STA442 Homework1

SongQi Wang 1003439442

2019-10-05

Fruitflies

Introduction

In the dataset from Faraway, one hundred twenty-five fruit flies were divided randomly into five groups of 25 each. The response was the lifetime of the fruit fly in days. One group was kept solitary, while another was kept individually with a virgin female each day. Another group was given eight virgin females per day. As an additional control the fourth and fifth groups were kept with one or eight pregnant females per day (pregnant fruit flies will not mate). The thorax length of each male was measured as this was known to affect lifetime.

Method

Here, we are using a Gamma generalized linear model to model the lifetimes as a function of the thorax length and activity, and our model would be:

$$\ln Longevity = \beta_0 + \beta_1 x_{Thorax} + \beta_2 I_{one} + \beta_3 I_{low} + \beta_4 I_{many} + \beta_5 I_{high}$$

Results

After we fit a Gamma generalized linear model, we found that the longer the thorax is, the longer the fly will live. Also, we can see that if a fly is isolated, it will live longer. The fruitflies with the hightest activity have lower longevity. This can be inferred from the exponentiated parameter estimates given in table 1. Variables are centered and rescaled before being fit to the model. Although the activity level of one and many has a positive beta, the coresponding p-values are higher than 0.1, which means they are not realiable.

Table 1: Estimated parameters of GLM model

	Exp. Estimate	Std. Error	t value	P-Value
Intercept	4.098	0.038	108.333	0.000
Thorax Length	0.204	0.017	11.804	0.000
Activityone	0.055	0.053	1.036	0.302
Activitylow	-0.116	0.053	-2.184	0.031
Activitymany	0.082	0.054	1.524	0.130
Activityhigh	-0.415	0.054	-7.687	0.000

Other obsivations are consistent with our medol.

Figure 1 is the Scatter plot of the dataset. Genrally, higher value of thorax leads to higher longevity. Also, we can see purple points, who are the fruitflies with highest activity, have lower longevity compered to other points. The scatter plot is consistent with our model.

Figure 2 shows the distribution of data compare to the Gamma generalized linear model we fit, it shows that our model fit the data pretty well.

