STA442 HW1

Depeng Ye 1002079500

2019-09-18

1 Flies

Short Report

Through the investigation of the data set, we can conclude that the length of thorax of a fruitfly has effect on its lifetime.

The data was collected from an experiment of 125 male fruitflies that were divided into 5 groups of equal size. Within the 5 groups, there is one solidary group, one given 1 virgin female per day, one given 8 virgin female per day, one given 1 pregnant female per day, the last one is given 8 pregnant female per day. All five groups were labeled many, isolated, one, low, and high, correspondingly where:

isolated = fly with no female fruitflies given
one = fly kept with one pregnant fruitfly
low = fly kept with one virgin female fruitfly
many = fly kept with eight pregnant fruitflies
high = fly kept with eight virgin fruitflies

The above experiment is designed and recorded in Faraway(2005).

The dataset includes 124 observations of 3 variables: thorax, longevity, and activity. Note that there is one subejct missing in the 'many' group while all other groups having 25 observations.
Ploting the histogram of the longevity and the thorax of the subject fruitflies. Notice that the longevity is likely normally distributed, while the thorax is left-screwed. Try to refine the data of thorax so that it is normalized (refined_thorax).

$$x_{norm} = \frac{x_i - \bar{x}}{\sigma_x}$$

Notice that the shape of histogram of subjects' longevity has a shape of Gamma distribution. Try to fit a gamma generalized linear model (MyFit) to longevity as a function of refined thorax and activity. Plot the GLM into the histogram of longevity can be found in the appendix as well. The mathematical description of the model is shown below:

$$X \sim Gamma(\phi, \nu)$$

$$f(x; \phi, \nu) = \frac{(x/\phi)x^{\nu-1}e^{-x/\phi}}{\Gamma(\nu)\phi}$$

$$log(\mu_i) = X_i\beta$$

where $\phi = \frac{\mu_i}{\nu}$ is $scale = 2.12$ and ν is $shape = 28.43$ in my model fitted.

Summary

By looking at the significance level of coefficients of MyFit, we can draw a conclusion: Considering the fruitfly data that has been collected and investigated, there is a significant link that the lifetime of a male fruitfly is correlated with the length of its thorax. When it comes to activity level, different levels of activity are having different effect to the life time.

• Isolated group shows a same level of significance as the length of thorax, meaning the control group is successful.

- One group with shows no significant relation to fruitflies' lifetime.
- Low group has a 95% confidence level that at this level of activity, note from the boxplot of longevity in each group, as well as the estimate of coefficient, the effect is negative.
- Many group again has no significant effect on he life time of fruitflies' lifetime.
- High group has a 99.9% conficence level that it has a strong negative effect on fruitflies' longevity.

