FMPH222 Final

Kevin Nguyen

2023-03-22

Libraries

```
library(tableone)
```

```
## Warning: package 'tableone' was built under R version 4.1.1
```

```
library(ggfortify)
```

```
## Warning: package 'ggfortify' was built under R version 4.1.1
```

```
## Loading required package: ggplot2
```

```
## Warning: package 'ggplot2' was built under R version 4.1.1
```

```
library(generalhoslem)
```

```
## Loading required package: reshape
```

```
## Warning: package 'reshape' was built under R version 4.1.1
```

```
## Loading required package: MASS
```

```
## Warning: package 'MASS' was built under R version 4.1.1
```

```
library(ggplot2)
library(lattice)
```

```
## Warning: package 'lattice' was built under R version 4.1.1
```

1

```
wcgsKM = read.csv("wcgsKM.csv")

wcgsKM$chd[wcgsKM$chd == "no"] = "0"
wcgsKM$chd[wcgsKM$chd == "yes"] = "1"
wcgsKM$chd = as.numeric(wcgsKM$chd)
```

1A

```
# Variables of Interest
variables_interest <- c("current.smoker", "age.cat", "bmi.cat",
              "dbp.cat", "sbp.cat","chol.cat", "chd")

# Table Creation
summary_table <- CreateTableOne(vars = variables_interest,
```

```
                           data = wcgsKM)
print(summary_table, formatOptions = list(big.mark = ","))
```

```
##
##                               Overall
##   n                           3,154
##   current.smoker (mean (SD))  0.48 (0.50)
##   age.cat (%)
##      [0,40)                    266 ( 8.4)
##      [40,45)                  1182 (37.5)
##      [45,50)                   801 (25.4)
##      [50,55)                   583 (18.5)
##      [55,60)                   322 (10.2)
##   bmi.cat (%)
##      healthy wt               1833 (58.1)
##      obese                      81 ( 2.6)
##      overweight               1217 (38.6)
##      underweight                23 ( 0.7)
##   dbp.cat = normal (%)        2505 (79.4)
##   sbp.cat (%)
##      elevated                  999 (31.7)
##      high                     1388 (44.0)
##      normal                    767 (24.3)
##   chol.cat (%)
##      high                     1205 (38.4)
##      normal                    831 (26.5)
##      very high                1105 (35.2)
##   chd (mean (SD))             0.08 (0.27)
```

Within the study population, we are able to see that about 48% of the study participants smoke, the majority of them, approximately 37.5% of them are 40-45 years old, 58.1% are a healthy weight, 79.4% of them have normal diastolic blood pressure, 44% of them have high systolic blood pressure, and 38.4% of them have high cholesterol. We are able to see that approximately 8.1% of the study population developed coronary heart disease.

1B

```
# saturated model
sat_model = glm(chd ~ current.smoker + age.cat + bmi.cat + dbp.cat +
                  chol.cat + sbp.cat,
                family = binomial(link="logit"), data = wcgsKM)
summary(sat_model)
```

```
##
## Call:
## glm(formula = chd ~ current.smoker + age.cat + bmi.cat + dbp.cat +
##     chol.cat + sbp.cat, family = binomial(link = "logit"), data = wcgsKM)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0307  -0.4597  -0.3314  -0.2357   3.0491
##
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
```

```
## (Intercept)         -3.10663    0.33962  -9.147  < 2e-16 ***
## current.smoker       0.62336    0.13894   4.486 7.24e-06 ***
## age.cat[40,45)      -0.46325    0.28060  -1.651 0.098755 .
## age.cat[45,50)       0.12750    0.27375   0.466 0.641379
## age.cat[50,55)       0.35444    0.27720   1.279 0.201024
## age.cat[55,60)       0.64201    0.29345   2.188 0.028683 *
## bmi.catobese         0.56398    0.35387   1.594 0.110993
## bmi.catoverweight    0.17031    0.14073   1.210 0.226203
## bmi.catunderweight -13.02211  292.50799  -0.045 0.964491
## dbp.catnormal       -0.06541    0.16959  -0.386 0.699729
## chol.catnormal      -0.58959    0.21786  -2.706 0.006804 **
## chol.catvery high    0.51920    0.14787   3.511 0.000446 ***
## sbp.cathigh          0.40262    0.17131   2.350 0.018761 *
## sbp.catnormal       -0.41411    0.22525  -1.838 0.065990 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1774.2  on 3140  degrees of freedom
## Residual deviance: 1629.6  on 3127  degrees of freedom
##   (13 observations deleted due to missingness)
## AIC: 1657.6
##
## Number of Fisher Scoring iterations: 14
```