Figure1: Longevity vs Thorax

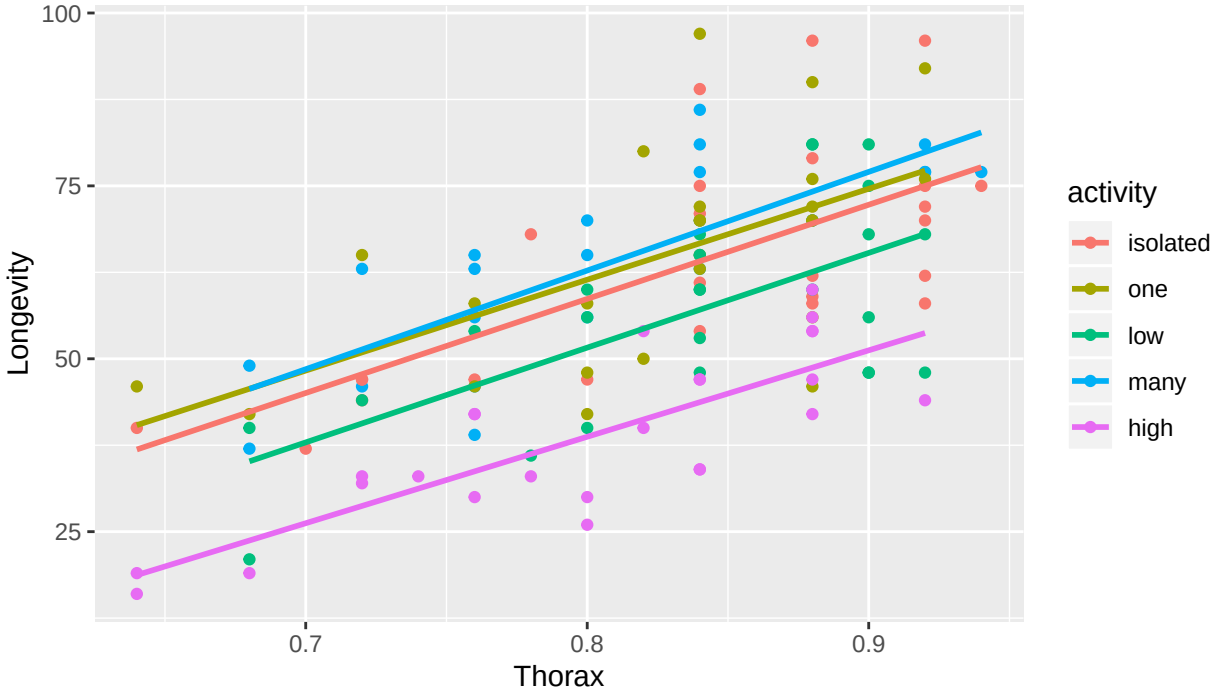
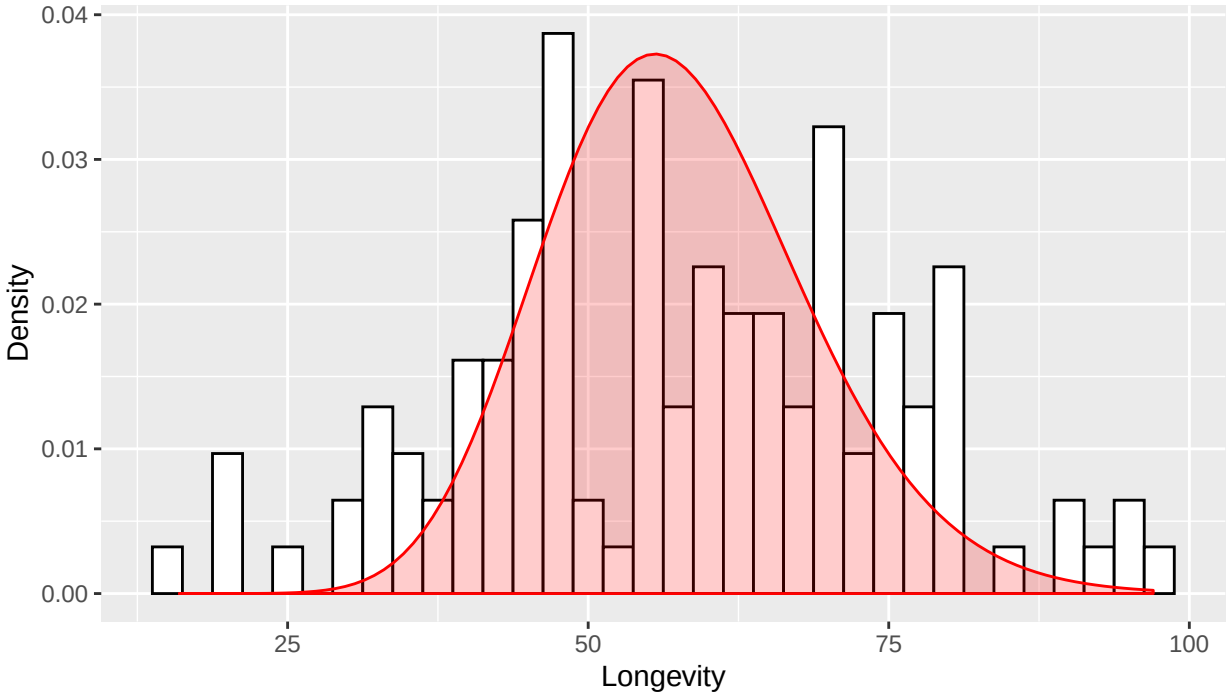


Figure2: Gamma generalized linear model



Smoking Report

Introduction

In the 2014 American National Youth Tobacco Survey, there are 22007 pieces of individuals' information. Based on this survey, following hypoheses are discussed in this report.

1. Regular use of chewing tobacco, snuff or dip is no more common amongst Americans of European ancestry than for Hispanic-Americans and African-Americans, once one accounts for the fact that white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon.
2. 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, ethnicity, and other demographic characteristics are similar.

Method

For our analysis we used a model that included the three races, Hispanic-Americans, African-Americans and White-Americans. We keep the other races to make the model more general.We removed the informations of 9 year olds because their data is suspicious. Since we are modeling probabilities,logistic regression model is the best choice. We considered the following model for each of the aforementioned analysis:

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

The *Odds* here is the ratio of probability of a person who is smoking vs non-smoking, π is the probability of a person who is smoking:

$$Odds = \frac{\pi}{1-\pi}$$

Therefore it can be effected by different factors, such as genderand race. Specifically, we tested whether white Americans more likely to live in rural areas and chewing tobacco is a rural phenomenon, compared to Hispanic-Americans, African-Americans and White-Americans . We also tested whether the likelihood of having used a hookah or waterpipe on at least one occasion is the same for two individuals of the different sexes. So, the null hypoheses we tested are:

$$H_0 \colon \beta_3 = \beta_4 = 0 \qquad H_0 \colon \beta_8 = 0 \qquad H_0 \colon \beta_2 = 0$$