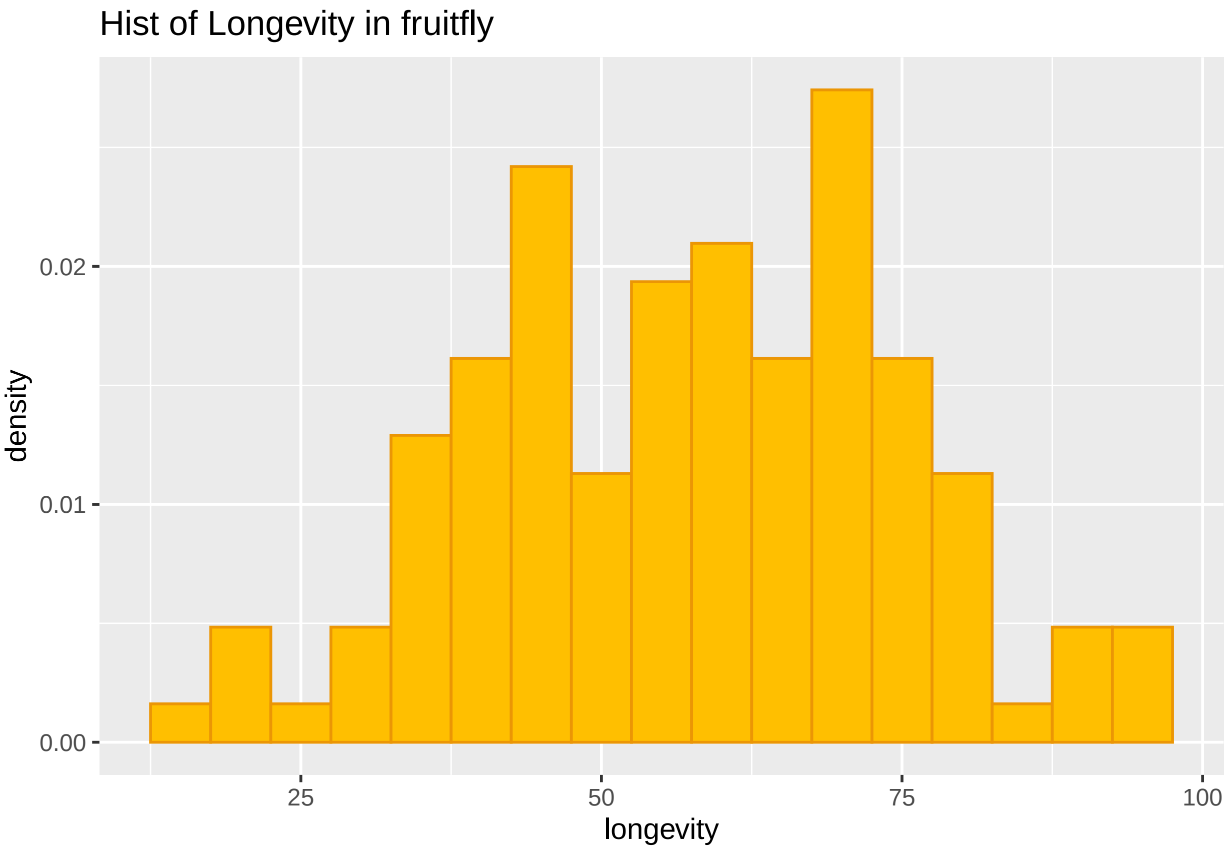
Table 1: Summary of fruitfly data

	thorax	longevity	activity
	Min. :0.6400	Min. :16.00	isolated:25
	1st Qu.:0.7600	1st Qu.:46.00	one :25
	Median :0.8400	Median :58.00	low :25
	Mean :0.8224	Mean :57.62	many :24
	3rd Qu.:0.8800	3rd Qu.:70.00	high :25
	Max. :0.9400	Max. :97.00	NA

