STAT 410 Final Exam

Charles Hwang

7 May 2022

This is a take-home exam. Submit this R Markdown pdf file and html file it compiles that includes your solutions -code, output, explanation/conclusions etc. Provide as much details and explanation as possible. If you ran multiple models/testing things out please show that and describe your process/choices. Make sure that your work is done by *you only and is completely original*. Anyone caught cheating will receive an automatic zero in the exam. Do not cheat.. Seriously please don't do it.

Problem Breakdown

- 310 Students must complete the first 5 questions (1-5) (each question is worth 20pts out of 100pts)
- 410 Students must complete all the questions (each question is worth 20pts out of 120 pts)

Data Description

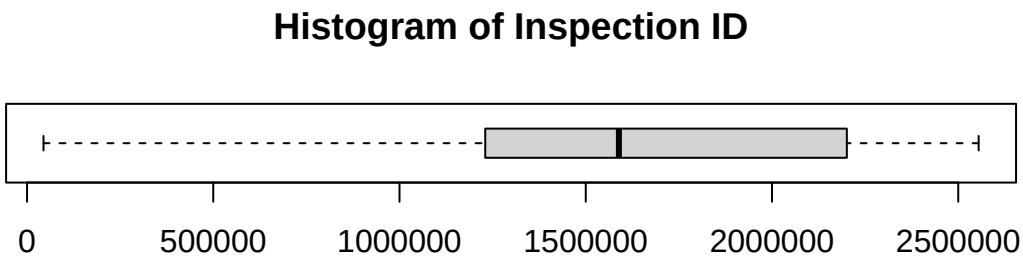
This data set comes from the City of Chicago Data Portal and is about food safety inspections throughout the city from 2010 to present. In the original file there are many variable and more information about them can be found here. I cleaned the dataset to its current form, found in data folder (inspections_clean.csv). Please note information about the following variables.

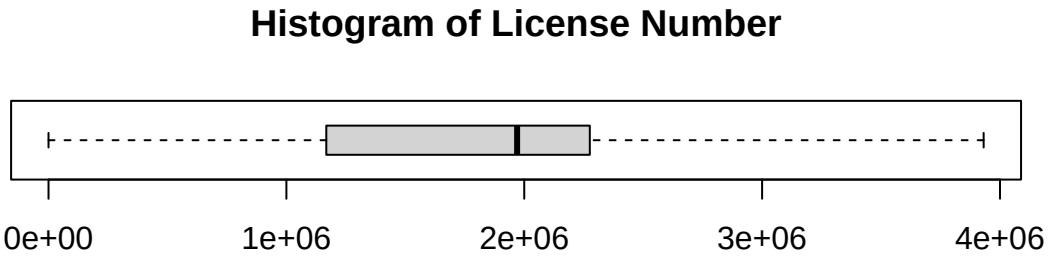
- `license_number` - a unique license number that corresponds to the business/establishment. This is different than `inspection_id` which is unique for each inspection.
- `facility_type` - this is what type of facility the business/establishment is. I cleaned to data to be only 9 categories with the most common being restaurant. For more information of the cleaning please see data_cleaning.R
- `risk` - a categorization made from Dept. of Public Health to assess the risk of adversely affecting the public's health, with 1 being the highest and 3 the lowest. Please see data description for more information.
- `inspection_type` - there are 3 categories that I cleaned the data to include; canvas, complaint, and license. Please see data description for more information.
- `results` - I cleaned this response to include three options: pass, conditional pass, and fail. The original had others but I selected these three as the only options.
- `violations` - a string vector of violations that were present at the inspection. Could be beyond the scope of the final but I thought you might find this interesting.

Please take some time to examine the dataset. There should be 127,903 observations with 11 variables in the original data. For you own final analysis, I would like you to work with a subsample of the data that will be specific to you. This will create a unique sample (n = 50,000) for each student and highlight your interpretations of your analysis. Please set a unique seed, like your birthday or last 4 of your phone number, etc. You will use this for the rest of the exam.