Results

Table 2: Modeling odds of regular use of Chewing tobacco

	Exp. Estimate	Std. Error	z value	P-Value
Intercept	0.000	0.347	-23.885	0.000
Age	1.416	0.021	16.301	0.000
Female	0.158	0.113	-16.390	0.000
Black	0.182	0.186	-9.166	0.000
Hispanic	0.441	0.109	-7.528	0.000
Asian	0.217	0.343	-4.461	0.000
Native	1.140	0.279	0.471	0.638
Pacific	2.858	0.363	2.894	0.004
Rural	2.689	0.090	10.990	0.000

In the first model, odd is based on the probability of a person chewing tobacco. From the exponentiated coefficients of our model ,we can see which group effect the odds most. Black-americans are 18% more likely to chew tobacoo than White-americans. Hispanic-americans are 44% more likely to chew tobacoo than white-americans. Clearly, we see that prople from rural areas are about 2.7 times more likely to chew tobacoo. We can conclude that chewing tobacco is a rural phenomenon. Additionally, every time a person grow up one year old, the probability of this person to start chewing tobacco increase 41%.

Table 3: Modeling odds of ever using a hookah or waterpipe

	Exp. Estimate	Std. Error	z value	P-Value
Intercept	0.000	0.189	-43.037	0.000
Age	1.531	0.012	36.290	0.000
Female	1.048	0.043	1.073	0.283
Black	0.525	0.071	-9.010	0.000
Hispanic	1.405	0.049	6.931	0.000
Asian	0.536	0.118	-5.267	0.000
Native	1.213	0.191	1.013	0.311
Pacific	2.683	0.275	3.591	0.000
Rural	0.678	0.045	-8.683	0.000

The second model is based on the probability of a person using a hookah or waterpipe. The odds of using a hookah are a little bit higher for female than male. However, the corresponding p-value is 0.28, which means it is not statistically significant, so we cannot conclude that 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, ethnicity, and other demographic characteristics are similar.

Appendix

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

thorax_mean = mean(fruitfly$thorax)
thorax_sd = sd(fruitfly$thorax)
thorax_scaled = (fruitfly$thorax - thorax_mean)/thorax_sd

gglm1 = glm(fruitfly$longevity ~ thorax_scaled + fruitfly$activity, family=Gamma(link='log'))

gglm1_table = summary(gglm1)$coef

colnames(gglm1_table) <- c("Exp. Estimate","Std. Error","t value","P-Value")
rownames(gglm1_table) <- c("Intercept",
                           "Thorax Length",
                           "Activityone",
                           "Activitylow",
                           "Activitymany",
                           "Activityhigh")


knitr::kable(gglm1_table, digits = 3, cap='Estimated parameters of GLM model' )
```

Table 4: Estimated parameters of GLM model

	Exp. Estimate	Std. Error	t value	P-Value
Intercept	4.098	0.038	108.333	0.000
Thorax Length	0.204	0.017	11.804	0.000
Activityone	0.055	0.053	1.036	0.302
Activitylow	-0.116	0.053	-2.184	0.031
Activitymany	0.082	0.054	1.524	0.130
Activityhigh	-0.415	0.054	-7.687	0.000