In this saturated model, before we proceed to model selection, we notice that bmi.catunderweight has a very low estimate, extremely high standard error and very high p value. This is indicative of quasi-complete separation which was also noticable when we were taking a look at the summary statistics since there were only 23 individuals in this catrgory when there were many more in the other catrgories. To address this, I would be combining the bmi.catunderweight with bmi.cathealthy wt. However, because the bmi.catobese category only has 81 individuals whereas the other categories would have over 1000 individuals, I will be adding this category to bmi.catoverweight as well. At the end, for bmi.cat, we will have healthy wt and overweight as the two categories.

```
# recategorizing bmi.cat
wcgsKM$bmi.cat2 = wcgsKM$bmi.cat
wcgsKM$bmi.cat2[wcgsKM$bmi.cat2 == "underweight"] = "healthy wt"
wcgsKM$bmi.cat2[wcgsKM$bmi.cat2 == "obese"] = "overweight"

# updated saturated model
sat_model2 = glm(chd ~ current.smoker + age.cat + bmi.cat2 + dbp.cat +
                   chol.cat + sbp.cat,
                 family = binomial(link="logit"), data = wcgsKM)
summary(sat_model2)
```

```
##
## Call:
## glm(formula = chd ~ current.smoker + age.cat + bmi.cat2 + dbp.cat +
##     chol.cat + sbp.cat, family = binomial(link = "logit"), data = wcgsKM)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -0.8950  -0.4556  -0.3346  -0.2388   3.0549
##
```

```
## Coefficients:
##                     Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.09924    0.33906  -9.141  < 2e-16 ***
## current.smoker      0.61429    0.13888   4.423 9.72e-06 ***
## age.cat[40,45)     -0.46099    0.28047  -1.644 0.100246
## age.cat[45,50)      0.12925    0.27365   0.472 0.636691
## age.cat[50,55)      0.35321    0.27708   1.275 0.202391
## age.cat[55,60)      0.64644    0.29321   2.205 0.027477 *
## bmi.cat2overweight  0.20958    0.13795   1.519 0.128704
## dbp.catnormal      -0.08548    0.16864  -0.507 0.612213
## chol.catnormal     -0.59730    0.21776  -2.743 0.006089 **
## chol.catvery high   0.51190    0.14769   3.466 0.000528 ***
## sbp.cathigh         0.40901    0.17122   2.389 0.016906 *
## sbp.catnormal      -0.41364    0.22506  -1.838 0.066073 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1774.2  on 3140  degrees of freedom
## Residual deviance: 1634.4  on 3129  degrees of freedom
##   (13 observations deleted due to missingness)
## AIC: 1658.4
##
## Number of Fisher Scoring iterations: 6
```

A backward selection model was used that took a look at the AIC for each variable and removed the variables with the lowest the model was reduced to the lowest AIC as possible. We then added the main variables of interest back into the model and proceeded with our analysis.

```
# saturated model without main variable of interest
sat_model3 = glm(chd ~ bmi.cat2 + dbp.cat + chol.cat + sbp.cat,
                 family = binomial(link="logit"), data = wcgsKM)
# backward selection
reduced_model = step(sat_model3, direction = "backward")
```

```
## Start:  AIC=1698.48
## chd ~ bmi.cat2 + dbp.cat + chol.cat + sbp.cat
##
##            Df Deviance    AIC
## - dbp.cat   1   1684.5 1696.5
## - bmi.cat2  1   1685.1 1697.1
## <none>          1684.5 1698.5
## - sbp.cat   2   1707.3 1717.3
## - chol.cat  2   1731.6 1741.6
##
## Step:  AIC=1696.54
## chd ~ bmi.cat2 + chol.cat + sbp.cat
##
##            Df Deviance    AIC
## - bmi.cat2  1   1685.3 1695.3
## <none>          1684.5 1696.5
## - sbp.cat   2   1714.4 1722.4
## - chol.cat  2   1731.9 1739.9
##
```

```
## Step:  AIC=1695.27
## chd ~ chol.cat + sbp.cat
##
##            Df Deviance    AIC
## <none>          1685.3 1695.3
## - sbp.cat   2   1718.0 1724.0
## - chol.cat  2   1732.8 1738.8
```

```
reduced_model = update(reduced_model, ~ . + current.smoker + age.cat)
```