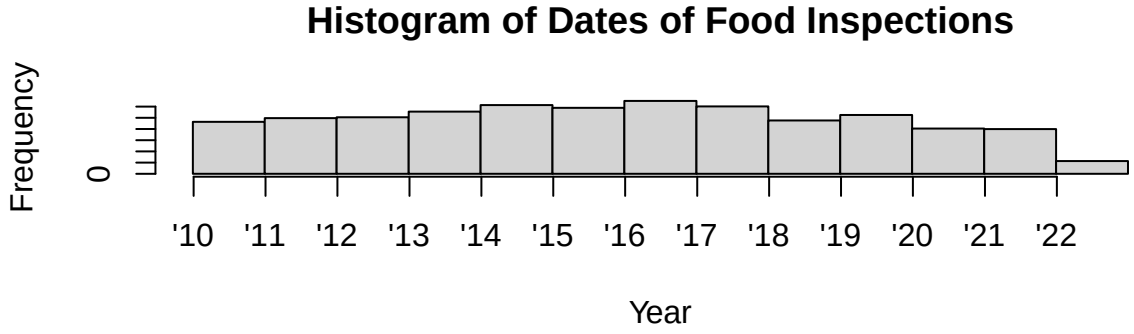
```
library(tidyverse)
inspection<-read_csv("/Users/newuser/Desktop/Notes/Graduate/STAT 410 - Categorical Data Analysis/final/
### data explore here
par(mfrow=c(2,1))
```

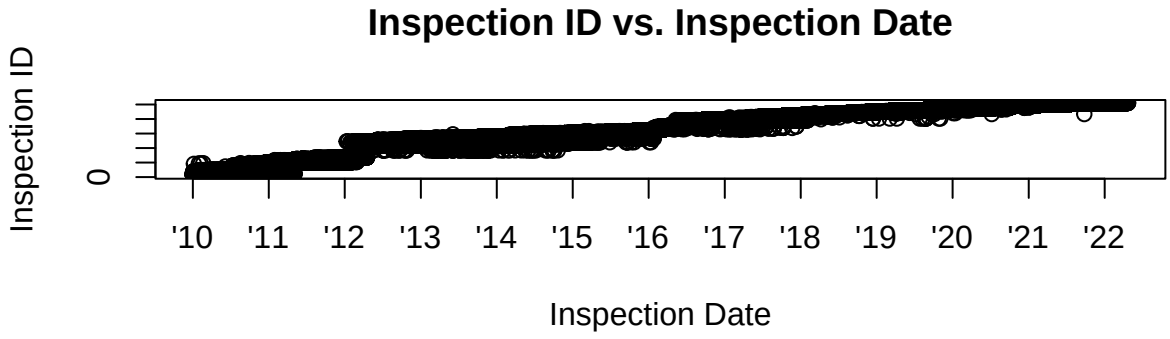
```
boxplot(inspection$inspection_id,main="Histogram of Inspection ID",horizontal=TRUE)
boxplot(inspection$license_number,main="Histogram of License Number",horizontal=TRUE)
```






```
hist(inspection$inspection_date,breaks="years",xaxt="n",xlab="Year",freq=TRUE,main="Histogram of Dates
axis.Date(1,at=seq(min(inspection$inspection_date),max(inspection$inspection_date),by="1 year"),format=
plot(inspection$inspection_date,inspection$inspection_id,xaxt="n",xlab="Inspection Date",ylab="Inspecti
axis.Date(1,at=seq(min(inspection$inspection_date),max(inspection$inspection_date),by="1 year"),format=
```





```r
summary(inspection$inspection_id)
length(unique(inspection$dba_name))
head(sort(table(inspection$license_number),decreasing=TRUE))
table(inspection$facility_type)
table(inspection$risk)
head(sort(table(inspection$address),decreasing=TRUE),36) # Midway Airport, Union Station, shopping cent
table(inspection$zip)
table(inspection$inspection_type)
table(inspection$results)
length(unique(inspection$violations)) # Over 86 percent of recorded entries in violations column are un
table(inspection$risk,inspection$results) # No apparent difference in probability of passing between ri
table(inspection$inspection_type,inspection$results)
### my random subsample
set.seed(7522,sample.kind="Rounding")
inspect_sub<-inspection[sample(nrow(inspection),50000),]
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   44247 1230168 1588972 1616234 2200962 2554825
## [1] 12325
##
##       0   25152   39623   60184 1095992   22811
##     124      45      41      41      41      39
##
##         bakery       catering    gas station  grocery store       hospital
##           1979           1698            102          14731            440
##         liquor long term care     restaurant         school
##            456           1309          89380          17808
##
##   Risk 1 (High) Risk 2 (Medium)    Risk 3 (Low)
##          100011           21665            6226
##
##          5700 S CICERO AVE           500 W MADISON ST
##                        493                        324
##           2300 S THROOP ST          7601 S CICERO AVE
##                        254                        213
##          100 W RANDOLPH ST           131 N CLINTON ST
##                        203                        201
## 222 W MERCHANDISE MART PLZ            600 E GRAND AVE
##                        159                        157
##            700 E GRAND AVE       2002 S WENTWORTH AVE
##                        142                        136
##             225 S CANAL ST           324 N LEAVITT ST
##                        132                        130
##         233 N MICHIGAN AVE         520 N MICHIGAN AVE
##                        128                        128
##            233 S WACKER DR           800 N KEDZIE AVE
##                        126                        120
##          200 E RANDOLPH ST             251 E HURON ST
##                        101                        100
##         225 N MICHIGAN AVE         175 W JACKSON BLVD
##                         99                         95
##         875 N MICHIGAN AVE            311 S WACKER DR
##                         85                         83
##         900 N MICHIGAN AVE             108 N STATE ST
```