```
shape = 1/summary(gglm1)$dispersion
scale = mean(fruitfly$longevity)/shape
```

Figure1: Longevity vs Thorax

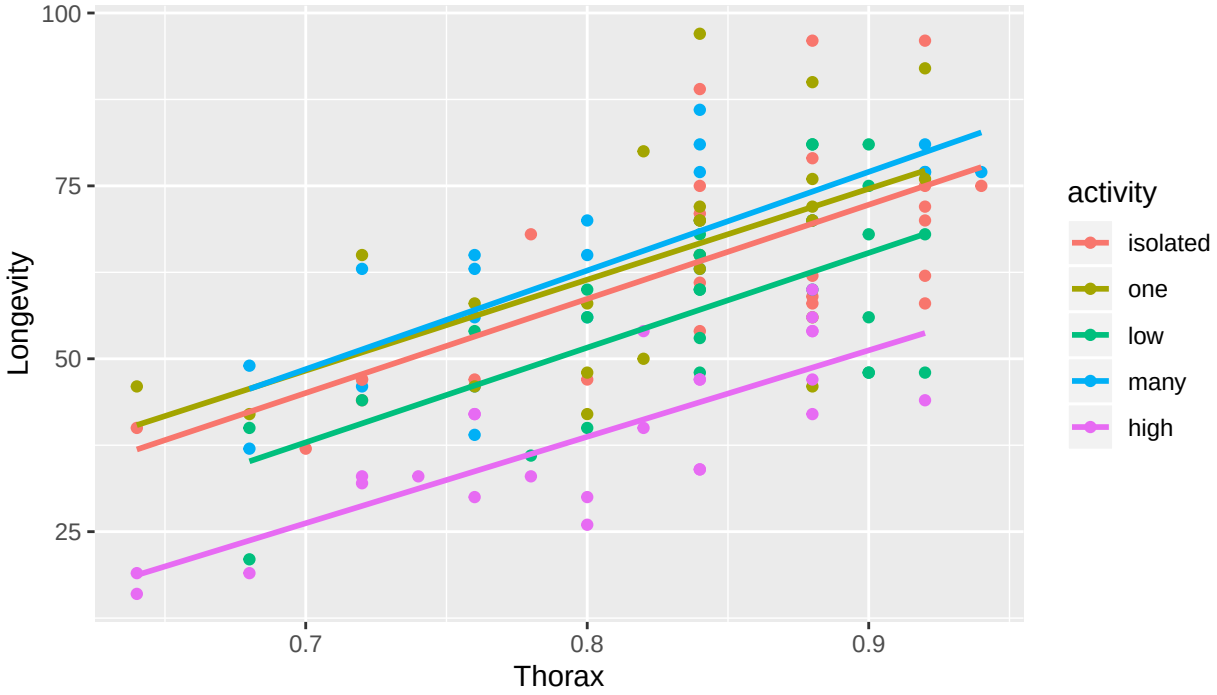
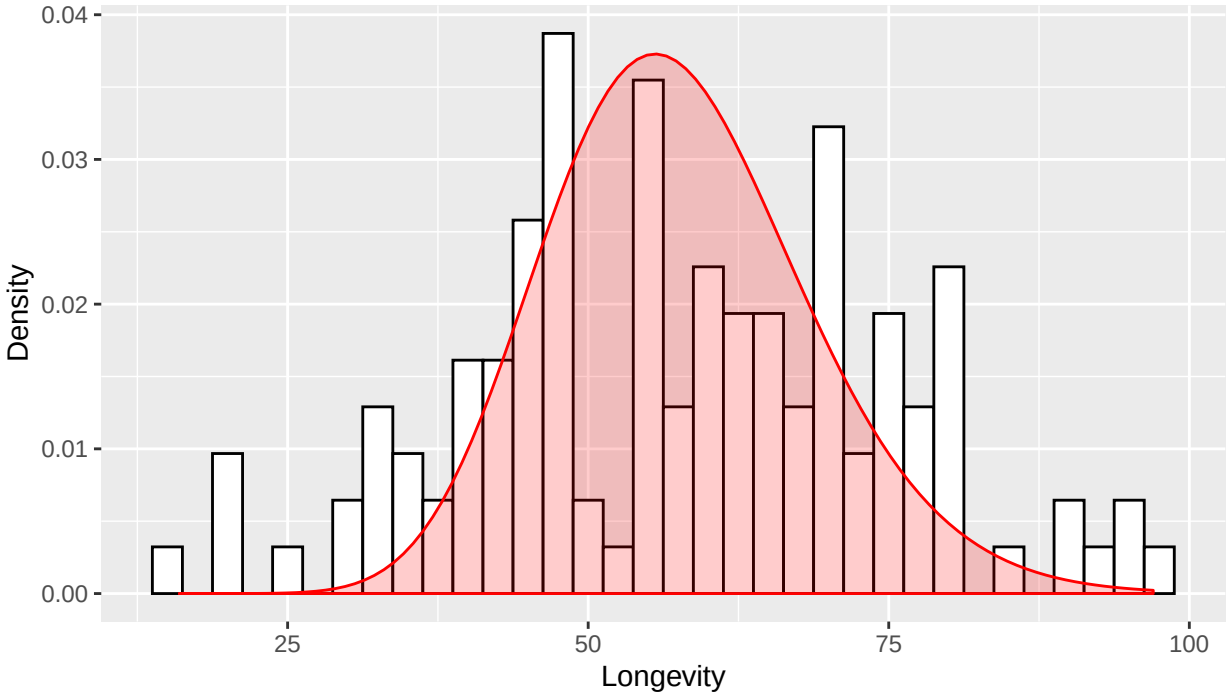


Figure2: Gamma generalized linear model




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

