Applied Statitistic - Assignment 1

Peiliang Guo

Jan 1, 2017

Short Answer

Simulation study

The following chunk of code is used to simulate 100 datasets of length 40, each generated from a **Poisson distribution** with $\lambda = 0.5 + 0.2x$, with the same x vector

```
x <- seq(-10, 10, len=40)
off <- rep(c(1,-1), c(25, length(x)-25))
n <- 100
count_beta_x <- 0
beta_x <- rep(0,n)
fit_lr <- rep(0,n)
for (i in 1:n){
  set.seed(i)
  y <- rpois(length(x), exp(off + 0.5 +0.2*x))
  fit1 <- glm(y~x+offset(off), family='poisson')
  beta_x[i] <- summary(fit1)[['coefficients']][[2]]
  se <- summary(fit1)[['coefficients']][[4]]
  if (beta_x[i]+2*se>0.2 & beta_x[i]-2*se<0.2) count_beta_x <- count_beta_x+1
  fit2 <- glm(y~offset(x*0.2)+offset(off),family='poisson')
  fit_lr[i] <- 2*(logLik(fit1)-logLik(fit2))
}
```

1. When Poisson regression model is fited on each of the 100 datasets, 96 have their estimation on x within the 2 standard error confidence interval.Therefore, the coverage probability of the 2 standard error confidence interval of the cofficient on x is 0.96. This can be a indicator that the 95% confidence interval can be constructed from the 2 standard error confidence interval.

2. To test the normality of the coefficient on x across different datasets, we plot the histogram and the normal-QQ plot below.

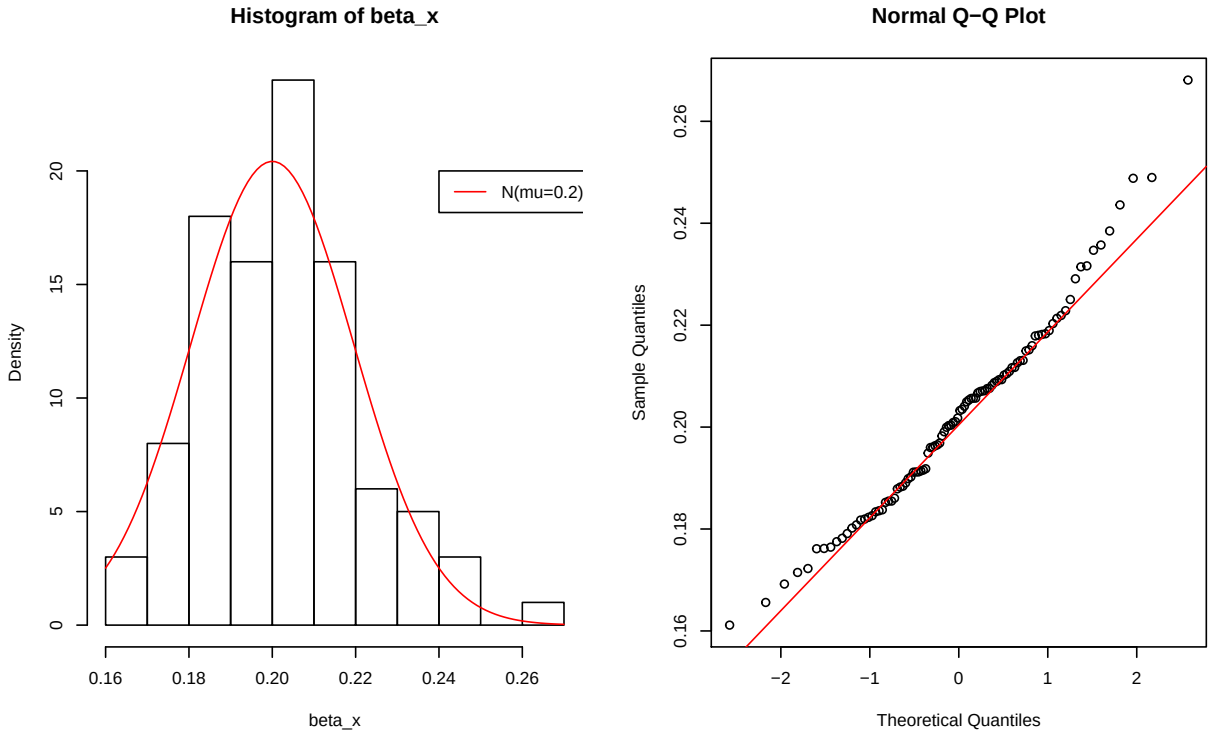

Figure 1: Histogram and Normal QQ plot of the coefficient on x from fitted Poisson Regression model on 100 simulated datasets.

From the histogram above, we observe that the distribution of coefficient on x is reasonably close the density of a normal distribution with mean at 0.2. Also, the Normal-QQ plot is close to a straight line, indicating that $\hat{\beta}$ coefficient for x is well approximated by a Normal distribution centred on 0.2.

3. To test the hypothesis of $\beta_1 = 0.2$, we perform a likelihood ratio test between the original Poisson regression model with $\lambda_i = \beta_1 x_i + \beta_0$ and the constrained Poisson regression model with $\lambda_i = 0.2x_i + \beta_0$. If the value of β_1 is indeed 0.2, the likelihood ratio should follow a χ^2_1 distribution. Within each iteration of the 100 loops, the likelihood ratio between the original model and the constrained model is calculated, and the distribution (histogram) of likelihood ratio is ploted below