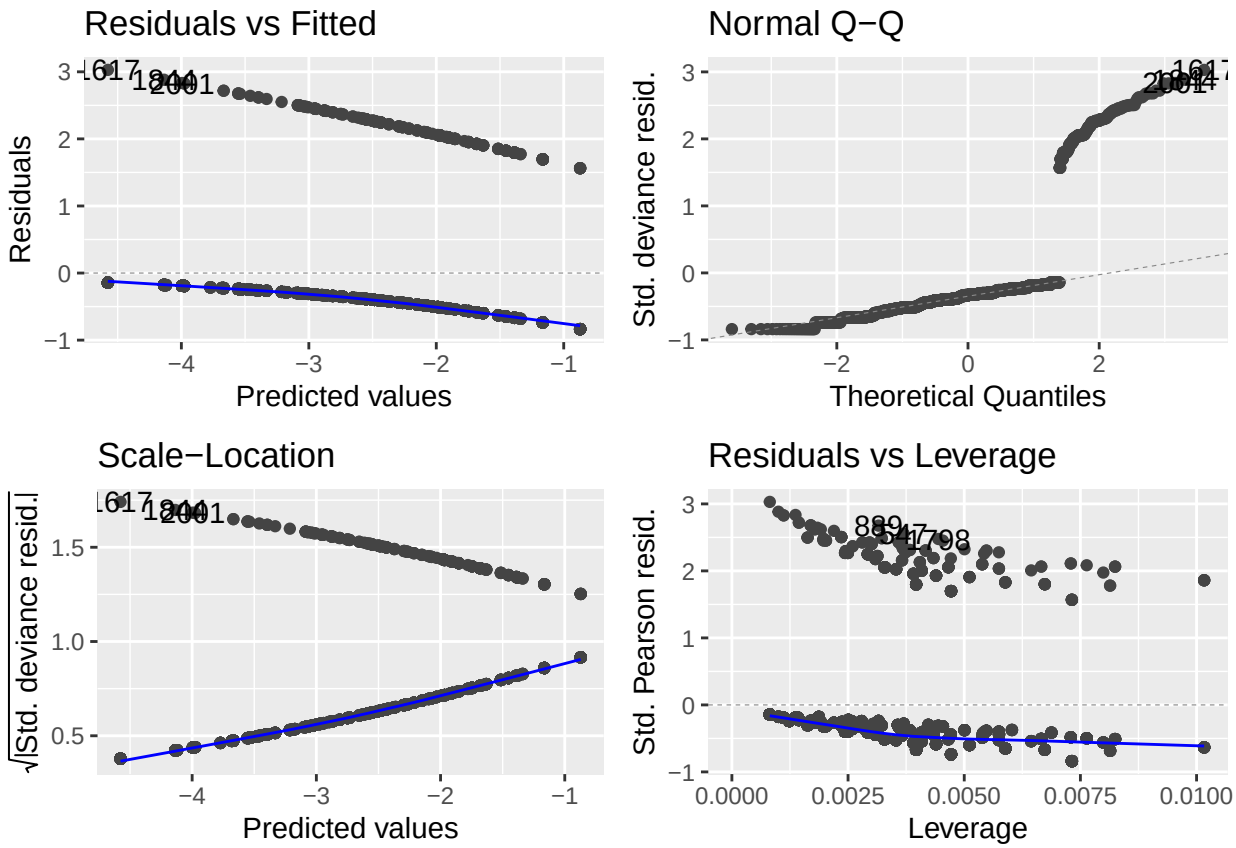
```
# Hosmer Lemeshow Test
hoslem = function (obs, pred, g = 0)
{
if (g == 0)
g = round(min(length(obs)/10, 10))
ord <- order(pred)
obs.o <- obs[ord]
pred.o <- pred[ord]
interval = cut(pred.o, quantile(pred.o, 0:g/g), include.lowest = TRUE)
counts = xtabs(formula = cbind(obs.o, pred.o) ~ interval)
centers <- aggregate(pred.o ~ interval, FUN = "mean")
pear.res <- (counts[, "obs.o"] - counts[, "pred.o"])/sqrt(counts[,
"pred.o"])
pearson <- sum(pear.res^2)
if (any(counts[, "pred.o"] < 5))
warning("Some expected counts are less than 5. Use smaller number of")
p = 1 - pchisq(pearson, g - 2)
results <- as.data.frame(cbind(counts[, "obs.o"], counts[,
"pred.o"], centers$pred.o, pear.res))
colnames(results) <- c("obs.o", "pred.o", "avg mean", "pearson resid")
cat("Hosmer-Lemeshow test with", g, "bins", "\n", "Pearson Stat = ",
pearson, "\n", "p = ", p, "\n \n")
return(results)
}

hoslem(obs = wcgsKM$chd, pred = reduced_model$fitted.values)
```

```
## Hosmer-Lemeshow test with 10 bins
##  Pearson Stat =  176.2338
##  p =  0
##
```

```
##                 obs.o    pred.o   avg mean pearson resid
## [0.0102,0.0225]    24  5.074950 0.01585922     8.4008095
## (0.0225,0.0305]    26  9.114271 0.02761900     5.5931808
## (0.0305,0.0437]    27 11.679206 0.03755372     4.4830605
## (0.0437,0.0496]    28 16.533691 0.04710453     2.8199325
## (0.0496,0.0697]    30 16.399785 0.05836223     3.3583556
## (0.0697,0.0768]    19 22.522083 0.07507361    -0.7421562
## (0.0768,0.0941]    19 26.796523 0.08616245    -1.5061276
## (0.0941,0.123]     31 38.953692 0.11662782    -1.2743668
## (0.123,0.164]      28 43.121131 0.14373710    -2.3027097
## (0.164,0.295]      25 65.804667 0.21717712    -5.0301564
```