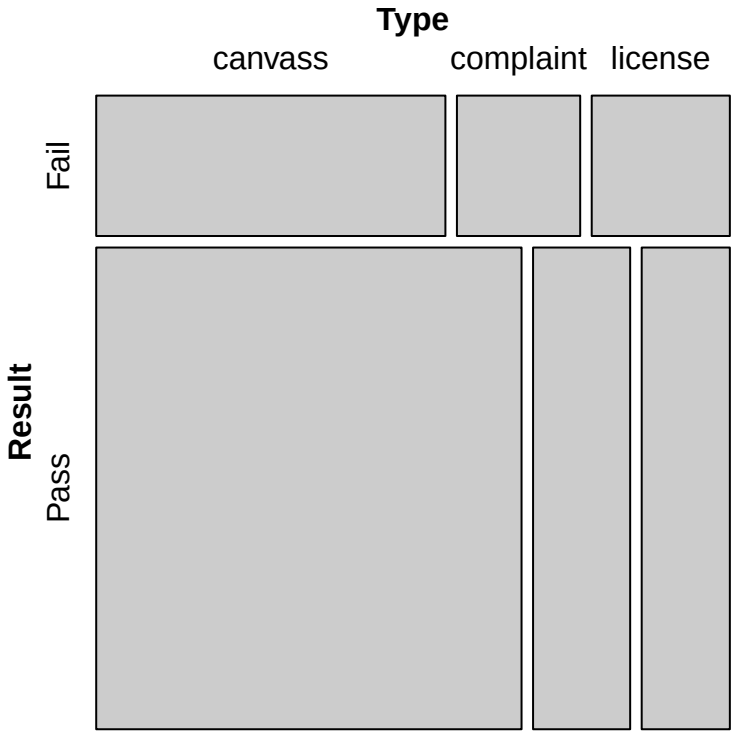
```
##                          82                          81
##       5401 S WENTWORTH AVE            1101 S CANAL ST
##                          79                          76
##             2101 E 71ST ST           2637 S THROOP ST
##                          73                          72
##          130 E RANDOLPH ST              100 W 87TH ST
##                          69                          64
##         216 W JACKSON BLVD           131 N Clinton ST
##                          64                          60
##            151 E WACKER DR             55 E GRAND AVE
##                          59                          58
##            333 E BENTON PL            100 E WALTON ST
##                          55                          53
##
## 60007 60018 60076 60077 60126 60153 60193 60201 60409 60429 60482 60501 60601
##     8     3     2     4     5     6    17     9     4     3     5     3  2098
## 60602 60603 60604 60605 60606 60607 60608 60609 60610 60611 60612 60613 60614
##   991  1345   923  1934  2253  4063  3674  2998  2369  3593  1803  2932  5002
## 60615 60616 60617 60618 60619 60620 60621 60622 60623 60624 60625 60626 60628
##  1732  3019  2504  3882  2617  2677  1263  4591  3033  1201  3347  1951  2587
## 60629 60630 60631 60632 60633 60634 60636 60637 60638 60639 60640 60641 60642
##  2817  1751  1113  2948   238  1864  1448  1465  2055  3110  3391  2112  1653
## 60643 60644 60645 60646 60647 60649 60651 60652 60653 60654 60655 60656 60657
##  2244  1596   772  1022  4316  1559  1682  1331  1129  2876   651   542  4441
## 60659 60660 60661 60666 60707 60714 60804 60827
##  2507  1902  2012    42   714     7     3   132
##
##   canvass complaint   license
##     85522     21645     20736
##
##               Fail               Pass Pass w/ Conditions
##              28854              76038              23011
## [1] 110023
##
##                    Fail  Pass Pass w/ Conditions
##   Risk 1 (High)   21785 59894              18332
##   Risk 2 (Medium)  5110 12609               3946
##   Risk 3 (Low)     1958  3535                733
##
##              Fail  Pass Pass w/ Conditions
##   canvass   16629 52097              16796
##   complaint  5847 11369               4429
##   license    6378 12572               1786
```