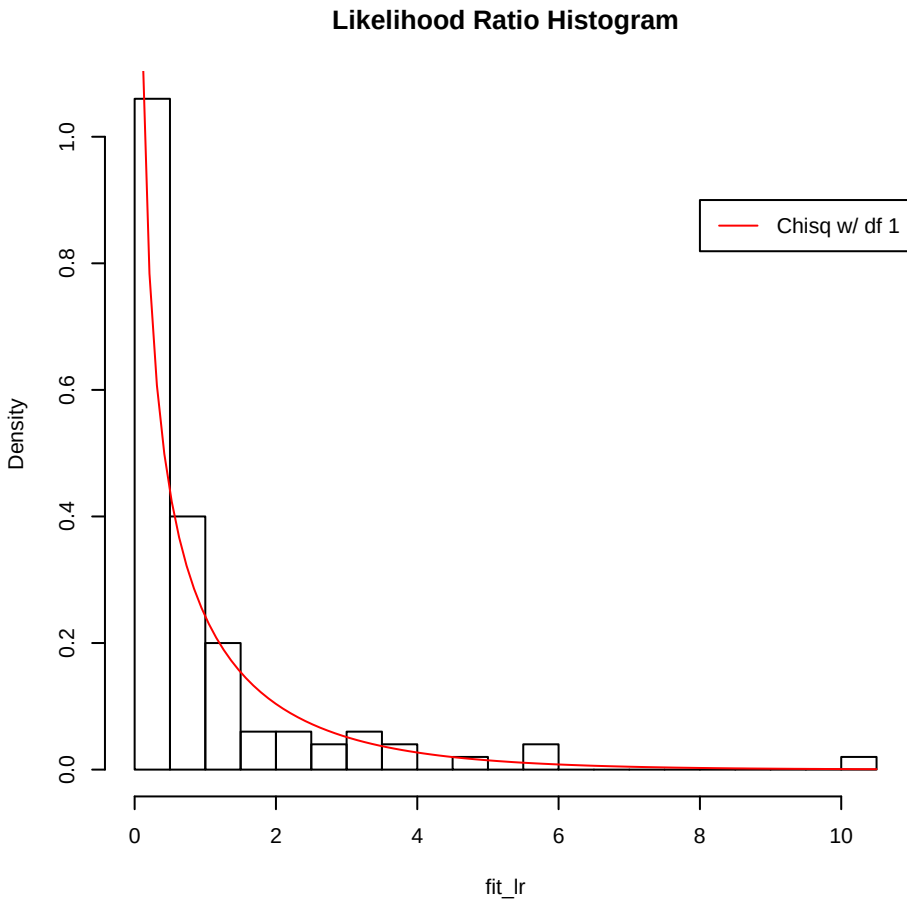

Figure 2: Histogram of the likelihood ratios between the original model with $\lambda_i = \beta_1 x_i + \beta_0$ vs. the constrained model with $\lambda_i = 0.2x_i + \beta_0$

From the figure above, the distribution of the likelihood ratio is very close to the density of the χ^2 distribution with 1 degree of freedom. This means that we accept the hypothesis of $\beta_1 = 0.2$.

Distribution functions

1.
- For zero-inflated Poisson with Poisson mean λ and extra zero proabability π,

$$\begin{aligned}
\mathbf{E}[ZIP(\pi,\lambda)] &= (1-\pi)\lambda = 2 \\
\text{Var}(ZIP(\pi,\lambda)) &= \mathbf{E}[ZIP(\pi,\lambda)^2] - \mathbf{E}[ZIP(\pi,\lambda)]^2 \\
&= (1-\pi)(\lambda^2+\lambda) - [(1-\pi)\lambda]^2 \\
&= \lambda(1-\pi)(1+\lambda\pi) = 3 \\
\Longrightarrow & 1+\lambda\pi = \frac{3}{2}, \lambda - \lambda\pi = 2 \\
\Longrightarrow & \lambda = \frac{5}{2}, \pi = \frac{1}{5}
\end{aligned}$$

Therefore, we have the Poisson mean equal to 2.5 and extra zero probability of 0.2.

- For Gamma distribution with shape α and rate β,

$$\begin{aligned}
\mathbf{E}[Gamma(\alpha,\beta)] &= \frac{\alpha}{\beta} = 2 \\
\text{Var}(Gamma(\alpha,\beta)) &= \frac{\alpha}{\beta^2} = 3 \\
\Longrightarrow & \frac{\frac{\alpha}{\beta}}{\frac{\alpha}{\beta^2}} = \frac{2}{3} \Longrightarrow \beta = \frac{2}{3} \\
\Longrightarrow & \alpha = \frac{4}{3}
\end{aligned}$$

Therefore, the shape of the Gamma distribution is 1.333 and with rate 0.667.

- For Weibull distribution with shape k and scale λ

$$\begin{aligned}
\mathbf{E}[Weibull(k,\lambda)] &= \lambda\Gamma\left(1+\frac{1}{k}\right) = 2 \\
\text{Var}(Weibull(k,\lambda)) &= \lambda^2\left[\Gamma\left(1+\frac{2}{k}\right) - \left(\Gamma\left(1+\frac{1}{k}\right)\right)^2\right] = 3 \\
\Longrightarrow & \lambda^2\Gamma\left(1+\frac{2}{k}\right) = 7 \\
\Longrightarrow & \frac{\left(\Gamma\left(1+\frac{1}{k}\right)\right)^2}{\Gamma\left(1+\frac{2}{k}\right)} = \frac{4}{7}
\end{aligned}$$

Therefore, we can use uniroot function to solve for the value of k numerically

```
f_wb <- function(k) gamma(1+1/k)^2/gamma(1+2/k)-4/7
wb_k <- uniroot(f_wb,c(0.1,2))$root
wb_lambda <- 2/gamma(1+1/wb_k)
```

From the numerical solution above, we obtain the shape k has value 1.158, and the scale λ has value of 2.106.

- For log-Normal distribution with location μ and scale σ, we have

$$\begin{aligned}
\mathbf{E}[\log \mathcal{N}(\mu, \sigma)] &= e^{\mu+\sigma^2/2} = 2 \\
\text{Var}(\log \mathcal{N}(\mu, \sigma)) &= (e^{\sigma^2} - 1)e^{2\mu+\sigma^2} = 3 \\
&\Longrightarrow e^{2\mu+\sigma^2} = 4, e^{2\mu+2\sigma^2} = 7 \\
&\Longrightarrow e^{\sigma^2} = \frac{7}{4}, \sigma = \sqrt{\log 7 - 2\log 2}, \mu = 2\log 2 - \frac{1}{2}\log 7
\end{aligned}$$

The location μ of the log-normal distribution is 0.413, and the scale σ is 0.748.

- For negative binomial distribution with r number of failures and succeed probability p, we have

$$\begin{aligned}
\mathbf{E}[NegBinom(r, p)] &= \frac{(1-p)r}{p} = 2 \\
\text{Var}(NegBinom(r, p)) &= \frac{(1-p)r}{p^2} = 3 \\
&\Longrightarrow p = \frac{2}{3} \Longrightarrow r = 4
\end{aligned}$$

The number of failures in the negative binomial distribution is 4 with success probability 0.667.

