regression

Antonio Javier González Ferrer

26 de diciembre de 2016

Introduction

The diamonds dataset contains the prices and other attributes such as the caratage, clarity or colour of 308 stones. The goal of this project is to come up with a multiple linear regression model and compare this initial model against different solutions to improve the initial simple approach.

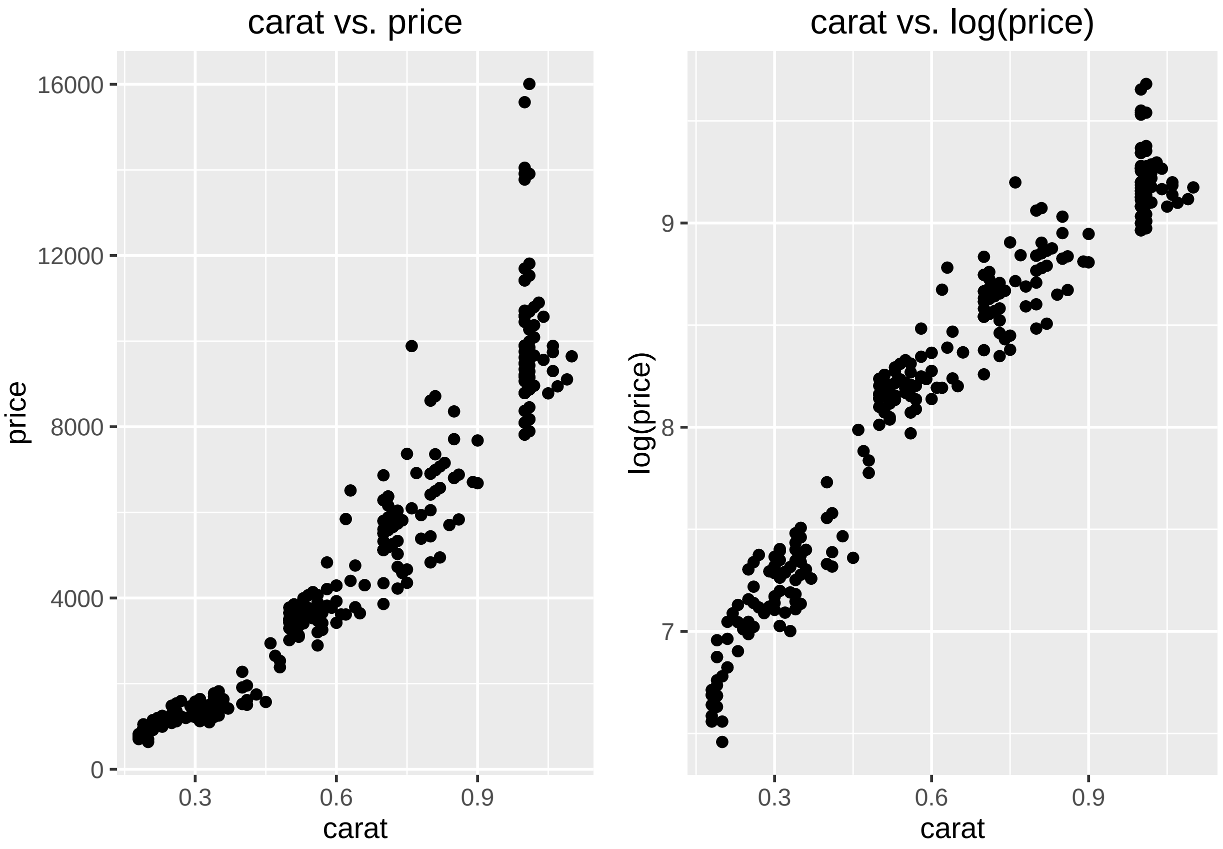
```
library(ggplot2)
library(gridExtra)

df <- read.table("data/HW-diamonds.txt", quote="\"", comment.char="")
colnames(df) <- c("carat", "color", "clarity", "institution", "price")

observations = nrow(df)
variables = ncol(df)
sprintf("observations: %s and variables: %s", observations, variables)
```

```
## [1] "observations: 308 and variables: 5"
```

Logarithmic transformation

When a linear regression study is performed, it is a good practice to transform some of the original variables if we could not achieve linearity formerly. The logarithmic transformation is one of the most popular. This transformation is commonly used to induce symmetry and linearity when the original data is non-linear and its variance is not constant and increases along the x-axis. For instance, compare the two plots of the below Figure, where the weight of the diamond is plotted against its original price. In the left-handed plot, the variance concentrated around 0.3 is significantly smaller than around 1, and besides the trend does not seem linear. On the other hand, if we consider the log(price), the plot looks linearly nicer.

```
p1 <- ggplot(df, aes(x=carat, y=price))+geom_point()+ ggtitle('carat vs. price')
p2 <- ggplot(df, aes(x=carat, y=log(price)))+geom_point() + ggtitle('carat vs. log(price)')
grid.arrange(p1, p2, ncol=2)
```



carat vs. price

carat vs. log(price)

From now on, we will consider as response variable the log(price). However, notice that the interpretation of the analysis will be notably different since we are not considering the original measure.

Multiple Linear Regression

We will start by performing a simple linear analysis for pricing the diamonds based on the caratage, colour purity, clarity and institution.

First of all, we will re-level some of the categorical variables as the reference category. Concretely, the new references categories will be I in colour, VS2 in clarity and HRD in institution.

```r
df$color=relevel(df$color, ref="I")
df$clarity=relevel(df$clarity, ref="VS2")
df$institution=relevel(df$institution, ref="HRD")
```

Then, we fit the linear model using the mentioned variables. In the log-linear model, the interpretation of the estimated coefficient $\hat{\beta}$ is that a one-unit increase in X will produce and expected increase in Y of $e^{\hat{\beta}0}$ units. The parameters for the fitted model and the fitted line are the following: