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
```

```
smoke$everSmoke = factor(smoke$Tried_cigarette_smkg_even, levels=1:2, labels=c('yes','no'))

# Create 2-way table, remove the missings in the process
smokeSub2 <- smoke %>%
  filter(!is.na(Race),
         !is.na(everSmoke),
         !is.na(Age),
         !is.na(Grade),
         Grade != 8,
         !(Age %in% c(9,10))) %>%
  mutate(Grade_cat = Grade+5)
#At here Grade+5, because for the data, 1 actually means grade 6, 2 means grade 7...

xtabs(~smokeSub2$Grade_cat+smokeSub2$Age)
```

```
##                     smokeSub2$Age
## smokeSub2$Grade_cat   11   12   13   14   15   16   17   18   19
```

```
##                   6  1181 1630  166    8    2    0    1    0    2
##                   7    10 1178 1868  173   10    0    0    0    1
##                   8     0    3 1294 1784  191    9    1    2    0
##                   9     0    0    6 1043 1498  164   11    2    2
##                   10    0    0    0    8 1068 1529  173   15    4
##                   11    0    0    0    1    3 1054 1459  170   11
##                   12    0    0    0    0    1   12 1072 1407  146
```

```
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

```
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

```
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
```

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

```
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
Female	minorites	fit's value a	re 2.3+2	.4=4.7,

the highest u

which divided by 1000 is smaller than the 0.5%. Thus it is reaso

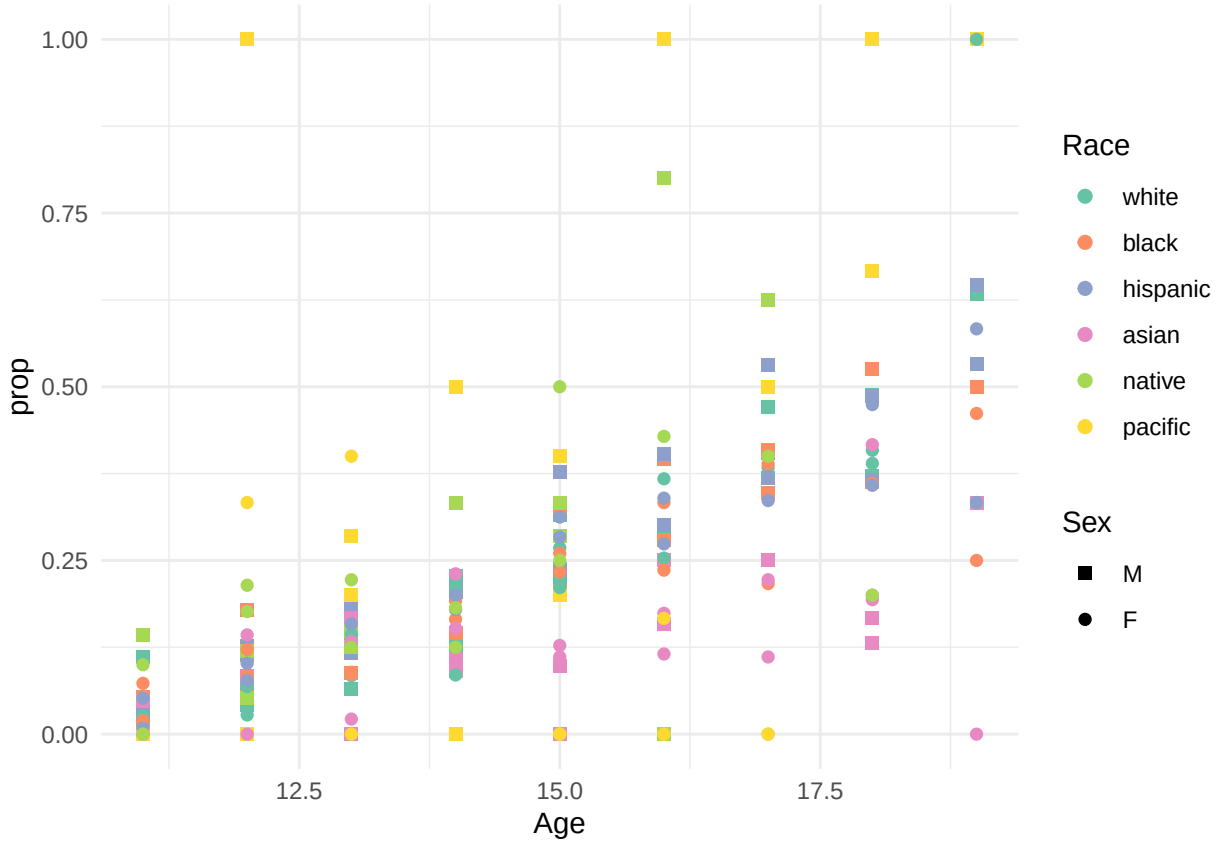
```
smokeAgg = reshape2::dcast(smokeSub,
    Age + Sex + Race + RuralUrban ~ everSmoke,
    length)
```

```
## Using ageC as value column: use value.var to override.
```

```
dim(smokeAgg)
```

```
## [1] 240   7
```

```
smokeAgg = na.omit(smokeAgg)

smokeAgg$total <- smokeAgg$yes + smokeAgg$no
smokeAgg$prop <- smokeAgg$yes/smokeAgg$total

smokeAgg %>%
  ggplot(aes(x = Age, y = prop, color = Race, shape = Sex)) +
  geom_point(size = 2) +
  scale_shape_manual(values = c(15, 16)) +
  scale_color_brewer(palette = "Set2") +
  theme_minimal()
```



