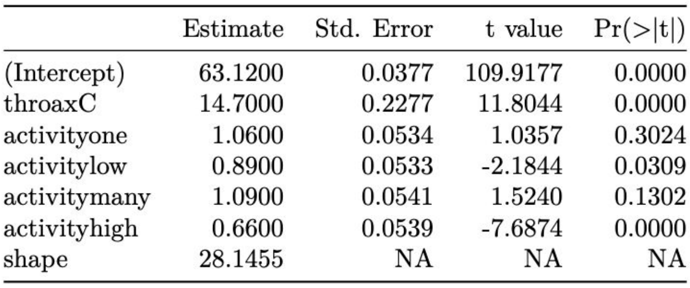
STA442 Homework 1

Question 1: Flies

In this experiment, we want to investigate the effect of thorax length and mating activity has on the expected lifespan of fruitflies. Fruitflies are placed in solitary, with one virgin female and eight virgin females. The experiment also contains two control groups, which are kept with either one pregnant female or eight pregnant females. The thorax lengths of fruitflies are also measured since it is known to affect the lifespan of fruitflies.

We use Gamma GLM to model the lifetimes as a function of the thorax length and mating activity. In the model, we have centred the predictor variable of thorax length by subtraction of 0.84 from every value in the variable. We have also scaled the variables using exp().

We get the estimated parameters below:



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	63.1200	0.0377	109.9177	0.0000
throaxC	14.7000	0.2277	11.8044	0.0000
activityone	1.0600	0.0534	1.0357	0.3024
activitylow	0.8900	0.0533	-2.1844	0.0309
activitymany	1.0900	0.0541	1.5240	0.1302
activityhigh	0.6600	0.0539	-7.6874	0.0000
shape	28.1455	NA	NA	NA

Fig 1. Estimated Parameters of the Gamma GLM

From Figure 1, we can see that:

- For the increase of every one unit length of thorax length, there will be an increase of approximately 14.70 days.
- Comparing flies kept in isolation, the flies kept with one virgin female have their lifespan reduced by 11%, whereas the flies kept with eight virgin females have their lifespan reduced by 34%.

I attach the modelled and empirical distribution below, to indicate the Gamma GLM is a good fit for the data:

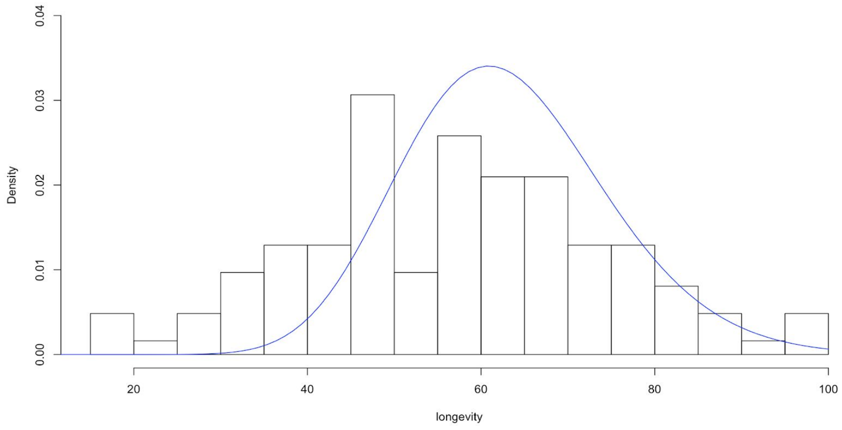


Fig 2. Empirical Distribution of the Data and the Model Fit

In conclusion, the lifespan of fruitflies is indeed affected by the presence and frequency of their mating activities. Fruitflies with no mating activities have a higher lifespan comparing to fruitflies with mating activities. Fruitflies with a lower frequency of mating activities also have a higher lifespan comparing wit fruitflies with a higher frequency of mating activities.

Question 2: Smoking

Summary

We have analyzed the 2014 American National Youth Tobacco survey, to see if race and gender are factors contributing to the increase in the use of substances such as chewing tobacco and hookah in the American youth population. From the analysis, we have found that race is a significant contributing factor in the use of chewing tobacco. The regular use of chewing tobacco, snuff or dip is around 2 times and 5 times more common amongst Americans of European ancestry than Hispanic Americans and African Americans respectively. On the other hand, gender is not a significant contributing factor in the use of hookah. The use of Hookah among women is only 4% more than men.

