STA 138: Final Project

Aman Singh & Jennifer Li
ID:915763034
ID:999339769
UC Davis
Instructor: Prabir Burman

3/09/2021

Question 1: Low Birth Rate

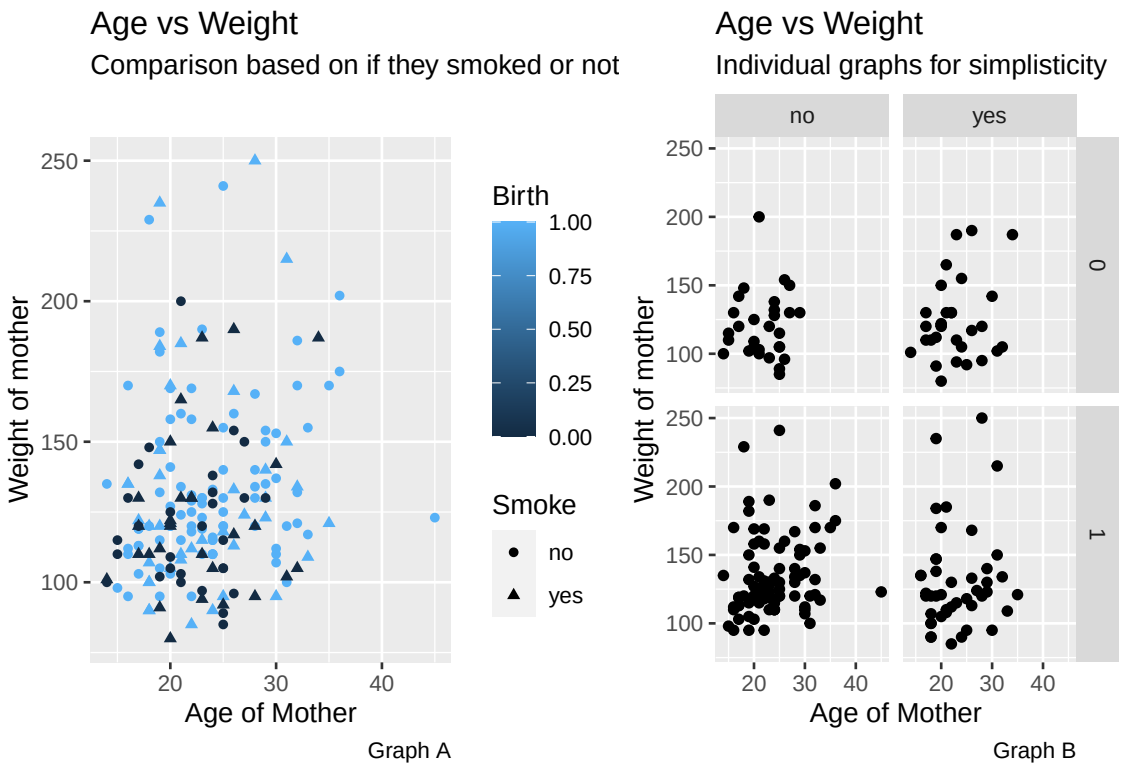
(I) Introduction

Smoking has been a huge issue since the 20^{th} century. There used to be advertisements all the time on the television where they would show as people having a great time and enjoying life. However, decades later, people found out that smoking leads to cancer and to horrible birth defects. Upon learning this, the government banned ads for smoking and started to limit the exposure of smoking to the public. Even when buying a cigarette, you have to be 18 years old as well as there is a huge caution warning on the box stating that it can lead to various defects in the body. The worst defect that we have seen so far has been when women who are pregnant are smoking at the same time. This can lead to some serious defects to the children and can affect them in the long run. While we do know that, what we want to investigate is whether the probability of low birth weight of infants is related to information on mother such as age, weight smoking status. We will be using the dataset called `Baby` provided by Professor Prabir Burman. `Baby` has 7 columns, `age`(age of the mother), `weight`(weight of the mother before pregnancy), `smoke`(smoking status during pregnancy), `pre`(history of premature labor), `hyp`(history of hypertension), `visits`(the number of visits during the first trimester), and `birth`(if the birth weight of the infant was low or not).