From the plots, we see that older students are more likely to have tried a cigarette than youger student.

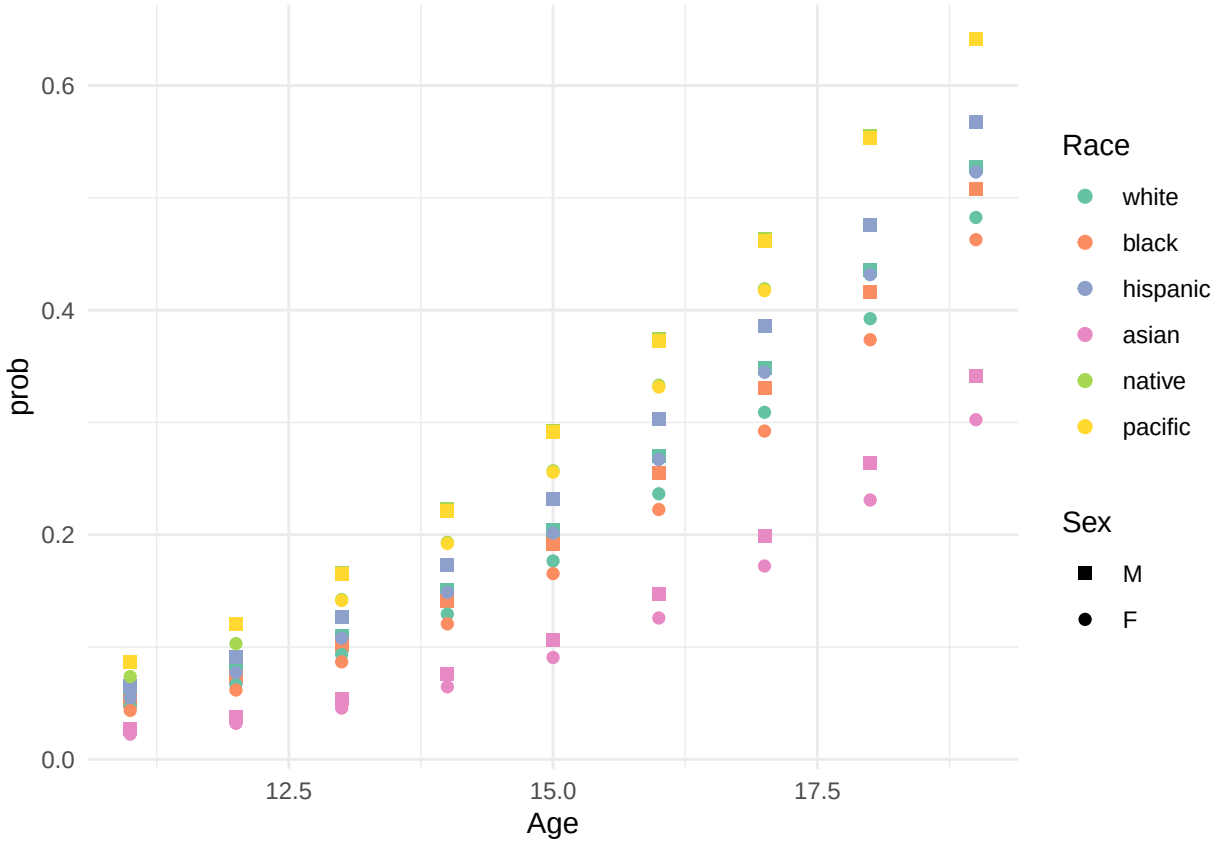
PREDICTION PLOT

```
smokeAgg$y <- cbind(smokeAgg$yes, smokeAgg$no)
smokeAgg$ageC <- smokeAgg$Age - 15

smokeFit2 <- glm(y ~ Race + Sex + Age + RuralUrban, family = binomial(link = "logit"), data = smokeAgg)
summary(smokeFit2)
```

```
##
## Call:
## glm(formula = y ~ Race + Sex + Age + RuralUrban, family = binomial(link = "logit"),
##     data = smokeAgg)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.3651  -1.0181  -0.1032   0.7826   3.0691
##
## Coefficients:
##                  Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -6.870994   0.147173 -46.686  < 2e-16 ***
## Raceblack       -0.079339   0.051246  -1.548 0.121578
## Racehispanic     0.162145   0.041854   3.874 0.000107 ***
```

```
## Raceasian       -0.765826   0.105799  -7.239 4.54e-13 ***
## Racenative       0.477840   0.138207   3.457 0.000545 ***
## Racepacific      0.470973   0.252767   1.863 0.062424 .
## SexF            -0.179244   0.035583  -5.037 4.72e-07 ***
## Age              0.367392   0.009239  39.766  < 2e-16 ***
## RuralUrbanRural  0.403191   0.036151  11.153  < 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: 2372.97  on 201  degrees of freedom
## Residual deviance:  340.03  on 193  degrees of freedom
## AIC: 1023.1
##
## Number of Fisher Scoring iterations: 4
```

