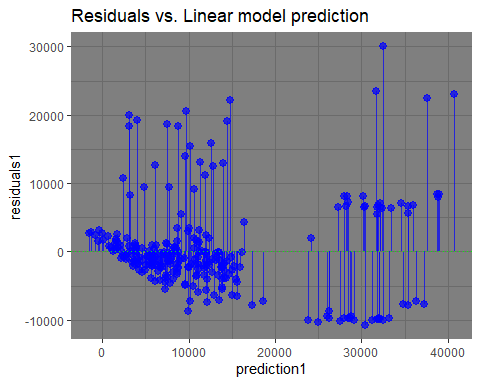
### Now we have to decide whether we go with second or third model

### Performance the second model

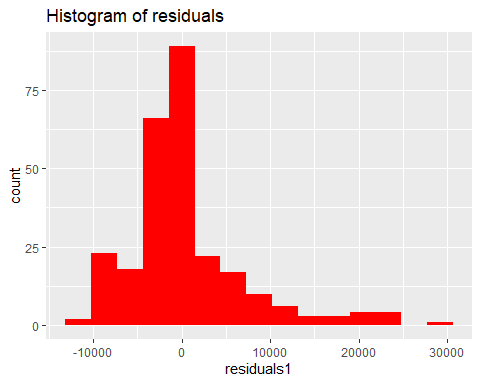
data\_test$prediction1 <- predict(model\_2, newdata = data\_test)  
ggplot(data\_test, aes(x = prediction1, y = charges)) +   
 geom\_point(color = "blue", alpha = 0.7) +   
 geom\_abline(color = "green") +  
 ggtitle("Prediction vs. Real values")+  
 theme\_dark()



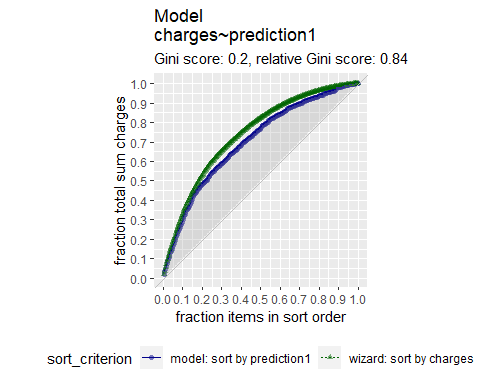
data\_test$residuals1 <- data\_test$charges - data\_test$prediction1  
  
ggplot(data = data\_test, aes(x = prediction1, y = residuals1)) +  
 geom\_pointrange(aes(ymin = 0, ymax = residuals1), color = "blue", alpha = 0.7) +  
 geom\_hline(yintercept = 0, linetype = 3, color = "green") +  
 ggtitle("Residuals vs. Linear model prediction")+  
 theme\_dark()



ggplot(data\_test, aes(x = residuals1)) +   
 geom\_histogram(bins = 15, fill = "red") +  
 ggtitle("Histogram of residuals")

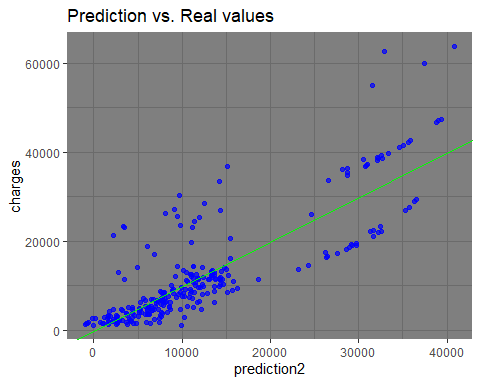


GainCurvePlot(data\_test, "prediction1", "charges", "Model")

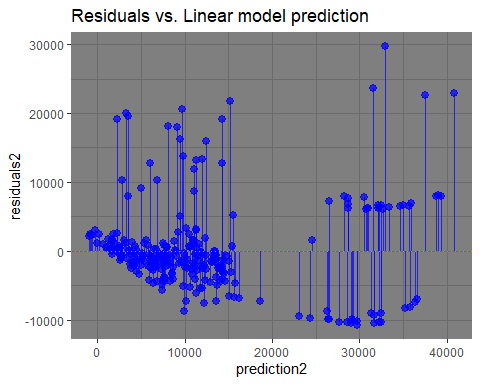


### Performance the third model

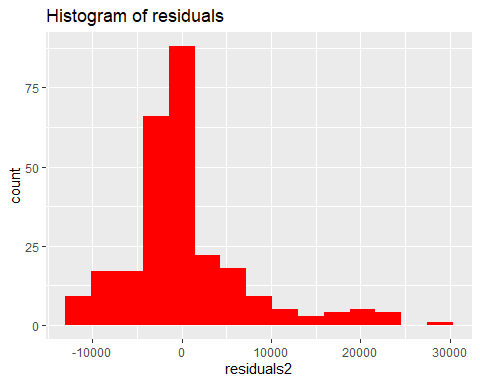
data\_test$prediction2 <- predict(model\_3, newdata = data\_test)  
ggplot(data\_test, aes(x = prediction2, y = charges)) +   
 geom\_point(color = "blue", alpha = 0.7) +   
 geom\_abline(color = "green") +  
 ggtitle("Prediction vs. Real values")+  
 theme\_dark()



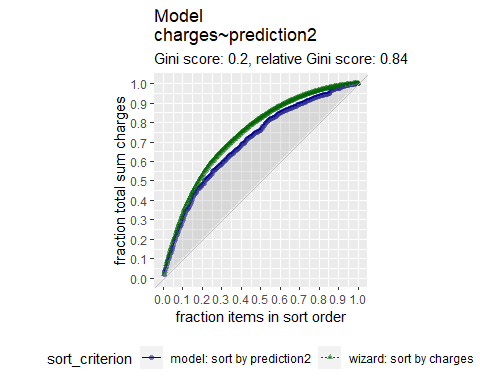
data\_test$residuals2 <- data\_test$charges - data\_test$prediction2  
  
ggplot(data = data\_test, aes(x = prediction2, y = residuals2)) +  
 geom\_pointrange(aes(ymin = 0, ymax = residuals2), color = "blue", alpha = 0.7) +  
 geom\_hline(yintercept = 0, linetype = 3, color = "green") +  
 ggtitle("Residuals vs. Linear model prediction")+  
 theme\_dark()



ggplot(data\_test, aes(x = residuals2)) +   
geom\_histogram(bins = 15, fill = "red") +  
ggtitle("Histogram of residuals")



GainCurvePlot(data\_test, "prediction2", "charges", "Model")



### We can see the errors in both models are close to zero so both models predicts quite well.

### We will take the third model since it’s simpler than the second model.

## Calculating coefficients of third model  
coef(model\_3)

## (Intercept) age bmi children smoker1   
## -12258.5792 252.8285 331.3097 479.2700 23708.1619

## So our predicted model is

### Charges = -12258.5792 + 252.8285*(Age) + 331.3097*(BMI) + 479.27*(Children) + 23708.1619*(Smoker)

## With R^2 = 0.7616

# Conclusion:

The aim of the project was to develop a model using multiple linear regression which can be used  
to predict health insurance charges.   
  
Our predicted model is   
 Charges = -12258.5792 + 252.8285\*Age + 331.3097\*BMI + 479.27\*(No of children) + 23708.1619\*S  
 where S=1 if the person is a smoker and S=0 if the person is not a smoker.