1. Please perform a Chi-squared test of independence between `result` and `inspection_type`. You will need to recategorize `result` to pass and fail (conditional pass is a pass). Interpret your results (include a residual plot). What is this test asking? Why would you be performing this test?

```
inspect_sub[inspect_sub$results=="Pass w/ Conditions","results"]<-"Pass"
Result<-inspect_sub$results
Type<-inspect_sub$inspection_type
table(Result,Type)
chisq.test(table(Result,Type))
library(vcd)
```

```
## Loading required package: grid
```

```
mosaic(xtabs(~Result+Type))
```




```
##       Type
## Result canvass complaint license
##   Fail    6453      2278    2553
##   Pass   26969      6156    5591
##
##  Pearson's Chi-squared test
##
## data:  table(Result, Type)
## X-squared = 657.81, df = 2, p-value < 2.2e-16
```

Answer 1:

H_0 : There is *no* association between type of inspection and result

H_A : There is *an* association between type of inspection and result

We reject H_0 at the $\alpha = 0.05$ level. There is strong statistical evidence ($\chi_2 = 657.8071383$, $p = 1.4420982 \times 10^{-143}$) that there is an association between type of inspection and result. The purpose of performing this test is to determine whether type of inspection and result may be related to one another to guide additional analysis.

2. Using the binary variable created in question 1 for result of the inspection (pass/fail), create a model using the variables `facility_type`, `risk`, `zip`, and `inspection_type`. Please describe your process of model selection (Chapter 5 tools) and interpret your final model result (at least one β). Some kind of LRT should be performed.