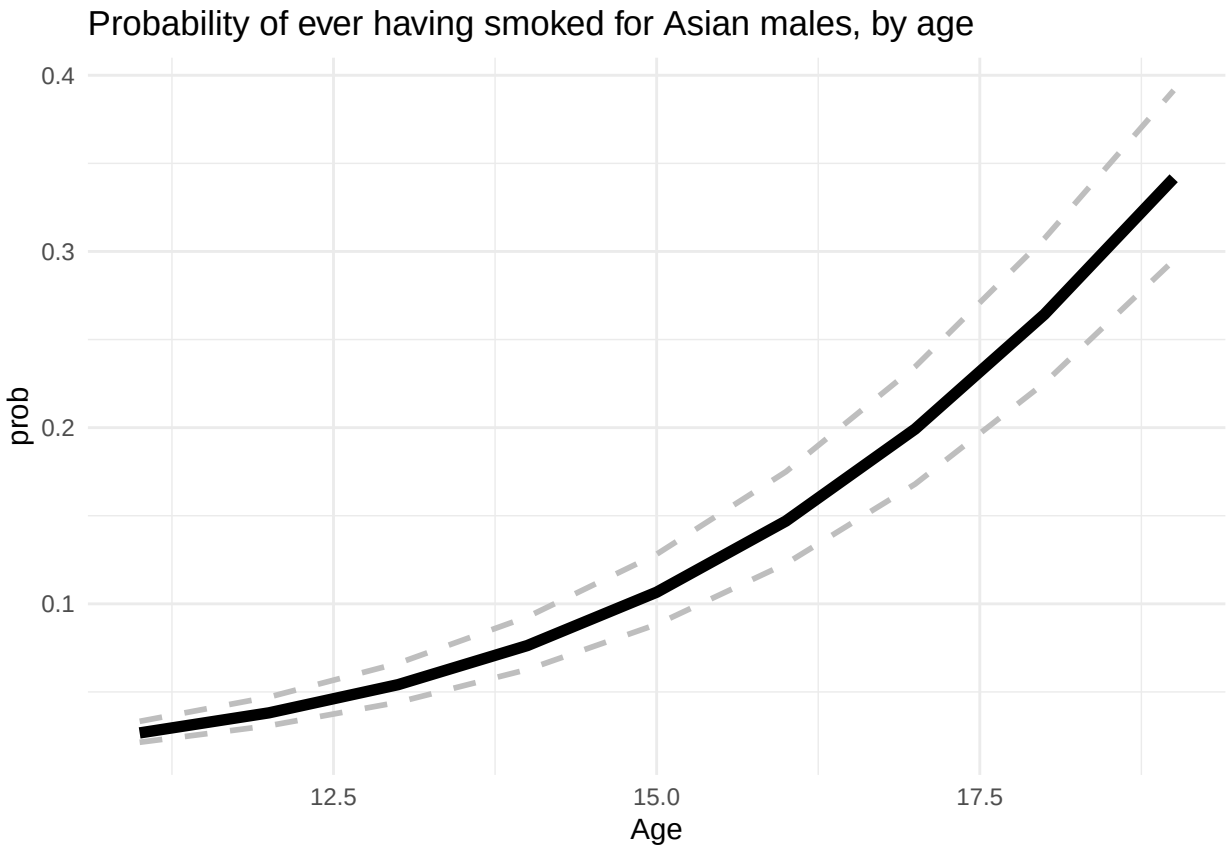
```
toPredict = smokeAgg[smokeAgg$RuralUrban == 'Urban', ] %>%
  ungroup() %>%
  mutate(id = row_number())

smokePred_tidy <- as_tibble(predict(smokeFit2, toPredict,  se.fit=TRUE)) %>%
  mutate(lower = fit - 2*se.fit,
         upper = fit + 2*se.fit) %>%
  select(fit, lower, upper) %>%
  sapply(exp) %>%
  as_tibble() %>%
  sapply(function(x) x/(1+x)) %>%
  as_tibble() %>%
  ungroup() %>%
  mutate(id = row_number()) %>%
  left_join(toPredict, by = "id")

smokePred_tidy %>%
  ggplot(aes(x = Age, y = fit, color = Race, shape = Sex)) +
  geom_point(size = 2) +
  scale_shape_manual(values = c(15, 16)) +
  scale_color_brewer(palette = "Set2") +
  ylab(label = "prob") +
  theme_minimal()
```



