Homework 3
CSCE 587
Fall 2016
Due: 10/04/2016 via Dropbox
Name: Sendurr Selvaraj
VIP ID: 00323540

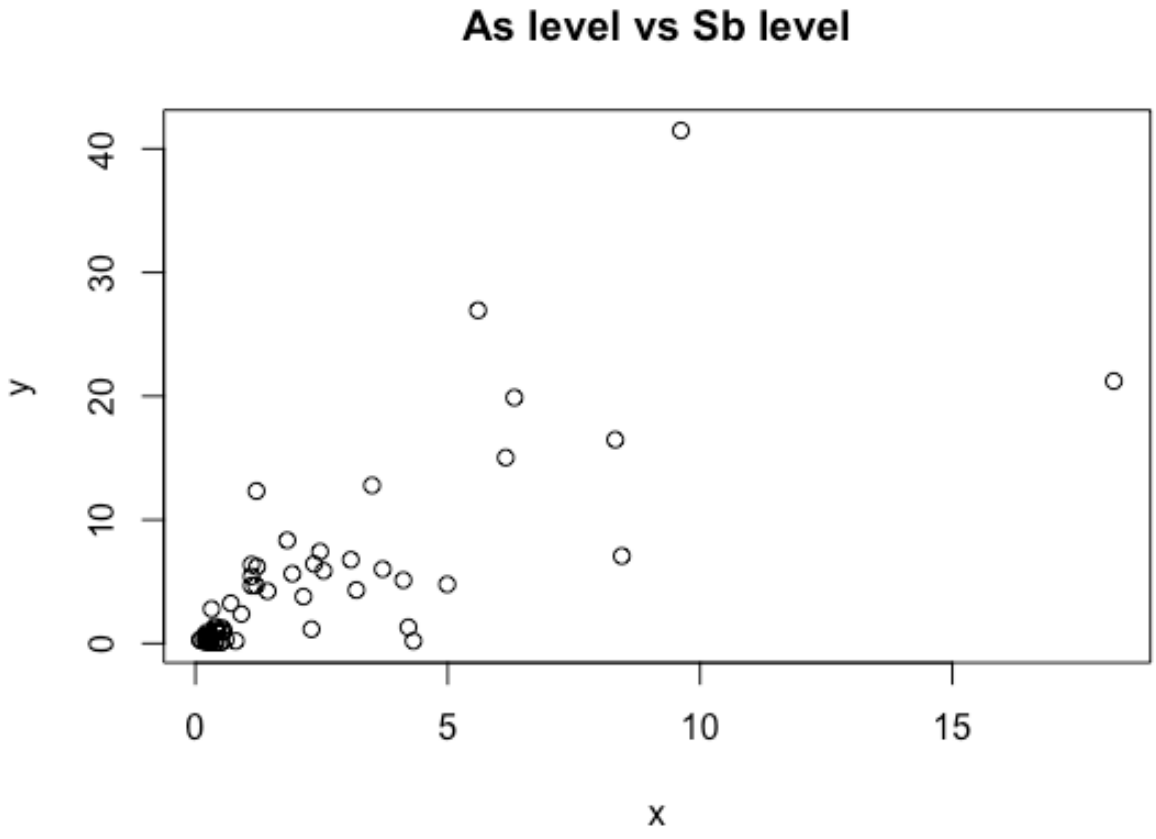
---

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

a. 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")
```



b. 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:

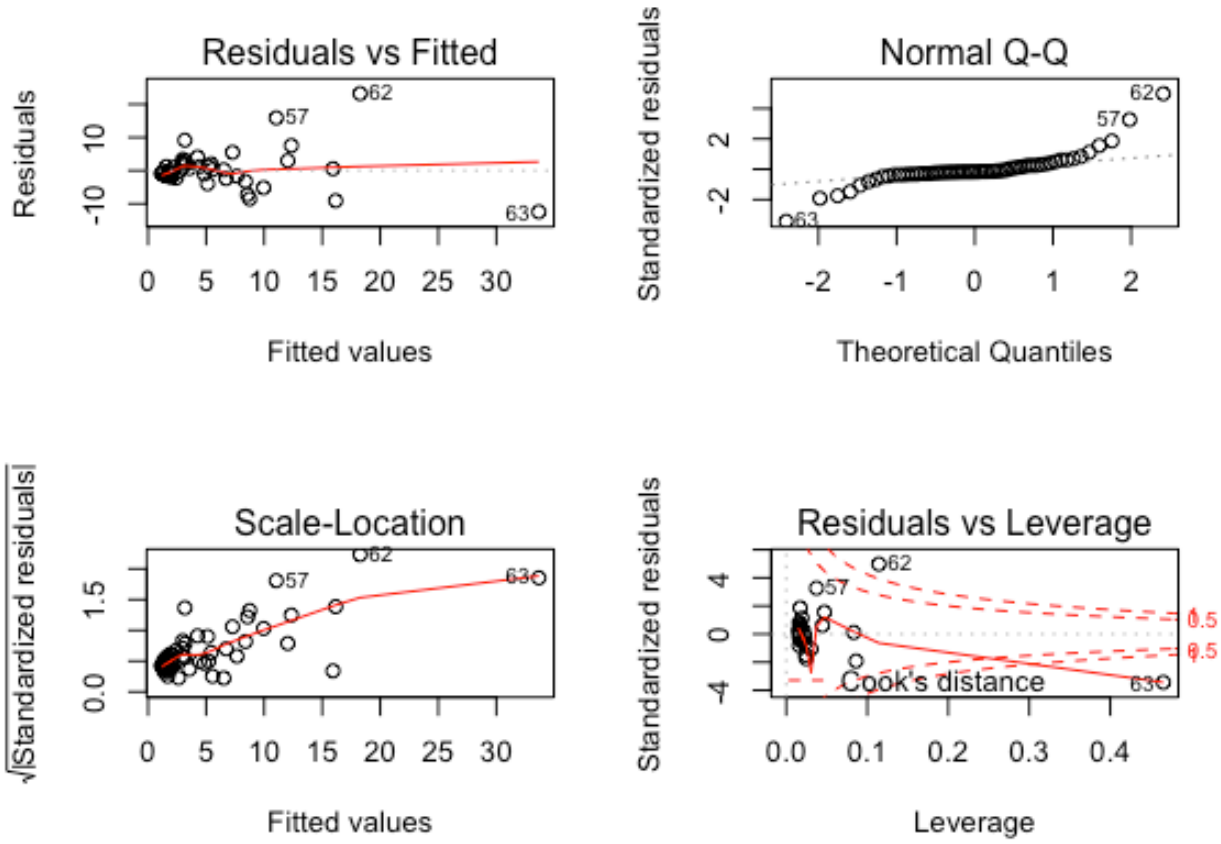
Call:
