Drills with R on Generalized Linear Models

Brayden Yates

University of the Cumberlands

MSDS-531: Statistics for Data Science

Dr. Danny Barnes

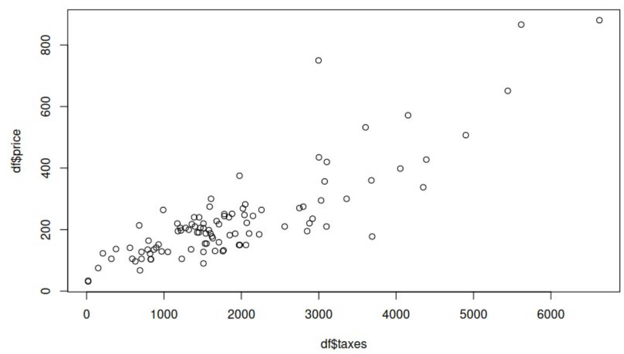
10 February 2024

Drills with R on Generalized Linear Models

In order to form a scatter plot with selling price as the dependent variable, and tax bill as the independent variable, I had to first import the dataset, and set the first row of data as the column names, then remove those column names from the dataset. Finally, I was able to plot the data on a scatter plot:

```
df <- read.table('~/Documents/DSIAN/stats/Houses.dat')

colnames(df) <- df[1, ]
```




```
df <- df[-1, ]

plot(df$taxes, df$price)
```

This returned the following scatter plot:

Based on the given scatter plot, there appears to be a strongly correlated linear relationship between taxation on a property and the selling price of the property. Based on this, the normal GLM structure of constant variability seems appropriate.

In order to fit the data using the identify link function with both a normal GLM and a gamma GLM, I used the following code:

```
normal_glm <- glm(y ~ x1 + x2, data = df, family = gaussian(link =
"identity"))
gamma_glm <- glm(y ~ x1 + x2, data = df, family = Gamma(link = "log"))
```

When running summary(normal_glm), we find a p-value of 0.47, which is significantly higher than our typical threshold significance level of 0.05. This indicates that the newness of a house does not have any statistical significance when examining the variation in price. Similar results appear when running summary(gamma_glm), which returns a value of 0.54.

Using the following code, I examined variability in selling prices:

```
x1_range <- seq(100000, 500000, by = 10000)
predicted_normal <- predict(normal_glm, newdata = data.frame(taxes =
x1_range, new = mean(df$new)))
predicted_gamma <- predict(gamma_glm, newdata = data.frame(taxes =
x1_range, new = mean(df$new)))


plot(x1_range, predicted_normal, type = "l", col = "blue", ylim = c(0,
max(predicted_normal, predicted_gamma)),
     xlab = "Mean Selling Price", ylab = "Estimated Variability", main =
"Variability in Selling Prices")
lines(x1_range, predicted_gamma, col = "red")
```

```
legend("topright", legend = c("Normal GLM", "Gamma GLM"), col = c("blue",
"red"), lty = 1)
```

In this chart, we can see that the estimated variability increases as mean selling price increases with the normal GLM, however the estimated variability remains at zero as mean selling price increases for the gamma GLM.

Examining A1C values for each of the models, we can identify that the Gamma GLM is preferred over the normal GLM with this particular dataset. The Normal GLM's A1C value is 1162.178, while the Gamma's A1C is 1115.106. Here is the code used to return these values:

```
AIC_normal <- AIC(normal_glm)

AIC_gamma <- AIC(gamma_glm)

AIC_normal

AIC_gamma
```