INDE546_HW7

Michael Shieh

2/25/2020

```
setwd("~/Documents/UW/2020Winter/INDE546_InferentialDataAnalysis/HW/HW7")
survey <- read.csv("Class_Survey_W20.csv", header = TRUE)
```

Exercise 1

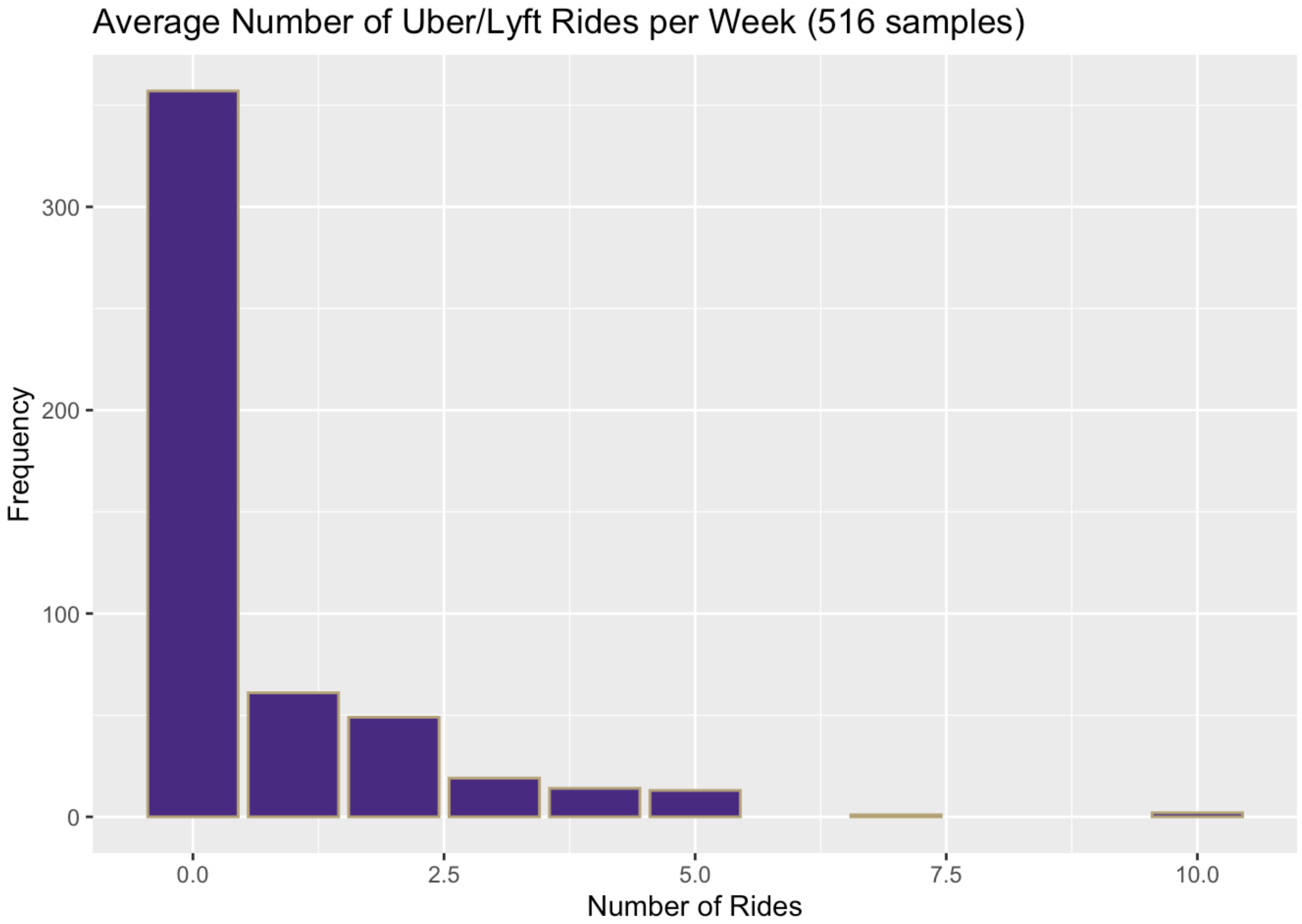
In this model, I want to see what could be the significant factors for affecting the frequency of the participants taking an Uber/Lyft. The data for this variable is from question 11 of the survey.