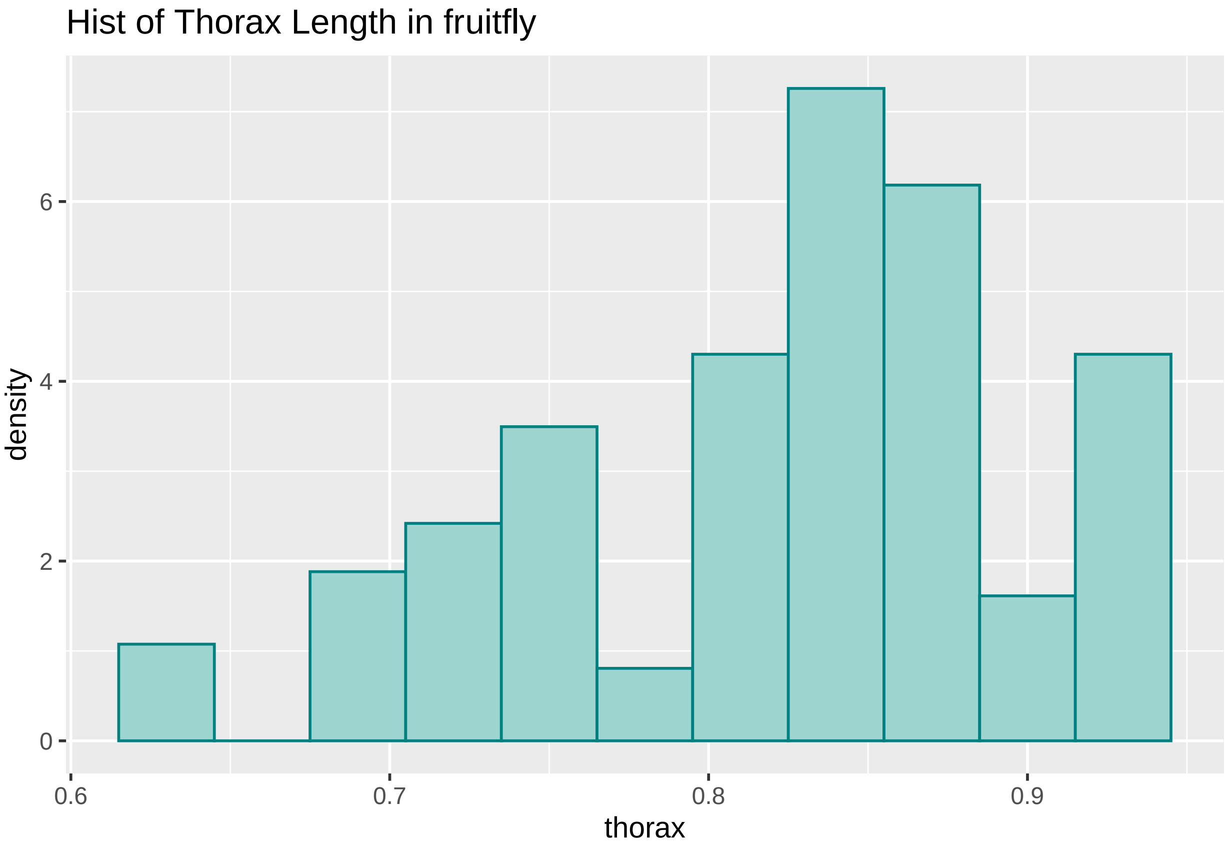
Code Appendix for Question 1

```r
# Calling fruitfly data from pkg faraway
data('fruitfly', package = 'faraway')
# Summary of data
knitr::kable(summary(fruitfly), format = 'latex', align = 'c', digits = 2,
              caption = 'Summary of fruitfly data')
```

