Project 2 Modeling, Testing, and Predicting: Health Related Data from Four Low-income Communities in the Valley of Atlixco, Mexico

#Veronica Remmert, VMR756

0.Introduction

The data from 'healthmx' was collected during Summer 2019 as part of UT Austin's President's Award for Global Learning. Three rural pueblos and one urban colonia popular surrounding Atlixco, Puebla, Mexico were selected to conduct a comprehensive community health needs assessment in order to understand and address health inequities that are present in these low-income communities.The dataset used for this assignment was created from the semi-quantitative household survey data, is heavily redacted, and much smaller than the actual dataset, which had over 200 questions. This research was approved by the IRB and was conducted under IRB guidelines. The data analysis for this project is exploratory and will not be used for publication or policy decisions. The 'number' variable relates to the relative timing of the survey, where 1 was the first household surveyed, and the variable 'rID' is the unique ID for the survey. The 'community' variable is the community where the survey was recorded and the 'housholdstucture' variable is the classification of the household such as nuclear or extended. The variable 'hp_visit' is the number of times someone in the houshold visited the household in the last three months from when the survey was taken. The 'transportation' variable is the type of transportation used to get to healthcare appointment and 'hp_time' is the amount of time it took to get to the healthcare appointment. The variable 'hp_payment' is the general amount paid for the healthcare services, where the options are none, part, or all. The 'payment_difficulty' is how difficult it was to absorb the costs related to healthcare in the household and 'confidence' was the amount of confidence in the provider the houshold had. The 'gender' variable is corresponds to the identified gender of the respondent, and the 'urban' variable was whether or not the community was classified as rural or urban. 'Hhsize' refers to the number of people living in the house and 'age' is the age of the respondent. The 'public_hp' is a variable to describe the primary healthcare utilization type of the household, where 1 is a public provider and 0 is private healthcare provider Overall, there are 242 observations or surveys in this dataset.

```
library(dplyr)
library(ggplot2)
library(tidyverse)
library(lmtest)
library(sandwich)
library(plotROC)
library(readr)
library(pROC)
healthmx = read.csv("healthmx.csv")
glimpse(healthmx)
```

```
## Observations: 242
## Variables: 18
## $ number            <int> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15...
## $ rID               <fct> R_cQGUwNi7fR2c4lI, R_4kPd7IAqbMnbwbQ, R_dmoHEjkje...
## $ community         <fct> San Fco Xochiteopan, San Fco Xochiteopan, San Fco...
## $ housholdstructure <fct> Nuclear, Extendida, Nuclear, Extendida, Nuclear y...
## $ hp_visit          <fct> 1-2 veces, 3-4 veces, 3-4 veces, 1-2 veces, Ningu...
```

```
## $ transportation    <fct> Caminando, Carro/caminioneta (propio), Carro/cami...
## $ hp_difficulty     <fct> Muy difícil, Un poco difícil, Un poco difícil, Mu...
## $ hp_payment        <fct> "Una parte", "Todo", "Todo", "Todo", "Todo", "Tod...
## $ pay_difficulty    <fct> Un poco difícil, Muy difícil, Muy difícil, Muy di...
## $ confidence        <fct> Mucha confianza, Mucha confianza, Mucha confianza...
## $ gender            <fct> F, F, M, F, F, F, F, F, M, M, F, M, M, M, M, M, F...
## $ urban             <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0...
## $ hp_time           <int> 60, 60, 30, 60, 30, 15, 2, 10, 15, 10, 10, 10, 10...
## $ hhsize            <int> 5, 5, 3, 6, 3, 3, 4, 2, 3, 3, 3, 3, 2, 4, 7, 3, 1...
## $ age               <int> 45, 64, 34, 59, 30, 57, 45, 59, 31, 78, 38, 60, 7...
## $ burden            <int> 1, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0...
## $ public_hp         <int> NA, 1, 0, 0, 0, 1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 0, ...
## $ hcu               <fct> NA, public, private, private, private, public, pu...
```

```
healthmx$female <- NULL
```

