Homework 1, Exploration of Generalized Linear Models

Jessica Chau

2019-01-30

1 Short Answer

1.1 Simulation Study

Part A

The computed coverage probability for a two standard error confidence interval is 0.96 (with 100 simulated data sets). Coverage probability is useful for assessing the validity of the confidence interval. A coverage probability that is approximate to the 95% CI provides a high validity. If the coverage probability is not approximate to the confidence interval, it means that there are biased estimates or overly-conservative/non-conservative standard errors (or both).

Part B

From Figure 1, it appears that the normal distribution is not a good fit for the $\hat{\beta}$ coefficient because Figure 1(a) shows the data is right-skewed and the QQ-plot in Figure 1(b) shows the $\hat{\beta}$ coefficient differing from the normal distribution.

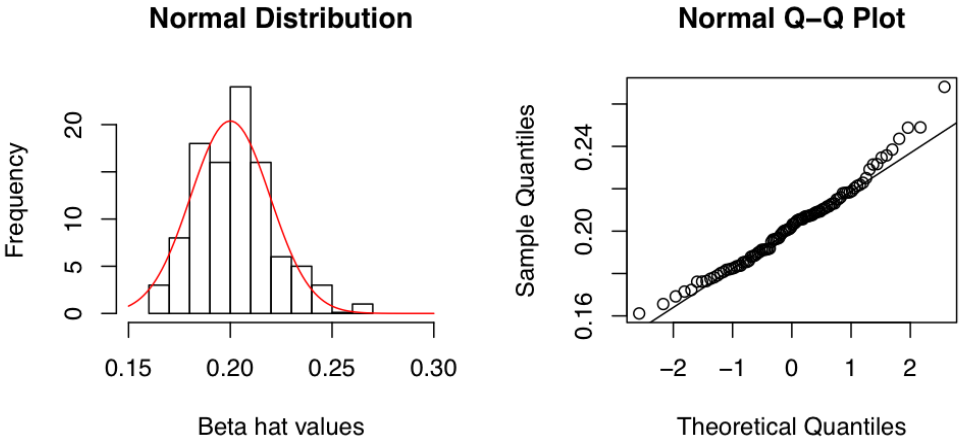


Figure 1: Evaluating Normality of Beta Coefficients

Part C

After calculating 100 likelihood ratio statistics for testing $\beta = 0.2$, it appears that with one degree of freedom, the data appears to follow a chi-squared distribution as displayed in Figure 2.

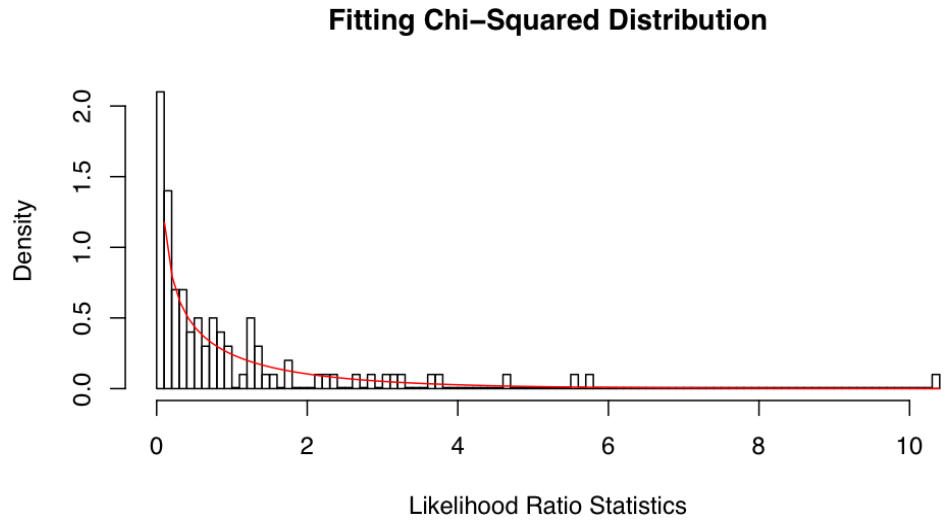


Figure 2: Fitting Chi-Squared Distribution on Likelihood Ratio Statistics

1.2 Distribution Functions

Part A: Solve for Distribution's Parameters (given mean=2 and variance=3)

Zero-inflated Poisson

Equation 1 (mean):

$$(1-\pi)\lambda = 2$$

Equation 2 (variance):

$$\lambda(1-\pi)(1+\pi\lambda) = 3$$

Solving for λ, π:

$$\lambda(1-1+2/\lambda)(1+(1-2/\lambda)\lambda) = 3$$

$$2(\lambda-1) = 3$$

$$\lambda = 5/2$$

$$\pi = 1/5$$

Gamma

Equation 1 (mean):

$$\frac{\alpha}{\beta} = 2$$

Equation 2 (variance):

$$\frac{\alpha}{\beta^2} = 3$$

Solving for β, α:

$$\frac{2\beta}{\beta^2} = 3$$
$$\beta = 2/3$$
$$\alpha = 4/3$$

Weibull

This requires the uniroot function in R to solve for scale λ and shape k. Equation 1 (mean):

$$\lambda\Gamma(1 + \frac{1}{k}) = 2$$
$$\lambda = \frac{2}{\Gamma(1 + \frac{1}{k})}$$

Equation 2 (variance):

$$\lambda^2[\Gamma(1 + \frac{2}{k}) - \Gamma(1 + \frac{1}{k})^2] = 3$$
$$3 = [\frac{4}{\Gamma(1 + \frac{1}{k})^2}][\Gamma(1 + \frac{2}{k}) - \Gamma(1 + \frac{1}{k})^2]$$
$$3 = 4(\frac{\Gamma(1 + \frac{2}{k})}{\Gamma(1 + \frac{1}{k})^2} - 1)$$
$$\frac{7}{4} = \frac{\Gamma(1 + \frac{2}{k})}{\Gamma(1 + \frac{1}{k})^2}$$

Now using the uniroot function in R to solve for k then substitute value of k to solve for λ.

