STA442 Assignment 1

Yiwen Yang (#1004244800)

Sep.23^{rd}, 2019

1. "Innocent" Fertile Female Flies

Setup of laboratory

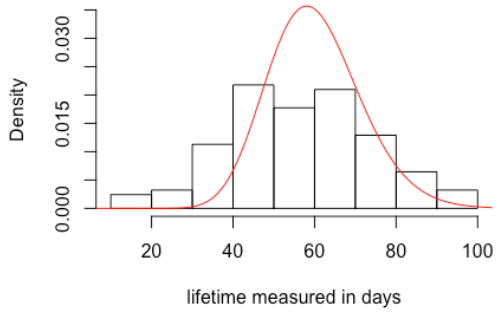
We analyzed the fruit fly data from the Faraway package, where 125 fruit flies were distributed to five groups with 25 flies in each group randomly. The first group was the control group with no female fruit flies given; whereas for the second and the third group, one and eight virgin female flies were given per day respectively. For the last two groups, one or eight pregnant females were given.

Our purpose is to find out the effect of thorax length and activity, in another word, the number of mating, to lifetime of male fruit flies measured in days. Note that pregnant females are less likely to mate comparing to virgin females. The reason why we included thorax length is due to the fact that it is highly correlated to lifetime.

Model and Interpretation

By drawing the histogram and checked its empirical distribution, we know the data roughly follow gamma distribution, please refer to Plot 1* down below.

Plot 1 (Empirical distribution of gamma distribution and histogram of longevity)*



We first normalize the thorax data, and then fit a gamma generalized linear model. Since the inverse default link is not commonly used, we change it to the log link.

$$ln\ (lifetime) = \beta_0 + \beta_1 \cdot Thoraxnew + \beta_2 I_{(activityone)} + \beta_3 I_{(activitylow)} + \beta_4 I_{(activitymany)} + \beta_5 I_{activity(high)}$$

Plot 2 (Summary of gamma GLM)*

Coefficients	Estimate	Std.error	t value	Pr(>|t|)
Intercept	4.098	0.0378	108.333	< 2e-16
thoraxnew	0.204	0.0173	11.804	< 2e-16
activityone	0.055	0.0534	1.036	0.302
activitylow	-0.116	0.0533	-2.184	0.0309
activitymany	0.083	0.0541	1.524	0.130
activityhigh	-0.415	0.0539	-7.687	4.93e-12
Null deviance: 13.2803 on 123 degrees of freedom				
Residual deviance: 4.3151 on 118 degrees of freedom				
Dispersion parameter for Gamma family taken to be 0.0355297				

From the fitted model, the 4 groups are considered as dummy variables, if a specific fly is in one of the five groups, then the value of *I* is equal to 1 with the remaining *I* = 0. We could write out five different fitted models corresponding to each group. For the control group, all *I* = 0.

To interpret, β_0 is the intercept; a unit increase in thoraxnew is associated with 1.23 times of base lifetime, which is 23% of increase. For males in group activity high, the lifetime is deduced by 34% and so on.

Turns out, the five fitted models are parallel to each other since β_1 is fixed, but with different intercepts.

The coefficients are generally statistically significant, their p-values are less than 0.05.

Plot 3 (Average lifetime of male fruit flies in different activity groups)*

1	activity	longevity
2	isolated	63.56
3	one	64.80
4	low	56.76
5	many	64.54
6	high	38.72

The mean of lifetime of males associated with different type or number of females is shown above. Not surprisingly, the lifetime of males given 1 virgin female and 8 virgin females have been deduced for about 11% and 40% respectively, comparing to the mean of lifetime of reference group with no sexual interaction. What's interesting is the 1.5% increase in lifetime from both group with 1 or 8 pregnant females.

Goodness of the fit

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$

$$H_a: \text{at least one of } \beta_i \neq 0, i = 1 \ldots, 5$$

From the summary table of gamma, we know that:

Null deviance: 13.2803 on 123 degrees of freedom

Residual deviance: 4.3151 on 118 degrees of freedom

Null deviance - Residual deviance ~ $\chi^2_{(5)}$

Since 1-pchisq (8.9652, df =5) = 0.11 > 0.05, we fail to reject H_0.

In another word, the null model may be adequate from this test.