1. MANOVA and ANOVA Testing

```
# MANOVA
healthmx1 <- healthmx %>% drop_na(hhsize, age, hp_time, hp_payment)
man <- manova(cbind(hhsize, age, hp_time) ~ hp_payment, data = healthmx1)
summary(man)
```

```
##             Df   Pillai approx F num Df den Df  Pr(>F)
## hp_payment   2 0.072046   2.8525      6    458 0.00975 **
## Residuals  230
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Homogeneity of (CO)Variances Assumption
covmats <- healthmx1 %>% group_by(hp_payment) %>% do(covs = cov(.[13:15]))
for (i in 1:3) {
    print(as.character(covmats$hp_payment[i]))
    print(covmats$covs[i])
}
```

```
## [1] "No, nada"
## [[1]]
##            hp_time     hhsize      age
## hp_time 1255.28831  10.366234 154.0130
## hhsize    10.36623   3.690584 -10.7289
## age      154.01299 -10.728896 319.4282
##
## [1] "Todo"
## [[1]]
##             hp_time     hhsize       age
## hp_time 26711.38730 -37.419962  15.56620
## hhsize    -37.41996   3.430393 -13.37143
## age        15.56620 -13.371433 267.98732
##
## [1] "Una parte"
## [[1]]
##            hp_time    hhsize       age
## hp_time 669.317358 -4.674459  51.22735
## hhsize   -4.674459  3.423729  -7.51315
```

```
## age      51.227352 -7.513150 237.02805
```

```
# ANOVA
summary(aov(hhsize ~ hp_payment, data = healthmx1))
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## hp_payment    2    2.1   1.062   0.304  0.738
## Residuals   230  802.9   3.491
```

```
summary(aov(age ~ hp_payment, data = healthmx1))
```

```
##              Df Sum Sq Mean Sq F value  Pr(>F)
## hp_payment    2   3198  1598.9   5.868 0.00327 **
## Residuals   230  62671   272.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
summary(aov(hp_time ~ hp_payment, data = healthmx1))
```

```
##              Df  Sum Sq Mean Sq F value Pr(>F)
## hp_payment    2   84425   42213   3.003 0.0516 .
## Residuals   230 3233094   14057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
healthmx %>% count(hp_payment)
```

```
## # A tibble: 4 x 2
##   hp_payment     n
##   <fct>      <int>
## 1 No, nada      59
## 2 Todo         118
## 3 Una parte     64
## 4 <NA>           1
```

```
healthmx %>% group_by(hp_payment) %>% summarize(mean = sd(age,
    na.rm = TRUE))
```

```
## # A tibble: 4 x 2
##   hp_payment  mean
##   <fct>      <dbl>
## 1 No, nada    17.8
## 2 Todo        16.4
## 3 Una parte   15.6
## 4 <NA>         NA
```

```
healthmx %>% group_by(hp_payment) %>% summarize(mean = sd(hhsize,
    na.rm = TRUE))
```

```
## # A tibble: 4 x 2
##   hp_payment  mean
##   <fct>      <dbl>
## 1 No, nada    1.92
## 2 Todo        1.85
## 3 Una parte   1.84
## 4 <NA>       NA
```

```
healthmx %>% group_by(hp_payment) %>% summarize(mean = sd(hp_time,
    na.rm = TRUE))
```

```
## # A tibble: 4 x 2
##   hp_payment  mean
##   <fct>      <dbl>
## 1 No, nada    34.7
## 2 Todo       163.
## 3 Una parte   25.4
## 4 <NA>        NA
```

```
# Post- Hoc T Tests
pairwise.t.test(healthmx1$hhsize, healthmx1$hp_payment, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  healthmx1$hhsize and healthmx1$hp_payment
##
##           No, nada Todo
## Todo      0.47     -
## Una parte 0.85     0.61
##
## P value adjustment method: none
```

```
pairwise.t.test(healthmx1$age, healthmx1$hp_payment, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  healthmx1$age and healthmx1$hp_payment
##
##           No, nada Todo
## Todo      0.001    -
## Una parte 0.202    0.061
##
## P value adjustment method: none
```

```
pairwise.t.test(healthmx1$hp_time, healthmx1$hp_payment, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  healthmx1$hp_time and healthmx1$hp_payment
##
##           No, nada Todo
## Todo      0.051    -
## Una parte 0.980    0.044
##
## P value adjustment method: none
```

```
# Probability of at least one Type 1 Error
type1 <- 1 - ((1 - 0.05)^13)
type1
```

```
## [1] 0.4866579
```

```
# Bonferroni Correction
bonferroni <- 0.05/13
bonferroni
```

```
## [1] 0.003846154
```

When performing a MANOVA test, we can see that there is a mean difference in healthcare payment level across household size, age, and time spent to get to healthcare services. Based on the significant p-value of 0.00975 of the MANOVA, there is a mean difference across healthcare payment levels with size, age, and time. With the MANOVA we are unsure as to which numeric variables show the difference, and therefore an ANOVA was performed. As for the assumptions of the MANOVA, the data did include random samples and independent observations, the multivariate normality of DV assumption was met because there were more than 25 samples in each payment group, and there appears to be linear relationship among DVs. However, there appears to be a lack of homogeneity of (co)variances according to the matrices, and there appears to be some univariate or multivariate outliers

When performing the three ANOVA tests, it was found the the mean difference in healthcare payment occurs for the response variables of both age of respondent and amount of time it takes to get to healthcare provider. For an ANOVA, the assumptions include random sample and independent observations, independent samples, normal distribution in each group or a large sample, and equal variance of each group. These likely have been met because the respondents were randomly selected, the sampling of each respondent was done independently, there was over 25 samples for each payment group, and the standard deviation is not more than 2x bigger in one group versus another, except for the case of 'hp_time'.

Using the pairwise.t.test function, it was found that there is a difference in 'No, nada' and 'Todo' when age is used in the pairwise.t.test. It was also found that there is a difference between 'Todo' and 'Una parte' when time to get to the healthcare provider was used.

There was 1 MANOVA, 3 ANOVAs, and 9 pairwise t-tests that were performed. The probability of at least one type I error is 0.4866579 if unadjusted. The boneferroni adjusted significance level that should be used is to keep the overall type I error rate at 0.05 is 0.003846154. After using this significance level, we cannot state that any of the payment groups vary by 'hp_time','age', or 'hhsize'.

2. Randomization Test

```
# DropNA
healthmx_nona <- healthmx %>% drop_na(gender) %>% drop_na(age)
# Mean Difference
healthmx_nona %>% group_by(gender) %>% summarize(means = mean(age)) %>%
    summarize(`mean_diff:` = diff(means))
```

