Homework 3

CSCE 587

Fall 2016

Due: 10/04/2016 via Dropbox

Name: Sendurr Selvaraj

VIP ID: 00323540

**1. Explore the first two columns which contain real numbers:**

1. **Plot first column (Y) against second column (X). Save the plot to a pdf file.**

**Answer:**

*data = read.csv("/Users/Sendurr/Dropbox/Transfer/CSCE587 - Big Data/gold\_target1.csv")*

*summary(data)*

*y=data[,1]*

*x=data[,2]*

*x*

*y*

*plot(y~x,main="As level vs Sb level")*

**

1. **Try fitting these two columns with a linear model lm(). Hint: You might want to review the linear regression lab.**

**Answer:**

*m=lm(y~x)*

*str(m)*

*print(m)*

Output:

Call:

lm(formula = y ~ x)

Coefficients:

(Intercept) x

0.9974 1.7948

1. **As in the linear regression lab, visualize the model with the commands, where m is the variable you used to hold the model: par(mfrow=c(2,2)) plot(m) Save this plot to a pdf file.**

**Answer:**

*par(mfrow=c(2,2))*

*plot(m)*



1. **Explain the top left figure. What does this tell us about the fit of our model?**

**Answer:**

The residuals are not evenly distributed, therefore, linear regression model is not suited to derive the dependency of Y and X.

1. **Visualize the predicted and observed y values similar to what we did in slide 6 of the linear regression lab. Save this graph to a pdf file.**

**Answer:**

*ypred=predict(m)*

*par(mfrow=c(1,1))*

*plot(y,ytype='l',xlab="observed y", ylab='predicted y')*

*points(y,ypred)*



**2. Explore column 4 versus columns 1 and 2.**

1. **Plot column 4 (Y) against column 1 (X). Save this plot to a pdf file.**

**Answer:**

*y=data[,4]*

*x=data[,1]*

*x*

*y*

*plot(y~x,main=" Gold deposit proximity vs As level")*



1. **Plot column 4 (Y) against column 2 (X). Save this plot to a pdf file.**

**Answer:**

*y=data[,4]*

*x=data[,2]*

*x*

*y*

*plot(y~x,main=" Gold deposit proximity vs Sb level")*



c. Try fitting column 4 versus column 2 with a logistic model glm(). Hint: You might want to review the logistic regression lab.

d. Visualize the fit of your model using: plot(gold\_target1$V4~gold\_target1$V2) lines(gold\_target1$V2,lrm1$fitted,type="l", col="red") Save this plot to a pdf.

e. Now try fitting column 4 versus columns 1 and 2 with the logistic model glm(). How can you accomplish this? When you only have Y versus X, you use Y~X as you did in step c. When you have X1 and X2 then you use Y~X1+X2. Note: RStudio will give a warning that glm fitted probabilities numerically 0 or 1 occurred. This is caused by the data in column 1.

f. Compare the models from step c with that of step e using the function summary(). In particular, compare the estimated coefficient for gold\_target$V2. What are the two values? How have the confidence values for these estimates changed? (Hint: look at the significance codes.)