$$k = 1.158$$
$$\lambda = 2.106$$

Log-Normal

Solving using R: Equation 1 (mean):

$$\mu = log(\frac{m^2}{\sqrt{\sigma^2 + \mu^2}})$$

Equation 2 (variance):

$$\sigma = \sqrt{log(\frac{\sigma^2}{(\mu^2 + 1)})}$$
$$\mu = 0.4133$$
$$\sigma = 0.7481$$

Negative Binomial

Equation 1 (mean):

$$\frac{pr}{(1-p)} = 2$$

Equation 2 (variance):

$$\frac{pr}{(1-p)^2} = 3$$

Solving for p, r:

$$r = 2(\frac{1-p}{p})$$

$$\frac{2p(\frac{(1-p)}{p})}{(1-p)^2} = 3$$

$$\frac{2}{(1-p)} = 3$$

$$p = 1/3$$

$$r = 4$$

Part B

The Figure below displays the 5 distributions with mean 2 and variance 3.

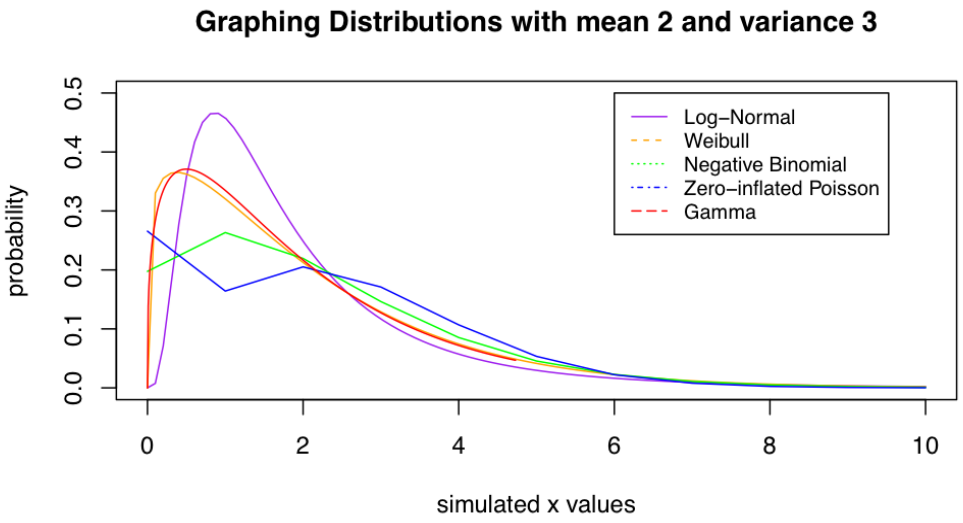


Figure 3: Plotting Distributions with Mean 2 Variance 3

Part C

The 99th percentile for the five distributions are in Table 1, calculated by taking the x-value when the CDF is equal to 0.99.

Table 1: 99th Percentile

Distribution	Percentile
Zero-inflated Poisson	7.000
Gamma	7.997
Weibull	7.874
Log-Normal	8.616
Negative Binomial	7.000

Part D

After computing 20 realisations from each distribution, the sample mean and variance are provided in Table 2. The 20 samples were randomly generated from the sample function.

Table 2: Sample Mean and Variance from 20 Randomly Simulated Realisations

Distribution	Sample.Mean	Sample.Variance
Zero-inflated Poisson	2.30	3.06
Gamma	2.32	3.36
Weibull	2.33	3.42
Log-Normal	2.27	2.71
Negative Binomial	2.20	3.01

1.3 Data Analysis

Interpretation of Coefficients

From the summary table computed, $e^{\beta_0+\beta_5 thorax}$ is the expected longevity of the average isolated fruitfly. The gamma model shows that the larger the thorax length, the longer the lifespan of the fruitfly. The coefficient: e^{β_1} or 1.057 represents a 5.7% increase in lifespan of a fruitfly provided with one pregnant fruitfly compared to the isolated fruitfly. However, since the p-value is large, this shows that β_1 is insignificant and that the fruitflies in this group have similar lifespans as the isolated group. The coefficient: e^{β_2} or 0.890 indicates that fruitflies that are given a virgin fruitfly experience a 11% decrease in longevity compared to isolated fruitflies. The coefficient: e^{β_3} or 1.086 shows that the group given many pregnant fruitflies experience a 8.6% increase in longevity. However, it is noted that the p-value is also large and we therefore cannot conclude that longevity increases. The coefficient: e^{β_4} or 0.661 shows that in the group given many virgin fruitflies, there is a significant decrease of 34% in lifespan compared to the isolated fruitfly.

Summary of Results: What kind of partners will increase the male fruitfly lifespan?

An analysis of whether longevity of male fruitflies depends on female fruitfly's state is conducted while controlling for thorax length, which is known to affect fruitfly longevity. It was discovered that male fruitflies' lifespan is dependent on the state of their female partner. Isolated male fruitflies had similar lifespan to those that were provided with pregnant female fruitflies. However, male fruitflies that were given virgin female fruitflies tended to have a significantly shorter lifespan. The more virgin female fruitflies provided, the shorter the male fruitflies' lifespan. These results indicate that there may be physiological downsides to reproduction in fruitflies (at the expense of a shorter lifespan).

Table 3: Summary of Fitted Gamma Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.887	0.194	9.726	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
thorax	2.688	0.228	11.804	0.000
shape	28.145	NA	NA	NA

Assessing Gamma Model Fit

From Figure 4, it appears that the Gamma distribution does not fit the data. 'Low' and 'Many' appear to follow the Gamma distribution while the other categories do not. The 'Isolated' graph does not have a mode around 60 days. Activity 'One' is left skewed whereas the 'High' appears to be right skewed.