Comments

1. Number of flies in group 'activitymany' has 24 flies, maybe that fly flies away.
2. β_2, β_4's p-values are greater than 0.05, the fitted model perhaps is not decent which is consistent to the result from the deviance test. However, from the empirical distribution, the data point roughly follows gamma, so our model should be fine.
3. The dispersion parameter for this gamma GLM is quite small (0.036), it is an under dispersion.

Summary of Result

From the analysis, the number of mating will have an enormous impact on average lifetime of male flies. In particular, total of 125 male fruit flies are stochastically assigned to 5 groups, with 25 of them in each group. The first group will be used for reference, and the remaining 4 groups are given one pregnant female, one virgin female, eight pregnant females and eight virgin females, which are named as activity one, activity low, activity many and activity high respectively. Clearly, these groups are distinguished by types of activities, since this is our main interest.

The lifetime of males decreased by 11% even with just one fertile female; as number increases to eight, the males lives 40% shorter compare to the males being isolated from females. Overwhelming number of mating may increase number of offspring, but too many interactions definitely deduces lifetime. What's surprising is the same 1.5% increase in lifetime from both activity one and activity many groups. Perhaps appropriate number of mating helps extending the lifetime.

To conclude, the words 'virgin' and 'innocence' are often being taken into consideration at the same time, but are fertile females really 'innocent'? Probably not, at least from this dataset.

Appendix

Please refer to the R-code at the very last several pages of the Appendix section.

2. Can Hookah and Water Pipe Become the Healthier Substitutes of Tobacco?

Summary of Result

We have analyzed the result from 2014 American National Youth Tobacco Survey to explore the correlation between ethnicity, age, sex and probability of chewing tobacco or trying its alternatives such as Hookah or water pipe.

Out of the three ethnic groups we are interested in, the probability for a grownup white people to have the habit of chewing tobacco is 20%, followed by Hispanic and Black people, which have approximately 8% and 5% of chance of having tobacco on a daily basis. Furthermore, as what we can expect from the trend, as age increases, the chance of chewing tobacco increases exponentially between males; whereas for females, since the average probability starts low and fluctuated only by around 2 percent for all ages, there is no clear pattern.

On the other hand, when the fancy water pipe or Hookah are everywhere on the street, it is much easier for urban youngsters and seniors to have a try. The odds of having used any of them between males and females within the same ethnical group is quite similar, the average probability of ever tried Hookah is nearly 40 percent for adults. At the same time, as age increases, odds of having alternatives of tobacco for all races and genders go up exponentially. To interpret from the ethnicity perspective, the odds of using Hookah for Hispanics and African-Americans are higher and indifferent compare to White-Americans correspondingly, which is the opposite to what we saw from the tobacco example.

Last but not least, the odds of chewing tobacco are 2.5 times higher for people living in rural area compare to urban individuals. However, the probability of trying Hookah for rural individuals is 32 percent less than urban population who tried Hookah, the reverse of the trend they have in chewing tobacco.

Introduction

We interpreted the result from 2014 American National Youth Tobacco Survey. The R version of the data is named as smokingData, which can be downloaded from pbrown.ca/appliedstats/astwo/data page. Our first main focus is to find out the relationship between race and the odds of chewing, snuff or dip tobacco. Specifically, for the following three races: White, Hispanic and African-Americans. Secondly, we would like to discover whether genders have potential impact on the probability of trying Hookah or water pipe for at least once. Since our only interest is to attain the effect of gender and race, we controlled the remaining variables, in particular, age and demographic features in order to receive the most accurate outcome of the analysis.

Methods and Interpretation

To clean up the data for enhanced analysis, we first removed the 9-year-old kids' data, since it is not a good representation of the overall dataset. Then we get rid of the missing values, and centered the variable 'Age', so the intercept is age 15.

When we are dealing with probabilities, logistic generalized linear model is the most appropriate. Down below is our fitted model.

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{Age} + \beta_2 I_{Female} + \beta_3 I_{black} + \beta_4 I_{hispanic} + \beta_5 I_{asian} + \beta_6 I_{native} + \beta_7 I_{pacific} + \beta_8 I_{Rural}$$

Note: $\frac{p}{1-p}$ is equivalent to the term odds.