- Comment on whether or not you factorized certain variables and why.

```
inspect_sub$results<-factor(inspect_sub$results,labels=0:1)
inspect_sub$facility_type<-as.factor(inspect_sub$facility_type)
```

```r
inspect_sub$risk<-as.factor(inspect_sub$risk)
inspect_sub$zip<-as.factor(inspect_sub$zip)
inspect_sub$inspection_type<-as.factor(inspect_sub$inspection_type)
f<-glm(results~facility_type,family=binomial,data=inspect_sub)                          # Step 1
r<-glm(results~risk,family=binomial,data=inspect_sub)
z<-glm(results~zip,family=binomial,data=inspect_sub)
i<-glm(results~inspection_type,family=binomial,data=inspect_sub)
1-pchisq(f$null.deviance-f$deviance,f$df.null-f$df.residual)                    # ***
1-pchisq(r$null.deviance-r$deviance,r$df.null-r$df.residual)                    # ***
1-pchisq(z$null.deviance-z$deviance,z$df.null-z$df.residual)                    # ***
1-pchisq(i$null.deviance-i$deviance,i$df.null-i$df.residual)                    # ***
fri<-glm(results~facility_type+risk+inspection_type,family=binomial,data=inspect_sub)   # Step 2
1-pchisq(fri$null.deviance-fri$deviance,fri$df.null-fri$df.residual)            # ***
fxri<-glm(results~facility_type*risk+inspection_type,family=binomial,data=inspect_sub)  # Step 3
fxir<-glm(results~facility_type*inspection_type+risk,family=binomial,data=inspect_sub)
frxi<-glm(results~facility_type+risk*inspection_type,family=binomial,data=inspect_sub)
1-pchisq(fri$deviance-fxri$deviance,fri$df.residual-fxri$df.residual)            # ***
1-pchisq(fri$deviance-fxir$deviance,fri$df.residual-fxir$df.residual)            # ***
1-pchisq(fri$deviance-frxi$deviance,fri$df.residual-frxi$df.residual)            # ***
fxrfxirxi<-glm(results~facility_type*risk+facility_type*inspection_type+risk*inspection_type,family=bino
1-pchisq(fri$deviance-fxrfxirxi$deviance,fri$df.residual-fxrfxirxi$df.residual) # ***
fxrxi<-glm(results~facility_type*risk*inspection_type,family=binomial,data=inspect_sub) # Step 5
```

```
## Warning: glm.fit: algorithm did not converge
```

```
## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```r
1-pchisq(fri$deviance-fxrxi$deviance,fri$df.residual-fxrxi$df.residual)         # " "
summary(step(fxrfxirxi,direction="both",trace=0))                               # Step 6
```

```
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 0
## [1] 4.403045e-09
## [1] 3.01922e-07
## [1] 5.202962e-07
## [1] 0
## [1] 1
##
## Call:
## glm(formula = results ~ facility_type * risk + facility_type *
##     inspection_type + risk * inspection_type, family = binomial,
##     data = inspect_sub)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.3272   0.5763   0.6495   0.7354   1.3753
##
## Coefficients: (5 not defined because of singularities)
##                                                 Estimate Std. Error
## (Intercept)                                     1.099329   0.127412
## facility_typecatering                           0.421705   0.207997
## facility_typegas station                       -9.529730  83.748838
```

```
## facility_typegrocery store                                  0.258932   0.140963
## facility_typehospital                                       0.569198   0.260593
## facility_typeliquor                                       -11.672508 119.468788
## facility_typelong term care                                 0.054290   0.167270
## facility_typerestaurant                                     0.349661   0.128400
## facility_typeschool                                         0.340937   0.132594
## riskRisk 2 (Medium)                                        -0.034835   0.163968
## riskRisk 3 (Low)                                            9.687640  83.746170
## inspection_typecomplaint                                   -0.352926   0.248010
## inspection_typelicense                                     -0.426612   0.248852
## facility_typecatering:riskRisk 2 (Medium)                   0.088373   0.294895
## facility_typegas station:riskRisk 2 (Medium)                9.376664  83.749288
## facility_typegrocery store:riskRisk 2 (Medium)             -0.223239   0.179069
## facility_typehospital:riskRisk 2 (Medium)                         NA         NA
## facility_typeliquor:riskRisk 2 (Medium)                    10.834565 119.469452
## facility_typelong term care:riskRisk 2 (Medium)                   NA         NA
## facility_typerestaurant:riskRisk 2 (Medium)                 0.297192   0.168214
## facility_typeschool:riskRisk 2 (Medium)                    -0.171363   0.217939
## facility_typecatering:riskRisk 3 (Low)                     -9.584929  83.746609
## facility_typegas station:riskRisk 3 (Low)                         NA         NA
## facility_typegrocery store:riskRisk 3 (Low)               -10.018582  83.746180
## facility_typehospital:riskRisk 3 (Low)                            NA         NA
## facility_typeliquor:riskRisk 3 (Low)                       1.018491 145.897497
## facility_typelong term care:riskRisk 3 (Low)                      NA         NA
## facility_typerestaurant:riskRisk 3 (Low)                   -9.612179  83.746221
## facility_typeschool:riskRisk 3 (Low)                       -9.320140  83.746604
## facility_typecatering:inspection_typecomplaint             -0.593575   0.445544
## facility_typegas station:inspection_typecomplaint          -0.379473   0.818318
## facility_typegrocery store:inspection_typecomplaint         0.109647   0.257004
## facility_typehospital:inspection_typecomplaint              1.323457   1.088399
## facility_typeliquor:inspection_typecomplaint                1.035556   0.506004
## facility_typelong term care:inspection_typecomplaint       -0.800692   0.543578
## facility_typerestaurant:inspection_typecomplaint           -0.077604   0.248180
## facility_typeschool:inspection_typecomplaint               -0.235726   0.361534
## facility_typecatering:inspection_typelicense               -0.347849   0.344587
## facility_typegas station:inspection_typelicense            -0.580735   0.926382
## facility_typegrocery store:inspection_typelicense          -0.149254   0.261409
## facility_typehospital:inspection_typelicense               -0.548767   0.929302
## facility_typeliquor:inspection_typelicense                  0.433764   0.468202
## facility_typelong term care:inspection_typelicense          0.271522   0.518808
## facility_typerestaurant:inspection_typelicense             -0.236373   0.248476
## facility_typeschool:inspection_typelicense                  0.204022   0.258315
## riskRisk 2 (Medium):inspection_typecomplaint               -0.111259   0.075301
## riskRisk 3 (Low):inspection_typecomplaint                  -0.333534   0.158868
## riskRisk 2 (Medium):inspection_typelicense                 -0.358114   0.083003
## riskRisk 3 (Low):inspection_typelicense                    -0.004022   0.144574
##                                                          z value Pr(>|z|)
## (Intercept)                                                8.628  < 2e-16 ***
## facility_typecatering                                      2.027  0.04262 *
## facility_typegas station                                  -0.114  0.90940
## facility_typegrocery store                                 1.837  0.06623 .
## facility_typehospital                                      2.184  0.02894 *
## facility_typeliquor                                       -0.098  0.92217
## facility_typelong term care                                0.325  0.74551
```

```
## facility_typerestaurant                                    2.723  0.00646 **
## facility_typeschool                                        2.571  0.01013 *
## riskRisk 2 (Medium)                                       -0.212  0.83176
## riskRisk 3 (Low)                                           0.116  0.90791
## inspection_typecomplaint                                  -1.423  0.15473
## inspection_typelicense                                    -1.714  0.08647 .
## facility_typecatering:riskRisk 2 (Medium)                  0.300  0.76442
## facility_typegas station:riskRisk 2 (Medium)               0.112  0.91085
## facility_typegrocery store:riskRisk 2 (Medium)            -1.247  0.21252
## facility_typehospital:riskRisk 2 (Medium)                     NA       NA
## facility_typeliquor:riskRisk 2 (Medium)                    0.091  0.92774
## facility_typelong term care:riskRisk 2 (Medium)               NA       NA
## facility_typerestaurant:riskRisk 2 (Medium)                1.767  0.07727 .
## facility_typeschool:riskRisk 2 (Medium)                   -0.786  0.43170
## facility_typecatering:riskRisk 3 (Low)                    -0.114  0.90888
## facility_typegas station:riskRisk 3 (Low)                     NA       NA
## facility_typegrocery store:riskRisk 3 (Low)               -0.120  0.90478
## facility_typehospital:riskRisk 3 (Low)                        NA       NA
## facility_typeliquor:riskRisk 3 (Low)                       0.007  0.99443
## facility_typelong term care:riskRisk 3 (Low)                  NA       NA
## facility_typerestaurant:riskRisk 3 (Low)                  -0.115  0.90862
## facility_typeschool:riskRisk 3 (Low)                      -0.111  0.91139
## facility_typecatering:inspection_typecomplaint            -1.332  0.18278
## facility_typegas station:inspection_typecomplaint         -0.464  0.64285
## facility_typegrocery store:inspection_typecomplaint        0.427  0.66964
## facility_typehospital:inspection_typecomplaint             1.216  0.22400
## facility_typeliquor:inspection_typecomplaint               2.047  0.04070 *
## facility_typelong term care:inspection_typecomplaint      -1.473  0.14075
## facility_typerestaurant:inspection_typecomplaint          -0.313  0.75451
## facility_typeschool:inspection_typecomplaint              -0.652  0.51439
## facility_typecatering:inspection_typelicense              -1.009  0.31275
## facility_typegas station:inspection_typelicense           -0.627  0.53073
## facility_typegrocery store:inspection_typelicense         -0.571  0.56803
## facility_typehospital:inspection_typelicense              -0.591  0.55485
## facility_typeliquor:inspection_typelicense                 0.926  0.35421
## facility_typelong term care:inspection_typelicense         0.523  0.60073
## facility_typerestaurant:inspection_typelicense            -0.951  0.34146
## facility_typeschool:inspection_typelicense                 0.790  0.42963
## riskRisk 2 (Medium):inspection_typecomplaint              -1.478  0.13954
## riskRisk 3 (Low):inspection_typecomplaint                 -2.099  0.03578 *
## riskRisk 2 (Medium):inspection_typelicense                -4.314  1.6e-05 ***
## riskRisk 3 (Low):inspection_typelicense                   -0.028  0.97781
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 53400  on 49999  degrees of freedom
## Residual deviance: 52458  on 49956  degrees of freedom
## AIC: 52546
##
## Number of Fisher Scoring iterations: 9
```