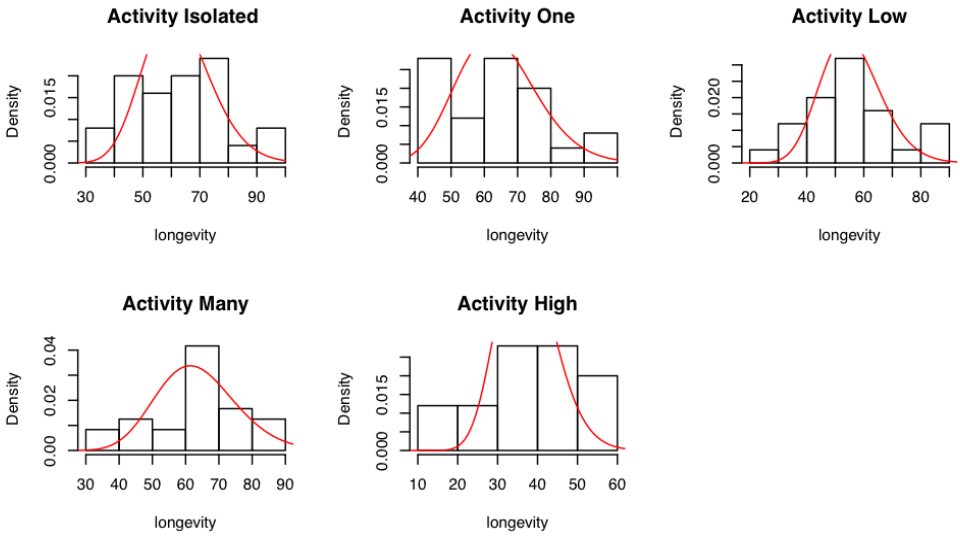


Figure 4: Evaluating Fit of Gamma Model on each Activity

Confidence Interval for Contrasts

In the fitted gamma model, the relevant contrast covariates are 'Low' and 'High' activity groups because their p-values are below the level of signficance (0.05). The 95% confidence interval is provided in Table 4.

Table 4: Confidence Intervals of the Significant Contrasts

Distribution	Lower.CI	Upper.CI
activitylow	0.8018	0.9881
activityhigh	0.5943	0.7342

1.4 Discussion

In Breiman's paper, "Statistical Modelling: The Two Cultures", it refutes the popular data model method and proposes the algorithmic modelling method instead. After reading this paper, I have re-evaluated the techniques used in the report and made suggestions to improve the analysis in the future.

Criticism of Hypothesis One and Two:

Breiman's paper comments on how R^2 is inflated because most models are overfitted with too many parameters. Because an inappropriate model was used to fit the data, any conclusions drawn based on any specific significance level would be erroneous. To evaluate the goodness of fit of the

logistic model, the R^2 was used. It would have been beneficial to fit the model graphically, use cross validation techniques to evaluate the accuracy of the proposed model and evaluate the residuals.

However, the algorithmic modelling approach is complex and suitable for predictability but not for interpretability. Since the obejective of the research question is to understand the impact of the variable on the response, black-box models would not be possible.

Criticism of the Secondary Problem:

The secondary problem asks to quantify the covariates. The covariates from an algorithmic model cannot be easily interpreted. However, Breiman criticizes that the goal is not interpretability, but accurate information. In my opinion, since the question asks for quantifiability, the algorithmic model would not be able to solve this even if the model is more accurate in predictability. I think it is important to understand the decrease in error rate from using a data model vs. algorithmic model; if the data model is slightly less accurate in predicting the response variable, but it can provide interpretability, a data model may be more useful (assuming appropriate goodness of fit).

2 A Study on American School Children and Their Likelihod of Smoking

2.1 Introduction

Smoking is highly addictive and is bad for the health. According to the National Health Service, smoking can cause several health issues such as lung cancer, increase risk of heart attacks, stomach cancer and more. It is important to understand the demograpgic profile of child smokers, in order to efficiently target health systems resources. To comprehend the demographics influencing youth to start smoking earlier, this report analyzes two hypotheses. The first hypothesis determines if it is more common for American children (ages 11 to 18) of European descent to chew tobacco, snuff or dip regularly compared to Hispanic and African-American children. The second hypothesis evaluates if a particular gender impacts the likelihood of using a hookah or waterpipe on at least one occasion, controlling for age, ethnicity, rural/urban and region. In addition to the two hypotheses, this report also quantifies how chewing tobacco changes with age, sex, and ethnic group.

The data used in this study is from the 2014 American National Youth Tobacco Survey. The survey is given to a sample of public and private school students enrolled in regular middle schools and high schools in the 50 US states.

2.2 Statistical Methods

Statistical Methods for Research Hypothesis 1

To evaluate if American youths of European descent tend to chew tobacco, snuff or dip regularly compared to Hispanic and African-American children, a logistic model is used. The logistic regression is useful when the response variable is binary (whether the youth chew tobacco, snuff or dip regularly or not).

To determine the appropriate covariates for the logistic model, the p-values are evaluated and scrutinized. A covariate is significant when the p-value of the predictor variable is less than 0.05.

Since European-Americans tend to live in rural areas and chewing tobacco tends to occur in the rural areas, the model must include the rural/urban covariate. The rural/urban variable is an effect modificer term (effect modifier is a variable that differently modifies the observed effect of a risk factor). To control for this, the interaction term: race and rural is added to the logistic model.

To evaluate the fit of the model, the McFadden's R^2 is used. McFadden's R^2 is defined as: $1 - \frac{ln(LM)}{ln(L0)}$.

LM is the fitted model and $L0$ is the model with only the intercept. Similar to R^2, McFadden's R^2 is also between 0 and 1, where values closer to 1 indicates that the model has a high predictive power.

To test the hypothesis: do European American youths tend to chew tobacco, snuff or dip regularly compared to Hispanic and African-American children, four test statistics are computed: To test for African-Americans and Urban:

$$H_0 : \beta_1 = 0$$

To test for African-Americans and Rural:

$$H_0 : \beta_1 + \beta_4 = 0$$

To test for Hispanic-Americans and Urban:

$$H_0 : \beta_2 = 0$$

To test for Hispanic-Americans and Rural:

$$H_0 : \beta_2 + \beta_8 = 0$$

With a significance level of 0.05, if the computed p-value is below 0.05, there is a high likelihood that there is difference in a particular race.