First, I replaced the NA value with 0, and I noticed there was a 100 in the responses, which is very unlikely, so I removed it as an outlier to make sure it wouldn’t affect the model.

```
survey <- survey %>%
  rename(UberRide = In.the.past.7.days..how.many.rides.did.you.hail.using.an.Uber.Lyf
t.app..) %>%
  mutate(UberRide = replace_na(UberRide, 0)) %>%
  filter(UberRide < 100)
```

Next, I plot the data to see how it is distributed.

```
ggplot(data = survey, aes(x = UberRide)) +
  geom_bar(fill = "#4B2E83", colour = "#b7a57a") +
  labs(title = 'Average Number of Uber/Lyft Rides per Week (516 samples)', y = "Frequ
ency", x = "Number of Rides")
```



As the histogram showed, the data is strongly skewed toward 0.

```
table <- cbind("Mean" = mean(survey$UberRide), "Var" = var(survey$UberRide), "Min" =
min(survey$UberRide), "Max" = max(survey$UberRide))
rownames(table) = c("UberRide")
print(xtable::xtable(table), type = "html", html.table.attributes = "border=3")
```

	Mean	Var	Min	Max
UberRide	0.71	1.88	0.00	10.00

Since the variance of the variable is larger the mean, due to overdispersion, I chose to use negative binomial model to analyze the model.

Exercise 2

The first independent variable I chose was whether the particiapnts are willing to share an Uber/Lyft with strangers. This variable is from question 13 of the survey. I chose this variable because I think that if you are not willing to share ride with strangers, you may not like sitting in stranger’s cars. Therefore, less likely to take an Uber/Lyft.

```
survey <- survey %>%
  rename(ShareUber = How.often.do.you...share.an.Uber.Lyft.ride.with.someone.you.did.
not.know.) %>%
  mutate(ShareUber = ifelse(ShareUber %in% c("Always", "Often", "Sometimes", "Rarely"
), "Yes", "No"))
```

```
survey1 <- survey %>% filter(ShareUber == "Yes")
survey2 <- survey %>% filter(ShareUber == "No")



obs <- c(nrow(survey1), nrow(survey2), nrow(survey))
mean <- c(mean(survey1$UberRide), mean(survey2$UberRide), mean(survey$UberRide))
var <- c(var(survey1$UberRide), var(survey2$UberRide), var(survey$UberRide))
min <- c(min(survey1$UberRide), min(survey2$UberRide), min(survey$UberRide))
max <- c(max(survey1$UberRide), max(survey2$UberRide), max(survey$UberRide))

table2 <- cbind("Obs" = obs, "Mean" = mean, "Var." = var, "Min" = min, "Max" = max)
rownames(table2) <- c("Yes", "No", "Total")

print(xtable::xtable(table2), type = "html", html.table.attributes = "border=3")
```

	Obs	Mean	Var.	Min	Max
Yes	184.00	1.45	2.53	0.00	10.00
No	332.00	0.30	1.05	0.00	10.00
Total	516.00	0.71	1.88	0.00	10.00