Table 1 (Summary of coefficients of fitted model for ever tried Hookah analysis)*

	Estimate	Std.Error	z value	Pr(>|z|)
Intercept	-1.724	0.044	-39.226	0.000
ageC	0.419	0.012	36.266	0.000
SexF	0.042	0.043	0.980	0.327
Raceblack	-0.635	0.070	-9.005	0.000
Racehispanic	0.346	0.048	7.138	0.000
Raceasian	-0.631	0.118	-5.362	0.000
Racenative	0.160	0.190	0.838	0.402
Racepacific	0.964	0.270	3.566	0.000
RuralUrbanRural	-0.388	0.044	-8.769	0.000

From the above '*I*' here are indicators, when the condition of the indicator variable is satisfied, its value equals to one, otherwise zero. Note, it is possible that both $I_{one\ of\ ethinicity\ mentioned\ in\ the\ model}$ and I_{rural} equate to one.

To interpret the coefficient, for a unit increase in age, it is associated with $e^{\beta_1}\ times$ increase in odds and $\frac{e^{\beta_1}}{1+e^{\beta_1}}$ increase in probability. The way of interpreting the remaining coefficients is the same. The 95% confidence interval for odds when there is a unit increase in age, is (1.485, 1.551) ($e^{0.419 - 1.96 \cdot 0.012}, e^{0.419 + 1.96 \cdot 0.012}$) and so on.

One of our interests is the effect of race to odds (directly proportional to probability) of chewing tobacco.

$$H_0: \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7$$

$$H_a: \text{at least one of } \beta_i \neq 0, \text{ i} = 3\ldots,7$$

Residual deviance: 375.9 on 198 degrees of freedom
Residual deviance: 625.42 on 203 degrees of freedom

Row 1 of the above table is the deviance of full fitted model, row 2 is deviance of the fitted model without variable 'race'.

Model without variable 'race' deviance - Full model deviance ~ $\chi^2_{(5)}$

p-value: 1-pchisq (249.52, df = 5) = 0

We have very strong evidence to reject the null hypothesis. Our fitted model is adequate and at least one of the 5 coefficients of race dummy variable is statistically significant, consistent to Plot 1* where the probability curves for each ethnicity are at distinct level.

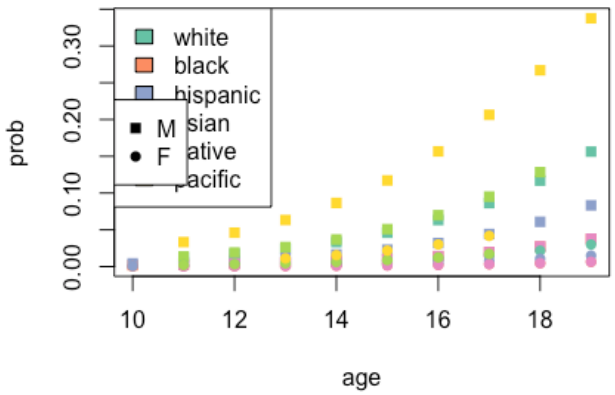
Another goal is to discover how the predictor 'gender', affect odds of trying Hookah.

$$H_0: \beta_2 = 0$$

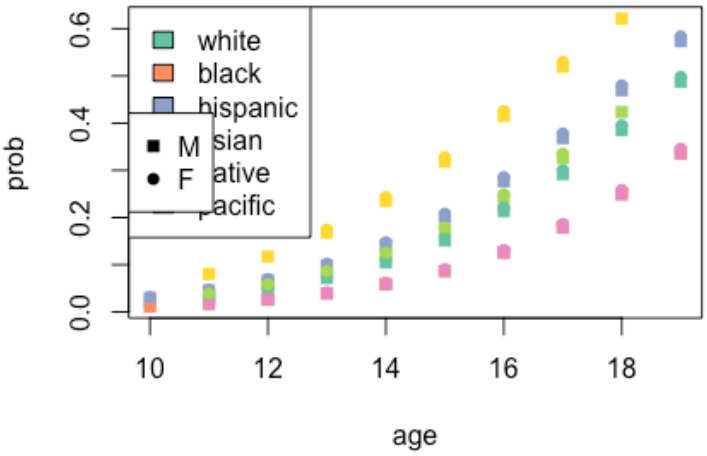
$$H_a: \beta_2 \neq 0$$

The p-value is 0.32, we fail to reject the null hypothesis. The gender variable is not statistically significant. It is consistent to what we have observed from Plot 2*, where the probability of trying Hookah between male and female for all races is indifferent.