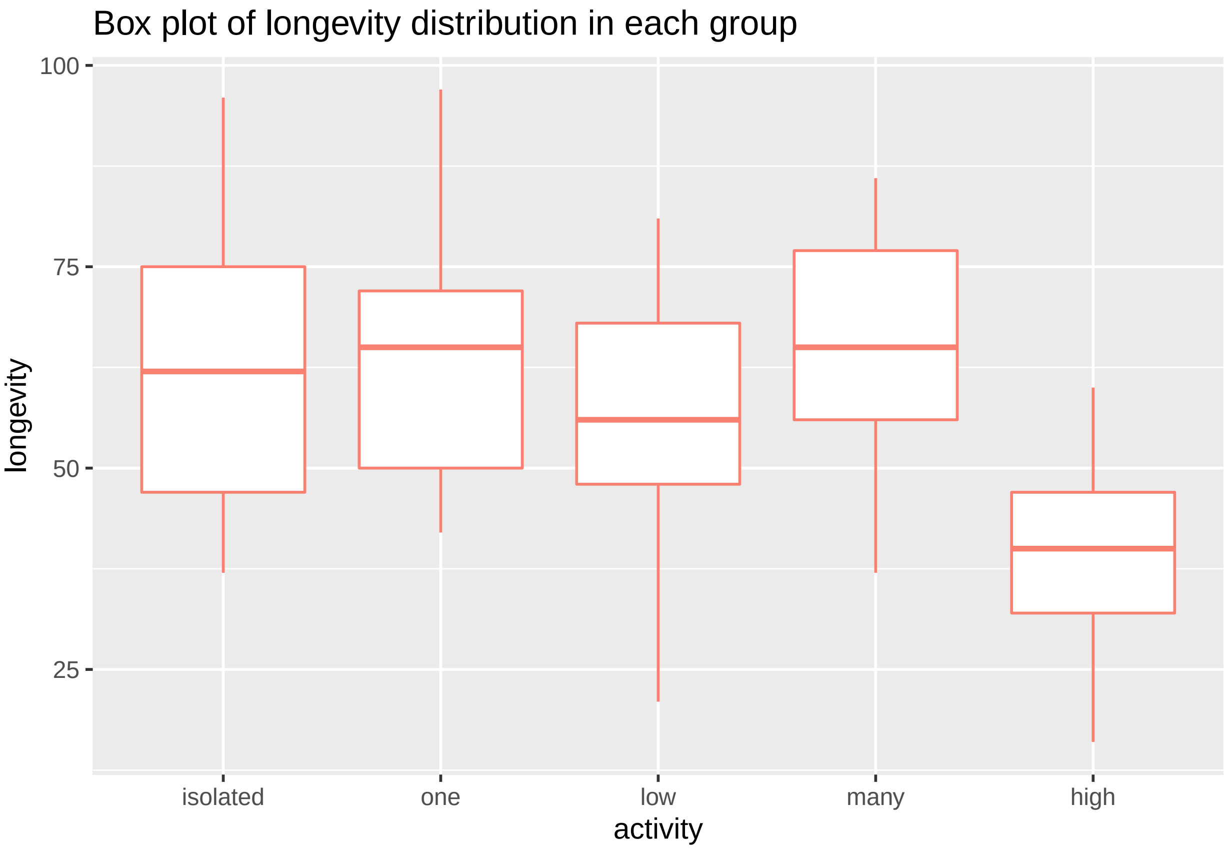
```r
# Histogram of longevity
ggplot(fruitfly, aes(x = longevity)) +
  geom_histogram(aes(y = ..density..), col = '#EB9605', fill = '#FFBF00', binwidth = 5) +
  labs(title = "Hist of Longevity in fruitfly")
```