Statistical Methods for Research Hypothesis 2

To understand if gender impacts using a hookah or waterpipe on at least one occasion, a logistic model is applied. Because the response variable is binary (whether hookah or waterpipe was used or not), the logistic model is used.

The predictor demographic variables are: age, race, rural/urban, and region. Region is a covariate calculated from grouping several US states according the US Census: South, West, Midwest, and Northeast. Region is calculated because there would have been too many states (50) to include as individual variables.

After identifying the demographic variables to include in the model, a correlation analysis of the numerical variables is computed to test for correlation. This is important because multicollinearity occurs when a predictor variable in the model can be predicted from another variable in the model. When this occurs, it is difficult to confirm the significance of the individual predictors.

Similar to research hypothesis one, the McFadden's R^2 is used to evaluate model fit.

To calculate the probability that a particular sex uses a hookah or waterpipe on at least one occasion, it requires us to convert the coefficients of the variables into probability. This can be easily computed with the logistic model.

Statistical Methods for Secondary Problem

To quantify how the use of chewing tobacco changes with age, sex, and ethnic group, the logistic model is implemented because the response variable (chewing tobacco) is binary.

Similar to hypothesis one and two, the McFadden's R^2 is used to evaluate model fit.

The variables included in the model are: age, sex, and ethnic group. Correlation analysis is conducted to understand if there is multicollinearity that may affect the magnitude of other variables.

2.3 Results

Results: Research Hypothesis 1

Four test statistics were performed to evaluate whether there is a difference in regular use of chewing tobacco between races. European-Americans are compared to African-Americans for rural and urban locations. The p-value of all test statistics is less than the significance level, 0.05.

Table 5: Covariates Impact on Chewing Tobacco

	Estimate	Std. Error	z value	Pr(>|z|)
(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
Raceasian	-1.590	0.510	-3.118	0.002
Racenative	0.280	0.465	0.603	0.546
Racepacific	1.076	0.607	1.773	0.076
RuralUrbanRural	1.151	0.108	10.643	0.000
Raceblack:RuralUrbanRural	-0.743	0.336	-2.208	0.027
Racehispanic:RuralUrbanRural	-0.572	0.200	-2.856	0.004
Raceasian:RuralUrbanRural	0.544	0.659	0.825	0.409
Racenative:RuralUrbanRural	-0.659	0.569	-1.157	0.247
Racepacific:RuralUrbanRural	0.028	0.726	0.038	0.970

Table 6: Testing Coefficients of Race (European as Reference) controlling for Rural/Urban

Race	Urban_P.value	Rural_P.value
African	<0.05	<0.05
Hispanic	<0.05	<0.05

Results: Research Hypothesis 2

From the summary statistics, the coefficient of sex is not statistically significant variable, with a p-value of 0.359 (greater than the signifiance level of 0.05).

McFadden's R^2 is 0.106 which shows that the model is better than the logistic model with only the intercept.

Table 7: Covariates Impact on Using Hookah or Water Pipe

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-7.991	0.187	-42.694	0.000
SexF	0.040	0.043	0.927	0.354
Age	0.407	0.011	35.408	0.000
Raceblack	-0.607	0.071	-8.510	0.000
Racehispanic	0.295	0.050	5.921	0.000
Raceasian	-0.702	0.118	-5.931	0.000
Racenative	0.153	0.188	0.815	0.415
Racepacific	0.904	0.271	3.338	0.001
RuralUrbanRural	-0.381	0.050	-7.596	0.000
regionnortheast	0.220	0.070	3.149	0.002
regionsouth	0.153	0.066	2.324	0.020
regionwest	0.391	0.067	5.787	0.000

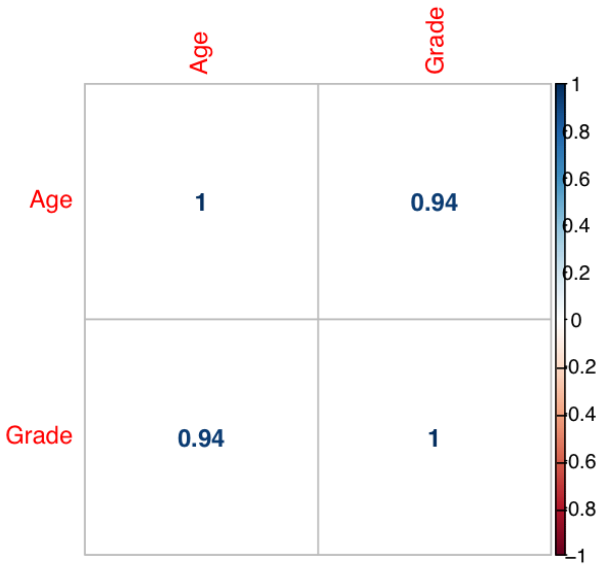


Figure 5: Correlation Analysis between Age and Grade

Results: Secondary Problem

The model shows the coefficients of the predictor variables in relation to the reference variables and the p-values represents the statistical significance.

To quantify how the use of chewing tobacco changes with age, sex and ethnic group, the probability

of the factor (π) affecting the response variable is solved for from the equation below:

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + ... + \beta_{11} x_{11}$$

Note that β_0 is the intercept term and represents the reference variables: European descent male Americans youths.

Table 8: Covariates Impact on Chewing Tobacco

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-6.876	0.316	-21.770	0.000
SexF	-1.714	0.104	-16.437	0.000
Age	0.300	0.020	15.031	0.000
Raceblack	-1.632	0.169	-9.680	0.000
Racehispanic	-0.803	0.098	-8.168	0.000
Raceasian	-1.756	0.323	-5.436	0.000
Racenative	0.088	0.274	0.321	0.748
Racepacific	0.840	0.357	2.354	0.019

2.4 Conclusion

From the analysis of the data, the following conclusions were generated about the research questions.

It was found that chewing tobacco, snuff or dip regularly is more common in European-Americans youths compared to Hispanic-Americans and African-American youths while controlling for rural/urban locations.