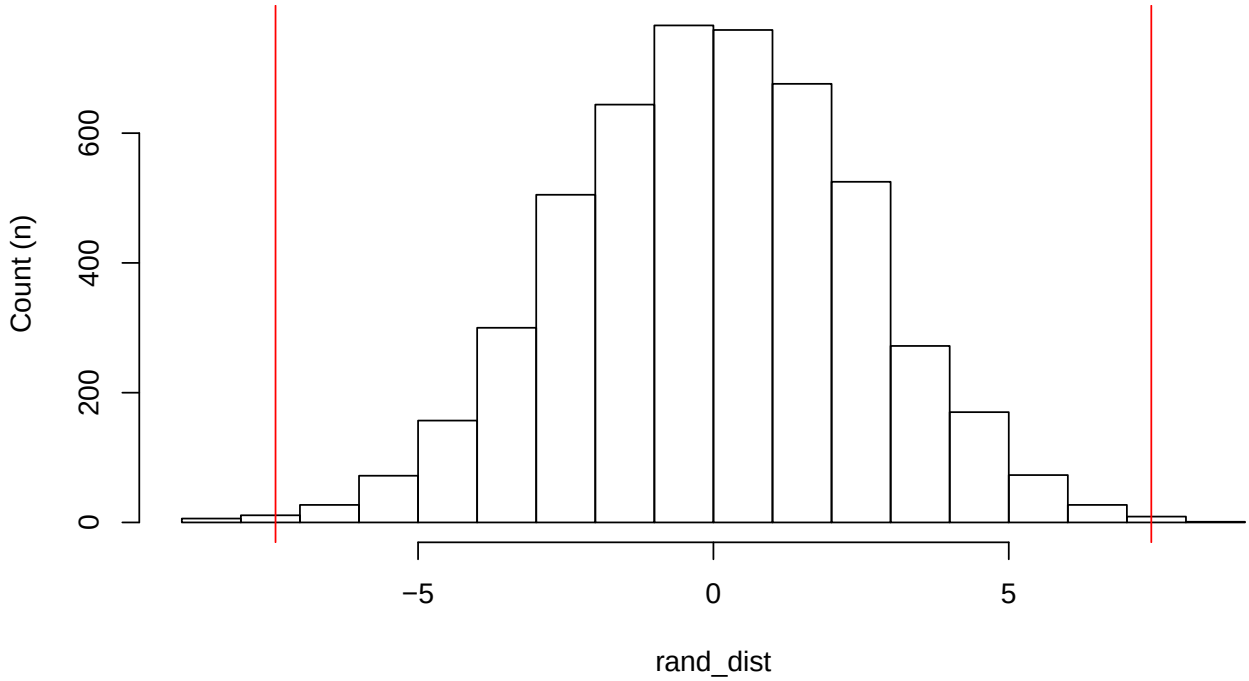
```
## # A tibble: 1 x 1
##   `mean_diff:`
##          <dbl>
## 1         7.41
```

```
# Randomization Test
set.seed(34)
rand_dist <- vector()
for (i in 1:5000) {
    new <- data.frame(age = sample(healthmx_nona$age), gender = healthmx_nona$gender)
    rand_dist[i] <- mean(new[new$gender == "F", ]$age) - mean(new[new$gender ==
        "M", ]$age)
}
mean(rand_dist > 7.413414 | rand_dist < -7.413414)
```

```
## [1] 0.0028
```

```
# Plot of Null Distribution and Test Statistic
{
    hist(rand_dist, main = "", ylab = "Count (n)")
```

```
    abline(v = -7.413414, col = "red")
    abline(v = 7.413414, col = "red")
}
```




```
# Comparison to t-test
t.test(data = healthmx_nona, age ~ gender)
```

```
##
##  Welch Two Sample t-test
##
## data:  age by gender
## t = -3.0901, df = 102.88, p-value = 0.002574
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -12.171436  -2.655392
## sample estimates:
## mean in group F mean in group M
##        48.74176        56.15517
```

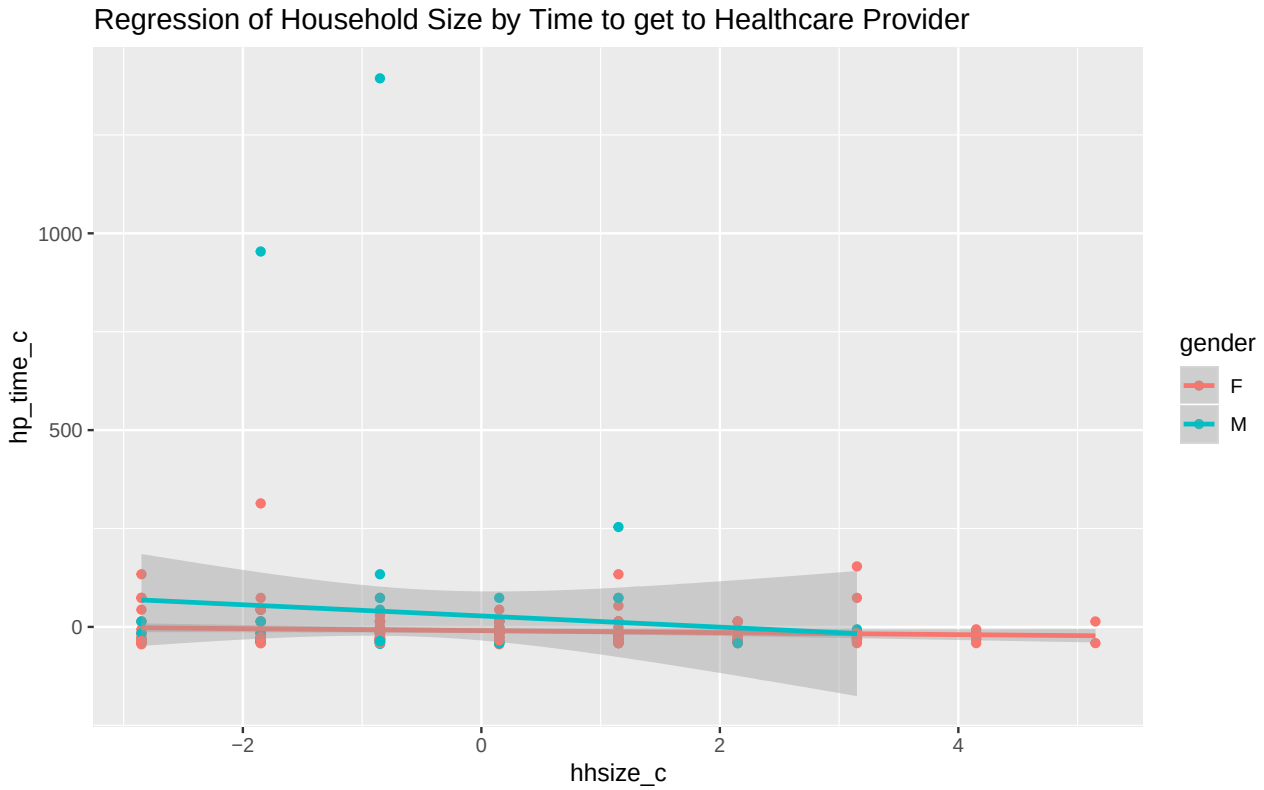
The null hypothesis of the randomization test is that the mean age is the same between male and female respondents. The alternative hypothesis is that the mean age is different between male and female respondents. The randomization test had a p-value of 0.0022 and therefore the results are significant and we can reject the null hypothesis that the mean age is the same between male and female respondents. WHen a two-sample t-test was run, the p-value was again significant at 0.002574, revealing that both the randiomization test and two-sample t-test arrived at similar conclusions. We reject te null hypothesis that the mean age is the same between male and female respondents.The plot was created to visualize the null distribution, where the red lines represent the positive and negative test statistic, which was 7.413414.