2. The density functions of log-Normal and Zero-Inflated Poisson are

```
dlognorm <- function(x,mu,sigma){
    dnorm(log(x),mu,sigma)/x
}

dzip <- function(x,lambda,pi){
  dpois(x,lambda)*(1-pi)+(x==0)*pi
}
```

The density functions of the above five distributions are plotted below

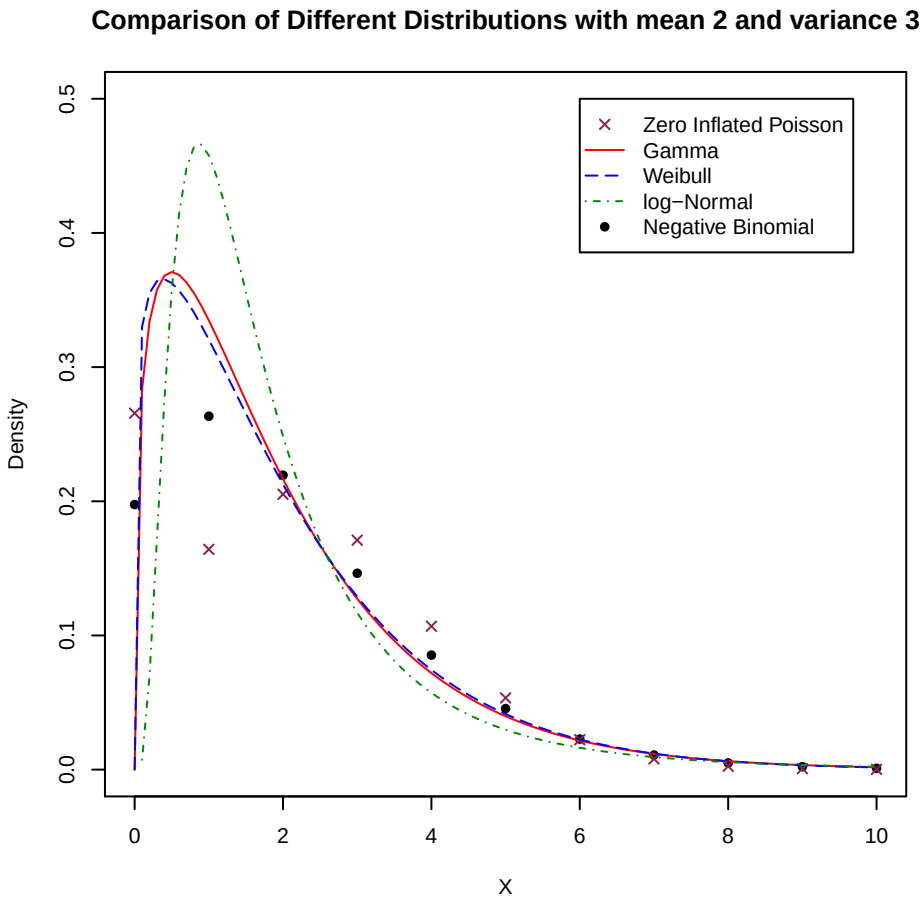

Figure 3: Comparison of ZIP, Gamma, Weibull, logNormal, and NegBinomial distributions that matches the first two moments, with mean=2 and variance=3

3. The following code will generate the 99% upper quantile for each distribution

```
qlognorm <- function(p,mu,sigma){
  exp(qnorm(p,mu,sigma))
}

qzip <- function(p,zip_lambda,zip_pi){
  p = max(1-(1-p)/(1-zip_pi),0)
  qpois(p,zip_lambda)
}

distr = c('Zero Inflated Poisson','Gamma', 'Weibull',
          'log-Normal','Negative Binomial')
q99_zip = qzip(0.99,zip_lambda,zip_pi)
q99_gamma = qgamma(0.99,gamma_alpha,gamma_beta)
q99_wb = qweibull(0.99,wb_k,wb_lambda)
```

```
q99_lnorm = qlognorm(0.99,lnorm_mu,lnorm_sigma)
q99_nbin = qnbinom(0.99,nbin_r,nbin_p)
q99 = data.frame(distr,c(q99_zip,q99_gamma,q99_wb,q99_lnorm,q99_nbin))
```

Distribution	99 Upper Quantile
Zero Inflated Poisson	7.00
Gamma	8.00
Weibull	7.87
log-Normal	8.62
Negative Binomial	7.00

Table 1: 99 Upper Quantile of Zero-Inflated Poisson, Gamma, Weibull, log-Normal, Negative Binomial distributions with mean 2 and variance 3

4. To gerate random sample from log-Normal and Zero-Inflated Poisson distributions, we define the following functions

```
rlognorm <- function(n,mu,sigma){
  exp(rnorm(n,mu,sigma))
}

rzip <- function(n,lambda,pi){
  (runif(n)>pi) * rpois(n,lambda)
}
```

The following chunk of code will randomly generate sample of size 20 from respective distributions. Sample means and variances are tabulated below.

```
n=20
zip_s = rzip(n,zip_lambda,zip_pi)
gamma_s = rgamma(n,gamma_alpha,gamma_beta)
wb_s = rweibull(n,wb_k,wb_lambda)
lnorm_s = rlognorm(n,lnorm_mu,lnorm_sigma)
nbin_s = rnbinom(n,nbin_r,nbin_p)
s_mat = rbind(zip_s,gamma_s,wb_s,lnorm_s,nbin_s)
s_mv = data.frame(distr,apply(s_mat,1,mean),apply(s_mat,1,var))
```

Distribution	Sample Mean	Sample Variance
Zero Inflated Poisson	2.60	2.57
Gamma	2.23	4.61
Weibull	2.88	5.05
log-Normal	2.21	4.80
Negative Binomial	1.80	1.33

Table 2: Sample mean and sample variance of randomly generated samples from respective distributions of size 20

Note that the sample means and sample variances are relatively far from the "true values" of mean 2 and variance 3.

Data Analysis

Before we fit Gamma glm model to the fruitfly dataset, we plot lifetimes vs. activity and thorax respectively to get a general sense of how sexual activity and thorax will effect longevity.

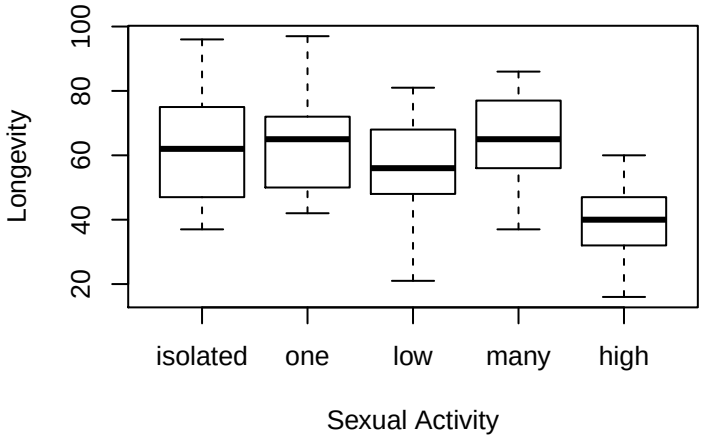

Figure 4: How longevity of fruitflies is affected by controlled sexual activities.
'isolated' = kept solitary, 'one' = one pregnant female per day, 'low' = one virgin female per day, 'many' = eight pregnant females per day, 'high' = eight virgin females per day