lm(formula = y ~ x)

Coefficients:
(Intercept) x
0.9974 1.7948

c. 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)



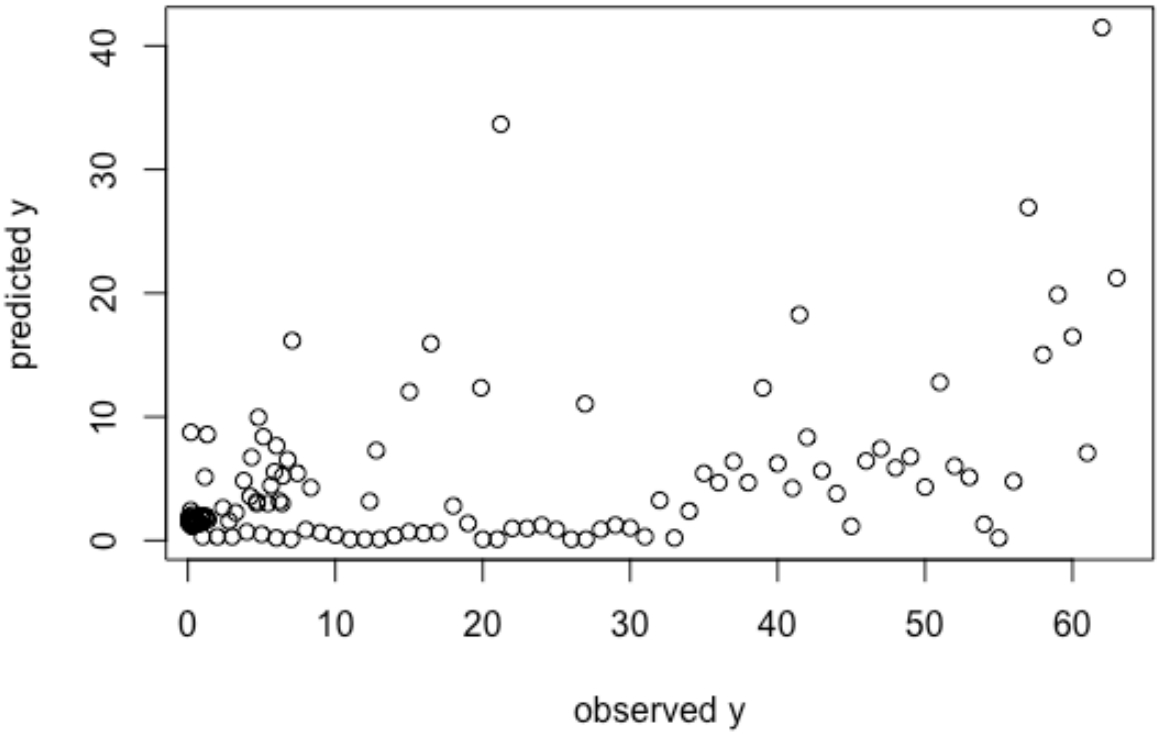
d. Explain the top left figure. What does this tell us about the fit of our model?

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

e. 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)
```

