# Tobacco_Analysis

Qizheng Wang

5/7/2020

### Tobacco Survey Analysis(Binary Generalized linear models)

```
smokeFile = 'smokeDownload.RData'
if(!file.exists(smokeFile)){
    download.file(
            'https://github.com/Roderickwqz/Tobacco_Analysis/blob/master/smoke.RData',
            smokeFile)
}
(load(smokeFile))
```

```
## [1] "smoke"        "smokeFormats"
```

```
smokeFormats[
  smokeFormats[,'colName'] == 'chewing_tobacco_snuff_or',
  c('colName','label')]
```

```
##                       colName
## 151 chewing_tobacco_snuff_or
##                                                                                label
## 151 RECODE: Used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days
```

```
smokeSub = smoke[which(smoke$Age > 10 & !is.na(smoke$Race)), ]
smokeSub$ageC = smokeSub$Age - 16
```

```
smokeModel = glm(chewing_tobacco_snuff_or ~ ageC + RuralUrban + Race + Sex, data=smokeSub, family=binom

knitr::kable(summary(smokeModel)$coef, digits=3)
```

| | Estimate | Std. Error | z value | Pr(>\|z\|) |
|---|---|---|---|---|
| (Intercept) | -2.700 | 0.082 | -32.843 | 0.000 |
| ageC | 0.341 | 0.021 | 16.357 | 0.000 |
| RuralUrbanRural | 0.959 | 0.088 | 10.934 | 0.000 |
| Raceblack | -1.557 | 0.172 | -9.068 | 0.000 |
| Racehispanic | -0.728 | 0.104 | -6.981 | 0.000 |
| Raceasian | -1.545 | 0.342 | -4.515 | 0.000 |
| Racenative | 0.112 | 0.278 | 0.404 | 0.687 |
| Racepacific | 1.016 | 0.361 | 2.814 | 0.005 |
| SexF | -1.797 | 0.109 | -16.485 | 0.000 |

```r
logOddsMat = cbind(est=smokeModel$coef, confint(smokeModel, level=0.99))
```

```
## Waiting for profiling to be done...
```

```r
oddsMat = exp(logOddsMat)
oddsMat[1,] = oddsMat[1,] / (1+oddsMat[1,])
rownames(oddsMat)[1] = 'Baseline prob'
knitr::kable(oddsMat, digits=3)
```

| | est | 0.5 % | 99.5 % |
|---|---|---|---|
| Baseline prob | 0.063 | 0.051 | 0.076 |
| ageC | 1.407 | 1.334 | 1.485 |
| RuralUrbanRural | 2.610 | 2.088 | 3.283 |
| Raceblack | 0.211 | 0.132 | 0.320 |
| Racehispanic | 0.483 | 0.367 | 0.628 |
| Raceasian | 0.213 | 0.077 | 0.466 |
| Racenative | 1.119 | 0.509 | 2.163 |
| Racepacific | 2.761 | 0.985 | 6.525 |
| SexF | 0.166 | 0.124 | 0.218 |

**Step 1: establish smokeModel**

$$\log(\frac{\mu_i}{1-\mu_i}) = \boldsymbol{X_i\beta}$$

For this GLM, I use logistic regression, where response is proportion of students using chewing tobacco, snuff or dip at least once in the last 30 days. The response is linked to a linear combination of covariates with logit link.
Covariates $X_i$ represents the age parameter(centered at 16), the rural or urban factor, and dummy variables for races, and sex(Malse as the reference level).

**Hypothesis based on TV**

If American TV is to believed, chewing tobacco is popular among cowboys, and cowboys are white, male and live in rural areas. Thus addressing the hypothesis that rural white males are the group most likely to use chewing tobacco, and there is reasonable certainty that less than half of one percent of ethnic-minority urban women and girls chew tobacco

```r
newData = data.frame(Sex = rep(c('M','F'), c(3,2)),
                     Race = c('white','white','hispanic','black','asian'),
                      ageC = 0, RuralUrban = rep(c('Rural','Urban'), c(1,4)))
smokePred = as.data.frame(predict(smokeModel, newData, se.fit=TRUE, type='link'))[,1:2]
smokePred$lower = smokePred$fit - 3*smokePred$se.fit
smokePred$upper = smokePred$fit + 3*smokePred$se.fit
smokePred
```

```
##         fit     se.fit     lower     upper
## 1 -1.740164 0.05471340 -1.904304 -1.576024
## 2 -2.699657 0.08219855 -2.946253 -2.453062
```

```
## 3 -3.427371 0.10692198 -3.748137 -3.106605
## 4 -6.053341 0.19800963 -6.647370 -5.459312
## 5 -6.041103 0.35209311 -7.097383 -4.984824
```

```
expSmokePred = exp(smokePred[,c('fit','lower','upper')])
knitr::kable(cbind(newData[,-3],1000*expSmokePred/(1+expSmokePred)), digits=1)
```

| Sex | Race | RuralUrban | fit | lower | |
|---|---|---|---|---|---|
| M | white | Rural | 149.3 | 129.6 | |
| M | white | Urban | 63.0 | 49.9 | |
| M | hispanic | Urban | 31.5 | 23.0 | |
| F | black | Urban | 2.3 | 1.3 | |
| F | asian | Urban | 2.4 | 0.8 | |
| Based | on the resu | lts, rural, w | hite mal | es have | the highest u |
| Female | minorites | fit's value a | re 2.3+2 | .4=4.7, | which divided by 1000 is smaller than the 0.5%. Thus it is reaso |