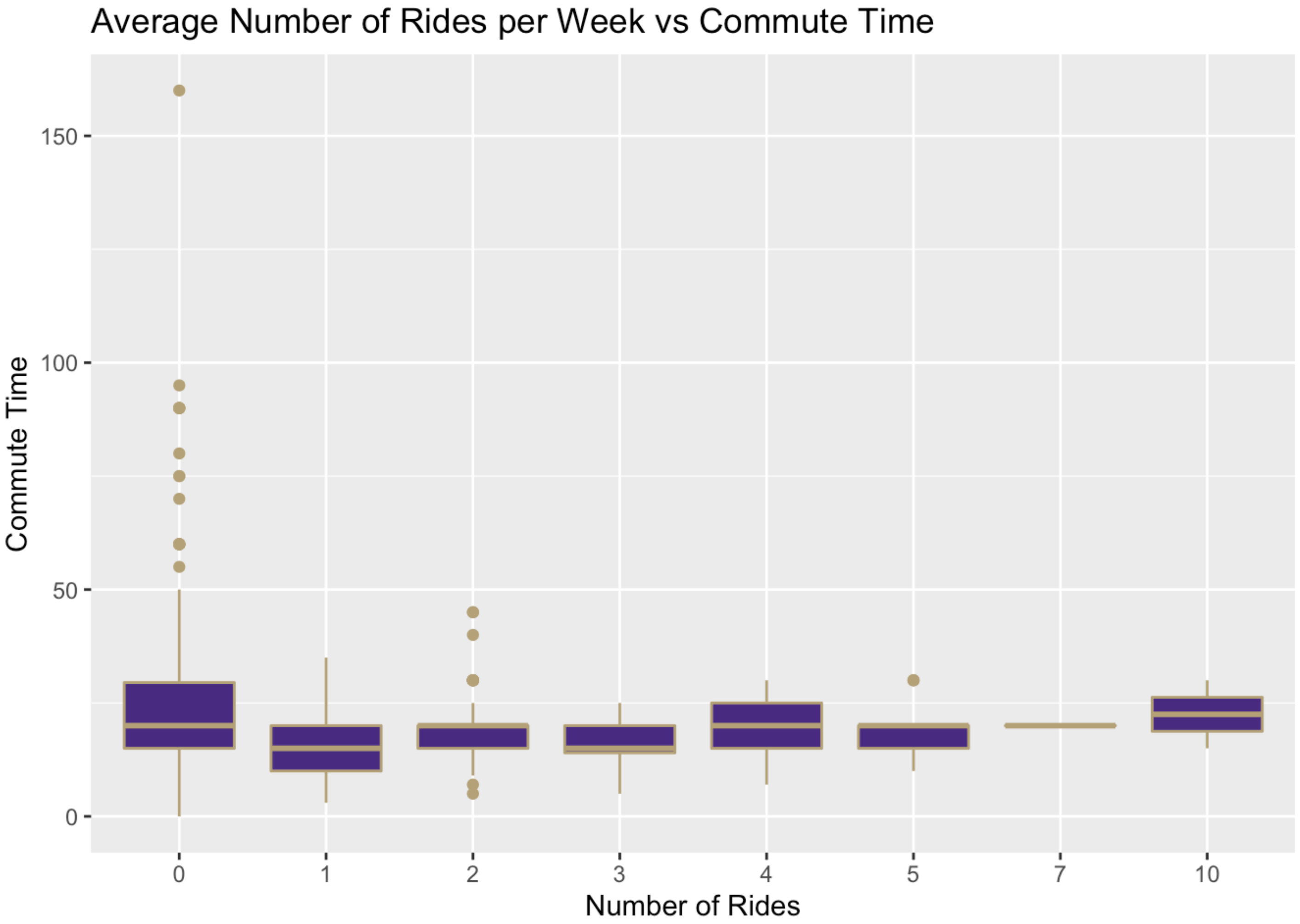
From the table above, it seems if the particapants are willing to share their Uber/Lyft, they are more likely to take more Uber/Lyft on average. Therefore, this is definitely a variable worth exploring in this model.

The second independent variable was the commute time of the participants. I chose this variable because that I’m curious to see what type of participants are more likely to take an Uber/Lyft. Is it the ones who live farther, or otherwise.

```
survey <- survey %>%
  rename(CommuteTime = On.average..how.many.minutes.does.it.take.you.to.get.to.the.U.
.Washington.from.your.home.) %>%
  filter(CommuteTime >= 0)
```

I draw a box plot to describe this variable to see if there's any initial trends I can tell.

```
survey2 <- survey
survey2$UberRide <- factor(survey2$UberRide, levels = c(0, 1, 2, 3, 4, 5, 6, 7, 8, 9,
10), labels = c("0", "1", "2", "3", "4", "5", "6", "7", "8", "9", "10"))
ggplot(data = survey2, aes(y = CommuteTime, x = UberRide)) +
  geom_boxplot(fill = "#4B2E83", color = "#b7a57a") +
  labs(title = "Average Number of Rides per Week vs Commute Time", x = "Number of Rid
es", y = "Commute Time")
```



From the boxplot, it seemed that people with longer commute time (30+ minutes) tend to not use Uber/Lyft.