3. Linear Regression Model

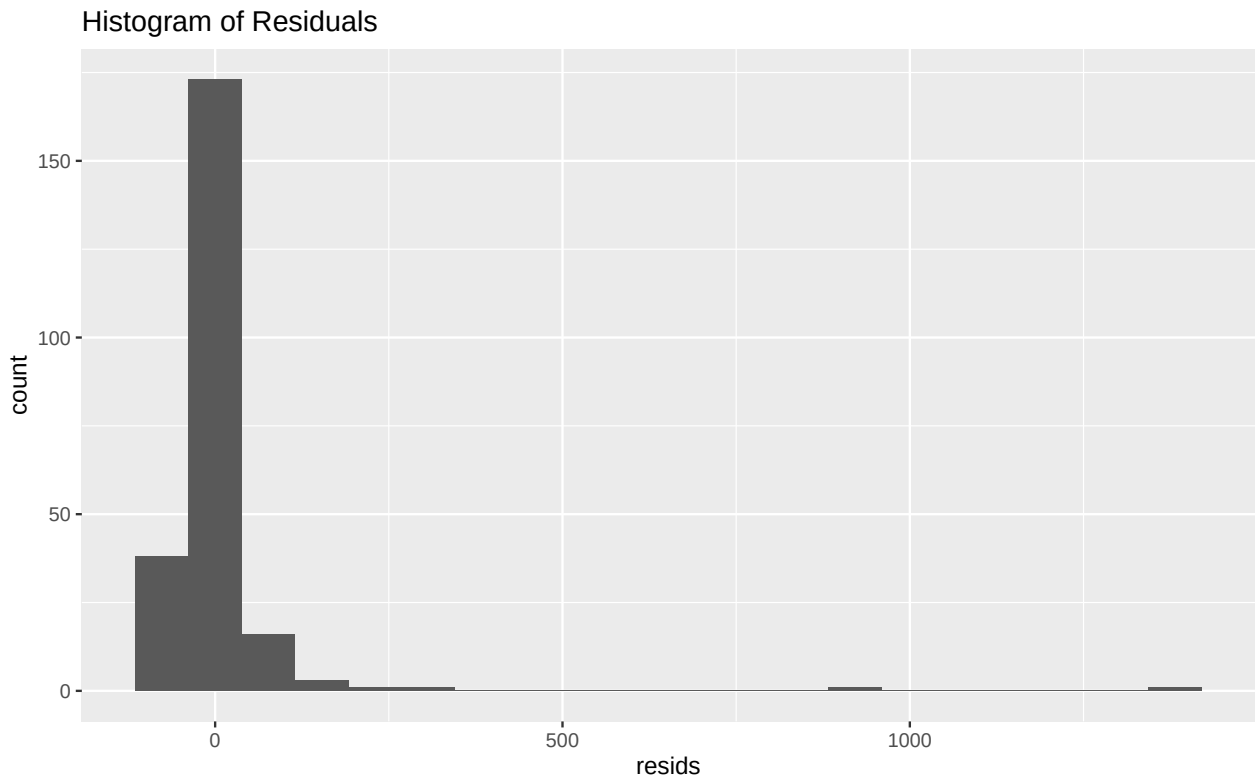
```r
healthmx$hhsize_c <- healthmx$hhsize - mean(healthmx$hhsize,
    na.rm = T)
healthmx$hp_time_c <- healthmx$hp_time - mean(healthmx$hp_time,
    na.rm = T)
fitlr <- lm(hp_time_c ~ gender * hhsize_c, data = healthmx)
summary(fitlr)
```

```
##
## Call:
## lm(formula = hp_time_c ~ gender * hhsize_c, data = healthmx)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -104.63  -28.86  -16.33    3.81 1353.82
##
## Coefficients:
##                  Estimate Std. Error t value Pr(>|t|)
## (Intercept)        -9.483      8.921  -1.063   0.2889
## genderM            37.379     18.437   2.027   0.0438 *
## hhsize_c           -2.535      4.561  -0.556   0.5789
## genderM:hhsize_c  -11.687     11.611  -1.007   0.3152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 118.1 on 230 degrees of freedom
##   (8 observations deleted due to missingness)
## Multiple R-squared:  0.03385,    Adjusted R-squared:  0.02125
## F-statistic: 2.686 on 3 and 230 DF,  p-value: 0.04731
```

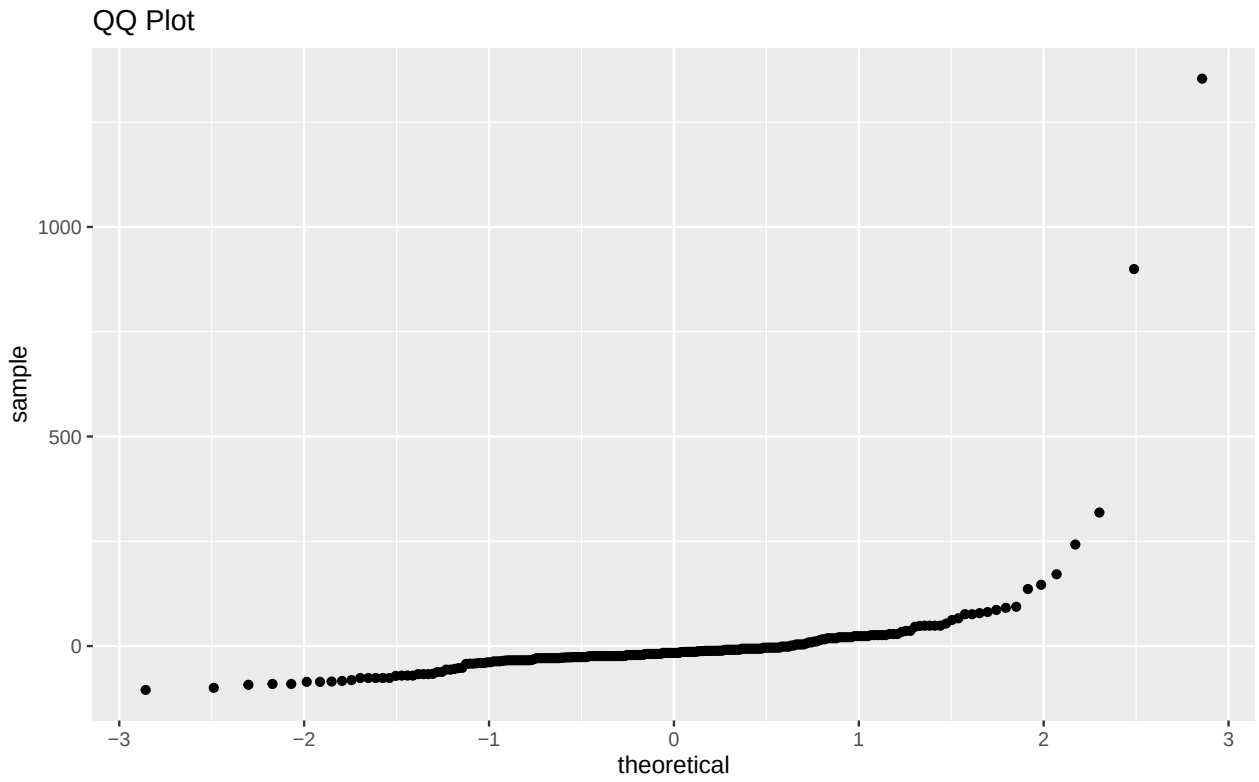
```r
# Regression
fitlr %>% ggplot(aes(hhsize_c, hp_time_c, color = gender)) +
    geom_point() + geom_smooth(method = "lm") + ggtitle("Regression of Household Size by Time to get to
```




```
# Normality
resids <- fitlr$residuals
fitvals <- fitlr$fitted.values
ggplot() + geom_histogram(aes(resids), bins = 20) + ggtitle("Histogram of Residuals")
```