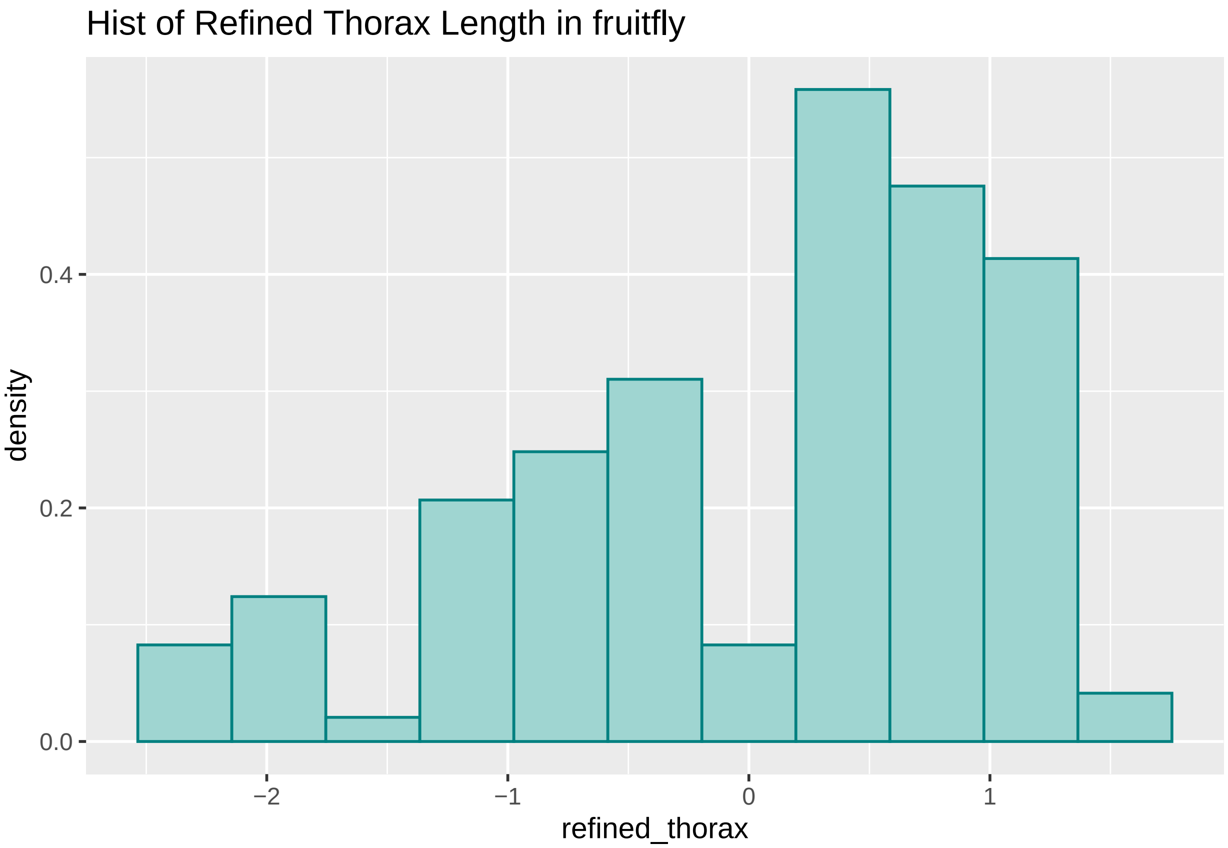
```r
# Histogram of thorax length
ggplot(fruitfly, aes(x = thorax)) +
  geom_histogram(aes(y = ..density..),
                 col = '#008080', fill = '#9FD5D1', binwidth = 0.03) +
    labs(title = "Hist of Thorax Length in fruitfly")
```



```
#Plot of activity level vs longevity
ggplot(fruitfly, aes(x = activity, y = longevity)) +
  geom_boxplot(col = '#FA8072') +
  labs(title = "Box plot of longevity distribution in each group")
```



```
# Processing thorax
attach(fruitfly)
refined_thorax = (thorax - mean(thorax)) / sd(thorax)
datare_thorax = data.frame(refined_thorax)
#plotting refined_thorax
ggplot(datare_thorax, aes(x = refined_thorax)) +
  geom_histogram(aes(y = ..density..), col = '#008080',
                 fill = '#9FD5D1', binwidth = 0.39) +
  labs(title = "Hist of Refined Thorax Length in fruitfly")
```



```r
# Fitting the GLM
MyFit = glm(longevity ~  refined_thorax + activity, family = Gamma(link = 'log'),
            data = fruitfly)
summary(MyFit)
```

```
## 
## Call:
## glm(formula = longevity ~ refined_thorax + activity, family = Gamma(link = "log"), 
##     data = fruitfly)
## 
## Deviance Residuals: 
##      Min        1Q    Median        3Q       Max  
## -0.50718  -0.15216  -0.02833   0.12434   0.39938  
## 
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)    
## (Intercept)     4.09771    0.03783 108.333  < 2e-16 ***
## refined_thorax  0.20433    0.01731  11.804  < 2e-16 ***
## activityone     0.05527    0.05337   1.036   0.3024    
## activitylow    -0.11646    0.05332  -2.184   0.0309 *  
## activitymany    0.08250    0.05413   1.524   0.1302    
## activityhigh   -0.41466    0.05394  -7.687 4.93e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## (Dispersion parameter for Gamma family taken to be 0.0355297)
## 
```