```
## [1] "C:\\Users\\82099\\AppData\\Local\\Temp\\RtmpCa3XjX\\file282042c42088.RData"
```

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

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

```
smoke[1:7,c('Age','Sex','Grade','RuralUrban','Race',
            'Tried_cigarette_smkg_even',
            'chewing_tobacco_snuff_or',
            'ever_tobacco_hookah_or_wa')]
```

```
## # A tibble: 7 x 8
##     Age Sex   Grade RuralUrban Race  Tried_cigarette~ chewing_tobacco~
##   <dbl> <fct> <dbl> <fct>      <fct>            <dbl> <lgl>
## 1    13 M         2 Urban      hisp~                2 FALSE
## 2    12 F         2 Urban      hisp~                2 FALSE
## 3    14 M         2 Urban      nati~                2 FALSE
## 4    13 M         2 Urban      hisp~                2 FALSE
## 5    14 M         2 Urban      nati~                1 FALSE
## 6    13 F         3 Urban      nati~                1 TRUE
## 7    14 M         3 Urban      hisp~                2 FALSE
## # ... with 1 more variable: ever_tobacco_hookah_or_wa <lgl>
```

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


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

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

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

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

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

```
smokeChew = na.omit(smokeChew)
smokeHookpipe = na.omit(smokeHookpipe)

smokeChew$y = cbind(smokeChew$yes, smokeChew$no)
smokeHookpipe$y = cbind(smokeHookpipe$yes, smokeHookpipe$no)


mod_chew = glm( y ~ Age + Sex + Race + RuralUrban, data = smokeChew,family=binomial(link='logit'))
```

```r
mod_Hookpipe = glm( y ~ Age + Sex + Race + RuralUrban, data = smokeHookpipe,family=binomial(link='logit

summary(mod_chew)
```

```
##
## Call:
## glm(formula = y ~ Age + Sex + Race + RuralUrban, family = binomial(link = "logit"),
##     data = smokeChew)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.1139  -0.7329  -0.3404   0.2955   3.4870
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.27999    0.34667 -23.885  < 2e-16 ***
## Age              0.34796    0.02135  16.301  < 2e-16 ***
## SexF            -1.84497    0.11257 -16.390  < 2e-16 ***
## Raceblack       -1.70279    0.18576  -9.166  < 2e-16 ***
## Racehispanic    -0.81931    0.10883  -7.528 5.15e-14 ***
## Raceasian       -1.52799    0.34251  -4.461 8.15e-06 ***
## Racenative       0.13106    0.27851   0.471   0.6379
## Racepacific      1.04977    0.36273   2.894   0.0038 **
## RuralUrbanRural  0.98906    0.09000  10.990  < 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: 1393.80  on 208  degrees of freedom
## Residual deviance:  279.29  on 200  degrees of freedom
## AIC: 585.96
##
## Number of Fisher Scoring iterations: 6
```

```r
summary(mod_Hookpipe)
```

```
##
## Call:
## glm(formula = y ~ Age + Sex + Race + RuralUrban, family = binomial(link = "logit"),
##     data = smokeHookpipe)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.4003  -0.9657  -0.1512   0.5154   3.2389
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)     -8.12442    0.18878 -43.037  < 2e-16 ***
## Age              0.42565    0.01173  36.290  < 2e-16 ***
## SexF             0.04654    0.04337   1.073 0.283205
## Raceblack       -0.64406    0.07148  -9.010  < 2e-16 ***
```

```
## Racehispanic     0.33974    0.04902   6.931 4.18e-12 ***
## Raceasian       -0.62377    0.11843  -5.267 1.39e-07 ***
## Racenative       0.19348    0.19096   1.013 0.310957
## Racepacific      0.98745    0.27497   3.591 0.000329 ***
## RuralUrbanRural -0.38904    0.04480  -8.683  < 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: 2145.46  on 206  degrees of freedom
## Residual deviance:  374.67  on 198  degrees of freedom
## AIC: 953.92
##
## Number of Fisher Scoring iterations: 4
```