The third independent variable was the primary means of transportation of the particiants. This variable is from question 1 of the survey. I chose this variable because I'm interested to see if there is a group of commuters wouldutilize Uber/Lyft more.

```
survey <- survey %>%
  rename(TransMode = What.is.your.primary.means.of.transportation.to.and.from.the.U..Washington.) %>%
  mutate(TransMode = case_when(TransMode == 'Bus' ~ 1, TransMode == 'Drive' ~ 2, TransMode == 'Walk' ~ 3, TransMode == 'Light rail' ~ 4)) %>%
  mutate(TransMode = replace_na(TransMode, 5))

survey$TransMode <- factor(survey$TransMode, levels = c(1, 2, 3, 4, 5), labels = c("Bus", "Drive", "Walk", "Light rail", "Others"))
survey$TransMode <- relevel(survey$TransMode, ref = "Others")
```

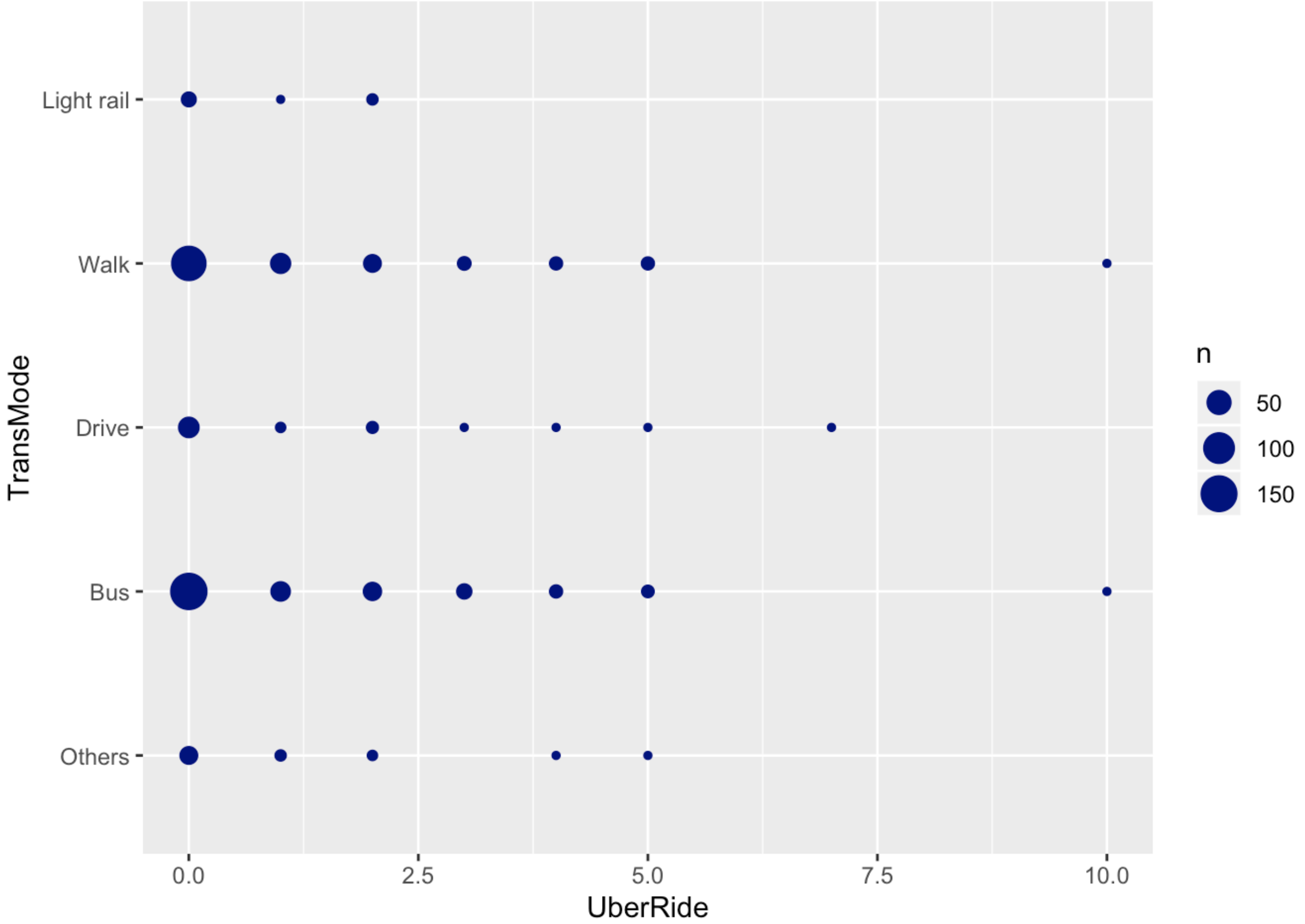
Next, I used a table to summarize this variable and its relationship to the dependent variable.

```
survey1 <- survey %>% filter(TransMode == "Bus")
survey2 <- survey %>% filter(TransMode == "Drive")
survey3 <- survey %>% filter(TransMode == "Walk")
survey4 <- survey %>% filter(TransMode == "Light rail")
survey5 <- survey %>% filter(TransMode == "Others")

obs <- c(nrow(survey1), nrow(survey2), nrow(survey3), nrow(survey4), nrow(survey5), nrow(survey))
mean <- c(mean(survey1$UberRide), mean(survey2$UberRide), mean(survey3$UberRide), mean(survey4$UberRide), mean(survey5$UberRide), mean(survey$UberRide))
var <- c(var(survey1$UberRide), var(survey2$UberRide), var(survey3$UberRide), var(survey4$UberRide), var(survey5$UberRide), var(survey$UberRide))
min <- c(min(survey1$UberRide), min(survey2$UberRide), min(survey3$UberRide), min(survey4$UberRide), min(survey5$UberRide), min(survey$UberRide))
max <- c(max(survey1$UberRide), max(survey2$UberRide), max(survey3$UberRide), max(survey4$UberRide), max(survey5$UberRide), max(survey$UberRide))

table3 <- cbind("Obs" = obs, "Mean" = mean, "Var." = var, "Min" = min, "Max" = max)
rownames(table3) <- c("Bus", "Drive", "Walk", "Light rail", "Others", "Total")


print(xtable::xtable(table3), type = "html", html.table.attributes = "border=3")
```