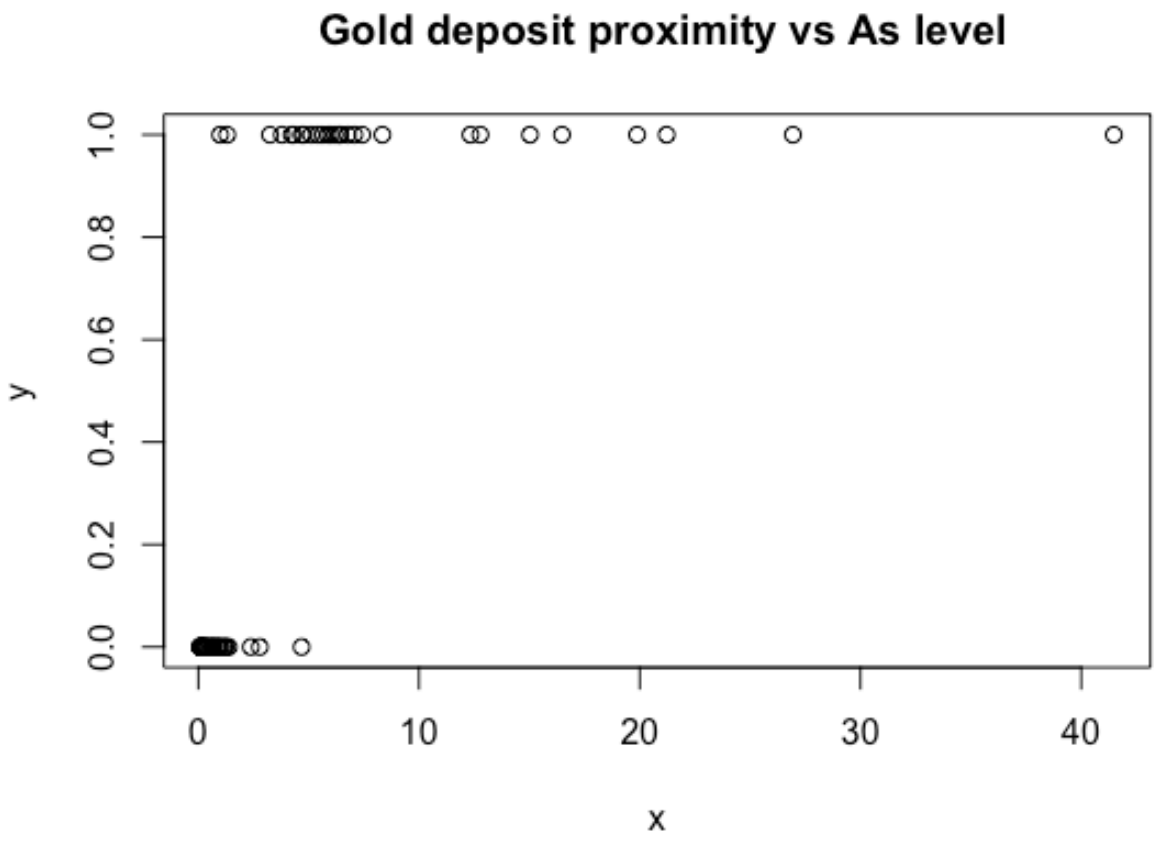


2. Explore column 4 versus columns 1 and 2.

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