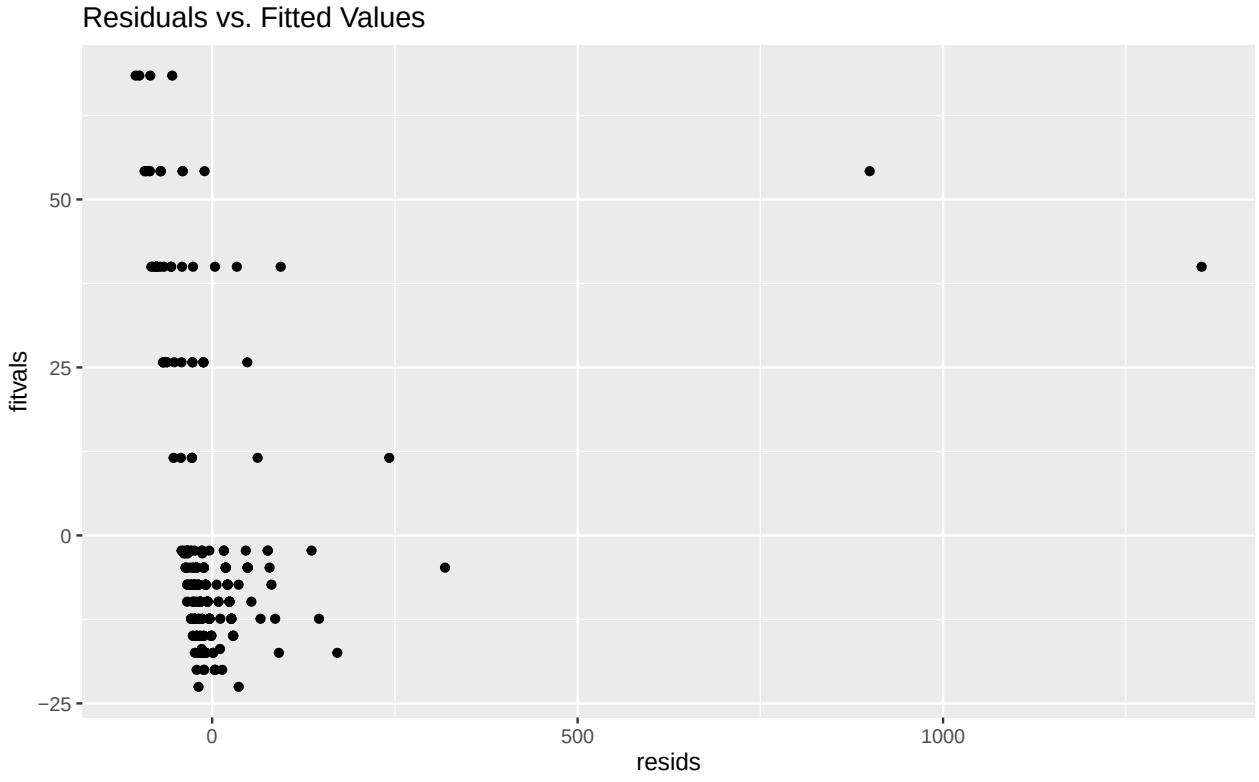

```
ggplot() + geom_qq(aes(sample = resids)) + geom_qq() + ggtitle("QQ Plot")
```



```
shapiro.test(resids)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resids
## W = 0.33953, p-value < 2.2e-16
```

```
# Linearity
ggplot() + geom_point(aes(resids, fitvals)) + ggtitle("Residuals vs. Fitted Values")
```




```
# Homoskedsaticity
bptest(fitlr)  #Reject the null hypothesis that the data is homoskedastic
```

```
##
##  studentized Breusch-Pagan test
##
## data:  fitlr
## BP = 7.9138, df = 3, p-value = 0.04783
```

```
# Redo the Regression Using Heteroskedasticity Robust
# Standard Errors
coeftest(fitlr, vcov = vcovHC(fitlr))
```

```
##
## t test of coefficients:
##
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -9.4826     3.2490 -2.9186 0.003865 **
## genderM         37.3788    26.4348  1.4140 0.158715
```

```
## hhsize_c          -2.5348     1.7971 -1.4105 0.159738
## genderM:hhsize_c -11.6873    13.5863 -0.8602 0.390559
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Variation Explained
summary(fitlr)$r.sq
```

```
## [1] 0.03385242
```

For people with an average household size, male respondents have a higher average time to get to healthcare provider that is 37.379 minutes greater than the centered mean time for female respondents. When controlling for gender, those with an average household size take 2.535 minutes less to get to healthcare provider. The slope of houshold size on time to get to healthcare provider is 11.687 less for males as compared to females.

The plot is the regression of centered household size and time to get to healthcare provider, colored by gender of the respondent.

The independent observations and random sampling assumption was met due to data collection design. To check normality, the Shapiro-Wilk normality test was run. Because the p-value of this test was very small, we reject the null hypothesis that the data is normally distributed and therefore the normality assumption was not met. This can also seen visually in the histogram, where the data does not appear to be normally distributed. To test for homoskedasticity, the Breuch-Pagan test was run and the p-value was significant, as a result we reject the null hypothesis that the data is homoskedastic and therefore the homoskedastic assumption was not met. Lastly, the linearity assumption appears to be met as seen by the linear relationship between the predictor and response variable.

The regression was recomputed with robust standard errors via coeftest(fitlr,vcov=vcovHC(fitlr)). After this was run, none of the coefficient variables (besides the intercept) were significant. Before this was run, the genderM variable coefficient was significant, but with robust standard errors it is no long signficant. When using the robust standard error, the standard errors increased in the genderM variable, going from 18.347 to 26.4348 with robust errors. The standard error interaction between genderM and household size also increased, where it was 11.611 previously and was 13.5863 with robust standard errors. There was little change in the coefficient estimates themselves between the linear regression and the linear regression with robust standard errors.

To determine the proportion of variation in the outcome that the model explains, the R squared value of the model was used. Based on the r-squared, the model explains 0.03385 proportion of variation, or 3.385%.

4.Linear Regression Model with Standard Errors

```
healthmx$hhsize_c <- healthmx$hhsize - mean(healthmx$hhsize,
    na.rm = T)
healthmx$hp_time_c <- healthmx$hp_time - mean(healthmx$hp_time,
    na.rm = T)

set.seed(1)
samp_distn <- replicate(5000, {
    boot_dat <- sample_frac(healthmx, replace = T)
    fit <- lm(hp_time_c ~ gender * hhsize_c, data = boot_dat)
    coef(fit)
})
samp_distn %>% t %>% as.data.frame %>% summarize_all(sd)
```

```
##   (Intercept)  genderM hhsize_c genderM:hhsize_c
## 1    3.200947 26.44833 1.780814         13.83077
```

The standard errors calculated with the boostrapped standard errors are very similar to the the standard errors calculated in the linear regression, but the errors are still smaller for bootstrapping. For the linear regression, the intercept SE was 3.2490, genderM SE was 26.4348, household size SE was 1.7971, and the interaction between household size and genderM SE was 13.5863. All of these values are larger than the SE calculated with the bootstrapping, where the Intercept SE was 3.20462, the genderM SE was 25.94308, household size SE was 1.790101, and the interaction between household size and genderM SE was 13.60298. This means that the smaller SE values from the bootstrapping have smaller p-values except for the interaction between household size and genderM. The standard errors are overall very similar.

##5. Logisitic Regression

```
dropna <- healthmx %>% drop_na(public_hp, community, hp_payment)
model <- glm(public_hp ~ community + hp_payment, data = dropna,
    family = binomial(link = "logit"))
summary(model)
```

```
##
## Call:
## glm(formula = public_hp ~ community + hp_payment, family = binomial(link = "logit"),
##     data = dropna)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.7100  -0.7109   0.2270   0.5230   1.9831
##
## Coefficients:
##                               Estimate Std. Error z value Pr(>|z|)
## (Intercept)                     3.4894     0.8050   4.335 1.46e-05 ***
## communityColonia Flores Magon  -0.2846     0.6024  -0.472   0.6366
## communitySan Fco Xochiteopan    0.1569     0.5166   0.304   0.7614
## communitySanta Ana Coatepec    -0.5690     0.5705  -0.997   0.3185
## hp_paymentTodo                 -4.7360     0.7630  -6.207 5.39e-10 ***
## hp_paymentUna parte            -1.5691     0.8087  -1.940   0.0524 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 321.50  on 233  degrees of freedom
## Residual deviance: 184.47  on 228  degrees of freedom
## AIC: 196.47
##
## Number of Fisher Scoring iterations: 6
```

```
coeftest(model)
```

```
##
## z test of coefficients:
##
##                               Estimate Std. Error z value  Pr(>|z|)
## (Intercept)                    3.48937    0.80498  4.3347 1.459e-05 ***
## communityColonia Flores Magon -0.28460    0.60236 -0.4725   0.63659
## communitySan Fco Xochiteopan   0.15686    0.51655  0.3037   0.76138
## communitySanta Ana Coatepec   -0.56904    0.57048 -0.9975   0.31854
## hp_paymentTodo                -4.73598    0.76298 -6.2072 5.393e-10 ***
## hp_paymentUna parte           -1.56907    0.80873 -1.9402   0.05236 .
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
exp(coef(model))
```

```
##                   (Intercept) communityColonia Flores Magon
##                   32.76529979                     0.75231717
##  communitySan Fco Xochiteopan   communitySanta Ana Coatepec
##                    1.16983120                     0.56607053
##                hp_paymentTodo           hp_paymentUna parte
##                    0.00877388                     0.20823823
```

```
# Confusion Matrix
prob <- predict(model, type = "response")
pred <- ifelse(prob > 0.5, 1, 0)
table(prediction = pred, truth = dropna$public_hp) %>% addmargins
```

```
##           truth
## prediction   0   1 Sum
##        0    93  24 117
##        1    11 106 117
##        Sum 104 130 234
```

```
# Accuracy
(93 + 106)/234
```

```
## [1] 0.8504274
```

```
# Sensitivity (TPR)
106/130
```

```
## [1] 0.8153846
```

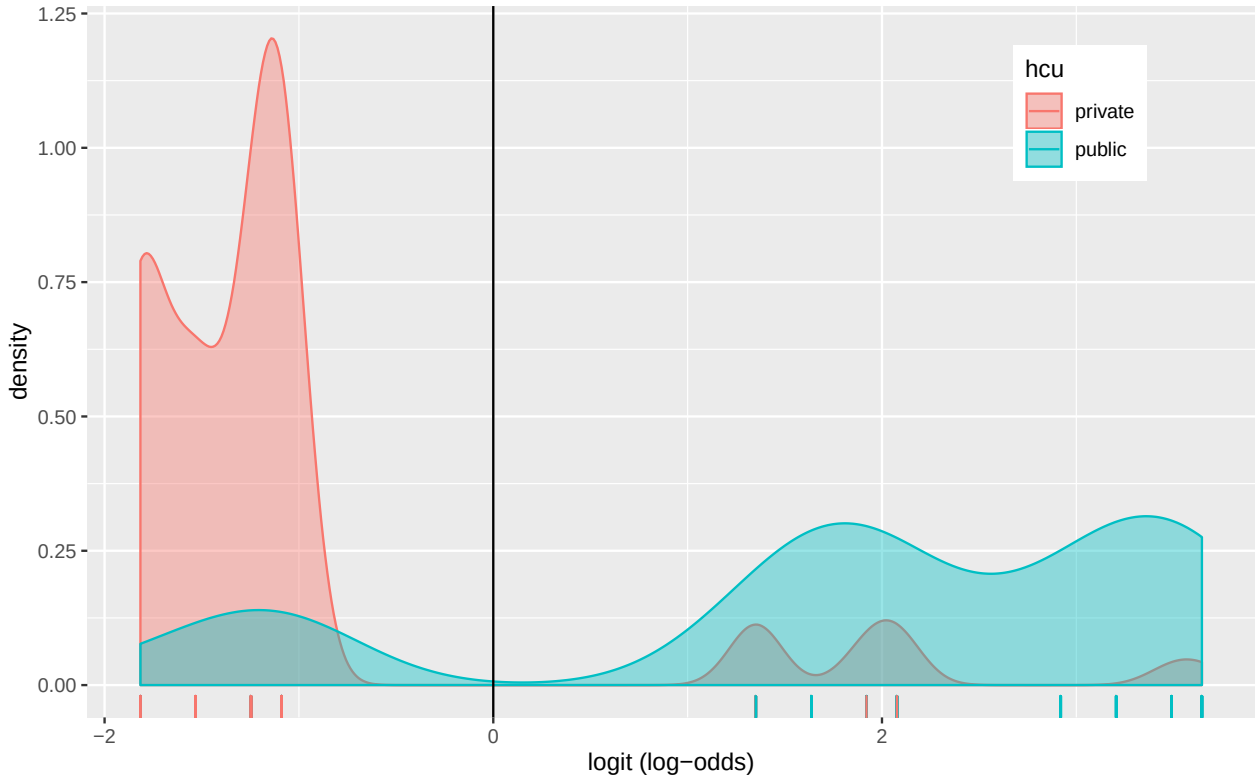
```
# Specificity (FPR)
93/104
```

```
## [1] 0.8942308
```

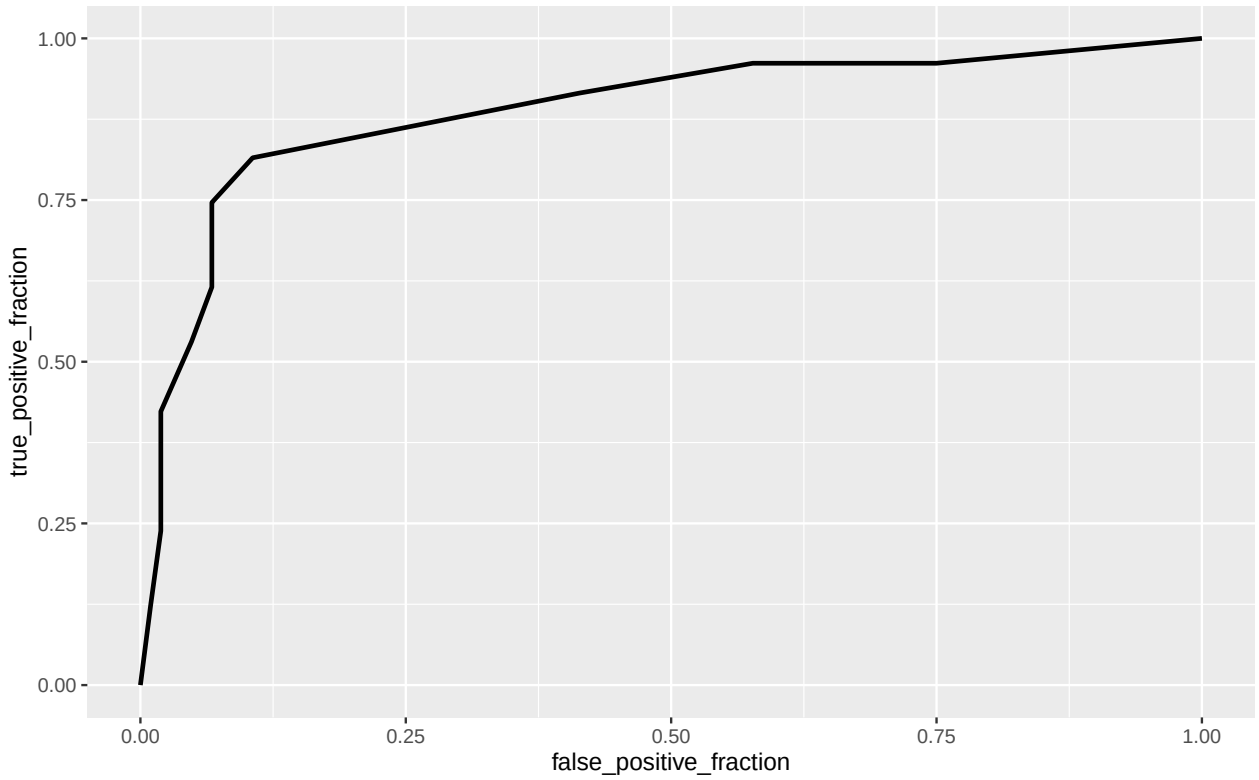
```
# Precision (PPV)
106/117
```

```
## [1] 0.9059829
```

```
# Density of Logodds
dropna$logit <- predict(model, type = "link")
dropna %>% ggplot() + geom_density(aes(logit, color = hcu, fill = hcu),
    alpha = 0.4) + theme(legend.position = c(0.85, 0.85)) + geom_vline(xintercept = 0) +
    xlab("logit (log-odds)") + geom_rug(aes(logit, color = hcu))
```




```
# ROC
tdat <- dropna %>% mutate(prob = predict(model, type = "response"),
    prediction = ifelse(prob > 0.5, 1, 0))
classify <- tdat %>% transmute(prob, prediction, truth = public_hp)
ROCplot <- ggplot(classify) + geom_roc(aes(d = truth, m = prob),
    n.cuts = 0)
ROCplot
```




```
calc_auc(ROCplot)
```

```
##   PANEL group       AUC
## 1     1    -1 0.8884246
```

```
# 10-Fold CV
set.seed(1234)
k = 10
data <- dropna[sample(nrow(dropna)), ]
folds <- cut(seq(1:nrow(dropna)), breaks = k, labels = F)
diags <- NULL
for (i in 1:k) {
    train <- data[folds != i, ]
    test <- data[folds == i, ]
    truth <- test$public_hp
    fit <- glm(public_hp ~ community + hp_payment, data = train,
        family = "binomial")
    probs <- predict(fit, newdata = test, type = "response")
    diags <- rbind(diags, class_diag(probs, truth))
}
summarize_all(diags, mean)
```

```
##         acc      sens      spec       ppv       auc
## 1 0.8507246 0.8200808 0.8892774 0.9096999 0.8761064
```