Two predictions:

For Asian males:

```
smokePred_tidy %>%
  filter(Sex == "M",
         Race == "asian") %>%
  ggplot(aes(x = Age, y = fit)) +
  geom_line(size = 2) +
  geom_line(aes(y = lower), lty = "dashed", color = "grey", size = 1) +
  geom_line(aes(y = upper), lty = "dashed", color = "grey", size = 1) +
  scale_shape_manual(values = c(15, 16)) +
  scale_color_brewer(palette = "Set2") +
  ylab(label = "prob") +
  theme_minimal() +
  ggtitle("Probability of ever having smoked for Asian males, by age") +
  theme(legend.position = "none")
```



For 17-year old urben men, differnt races' probability of tried smoking

```
newData = data.frame(Sex = rep("M", 5),
                     Race = c('white','native','hispanic','black','asian'),
                      Age =17, RuralUrban = rep('Urban', 5)) %>%
  mutate(id = row_number())

smokePred = as.data.frame(predict(smokeFit2, newData, se.fit=TRUE, type='link'))[,1:2]

predict(smokeFit2, newData, se.fit = TRUE, type="response")
```

```
## $fit
##         1         2         3         4         5
## 0.3485689 0.4631929 0.3862292 0.3307767 0.1992223
##
## $se.fit
##           1           2           3           4           5
## 0.009067587 0.034931049 0.010196797 0.011490797 0.016644402
##
## $residual.scale
## [1] 1
```

```
smokePred$lower = smokePred$fit - 2*smokePred$se.fit
smokePred$upper = smokePred$fit + 2*smokePred$se.fit
smokePred
```

```
##          fit     se.fit      lower      upper
```

```
## 1 -0.6253355 0.03993324 -0.7052020 -0.5454690
## 2 -0.1474954 0.14048549 -0.4284663  0.1334756
## 3 -0.4631908 0.04301426 -0.5492193 -0.3771623
## 4 -0.7046741 0.05190918 -0.8084924 -0.6008557
## 5 -1.3911620 0.10433217 -1.5998263 -1.1824976
```

```
expSmokePred = exp(smokePred[,c('fit','lower','upper')]) %>%
  mutate(id = row_number())

new_pred <- expSmokePred %>%
  left_join(newData, by = "id")

new_pred %>%
  ggplot(aes(x = Race, y = fit)) +
  geom_point() +
  geom_errorbar(aes(ymin = lower, ymax = upper)) +
  ggtitle("17-year old urben men prediction")
```