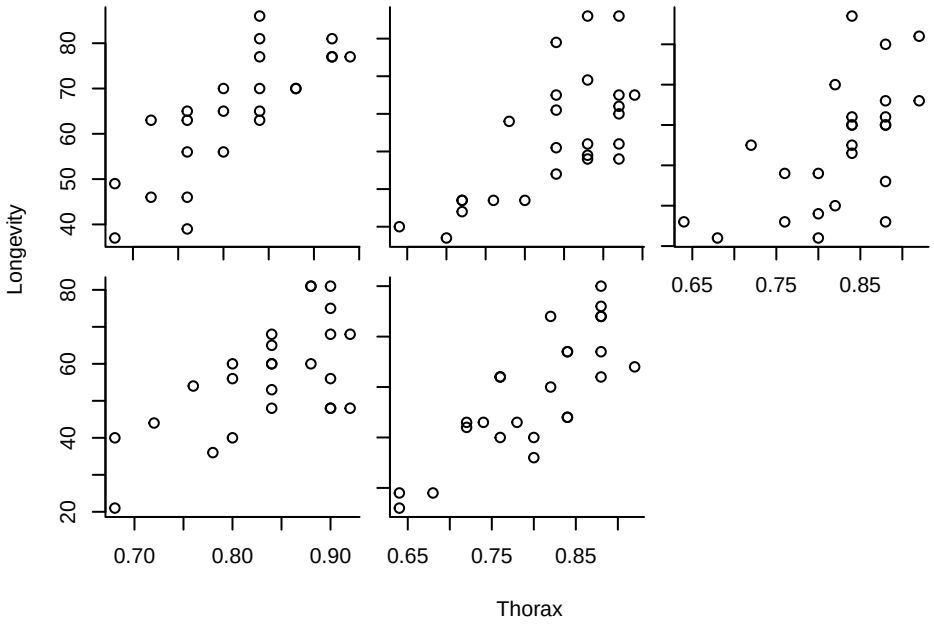

Figure 5: How longevity of fruitflies is affected by thorax, controlling for sexual activities.

Clearly, sexual activity of fruitflies will affect their longevity. To be specific, higher sexual activity will shorten longevity of fruitflies. From Figure 5 above, it is also clear that controlling for sexual activities, the longevity of fruitflies is positively correlated to thorax lenghts. The effect of interaction term between sexual activity and thorax length on longevity of fruitflies is not clear from the plots.
Since the effect of thorax length on longevity of fruitflies is already known, naturally, we are interested in the effect of sexual activity on their longevity controlling for thorax length. To test the effect formally, we first fit a Gamma glm

```
#centering thorax
ff_mthorax <- mean(fruitfly$thorax)
fruitfly$thorax <- fruitfly$thorax - ff_mthorax
ff_fit <- glm(longevity~activity + thorax,
              data=fruitfly, family = Gamma(link='log'))
```

	Estimate	Std. Error	t value	P val
(Intercept)	4.0977	0.0378	108.3326	0.0000
activityone	0.0553	0.0534	1.0357	0.3024
activitylow	-0.1165	0.0533	-2.1844	0.0309
activitymany	0.0825	0.0541	1.5240	0.1302
activityhigh	-0.4147	0.0539	-7.6874	0.0000
thorax	2.6878	0.2277	11.8044	0.0000
shape	28.1455	NA	NA	NA

Table 3: Gamma glm fit coefficient results

From the tabulated result above, we see that the intercept and thorax length are highly significant as expeccted. What is interesting is that both low sexual activity and high sexual activity are significant at $\alpha = 0.05$ level. This matches with our initial observation.
The coefficients of the Gamma glm can be interpreted as following

$\beta_0 = (Intercept)$	4.098	The life expectancy (expected longevity) of a solidary (reference case) fruitfly with average thorax length (0.822) is $e^{4.098} = 60.2$ days.
$\beta_{\text{one}} = activityone$	0.055	The life expectancy of a fruitfly kept with one pregnant female each day will increase by $e^{0.055} = 1.057$ (multiplier), when controlling for thorax length.
$\beta_{\text{low}} = activitylow$	-0.116	The life expectancy of a fruitfly kept with one virgin female each day (low sexual activity) is $e^{-0.116} = 0.89$ of the life expectancy of a fruitfly kept solitary, when controlling for thorax length.

$\beta_{\text{many}} = activitymany$	0.082	The life expectancy of a fruitfly kept with eight pregnant female each day will increase by $e^{0.082} = 1.086$, when controlling for thorax length.
$\beta_{\text{high}} = activityhigh$	-0.415	The life expectancy of a fruitfly kept with eight virgin female each day (high sexual activity) is $e^{-0.415} = 0.66057$ of the life expectancy of a fruitfly kept solitary, when controlling for thorax length.
$\beta_{\text{thorax}} = thorax$	2.688	The life expectancy of a fruitfly will increase by $e^{2.687781} = 14.699$ for per unit increase in the thorax length, when controlling for sexual activities.

To see if the Gamma glm is a good fit to the fruitfly data, we first observe that Residual Deviance of the fitted model is 4.315 where the null deviance is as large as 13.28. Also, from the boot library, we can get some diagnostic information

```
ff_diag = glm.diag(ff_fit)
```

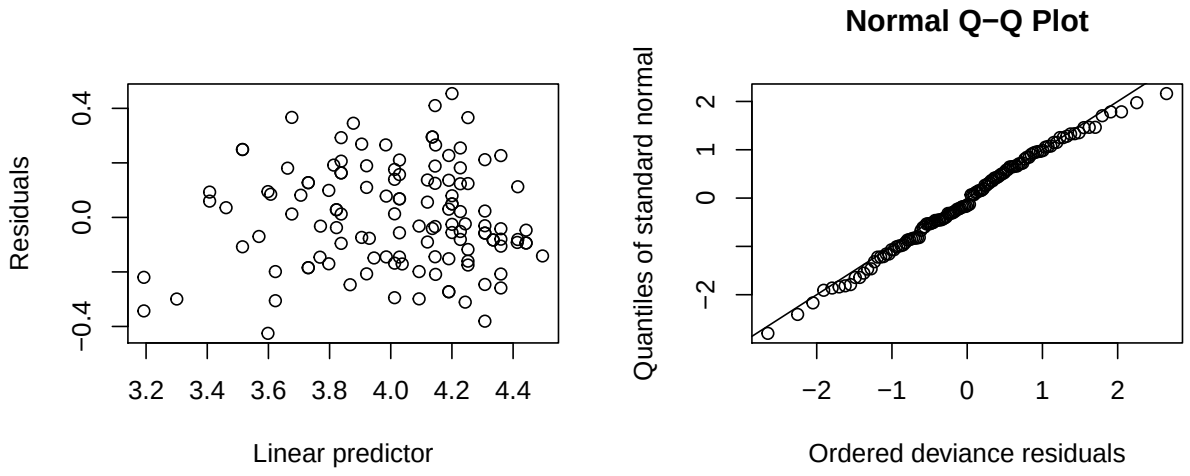

Figure 6: Diagnostic plots of Gamma glm fit on fruitfly data

From the Residual vs. Linear predictor plot above, we can see that there is no clear pattern in the plot, also the normal-QQ plot shows a very linear line. Both observations confirms that the Gamma glm model is a good fit to the fruitfly data set.
To formally test the hypothesis that *increased sexual activity will decrease life expectancy (expected longevity)*, we need to test the following contrasts simultaneously