The Hosmer Lemeshow test gave us a p-value of 0 which indicates that our model is not a good fit of the data and that a more complex model is needed.

```
autoplot(reduced_model)
```



If we take a look at the residuals and normal plot, we are able to see a separation of the data which indicates that we should aggregate the data before we can proceed with further analysis.

```
# aggregating data
new_dt = wcgsKM[-c(4,6,8,9)]

new_dt[] = lapply(new_dt, as.character)
new_dt$chd = as.numeric(as.character(new_dt$chd))

agg_data = aggregate(chd ~ current.smoker + sbp.cat +
                         chol.cat + age.cat,
                     data = new_dt,
                     FUN = sum)
agg_data = cbind(agg_data,
                 aggregate(chd ~ current.smoker + sbp.cat +
                         chol.cat + age.cat,
                            data = new_dt,
                            FUN = length))

names(agg_data)[10] = "chd_tot"
agg_data = agg_data[, !duplicated(colnames(agg_data))]
head(agg_data)
```

```
##   current.smoker  sbp.cat chol.cat age.cat chd chd_tot
## 1              0 elevated     high  [0,40)   0      19
```

```
## 2              1 elevated     high  [0,40)   1      10
## 3              0     high     high  [0,40)   1      16
## 4              1     high     high  [0,40)   1      20
## 5              0   normal     high  [0,40)   0      17
## 6              1   normal     high  [0,40)   1      18
```

```
# aggregated model
sat_agg_model = glm(chd/chd_tot ~ age.cat + current.smoker + chol.cat + sbp.cat,
                family = binomial(link="logit"), data = agg_data,
                weight = chd_tot)
summary(sat_agg_model)
```

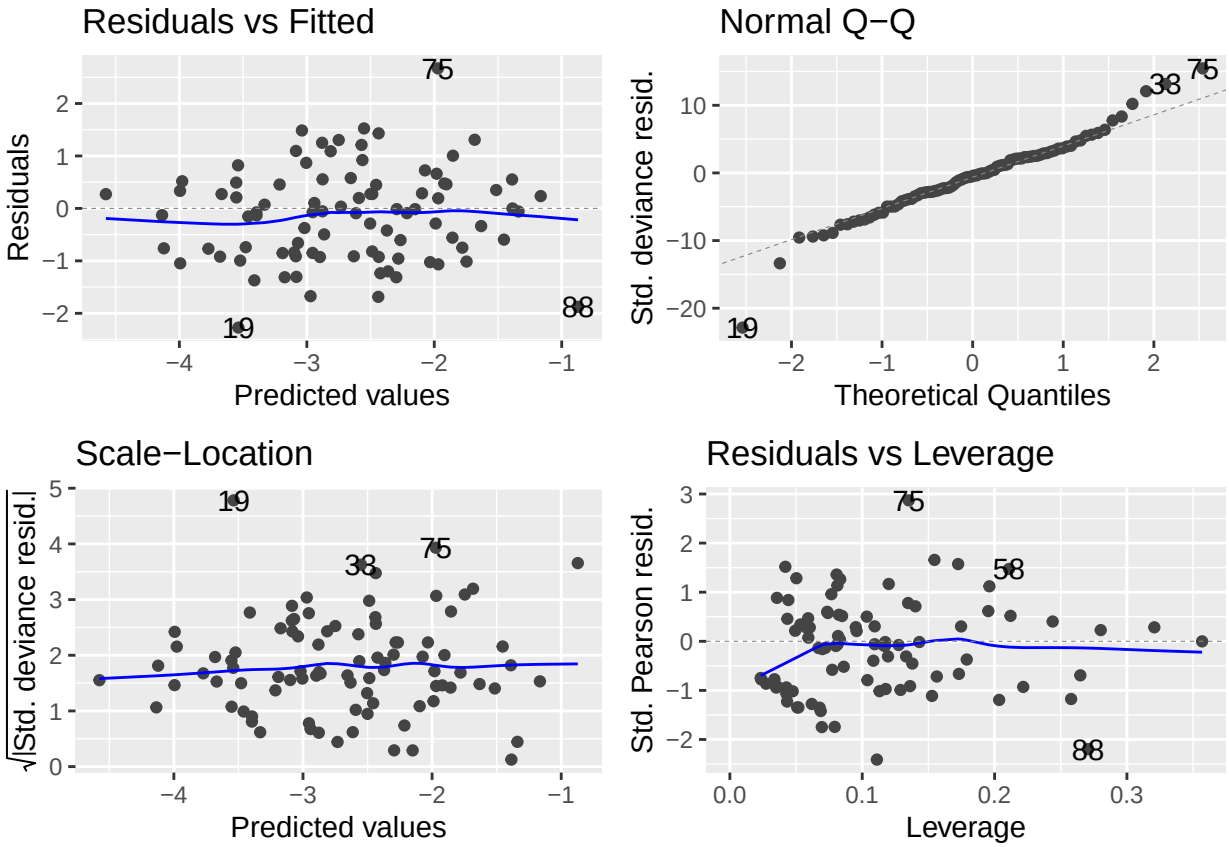
```
##
## Call:
## glm(formula = chd/chd_tot ~ age.cat + current.smoker + chol.cat +
##     sbp.cat, family = binomial(link = "logit"), data = agg_data,
##     weights = chd_tot)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -2.27249  -0.84691  -0.07389   0.45875   2.66971
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)        -3.0819     0.2917 -10.564  < 2e-16 ***
## age.cat[40,45)     -0.4550     0.2802  -1.624 0.104472
## age.cat[45,50)      0.1281     0.2734   0.469 0.639319
## age.cat[50,55)      0.3494     0.2769   1.262 0.206888
## age.cat[55,60)      0.6426     0.2930   2.193 0.028284 *
## current.smoker1     0.5820     0.1372   4.242 2.21e-05 ***
## chol.catnormal     -0.5990     0.2174  -2.756 0.005855 **
## chol.catvery high   0.5183     0.1475   3.513 0.000444 ***
## sbp.cathigh         0.4675     0.1568   2.982 0.002867 **
## sbp.catnormal      -0.4408     0.2243  -1.965 0.049429 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 209.176  on 89  degrees of freedom
## Residual deviance:  72.171  on 80  degrees of freedom
## AIC: 287.57
##
## Number of Fisher Scoring iterations: 4
```

1C

```
pchisq(summary(sat_agg_model)$deviance,
       df = summary(sat_agg_model)$df[2],
       lower.tail = F)
```

```
## [1] 0.721454
```

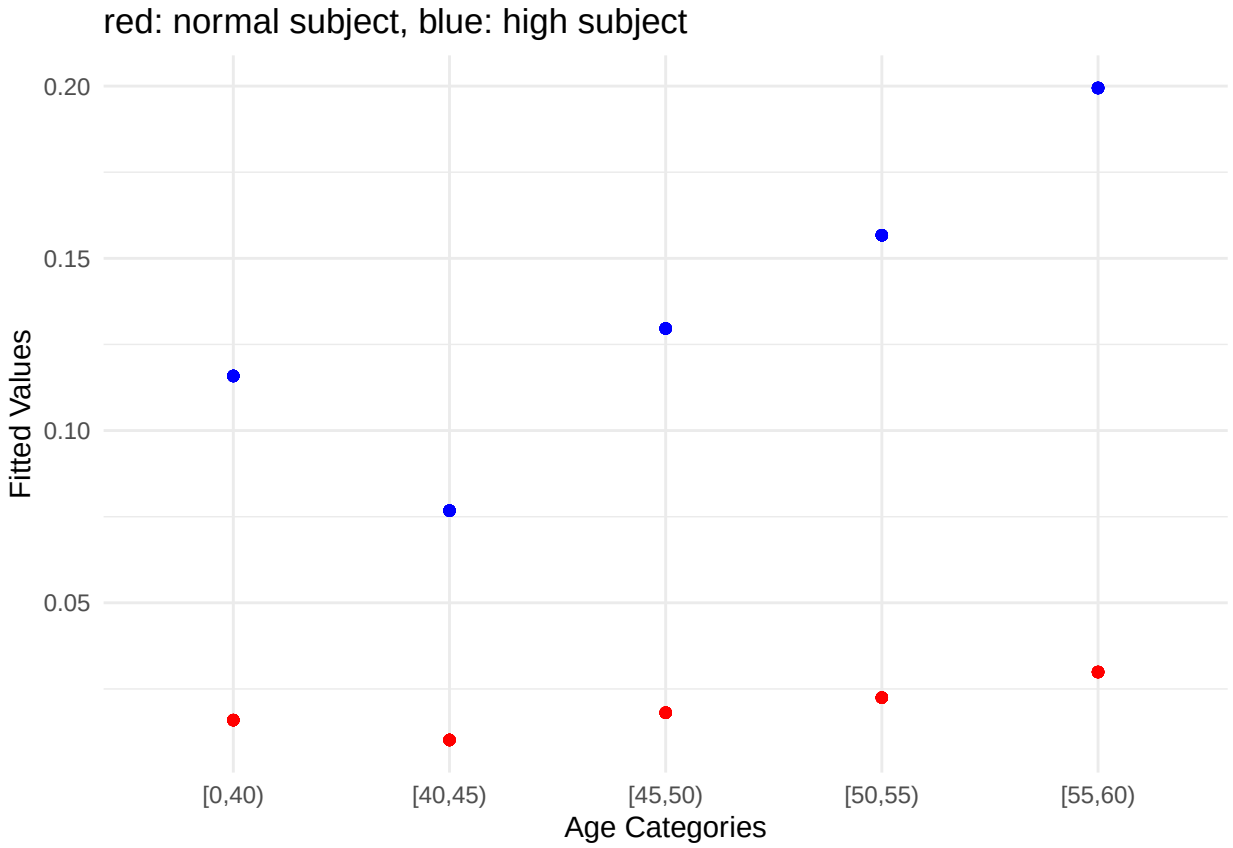
```
autoplot(sat_agg_model)
```



I performed a residual deviance goodness-of-fit test that returned a p value greater than 0.05. This indicates that this is a good model and that there is no evidence that a more complex model is needed. The residuals plot here shows that the data is relatively stable and the Normal Q-Q plot shows that the data is mostly normal and that the ends, because they diverge, indicates some small skewness.

1D

```
norm_subject = data.frame(current.smoker = "0", sbp.cat = "normal",
                          chol.cat = "normal", age.cat = agg_data$age.cat)
high_subject = data.frame(current.smoker = "1", sbp.cat = "high",
                          chol.cat = "high", age.cat = agg_data$age.cat)

norm_prediction = predict(object = sat_agg_model, newdata = norm_subject,
                          type = "response")
high_prediction = predict(object = sat_agg_model, newdata = high_subject,
                          type = "response")

ggplot(agg_data, aes(x = age.cat, y = sat_agg_model$fitted.values)) +
  geom_point(aes(x = age.cat, y = norm_prediction), color = "red") +
  geom_point(aes(x = age.cat, y = high_prediction), color = "blue") +
  labs(y = "Fitted Values", x = "Age Categories") + theme_minimal() +
  ggtitle("red: normal subject, blue: high subject")
```



1E

```
final_table = exp(cbind(sat_agg_model$coefficients,
                    confint(sat_agg_model)))
```

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

```
colnames(final_table) = c("Estimate", "2.5%", "97.5%")
print(final_table)
```

```
##                     Estimate       2.5%      97.5%
## (Intercept)       0.04587343 0.02521045 0.07940289
## age.cat[40,45)    0.63447714 0.37259699 1.12424092
## age.cat[45,50)    1.13671551 0.67800495 1.99143123
## age.cat[50,55)    1.41826843 0.83958437 2.49939004
## age.cat[55,60)    1.90139284 1.08613380 3.44501538
## current.smoker1   1.78957806 1.37033868 2.34772140
## chol.catnormal    0.54933918 0.35388327 0.83216540
## chol.catvery high 1.67914242 1.26022254 2.24876655
## sbp.cathigh       1.59603694 1.17902286 2.18214104
## sbp.catnormal     0.64352461 0.41004906 0.99098694
```

