Modeling the National Youth Tobacco Survey Data

Junru Zhang

06/07/2020

Loading and Cleaning the Data

```
smokeFile = "smokeDownload.RData"
if (!file.exists(smokeFile)) { download.file("https://github.com/junruzhang/tobacco_survey/blob/master/s
 }
(load(smokeFile))
```

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

The `smoke` object is a `data.frame` containing the data. The `smokeFormats` gives some explanation of the variables. The `colName` and `label` columns of `smokeFormats` contain variable names in `smoke` and descriptions respectively.

```
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
```

```
# get rid of 9, 10 year olds and missing age and race
smokeSub = smoke[which(smoke$Age > 10 & !is.na(smoke$Race)), ]
smokeSub$ageC = smokeSub$Age - 16
```

Fitting a Generalized Linear Model

Consider the following model and set of results

```
smokeModel = glm(chewing_tobacco_snuff_or ~ ageC + RuralUrban + Race + Sex,
data=smokeSub, family=binomial(link='logit'))
summary(smokeModel)
```

```
##
## Call:
## glm(formula = chewing_tobacco_snuff_or ~ ageC + RuralUrban +
##     Race + Sex, family = binomial(link = "logit"), data = smokeSub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0196  -0.2833  -0.1677  -0.1004   3.9397
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -2.69966    0.08220 -32.843  < 2e-16 ***
## ageC             0.34134    0.02087  16.357  < 2e-16 ***
## RuralUrbanRural  0.95949    0.08775  10.934  < 2e-16 ***
## Raceblack       -1.55707    0.17171  -9.068  < 2e-16 ***
## Racehispanic    -0.72771    0.10424  -6.981 2.93e-12 ***
## Raceasian       -1.54483    0.34218  -4.515 6.34e-06 ***
## Racenative       0.11209    0.27775   0.404  0.68654
## Racepacific      1.01557    0.36089   2.814  0.00489 **
## SexF            -1.79661    0.10899 -16.485  < 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: 6235.9  on 20393  degrees of freedom
## Residual deviance: 5148.4  on 20385  degrees of freedom
##   (322 observations deleted due to missingness)
## AIC: 5166.4
##
## Number of Fisher Scoring iterations: 7
```

```
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

Interpreting the Model

The smokeModel corresponds to

$$Y_i \sim Binomial(N_i, \mu_i)$$

with the link function

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

where the response Y_i is the number of people who chew tobacco (failure) and the number of people who do not (success) out of a fixed number of trials given X_i
N_i is the total number of people (trials) with the given X_i
μ_i is the proportion of the N_i people who chew tobacco

Covariates X_i includes
X_{i1}: age of individual i, $=0$ if age $=16$, otherwise $=$ age of individual i -16

X_{i2}: the residence area of individual i, $=1$ if rural and $=0$ if urban,

$X_{i3}, ..., X_{i7}$: the race of individual i, $=1$ if race of individual i is black, hispanic, asian, native, pacific **respectively**, otherwise 0.

X_{i8}: the sex of individual i, $=1$ if female, $=0$ if male.

Transforming the Data

The data were in the log scale before the transformation.

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

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

```
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

It is worthy noting that estimate of `baseline prob` our estimated probability that a **16-year-old urban, white, male** has used chewing tobacco, snuff or dip at least once in the last 30 days. The confidence interval of baseline prob indicate that the true probability of white males who live in urban area of age 16 have used chewing tobacco, snuff or dip at least once in the last 30 days is between 5.1% and 7.6%.

Exploring the Data

If American TV is to believed, chewing tobacco is popular among cowboys, and cowboys are white, male and live in rural areas. In the early 1980s, the only Asian woman ever on North American TV was Yoko Ono, and Yoko Ono lived in a city and was never seen chewing tobacco. Is it true that rural white males are the group most likely to use chewing tobacco, and there is reasonable certainty that less than 0.5% of ethnic-minority urban women and girls chew tobacco?

Let's manipulate the data to find out more!

```
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]
# a rough 99% confidence interval
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
```

```r
expSmokePred = exp(smokePred[,c('fit','lower','upper')])

knitr::kable(cbind(newData[,-3],1000*expSmokePred/(1+expSmokePred)), digits=1)
```

Sex	Race	RuralUrban	fit	lower	upper
M	white	Rural	149.3	129.6	171.4
M	white	Urban	63.0	49.9	79.2
M	hispanic	Urban	31.5	23.0	42.8
F	black	Urban	2.3	1.3	4.2
F	asian	Urban	2.4	0.8	6.8

```r
expSmokePred
```

```
##           fit        lower       upper
## 1 0.175491596 0.1489262118 0.206795700
## 2 0.067228551 0.0525361951 0.086029795
## 3 0.032472186 0.0235615928 0.044752614
## 4 0.002349997 0.0012974296 0.004256481
## 5 0.002378933 0.0008272674 0.006840982
```

According to the output, rural white males are the group most likely to use chewing tobacco since it has the highest probability of using chewing tobacco, which is 14.93%, among all groups that are listed.

Ethnic-minority urban women and girls refers to the last two rows of the table output generated by the second-last line of code above. These two rows represent urban black females and urban asian females respectively. This implies that ethnic-minority urban women and girls have a total of $(\frac{2.3+2.4}{1000}) = 0.47\%$ chance of chewing tobacco, which is less than half of one percent $(< 0.50\%)$.

Furthermore, the 99% confidence interval of urban black females is (1.3, 4.2), which means most likely the probability of chewing tobacco is between 0.13% and 0.42% (which does not include 0.5%). However, for urban asian females, the 99% confidence interval is (0.08%, 0.68%), which does include 0.5%, so the probability for them to chew tobacco **can excceed 0.5%**.

Thus, there is only some certainty that less than half of one percent of ethnic-minority urban women and girls chew tobacco, but again, it might be more than 0.5%. Also note that we are also limited in that these results are for 16 year-olds and so should be cautious in our generalisations to the whole population.