(II) Materials and Methods

Table 1: Summary Statistics

	Age	Weight	Visits	Birth
Min:	14	80	0	0
1st Qt.:	19	110	0	0
Median:	23	121	0	1
Mean:	23.24	129.8	.7937	.6878
3rd Qt:	26	140	1	1
Max:	45	250	6	1



We notice right away from the summary of the table that there seems to be multiple binomial variables in

this dataset. We see that `smoke`, `pre`, and `hyp` are all categorical variables with the 2 levels being `yes` or `no`. We wanted to make a way so we can start off with a regular logistic regression without any transformations or any interaction terms. What we will do next is to compare it to an interaction regression. We will then use an Goodness-of-Fit test to see if we can eliminate any variables and then proceed to use an stepwise regression to confirm if our regression is valid. We will then use that final regression to predict the birth column of the `Baby` dataset. After that we will compare our model with the dataset and see how accurate our model was.

(III) Results

We made the categorical variables (`Smoke`,`Pre`,`Hyp`) an factor to accurately account for the response. The regression we started off with is the following:

$$Birth = Age + Weight + Factor(Smoke) + Factor(Pre) + Factor(Hyp) + Visits + \epsilon_t$$

Table 2: Logistic Regression 1 Results:

	Estimate	**Standard Error**	**Z-Value**	**P.Value**
Intercept:	-2.021488	1.113152	-1.816	0.06937
Age:	0.059091	0.036965	1.599	0.10992
Weight:	0.016086	0.006943	2.317	0.02051
Factor.Smoke.Yes:	-0.513740	0.349295	-1.471	0.14135
Factor.Pre.Yes:	-1.798908	0.510014	-3.527	0.00042
Factor.Hyp.Yes:	-1.772643	0.717756	-2.470	0.01352
Visits:	0.032113	0.178906	0.179	0.85755

Null deviance: 234.67 on 188 degrees of freedom
Residual deviance: 202.15 on 182 degrees of freedom
AIC: 216.15

Based on the rule of thumb, we see that `Weight`,`factor.pre.yes`,`factor.hyp.yes`, all seem to be statistically significant. We see here that the `factor.smoke.yes`, `factor.pre.yes`,and `factor.hyp.yes` all are accounted in the regression if the mother smokes,has a history of pre-mature labor, and has an history of hypertension.

We than try out an regression including the interaction terms stated in the problem. We add Weight & Pre together, Age & Weight together, and Weight & Hyp. We build an logistic regression that looks like the following:

$Birth = Age + Weight + Factor(Smoke) + Factor(Pre) + Factor(Hyp) + Visits + Age : Weight + Weight : Factor(Pre) + Weight : Factor(Hyp) + \epsilon_t$

Table 3: Logistic Regression 2 Results:

	Estimate	Standard Error	Z-Value	P.Value
Intercept:	-3.2742473	3.9295522	-0.833	0.405
Age:	0.1165393	0.1682394	0.693	0.488
Weight:	0.0258898	0.0304761	0.850	0.396
Factor.Smoke.Yes:	-0.5087686	0.3545077	-1.435	0.151
Factor.Pre.Yes:	-2.6303649	2.8680721	-0.917	0.359
Factor.Hyp.Yes:	-1.6168753	2.6785514	-0.604	0.546
Visits:	0.0292054	0.1802465	0.162	0.871
Age:Weight	-0.0004465	0.0012718	-0.351	0.726
Weight:Factor(pre)yes	0.0064942	0.0222361	0.292	0.770
Weight:Factor(hyp)yes	-0.0009711	0.0171863	-0.057	0.955

Null deviance: 234.67 on 188 degrees of freedom
Residual deviance: 201.96 on 179 degrees of freedom
AIC: 221.96

We immediately notice between the 2 regressions is that the AIC immediately increases between the 2 regressions.This is not a good look, which seems to be that we would have to drop some variables to lower the AIC. The AIC is basically an mathematical method to evaluate how well an model fits the data. We also notice that the Residual deviance goes down which means that we should also look at smaller variables in the regression. We also see that none of the variables are statistically significant.