1F

The odds of developing cardiovascular disease among people who smoke is 1.79 (95% CI: 1.37, 2.35) times higher than among those who do not smoke if all other predictors are held constant.

1G

```
releveled_agg = agg_data
releveled_agg$age.cat = factor(releveled_agg$age.cat,
                               levels = c("[45,50)", "[0,40)",
                                          "[40,45)", "[50,55)", "[55,60)"))
releveled_agg$chol.cat = factor(releveled_agg$chol.cat,
                               levels = c("normal", "high", "very high"))
releveled_agg$sbp.cat = factor(releveled_agg$sbp.cat,
                              levels = c("normal", "elevated", "high"))

releveled_agg_reduced_model = glm(formula = chd/chd_tot ~
                                    chol.cat + sbp.cat + current.smoker +
    age.cat, family = binomial(link = "logit"), data = releveled_agg,
    weights = chd_tot)

final_table1 = exp(cbind(releveled_agg_reduced_model$coefficients,
                   confint(releveled_agg_reduced_model)))
```

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

```
colnames(final_table1) = c("Estimate", "2.5%", "97.5%")
print(final_table1)
```

```
##                     Estimate       2.5%      97.5%
## (Intercept)       0.01843396 0.01037052 0.03152094
## chol.cathigh      1.82036899 1.20168418 2.82579054
## chol.catvery high 3.05665878 2.06438512 4.66119324
## sbp.catelevated   1.55394212 1.00909503 2.43873256
## sbp.cathigh       2.48014903 1.68936674 3.74850150
## current.smoker1   1.78957806 1.37033868 2.34772140
## age.cat[0,40)     0.87972760 0.50215141 1.47491548
## age.cat[40,45)    0.55816705 0.38360380 0.80811996
## age.cat[50,55)    1.24768988 0.86921845 1.78880384
## age.cat[55,60)    1.67270775 1.10935308 2.50232508
```

The odds of developing cardiovascular disease among people who are in age category [55,60) is 1.67 (95% CI: 1.11, 2.50) times higher than among those who are in age category [45,50) if the individual has normal levels of cholesterol, normal levels of systolic blood pressure,and is not a smoker.