Plot 1 (Probability of trying Tobacco among all genders and ethnicities as Age increases)*



Plot 2 (Probability of chewing Hookah among all genders and ethnicities as Age increases)*



Comments and Limitation

1. In our fitted model, even though Asians, Natives and Pacific-Americans are not the races we targeted for the statistical analysis, but they are still included in the model.

2. From the interaction model (refer from the appendix), all the p-value associated with interaction coefficient is greater than 0.05, so they are statistically insignificant. Some exceptions though, the effect of age to the race Black, Hispanic and Asian is nonnegligible (p-value = 0)； but since it is not our biggest interest, the summary of coefficients will not be shown in the report.

3. Logistic model requires independence of data. Our definition of regularly chewing tobacco is having chewed for once or more in the last month, it is likely for a person to try Hookah after tobacco because they don't like the flavor of tobacco. Resulting the data point to be correlated. Additionally, the probability of ever tried water pipe in urban regions is much higher than probability of chewing tobacco by prior analysis, so the urban area data are tending to be more dependent by the reason I stated.

Result

Our predominant interest is to discover the effect of different race to odds of chewing tobacco regularly. Refer to Plot 1*, it is clear that odds of chewing tobacco for Hispanics and Blacks is half and 1/5 respectively to odds of Whites dipping tobacco. This is consistent to the result from other researches, where White-Americans are tending to live in rural areas, and chewing tobacco is quite common around there. From this fact, the 2.5 times of odds of chewing tobacco in rural regions compare to urban area is expected. Female population that regularly chews tobacco only consists 1/5 of males.

The conclusion for the question on how gender correlates to probability of ever try Hookah is obvious from plot 2. Gender is indeed not statistically significant, trends of trying water pipe is similar to both genders. Hispanics prefer Hookah much more than tobacco, nearly 40 percent more of odds to Whites, opposite to the trends in tobacco scenario between the three races. Astonishingly, rural individuals may not be able to try water pipes that frequently, there is a 32 percent less of the population compare to urban public.

Even though we tried to minimize the effect of 'Age', the odds still dramatically increase as age goes up by one unit with 1.34 times and 1.51 times in tobacco scenario and Hookah situation correspondingly. And to answer the title question, maybe yes for urban population, but not yet in rural areas.

Appendix

Code for analyzing how male fruit flies' lifetime can be effected by activities

```
data('fruitfly', package='faraway')

summary(fruitfly)

# Here is the model without rescaling and centering variables

glm(thorax ~ longevity + activity, family=Gamma(), data=fruitfly)

# However, there are several problems.

# 1. The fitted model recognizes the longevity (lifetime) and activity as covariates,

# but our interest is the effect of thorax and activity (as covariates) to longevity (response).

# 2. Since thorax is fixed, we need to normalize it.

# Centering and rescaling the variables:

meanthorax <- mean(fruitfly$thorax)

sdthorax <- sd(fruitfly$thorax)

thoraxnew <- (fruitfly$thorax - meanthorax)/ sdthorax

# Consider the gamma model for our interest and with normalized thorax:

gamma <- glm(longevity ~ thoraxnew + activity,family = Gamma(link = log), data=fruitfly)

summary(gamma)

aggregate(longevity~activity,fruitfly,mean)

shape=1/summary(gamma)$dispersion

scale = exp(gamma$coef['(Intercept)'])/shape

hist(fruitfly$longevity,prob=TRUE,main='',xlab='lifetime measured in days', ylim = c(0,0.035))

xSeq = seq(0,120,len=200)

lines(xSeq,dgamma(xSeq,shape=shape,scale=scale),col='red')
```