```
TabChew = round(summary(mod_chew)$coefficients, 3)
TabHookpipe = round(summary(mod_Hookpipe)$coefficients, 3)
TabChew[,1] = round(exp(TabChew[,1]),3)
TabHookpipe[,1] = round(exp(TabHookpipe[,1]),3)
colnames(TabChew) = c("Exp. Estimate","Std. Error","z value","P-Value")
colnames(TabHookpipe) = c("Exp. Estimate","Std. Error","z value","P-Value")
rownames(TabChew) = c("Intercept","Age","Female","Black",
                      "Hispanic","Asian","Native","Pacific","Rural")
rownames(TabHookpipe) = c("Intercept","Age","Female","Black",
                          "Hispanic","Asian","Native","Pacific","Rural")

knitr::kable(TabChew, cap="Modeling odds of regular use of Chewing tobacco")
```

Table 5: Modeling odds of regular use of Chewing tobacco

	Exp. Estimate	Std. Error	z value	P-Value
Intercept	0.000	0.347	-23.885	0.000
Age	1.416	0.021	16.301	0.000
Female	0.158	0.113	-16.390	0.000
Black	0.182	0.186	-9.166	0.000
Hispanic	0.441	0.109	-7.528	0.000
Asian	0.217	0.343	-4.461	0.000
Native	1.140	0.279	0.471	0.638
Pacific	2.858	0.363	2.894	0.004
Rural	2.689	0.090	10.990	0.000

```
knitr::kable(TabHookpipe, cap="Modeling odds of ever using a hookah or waterpipe")
```

Table 6: Modeling odds of ever using a hookah or waterpipe

	Exp. Estimate	Std. Error	z value	P-Value
Intercept	0.000	0.189	-43.037	0.000
Age	1.531	0.012	36.290	0.000
Female	1.048	0.043	1.073	0.283
Black	0.525	0.071	-9.010	0.000

	Exp. Estimate	Std. Error	z value	P-Value
Hispanic	1.405	0.049	6.931	0.000
Asian	0.536	0.118	-5.267	0.000
Native	1.213	0.191	1.013	0.311
Pacific	2.683	0.275	3.591	0.000
Rural	0.678	0.045	-8.683	0.000

```
chewTable = as.data.frame(summary(mod_chew)$coef)
chewTable
```

```
##                   Estimate Std. Error     z value      Pr(>|z|)
## (Intercept)     -8.2799904 0.34666695 -23.8845680 4.431143e-126
## Age              0.3479621 0.02134544  16.3014709  9.634531e-60
## SexF            -1.8449668 0.11256763 -16.3898523  2.259932e-60
## Raceblack       -1.7027858 0.18576483  -9.1663519  4.892976e-20
## Racehispanic    -0.8193086 0.10883461  -7.5280149  5.151752e-14
## Raceasian       -1.5279918 0.34251077  -4.4611496  8.152117e-06
## Racenative       0.1310622 0.27851248   0.4705792  6.379413e-01
## Racepacific      1.0497737 0.36272899   2.8940991  3.802481e-03
## RuralUrbanRural  0.9890592 0.08999631  10.9899974  4.269384e-28
```

