Regression Analysis

Logistic Regression

Nicoleta Serban, Ph.D.
Professor
Stewart School of Industrial and Systems Engineering

The Demographics of Obesity:
Modeling and Prediction

Georgia Tech

1

About This Lesson

Georgia Tech

2

Model Estimation

Fit a logistic regression model

model = glm(Obesity~agegr+gender+edu, family=binomial)
summary(model)

```
…
                 Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)      -1.20581  0.15730     -7.666   1.78e-14 ***
agegr25to34      0.47271   0.14428     3.276    0.001052 **
agegr35to44      0.76486   0.14196     5.388    7.13e-08 ***
agegr45to64      0.84815   0.13240     6.406    1.49e-10 ***
agegr65+         0.60086   0.13751     4.370    1.24e-05 ***
genderFemale     0.23041   0.06363     3.621    0.000293 ***
edu9to11Grade    0.05632   0.12229     0.461    0.645110
eduHighSchool    -0.03440  0.11436     -0.301   0.763579
eduSomeCollege   0.13947   0.11036     1.264    0.206301
eduCollege+      -0.40077  0.11757     -3.409   0.000653 ***
…
Null deviance: 5739.9 on 4313 degrees of freedom
Residual deviance: 5641.3 on 4304 degrees of freedom
…
```

3

Model Estimation

Fit a logistic regression model

model = glm(Obesity~agegr+gender+edu, family=binomial)
summary(model)

```
…
                 Estimate  Std. Error  z value  Pr(>|z|)
(Intercept)      -1.20581  0.15730     -7.666   1.78e-14 ***
agegr25to34      0.47271   0.14428     3.276    0.001052 **
agegr35to44      0.76486   0.14196     5.388    7.13e-08 ***
agegr45to64      0.84815   0.13240     6.406    1.49e-10 ***
agegr65+         0.60086   0.13751     4.370    1.24e-05 ***
genderFemale     0.23041   0.06363     3.621    0.000293 ***
edu9to11Grade    0.05632   0.12229     0.461    0.645110
eduHighSchool    -0.03440  0.11436     -0.301   0.763579
eduSomeCollege   0.13947   0.11036     1.264    0.206301
eduCollege+      -0.40077  0.11757     -3.409   0.000653 ***
…
Null deviance: 5739.9 on 4313 degrees of freedom
Residual deviance: 5641.3 on 4304 degrees of freedom
…
```


$$\widehat{\beta}_{agegr25to34} = 0.4727$$

The ratio of the odds of obesity for age group 25-34 versus the age group 18-24 is 1.604 (or, equivalently the log odds ratio is 0.4727), holding all other predicting variables fixed. Odds of obesity for age group 25-34 are 60.4% higher than for age group 18-24 (baseline group).

$$\widehat{\beta}_{genderfemale} = 0.2304$$

The ratio of the odds of obesity for females versus males is 1.259, holding all other predicting variables fixed. Odds of obesity for females is 26% higher than for males.

4

Statistical Inference

Test for overall regression
```
gstat = model$null.deviance - deviance(model)
cbind(gstat, 1-pchisq(gstat,length(coef(model))-1))
     gstat
[1,] 98.63672 0
```

Test for overall regression: ***p-value* ≈ 0 (< 0.01).** Reject the null hypothesis that all regression coefficients are zero. Conclude there are predicting variables that explain the variability in obesity.

```
round(coefficients(summary(model))[,4],4)
     (Intercept)   agegr25to34   agegr35to44   agegr45to64   agegr65+
     0.0000        0.0011        0.0000        0.0000        0.0000
     genderFemale  edu9to11Grade eduHighSchool eduSomeCollege eduCollege+
     0.0003        0.6451        0.7636        0.2063         0.0007
```

Except for one, education regression coefficients are not statistically significant given that we account for age and gender.