Answer:

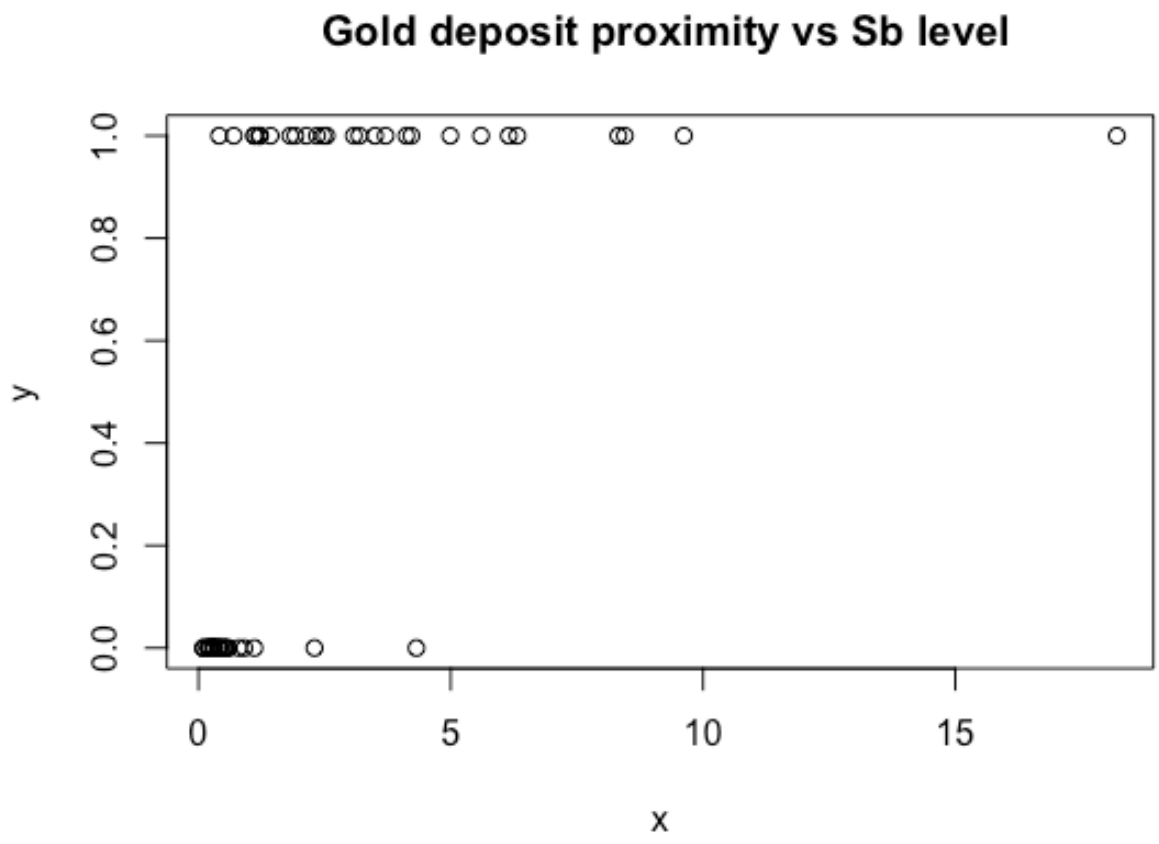
```
y=data[,4]
x=data[,1]
x
y
plot(y~x,main=" Gold deposit proximity vs As level")
```