Introduction

In this report, we will be analyzing the 2014 American National Youth Tobacco survey, to investigate if demographics have effects on the smoking habits of the American population. The first topic we want to investigate is whether race is a factor contributing to the regular use of chewing tobacco, snuff or dip, regular use meaning the youths have used these substances on 1 or more days in the past 30 days. We are especially interested in seeing the comparison between Americans or European ancestry, Hispanic-Americans and African-Americans. The second topic we want to investigate is whether gender is a factor affecting the likelihood of having used a hookah or waterpipe on at least one occasion, given the other demographic characteristics are similar.

Methods

We will be using the logistic regression in this research, as in both topics, we want to model probabilities. For the variables in our model, we will be including age, gender, race and living area (urban or rural) of the youth.

$$\ln Odds = \beta_0 + \beta_1 x_{Age} + \beta_2 I_{Female} + \beta_3 I_{Black} + \beta_4 I_{Hisp} + \beta_5 I_{Asian} + \beta_6 I_{Native} + \beta_7 I_{Pacif} + \beta_8 I_{Rural}$$

In the first topic, the “Odds” refers to the odds of regular use of chewing tobacco, snuff or dip. Since we want to see if the race is a significant factor contributing to that odds, the null hypothesis of this topic is:

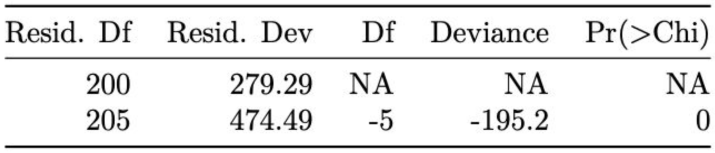
$$H_0 \colon \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7 = 0$$

In the second topic, the “Odds” refers to the odds of having used a hookah or waterpipe on at least one occasion. Given that all demographic characteristics except gender are similar, and we want to see if gender is a significant factor contributing to that odds, the null hypothesis of this topic is:

$$H_0 \colon \beta_2 = 0$$

Then, using the likelihood ratio test, we compare the likelihood of the data under the full model against the likelihood of the data under a model without race or gender respectively. From that, we can test whether the observed difference in model fit is statistically significant.

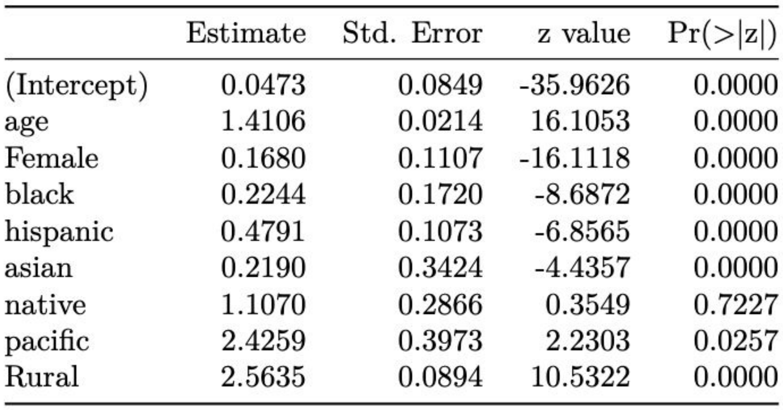
Results



Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
200	279.29	NA	NA	NA
205	474.49	-5	-195.2	0

Fig 3. Likelihood Ratio Test for Regular Use of Chewing Tobacco, Snuff or Dip

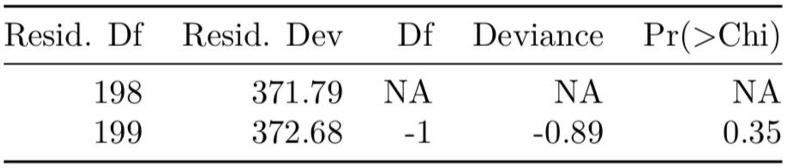
From the LRT in Figure 3, given the p-value <0.05, we can reject the null hypothesis, and conclude that race is a significant factor contributing to the odd of regular use of chewing tobacco.



