SESSION 12: Generalized Linear Models

Assignment 2

Problem Statement

1. Use the given link below:

https://archive.ics.uci.edu/ml/machine-learning-databases/communities/

Perform the below operations:

a. Visualize the correlation between all variable in a meaningful way, clear representation of correlations. Find out top 3 reasons for having more crime in a city.

Simple Linear Regression

```
dataset <- read.csv(file.choose())
View(dataset)
install.packages("caTools")
library(caTools)
set.seed(123)
split <- sample.split(dataset$Sales,SplitRatio = 2/3)
trainingdata <- subset(dataset,split == T)
testdata <- subset(dataset,split == F)
#fit the simple Linear regression mode into training data

regressor <- lm(formula = Sales ~ TV,data = trainingdata)
summary(regressor)
pred_value <- predict(regressor, newdata = testdata)
View(testdata)
library(ggplot2)
ggplot()+ geom_point(aes(x=trainingdata$TV,y = trainingdata$Sales),colour = 'red')+
                geom_line(aes(x = trainingdata$TV,y =predict(regressor, newdata = trainingdata)),
                colour ='blue')+
```

ggtitle('TV vs Sales')+

xlab("TV Spend")+

ylab('Sales')

b. What is the difference between covariance and correlation,

take an example from this dataset and show the differences if

any?

The problem with covariances is that they are hard to compare: when you calculate the covariance of a set of heights and weights, as expressed in (respectively) meters and kilograms, you will get a different covariance from when you do it in other units (which already gives a problem for people doing the same thing with or without the metric system!), but also, it will be hard to tell if (e.g.) height and weight 'covary more' than, say the length of your toes and fingers, simply because the 'scale' the covariance is calculated on is different.

The solution to this is to 'normalize' the covariance: you divide the covariance by something that represents the diversity and scale in both the covariates, and end up with a value that is assured to be between -1 and 1: the correlation. Whatever unit your original variables were in, you will always get the same result, and this will also ensure that you can, to a certain degree, compare whether two variables 'correlate' more than two others, simply by comparing their correlation.

Note: the above assumes that the reader already understands the concept of covariance.

Are you sure you can't compare covariances with different units? The units pass through covariance multiplied - if your X is in `cm`, and your Y is in `s`, then your $\mathrm{cov}(X,Y) = z\ \mathrm{cm} \cdot s$cov(X,Y)=z cm⋅s. And then you can just multiply by the result by the unit conversion factor. Try it in R: `cov(cars$speed,cars$dist) == cov(cars$speed/5,cars$dist/7)*(7*5)`