```
chewTable$lower = chewTable$Estimate - 2*chewTable$'Std. Error'
chewTable$upper = chewTable$Estimate + 2*chewTable$'Std. Error'
chewoddsRatio = exp(chewTable[,c('Estimate','lower','upper')])
rownames(chewoddsRatio)[1] = 'baseline prob'
chewoddsRatio[1,] = chewoddsRatio[1,]/(1+chewoddsRatio[,1])
chewoddsRatio
```

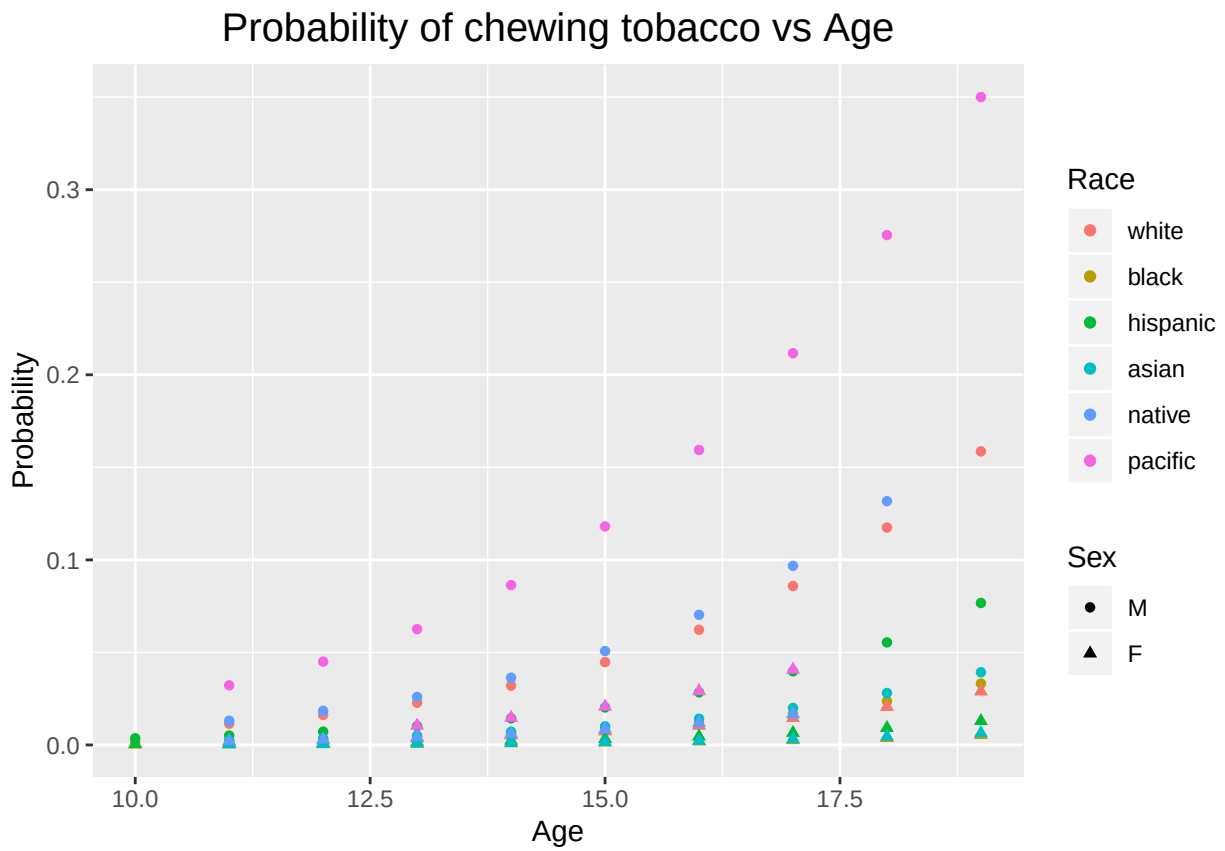
```
##                      Estimate        lower        upper
## baseline prob   0.0002534754 5.245727e-05 0.0004379625
## Age             1.4161786369 1.356993e+00 1.4779456252
## SexF            0.1580305741 1.261729e-01 0.1979320466
## Raceblack       0.1821753172 1.256424e-01 0.2641452544
## Racehispanic    0.4407362861 3.545245e-01 0.5479126782
## Raceasian       0.2169709541 1.093706e-01 0.4304301416
## Racenative      1.1400386640 6.531407e-01 1.9899054714
## Racepacific     2.8570044019 1.383084e+00 5.9016480718
## RuralUrbanRural 2.6887037252 2.245811e+00 3.2189390150
```

