2020 Applied Exam Q2

November

(Note: For clarity of the report, the output results are partially presented. R packages, self-defined functions, part of the R results, and figures are shown in the Appendix. Through out the report, I will describe the methods and summarize the findings in the beginning of each section, and include relevant R code and results afterwards. Some hidden R results are summarized in the comment for the related code.)

Part 1

```r
load('data/chol_baseline.RData')
dataset <- data.frame(chol_data_baseline)
# define a binary variable for undesirable
dataset <- within(dataset, {undesirable <- ifelse(chol_level >= 2, 1, 0)})
```

(a)

The study is cross-sectional, and the count in each cell is large enough, so it is natural to use a multinomial distribution to model the response. The null hypothesis is that `undesirable` and `Gender` are independent. Let p_i for $i = 1, 2$ be the probabilities of the two desirableness outcome and p_j for $j = 1, 2$ be the probabilities of the two genders. Let p_{ij} be the probability of a particular joint outcome and y_{ij} be the observed response in cell (i, j). Suppose the total count is n, and let $\hat{\mu}_{ij} = \sum_i y_{ij} \cdot \sum_j y_{ij}/n$ for $i = 1, 2, j = 1, 2$.

The hypotheses can then be specified as: $H_0 : p_{ij} = p_i p_j$ for $\forall i, j$ vs. $H_a : p_{ij} \neq p_i p_j$ for some i, j. Pearson X^2 statistic can be used as the test statistic: $X^2 = \sum_{i,j} \frac{(y_{ij} - \hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$. And I use Yates' continuity correction in the calculation. Under H_0, $X^2 \sim \chi^2_1$ approximately. There are two assumptions for this test: (1) The total number of counts is fixed; (2) The observations are random and independent.

As shown below, the test statistic is 10.799, and the p-value is 0.001015. Also, the 95% CI of the difference in proportions is [-0.4384668 -0.1053928], not including 0. Therefore, we should reject the null hypothesis, and conclude that having an undesirable level of cholesterol and gender are dependent.

```r
d1a <- data.frame(dataset %>% group_by(undesirable, Gender) %>% summarise(y=n()))
ov <- xtabs(y ~ undesirable + Gender, data = d1a)
prop.test(ov)
```

```
...
X-squared = 10.799, df = 1, p-value = 0.001015
alternative hypothesis: two.sided
95 percent confidence interval:
 -0.4384668 -0.1053928
sample estimates:
   prop 1    prop 2
0.4385965 0.7105263
...
```

(b)

For this case, the resulting table has small counts, and hence the Chi-square approximation may not be accurate. Therefore, we should use Fisher's exact test. The null hypothesis is that in center B, `undesirable` and `Gender` are independent. In Fisher's exact test, we can express the hypotheses in terms of odds ratio. Use the same notations as (a), and let OR be the odds ratio $\frac{p_{11}p_{22}}{p_{12}p_{21}}$.

The hypotheses can then be specified as: $H_0 : OR = 1$ vs. $H_a : OR \neq 1$. Unlike most other tests, Fisher's exact test doesn't use a mathematical function to estimate the probability of a test statistic. Instead, it gets the p-value by calculating the probability of observing "as extreme or more extreme" data than the current observed. Not strictly speaking, we may say y_{11} is the test statistic, and it will have a hypergeometric distribution conditioned on the margins under H_0, and we can use such information to calculate the p-value(sum of probabilities of the tables whose odds ratio is in the direction of H_a). There are two assumptions for this test: (1) the row and column totals are fixed; (2) the observations are random and independent.

As shown below, the p-value is 1 and the 95% CI for the odds ratio is [0.1255997 6.3908991], containing 1. Therefore, we should accept the null hypothesis and conclude that in center B, having an undesirable level of cholesterol is independent of gender.

```
d1b <- dataset %>% mutate(center = as.factor(substr(as.character(id), 1, 1))) %>%
  filter(center == 'B') %>% group_by(undesirable, Gender) %>% summarise(y=n())
ov <- xtabs(y ~ undesirable + Gender, data = d1b)
fisher.test(ov)
```

```
...
p-value = 1
alternative hypothesis: true odds ratio is not equal to 1
95 percent confidence interval:
 0.1255997 6.3908991
sample estimates:
odds ratio
 0.9177039
...
```

Part 2

To start with, I fit a large model, including all two-way interaction terms, as well as a quadratic effect of `bmi`. Here, I treat `age_group` as factors to make the model more general. Then I perform model diagnostics by checking the residuals(shown in Figure 1), and check outliers by checking the halfnormal plot of the hatvalues, boxplot, and histogram of `bmi`(shown in Figure 2). Two possible outliers are suggested by the halfnormal plot, but after checking the histogram, I think these two points are not too extreme, and given the relatively large size of the dataset, I decide to keep them in the data.

Next, I compare the model with `age_group` as factors and as numerical, and it suggests that we should use the factorized `age_group`. Then I use AIC to select models and use `drop1()` and `anova()` to manually further eliminate features. The final chosen model is `m4`, with the formula shown in Equation (1) and the estimated coefficients are shown in Result 1 in the Appendix. Model diagnostics for `m4` are shown in Figure 3 and they look fine.

Finally, I check the goodness of fit the selected model: (1) Plot the predicted probability vs. observed proportion(Figure 5a); (2) Perform Hosmer-Lemeshow test and it shows no lack of fit(p-value = 0.7419)(in Result 2); (3) Check the confusion matrix and the misclassification rate is 0.2397661(in Result 2).

$$\log(\frac{p}{1-p}) \sim factor(age\_group) + bmi + bmi^2 + Gender + factor(age\_group) : bmi + bmi : Gender + bmi^2 : Gender \quad (1)$$

```
m0 <- glm(undesirable ~ (factor(age_group) + bmi + I(bmi^2) + Gender)^2, family = binomial, data = dataset)
sumary(m0)
drop1(m0, test = "Chi")
```

```
# check the implied outliers by halfnormal plot
filter(dataset, hatvalues(m0) > 0.55)
```

```
     id Gender      bmi chol_level age_group undesirable
1 B0429      M 18.45952          3         2           1
2 E0457      F 40.30382          2         1           1
```

```
# test factor or numeric age
m1 <- glm(undesirable ~ (age_group + bmi + I(bmi^2) + Gender)^2, family = binomial, data = dataset)
anova(m1, m0, test = "Chi")  # should use factorized age_group
```

```
...
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       160     178.76
```

```
2       156     168.30  4   10.465  0.03328 *
...
```

```
# use AIC to select models
m2 <- step(m0, trace = 0)
anova(m2, m0, test = "Chi")  # compare m2 with m0, shows m2 is good enough
sumary(m2)
drop1(m2, test = "Chi")  # suggests removing bmi:I(bmi^2)
# perform manual selection
m3 <- update(m2, .~.-bmi:I(bmi^2)) # remove bmi:I(bmi^2)
drop1(m3, test = "Chi")  # suggests removing factor(age_group):I(bmi^2)
m4 <- update(m3, .~.-factor(age_group):I(bmi^2)) # remove factor(age_group):I(bmi^2)
```

```
drop1(m4, test = "Chi") # no further elimination need to be done
```

```
...
                      Df Deviance    AIC     LRT Pr(>Chi)
<none>                     178.89 198.89
factor(age_group):bmi  2   189.46 205.46 10.5722 0.005061 **
bmi:Gender             1   183.46 201.46  4.5687 0.032561 *
I(bmi^2):Gender        1   183.28 201.28  4.3927 0.036092 *
...
```

```
anova(m4, m0, test = "Chi") # compare m4 with m0, shows m4 is good enough
```

```
...
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1       161     178.89
2       156     168.30  5   10.592  0.06009 .
...
```

Part 3

To start with, I investigate the relationship between the covariates and the response(shown in Figure 4) to get some basic idea. Then I fit a large model with all two-way interaction terms and treat `age_group` as factors. For better interpretation of the model, I centralize the numerical covariates. As in part 2, I first check whether we should treat `age_group` as factors, and the results show that we can model it as numerical. Then I use AIC to select models and manually check the significance of variables. The final model is `mod3`, with the formula shown in Equation (2), and the estimated coefficients are shown in Result 3. As for model diagnostics, it is not straightforward for multinomial models. To identify outliers or influential points, we can run separate logit models, and as mentioned in part 2, no extreme points are detected. So I would just use all data points in this part. Goodness of fit and model interpretations are done in part 4.

$$\log(\frac{p_i}{p_{desirable}}) \sim I(age\_group-1) + I(bmi-27) + Gender + I(age\_group-1):I(bmi-27) \tag{2}$$

```
# set baseline level
dataset <- within(dataset, {chol_level <- relevel(as.factor(chol_level), ref = 1)})
# fit a large model
mod1 <- multinom(chol_level ~ (factor(age_group) + I(bmi-27) + Gender)^2, data = dataset)
# check numeric or factorized age
mod2 <- multinom(chol_level ~ (I(age_group-1) + I(bmi-27) + Gender)^2, data = dataset)
```

```
model.comp(mod1, mod2) # suggests that it is fine to include age_group as numeric
```

```
[1] 0.6016385
```

```
# AIC model selection
mod3 <- step(mod2, trace = 0)
```

```r
model.comp(mod2, mod3) # suggests m3 is good enough
```

```
[1] 0.6718014
```

```r
# manual model selection
mod4 <- update(mod3, .~.-I(age_group-1):I(bmi-27))
```

```r
model.comp(mod3, mod4) # suggests that we should keep age_group:bmi and choose m3 as the final model
```

```
[1] 0.001004676
```

Part 4

Goodness of Fit

I consider two methods to check the goodness of fit of `mod3`. (1) Use `logitgof{generalhoslem}`(Jay 2019) to perform the Hosmer-Lemeshow goodness of fit test. It compares the observed with the expected frequencies of the outcome and compute a test statistic which is chi-square distributed under null hypothesis. A non-significant p-value would indicate evidence of good fit. A p-value of 0.6425 suggests no lack of fit of `mod3`. (2) Check the confusion matrix and the total misclassification rate is 0.497076, which is not good though. A closer look at each level tells us that the high misclasfication rate may come from level 2(about 87.7% are misclassified). The detailed results are shown in Result 4 in the Appendix.

Summary and Interpretation of the Fitted Model

Model Summary: Let b_2 denote the coefficients for 2, and b_3 denote the coefficients for 3. Then we can write the model equations as follows. The estimated coefficients are shown in Result 3. Moreover, to better understand the fitted model, I (1) calculate the predicted probabilities at each level of `(age_group, Gender)` combination with `bmi` fixed at the median level 26.94376(in Result 5); (2) calculate the averaged predicted probabilities at each level of `(age_group, Gender)` for `bmi` ranging from 18 to 41(in Result 5); (3) plot out how the predicted probabilities change along with `bmi` for each level of `(age_group, Gender)`(Figure 5b). One pattern to note in Figure 5b is that, for level 1(desirable), the predicted probabilities for Males are larger than those for Females in general, while for level 3(high), it is the other way round.

$$\log(\frac{P(chol\_level = borderline\ high)}{P(chol\_level = desirable)}) = b_{20}+b_{21}(age\_group-1)+b_{22}(bmi-27)+b_{23}I(Gender == M)+b_{24}(age\_group-1)*(bmi-27)$$

$$\log(\frac{P(chol\_level = high)}{P(chol\_level = desirable)}) = b_{30}+b_{31}(age\_group-1)+b_{32}(bmi-27)+b_{33}I(Gender == M)+b_{34}(age\_group-1)*(bmi-27)$$

Interpretation of the intercepts: Since I have centralized the numeric covariates, the intercept terms have practical meanings, and they model the probabilities of the cholesterol level identification for an individual with `age_group=1, bmi=27, Gender="F"`, namely, a female younger than 55 with bmi 27. We can see from the calculation below that this individual is most likely to have high cholesterol level.

```r
s <- summary(mod3)
cc <- c(0, c(s$coefficients[, 1]))
exp(cc) / sum(exp(cc))
```

```
                  2         3 
0.1738168 0.3901974 0.4359858 
```

Interpretation of the other coefficients: I would interpret b_2 in detail and the interpreation for b_3 can be done in a similar pattern. Since there is an interaction term in the model, we need to specify the fixed levels of some covariates to interpret the coefficients.

- $e^{b_{21}}$: Odds ratio is estimated as 0.7171535, with 95% CI being [0.4059556, 1.266910]. For an individual with `bmi` being 27 and `Gender` held fixed, a unit increase in `age_group` will increase the odds of having borderline high vs. desirable level by a factor of 0.7171535. Note that this confidence interval contains 1, so the effect is not significant.

- $e^{b_{22}}$: Odds ratio is estimated as 1.285246, with 95% CI being [1.083335, 1.524788]. For an individual with `age_group=1`(under 55) and `Gender` held fixed, a unit increase in `bmi` will increase the odds of having borderline high vs. desirable level by a factor of 1.285246, implying that people(under 55) with higher bmi are at higher relative risk to have borderline high cholesterol level vs. desirable level.
- $e^{b_{23}}$: Odds ratio is estimated as 0.4450899, with 95% CI being [0.19925462, 0.9942306]. With `bmi` and `age_group` held fixed, switching from Female to Male will increase the odds of having borderline high vs. desirable level by a factor of 0.4450899, implying that males tend to have lower relative risk to have borderline high cholesterol level vs. desirable level.
- b_{24}: This interaction effect is estimated as -0.1790500. There are two ways to interpret this interaction effect. (1) It measures the difference between the log-odds ratios corresponding to a unit increase in `age_group` for two `bmi` homogeneous groups which differ by 1 unit, with `Gender` held fixed. (2) It also measures the difference between the log-odds ratios corresponding to a unit increase in `bmi` for two `age_group` homogeneous groups which differ by 1 unit, with `Gender` held fixed. As an illustration, let us consider 4 individuals with the same gender and `(age_group=2, bmi=18)`, `(age_group=2, bmi=19)`, `(age_group=3, bmi=18)`, `(age_group=3, bmi=19)` respectively. Let o_i denote the odds of having borderline high vs. desirable level for individual i. Then we have $\log(\frac{o_2}{o_1}) = b_{22} + b_{24}; \quad \log(\frac{o_4}{o_3}) = b_{22} + 2b_{24}; \Rightarrow b_{24} = \log(\frac{o_4}{o_3}) - \log(\frac{o_2}{o_1}) = \log(\frac{o_4 o_1}{o_3 o_2}) = \log(\frac{o_4}{o_2}) - \log(\frac{o_3}{o_1})$.

```r
# estimate odds ratio
exp(s$coefficients)
```

```
  (Intercept) I(age_group - 1) I(bmi - 27)   GenderM I(age_group - 1):I(bmi - 27)
2    2.244877        0.7171535    1.285246 0.4450899                    0.8360641
3    2.508307        0.9896578    1.293501 0.1277369                    0.7805774
```

```r
# 95% CI for the odds ratio
(ci.lower <- exp(s$coefficients - 1.96 * s$standard.errors))
```

```
  (Intercept) I(age_group - 1) I(bmi - 27)    GenderM I(age_group - 1):I(bmi - 27)
2    1.049911        0.4059556    1.083335 0.19925462                    0.7297765
3    1.161942        0.5499931    1.083187 0.05138896                    0.6774416
```

```r
(ci.upper <- exp(s$coefficients + 1.96 * s$standard.errors))
```

```
  (Intercept) I(age_group - 1) I(bmi - 27)   GenderM I(age_group - 1):I(bmi - 27)
2    4.799904         1.266910    1.524788 0.9942306                    0.9578318
3    5.414731         1.780791    1.544650 0.3175139                    0.8994150
```

Part 5

In this case, the individuals inside the same center have some form of group structure, and so analyses that assume independence of the observations will be inappropriate. Also, in this scenario, we care less about the effect of specific levels of `center`, but more about the entire population of centers. To model such a within-center grouping structure, we can consider mixed effect models, treating `center` as the random effects and the other covariates as fixed effects. We can hence fit generalized linear mixed models to generalize the work in part 2. Suppose we have Y_{ij} as the response of the jth individual in the ith center for $i = 1, \cdots, 200, j = 1, \cdots, n_i$, which takes the values zero or one with $P(Y_{ij}) = p_{ij}$. The link function is logit link, $\eta_{ij} = \log(\frac{p_{ij}}{1-p_{ij}})$. Let $\eta_i = (\eta_{i1}, \cdots, \eta_{in_i})^T$, β_i denote the fixed effects, γ_i denote the random effects, X_i denote the design matrix for the fixed effects, and Z_i denote the design matrix for the random effects. Assume $\gamma_i \sim N(0, D)$ $i.i.d.$ and D is some positive semi-definite symmetric matrix. Conditional on the random effects, the model is specified as $\eta_i = X_i\beta + Z_i\gamma_i$. As an illustration, if we consider random slope of `bmi` and the random intercept, then we can start with `m0.glmm` as shown below and do model selection thereafter.(suppose the dataset is named as `newdataset`.)

```r
# do not evalute the following
m0.glmm <- lme4::glmer(undesirable~(factor(age_group) + bmi +I(bmi^2) + Gender)^2 + (bmi|center),
                       family = binomial, data = newdataset)
```

References

Jay, Matthew. 2019. *Generalhoslem: Goodness of Fit Tests for Logistic Regression Models*. https://CRAN.R-project.org/package=generalhoslem.

Appendix

Packages

All R packages and self-defined functions used in this project are listed below.

```
library(dplyr)
library(faraway)
library(nnet)
library(ggplot2)
library(reshape2)
library(generalhoslem)
# a self-defined function using deviance to compare models
model.comp <- function(mod, modr) {
  deviance.diff <- deviance(modr) - deviance(mod)
  pchisq(deviance.diff, mod$edf - modr$edf, lower = F)
}
```

R Results

Result 1

```
# summary of m4 in part 2
sumary(m4)
```

```
                         Estimate  Std. Error z value  Pr(>|z|)
(Intercept)            -29.0648703   8.0483664 -3.6113 0.0003047
factor(age_group)2       7.9943699   3.0045303  2.6608 0.0077962
factor(age_group)3       7.9426441   3.4990296  2.2700 0.0232102
bmi                      1.9477224   0.5584828  3.4875 0.0004875
I(bmi^2)                -0.0292181   0.0095396 -3.0628 0.0021927
GenderM                 28.3881601  14.4697878  1.9619 0.0497751
factor(age_group)2:bmi  -0.2942543   0.1125796 -2.6137 0.0089556
factor(age_group)3:bmi  -0.3167667   0.1265159 -2.5038 0.0122878
bmi:GenderM             -2.1189331   1.0317128 -2.0538 0.0399949
I(bmi^2):GenderM         0.0365592   0.0180970  2.0202 0.0433652

n = 171 p = 10
Deviance = 178.88845 Null Deviance = 217.68785 (Difference = 38.79940)
```

Result 2

```
# Goodness of Fit for m4 in part 2: Hosmer-Lemeshow test
logitgof(dataset$undesirable, fitted(m4))
```

```

    Hosmer and Lemeshow test (binary model)

data:  dataset$undesirable, fitted(m4)
X-squared = 5.1456, df = 8, p-value = 0.7419
```

```
# Goodness of Fit for m4 in part 2: confusion matrix
preddf <- dataset %>% mutate(predprob=predict(m4, type="response"), predout=ifelse(predprob < 0.5, 0, 1))
(t <- xtabs( ~ undesirable + predout, preddf))
```

```
           predout
undesirable   0   1
```

```
          0  29  28
          1  13 101
```

```
# misclassification rate
(t[1, 2] + t[2, 1])/(t[1, 1] + t[1, 2] + t[2, 1] + t[2, 2])
```

```
[1] 0.2397661
```

Result 3

```
# summary of mod3 in part 3
summary(mod3)
```

```
...
Coefficients:
  (Intercept) I(age_group - 1) I(bmi - 27)   GenderM I(age_group - 1):I(bmi - 27)
2   0.8086508      -0.33246543   0.2509499 -0.809479                   -0.1790500
3   0.9196079      -0.01039604   0.2573525 -2.057783                   -0.2477214

Std. Errors:
  (Intercept) I(age_group - 1) I(bmi - 27)   GenderM I(age_group - 1):I(bmi - 27)
2   0.3877270        0.2903297  0.08719683 0.4100474                   0.06937091
3   0.3926099        0.2997211  0.09053324 0.4645659                   0.07230131

Residual Deviance: 337.8136
AIC: 357.8136
...
```

Result 4

```
# Goodness of Fit for mod3 in part 3: Hosmer-Lemeshow test
logitgof(dataset$chol_level, fitted(mod3))
```

```

    Hosmer and Lemeshow test (multinomial model)

data:  dataset$chol_level, fitted(mod3)
X-squared = 13.412, df = 16, p-value = 0.6425
```

```
# Goodness of Fit for mod3 in part 3: confusion matrix
(t <- xtabs( ~ predict(mod3) + dataset$chol_level))
```

```
             dataset$chol_level
predict(mod3)  1  2  3
            1 38 22 12
            2  5  7  4
            3 14 28 41
```

```
1 - (t[1, 1] + t[2, 2] + t[3, 3]) / nrow(dataset) # total misclassification rate
```

```
[1] 0.497076
```

```
1 - c(t[1, 1], t[2, 2], t[3, 3]) / colSums(t) # misclassification rate for each level
```

```
        1         2         3 
0.3333333 0.8771930 0.2807018 
```

Result 5

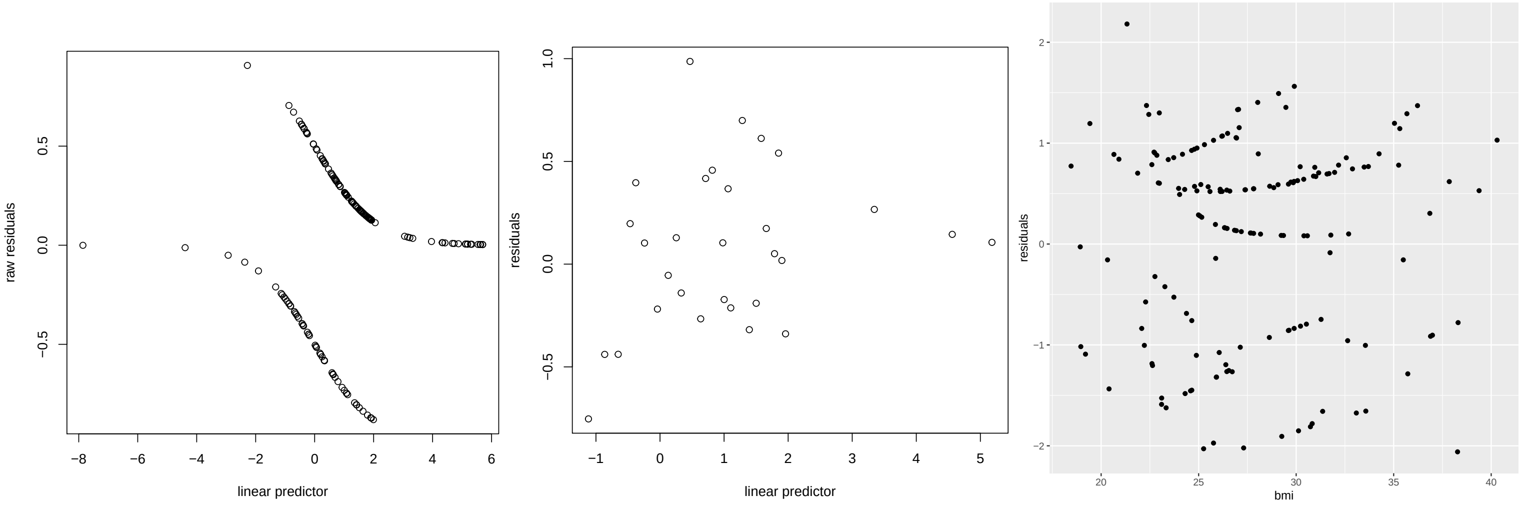
```
# at median bmi level, the predicted probabilities for all (age_group, Gender) combinations
df.ag <- data.frame(age_group=rep(c(1, 2, 3), each = 2), Gender = rep(c("F", "M"), 3),
                    bmi = median(dataset$bmi))
pp.ag <- cbind(df.ag, predict(mod3, newdata = df.ag, type = "probs"))
names(pp.ag)[4:6] <- c('p1', 'p2', 'p3')
pp.ag
```

```
  age_group Gender      bmi        p1        p2        p3
1         1      F 26.94376 0.1758804 0.3892967 0.4348229
2         1      M 26.94376 0.4345994 0.4281543 0.1372463
3         2      F 26.94376 0.1966783 0.3153588 0.4879629
4         2      M 26.94376 0.4924686 0.3514592 0.1560721
5         3      F 26.94376 0.2149915 0.2497210 0.5352875
6         3      M 26.94376 0.5449503 0.2817335 0.1733162
```

```
# mean probabilities within each level of (age_group, Gender)
df.bmi <- data.frame(age_group=rep(rep(c(1, 2, 3), each = 2), 40), Gender = rep(c("F", "M"), 3 * 40),
                     bmi = rep(seq(18, 41, length.out = 40), each = 6))
pp.bmi <- cbind(df.bmi, predict(mod3, newdata = df.bmi, type = "probs"))
names(pp.bmi)[4:6] <- c("prob1", "prob2", "prob3")
data.frame(pp.bmi %>% group_by(age_group,Gender) %>% summarise(p1=mean(prob1),p2=mean(prob2),p3=mean(prob3)))
```

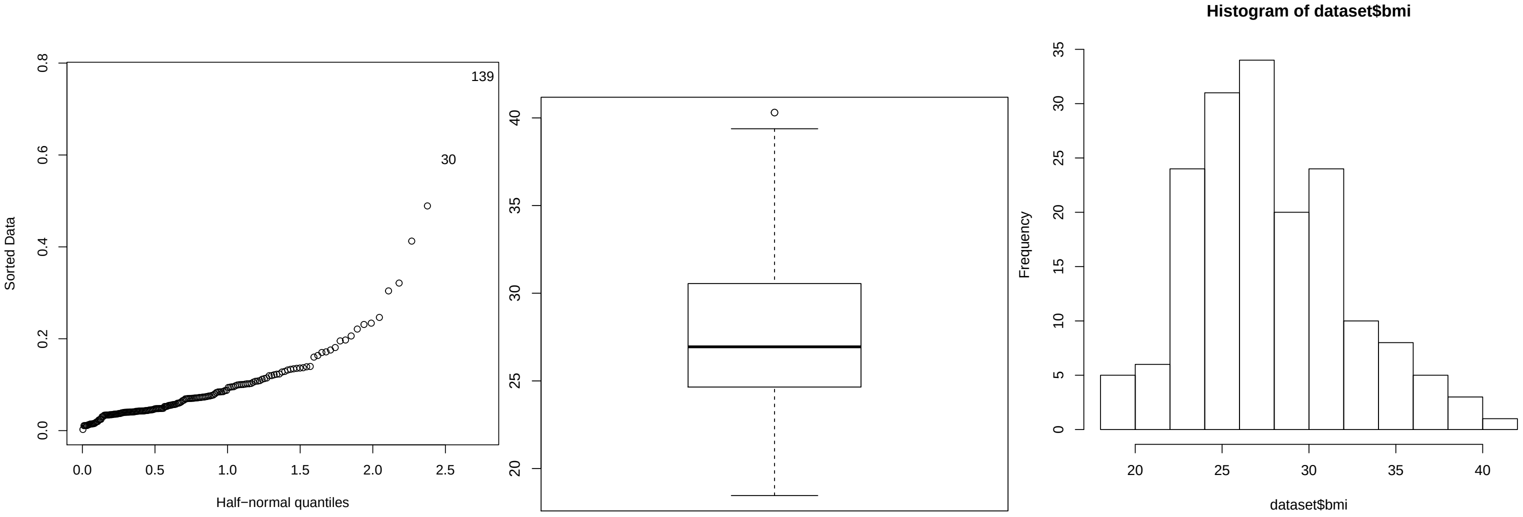
```
  age_group Gender        p1        p2        p3
1         1      F 0.1948176 0.3750107 0.4301717
2         1      M 0.3639868 0.4773375 0.1586757
3         2      F 0.1825091 0.3584708 0.4590201
4         2      M 0.4554066 0.3982732 0.1463201
5         3      F 0.3428579 0.2254903 0.4316519
6         3      M 0.5992108 0.2284348 0.1723544
```

Figures



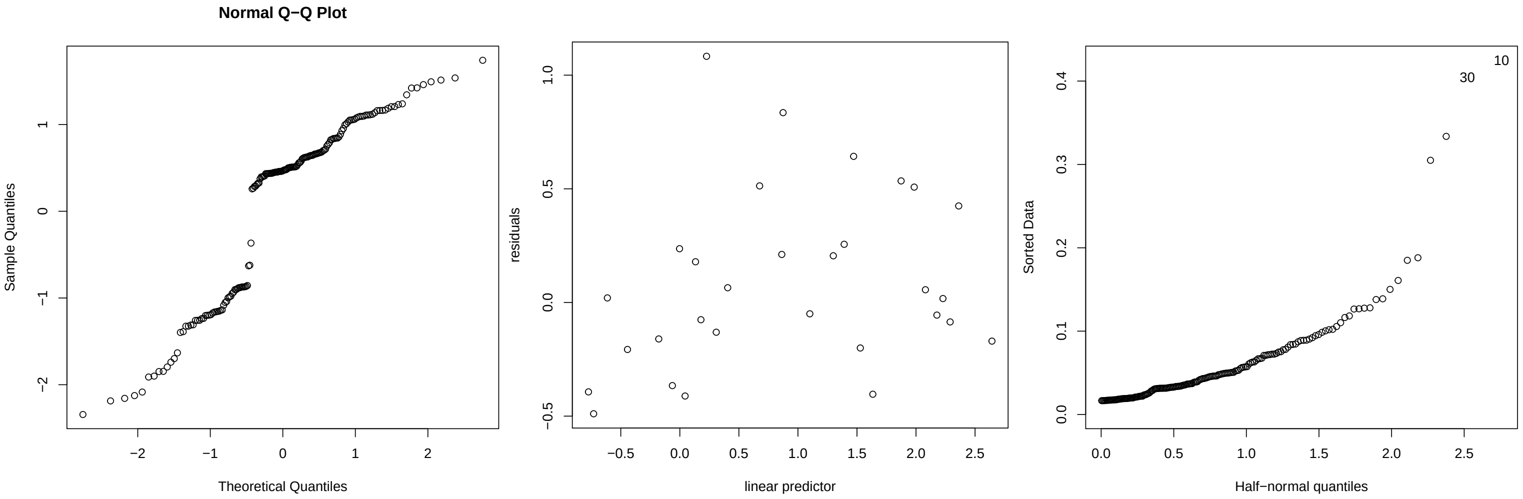
(a) raw residual vs. linpred (b) binned residual vs. linpred (c) residuals vs. bmi

Figure 1: m0 Model Diagnostics



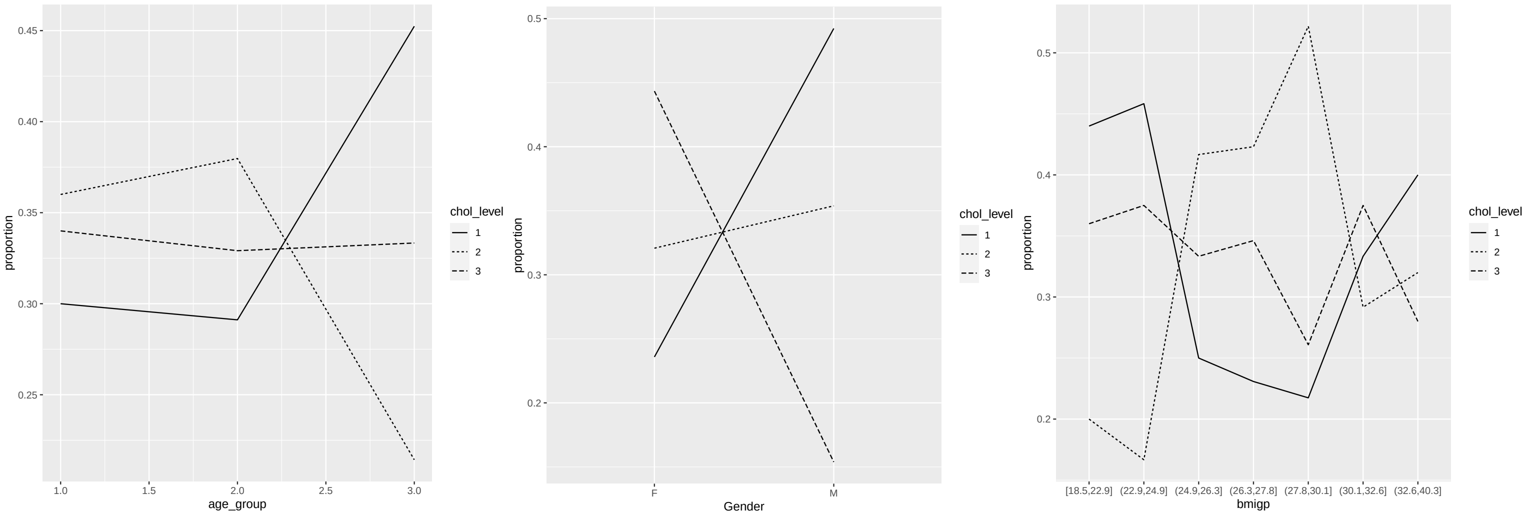
(a) halfnorm plot of hatvalues (b) boxplot of bmi (c) histogram of bmi