	Obs	Mean	Var.	Min	Max
Bus	225.00	0.71	1.86	0.00	10.00
Drive	41.00	0.71	2.41	0.00	7.00
Walk	203.00	0.75	1.98	0.00	10.00
Light rail	14.00	0.50	0.73	0.00	2.00
Others	26.00	0.62	1.69	0.00	5.00
Total	509.00	0.72	1.90	0.00	10.00

The variances between each group are consistently larger than the mean, but there’s no indications that the average number of Uber/Lyft rides per week between groups differ drastically. Next, I draw a plot to see how did the observations spreaded between each average number of rides.

```
ggplot(survey, aes(x = UberRide, y = TransMode))+
  geom_count(color = "Navy Blue")
```



From the plot above, it seems that some of the commuters who utilize buses or walk primarily take Uber/Lyft more often. However, no relationships can be observed directly.

Exercise 3 and 4

First, I inputed all independent variables into the model.

```
model_nb <- glm.nb(UberRide ~ CommuteTime + TransMode + ShareUber, data = survey)
#summary(model_nb)

table4 <- coef(summary(model_nb))
table4 <- round(cbind(table4, "Estimate" = exp(coef(model_nb)), exp(confint(model_nb)
)),3)
```

```
## Waiting for profiling to be done...
```

```
rownames(table4) <- c("Intercept", "Commute Time", "Take Bus", "Drive", "Walk", "Ligh
t Rail", "Willing to share Uber/Lyft")
print(xtable::xtable(table4), type = "html", html.table.attributes = "border = 2")
```