Code for analyzing ever tried Hookah or water pipe

```
#' load the smoking data
#+ smokeData
```

```
smokeUrl = 'http://pbrown.ca/teaching/appliedstats/data/smoke.RData'
(smokeFile = tempfile(fileext='.RData'))
download.file(smokeUrl, smokeFile, mode='wb')
(load(smokeFile))
dim(smoke)
#'
#'
#'
#'
#+ exploreSmoke
smoke[1:5,c('Age','Sex','Grade','RuralUrban','Race', ' ever_tobacco_Hookah_or_wa')]
smokeFormats[smokeFormats$colName == ' ever_tobacco_Hookah_or_wa', ]
smoke$everSmoke = factor(smoke$ ever_tobacco_Hookah_or_wa, levels=c('TRUE','FALSE'), labels=c('yes','no'))
table(smoke$Grade, smoke$Age, exclude=NULL)
table(smoke$Race, smoke$everSmoke, exclude=NULL)
#'
#' nine year olds look suspicious
#' get rid of missings and age 9
#+ smokeSub
smokeSub = smoke[smoke$Age != 9 & !is.na(smoke$Race) &
                 !is.na(smoke$everSmoke), ]
dim(smokeSub)
#'
#'
#+
smokeAgg = reshape2::dcast(smokeSub,
                           Age + Sex + Race + RuralUrban ~ everSmoke,
                           length)
dim(smokeAgg)
smokeAgg = na.omit(smokeAgg)
dim(smokeAgg)

smokeAgg[which(smokeAgg$Race == 'white' &
               smokeAgg$Sex == 'M' & smokeAgg$RuralUrban == 'Urban'),]
smokeAgg$total = smokeAgg$no + smokeAgg$yes
smokeAgg$prop = smokeAgg$yes / smokeAgg$total
#'

#+ smokeExplPlot
Spch = c('M' = 15, 'F'=16)
Scol = RColorBrewer::brewer.pal(nlevels(smokeAgg$Race), 'Set2')
names(Scol) = levels(smokeAgg$Race)
plot(smokeAgg$Age, smokeAgg$prop, pch = Spch[as.character(smokeAgg$Sex)],
    col = Scol[as.character(smokeAgg$Race)])
legend('topleft', fill=Scol, legend=names(Scol))
legend('left', pch=Spch, legend=names(Spch))


#'
#' Which races smoke the least?
#' ... age is a confounder
#' ... as is urban/rural.
#+ smokeModel
smokeAgg$y = cbind(smokeAgg$yes, smokeAgg$no)
smokeFit = glm(y ~ Age + Sex + RuralUrban,
               family=binomial(link='logit'), data=smokeAgg)
summary(smokeFit)
knitr::kable(summary(smokeFit)$coef, digits=3)
#'
#'
#' Intercept is age zero
#' center Age so intercept is age 15
```

```
#+ smokeFit2
smokeAgg$ageC = smokeAgg$Age - 15
smokeFit2 = glm(y ~ ageC + Sex + Race + RuralUrban,
            family=binomial(link='logit'), data=smokeAgg)
knitr::kable(summary(smokeFit2)$coef, digits=3)
#'
#'
#' convert to baseline prob and odds
#+ smokeConvert
smokeTable = as.data.frame(summary(smokeFit2)$coef)
smokeTable$lower = smokeTable$fit - 2*smokeTable$se.fit
smokeTable$upper = smokeTable$fit + 2*smokeTable$se.fit

smokeOddsRatio = exp(smokeTable[,c('Estimate','lower','upper')])
rownames(smokeOddsRatio)[1] = 'baseline prob'
smokeOddsRatio[1,] = smokeOddsRatio[1,]/(1+smokeOddsRatio[,1])
smokeOddsRatio
#'
#' make row names nicer
#+ newNames
rownames(smokeOddsRatio) = gsub("Race|RuralUrban|C$", "",
                        rownames(smokeOddsRatio) )
rownames(smokeOddsRatio) = gsub("SexF","Female",
                        rownames(smokeOddsRatio))
knitr::kable(smokeOddsRatio, digits=3)

#'
#+ smokeFitted
toPredict = smokeAgg[smokeAgg$RuralUrban == 'Urban', ]
smokePred = as.data.frame(predict(smokeFit2, toPredict,  se.fit=TRUE))
smokePred$lower = smokePred$fit - 2*smokePred$se.fit
smokePred$upper = smokePred$fit + 2*smokePred$se.fit
smokePredExp = exp(smokePred[,c('fit','lower','upper')])
smokePredProb = smokePredExp / (1+smokePredExp)
#'
#+ plotFitted
plot(toPredict$Age, smokePredProb$fit,
    pch = Spch[as.character(toPredict$Sex)],
    col = Scol[as.character(toPredict$Race)],
    xlab='age', ylab='prob')
legend('topleft', fill=Scol, legend=names(Scol))
legend('left', pch=Spch, legend=names(Spch))

#'

#' asian males with error bars
#+ plotAsian
isAsianMale = toPredict$Sex == 'M' &
  toPredict$Race == 'asian'
matplot(
  toPredict[isAsianMale, 'Age'],
  smokePredProb[isAsianMale,
            c('fit','lower','upper')],
  type='l', lty=c(1,2,2), lwd=3,
  col=c('black','grey','grey'),
  xlab='Age of an Asian Male', ylab='probability of '
)


#'
```

```
#+ smokeFitInteraction
smokeFitInt = glm(y ~ ageC * Sex * Race + RuralUrban,
                  family=binomial(link='logit'), data=smokeAgg)
knitr::kable(summary(smokeFitInt)$coef, digits=3)
#'
#'
#'
#+ smokeFittedInt
smokePred = as.data.frame(predict(smokeFitInt, toPredict,  se.fit=TRUE))
smokePred$lower = smokePred$fit - 2*smokePred$se.fit
smokePred$upper = smokePred$fit + 2*smokePred$se.fit
smokePredExp = exp(smokePred[,c('fit','lower','upper')])
smokePredProb = smokePredExp / (1+smokePredExp)
#'
#+ plotFitted
plot(toPredict$Age, smokePredProb$fit,
    pch = Spch[as.character(toPredict$Sex)],
    col = Scol[as.character(toPredict$Race)],
    xlab='age', ylab='prob')
legend('topleft', fill=Scol, legend=names(Scol))
legend('left', pch=Spch, legend=names(Spch))

#'
#'
#'
#' asian males with error bars
#+ plotAsianInt
matplot(
  toPredict[isAsianMale, 'Age'],
  smokePredProb[isAsianMale,
                c('fit','lower','upper')],
  type='l', lty=c(1,2,2), lwd=3,
  col=c('black','grey','grey'),
  xlab='Age', ylab='prob'
)
```