Figure 2: m0 Outlier Analysis



(a) Q-Q plot for residuals (b) binned residual vs. linpred (c) halfnormal plot of hatvalues

Figure 3: m4 Model Diagnostics

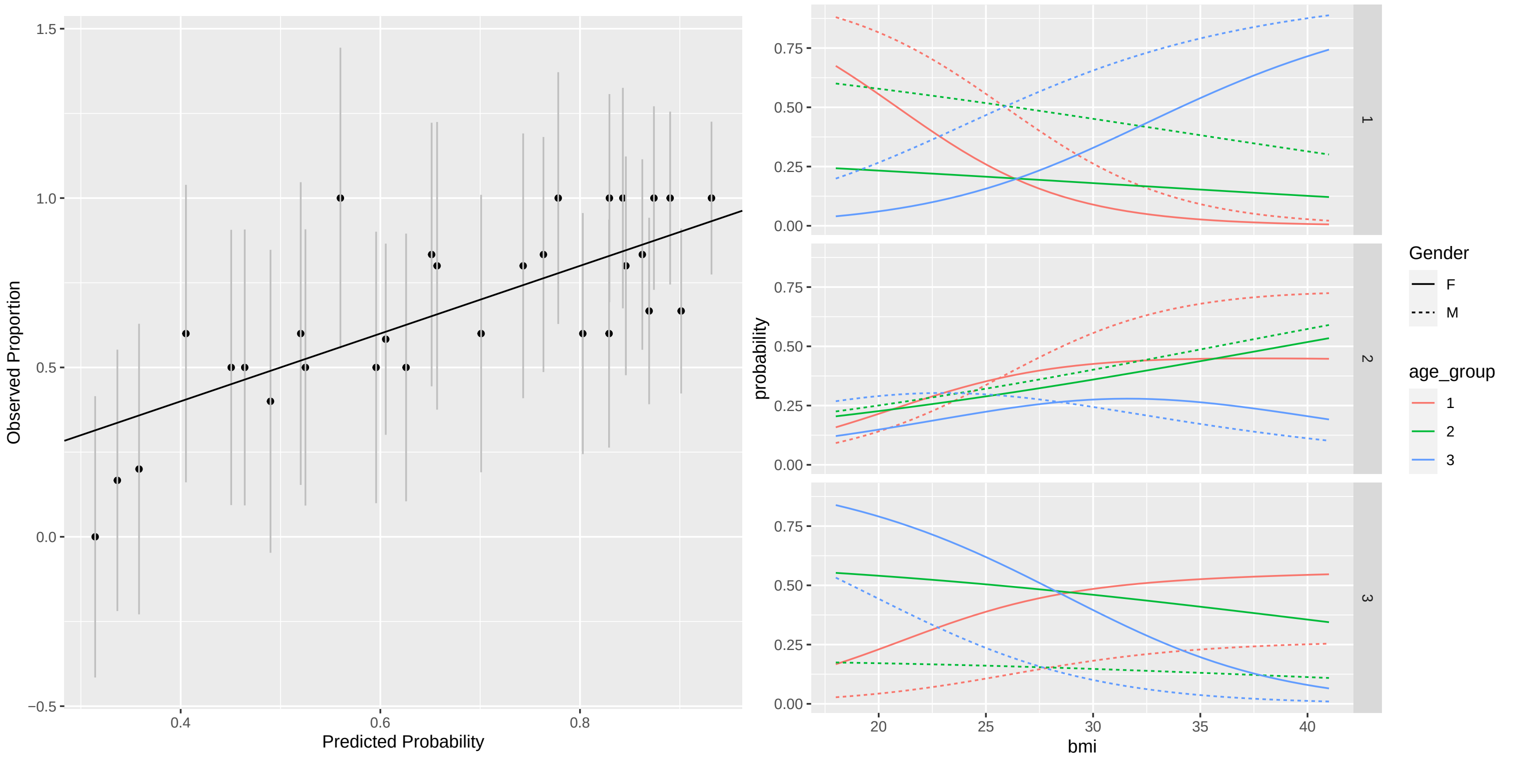


(a) age vs. chol level

(b) Gender vs. chol level

(c) bmi vs. chol level

Figure 4: Data Exploration for part 3



(a) m4 Probability Calibration Curve

(b) Predicted Probabilities by mod3 vs. bmi

Figure 5: m4 Probability Calibration Plot and mod3 Summary Plot