$$H_0 : \beta_{\text{low}} - \beta_{\text{one}} = 0 \text{ and } \beta_{\text{high}} - \beta_{\text{many}} = \beta_{\text{low}} - \beta_{\text{one}} \, vs.$$

$$H_a : \beta_{\text{low}} - \beta_{\text{one}} < 0 \text{ or } \beta_{\text{high}} - \beta_{\text{many}} < \beta_{\text{low}} - \beta_{\text{one}}$$

or equivalently,

$$H_0 : \beta_{\text{low}} - \beta_{\text{one}} = 0 \text{ and } \beta_{\text{high}} - \beta_{\text{many}} - \beta_{\text{low}} + \beta_{\text{one}} = 0 \, vs.$$

$$H_a : \beta_{\text{low}} - \beta_{\text{one}} < 0 \text{ or } \beta_{\text{high}} - \beta_{\text{many}} - \beta_{\text{low}} + \beta_{\text{one}} < 0$$

The following chunck of code will perform the simultaneous contrasts in the hypothesis

```
K <- rbind(matrix(c(0,-1,1,0,0,0),1),matrix(c(0,1,-1,-1,1,0),1))
ff_ctrst <- glht(ff_fit,K,alternative='less')
ff_test <- summary(ff_ctrst)$test
```

	Estimate	Std. Error	t Statistic	P value
low - one	-0.172	0.053	-3.217	0.001
high - many - low + one	-0.325	0.076	-4.282	0.000

Table 5: contrast results of testing whether increased sexual activity will result in decreased life expectancy of fruitflies

With Bonfferroni correction, we set the significance of individual contrast to $\frac{\alpha}{2} = 2.5\%$. For each inidvidual contrast, the p-value is well below the corrected significance level. Therefore, we can reject the null hypothesis with confidence, meaning that $\beta_{\text{high}} - \beta_{\text{many}} < \beta_{\text{low}} - \beta_{\text{one}} < 0$.
To give a 95% confidence interval for each of the contrast, we use

```
ff_95ci <- ff_test$coefficient+qnorm(0.975)*matrix(c(-1,-1,1,1),2)*ff_test$sigma
```

to get that the 95% confidence interval for $\beta_{\text{low}} - \beta_{\text{one}}$ is $(-0.276, -0.067)$, and the 95% confidence interval for $\beta_{\text{high}} - \beta_{\text{many}} - \beta_{\text{low}} + \beta_{\text{one}}$ is $(-0.474, -0.176)$.
To summarize, we have shown that sexual activities will decrease life expectancy of fruitflies. In particular, after fixing the effect of thorax length, even low sexual activity will decrease the life expectancy of fruitflies by 15.8% on average. Highly frequent sexual activities will decrease the life expectancy of fruitflies by a stunning 27.8%.

Discussion

In the paper "Statistical Modelling: The Two Cultures", Breiman describes two distinct approaches towards the analysis of data. The first approach leans towards traditional statistics, which utilizes model's mechanism as an approximation of nature's mechanism, to delineate the fundamental underlying relationship of how the data is generated, and makes predictions based on the hypotheses. The other approach is through algorithmic modelling, which tries to find a useful algorithm to predict the response from a black-box approach, instead of approximating the actual relationship between predictor and response.

The question we may be asked about is that, as statisticians, under what situations, should we consider the first approach over the second. For our three research questions in the tobacco usage study, we focus on the relationships between response variables and predictors. While the algorithmic modelling approach only concerns about the prediction, it fails to provide us with the understanding of how different characteristics of an individual will effect his/her likelihood of tobacco usage. On the other hand, the data modelling approach will aid us to capture this relationship, which makes it more suitable for these research questions.

As for model selection, since the three response variables in our questions are all binary variables, it is natural to use logistic regression in all three models. To avoid overfitting to in-sample data, and to capture the real underlying relationship without being susceptible to erros, we limit the independent variables to only the ones the research questions concern. For the first research hypothesis, since the exploratory study on the dataset shows an increased effect of ethnic group in rural areas in comparison to that in urban areas, we include the interaction term between race and living area in our model, along with the two original variables. While our three research questions involve only a few predictors, according to Breiman, logistic regression should "present a simple and clear picture of nature's mechanism".

The next question that Breiman may have is how good is the fit of the model. In the plot of deviance residuals against the linear predictors, there is no evidence of nonlinearity. And the variances are also roughly constant. So the logistic models we have explored in the paper seem to be a good representation of thhe data. However, we have to keep in mind that, according to Breiman, goodness-of-fit tests and residual analysis are not quite reliable in the high dimensional case. On the other hand, Cox thinks that, sometimes, the goodness-of-fit of models may be of paramount importance, as he says that "quite often, the limitations of conclusions lie more in weakness of data quality and study design than in ineffective analysis".

Report on American National Youth Tobacco Usage

Abstract

Smoking among youths of America has been an increasingly serious issue. In this paper, we study the usage of tobacco in alternative forms other than cigarrettes and cigars. In particular, we are interested in how sex, ethnic group, residential type, and age affect the usage of chewing tobacco and hookah among junior Americans. In this paper we have shown that white Americans are heavy chewing tobacco users, and hookah usage is not related to sex.

1 Introduction

Smoking has been a major health concern for decades for its close association to more than two dozen diseases and conditons, including cancer and heart disease. More vulnerable to its deadly effects, tobacco usage among the youths is even more detrimental to American's well-being. Tobacco use is started and established primarily during adolescence - nearly 9 out of 10 cigarette smokers first tried smoking by age 18, and 99% first tried smoking by age 26. Preventing tobacco use among youth is critical to ending the tobacco epidemic in the U.S. Other than cigarrettes, many alternative forms of use of tobacco exist, each proven to lead to serious health issues. In this paper, we will investigate the usage of chewing tobacco and water pipe, also known as 'hookah' among American youths. In particular, we are intersted in the following research questions:

- Regular use of chewing tobacco is no more common amongst Americans of European ancestry than for Hispanic-Americans and African-Americans, once one accouts for the fact that white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon.
- The likelihood of having used a hookah or waterpipe on at least one occasion is the same for two individuals of the different sexes, provided their age, eithnicity, and other demographic characteristics are similar.

As a secondary problem, this paper is also interested in quantifying how the use of chewing tobacco changes with age, sex, and ethnic group.

2 Methods and Exploratory Analysis

2.1 Data Description

The data of this research paper comes from the 2014 American National Youth Tobacco Survey. The raw dataset contains observations from 22007 youths aged from 9 to 19, each with 162 variables, including basic demographics data such as age, race, sex, as well as information regarding usage of tobacco. Since the dependent variables (usage of chewing tobacco and hookah) are binary categorical in both primary research hypotheses and the secondary problem of interest, we use binomial GLM, logistic regression in particular, in all cases.