We initially decided to try out an Goodness-of-Fit test on this data using categorical data, however we saw that the data doesn't seem to be good because the `age` variable is very widely dispersed as well as the `weight` variable leading us to perform the Likelihood ratio test instead.We then decide to carry out an likelihood ratio test to decide if we should drop the interaction terms. We will than reject or fail to reject based off the following hypothesis:

$$H_0 : B_{age:weight+weight:factor(pre)+weight:factor(hyp)} = 0$$
$$H_A : B_{age:weight+weight:factor(pre)+weight:factor(hyp)} \neq 0$$

As we can see testing the full model with the interactions compared to the reduced model, since the p-value is `.9789857`, we end up dropping the interaction variables as they do not provide any use to us. I also believe that we should test one more time for the variable `visits` as it doesn't seem to be that important of a variable. We use the following likelihood ratio hypothesis test:

$$H_0 : B_{visits} = 0$$

$$H_A : B_{visits} \neq 0$$

We also see that since the P-value is `.8572994`, we also fail to reject the Null Hypothesis which lets us drop the variable `visits` from the regression.

We will now use Stepwise regression to see if the model that we fit and tested is accurate to the Stepwise Regression fit, which uses the AIC values to calculate best fit. The backward model starts with the full model and goes to the null model to analyze best fit, while the forward model starts with the null model and goes to the full model. We will use both of them combined just for better analysis.

After running the model,we notice that the Front/Backward Step-wise regression function also provided the same regression as what we originally thought it was with all the interactions dropped as well as the variable `visits`. Our final regression looks like:

$$Birth = Age + Weight + Factor(Smoke) + Factor(Pre) + Factor(Hyp) + \epsilon_t$$

with the summary statistics being:

Table 4: Logistic Regression 3 Results:

	Estimate	**Standard Error**	**Z-Value**	**P.Value**
Intercept:	-3.2742473	3.9295522	-0.833	0.405
Age:	0.1165393	0.1682394	0.693	0.488
Weight:	0.0258898	0.0304761	0.850	0.396
Factor.Smoke.Yes:	-0.5087686	0.3545077	-1.435	0.151
Factor.Pre.Yes:	-2.6303649	2.8680721	-0.917	0.359
Factor.Hyp.Yes	-1.782710	0.716698	-2.487	0.012868

Null deviance: 234.67 on 188 degrees of freedom
Residual deviance: 202.19 on 183 degrees of freedom
AIC: 214.19

Comparing the last 3 regressions, we notice that this is the best fit model that we have tried so far. The reasons being that the AIC is at its lowest of the 3.

Using this final regression, we will then try to approximate our results by making a test dataset which has all the same columns as the `Baby` dataset except for the column Birth. We will then compare our results to the real world data to see how accurate our model is.

We made our column into a logical class where it displayed true if the prediction was equal to the Birth column in `Baby` and we found the length of the True which ended up being $\frac{141}{189} = 74.6\%$. This means our model accurately explains 74.6% of the data in the `Baby` dataset.