The likelihood of male youths using a hookah or waterpipe on at least one occasion is similar to female youths, while controlling for age, region, rural/urban and race.

In the secondary problem, there were a few findings. The logistic model shows that males are more likely to chew tobacco compared to females by 0.085%. The results also found that the older the youth, the higher the likelihood of chewing tobacco by 0.036%. The probability that a youth who is African American or hispanic to chew tobacco is respectively 0.083% and 0.057% less than European-American youths. The McFadden's R^2 for this model is 0.134, which indicates low predictability. Further covariates should be investigated to improve quantifiability of changes to age, sex, and ethnicity.

3 Appendix:

3.1 References

- Geography Division of the United States Census Bureau (2013), Census Regions and Division of the United States [data file]. Retrieved from https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
- National Health Service (January 20, 2019), Effects of Smoking on the Body, Retrieved from https://www.nhs.uk/smokefree/why-quit/smoking-health-problems
- Centers for Disease Control and Prevention (January 20, 2019), National Youth Tobacco Survey (NYTS), Retrieved from https://www.cdc.gov/tobacco/data_statistics/surveys/nyts/index.htm
- Matthew Analytics (August 17, 2015), Evaluating Logistic Regression Models [blog post], retrieved January 20, 2019, from https://www.r-bloggers.com/evaluating-logistic-regression-models/

3.2 R Code

Setup Code

```
knitr::opts_chunk$set(echo = TRUE)
library(MASS)
library(lmtest)
library(ggplot2)
library(faraway)
library(sqldf)
library(knitr)
library(pscl)
library(corrplot)
options(digits=4)
options(scipen=999)
```

1.1 Part A

```
correct_sum = 0
for (i in 1:100) {
  set.seed(i)
  x = seq(-10, 10, len=40)
  off = rep(c(1,-1), c(25, length(x)-25))
  y = rpois(length(x), exp(off + 0.5 +0.2*x))
  summary_mod = summary(glm(y~x+offset(off), family='poisson'))
  se = summary_mod$coefficients[2,2]
  estimate = summary_mod$coefficients[2,1]
  upper_bound = estimate + 2*se
  lower_bound = estimate - 2*se
  if ((0.2 <= upper_bound) &  (0.2 >=lower_bound)) {
    correct_sum = correct_sum+1
```

```
  }
}
p_coverage = correct_sum/100
```

1.1 Part B

```
lst = c()
for (i in 1:100) {
  set.seed(i)
  x = seq(-10, 10, len=40)
  off = rep(c(1,-1), c(25, length(x)-25))
  y = rpois(length(x), exp(off + 0.5 +0.2*x))
  summary_mod = summary(glm(y~x+offset(off), family='poisson'))
  coef = summary_mod$coefficients[2,1]
  lst = c(lst, coef)
}
beta_values = lst

sum_of_dat = 0
for (i in beta_values) {
  val = (i-0.2)^2
  sum_of_dat = sum_of_dat + val
}
sd_beta=sqrt(sum_of_dat/100)

#Plot:
x = seq(-10, 10, len=40)
#par(mfrow=c(1,2))
#hist(beta_values,
#     main="Normal Distribution",
#     xlab="Beta hat values", ylab="Frequency", xlim=c(0.15,0.30))
#curve(dnorm(x, mean=0.2, sd=sd_beta),  col="red", add=TRUE)
#qqnorm(beta_values)
#qqline(beta_values)
```

1.1 Part C

```
lrt_vec = c()
set.seed(0)
for (i in 1:100) {
  set.seed(i)
  x = seq(-10, 10, len=40)
  off = rep(c(1,-1), c(25, length(x)-25))
  y = rpois(length(x), exp(off + 0.5 +0.2*x))
  model_1 = glm(y~ offset(off + 0.2*x), family='poisson')
  model_full = glm(y~ x+offset(off), family='poisson')
```

```
  lrtest(model_1, model_full)
  value = model_1$deviance-model_full$deviance
  lrt_vec = c(lrt_vec, value)
}
#hist(lrt_vec, breaks=100, freq=FALSE,
#     main="Fitting Chi-Squared Distribution", xlab = "Likelihood Ratio Statistics")
#curve(dchisq(x, df=1,ncp = 0, log = FALSE),col="red", add=TRUE)
```

1.2 Part A

```
optimize_weibull <- function(k){
    return(gamma(1+2/k)/(gamma(1+1/k))^2 - 7/4)
}
k = uniroot(optimize_weibull, c(1,3))$'root'
lambda = 2/gamma(1+1/k)

v = 3
m = 2
mu = log((m^2)/sqrt(v+(m^2)))
sig = sqrt(log(v/(m^2)+1))
```

1.2 Part B

```
#Gamma
alpha=4/3
beta = 2/3
avrg = alpha*beta
std.dv=sqrt(alpha*beta^2)
x=50
x_gam=seq(0,avrg+5*std.dv,0.01)
y_gam=dgamma(x_gam, alpha,rate=beta)
#plot(x_gam, y_gam, type='l', col="red")

#Zero inflated Poisson
p_i=1/5
lambda=5/2
poi_0 = p_i + (1-p_i)*exp(-lambda)
x_poi=seq(0,10)
y_poi = c(poi_0)
for (i in 1:10) {
  val = ((1-p_i)*(lambda^i)*exp(-lambda))/factorial(i)
  y_poi=c(y_poi,val)
}
#plot(x_poi, y_poi, type='l', ylim=c(0,max(y_poi)+0.01), col="blue")

# Negative Binomial
```

```r
p=2/3
r=4
x_nbin=seq(0,10)
y_nbin = dnbinom(x_nbin, r, p, log = FALSE)
#plot(x_nbin, y_nbin, type='l', ylim=c(0,max(y_nbin)+0.01), col="green")

#Weibull
k = 1.158 #shape
c = 2.106 #scale
x_wb=seq(0,10,length=100)
y_wb = dweibull(x_wb, k, c, log = FALSE)
#plot(x_wb, y_wb, type='l', col="orange")

#Log normal
mu=0.4133393
sigma=0.7480747
x_ln=seq(0.0001,10,length=100)
y_ln = c()
for (i in x_ln) {
  val = (1/(i*sigma*sqrt(2*pi)))*exp(-((log(i)-mu)^2)/(2*sigma^2))
  y_ln=c(y_ln,val)
}
#plot(x_ln, y_ln, type='l', col="purple")

# All 5 distributions
# plot(x_ln, y_ln, type='l',xlim=c(0,10), ylim=c(0,0.5), xlab="x", ylab="y", col="purple", mai
# lines(x_wb, y_wb, type='l', col="orange")
# lines(x_nbin, y_nbin, type='l', col="green")
# lines(x_poi, y_poi, type='l', col="blue")
# lines(x_gam, y_gam, type='l', col="red")
# legend(6, 0.5, legend=c("Log-Normal", "Weibull", "Negative Binomial", "Zero-inflated Poisson
#        col=c("purple", "orange", "green", "blue", "red"), lty=1:5, cex=0.8)
```