b. 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.

Answer:

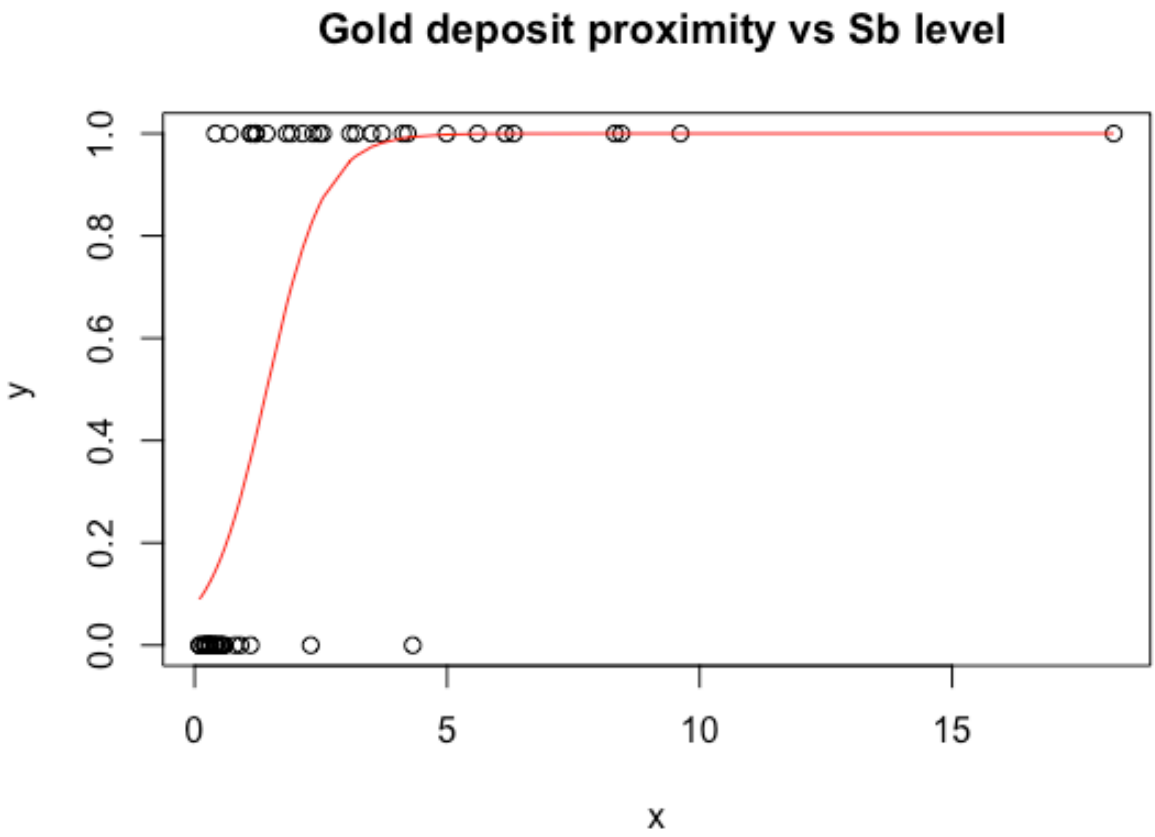
glm_out1 = glm(y~x,family=binomial(logit))
glm_out1

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.

