Statistical Modeling: Effects on Medical Expenses in Vietnam

Hieu Pham

4/19/2021

1.) Introduction

```
# Import data/library function
library(readxl)
```

```
## Warning: package 'readxl' was built under R version 3.6.3
```

```
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.6.3
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
VN <- read_excel("~/Bioinformatics Data/VN_Individual.xlsx")
```

```
## New names:
## * `` -> ...1
```

```
# Rename variables in dataset
VN_Data <- VN %>%
rename(Individual = "...1", MExpense = "lnhhexp") %>%
    mutate(Insurance = case_when(insurance>0 ~ "Yes",
                              insurance<1 ~ "No"))

# View the first 6 observation of the dataset
head(VN_Data)
```

```
## # A tibble: 6 x 14
##   Individual pharvis MExpense   age sex   married  educ illness injury illdays
##        <dbl>   <dbl>    <dbl> <dbl> <chr>   <dbl> <dbl>   <dbl>  <dbl>   <dbl>
## 1          1       0     2.73  3.76 male        1     2       1      0       7
## 2          2       0     2.74  2.94 fema~       0     0       1      0       4
## 3          3       0     2.27  2.56 male        0     4       0      0       0
## 4          4       1     2.39  3.64 fema~       1     3       1      0       3
## 5          5       1     3.11  3.30 male        1     3       1      0      10
## 6          6       0     3.76  3.37 male        1     9       0      0       0
## # ... with 4 more variables: actdays <dbl>, insurance <dbl>, commune <dbl>,
## #   Insurance <chr>
```

```
# count number of observations
nrow(VN_Data)
```

```
## [1] 27765
```

This project was conducted to observe how medical expenses varied amongst individuals in Vietnam based on their circumstances and status. This dataset was acquired through vincentarebundock.The data was collected from a cross sectional study which observed a number of 27,765 participants ranging from multiple regions in Vietnam (Cameron, A.C. and P.K. Trivedi).

The dataset contains information regarding particpants' pharmacy visits, annual medical expenses, age, sex, marriage status, education level, annual number of illnesses/illness days/injuries, number of days inactive, commune population, and insurance status. The dataset did not have to be tidied as each observation (each representing an individual) had their own columns (representing the variables).

This assigment is a continuation of the first project which investigated how medical expense was tied to the insurance and education level of the particpants. However, this project in particular, will expand on the ways that medical expenses is affected by secondary and tertiary factors including annual pharmacy visits, age, and living commune size. I think that it's important to assess these aspects because it provides a broader understanding of healthcare conditions in underpriveleged countries. I hypothesize that some of these factors (pharmacy visits, age, education, and commune size) are proportional to an individual's annual medical expenses and are suggestive of their insurance status.

2.) Exploratory Data Analysis (EDA)

```
# Library
library(ggplot2)
```

```
## Warning: package 'ggplot2' was built under R version 3.6.3
```

```
library(tidyverse)
```

```
## Warning: package 'tidyverse' was built under R version 3.6.3
```

```
## -- Attaching packages ----------------------------------------------------------------------
------------------------------------------------- tidyverse 1.3.0 --
```

```
## v tibble  3.0.5     v purrr   0.3.3
## v tidyr   1.1.2     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## Warning: package 'tibble' was built under R version 3.6.3
```

```
## Warning: package 'tidyr' was built under R version 3.6.3
```

```
## Warning: package 'readr' was built under R version 3.6.3
```

```
## Warning: package 'purrr' was built under R version 3.6.3
```

```
## Warning: package 'stringr' was built under R version 3.6.3
```

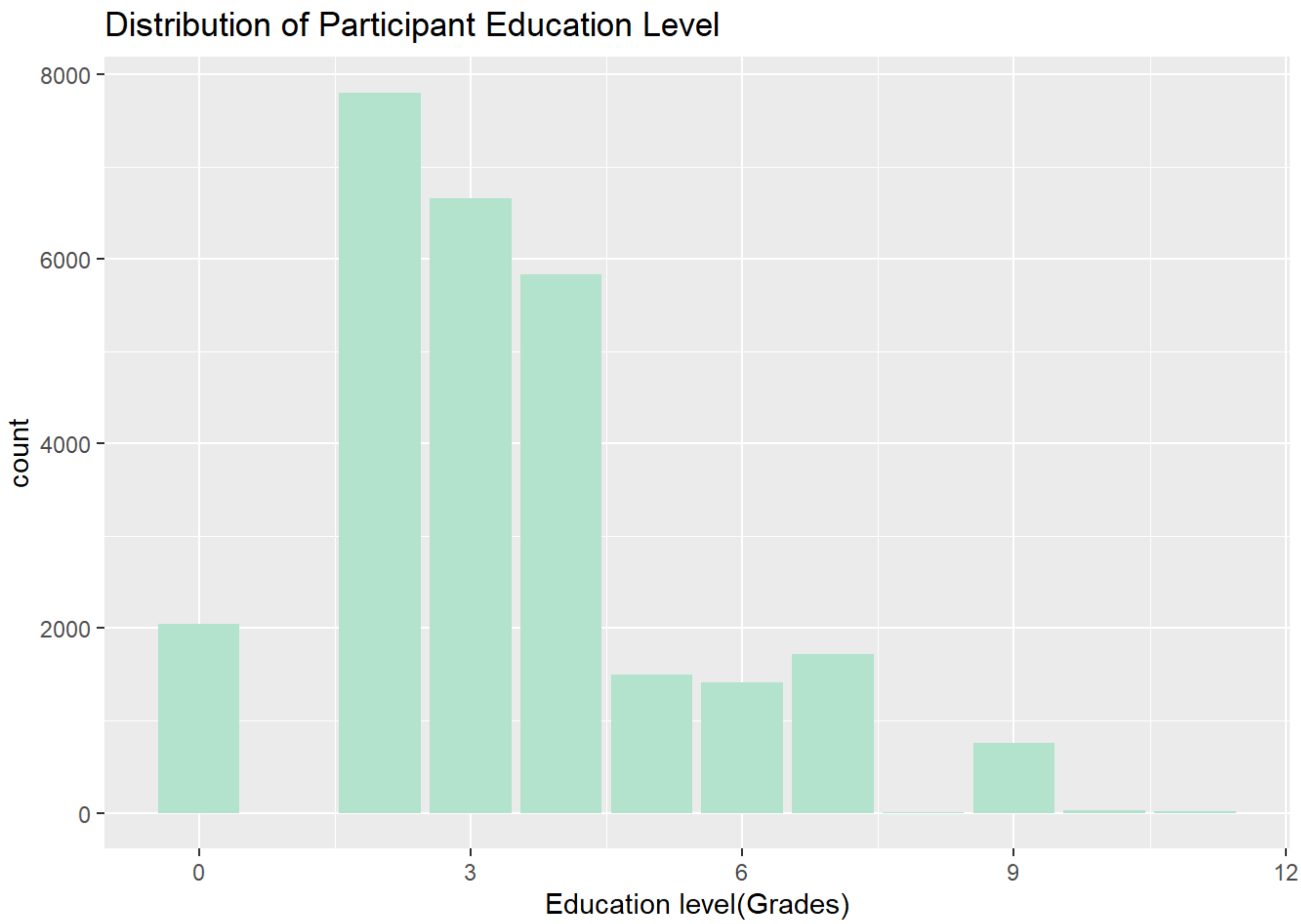
```
## Warning: package 'forcats' was built under R version 3.6.3
```

```
## -- Conflicts ----------------------------------------------------------------------------------------------------------------------------- tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
# Correlation Matrix with all numeric variables
VN_num <- VN_Data %>%
  select(illness, injury, illdays, actdays, commune, MExpense, pharvis, age, educ)
VN_Matrix <- VN_Data %>%
  select_if(is.numeric)
cor(VN_num, use = "pairwise.complete.obs")
```