	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	2.5 %	97.5 %
Intercept	-1.05	0.45	-2.33	0.02	0.35	0.15	0.85
Commute Time	-0.02	0.01	-2.57	0.01	0.98	0.96	0.99
Take Bus	0.42	0.43	0.99	0.32	1.53	0.67	3.44
Drive	0.39	0.52	0.75	0.45	1.47	0.56	3.89
Walk	0.15	0.43	0.34	0.73	1.16	0.51	2.61
Light Rail	0.10	0.71	0.14	0.89	1.10	0.28	4.35
Willing to share Uber/Lyft	1.62	0.17	9.55	0.00	5.04	3.63	7.05

From the result above, I decided to remove the insignificant variable “TransMode”, and re-run the model.

```
model_nb2 <- glm.nb(UberRide ~ CommuteTime + ShareUber, data = survey)
#summary(model_nb)

table5 <- coef(summary(model_nb2))
table5 <- round(cbind(table5, "Estimate" = exp(coef(model_nb2)), exp(confint(model_nb
2))),3)
```

```
## Waiting for profiling to be done...
```

```
rownames(table5) <- c("Intercept", "Commute Time", "Willing to share Uber/Lyft")
print(xtable::xtable(table5), type = "html", html.table.attributes = "border = 2")
```

	Estimate	Std. Error	z value	Pr(>|z|)	Estimate	2.5 %	97.5 %
Intercept	-0.81	0.20	-3.97	0.00	0.45	0.30	0.67
Commute Time	-0.02	0.01	-2.46	0.01	0.98	0.96	1.00
Willing to share Uber/Lyft	1.58	0.17	9.39	0.00	4.86	3.51	6.77

The final model shows that both of the remaining variables are significant. It showed that for each one-unit increases in commute time, the expected log count of average number of Uber/Lyft rides decreases by 0.02. This result match the original observation that participants who have longer commute time tend to take less Uber/Lyft ride. Moreover, The expected log count is 1.58 higher for the participants who are willing to share an Uber/Lyft ride comparing to those who aren't. This also match what I saw earlier that average number of Uber/Lyft rides is higher for the willing-to-share case. The confidence interval also confirmed the result above.

```
ini_conv <- c(model_nb$null.deviance, model_nb2$null.deviance)
fin_conv <- c(model_nb$deviance, model_nb2$deviance)
aic <- c(model_nb$aic, model_nb2$aic)
df <- c(model_nb$df.residual, model_nb2$df.residual)

table6 <- cbind("Initial Convergence" = ini_conv, "Final Convergence" = fin_conv, "AI
C" = aic, "df" = df)
rownames(table6) <- c("Model w/ TransMode", "Model w/o TransMode")
print(xtable::xtable(table6), type = "html", html.table.attribute = "border = 2")
```

	Initial Convergence	Final Convergence	AIC	df
Model w/ TransMode	492.01	382.37	1062.35	502.00
Model w/o TransMode	488.60	382.75	1057.27	506.00

Lastly, I wanted to compare this model with others to make sure I select the better approach, so I ran the Poisson model.

```
model_pois <- glm(UberRide ~ CommuteTime + ShareUber, data = survey, family = poisson
)

table7 <- cbind("Negative Binomial" = c(model_nb2$aic, logLik(model_nb2)), "Poisson"
= c(model_pois$aic, logLik(model_pois)))
rownames(table7) <- c("AIC", "LogLikelihood")

print(xtable::xtable(table7), type = "html", html.table.attribute = "border = 1")
```

	Negative Binomial	Poisson
AIC	1057.27	1187.76
LogLikelihood	-524.63	-590.88

```
table8 <- cbind("Ratio of Log Likelihoods" = 2*(logLik(model_nb2) - logLik(model_pois
)),
                "Test for Significance" = pchisq(2*(logLik(model_nb2) - logLik(model_
pois)), df = 1, lower.tail = FALSE))

print(xtable::xtable(table8), type = "html", html.table.attribute = "border = 1")
```

	Ratio of Log Likelihoods	Test for Significance
1	132.50	0.00

With the lower AIC, I can say that negative binomial model is the better approach. However, with the slight difference in the ratio of the log likelihoods, the negative binomial model is only slightly better than the poisson model.