Code for analyzing regularly chewing tobacco

```
#' load the smoking data
#+ smokeData
smokeUrl = 'http://pbrown.ca/teaching/appliedstats/data/smoke.RData'
(smokeFile = tempfile(fileext='.RData'))
download.file(smokeUrl, smokeFile, mode='wb')
(load(smokeFile))
dim(smoke)

#+ exploreSmoke
smoke[1:5,c('Age','Sex','Grade','RuralUrban','Race', 'chewing_tobacco_snuff_or')]
# or change chewing_tobacco_snuff_or to ever_tobacco_Hookah_or_wa
smokeFormats[smokeFormats$colName == 'chewing_tobacco_snuff_or', ]
smoke$everSmoke = factor(smoke$chewing_tobacco_snuff_or, levels=c('TRUE','FALSE'), labels=c('yes','no'))
table(smoke$Grade, smoke$Age, exclude=NULL)
table(smoke$Race, smoke$everSmoke, exclude=NULL)
#'
#' nine year olds look suspicious
#' get rid of missings and age 9
#+ smokeSub
smokeSub = smoke[smoke$Age != 9 & !is.na(smoke$Race) &
                   !is.na(smoke$everSmoke), ]
dim(smokeSub)
```

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

smokeAgg[which(smokeAgg$Race == 'white' &
               smokeAgg$Sex == 'M' & smokeAgg$RuralUrban == 'Urban'),]
smokeAgg$total = smokeAgg$no + smokeAgg$yes
smokeAgg$prop = smokeAgg$yes / smokeAgg$total
#'get rid of NA points