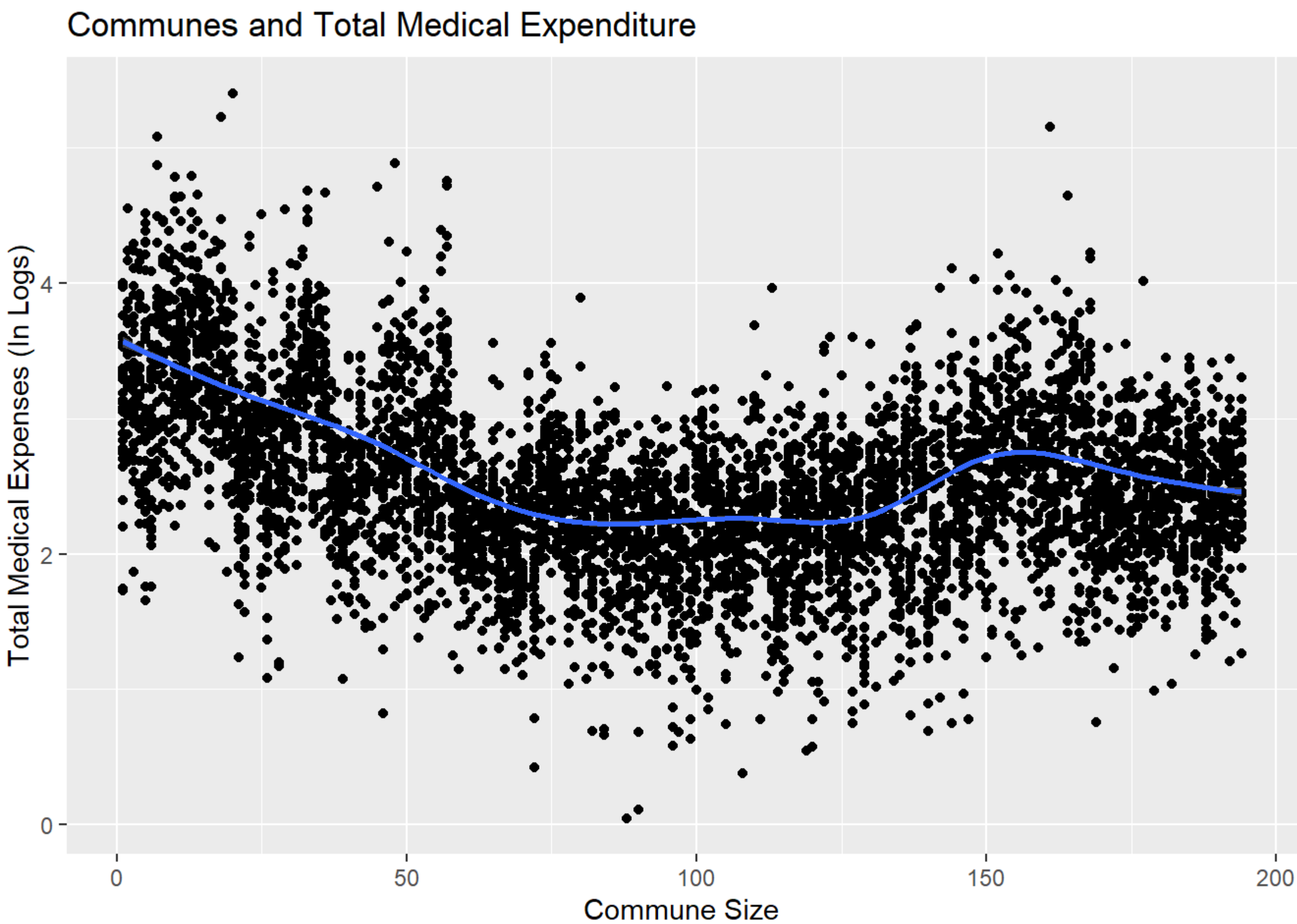
```
##              illness        injury     illdays      actdays       commune
## illness   1.00000000  0.0342110387  0.58251926  0.014887487  0.0393017837
## injury    0.03421104  1.0000000000  0.06081290  0.595513059 -0.0006967836
## illdays   0.58251926  0.0608128964  1.00000000  0.081789848  0.0071653205
## actdays   0.01488749  0.5955130586  0.08178985  1.000000000 -0.0097280785
## commune   0.03930178 -0.0006967836  0.00716532 -0.009728078  1.0000000000
## MExpense -0.10070033 -0.0028277489 -0.06495466 -0.009907336 -0.2883442105
## pharvis   0.42627527  0.0482468328  0.35452961  0.045659014  0.0574551630
## age       0.08107781  0.0248544624  0.14656448  0.031475407 -0.0813954238
## educ     -0.04506705 -0.0028733600 -0.02207188 -0.004395189 -0.3294988982
##              MExpense     pharvis         age         educ
## illness  -0.100700333  0.42627527  0.08107781 -0.045067052
## injury   -0.002827749  0.04824683  0.02485446 -0.002873360
## illdays  -0.064954659  0.35452961  0.14656448 -0.022071880
## actdays  -0.009907336  0.04565901  0.03147541 -0.004395189
## commune  -0.288344210  0.05745516 -0.08139542 -0.329498898
## MExpense  1.000000000 -0.03127047  0.06171122  0.255619678
## pharvis  -0.031270466  1.00000000  0.08339587 -0.052768910
## age       0.061711223  0.08339587  1.00000000  0.025132618
## educ      0.255619678 -0.05276891  0.02513262  1.000000000
```

```
# Univariate Graph
ggplot(VN_Data, aes(x = educ, fill = "educ"))+
  geom_bar() +
  scale_fill_brewer(palette = "Pastel2") +
  ggtitle("Distribution of Participant Education Level") +
  theme(legend.position = "none") +
  xlab("Education level(Grades)")
```




```
# Bivariate Graph
## A scatterplot was graphed to indicate the education levels and total medical expenses of indi
viduals within communes.
ggplot(VN_Data, aes(commune,MExpense)) +
  geom_point() +
  geom_smooth() +
  ggtitle("Communes and Total Medical Expenditure") +
  xlab("Commune Size") +
  ylab("Total Medical Expenses (In Logs)")
```

```
## `geom_smooth()` using method = 'gam' and formula 'y ~ s(x, bs = "cs")'
```




```
# Correlation coefficient for Bivariate relationship
cor(VN_Data$MExpense,VN_Data$commune)
```

```
## [1] -0.2883442
```

For summary statistics, I created a correlation matrix which conveyed the coeffient, or strength, of each numeric variables' relation to each other. Commune and annual medical expense had a relatively weak negative correlation so I made a bivariate scatterplot graph to illustrate this relationship. The trendline suggests that as commune size increases, an individuals' total medical expenditure will decrease as a result. This may suggests that as living commune size increases, the need for healthcare service will decrease (as communes tend to take care of each other), and therefore the total medical expediture will also decrease.

The univariate histogram depicts the distribution of education level amongst the sample of participants. In this graph, it appears that there is a high proportion of people who finished grades 2-4, however this number significantly drops at around 5th grade, and becomes seldom at the higher end of the spectrum. The graph suggests that a majority of the sample has a low level of education which may affect individual insurance status and medical service affordibility.

3.) MANOVA

```
# Mean difference of annual medical expense (log) across insurance status.
VN_Data %>%
  group_by(insurance) %>%
  summarize(mean(MExpense)) %>%
  mutate(Insurance = case_when(insurance>0 ~ "Yes",
                                insurance<1 ~ "No"))
```

```
## # A tibble: 2 x 3
##   insurance `mean(MExpense)` Insurance
## *     <dbl>            <dbl> <chr>
## 1         0             2.56 No
## 2         1             2.82 Yes
```

```
# MANOVA of individuals' annual medical expense, days of limited activity, pharmacy visits and l
iving commune size across insurance status
VN_manova <- manova(cbind(MExpense, actdays , commune, pharvis) ~ insurance, data = VN_Data)
summary(VN_manova)
```

```
##              Df   Pillai approx F num Df den Df    Pr(>F)
## insurance     1 0.045826    333.3      4  27760 < 2.2e-16 ***
## Residuals 27763
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Univariate ANOVA of variables
summary.aov(VN_manova)
```

```
##  Response MExpense :
##                 Df  Sum Sq Mean Sq F value    Pr(>F)
## insurance       1   247.6 247.643     650 < 2.2e-16 ***
## Residuals   27763 10577.4   0.381
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response actdays :
##                 Df Sum Sq Mean Sq F value Pr(>F)
## insurance       1      1 0.52873  0.4246 0.5147
## Residuals   27763  34575 1.24535
##
##  Response commune :
##                 Df   Sum Sq Mean Sq F value    Pr(>F)
## insurance       1  3070383 3070383  1004.3 < 2.2e-16 ***
## Residuals   27763 84880827    3057
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
##  Response pharvis :
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## insurance       1     62  61.992   35.98 2.018e-09 ***
## Residuals   27763  47833   1.723
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Post-hoc t test for annual medical expenses (significant)
pairwise.t.test(VN_Data$MExpense,VN_Data$insurance, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  VN_Data$MExpense and VN_Data$insurance
##
##   0
## 1 <2e-16
##
## P value adjustment method: none
```

```
# Post-hoc t test for living commune size (significant)
pairwise.t.test(VN_Data$commune,VN_Data$insurance, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  VN_Data$commune and VN_Data$insurance
##
##   0
## 1 <2e-16
##
## P value adjustment method: none
```

```
# Post-hoc t test for annual pharmacy visits (significant)
pairwise.t.test(VN_Data$pharvis,VN_Data$insurance, p.adj="none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  VN_Data$pharvis and VN_Data$insurance
##
##   0
## 1 2e-09
##
## P value adjustment method: none
```

```
# Probability of one type I error
# 1 MANOVA + 4 ANOVAs (summary.aov) + 3 post-hoc t. test  = 8 tests in total
Probability_of_Type_I_error  <- 1 - (0.95^8)
Probability_of_Type_I_error * 100
```

```
## [1] 33.65796
```

```
# (Bonferroni's Correction)
0.05/8
```

```
## [1] 0.00625
```

A MANOVA was conducted in order to determine whether individuals' annual medical expense, days of limited activity, pharmacy visits and living commune size differed across insurance status. Since the p-value received was lower than the critical value of 0.05, we reject the null hypothesis and confirm that there is a signifcant difference in annual medical expense, days of limited activity, pharmacy visits and living commune size between insurance status.

Individual ANOVAs were then conducted on each of the variables, and post-hoc analyses were applied to variables that yielded a significantly low p-value (which included medical expenses, commune size, and pharmacy visits). Overall, a total of 8 tests were conducted (1 MANOVA + 4 ANOVAs + 3 post-hoc t. test). There is a 33.6 % chance that at least one type I error had occured within the 8 tests conducted. Therefore, the p-value was adjusted from 0.05 to 0.0063, using bonferroni's correction, and was reapplied to the tests. However, all three factors remained significant and suggests a difference from the null hypothesis. Since the observations in this study are collected in groups of individuals, its unlikely that the sample is random and are independent of each other. Moreover, the data is not centered around a mean and is likely to fail the multivariate normality assumption. Thus, we can assume that the assumptions for this MANOVA are unlikely to be met.

4 Randomization test

```
# ANOVA on medical expenses between different education levels & f statistic
summary(aov(MExpense ~ educ, data = VN_Data))
```

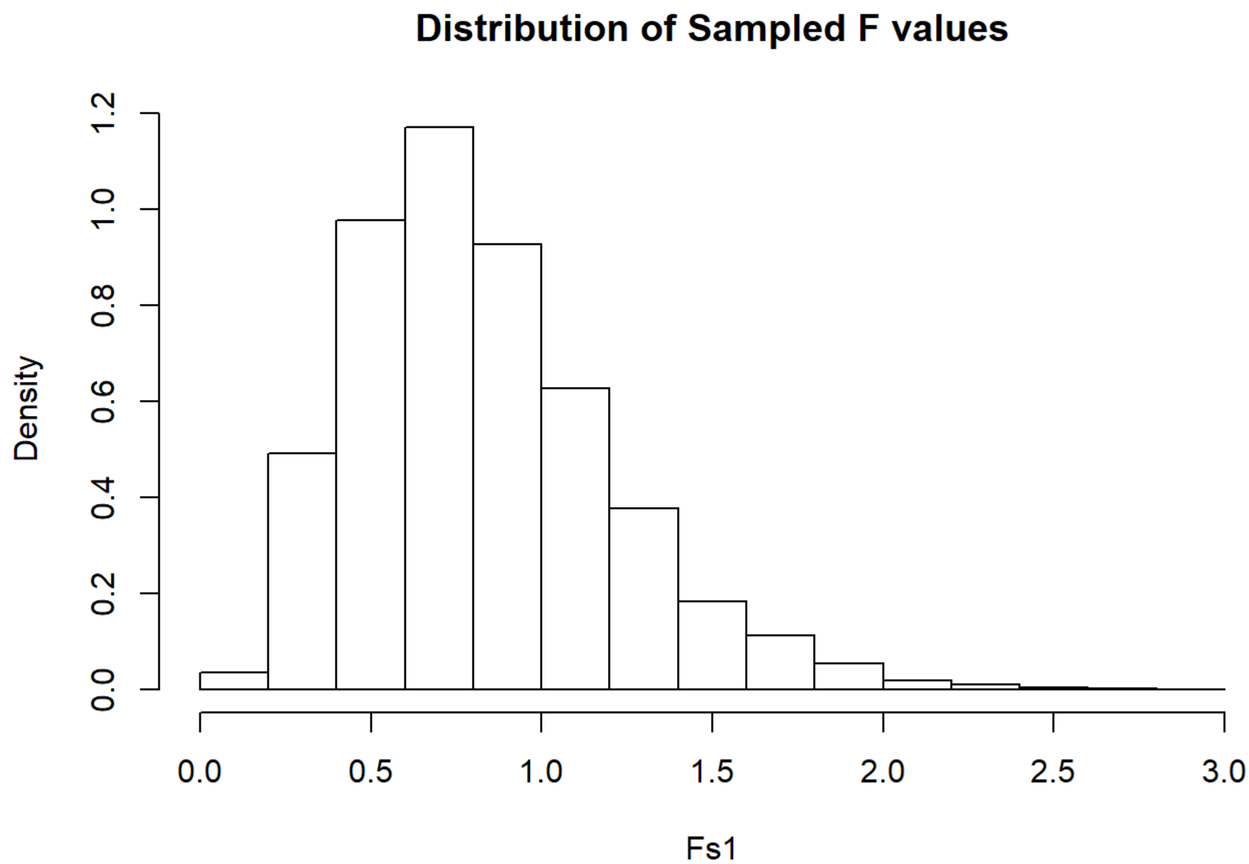
```
##                Df Sum Sq Mean Sq F value Pr(>F)
## educ            1    707   707.3    1941 <2e-16 ***
## Residuals   27763  10118     0.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
F_stat1 <- 1941

# Post hoc for ANOVA
pairwise.t.test(VN_Data$MExpense, VN_Data$educ, p.adj = "none")
```

```
##
##  Pairwise comparisons using t tests with pooled SD
##
## data:  VN_Data$MExpense and VN_Data$educ
##
##    0       2       3       4       5       6       7       8       9
## 2  < 2e-16 -       -       -       -       -       -       -       -
## 3  < 2e-16 < 2e-16 -       -       -       -       -       -       -
## 4  < 2e-16 2.5e-06 < 2e-16 -       -       -       -       -       -
## 5  < 2e-16 < 2e-16 < 2e-16 < 2e-16 -       -       -       -       -
## 6  < 2e-16 < 2e-16 9.8e-07 < 2e-16 2.4e-06 -       -       -       -
## 7  < 2e-16 < 2e-16 < 2e-16 < 2e-16 0.64339 9.7e-08 -       -       -
## 8  2.4e-07 1.0e-05 0.00014 2.7e-05 0.00308 0.00062 0.00353 -       -
## 9  < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 < 2e-16 0.47552 -
## 10 < 2e-16 1.7e-14 6.8e-11 3.7e-13 9.4e-07 7.9e-09 1.4e-06 0.75752 0.49438
## 11 1.0e-14 2.2e-11 5.4e-09 1.7e-10 3.7e-06 1.3e-07 4.9e-06 0.79356 0.14606
##    10
## 2  -
## 3  -
## 4  -
## 5  -
## 6  -
## 7  -
## 8  -
## 9  -
## 10 -
## 11 0.44822
##
## P value adjustment method: none
```

```r
# Randomization test for F-statistic
set.seed(123)
Fs1 <- replicate(5000,{
  new <- VN_Data %>%
    mutate(ME = sample(MExpense))
  SSW <- new %>%
    group_by(educ) %>%
    summarize(SSW = sum((ME - mean(ME))^2)) %>%
    summarize(sum(SSW)) %>%
    pull
  SSB <- new %>%
    mutate(mean = mean(ME)) %>%
    group_by(educ) %>%
    mutate(groupmean = mean(ME)) %>%
    summarize(SSB = sum((mean - groupmean)^2)) %>%
    summarize(sum(SSB)) %>%
    pull
  # Compute the F-statistic (ratio of MSB and MSW)
  # df for SSB is 13 groups - 1 = 12
  # df for SSW is 27765 observations - 13 groups = 27752
  (SSB/12)/(SSW/27752)
})

# Null distribution and F-statistic graph
hist(Fs1, prob=T, main = "Distribution of Sampled F values");
     abline(v = F_stat1, col="red",add = T)
```

```
## Warning in int_abline(a = a, b = b, h = h, v = v, untf = untf, ...): "add" is
## not a graphical parameter
```



The null hypothesis for the ANOVA states that there is no significant difference in medical expenses by education level. However, since the p-value turned out to be less than 0.05, we reject the null hypothesis and conclude that there is a significant difference between medical expenses and education. A looped randomization test was then conducted to get a many sample statistics. The graph shows the distribution of sampled values compared to the observed f statistic value acquired through ANOVA. Since the observed f-statistic (1,941) surpasses the x-axis domain, we are unable to see its indicator in this graph. As shown, the histogram is moderately skewed to the right meaning that the mean of sampled f-values is likely to be greater than the median value.

5.) Linear regression model

```
# Update VN dataset

# Linear Regression Model
VN_fit <- lm(MExpense ~ educ + Insurance + educ*Insurance, data = VN_Data)
summary(VN_fit)
```

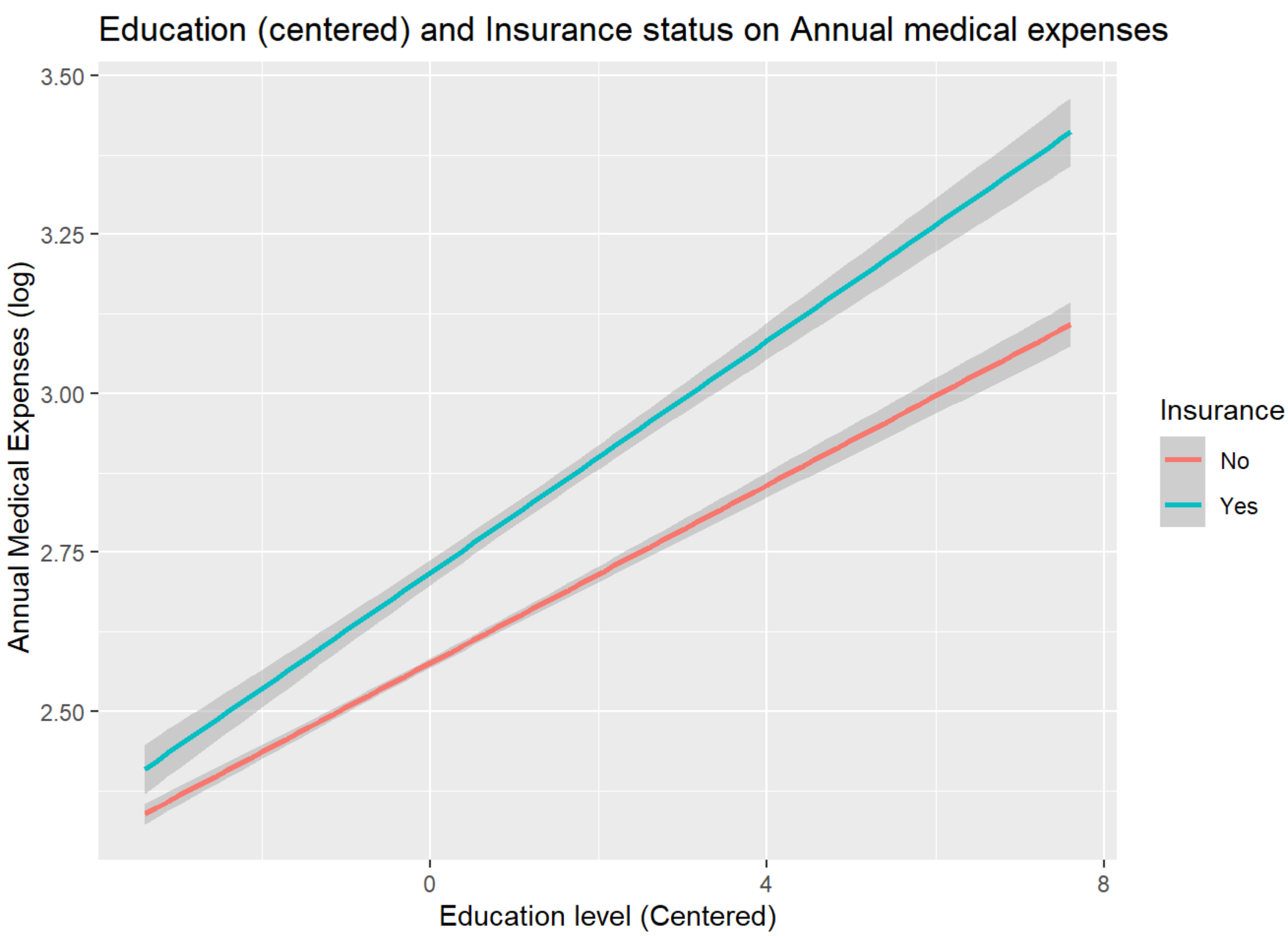
```
##
## Call:
## lm(formula = MExpense ~ educ + Insurance + educ * Insurance,
##     data = VN_Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50174 -0.40133 -0.05707  0.35388  2.60332
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)        2.338425   0.008072 289.698  < 2e-16 ***
## educ               0.070005   0.002216  31.593  < 2e-16 ***
## InsuranceYes       0.070423   0.021248   3.314  0.00092 ***
## educ:InsuranceYes 0.021111   0.004492   4.699 2.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6008 on 27761 degrees of freedom
## Multiple R-squared:  0.07429,    Adjusted R-squared:  0.07419
## F-statistic: 742.6 on 3 and 27761 DF,  p-value: < 2.2e-16
```

```
# Mean-center numeric explanatory variables
VN_Data$educ_c <- VN_Data$educ - mean(VN_Data$educ, na.rm = T)

# New mean-centered LRM
VN_fit_c <- lm(MExpense ~ educ_c * Insurance, data = VN_Data)
summary(VN_fit_c)
```

```
##
## Call:
## lm(formula = MExpense ~ educ_c * Insurance, data = VN_Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50174 -0.40133 -0.05707  0.35388  2.60332
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)          2.575788   0.003968 649.161  < 2e-16 ***
## educ_c               0.070005   0.002216  31.593  < 2e-16 ***
## InsuranceYes         0.142003   0.010668  13.312  < 2e-16 ***
## educ_c:InsuranceYes 0.021111   0.004492   4.699 2.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6008 on 27761 degrees of freedom
## Multiple R-squared:  0.07429,    Adjusted R-squared:  0.07419
## F-statistic: 742.6 on 3 and 27761 DF,  p-value: < 2.2e-16
```

```
# Line graph depicting interaction between annual medical expenses (log) and mean-centered educa
tion
ggplot(data = VN_Data, aes(x= educ_c, y= MExpense, col = Insurance)) +
 geom_smooth(method = "lm") +
  ggtitle("Education (centered) and Insurance status on Annual medical expenses") +
  ylab(" Annual Medical Expenses (log)") +
  xlab("Education level (Centered)")
```

```
## `geom_smooth()` using formula 'y ~ x'
```




```
# coeffiicient estimates for mean-centered LRM
coef(VN_fit_c)
```

```
##        (Intercept)           educ_c        InsuranceYes educ_c:InsuranceYes
##         2.57578794       0.07000464          0.14200280          0.02111094
```

```
2.6 # Intercept
```

```
## [1] 2.6
```

```
0.07 # Centered education
```

```
## [1] 0.07
```

```
0.14 # Individuals with Insurance
```

```
## [1] 0.14
```

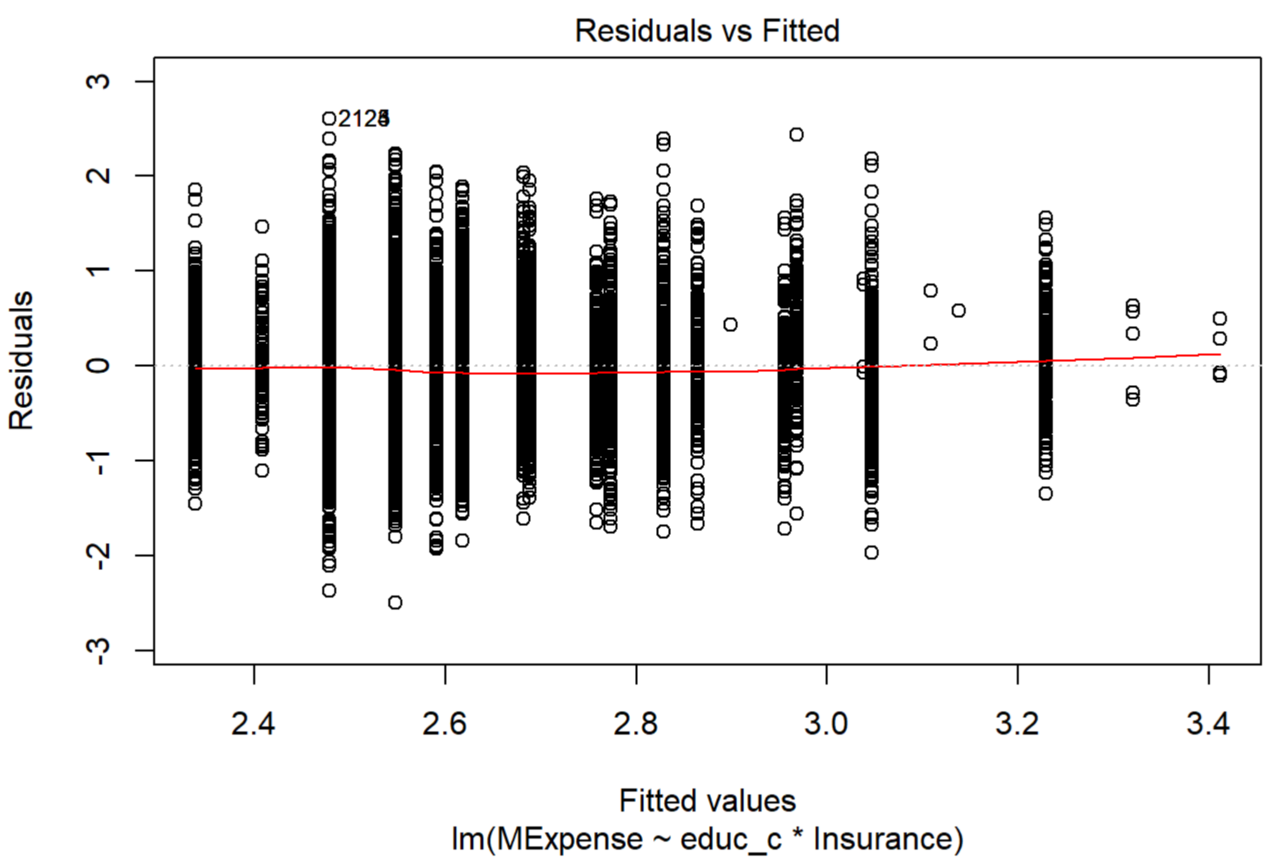
```
.021 # Interaction between centered education and individuals with insurance
```

```
## [1] 0.021
```

```
# Proportion of variation in response explained by the model (r^2)
Adjusted_R_Squared <- 0.07419
Adjusted_R_Squared * 100
```

```
## [1] 7.419
```

```
# Linearity assumption
plot(VN_fit_c, which = 1)
```



```r
# Normality assumption
ks.test(VN_fit_c$residuals, "pnorm", mean=0, sd(VN_fit_c$residuals))
```

```
## Warning in ks.test(VN_fit_c$residuals, "pnorm", mean = 0,
## sd(VN_fit_c$residuals)): ties should not be present for the Kolmogorov-Smirnov
## test
```

```
##
##  One-sample Kolmogorov-Smirnov test
##
## data:  VN_fit_c$residuals
## D = 0.041398, p-value < 2.2e-16
## alternative hypothesis: two-sided
```

```r
# Homoscedasticity assumption
install.packages("lmtest", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/hpham/OneDrive/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'lmtest' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'lmtest'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE):
## problem copying C:\Users\hpham\OneDrive\Documents\R\win-
## library\3.6\00LOCK\lmtest\libs\x64\lmtest.dll to C:
## \Users\hpham\OneDrive\Documents\R\win-library\3.6\lmtest\libs\x64\lmtest.dll:
## Permission denied
```

```
## Warning: restored 'lmtest'
```

```
##
## The downloaded binary packages are in
##  C:\Users\hpham\AppData\Local\Temp\Rtmp6BwTT2\downloaded_packages
```

```r
library(lmtest)
```

```
## Warning: package 'lmtest' was built under R version 3.6.3
```

```
## Loading required package: zoo
```

```
## Warning: package 'zoo' was built under R version 3.6.3
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(VN_fit_c)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  VN_fit_c
## BP = 65.881, df = 3, p-value = 3.25e-14
```

```
# Robust standard errors
install.packages("sandwich", repos = "http://cran.us.r-project.org")
```

```
## Installing package into 'C:/Users/hpham/OneDrive/Documents/R/win-library/3.6'
## (as 'lib' is unspecified)
```

```
## package 'sandwich' successfully unpacked and MD5 sums checked
##
## The downloaded binary packages are in
##  C:\Users\hpham\AppData\Local\Temp\Rtmp6BwTT2\downloaded_packages
```

```
library(sandwich)
```

```
## Warning: package 'sandwich' was built under R version 3.6.3
```

```
coeftest(VN_fit_c, vcov = vcovHC(VN_fit_c))
```

```
##
## t test of coefficients:
##
##                      Estimate Std. Error  t value  Pr(>|t|)
## (Intercept)         2.5757879  0.0039991 644.0864 < 2.2e-16 ***
## educ_c              0.0700046  0.0023189  30.1892 < 2.2e-16 ***
## InsuranceYes        0.1420028  0.0109542  12.9633 < 2.2e-16 ***
## educ_c:InsuranceYes 0.0211109  0.0045323   4.6579  3.21e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
# Compare robust SEs to original SEs
summary(VN_fit_c)
```

```
##
## Call:
## lm(formula = MExpense ~ educ_c * Insurance, data = VN_Data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.50174 -0.40133 -0.05707  0.35388  2.60332
##
## Coefficients:
##                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)         2.575788   0.003968 649.161  < 2e-16 ***
## educ_c              0.070005   0.002216  31.593  < 2e-16 ***
## InsuranceYes        0.142003   0.010668  13.312  < 2e-16 ***
## educ_c:InsuranceYes 0.021111   0.004492   4.699 2.62e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6008 on 27761 degrees of freedom
## Multiple R-squared:  0.07429,    Adjusted R-squared:  0.07419
## F-statistic: 742.6 on 3 and 27761 DF,  p-value: < 2.2e-16
```

```
# Bootstrapped errors
set.seed(123)
VN_samp_SEs <- replicate(5000, {
  boot_data <- sample_frac(VN_Data, replace = TRUE)
  fitboot <- lm(MExpense ~ educ_c * Insurance, data = boot_data)
  coef(fitboot)
})

# Bootstrapped confidence interval
VN_samp_SEs  %>%
  t %>%
  # Consider the matrix as a data frame
  as.data.frame %>%
  pivot_longer(everything(), names_to = "estimates", values_to = "value") %>%
  group_by(estimates) %>%
  summarize(lower = quantile(value,.025), upper = quantile(value,.975))
```

```
## # A tibble: 4 x 3
##   estimates            lower  upper
## * <chr>                <dbl>  <dbl>
## 1 (Intercept)         2.57   2.58
## 2 educ_c              0.0655 0.0745
## 3 educ_c:InsuranceYes 0.0124 0.0298
## 4 InsuranceYes        0.120  0.163
```

```
# Compare bootstraps to original CI
confint(VN_fit_c, level = 0.95)
```

```
##                           2.5 %     97.5 %
## (Intercept)          2.56801072 2.58356516
## educ_c               0.06566152 0.07434776
## InsuranceYes         0.12109396 0.16291165
## educ_c:InsuranceYes 0.01230593 0.02991595
```

```
# Bootstrapped standard errors
VN_samp_SEs %>%
  t %>%
  as.data.frame %>%
  summarize_all(sd)
```

```
##   (Intercept)      educ_c InsuranceYes educ_c:InsuranceYes
## 1 0.003975099 0.002298117   0.01095353         0.004453165
```

```
# Compare bootstrapped SEs to original SEs
coeftest(VN_fit_c)[,1:2]
```

```
##                        Estimate  Std. Error
## (Intercept)          2.57578794 0.003967870
## educ_c               0.07000464 0.002215823
## InsuranceYes         0.14200280 0.010667510
## educ_c:InsuranceYes 0.02111094 0.004492237
```

```
# Compare bootstrapped SEs to robust SEs
coeftest(VN_fit_c, vcov = vcovHC(VN_fit_c))[,1:2]
```

```
##                        Estimate  Std. Error
## (Intercept)          2.57578794 0.003999134
## educ_c               0.07000464 0.002318861
## InsuranceYes         0.14200280 0.010954230
## educ_c:InsuranceYes 0.02111094 0.004532301
```

A linear regression model was conducted on mean centered education level, insurance, and their interactions on the annual medical expenditure. At an education level of 0, the annual medical expense (log) would be 0.0040 units in currency. While holding all other variables constant, a one grade increase in education will result in a 0.0023 currency increase in annual medical expenses.Moreover, while holding education level constant,people with insurance have a .011 currency increase in annual medical expenses. Finally, people with education and insurance have a 0.0045 increase in rate of change in annual medical expenses than educated people who do not have insurance. The model explains for approximately 7.419 % of variance in the response.

The residuals on the residual plot indicates a megaphone-like distribution, where a majority of the data is on the left end (growing from the right), thus failing the linearity assumption. Moreover, by performing the KS test for normality, I had obtained a p-value of less than 0.05, therefore I reject the null hypothesis that there is no significant difference in normality (normality assumption fails). With the addition of a significant difference found in the homoscedasticity test (p-value < 0.05), I can confirm that all three assumptions (lineraity, normality, and homoscedasticity) failed for this dataset.

By comparing the robust SEs to the original SEs, we do not see much of a change, indicating that the robust method has not helped to adjust the precision in the estimate. Moreover, by comparing the bootstrap SE to the original and robust models,we do not see much of a change in standard error suggesting that both tests were ineffective at normalizing the data. I further compared the bootstrap confidence interval to the original model interval and there was still little change in the difference in the lower and upper bounds (except for an extremely miniscule increase in people who have insurance in the bootstrap model).

6.) Logistic Regression

```
# Logistic Regression Model observing the effect of pharmacy visits and age on insurance
VN_fit2 <- glm(insurance ~ pharvis + age, data = VN_Data, family = "binomial")
summary(VN_fit2)
```

```
##
## Call:
## glm(formula = insurance ~ pharvis + age, family = "binomial",
##     data = VN_Data)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.7167  -0.6315  -0.5849  -0.5046   2.7190
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.28710    0.05945 -38.473  < 2e-16 ***
## pharvis     -0.11619    0.01639  -7.087 1.37e-12 ***
## age          0.23043    0.01849  12.462  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 24651  on 27764  degrees of freedom
## Residual deviance: 24443  on 27762  degrees of freedom
## AIC: 24449
##
## Number of Fisher Scoring iterations: 4
```

```
# Coefficient of Logistic Regression Model
coef(VN_fit2)
```

```
## (Intercept)     pharvis         age
##  -2.2870977  -0.1161910   0.2304348
```

```
# Exponentiate coefficients
exp_VN_fit2 <- exp(coef(VN_fit2))
exp_VN_fit2
```

```
## (Intercept)     pharvis         age
##   0.1015608   0.8903051   1.2591473
```

```
# Confusion matrix of LRM: Create predicted probability variable
VN_Data$prob <- predict(VN_fit2, type = "response")

# Confusion matrix of LRM: Classifying predicted outcome
VN_Data$predicted <- ifelse(VN_Data$prob > .20, "insured", "uninsured")

# Confusion matrix of LRM: Table
 table(truth = VN_Data$Insurance, prediction = VN_Data$predicted) %>%
  addmargins
```

```
##      prediction
## truth insured uninsured   Sum
##   No     2079     21172 23251
##   Yes     612      3902  4514
##   Sum    2691     25074 27765
```

```
# Accuracy
(21172 + 612)/27765
```

```
## [1] 0.7845849
```

```
# Sensitivity - True Positive Rate
612/4514
```

```
## [1] 0.1355782
```

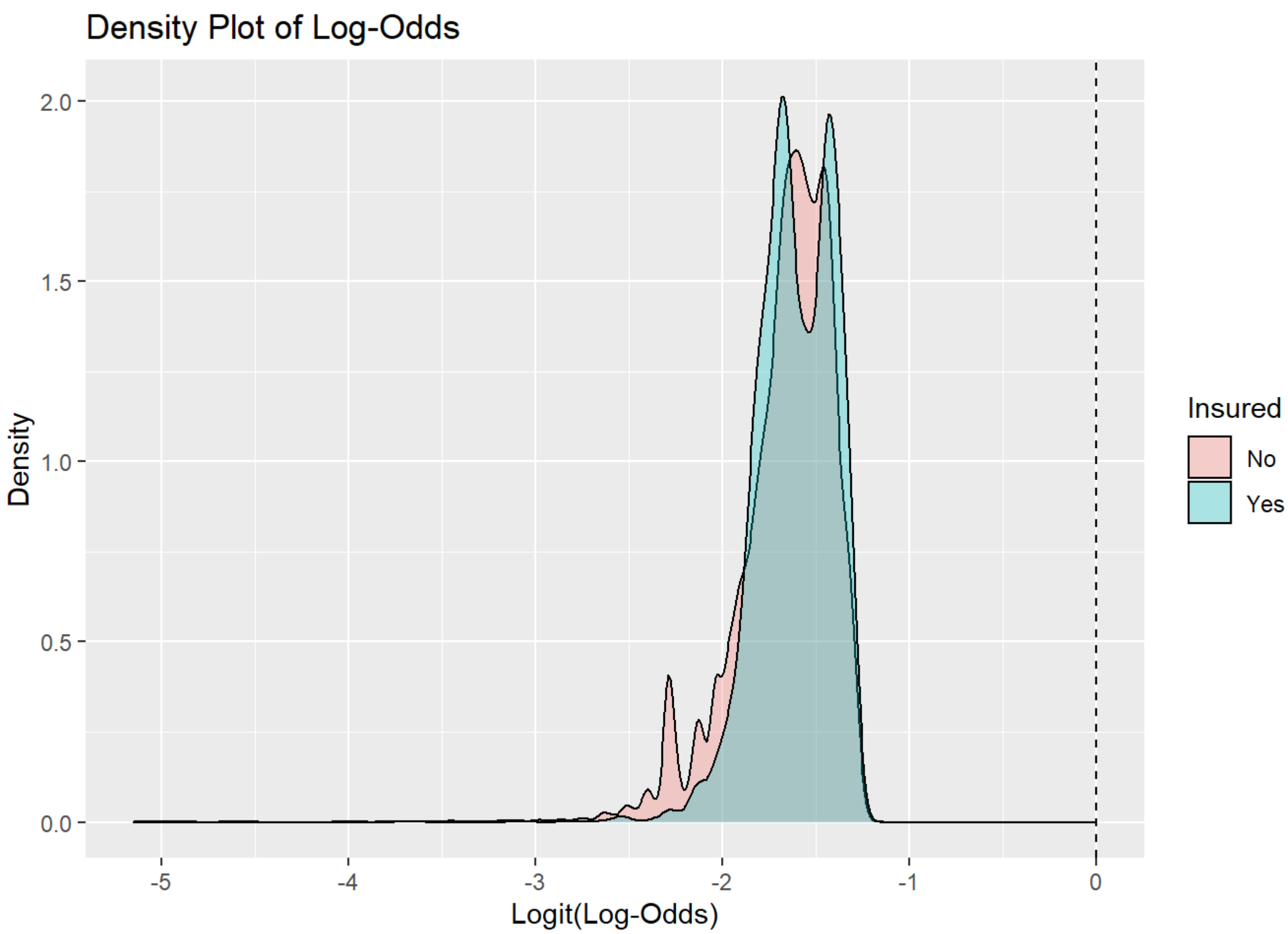
```
# Specificity - True Negative Rate
21172/23251
```

```
## [1] 0.9105845
```

```
# Recall - Positive Predictive Value
612/2691
```

```
## [1] 0.2274247
```

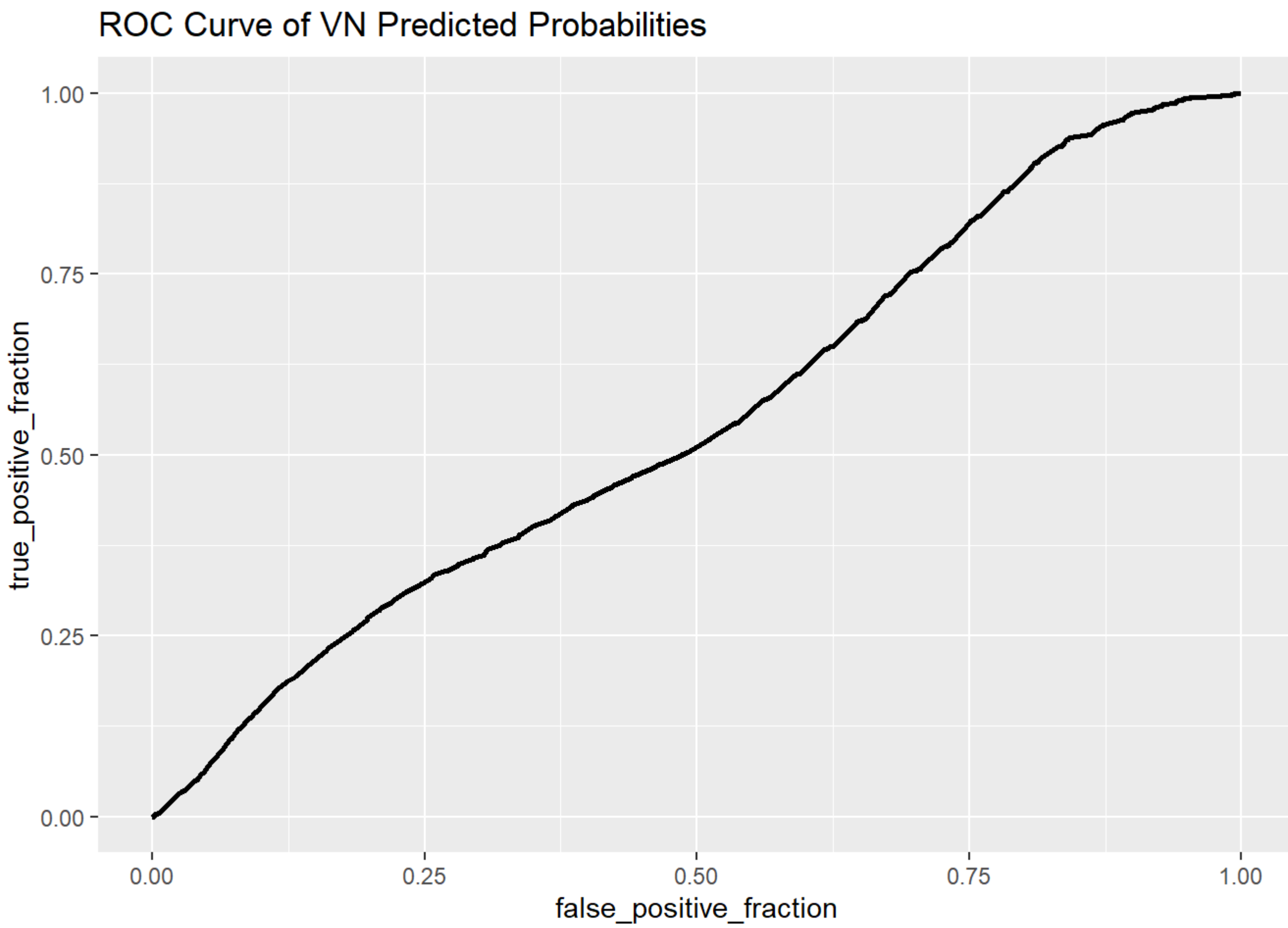
```
# Preparing the Density plot with Predicted log odds
VN_Data$logit <- predict(VN_fit2, type = "link")

# Density plot of Log-odds
ggplot(VN_Data, aes(logit, fill = as.factor(Insurance))) +
  geom_density(alpha = .3) +
  geom_vline(xintercept = 0, lty = 2) +
  labs(fill = "Insured") +
  ggtitle("Density Plot of Log-Odds") +
  xlab("Logit(Log-Odds)") +
  ylab("Density")
```




```
# ROC-curve plot
library(plotROC)
```

```
## Warning: package 'plotROC' was built under R version 3.6.3
```

```
VN_Data$prob <- predict(VN_fit2, type = "response")
VN_ROCplot <- ggplot(VN_Data) +
  geom_roc(aes(d = insurance , m = prob), n.cuts=0) +
  ggtitle("ROC Curve of VN Predicted Probabilities")
VN_ROCplot
```




```
# AUC calculations
calc_auc(VN_ROCplot)
```

```
##   PANEL group       AUC
## 1     1    -1 0.5478776
```

A logistic regression model was conducted to observe whether there was a difference between pharmacy visits and age on insurance. The coefficients were exponentiated in order to increase the normality of the data.

While controlling for age, for every 1 unit increase in pharmacy visits increases the odds of insurance by a factor of 0.8903.

While controlling for pharmacy visits, for every 1 unit increase in age (year), increases the odds of insurance by a factor of 1.259.

The accuracy of this model was determined to be 47.4 %, while the sensitivity was determined to be 13.5%. Likewise, the specificity of the model was determined to be 91.0% while the recall was determined to be 22.7%. Using all of this information derived from the confusion matrix, a roc curve plot was drawn and its AUC was calculated. Both the diagonal slope of the graph and the AUC caluclation (0.58) suggests that the model has a fairly weak power of prediction. Thus from this model, we cannot reaffirm the significance of the p-values from the logistic regression model and therefore fail to conclude that there is a signifcant difference between pharmacy visits and age on insurance status.