2.2 First Primary Research Hypothesis

To test the first research hypothesis, the mean responses for each combination of races of interest and residential conditons are plotted.

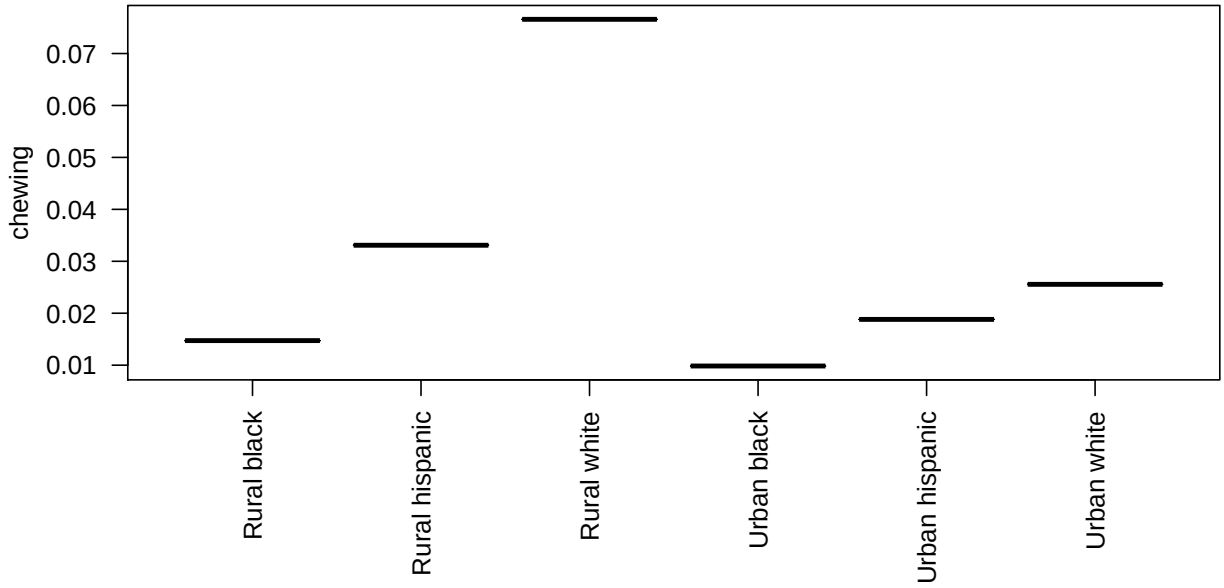

Figure 7: Proportion of chewing tobacco usage in each combination of racial groups and residential conditions among U.S. youths.

From the figure above, it is obvious that Caucasions are more likely to use chewing tobacco in both rural and urban areas. We also see that not only residential type will affect the overall usage of chewing tobacco, but it also has different effects on different racial groups. Therefore, the regression model will need to incorporate an interaction term between race and residential type. To test the hypothesis formally, we fit the following GLM, denoted as Model 1,

$$Y_i \sim Bernoulli(\mu_i), \text{where}$$
$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 I_b + \beta_2 I_h + \beta_3 I_r + \beta_4 I_b I_r + \beta_5 I_h I_r$$

where the indicator variables are elucidated in the following table

Race	I_b	I_h
White	0	0
Black	1	0
Hispanic	0	1

Residential Type	I_r
Urban	0
Rural	1

Table 6: Encoding table of dummy variables for racial groups and residential types in logistic regression model

Under this encoding scheme, the conclusion of white American youths are more likely to use chewing tobacco can be tested if each of the following one-sided hypotheses is rejected: $\beta_1 = 0$, $\beta_2 = 0$, $\beta_1 + \beta_4 = 0$, and $\beta_2 + \beta_5 = 0$.

2.3 Second Primary Research Hypothesis

To test the second hypothesis of effect of sex on usage of hookah when fixing other demographic characteritics, we first plot the following exploratory analysis to spot any obvious interactions

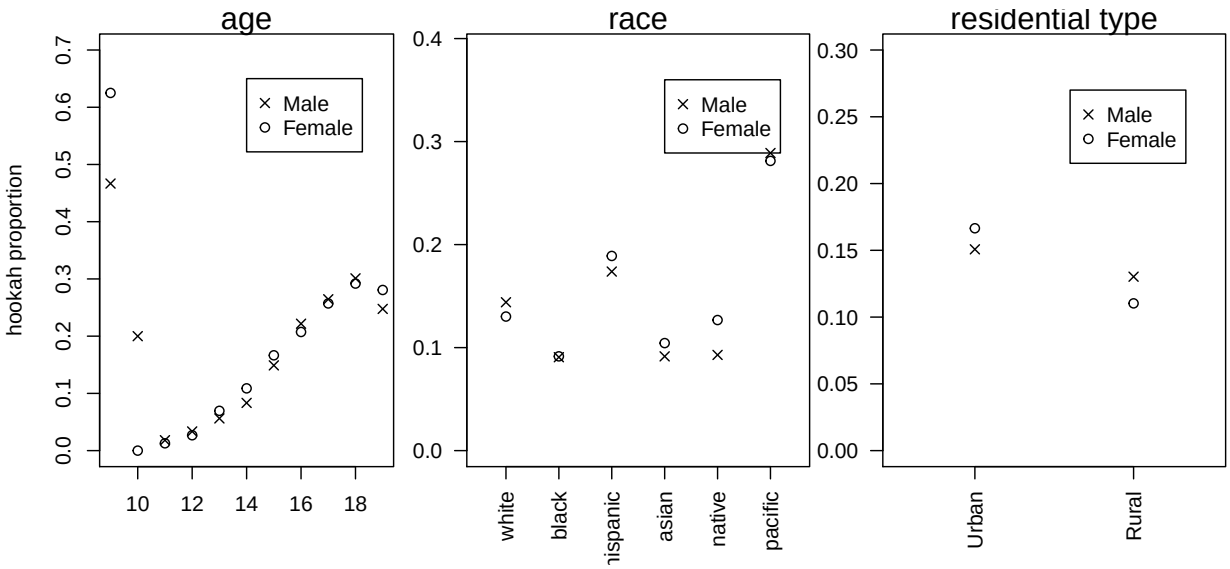

Figure 8: Proportion of hookah usage in different age groups, racial groups, and residential types among U.S. youths.