1.2 Part C

```r
tbl_percentile <- data.frame(
  "Distribution" = c("Zero-inflated Poisson", "Gamma", "Weibull","Log-Normal","Negative Binomi
  "Percentile" = c(7,7.997, 7.874,8.616,7))
knitr::kable(
  tbl_percentile, caption = 'Percentile 99th'
)
```

Table 9: Percentile 99th

Distribution	Percentile
Zero-inflated Poisson	7.000
Gamma	7.997

Distribution	Percentile
Weibull	7.874
Log-Normal	8.616
Negative Binomial	7.000

```
gam_per = qgamma(0.99, alpha, rate=beta)
wb_per = qweibull(0.99, k, c, lower.tail = TRUE, log.p = FALSE);
nbin_per = qnbinom(0.99, r, p, log.p = FALSE);
zpoi_per = 0.26566800+0.16417000+0.20521250+0.17101041+0.10688151+0.0534407543+0.0222669810
lnorm_per = exp(qnorm(0.99)*sigma+mu)
```

1.2 Part D

```
tbl_percentile <- data.frame(
  "Distribution" = c("Zero-inflated Poisson", "Gamma", "Weibull","Log-Normal","Negative Binomi
  "Sample Mean" = c(2.30, 2.32, 2.33, 2.27, 2.20),
  "Sample Variance" = c(3.06, 3.36, 3.42, 2.71, 3.01)
  )
knitr::kable(
  tbl_percentile, caption = 'Sample Mean and Variance from 20 Random Realisations'
)
```

Table 10: Sample Mean and Variance from 20 Random Realisations

Distribution	Sample.Mean	Sample.Variance
Zero-inflated Poisson	2.30	3.06
Gamma	2.32	3.36
Weibull	2.33	3.42
Log-Normal	2.27	2.71
Negative Binomial	2.20	3.01

```
#Random Sample
set.seed(123)
x <- sample(0:999,20,TRUE)
sample_x =x/1000

#Weibull
k = 1.158 #shape
c = 2.106 #scale
wb_y_sample=qweibull(sample_x, k, c, lower.tail = TRUE, log.p = FALSE)
wb_sample_mean=mean(wb_y_sample)
wb_sample_var = var(wb_y_sample)

#Gamma
gamma_y_sample= qgamma(sample_x, alpha, rate=beta);
```

```r
gamma_sample_mean=mean(gamma_y_sample)
gamma_sample_var=var(gamma_y_sample)

#Zero-inflated Poisson
p_i=1/5
lambda=5/2
x_poi=seq(0,15)
y_poi = c(poi_0)
for (i in 1:15) {
  val = ((1-p_i)*(lambda^i)*exp(-lambda))/factorial(i)
  y_poi=c(y_poi,val)
}

poisson_sample=c()
for (i in sample_x) {
  prob = i
  n = 1
  while (prob > y_poi[n]) {
    prob = prob - y_poi[n]
    n=n+1
    }
  poisson_sample = c(poisson_sample,n-1)
}
poi_mean = mean(poisson_sample)
poi_var = var(poisson_sample)

#Lognormal
lnorm_y_sample = exp(qnorm(sample_x)*sigma+mu);
lnorm_sample_mean=mean(lnorm_y_sample)
lnorm_sample_var=var(lnorm_y_sample)

# Negative Binomial
p=2/3
r=4
qnbinom_y_sample=qnbinom(sample_x, r, p, log.p = FALSE); # qnbinom_y_sample
qnbinom_sample_mean=mean(qnbinom_y_sample)
qnbinom_sample_var=var(qnbinom_y_sample)
```

1.3

```r
d = data('fruitfly', package='faraway')
sum_fruitfly = summary(fruitfly)
apply(fruitfly, 2, function(x) any(is.na(x)))
```

```
##    thorax longevity  activity
##     FALSE     FALSE     FALSE
```

```r
model_0 = glm(longevity~activity+thorax,family=Gamma(link='log'),data=fruitfly)
sum_mod0 = summary(model_0)
knitr::kable(rbind(summary(model_0)$coef, shape=c(1/summary(model_0)$dispersion, NA, NA,NA)), 
```