Controlling for community and paying for a part of healthcare services, paying for all of healthcare services has a significant negative impact on odds of going to a public healthcare provider. Controlling for community and paying for a part of healthcare services, the odds of going to a public healthcare provider is 0.00877388 times the odds of someone who pays for none of their healthcare services. Controlling for community and paying for all of healthcare services, paying for part of healthcare services does not have a significant negative impact on odds of going to a public healthcare provider. Controlling for payment for healthcare services, community does

not significantly impact if someone goes to public healthcare provider.

To compute the accuracy, sensitivity, specificity, and recall a confusion matrix of the model was produced. The accuracy was computed to be 0.8504274, which is a good accuracy at which the model predicts public healthcare utilization. Sensitivity is the true positive rate which is the probability of predicting public healthcare utilization for people who actually use public healthcare services and was calculated to be 0.8153846. Specificity is the true negative rate and is the proability of a person who uses private services been predicted as using public healthcare services and was calculated to be 0.8942308. The precision is the proportion of people who were classified as public healthcare utilizers who actually are and was calculated to be 0.9059829.

Based on the density plot we can visualize the misclassified area, which is gray. To the right of 0, the gray area is the proportion of false positives and to the left of 0, the gray area is the proportion of false negatives of the model prediciting public healthcare utilization

The ROC plot was generated and a AUC was calculated. THe ROC curve alllows for visualization of the trade-off between sensitivity and specificity. The AUC was calculated to be 0.8884246, which is a good AUC and means that payment level and community are a good predictor of healthcare utilization type.

After performing a 10-fold CV, the average accuracy was calculcated to be 0.85, the average sensitivity was calculated to be 0.8048194, the precision was calculcated to be 0.9015287, and the auc was calculated to be 0.8771041. These values, while slightly different, are very similar to the others calculated above. Overall, the model appears to be good at predicting healthcare utilization type.

##6. LASSO

```
library(glmnet)
set.seed(1234)
healthmx1 <- healthmx %>% na.omit(public_hp, community, housholdstructure,
    hp_visit, transportation, hp_difficulty, hp_payment, pay_difficulty,
    confidence, gender, hp_time, hhsize, age)
y <- as.matrix(healthmx1$public_hp)
x <- model.matrix(public_hp ~ community + housholdstructure +
    hp_visit + transportation + hp_difficulty + hp_payment +
    pay_difficulty + confidence + gender + hp_time + hhsize +
    age, data = healthmx1)
cv.lasso1 <- cv.glmnet(x, y, family = "binomial")
lasso1 <- glmnet(x, y, family = "binomial", lambda = cv.lasso1$lambda.1se)
coef(lasso1)
```

```
## 33 x 1 sparse Matrix of class "dgCMatrix"
##                                                      s0
## (Intercept)                                 7.202239e-01
## (Intercept)                                 .
## communityColonia Flores Magon               .
## communitySan Fco Xochiteopan                .
## communitySanta Ana Coatepec                 .
## housholdstructureExtendida y Compuesta      .
## housholdstructureNuclear                    .
## housholdstructureNuclear y compuesta        .
## housholdstructureSoltero/a                  .
## hp_visit3-4 veces                           .
## hp_visit5+ veces                            .
## hp_visitNinguna vez                         .
## transportationCaminando                     6.323514e-01
## transportationCarro (alguien me lleva)      .
## transportationCarro/caminioneta (propio)    .
## transportationCombi                         .
```

```
## transportationTransporte rentado              .
## hp_difficultyMuy difícil                      .
## hp_difficultyNada difícil                     .
## hp_difficultyUn poco difícil                  .
## hp_paymentTodo                               -1.891072e+00
## hp_paymentUna parte                           7.931505e-16
## pay_difficultyMuy difícil                     .
## pay_difficultyNada difícil                    .
## pay_difficultyNo sé o prefiero no responder   .
## pay_difficultyUn poco difícil                 .
## confidenceNada de confianza                   .
## confidenceNo sé o prefiero no responder       .
## confidencePoca confianza                      .
## genderM                                       .
## hp_time                                       4.019132e-05
## hhsize                                        .
## age                                           .
```

```
# CV Lasso
healthmx_1 <- healthmx1 %>% mutate(Caminando = ifelse(transportation ==
    "Caminando", 1, 0)) %>% mutate(parte = ifelse(hp_payment ==
    "Una parte", 1, 0)) %>% select(public_hp, Caminando, parte)

set.seed(1234)
k = 10
data <- healthmx_1 %>% sample_frac
folds <- ntile(1:nrow(data), n = 10)
diags <- NULL
for (i in 1:k) {
    train <- data[folds != i, ]
    test <- data[folds == i, ]
    truth <- test$public_hp
    fit <- glm(public_hp ~ Caminando + parte, data = train, family = "binomial")
    probs <- predict(fit, newdata = test, type = "response")
    diags <- rbind(diags, class_diag(probs, truth))
}
diags %>% summarize_all(mean)
```

```
##         acc     sens      spec       ppv       auc
## 1 0.8117647 0.650754 0.9197547 0.8538889 0.8329095
```

The 10-fold CV was performed using the LASSO and revealed a lower accuracy at 0.8 as compared to the accuracy from the logisitic regression of 0.85. This means that the LASSO model has lower accuracy than the logisitc regression model when predicting public healthcare provider utilization. THe 10-fold CV using the LASSO had an AUC of 0.8551299 whereas the AUC from the logistic regression was 0.8771041. This means that the logistic regression model is better at predicting than the model produced from LASSO.