1H

```
(risk_diff_n = agg_data[agg_data$age.cat == "[45,50)" &
                          agg_data$sbp.cat == "normal" &
                          agg_data$chol.cat == "normal",])
```

```
##    current.smoker sbp.cat chol.cat age.cat chd chd_tot
## 47              0  normal   normal [45,50)   0      30
## 48              1  normal   normal [45,50)   0      29
```

```
table_n = array(c(0,0,30,29), dim = c(2,2))
dimnames(table_n) = list(current.smoker = c("yes", "no"),
                         chd = c("yes", "no"))
print(table_n)
```

```
##               chd
```

```
## current.smoker yes no
##            yes   0 30
##            no    0 29
```

```
prop.table(table_n, 1)
```

```
##               chd
## current.smoker yes no
##            yes   0  1
##            no    0  1
```

```
prop.test(table_n, 1, correct = F)
```

```
## Warning in prop.test(table_n, 1, correct = F): Chi-squared approximation may be
## incorrect
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  table_n
## X-squared = NaN, df = 1, p-value = NA
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0 0
## sample estimates:
## prop 1 prop 2
##      0      0
```

An individual has a 0% likelihood of obtaining coronary heart disease regardless of current smoker status if they are between the ages of 45 and 50 years, with normal blood pressure and cholesterol level. The risk difference for the risk of CHD for a smoker vs a non-smoker, for a subject between the ages of 45 and 50 years, with normal blood pressure and cholesterol level is 0.

```
(risk_diff1 = agg_data[agg_data$age.cat == "[45,50)" &
                         agg_data$sbp.cat == "high" &
                         agg_data$chol.cat == "very high",])
```

```
##    current.smoker sbp.cat  chol.cat age.cat chd chd_tot
## 51              0    high very high [45,50)   5      62
## 52              1    high very high [45,50)  18      90
```

```
table_1 = array(c(18,5,62-5,90-18), dim = c(2,2))
dimnames(table_1) = list(current.smoker = c("yes", "no"),
                         chd = c("yes", "no"))
print(table_1)
```

```
##               chd
## current.smoker yes no
##            yes  18 57
##            no    5 72
```

```
prop.table(table_1, 1)
```

```
##               chd
## current.smoker        yes        no
##            yes 0.24000000 0.7600000
##            no  0.06493506 0.9350649
```

```r
prop.test(table_1, 1, correct = F)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  table_1
## X-squared = 9.0673, df = 1, p-value = 0.002602
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  0.06383716 0.28629271
## sample estimates:
##     prop 1     prop 2
## 0.24000000 0.06493506
```

```r
0.24-0.06493
```

```
## [1] 0.17507
```

An individual has a 24% likelihood of obtaining coronary heart disease if they are a current smoker if they are between the ages of 45 and 50 years, with high blood pressure and very high cholesterol. However, if the same individual was not a current smoker, their likelihood of obtaining coronary heart disease is 6.49%. The risk difference for the risk of CHD for a smoker vs a non-smoker, for a subject between the ages of 45 and 50 years, with high blood pressure and very high cholesterol is 17.5%.

1I

Two different statistical techniques that could be used to make confidence intervals for the risk differences include the Wald Confidence Interval and the Agresti-Caffo Confidence Interval. The Wald CI is just the prop.test of the table and the Agresti-Caffo CI is just adding a count of 1 to each cell of the table and calculating the prop.test.

1J

I performed a simple logistic regression on the wcgsKM data and used a backward selection based on AIC to reduce the model. I checked the residuals and the goodness-of-fit and was inclined to aggregate the data since there was a separation in the residuals and the Hosmer Lemeshow test indicated that our model was not a good fit of the model. I then aggregated the data and repeated the analysis which then gave me a model that was a good fit of the data based off the residual deviance goodness-of-fit test and residuals plot. We find that the odds of developing coronary heart disease among people who smoke is 1.79 (95% CI: 1.37, 2.35) times higher than among those who do not smoke if all other predictors are held constant. However, if the other constants are all normal and within the 45 to 50 years old age range, smoking status does not have an effect on the likelihood of developing coronary heart disease.

