Christina Wang

Econ 108

November 4, 2022

Problem Set 4

1)

a)

t-val: 19.645, 8,874, 2.332

p-val : <2e-16, <2e-16, .01

b)

The regression output indicates rejection of the null hypothesis.

c)

Now, the regression output indicates nonrejection of the null hypothesis.

2)

a)

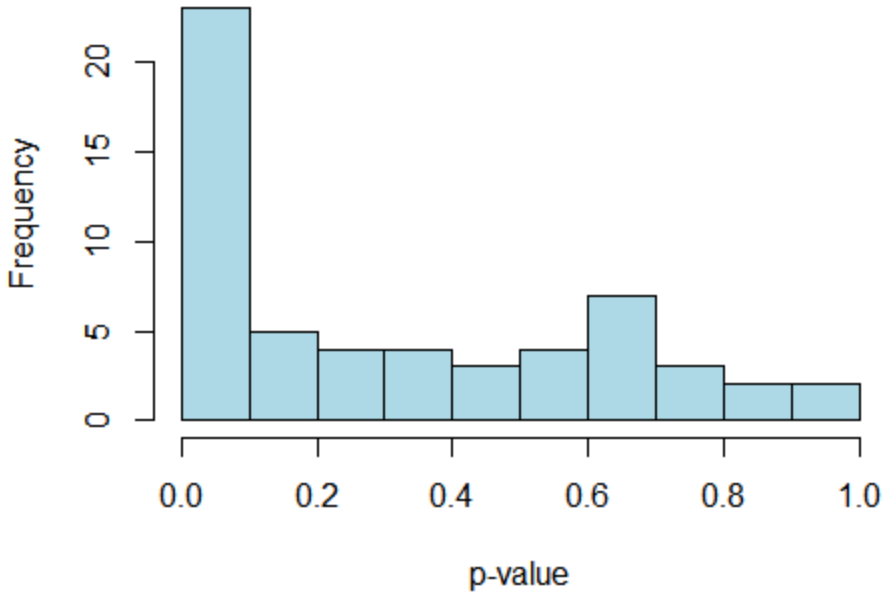
To do this, we will find the chance that the email is not spam by using word_george, as spammers are not likely to use your name.

```
full <- glm(word_george ~ ., data=email, family=binomial)
```

```
1 - full$deviance/full$null.deviance
```

R^2 = 0.4419095

b)



Yes, the histogram indicates that some of the features might be useful signals for predicting spam probabilities.

c)

0.005297195 is the resulting p-value threshold for declaring that a feature is a statistically significant signal. 57 features a found by the B-H procedure. Yes, they make sense as significant features for predicting spams.

d)

The in sample R^2 of the cut model (0.4239541) is less than than the in sample R^2 of the full model.

e)

full cut

-14.73661 -14.22485

f) 28 features are discovered by the forward stepwise regression. They are a subset of the features found by the FDR method. Together, they have 57 features.