Georgia Tech

5

Predictive Power

Prediction Accuracy
```
library(boot)
cost0.5 = function(y, pi){
    ypred=rep(0,length(y))
    ypred[pi>0.5] = 1
    err = mean(abs(y-ypred))
    return(err)}
obdata.fr = data.frame(cbind(Obesity, agegr, gender, edu))
```

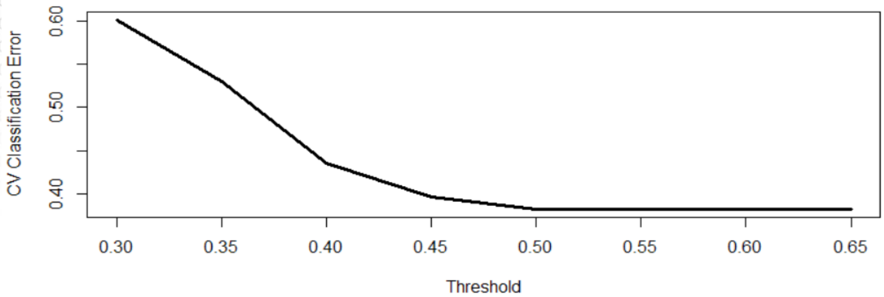
classification error for 10-fold cross-validation
```
cv.err = cv.glm(obdata.fr, model, cost=cost0.5, K=10)$delta[1]
⋮
cv.err = c(cv.err0.3, cv.err0.35, cv.err0.4, cv.err0.45, cv.err0.5,
    cv.err0.55, cv.err0.6, cv.err0.65)
```

Smallest prediction error is 0.3824
```
plot(c(0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65), cv.err,
    type="l", lwd=3, xlab="Threshold", ylab="CV Classification Error")
```

Georgia Tech

6

Predictive Power



Prediction accuracy is highest and equal for thresholds above 0.5. ***Why?***

- It is the same prediction accuracy as if we were to replace all predictions with 0 (that is, predict everyone is not obese).
- The model has no predictive power since it performs worse than prediction without modeling.

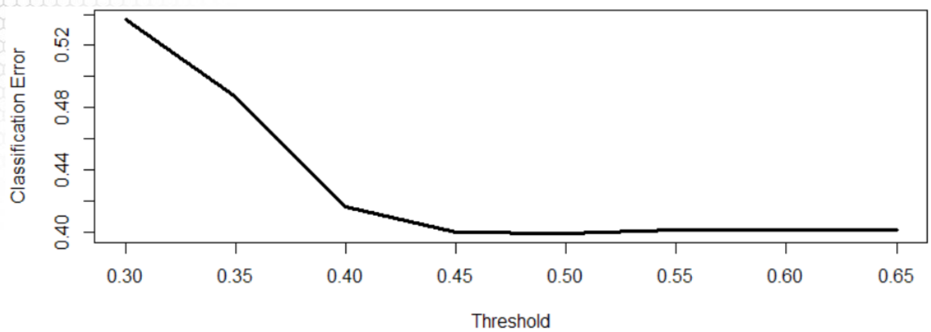
7

Prediction for Test Data

```
## Prediction given a set of new observations
## Prepare the test data
testobdata = read.table("testobesitydata.txt", h=T)
agegr.t = factor(testobdata$AgeGroup, labels=c("18to24", "25to34", "35to44",
    "45to64", "65+"))
gender.t = factor(testobdata$Gender, labels=c("Male","Female"))
edu.t = factor(testobdata$Education, labels=c("<9thGrade","9to11Grade",
    "HighSchool", "SomeCollege", "College+"))
pred.data = data.frame(agegr=agegr.t, gender=gender.t, edu=edu.t)
### Predict
pred.test = predict.glm(model,pred.data,type="response")
### Prediction Accuracy for multiple thresholds
err0.3 = cost0.3(testobdata$Obesity, pred.test)
⋮
err0.65 = cost0.65(testobdata$Obesity, pred.test)
err = c(err0.3, err0.35, err0.4, err0.45, err0.5, err0.55, err0.6, err0.65)
plot(c(0.3, 0.35, 0.4, 0.45, 0.5, 0.55, 0.6, 0.65), err,
    type="l", lwd=3, xlab="Threshold", ylab="Classification Error")
```

8

Prediction for Test Data



- Prediction accuracy is highest at 0.5; it is similar as the prediction accuracy if we were to predict everyone is not obese.
- The prediction accuracy using the fitted model did not improve for the test data.

9

Summary

10