(IV) Conclusion

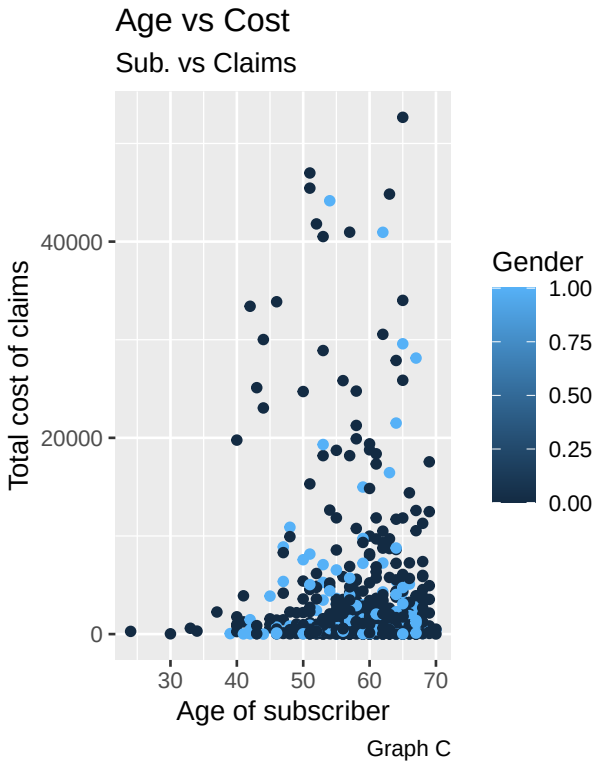
Seeing that our data accurately predicted 74.6% of the data in the dataset, we think that it's a result of a pretty accurate and good model. It is possible to increase the validity and accuracy of this model maybe if we added transformations and of course other variables such as whether they attended college, if their parents smoked, etc. While our model works pretty well on this dataset, it remains to be seen if it can do the same on another dataset with different variables. Another reasoning behind the predictability of it being 74.6% is largely because of outside influences on people. Everyone is different and is not bound by the variables described in the dataset. Thus, no model can 100% accurately describe a dataset, thus describing the error term at the end of a regression, ϵ_t. We also highly recommend not to base any claims off of this final regression model as the dataset is not that big as well as not that good in the amount of variables they should account for.

Question 2: Ischemic heart disease

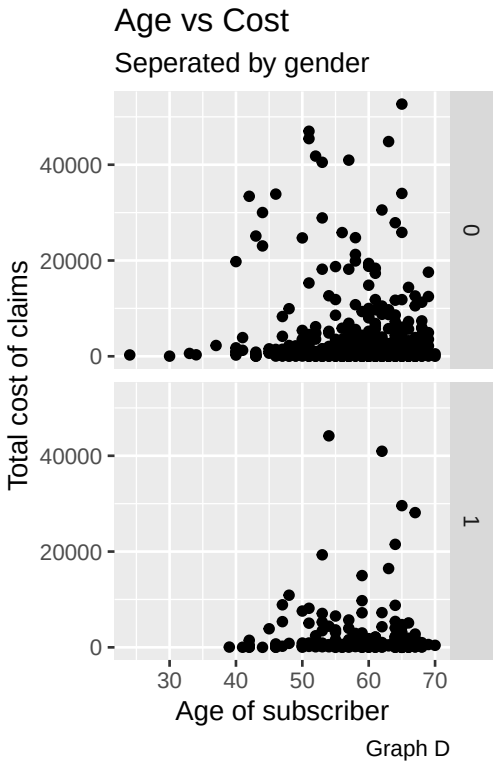
(I) Introduction

Heart disease is one of the most common diseases which can lead to death. It is in fact a very serious disease and is a type of disease which starts off by having a buildup of plaque. This leads to the coronary arteries thus having to narrow, which limits the blood flow to the heart. Some symptoms of coronary artery disease can really range from no symptoms, to chest pain and even a heart attack depending on the person. Some treatment can help but ultimately there is no cure and would have to deal with for the rest of your life. In this paper, we are looking into a dataset called `ischemic` given to us by Professor Prabir Burman which contains 9 columns. The 9 columns are `cost`(the total cost of claims made by the subscriber), `age`(age of the subscriber), `gender`(gender of subscriber),`inter`(total number of interventions or procedures carried out),`drugs`(number of tracked drugs prescribed), `complications`(number of other complications that came from the heart treatment), `comorbidities` (number of other diseases that the subscriber had during the period), `duration` (number of days of duration of treatment condition), and `visits` (number of emergency room visits). We will be using this dataset to perform a poisson regression that will perform a data summary, goodness-of-fit and model selection to model the mean as a function of 8 other variables.

