Christina Wang

Econ 108

21 October 2022

Econ 108 Pset 3

1)

a)

b) p(y=1 | x) = e^x'B / (1 + e^x'B).

exp(b["char_dollar"]) = 6.492946

exp(b["word_george"]) = .003089205

(6.492946+.003089205)/ (6.492946+.003089205+1)=.8666

The in-sample R^2 = 1 – (1548.7/6170.2) = .75

.8666 > .75, so the in-sample R^2 of the simplified model is larger than the in-sample R^2 of the full regression.

c) Having the word "george" in an email multiplies the odds of spam by exp(-5.8) or around .003. It is statistically significant because it greatly reduces the chances of this email being spam.

d) Two feature logit model using training subsample:

(6.160852 + .001257)/ (6.160852 + .001257 + 1) = .86037

R^2 on leave-out subsample:

1 – (223.13/1310.25) = .8297

R^2 on leave-out subsample is greater than in-sample R^2.

e)

exp(b["word_make"]) + exp(b["word_address"]) + exp(b["word_all"]) + exp(b["word_3d"]) + exp(b["word_our"]) + exp(b["word_over"]) + exp(b["word_remove"]) + exp(b["word_internet"]) + exp(b["word_people"]) + exp(b["word_report"]) + exp(b["word_addresses"]) + exp(b["word_free"]) + exp(b["word_order"]) + exp(b["word_mail"]) + exp(b["word_receive"]) + exp(b["word_will"]) + exp(b["word_business"]) + exp(b["word_email"]) + exp(b["word_you"]) + exp(b["word_credit"]) + exp(b["word_your"]) + exp(b["word_font"]) + exp(b["word_000"]) + exp(b["word_money"]) + exp(b["word_hp"]) + exp(b["word_hpl"]) + exp(b["word_650"]) + exp(b["word_lab"]) + exp(b["word_labs"]) + exp(b["word_telnet"]) + exp(b["word_857"]) + exp(b["word_data"]) + exp(b["word_415"]) + exp(b["word_85"]) + exp(b["word_technology"]) + exp(b["word_1999"]) + exp(b["word_parts"]) + exp(b["word_pm"]) + exp(b["word_direct"]) + exp(b["word_cs"]) + exp(b["word_meeting"]) + exp(b["word_original"]) + exp(b["word_project"]) + exp(b["word_re"]) + exp(b["word_edu"]) + exp(b["word_table"]) + exp(b["word_conference"]) + exp(b["char_semicolon"]) + exp(b["char_leftbrac"]) + exp(b["char_leftsquarebrac"]) + exp(b["char_exclaim"]) + exp(b["char_pound"]) = 89.5785

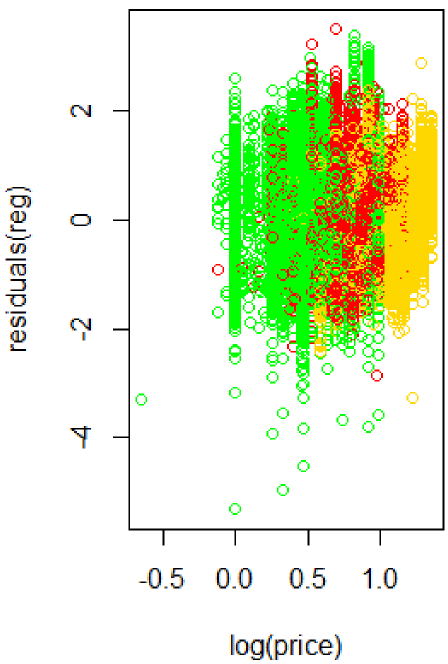
89.5785/(89.5785+1) = .989

.989 > .75, so this out of sample R^2 is much greater than the in-sample R^2.

2)

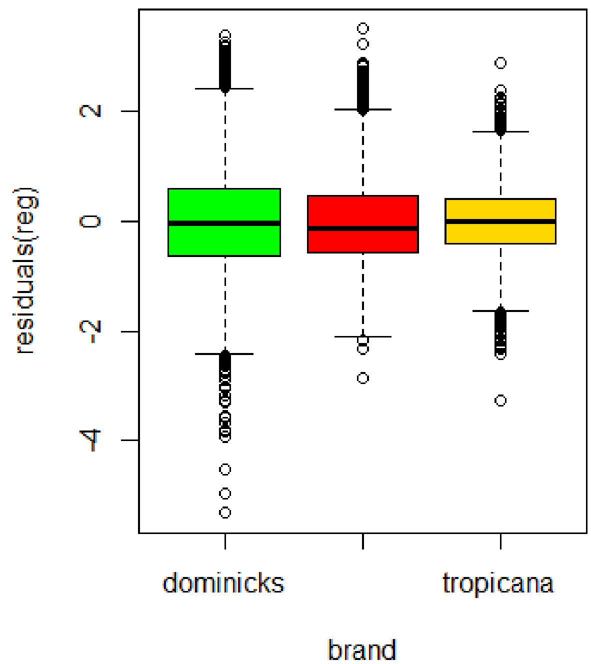
a) glm(formula = log(sales) ~ log(price) + brand, data = oj)

b) plot(residuals(reg) ~ log(price), data=oj, col=brandcol[oj$brand])



The residual plots are mostly consistent, so I don’t think they suggest possible conditional heteroskedasticity.

c) plot(residuals(reg) ~ brand, data=oj, col=brandcol)



The residual plots mostly center around 0, so I don't think they suggest possible conditional heteroskedasticity.

d) summary(reg) shows us std. error of log(price) = .022, brandminute.maid = .012, brandtropicana = .016 as shown below.

```
Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)       10.82882    0.01453  745.04   <2e-16
log(price)        -3.13869    0.02293 -136.89   <2e-16
brandminute.maid   0.87017    0.01293   67.32   <2e-16
brandtropicana     1.52994    0.01631   93.81   <2e-16
```

When computing the HC-robust standard eroors using the AER package, we get very similar numbers; however, each of these standard errors are slightly larger.

```
> sqrt(bvar["log(price)","log(price)"])
[1] 0.02494664
> sqrt(bvar["brandminute.maid","brandminute.maid"])
[1] 0.01441403
> sqrt(bvar["brandtropicana","brandtropicana"])
[1] 0.01740783
> 
```