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.8872	0.1940	9.726	0.0000
activityone	0.0553	0.0534	1.036	0.3024
activitylow	-0.1165	0.0533	-2.184	0.0309
activitymany	0.0825	0.0541	1.524	0.1302
activityhigh	-0.4147	0.0539	-7.687	0.0000
thorax	2.6878	0.2277	11.804	0.0000
shape	28.1455	NA	NA	NA

```r
shape = 1/summary(model_0)$dispersion
scale = exp(model_0$coef["(Intercept)"])/shape

scale2 = exp(model_0$coef["(Intercept)"]+model_0$coef["activityone"]+
               (model_0$coef["thorax"]*mean(fruitfly$thorax[fruitfly$activity=='one'])))/shape

scale3 = exp(model_0$coef["(Intercept)"]+model_0$coef["activitylow"]+
               (model_0$coef["thorax"]*mean(fruitfly$thorax[fruitfly$activity=='low'])))/shape

scale4 = exp(model_0$coef["(Intercept)"]+model_0$coef["activitymany"]+
               (model_0$coef["thorax"]*mean(fruitfly$thorax[fruitfly$activity=='many'])))/shap

scale5 = exp(model_0$coef["(Intercept)"]+model_0$coef["activityhigh"]+
             (model_0$coef["thorax"]*mean(fruitfly$thorax[fruitfly$activity=='high'])))/shape

xSeq = seq(0,100,len=1000)
# hist(fruitfly$longevity[fruitfly$activity=='isolated'], xlab='longevity', main='Activity Iso
# lines(xSeq, dgamma(xSeq, shape = shape, scale = scale), col = "red")
#
# hist(fruitfly$longevity[fruitfly$activity=='one'], xlab='longevity', main='Activity One', pr
# lines(xSeq, dgamma(xSeq, shape = shape, scale = scale2), col = "red")
#
# hist(fruitfly$longevity[fruitfly$activity=='low'], xlab='longevity', main='Activity Low', pr
# lines(xSeq, dgamma(xSeq, shape = shape, scale = scale3), col = "red")
#
# hist(fruitfly$longevity[fruitfly$activity=='many'], xlab='longevity', main='Activity Many', 
# lines(xSeq, dgamma(xSeq, shape = shape, scale = scale4), col = "red")
#
# hist(fruitfly$longevity[fruitfly$activity=='high'], xlab='longevity', main='Activity High', 
# lines(xSeq, dgamma(xSeq, shape = shape, scale = scale5), col = "red")

# CI:
CI = c()
```

```r
for (i in 2:5) {
  estimate =summary(model_0)$coefficients[i,1]
  se = summary(model_0)$coefficients[i,2]
  upper = exp(estimate + 1.96*se)
  lower = exp(estimate - 1.96*se)
  u = (upper)
  l = (lower)
  interval = c(l,u)
  CI = c(CI, interval)
}
tbl_CI <- data.frame(
  "Distribution" = c("activitylow", "activityhigh"),
  "Lower CI" = c(0.8018,0.5943),
  "Upper CI" = c(0.9881, 0.7342)
  )
# knitr::kable(
#   tbl_CI, caption = 'Confidence Intervals of the Significant Contrasts'
# )
```

2 Setup

```r
dataDir = "/Users/jchau/Documents/Learnings/STA 2102 Applied Stats 2/"
smokeFile = file.path(dataDir, "smokeDownload.RData")
if (!file.exists(smokeFile)) {
download.file("http://pbrown.ca/teaching/astwo/data/smoke.RData", smokeFile)
}
(load(smokeFile))
```

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

```r
for(D in c('chewing_tobacco_snuff_or', 'ever_tobacco_hookah_or_wa')){
cat("- `", D, "`: ",
as.character(smokeFormats[match(D, smokeFormats[,'colName']), 'label']),
'\n\n', sep='')
}
```

```
## - `chewing_tobacco_snuff_or`: RECODE: Used chewing tobacco, snuff, or dip on 1 or more days
##
## - `ever_tobacco_hookah_or_wa`: RECODE: Ever smoked tobacco out of a hookah or waterpipe
```

2.1 Setup

```r
df <- sqldf("
select
  student
  , RuralUrban
  , Race
  , case when RuralUrban = 'Urban' and Race = 'black' then 1 else 0 end as p11
```

```
  , case when RuralUrban = 'Urban' and Race = 'hispanic' then 1 else 0 end as p12
  , case when RuralUrban = 'Urban' and Race = 'white' then 1 else 0 end as p13
  , case when RuralUrban = 'Rural' and Race = 'black' then 1 else 0 end as p21
  , case when RuralUrban = 'Rural' and Race = 'hispanic' then 1 else 0 end as p22
  , case when RuralUrban = 'Rural' and Race = 'white' then 1 else 0 end as p23
  , case when chewing_tobacco_snuff_or=1 then 1 else 0 end as y_out
from smoke
where Race in ('white', 'black', 'hispanic')
and Race is not null
and student is not null
and RuralUrban is not null
and chewing_tobacco_snuff_or is not null
")

model_1 = glm(y_out ~ Race + RuralUrban + Race*RuralUrban, family=binomial, data=df)
sum_mod1 = summary(model_1)
model_1fit = pR2(model_1)


Wtest = function(L,Tn,Vn,h=0) # H0: L theta = h
# Tn is estimated theta, usually a vector.
# Vn is the estimated asymptotic covariance matrix of Tn.
# For Wald tests based on numerical MLEs, Tn = theta-hat,
# and Vn is the inverse of the Hessian of the minus log
# likelihood.
     {
     Wtest = numeric(3)
     names(Wtest) = c("W","df","p-value")
     r = dim(L)[1]
     W = t(L%*%Tn-h) %*% solve(L%*%Vn%*%t(L)) %*%
          (L%*%Tn-h)
     W = as.numeric(W)
     pval = 1-pchisq(W,r)
     Wtest[1] = W; Wtest[2] = r; Wtest[3] = pval
     Wtest
     } # End function Wtest

# Testing for Rural Africans vs Europeans H0 = beta1 + beta4
betahat1 = model_1$coefficients;betahat1
```

```
##                  (Intercept)                    Raceblack
##                      -3.6408                      -0.9706
##                  Racehispanic               RuralUrbanRural
##                      -0.3135                       1.1511
##    Raceblack:RuralUrbanRural Racehispanic:RuralUrbanRural
##                      -0.7429                      -0.5718
```

```r
L0=rbind(c(0,1,0,0,1,0))
Vhat = vcov(model_1)
Wtest(L0, betahat1, Vhat)
```

```
##                   W                  df              p-value
## 57.63360975774693884  1.00000000000000000  0.00000000000003153
```

```r
# Testing for Rural Hispanics vs Europeans H0 = beta1 + beta4
betahat1 = model_1$coefficients;betahat1
```

```
##                  (Intercept)                    Raceblack
##                      -3.6408                      -0.9706
##                 Racehispanic               RuralUrbanRural
##                      -0.3135                       1.1511
##    Raceblack:RuralUrbanRural Racehispanic:RuralUrbanRural
##                      -0.7429                      -0.5718
```

```r
L1=rbind(c(0,0,1,0,0,1))
Vhat = vcov(model_1)
Wtest(L1, betahat1, Vhat)
```

```
##                   W                 df             p-value
## 48.303600800316907  1.000000000000000  0.000000000003651
```

```r
tbl_race <- data.frame(
  "Race" = c("African", "Hispanic"),
  "Urban_P-value" = c('<0.05','<0.05'),
  "Rural_P-value" = c('<0.05','<0.05'))
# knitr::kable(
#  tbl_race, caption = 'Testing Coefficients of Race (European as Reference) controlling for R
#)
```