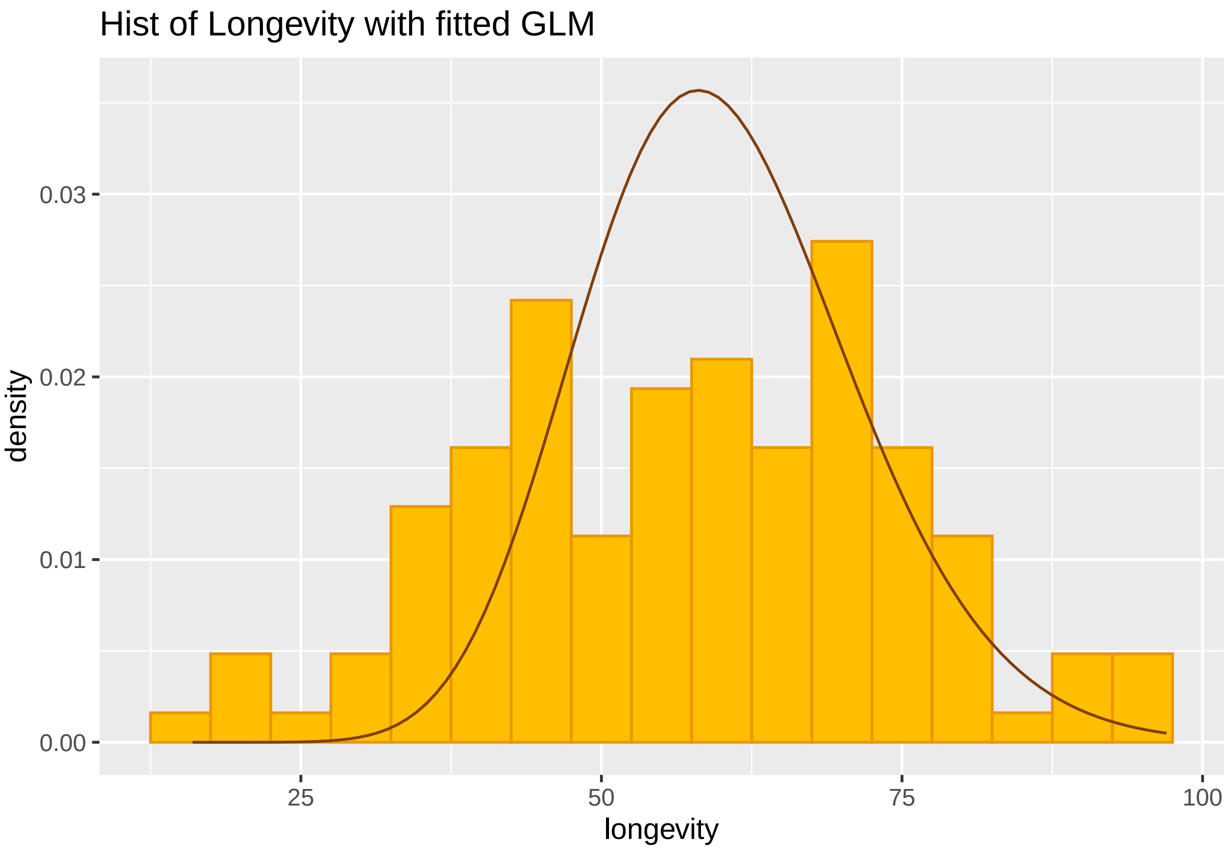
Table 2: Coefftients of Fitted GLM

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	4.10	0.04	108.33	0.00
refined_thorax	0.20	0.02	11.80	0.00
activityone	0.06	0.05	1.04	0.30
activitylow	-0.12	0.05	-2.18	0.03
activitymany	0.08	0.05	1.52	0.13
activityhigh	-0.41	0.05	-7.69	0.00

```
##     Null deviance: 13.2803  on 123  degrees of freedom
## Residual deviance:  4.3151  on 118  degrees of freedom
## AIC: 942.29
##
## Number of Fisher Scoring iterations: 4
```

```r
# summary of MyFit
knitr::kable(summary(MyFit)$coef, digits = 2, caption = 'Coefftients of Fitted GLM')
```

```r
shape = 1/summary(MyFit)$dispersion
scale = exp(MyFit$coef["(Intercept)"]) / shape
#visualization of MyFit
ggplot(fruitfly, aes(x = longevity)) +
  geom_histogram(aes(y = ..density..), col = '#EB9605', fill = '#FFBF00', binwidth = 5) +
  stat_function(fun = dgamma, args = list(shape = shape, scale = scale), col = '#813F0B') +
  labs(title = "Hist of Longevity with fitted GLM")
```



2 Smoke

Summary

Based on the investigation and analysis of the given data collected by American National Youth Tobacco Survey in 2014, regarding the first hypothesis, there is sufficient evidence to show that the chewing tobacco, snuff, or dip consumption of tobacco 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.

Regarding the second hypothesis, there is sufficient evidence to conclude that

write-up

In the recent years, smoking is gradually becoming a major concern of people's health threat within many countries, especially the US. Within all the humanbeing, youngesters' health are concerned the most. This data investigation is based on a smoke data published by American National Youth Tobacco Survey in 2014. and we made two hypothesises to check whether they are true or not.

The first focus of our study is to examine that whether 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.

Another point of investigation is to determine 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.

Hypothesis I is tested by fitting a binomial (logistic) model. According to what we have learned in the lecture, binomial model with link = 'logit' is the go to when it comes to investigating 'yes' or 'no' data.

$$y = \beta_0 + \beta_1 x_{Age} + \beta_2 I_{Male} + \beta_3 I_{Black} + \beta_4 I_{Hips} + \beta_5 I_{Asian} + \beta_6 I_{Native} + \beta_7 I_{Pacific} + \beta_8 I_{Rural}$$

Hypothesis II is tested by a similar method of fitting a logistic model. Notice that for the second hypothesis,the exponential of the coefficients of PipeFit shows that older teenagers are more likely to use Hookah than those younger kids. The rank of most likely to use Hookah based on race is Hispanic, white, black. Moreover, people in the cities are more likely to use Hookahs. This could coincide with that mose Hookah stores are located in the more developed areas instead of rural area.