(II) Materials and Methods



Graph C



Graph D

(III) Results

(IV) Conclusion

Ⓥ

Code Appendix

```
# cuttingoffcode
library(knitr)
opts_chunk$set(tidy.opts = list(width.cutoff = 70), tidy = TRUE)
# library & importing data
library(readxl)
library(tidyverse)
library(patchwork)
library(MASS)
baby <- read_excel("baby.xls")
summary(baby)
# analysis for question 1
p = baby %>% ggplot() + geom_point(mapping = aes(x = age, y = weight, color = birth,
    shape = smoke)) + labs(title = "Age vs Weight", subtitle = "Comparison based on if they smoked or n
    x = "Age of Mother", y = "Weight of mother", color = "Birth", caption = "Graph A",
    shape = "Smoke")

b = baby %>% ggplot(aes(age, weight)) + geom_point() + facet_grid(vars(birth),
    vars(smoke)) + labs(title = "Age vs Weight", subtitle = "Individual graphs for simplisticity",
    x = "Age of Mother", y = "Weight of mother", caption = "Graph B")

p + b
z = glm(birth ~ age + weight + factor(smoke) + factor(pre) + factor(hyp) +
    visits, data = baby, family = "binomial")
summary(z)
x = glm(birth ~ age + weight + factor(smoke) + factor(pre) + factor(hyp) +
    visits + age:weight + weight:factor(pre) + weight:factor(hyp), data = baby,
    family = "binomial")
summary(x)
fit_full = glm(birth ~ age + weight + factor(smoke) + factor(pre) + factor(hyp) +
    visits + age:weight + weight:factor(pre) + weight:factor(hyp), data = baby,
    family = "binomial")
fit_reduced = glm(birth ~ age + weight + factor(smoke) + factor(pre) +
    factor(hyp) + visits, data = baby, family = "binomial")
G2 = fit_reduced$deviance - fit_full$deviance
1 - pchisq(G2, df = 3)
fit_full = glm(birth ~ age + weight + factor(smoke) + factor(pre) + factor(hyp) +
    visits, data = baby, family = "binomial")
fit_reduced = glm(birth ~ age + weight + factor(smoke) + factor(pre) +
    factor(hyp), data = baby, family = "binomial")
G2 = fit_reduced$deviance - fit_full$deviance
1 - pchisq(G2, df = 1)
step.model <- stepAIC(x, direction = "both", trace = FALSE)
summary(step.model)
final = glm(formula = birth ~ age + weight + factor(smoke) + factor(pre) +
    factor(hyp), family = "binomial", data = baby)

summary(final)
test <- data.frame(baby[1:189, 1:6])
prediction = predict(final, test, type = "response")
lol = ifelse(prediction > 0.5, 1, 0)
```

```
cbind(prediction, lol, baby$birth)
cool = lol == baby$birth
length(cool[cool == TRUE])
# Question 2
ischemic <- read_excel("ischemic.xlsx")
a = ischemic %>% ggplot() + geom_point(mapping = aes(x = age, y = cost,
    color = gender)) + labs(title = "Age vs Cost", subtitle = "Sub. vs Claims",
    x = "Age of subscriber", y = "Total cost of claims", color = "Gender",
    caption = "Graph C", shape = "Smoke")

b = ischemic %>% ggplot(aes(age, cost)) + geom_point() + facet_grid(vars(gender)) +
    labs(title = "Age vs Cost", subtitle = "Seperated by gender", x = "Age of subscriber",
        y = "Total cost of claims", caption = "Graph D")


a + b
```