#'
#'
#+ smokeExplPlot
Spch = c('M' = 15, 'F'=16)
Scol = RColorBrewer::brewer.pal(nlevels(smokeAgg$Race), 'Set2')
names(Scol) = levels(smokeAgg$Race)
plot(smokeAgg$Age, smokeAgg$prop, pch = Spch[as.character(smokeAgg$Sex)],
    col = Scol[as.character(smokeAgg$Race)])
legend('topleft', fill=Scol, legend=names(Scol))
legend('left', pch=Spch, legend=names(Spch))
#'
#' Which races smoke the least?
#' ... age is a confounder
#' ... as is urban/rural.
#+ smokeModel
smokeAgg$y = cbind(smokeAgg$yes, smokeAgg$no)
smokeFit = glm(y ~ Age + Sex + Race + RuralUrban,
               family=binomial(link='logit'), data=smokeAgg)
knitr::kable(summary(smokeFit)$coef, digits=3)
#'
#'
#' Intercept is age zero
#' center Age so intercept is age 15
#+ smokeFit2
smokeAgg$ageC = smokeAgg$Age - 15
smokeFit2 = glm(y ~ ageC + Sex + Race + RuralUrban,
                family=binomial(link='logit'), data=smokeAgg)
knitr::kable(summary(smokeFit2)$coef, digits=3)
#'
#'
#' convert to baseline prob and odds
#+ smokeConvert
smokeTable = as.data.frame(summary(smokeFit2)$coef)
smokeTable$lower = smokeTable$fit - 2*smokeTable$se.fit
smokeTable$upper = smokeTable$fit + 2*smokeTable$se.fit

smokeOddsRatio = exp(smokeTable[,c('Estimate','lower','upper')])
rownames(smokeOddsRatio)[1] = 'baseline prob'
smokeOddsRatio[1,] = smokeOddsRatio[1,]/(1+smokeOddsRatio[,1])
smokeOddsRatio
#'
#' make row names nicer
#+ newNames
rownames(smokeOddsRatio) = gsub("Race|RuralUrban|C$", "",
                           rownames(smokeOddsRatio) )
rownames(smokeOddsRatio) = gsub("SexF","Female",
                           rownames(smokeOddsRatio))
knitr::kable(smokeOddsRatio, digits=3)
#'
```

```
#+ smokeFitted
toPredict = smokeAgg[smokeAgg$RuralUrban == 'Urban', ]
smokePred = as.data.frame(predict(smokeFit2, toPredict,  se.fit=TRUE))
smokePred$lower = smokePred$fit - 2*smokePred$se.fit
smokePred$upper = smokePred$fit + 2*smokePred$se.fit
smokePredExp = exp(smokePred[,c('fit','lower','upper')])
smokePredProb = smokePredExp / (1+smokePredExp)
#'
#+ plotFitted
plot(toPredict$Age, smokePredProb$fit,
    pch = Spch[as.character(toPredict$Sex)],
    col = Scol[as.character(toPredict$Race)],
    xlab='age', ylab='prob')
legend('topleft', fill=Scol, legend=names(Scol))
legend('left', pch=Spch, legend=names(Spch))

#'

#' asian males with error bars
#+ plotAsian
isAsianMale = toPredict$Sex == 'M' &
  toPredict$Race == 'asian'
matplot(
  toPredict[isAsianMale, 'Age'],
  smokePredProb[isAsianMale,
                c('fit','lower','upper')],
  type='l', lty=c(1,2,2), lwd=3,
  col=c('black','grey','grey'),
  xlab='Age', ylab='prob'
)


#+ smokeFitInt
smokeFitInt = glm(y ~ ageC * Sex * Race + RuralUrban,
                  family=binomial(link='logit'), data=smokeAgg)
knitr::kable(summary(smokeFitInt)$coef, digits=3)
#'
#'
#+ smokeFittedInt
smokePred = as.data.frame(predict(smokeFitInt, toPredict,  se.fit=TRUE))
smokePred$lower = smokePred$fit - 2*smokePred$se.fit
smokePred$upper = smokePred$fit + 2*smokePred$se.fit
smokePredExp = exp(smokePred[,c('fit','lower','upper')])
smokePredProb = smokePredExp / (1+smokePredExp)
#'
#+ plotFitted
plot(toPredict$Age, smokePredProb$fit,
    pch = Spch[as.character(toPredict$Sex)],
    col = Scol[as.character(toPredict$Race)],
    xlab='age', ylab='prob')
legend('topleft', fill=Scol, legend=names(Scol))
legend('left', pch=Spch, legend=names(Spch))
#'
#' asian males with error bars
#+ plotAsianInt
matplot(
  toPredict[isAsianMale, 'Age'],
  smokePredProb[isAsianMale,
                c('fit','lower','upper')],
  type='l', lty=c(1,2,2), lwd=3,
  col=c('black','grey','grey'),
  xlab='Age', ylab='prob')
```