Appendix

Data exploration

```r
#Loading/Downloading smoke.RData
dataDir = "../HW1"
smokeFile = file.path(dataDir, "smoke.RData")
if (! file.exists(smokeFile)) {
  download.file("http://pbrown.ca/teaching/appliedstats/data/smoke.RData",
              smokeFile)
  }
(load(smokeFile))
```

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

Table 3: Abstract of smoke.RData

Age	Sex	Grade	RuralUrban	Race	chewing_tobacco_snuff_or
13	M	2	Urban	hispanic	FALSE
12	F	2	Urban	hispanic	FALSE
14	M	2	Urban	native	FALSE
13	M	2	Urban	hispanic	FALSE
14	M	2	Urban	native	FALSE
13	F	3	Urban	native	TRUE
14	M	3	Urban	hispanic	FALSE
14	F	3	Urban	native	FALSE
14	F	3	Urban	NA	FALSE
14	F	3	Urban	native	FALSE
13	F	3	Urban	native	FALSE
14	F	3	Urban	hispanic	FALSE

Table 4: Grade vs Age Distributino Table

	9	10	11	12	13	14	15	16	17	18	19	NA
1	13	8	1311	1806	192	9	3	2	3	0	2	8
2	6	2	13	1267	2029	201	12	1	0	0	1	9
3	2	0	0	3	1379	1907	211	10	2	3	1	3
4	4	1	0	0	6	1085	1581	181	16	3	2	6
5	0	0	0	1	1	10	1114	1593	188	18	4	4
6	3	0	0	0	0	1	3	1089	1524	183	11	3
7	10	0	0	0	1	1	1	13	1109	1471	153	5
8	2	1	1	0	2	0	0	1	3	1	6	1
NA	0	0	2	5	7	4	14	10	8	3	0	118

```r
# Explore smoke.RData
knitr::kable(smoke[1:12,
                   c('Age', 'Sex', 'Grade', 'RuralUrban', 'Race',
                     'chewing_tobacco_snuff_or')], digits = 2, caption = "Abstract of smoke.RData")
```

```r
# Indexing the column investigated in smokeFormat dataframe
smokeFormats[smokeFormats$colName == 'chewing_tobacco_snuff_or',]
```

```
##         ID
## 151 cslt_r
##                                                                                  label
## 151 RECODE: Used chewing tobacco, snuff, or dip on 1 or more days in the past 30 days
##                                                              shortLabel
## 151 chewing tobacco snuff or dip on 1 or more days in the past 30 days
##                       colName
## 151 chewing_tobacco_snuff_or
```

```r
# Relabelling the investigated column
smoke$everSmoke = factor(smoke$chewing_tobacco_snuff_or, levels =
                           c('TRUE', 'FALSE'), labels = c('Yes', 'No'))
#looking at the age ~ grade distribution of subjects
knitr::kable(table(smoke$Grade, smoke$Age, exclude = NULL),
                  caption = "Grade vs Age Distributino Table", digits = 2)
```

Table 5: Race vs Smoke Experience Distribution Table

	Yes	No	NA
white	527	9300	66
black	40	3317	73
hispanic	145	5820	116
asian	10	954	9
native	15	320	3
pacific	11	71	3
NA	47	1054	106

```
#Looking at the Race ~ everSmoke density of subjects
knitr::kable(table(smoke$Race, smoke$everSmoke, exclude= NULL),
             caption = "Race vs Smoke Experience Distribution Table", digits = 2)
```

Code for Hypothesis I

```
# Removing 9 years old because their data is suspicious
# smokeSub = smoke[smoke$Age >= 10, ]
smokeSub = smoke[smoke$Age >= 10 & !is.na(smoke$Race) &
                   !is.na(smoke$chewing_tobacco_snuff_or), ]
smokeAgg_chew = reshape2::dcast(smokeSub, Age + Sex + Race +
                             RuralUrban ~ chewing_tobacco_snuff_or, length)
```