2

2A

```r
# aggregating data
new_dt2 = wcgsKM[wcgsKM$chd == 1,]
new_dt2 = new_dt2[-c(4,6,9)]

agg_data2 = aggregate(chd ~ current.smoker + sbp.cat +
                        chol.cat + age.cat,
                      data = new_dt2,
```

```
                     FUN = sum)
agg_data2 = cbind(agg_data2,
                  aggregate(time.years ~ current.smoker + sbp.cat +
                        chol.cat + age.cat,
                           data = new_dt2,
                           FUN = sum))
agg_data2 = agg_data2[-c(6:9)]
head(agg_data2)
```

```
##   current.smoker  sbp.cat chol.cat age.cat chd time.years
## 1              1 elevated     high  [0,40)   1   4.279452
## 2              0     high     high  [0,40)   1   7.887671
## 3              1     high     high  [0,40)   1   3.663014
## 4              1   normal     high  [0,40)   1   4.027397
## 5              0     high   normal  [0,40)   1   5.284932
## 6              1   normal   normal  [0,40)   1   2.917808
```

2B

```
fit_pois = glm(chd ~ current.smoker + sbp.cat + chol.cat + age.cat + offset(log(time.years)), data = agg
summary(fit_pois)
```

```
##
## Call:
## glm(formula = chd ~ current.smoker + sbp.cat + chol.cat + age.cat +
##     offset(log(time.years)), family = poisson, data = agg_data2)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -0.96232  -0.33867  -0.01611   0.32535   1.31625
##
## Coefficients:
##                   Estimate Std. Error z value Pr(>|z|)
## (Intercept)       -1.72981    0.28897  -5.986 2.15e-09 ***
## current.smoker     0.04655    0.13097   0.355    0.722
## sbp.cathigh       -0.04663    0.14958  -0.312    0.755
## sbp.catnormal      0.16134    0.21832   0.739    0.460
## chol.catnormal    -0.02067    0.21456  -0.096    0.923
## chol.catvery high  0.15459    0.14044   1.101    0.271
## age.cat[40,45)     0.03604    0.26780   0.135    0.893
## age.cat[45,50)     0.14946    0.26048   0.574    0.566
## age.cat[50,55)     0.09362    0.26147   0.358    0.720
## age.cat[55,60)     0.28649    0.27858   1.028    0.304
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 20.309  on 70  degrees of freedom
## Residual deviance: 16.428  on 61  degrees of freedom
## AIC: 239.41
##
## Number of Fisher Scoring iterations: 4
```

```
pois_table = exp(cbind(fit_pois$coefficients,
                   confint(fit_pois)))
```

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

```
colnames(pois_table) = c("Estimate", "2.5%", "97.5%")
print(pois_table)
```

```
##                    Estimate       2.5%     97.5%
## (Intercept)       0.1773175 0.09797569 0.3052303
## current.smoker    1.0476554 0.81227087 1.3581343
## sbp.cathigh       0.9544405 0.71554655 1.2874677
## sbp.catnormal     1.1750869 0.75720481 1.7876075
## chol.catnormal    0.9795460 0.63425703 1.4751245
## chol.catvery high 1.1671777 0.88859299 1.5421995
## age.cat[40,45)    1.0366969 0.62487861 1.7966350
## age.cat[45,50)    1.1612026 0.71204209 1.9885011
## age.cat[50,55)    1.0981430 0.67169674 1.8834105
## age.cat[55,60)    1.3317500 0.78377319 2.3509607
```

If you are a current smoker, the rate of catching coronary heart disease is 1.05 (95% CI: 0.81, 1.36) times higher than for a non smoker.

If you have high systolic blood pressure, the rate of catching coronary heart disease is 0.95 (95% CI: 0.72, 1.29) times that of a person who has elevated systolic blood pressure.

If you have normal systolic blood pressure, the rate of catching coronary heart disease is 1.18 (95% CI: 0.76, 1.29) times that of a person who has elevated systolic blood pressure.

If you have normal cholesterol, the rate of catching coronary heart disease is 0.98 (95% CI: 0.63, 1.48) times that of a person who has high cholesterol.

If you have very high cholesterol, the rate of catching coronary heart disease is 1.17 (95% CI: 0.89, 1.54) times that of a person who has high cholesterol.

If you are in the [40,45) age category, the rate of catching coronary heart disease is 1.04 (95% CI: 0.62, 1.80) times higher than someone in age category [0,40).

If you are in the [45,50) age category, the rate of catching coronary heart disease is 1.16 (95% CI: 0.71, 1.99) times higher than someone in age category [0,40).

If you are in the [50,55) age category, the rate of catching coronary heart disease is 1.09 (95% CI: 0.67, 1.88) times higher than someone in age category [0,40).

If you are in the [55,60) age category, the rate of catching coronary heart disease is 1.33 (95% CI: 0.78, 2.35) times higher than someone in age category [0,40).

2C

I would pick the poisson regression if I was a public health official because it is easier to understand and comprehend if you are trying to explain it to the public. There is nothing you have to keep constant which keeps the interpretation simple and easy to understand. With the logistic regression, when you make the interpretations, you have to mention that you are keeping the other variables constant which makes your interpretation very specific to a specific group of people.