First, from the figure above, we see anormallies at age of 9 and 10. This might be caused by juniors at a very young age are not answering the surveys responsbily, and since there are only 32 observations from youths between age 9 and 10, we decide to take them out in our study. Furthermore, from the figure above, we do not observe significant interactions between sex and other demographic information, so we do not consider interaction terms in our logistic regression model, denoted as Model 2:

$$Y_i \sim Bernoulli(\mu_i), \text{where}$$
$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 I_f + \beta_2 X_a + \beta_3 I_b + \beta_4 I_h + \beta_5 I_a + \beta_6 I_n + \beta_7 I_p + \beta_8 I_r$$

where X_a is the covariate variable of Age, and the indicator variables are elucidated in the following table

Sex	I_f
Male	0
Female	1

Race	I_b	I_h	I_a	I_n	I_p
White	0	0	0	0	0
Black	1	0	0	0	0
Hispanic	0	1	0	0	0
Asian	0	0	1	0	0
Native	0	0	0	1	0
Pacific	0	0	0	0	1

Residential Type	I_r
Urban	0
Rural	1

Table 7: Encoding table of dummy variables for racial groups and residential types in logistic regression model

Under this model, our hypothesis of interest can be tested with $H_0 : \beta_1 = 0$ vs. $H_a : \beta_1 \neq 0$.

2.4 Quantifying Chewing Tobacco Usage

To quantify how chewing tobacco changes with age, sex, and racial group, we use a very similar model as above, denoted as Model 3:

$$Y_i \sim Bernoulli(\mu_i), \text{where}$$
$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 I_f + \beta_2 X_a + \beta_3 I_b + \beta_4 I_h + \beta_5 I_a + \beta_6 I_n + \beta_7 I_p$$

where X_a is the covariate variable of Age, and the indicator variables follow exactly the same definition from the the previous model. Also note that since we are interested in the actual level of odds ratios, we shift Age to have zero at 11.

3 Results and Discussion

3.1 First Primary Research Hypothesis

The tabulated R results are listed below

	Estimate	Std. Error	z value	p val
(Intercept)	-3.641	0.095	-38.205	0.000
raceblack	-0.971	0.249	-3.891	0.000
racehispanic	-0.313	0.154	-2.030	0.042
ruralRural	1.151	0.108	10.643	0.000
raceblack:ruralRural	-0.743	0.336	-2.208	0.027
racehispanic:ruralRural	-0.572	0.200	-2.856	0.004

Table 8: Coefficient summary of Model 1 fit

From the table above, we notice that every term is significant with at a level of $\alpha = 0.05$, which means the usage of tobacco is significantly related to both racial group (among White,

Black and Hispanic population) and residential type. In particular, since the coefficients of interaction terms are also negative, living in rural areas will enlarge the effect of racial identities on tobacco usage.
For Caucasians living in urban area, the odds of using chewing tobacco is 0.026. More importantly, the odds ratio between urban Black and Hispanic Americans compared to White Americans is 0.379 and 0.731, repectively. In rural areas, the odds ratio between Black and Hispanic Americans compared to White Americans is 0.18 and 0.413, repectively. In both cases, the odds ratios are significantly less than 1. The odds of using chewing tobacco of possible combinations of racial groups and residential types are listed below:

	White	Black	Hispanic
Urban	0.026	0.010	0.019
Rural	0.083	0.015	0.034

Table 9: Odds of chewing tobacco usage of American youths

The results of contracts outlined in section 2.2 are tabulated below

	Estimate	Std. Error	t stat.	p val.	95% CI lower	95% CI upper
β_1	-0.971	0.249	-3.891	0	-1.46	-0.482
β_2	-0.313	0.154	-2.03	0.042	-0.616	-0.011
$\beta_1 + \beta_4$	-1.714	0.226	-7.592	0	-2.156	-1.271
$\beta_2 + \beta_5$	-0.885	0.127	-6.95	0	-1.135	-0.636

Table 10: Linear contrasts of model pamameters showing Caucasian Americans are more likely to use chewing tobacco in comparison to African and Hispanic Americans, in both urban and rural settings

From table above, all the test contrasts are significantly lower than 0. This concludes that we can safely reject the first research hypothesis, i.e. usage of chewing tobacco is in fact more common amongst Americans of European ancestry than for Hispanic-Americans and African-Americans, even after accounting for the fact that white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon.

3.2 Second Primary Research Hypothesis

The coefficients of glm in Model 2 are summarized in the table below

	Estimate	Std. Error	t value	p val.
(Intercept)	-0.490	0.017	-28.357	0.000
age	0.045	0.001	39.129	0.000
sexF	0.004	0.005	0.913	0.361
raceblack	-0.058	0.007	-8.604	0.000
racehispanic	0.044	0.006	7.899	0.000
raceasian	-0.062	0.011	-5.384	0.000
racenative	0.016	0.019	0.876	0.381
racepacific	0.138	0.038	3.633	0.000
ruralRural	-0.043	0.005	-8.978	0.000

Table 11: Coefficient summary of Model 2 fit

Our second search hypothesis focuses on the effect of gender on usage of hookah. From the results of Model 2 coefficients above, we can see that while controlling for other demographic variables, the p-value of β_1 is 0, which is substantially larger than the normal significance level of $\alpha = 0.05$. The 95% confidence interval for β_1 is $(0.042, 0.047)$. This means that the effect of gender is not significant statistically. Therefore, reject the null hypothesis of $\beta_1 = 0$ at 0.05 significance level. This means that the likelihood of having used a hookah on at least one occasion is not significantly different for two individuals of different sexes, provided their age, ethnicity, and residential conditions are similar.

3.3 Quantifying Chewing Tobacco Usage

In previous section, we have already shown that white Americans are more likely to use chewing tobacco relative to African and Hispanic Americans. In this section, we will include more ethnic groups as well as the effect of gender and age to quantify the use of chewing tobacco. The effect of each variable on the usage of chewing tobacco is plotted first. Since chewing tobacco data suffer from the same anomalies with youths of age 9 and 10, we exclude them from our analysis