2.2

```r
# The likelihood of using a hookah or waterpipe on at least one occasion
# https://www2.census.gov/geo/pdfs/maps-data/maps/reference/us_regdiv.pdf
# method 1 Logistic

df_test <- sqldf("
  select
  case when state in ('WA', 'OR', 'MT', 'WY', 'ID', 'CA', 'NV','UT','CO', 'AZ', 'NM') then 'we
     when state in ('ND', 'SD', 'NE', 'KS', 'MN', 'IA', 'MO','WI','IL', 'IN', 'MI', 'OH') then
     when state in ('PA', 'NY', 'NJ', 'CT', 'MA', 'VT', 'NH', 'ME') then 'northeast'
     when state in ('TX', 'OK', 'AR', 'LA', 'MS', 'AL', 'TN', 'KY','GA','FL','SC','NC','VA','W
     ELSE state end as region
  , count(student) as count_total
  , sum(ever_tobacco_hookah_or_wa) as tob
from smoke
group by 1
```

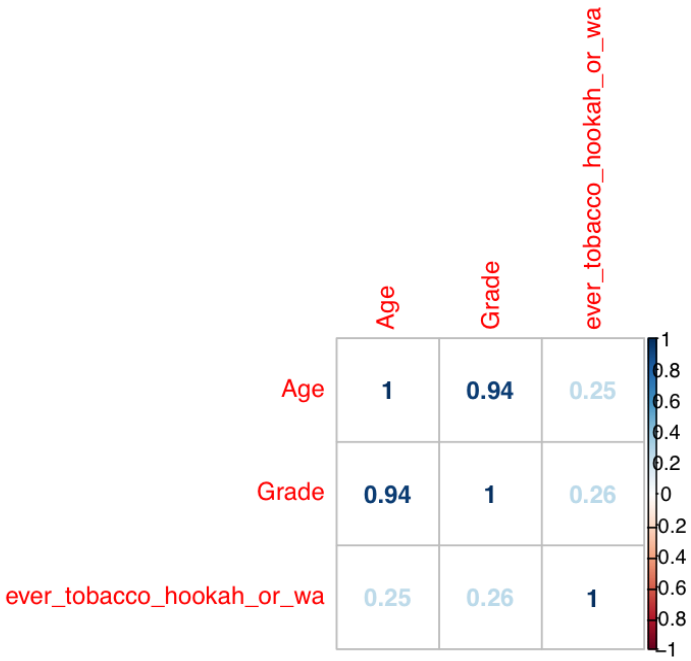
```
")
# df_test$prop = df_test$tob/df_test$count_total

# what is the null rate?
df2 <- sqldf("
select
  student
  , sex
  , age
  , Race
  , RuralUrban
  , ever_tobacco_hookah_or_wa
from smoke
where sex is not null
and age is not null
and race is not null
and ever_tobacco_hookah_or_wa is not null
and RuralUrban is not null
and student is not null
")

c <- sqldf("
select
  age
  , grade
  , case when ever_tobacco_hookah_or_wa=1 then 1 else 0 end as ever_tobacco_hookah_or_wa
from smoke
where age is not null
and grade is not null
")


model_2 = glm(ever_tobacco_hookah_or_wa~Sex+Age+Race+RuralUrban, family=binomial, data=df2)
sum_mod2 = summary(model_2)

# Correlation of Variables
M<-cor(c)
corrplot(M, method="number")
```




```
# It appears that sex is not a factor when students uses a hookah, provided their age, ethnici
model_2fit = pR2(model_2)
```

2.3

```
df3 <- sqldf("
select
  student
  , sex
  , age
  , Race
  , chewing_tobacco_snuff_or
from smoke
where sex is not null
and age is not null
and race is not null
and chewing_tobacco_snuff_or is not null
")

model_3 = glm(chewing_tobacco_snuff_or~Sex+Age+Race, family=binomial, data=df3)
sum_mod3 = summary(model_3)

pi_0 = exp(summary(model_3)$coefficients[1,1])/(1+exp(summary(model_3)$coefficients[1,1]))
```

```
pi_sex = exp(summary(model_3)$coefficients[1,1]+summary(model_3)$coefficients[2,1])/
  (1+exp(summary(model_3)$coefficients[1,1]+summary(model_3)$coefficients[2,1]))
(pi_sex - pi_0)*100
```

```
## [1] -0.08449
```

```
pi_age = exp(summary(model_3)$coefficients[1,1]+summary(model_3)$coefficients[3,1])/
  (1+exp(summary(model_3)$coefficients[1,1]+summary(model_3)$coefficients[3,1]))
(pi_age - pi_0)*100
```

```
## [1] 0.03596
```

```
pi_blk = exp(summary(model_3)$coefficients[1,1]+summary(model_3)$coefficients[4,1])/
  (1+exp(summary(model_3)$coefficients[1,1]+summary(model_3)$coefficients[4,1]))
(pi_blk - pi_0)*100
```

```
## [1] -0.08291
```

```
pi_his = exp(summary(model_3)$coefficients[1,1]+summary(model_3)$coefficients[5,1])/
  (1+exp(summary(model_3)$coefficients[1,1]+summary(model_3)$coefficients[5,1]))
(pi_his - pi_0)*100
```

```
## [1] -0.05687
```

```
model_3fit = pR2(model_3)
```