Modeling HW3 Part3

Kerui Cao

9/30/2019

Building a logistic regression model:

the folder rodents contains data on rodents in a sample of New York City apartments.

```
##
## Attaching package: 'data.table'
```

```
## The following objects are masked from 'package:dplyr':
##
##     between, first, last
```

```
## The following object is masked from 'package:purrr':
##
##     transpose
```

Please read for the data details. http://www.stat.columbia.edu/~gelman/arm/examples/rodents/rodents.doc

First we transform the data in race, let 2 represents Asian,3 represents Black,4 represents Hispanish, and 1 represents White.

```
apt_dt$race[apt_dt$asian==T]=2
apt_dt$race[apt_dt$black==T]=3
apt_dt$race[apt_dt$hisp==T]=4
```

1. Build a logistic regression model to predict the presence of rodents (the variable y in the dataset) given indicators for the ethnic groups (race). Combine categories as appropriate. Discuss the estimated coefficients in the model.

Since we don‘t know the real meaning of each variables, so we just use some information criteria as model selection standards, such as R^2 and the significance of coefficients.

```
##
## Call:
## glm(formula = y ~ factor(race), family = binomial(link = "logit"),
##     data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.9922  -0.9293  -0.4690  -0.4690   2.1270
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -2.1521     0.1281 -16.798   <2e-16 ***
## factor(race)2   0.5518     0.2665   2.070   0.0384 *
## factor(race)3   1.5361     0.1687   9.108   <2e-16 ***
## factor(race)4   1.6995     0.1664  10.212   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1672.2  on 1521  degrees of freedom
## Residual deviance: 1526.3  on 1518  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1534.3
##
## Number of Fisher Scoring iterations: 4
```

The result shows that, if the race is 0, the probability of presence of rodents will decrease , if the race is 1,2,and 3, the probability of presence of rodents will significantly increase.

2. Add to your model some other potentially relevant predictors describing the apartment, building, and community district. Build your model using the general principles explained in Section 4.6 of the Gelman and Hill. Discuss the coefficients for the ethnicity indicators in your model.

```
##
## Call:
## glm(formula = y ~ factor(race) + defects + dist + bldg, family = binomial(link = "logit"),
##     data = apt_dt)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.0007  -0.7000  -0.4152  -0.3055   2.4569
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -2.223947   0.226320  -9.827  < 2e-16 ***
## factor(race)2  0.464621   0.285743   1.626    0.104
## factor(race)3  1.152582   0.183535   6.280 3.39e-10 ***
## factor(race)4  1.395581   0.182577   7.644 2.11e-14 ***
## defects        0.494955   0.043147  11.471  < 2e-16 ***
## dist           0.036348   0.045607   0.797    0.425
## bldg          -0.002857   0.002504  -1.141    0.254
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1672.2  on 1521  degrees of freedom
## Residual deviance: 1349.2  on 1515  degrees of freedom
##   (225 observations deleted due to missingness)
## AIC: 1363.2
##
## Number of Fisher Scoring iterations: 5
```

Compared to previous model, the main change in ethnicity coefficients is that if the race is 2, it no longer have significant influence on the probability of the presents of rodents, we can deduce that the bias in previous model generated by missing variables “defects”.