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

```
smokeAgg_chew = na.omit(smokeAgg_chew)
smokeAgg_chew$Age = smokeAgg_chew$Age - mean(smokeAgg_chew$Age)

colnames(smokeAgg_chew)[colnames(smokeAgg_chew) == "FALSE"] = "no"
colnames(smokeAgg_chew)[colnames(smokeAgg_chew) == "TRUE"] = "yes"

#select the white
smokeAgg_chew[which (smokeAgg_chew$Race == 'white' &
                       smokeAgg_chew$Sex == "M" &
                       smokeAgg_chew$RuralUrban == 'Rural'),]
```

```
##            Age Sex  Race RuralUrban  no yes NA
## 9   -3.722488   M white      Rural 129   7  0
## 31  -2.722488   M white      Rural 320   5  0
## 59  -1.722488   M white      Rural 391  13  0
## 88  -0.722488   M white      Rural 350  26  0
## 117  0.277512   M white      Rural 348  58  0
## 145  1.277512   M white      Rural 321  72  0
## 174  2.277512   M white      Rural 339  83  0
## 204  3.277512   M white      Rural 201  91  0
## 230  4.277512   M white      Rural  22   8  0
```

```
smokeAgg_chew$y = cbind(smokeAgg_chew$yes, smokeAgg_chew$no)
# Fit the model
smokeFit = glm(y ~ Age + Sex + Race + RuralUrban,
               family=binomial(link='logit'), data=smokeAgg_chew)
```

```
knitr::kable(summary(smokeFit)$coef, digits=4)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.1254	0.0843	-37.0546	0.000
Age	0.3366	0.0208	16.2043	0.000
SexF	-1.7884	0.1085	-16.4810	0.000
Raceblack	-1.5561	0.1717	-9.0644	0.000
Racehispanic	-0.7131	0.1036	-6.8843	0.000
Raceasian	-1.5463	0.3421	-4.5194	0.000
Racenative	0.1069	0.2776	0.3853	0.700
Racepacific	1.0121	0.3605	2.8072	0.005
RuralUrbanRural	0.9508	0.0874	10.8757	0.000

Code for Hypothesis II

```
# Using similar method as the previous session
smokeSub1 = smoke[smoke$Age >= 10 & !is.na(smoke$Race) &
                    !is.na(smoke$ever_tobacco_hookah_or_wa), ]
smokeAgg_pipe = reshape2::dcast(smokeSub1, Age + Sex + Race +
                                  RuralUrban ~ ever_tobacco_hookah_or_wa, length)
```

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

```
smokeAgg_pipe = na.omit(smokeAgg_pipe)
smokeAgg_pipe$Age = smokeAgg_pipe$Age - mean(smokeAgg_pipe$Age)
colnames(smokeAgg_pipe)[colnames(smokeAgg_pipe) == "TRUE"] = "yes"
colnames(smokeAgg_pipe)[colnames(smokeAgg_pipe) == "FALSE"] = "no"
smokeAgg_pipe[which (smokeAgg_pipe$Sex == "M" & smokeAgg_pipe$RuralUrban == "Rural" &
                     smokeAgg_pipe$Race == "black"), ]
```

```
##             Age Sex  Race RuralUrban no yes NA
## 12  -3.7149758   M black      Rural 35   0  0
## 34  -2.7149758   M black      Rural 87   4  0
## 61  -1.7149758   M black      Rural 84   2  0
## 90  -0.7149758   M black      Rural 95   4  0
## 119  0.2850242   M black      Rural 76   4  0
## 147  1.2850242   M black      Rural 87  16  0
## 176  2.2850242   M black      Rural 97   8  0
## 206  3.2850242   M black      Rural 50   9  0
## 232  4.2850242   M black      Rural 15   2  0
```

```
smokeAgg_pipe$y = cbind(smokeAgg_pipe$yes, smokeAgg_pipe$no)
# fitting logestic model for the second hypothesis
PipeFit = glm(y ~ Age + Sex + Race + RuralUrban,
              family = binomial(link = 'logit'), data = smokeAgg_pipe)
knitr::kable(exp(summary(PipeFit)$coef), digits = 4)
```

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	0.1584	1.0456	0.000000e+00	1.0000
Age	1.5199	1.0116	5.623341e+15	1.0000
SexF	1.0429	1.0438	2.665600e+00	1.3866
Raceblack	0.5301	1.0730	1.000000e-04	1.0000
Racehispanic	1.4129	1.0496	1.258299e+03	1.0000
Raceasian	0.5321	1.1249	4.700000e-03	1.0000
Racenative	1.1731	1.2097	2.312500e+00	1.4946
Racepacific	2.6213	1.3102	3.538700e+01	1.0004
RuralUrbanRural	0.6781	1.0453	2.000000e-04	1.0000