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0473	0.0849	-35.9626	0.0000
age	1.4106	0.0214	16.1053	0.0000
Female	0.1680	0.1107	-16.1118	0.0000
black	0.2244	0.1720	-8.6872	0.0000
hispanic	0.4791	0.1073	-6.8565	0.0000
asian	0.2190	0.3424	-4.4357	0.0000
native	1.1070	0.2866	0.3549	0.7227
pacific	2.4259	0.3973	2.2303	0.0257
Rural	2.5635	0.0894	10.5322	0.0000

Fig 4. Odds of Regular Use of Chewing Tobacco, Snuff or Dip

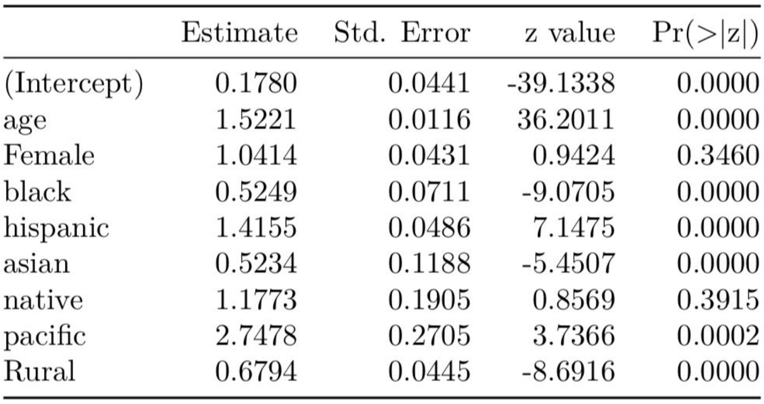
To compare the regular use of these substances amongst Americans of European ancestry, Hispanic-Americans and African-Americans, we can look at the coefficients of our model in Figure 4. The odds of regular use of these substances among Hispanic-Americans and African-Americans are 0.22 and 0.48 times, the odds of that among Americans of European ancestry respectively. It also seems that youth living in rural areas tends to use these substances more regularly than youth living in the urban area.



Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
198	371.79	NA	NA	NA
199	372.68	-1	-0.89	0.35

Fig 5. Likelihood Ratio Test for Having Used Hookah or Waterpipe

From the LRT in Figure 4, given the p-value >0.05, which is statistically not significant, and hence we cannot reject the null hypothesis. Hence it seems that the reduced model without the gender variable is adequate in fitting the data.



	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.1780	0.0441	-39.1338	0.0000
age	1.5221	0.0116	36.2011	0.0000
Female	1.0414	0.0431	0.9424	0.3460
black	0.5249	0.0711	-9.0705	0.0000
hispanic	1.4155	0.0486	7.1475	0.0000
asian	0.5234	0.1188	-5.4507	0.0000
native	1.1773	0.1905	0.8569	0.3915
pacific	2.7478	0.2705	3.7366	0.0002
Rural	0.6794	0.0445	-8.6916	0.0000

Fig 6. Odds of Having Used Hookah or Waterpipe

From Figure 6, we can see that the odds of having used Hookah or Waterpipe for a female is 1.04 times than that of a male, given all other demographic characteristics stay the same. But again given p-value >0.05, we cannot determine whether or not gender is a significant factor contributing to the odds of having used hookah or water pipes.