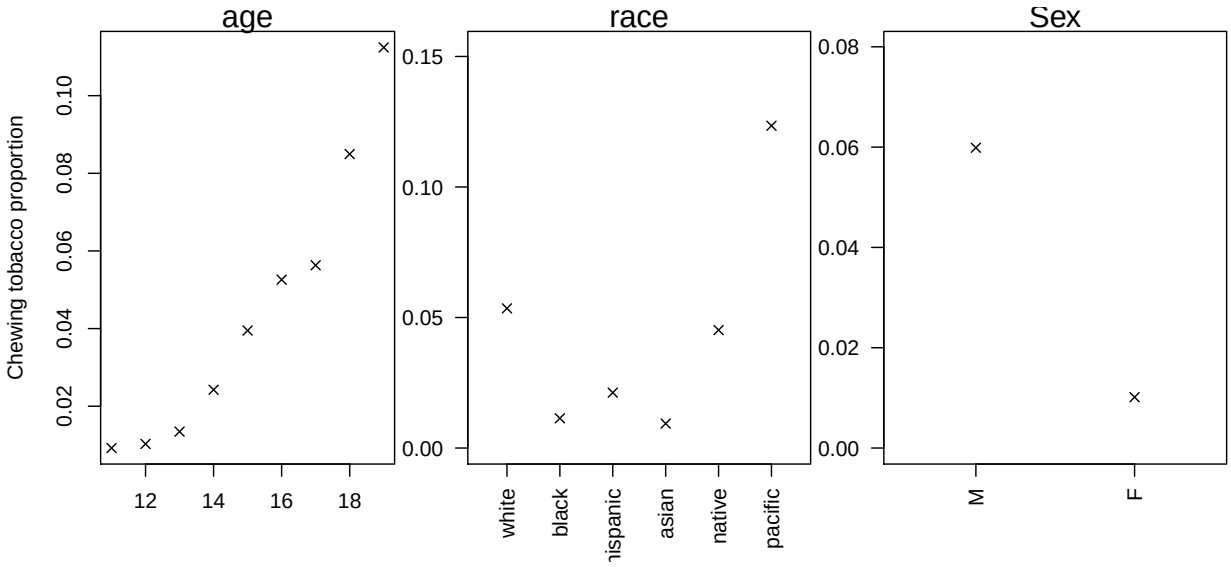

Figure 9: Proportion of chewing tobacco usage in different age groups, racial groups, and sexes among U.S. youths.

From the figure above, we can identify that the usage of chewing tobacco increases as age increases; amongst different ethnic groups, Pacific Islander Americans are the heaviest chewing tobacco users, and Caucasian Americans and Native Americans also use more chewing tobacco than Black, Hispanic, and Asian Americans. Also, chewing tobacco usage is much more common among males than females. To further quantify the odds of chewing tobacco usage, the coefficients of Model 3 fit are tabulated below in Table 12

	Estimate	Std. Error	z value	p val.
(Intercept)	-3.819	0.112	-34.220	0.000
sexF	-1.793	0.109	-16.499	0.000
age	0.351	0.021	16.850	0.000
raceblack	-1.672	0.171	-9.777	0.000
racehispanic	-0.904	0.103	-8.806	0.000
raceasian	-1.877	0.340	-5.517	0.000
racenative	0.157	0.276	0.569	0.569
racepacific	0.858	0.361	2.381	0.017

Table 12: Coefficient summary of Model 3 fit

From the table above, we can calculate that a 11 year-old White male has an odds of 0.022 of using chewing tobacco. Controlling for other variables, females have an odds ratio of 0.166 to males; the odds will increase by a ratio of 1.421 per year increase in age. As for racial groups, the odds ratio of using chewing tobacco of a Black, a Hispanic, an Asian, and a Pacific Islander American raltive to a White American is 0.188, 0.405, 0.153, and 2.359 respectively. There is not a significant difference between a White American and a Native American in terms of chewing tobacco usage.

Summary

Chewing tobacco has been proven to be a rural phenomenon, and white Americans are more likely to live in rural areas. This has been the explaination of why white Americans have been pictured as heavier chewing tobacco users. However, in this paper, we have shown that even after considering the difference between living in urban or rural areas, white Americans are still more likely to use chewing tobacco in comparason to Hispanic and black Americans. In this paper, we have a also discovered that the usage of hookah, a type of water pipe tobacco, is approximately the same among male and female youth population.
Last but not least, males are more likely to use chewing tobacco, and the usage of chewing tobacco increases with age. Also, Pacific Islanders are the heaviest chewing tobacco users, and then it comes white Americans and Native Americans. Asian Americans, Black Americans, and Hispanic Americans use the least amount of chewing tobacco.

Appendix

```
library(multcomp)
load('smoke.RData')
smoke <- smoke[,c('Age','Sex','Race','RuralUrban',
    'chewing_tobacco_snuff_or','ever_tobacco_hookah_or_wa')]
colnames(smoke) = c('age','sex','race','rural','chewing','hookah')

#testing if race (white over hispanic and black) has effect
#on usage of chewing tobacco, after fixing residential condition
smoke1 <- smoke[smoke$race %in% c('white','hispanic','black') &
              !is.na(smoke$chewing),c('race','rural','chewing')]
chewing_group_mean <- aggregate(chewing~race*rural, data=smoke1, FUN=mean)
chewing_group_mean$group <-
  as.factor(paste(chewing_group_mean$rural,chewing_group_mean$race))

smoke1_fit <- glm(chewing ~ race+rural+race*rural,
                  family="binomial", data=smoke1)
smoke1_M <- rbind(matrix(c(1,0,0,0,0,0),1),matrix(c(1,1,0,0,0,0),1),
                  matrix(c(1,0,1,0,0,0),1),matrix(c(1,0,0,1,0,0),1),
                  matrix(c(1,1,0,1,1,0),1),matrix(c(1,0,1,1,0,1),1))
smoke1_odds <- as.data.frame(t(matrix(exp(smoke1_M %*%
                          smoke1_fit$coefficients),ncol=2,nrow=3)))
rownames(smoke1_odds)=c('Urban','Rural')
colnames(smoke1_odds)=c('White','Black','Hispanic')

smoke1_K <- rbind(matrix(c(0,1,0,0,0,0),1),matrix(c(0,0,1,0,0,0),1),
                  matrix(c(0,1,0,0,1,0),1),matrix(c(0,0,1,0,0,1),1))
smoke1_ctrst <- glht(smoke1_fit,smoke1_K)
smoke1_95ci <- summary(smoke1_ctrst)$test$coefficient+qnorm(0.975)*
```

```
      matrix(c(-1,-1,-1,-1,1,1,1,1),4)*summary(smoke1_ctrst)$test$sigma

#testing if sex has effect on usage of hookah, after fixing other
#demographic info
smoke2 <- smoke[,c('age','sex','race','rural','hookah')]
smoke2 <- smoke2[complete.cases(smoke2),]
hookah_age_mean <- aggregate(hookah~sex*age, data=smoke2, FUN=mean)
hookah_race_mean <- aggregate(hookah~sex*race, data=smoke2, FUN=mean)
hookah_rural_mean <- aggregate(hookah~sex*rural, data=smoke2, FUN=mean)
smoke2_fit <- glm(hookah ~ age + sex + race + rural, data=smoke2)

smoke3 <- smoke[,c('sex','age','race','chewing')]
smoke3 <- smoke3[complete.cases(smoke3) & smoke3$age>10,]
chewing_age_mean <- aggregate(chewing~age, data=smoke3, FUN=mean)
chewing_race_mean <- aggregate(chewing~race, data=smoke3, FUN=mean)
chewing_sex_mean <- aggregate(chewing~sex, data=smoke3, FUN=mean)
smoke3$age <- smoke3$age-11
smoke3_fit <- glm(chewing~sex+age+race,data=smoke3,family = 'binomial')
```