Answer:

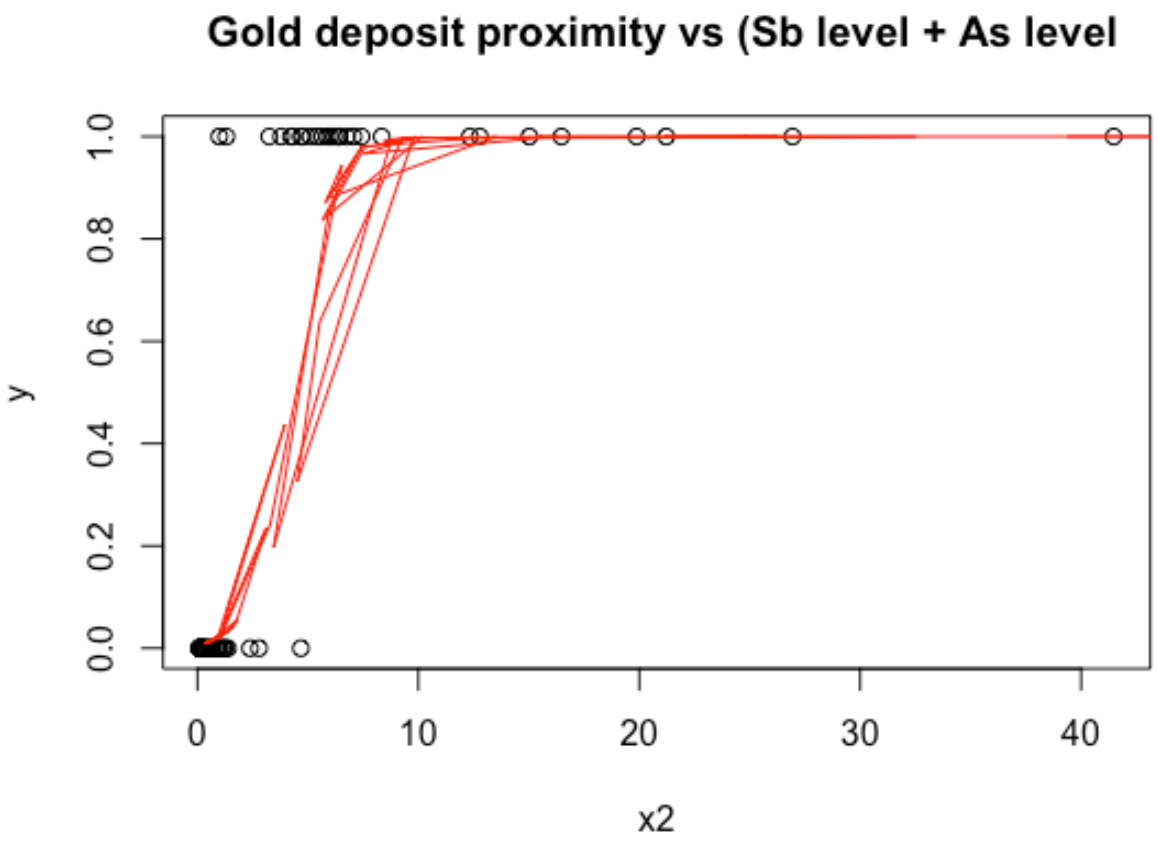
plot(y~x)
lines(x,glm_out1$fitted,type="l", col="red")



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.**

Answer:

```
y=data[,4]
x1=data[,2]
x2=data[,1]
x1
x2
y
plot(y~x1+x2,main=" Gold deposit proximity vs (Sb level + As level")
glm_out2 = glm(y~x1+x2,family=binomial(logit))
glm_out2
lines(x1+x2,glm_out2$fitted,type="l", col="red")
```



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.)

Answer:

summary(glm_out1)
summary(glm_out2)

Estimated co efficient of v2 in step c= 1.7427
Estimated co efficient of v2 in step e= 0.9190

Signif. codes: 0 '***'
Signif. codes: 0 '***'