```
rownames(chewoddsRatio) = gsub("Race|RuralUrban|C$", "",
                               rownames(chewoddsRatio) )
rownames(chewoddsRatio) = gsub("SexF","Female",
                               rownames(chewoddsRatio))
knitr::kable(chewoddsRatio, digits=3)
```

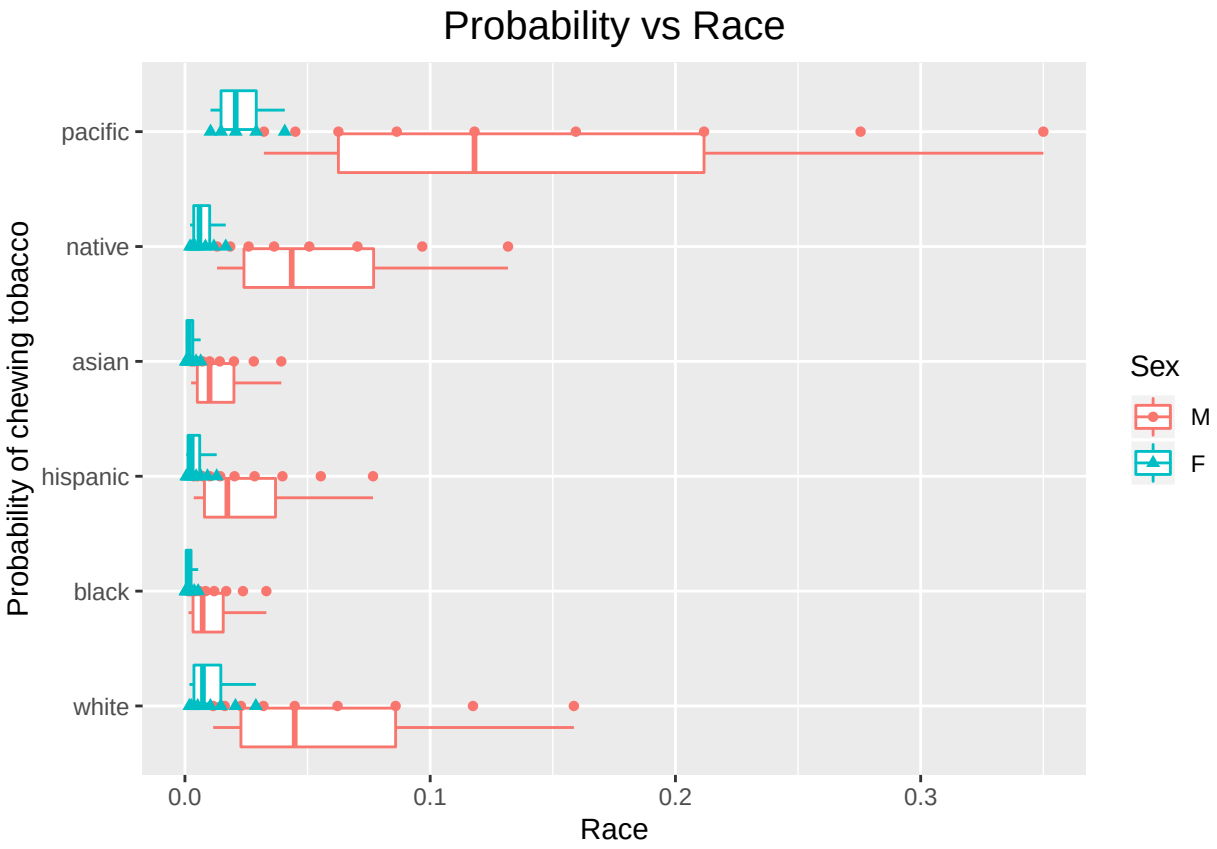
	Estimate	lower	upper
baseline prob	0.000	0.000	0.000
Age	1.416	1.357	1.478
Female	0.158	0.126	0.198
black	0.182	0.126	0.264

	Estimate	lower	upper
hispanic	0.441	0.355	0.548
asian	0.217	0.109	0.430
native	1.140	0.653	1.990
pacific	2.857	1.383	5.902
Rural	2.689	2.246	3.219

```
toPredict = smokeChew[smokeChew$RuralUrban == 'Urban', ]

chewPred = as.data.frame(predict(mod_chew, toPredict,  se.fit=TRUE))
chewPred$lower = chewPred$fit - 2*chewPred$se.fit
chewPred$upper = chewPred$fit + 2*chewPred$se.fit
chewPredExp = exp(chewPred[,c('fit','lower','upper')])
chewPredProb = chewPredExp / (1+chewPredExp)

plot_1 = data.frame(toPredict$Age,chewPredProb$fit,toPredict$Sex,toPredict$Race)

colnames(plot_1)[1]='Age'
colnames(plot_1)[2]='Prob'
colnames(plot_1)[3]='Sex'
colnames(plot_1)[4]='Race'

ggplot(plot_1, aes(x= Age,y=Prob, col = Race, pch = Sex) )+
  labs(title="Probability of chewing tobacco vs Age") +
  labs(x="Age", y="Probability") +
  geom_point()+
  theme(plot.title=element_text(size=15,
                                hjust=0.5,
                                lineheight=1.2))
```




```
ggplot(plot_1, aes(x= Race,y=Prob, col = Sex, pch = Sex) )+
  geom_boxplot()+ coord_flip()+
  geom_jitter(position = position_jitter(width=0, height=0))+
  labs(title="Probability vs Race") +
  labs(x="Probability of chewing tobacco", y="Race") +
  theme(plot.title=element_text(size=15,
                                hjust=0.5,
                                lineheight=1.2))
```


Probability vs Race
Probability of chewing tobacco
pacific
native
asian
hispanic
black
white
0.0
0.1
0.2
0.3
Race
Sex
M
F