STA442 Homework 1 Appendix

Question 1

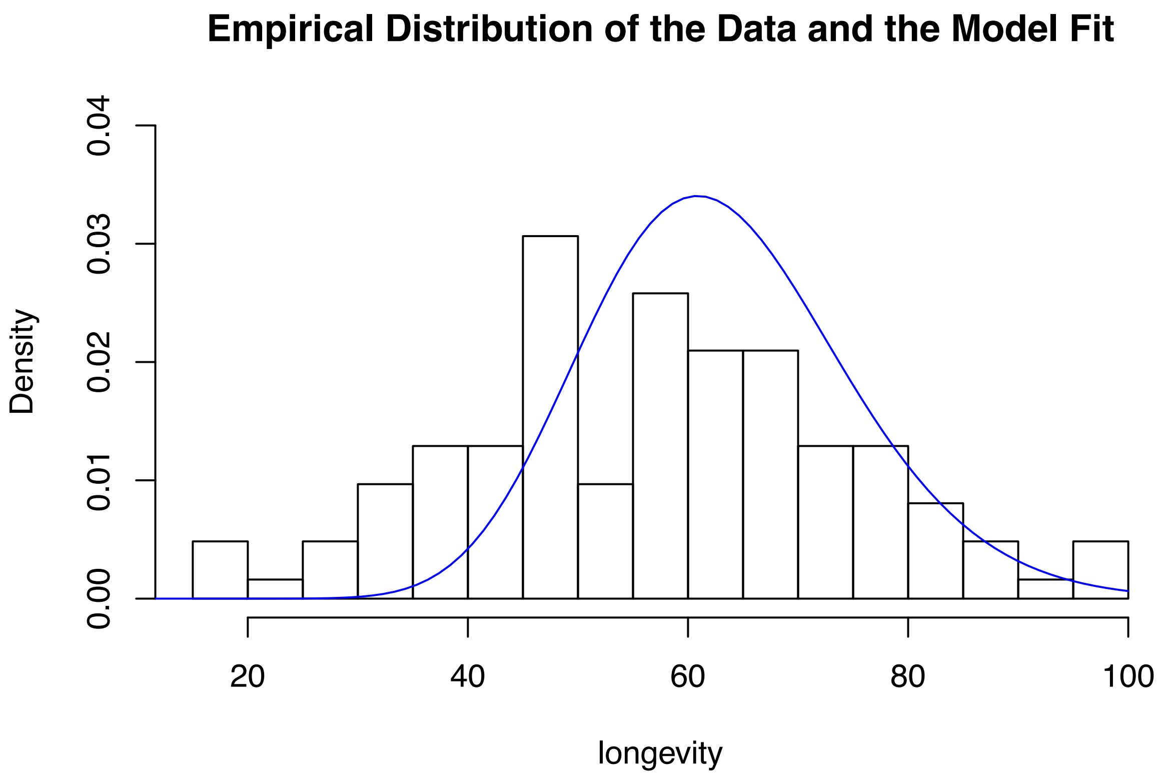
```
data('fruitfly', package='faraway')
```

```
throaxC = fruitfly$thorax - 0.84 # Use quantile(fruitfly$thorax) to center variable
mod1 = glm(longevity ~ throaxC + activity, family=Gamma(link = 'log'), data=fruitfly)
mod1coeff = summary(mod1)$coef
mod1coeff[,1] = round(exp(mod1coeff[,1]),2) # Use to rescale variables
knitr::kable(rbind(mod1coeff,shape=c(1/summary(mod1)$dispersion, NA, NA, NA)), digits=4,
             caption = "Estimated Parameters of the Gamma GLM")
```

Table 1: Estimated Parameters of the Gamma GLM

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	63.1200	0.0377	109.9177	0.0000
throaxC	14.7000	0.2277	11.8044	0.0000
activityone	1.0600	0.0534	1.0357	0.3024
activitylow	0.8900	0.0533	-2.1844	0.0309
activitymany	1.0900	0.0541	1.5240	0.1302
activityhigh	0.6600	0.0539	-7.6874	0.0000
shape	28.1455	NA	NA	NA

```
shape = 1/summary(mod1)$dispersion
scale = exp(mod1$coef["(Intercept)"])/shape
xSeq = seq(0,100,len=100)
hist(fruitfly$longevity, prob = TRUE, breaks = 20, xlab = "longevity",ylim = c(0,0.040),
     main="Empirical Distribution of the Data and the Model Fit")
lines(xSeq, dgamma(xSeq, shape = shape,scale = scale), col = "blue")
```



Question 2

```
smokeUrl = 'http://pbrown.ca/teaching/appliedstats/data/smoke.RData'
(smokeFile = tempfile(fileext='.RData'))
```

```
## [1] "/var/folders/d6/69h871p92l59n9sx36t_bz000000gn/T//RtmpDV0x4b/file33cb37bbaf98.RData"
```

```
download.file(smokeUrl, smokeFile, mode='wb')
(load(smokeFile))
```

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

```
smokeFormats[smokeFormats$colName == 'Tried_cigarette_smkg_even', ]
```

```
##     ID                                   label
## 23 qn7 Tried cigarette smkg, even 1 or 2 puffs
##                                 shortLabel                   colName
## 23 Tried cigarette smkg even 1 or 2 puffs Tried_cigarette_smkg_even
```

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

```
smokeSub = smoke[smoke$Age != 9 & !is.na(smoke$Race)
                 & !is.na(smoke$ever_tobacco_hookah_or_wa)
                 & !is.na(smoke$chewing_tobacco_snuff_or), ]
```

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

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

```
smokeAgg = na.omit(smokeAgg)
smokeAgg = smokeAgg[-7]
dim(smokeAgg)
```

```
## [1] 207   6
```

```
# smokeModel
smokeAgg$y = cbind(smokeAgg$'TRUE', smokeAgg$'FALSE')
smokeFit = glm(y ~ Age + Sex + Race + RuralUrban,
    family=binomial(link='logit'), data=smokeAgg)

# We want to scale the variable Age,
# since the center age of intercept is 15, we substract 15 from values of variable
smokeAgg$ageC = smokeAgg$Age - 15
smokeFit = glm(y ~ ageC + Sex + Race + RuralUrban,
    family=binomial(link='logit'), data=smokeAgg)
smokeTable = as.data.frame(summary(smokeFit)$coef)

# LRT
smokeFitReduced = glm(y ~ ageC + Sex + RuralUrban,
    family=binomial(link='logit'), data=smokeAgg)
knitr::kable(anova(smokeFit,smokeFitReduced,test = "Chisq"),
digits = 2,caption = "Likelihood Ratio Test of
             Regular Use of Chewing Tobacco, Snuff or Dip")
```

Table 2: Likelihood Ratio Test of Regular Use of Chewing Tobacco, Snuff or Dip

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
198	265.33	NA	NA	NA
203	427.36	-5	-162.04	0

```
# After renaming the variables and using knitr
rownames(smokeTable) = gsub("Race|RuralUrban|C$", "",
                                rownames(smokeTable) )
rownames(smokeTable) = gsub("SexF","Female",
                                rownames(smokeTable))
smokeTable[,1] = exp(smokeTable[,1])
knitr::kable(smokeTable, digits=4,
             caption = "Odds of Regular Use of Chewing Tobacco, Snuff or Dip")
```

Table 3: Odds of Regular Use of Chewing Tobacco, Snuff or Dip

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.0473	0.0849	-35.9626	0.0000
age	1.4106	0.0214	16.1053	0.0000
Female	0.1680	0.1107	-16.1118	0.0000
black	0.2244	0.1720	-8.6872	0.0000
hispanic	0.4791	0.1073	-6.8565	0.0000
asian	0.2190	0.3424	-4.4357	0.0000
native	1.1070	0.2866	0.3549	0.7227
pacific	2.4259	0.3973	2.2303	0.0257

	Estimate	Std. Error	z value	Pr(>|z|)
Rural	2.5635	0.0894	10.5322	0.0000

```
smokeAgg1 = reshape2::dcast(smokeSub,
    Age + Sex + Race + RuralUrban ~ ever_tobacco_hookah_or_wa,
    length)
```

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

```
smokeAgg1 = na.omit(smokeAgg1)
smokeAgg1 = smokeAgg1[-7]
dim(smokeAgg1)
```

```
## [1] 207   6
```

```
# smokeModel
smokeAgg1$y = cbind(smokeAgg1$'TRUE', smokeAgg1$'FALSE')
smokeFit1 = glm(y ~ Age + Sex + Race + RuralUrban,
    family=binomial(link='logit'), data=smokeAgg1)

# We want to scale the variable Age,
# since the center age of intercept is 15, we substract 15 from values of variable
smokeAgg1$ageC = smokeAgg1$Age - 15
smokeFit1 = glm(y ~ ageC + Sex + Race + RuralUrban,
    family=binomial(link='logit'), data=smokeAgg1)
smokeTable1 = as.data.frame(summary(smokeFit1)$coef)

smokeFitReduced1 = glm(y ~ ageC + Sex + RuralUrban,
    family=binomial(link='logit'), data=smokeAgg1)
knitr::kable(anova(smokeFit1,smokeFitReduced1,test = "Chisq"),
             digits = 2,caption = "Likelihood Ratio Test for Having Used Hookah or Waterpipe")
```

Table 4: Likelihood Ratio Test for Having Used Hookah or Waterpipe

Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
198	371.79	NA	NA	NA
203	625.40	-5	-253.61	0

```
# After renaming the variables and using knitr
rownames(smokeTable1) = gsub("Race|RuralUrban|C$", "",
                             rownames(smokeTable1) )
rownames(smokeTable1) = gsub("SexF","Female",
                             rownames(smokeTable1))
smokeTable1[,1] = exp(smokeTable1[,1])
knitr::kable(smokeTable1, digits=4,
             caption = "Odds of Having Used Hookah or Waterpipe")
```

Table 5: Odds of Having Used Hookah or Waterpipe

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.1780	0.0441	-39.1338	0.0000

	Estimate	Std. Error	z value	Pr(>|z|)
age	1.5221	0.0116	36.2011	0.0000
Female	1.0414	0.0431	0.9424	0.3460
black	0.5249	0.0711	-9.0705	0.0000
hispanic	1.4155	0.0486	7.1475	0.0000
asian	0.5234	0.1188	-5.4507	0.0000
native	1.1773	0.1905	0.8569	0.3915
pacific	2.7478	0.2705	3.7366	0.0002
Rural	0.6794	0.0445	-8.6916	0.0000