Answer 2:

I set all four variables as factors. ZIP code was set as a factor because the ZIP codes of Chicago do not have any geographic order. The other three variables were set as factors due to being qualitative. The model selection began by comparing the null model to the four single-variable models (Step 1) and the LRT found they were all strongly significant ($p = 0$, $p = 0$, $p = 0$, $p = 0$). However, I chose to exclude the ZIP code variable because it had 72 levels which would be impractical to interpret, especially with interaction terms, and additionally none of the individual levels were significant at the $\alpha = 0.2$ level (the most significant was 60604 with $p = 0.2906645$).

I fit a model with all three remaining variables (Step 2) and the LRT again found the deviance was significantly less than that in the null model ($\chi_{12} = 795.0736886$, $p = 0$). I fit models that included the three possible two-variable interaction terms (Step 3) and the LRT found them all to be significantly better than the three-variable model without interaction ($p = 4.4030454 \times 10^{-9}$, $p = 3.0192195 \times 10^{-7}$, $p = 5.2029617 \times 10^{-7}$). A model that included all three of these interaction pairs (Step 4) was also significant ($\chi_{31} = 147.0872601$, $p = 0$). However, when I fit the same model including a three-way interaction (Step 5), it was not significant ($\chi_{50} =$ -1056890, $p = 1$) and actually increased the deviance by more than twofold. I verified my results by using the `step()` function (Step 6) on the model from **Step 4** and it concurred that that model appeared to be the "best" model with a Akaike information criterion (AIC) value of 52546.19.

The interpretation of the model is difficult as there are $f + r + t + fr + ft + rt - 5 = 43$ coefficients (where $f = 9 - 1 = 8$ variables for facility type, $r = 3 - 1 = 2$ variables for risk, and $t = 3 - 1 = 2$ variables for inspection type, and 5 interaction terms were excluded for having only one value), but we can see most coefficients are not significant in the model. However, at least one coefficient from every variable and interaction is significant at the $\alpha = 0.1$ level. The most significant is the interaction term between medium risk and license inspection ($p = 0.00001599997$), followed by restaurant ($p = 0.0064649$) and school facility ($p = 0.010132$). Hospital facility, low risk/complaint inspection, liquor store/complaint inspection, and gas station were also significant at the $\alpha = 0.05$ level, and grocery store, restaurant/medium risk, and license inspection were significant at the $\alpha = 0.1$ level.

Among the single levels significant at the $\alpha = 0.05$ level, hospital facility had the greatest effect (0.5691979). The **estimated odds** of a hospital facility passing an inspection are approximately 1.7668493 times greater than the estimated odds of a bakery facility (null facility variable) passing an inspection. There were several variables that had large effect sizes, including liquor store, liquor store/medium risk, and grocery/low risk, but with no statistical significance ($p > 0.9$).

3. Either perform a baseline-category logit model or a cumulative logit model for the outcome of `result` in its original form (using previous explanatory variable from question 2 except using `zip`). Describe why you picked the model you picked, write out what the model would look like ($\alpha's$ and $\beta's$), and interpret the results of your model (at lease one β parameter and the overall fit of the model through a hypothesis test).

```
set.seed(7522,sample.kind="Rounding")
```

```
## Warning in set.seed(7522, sample.kind = "Rounding"): non-uniform 'Rounding'
## sampler used
```

```
inspect_sub<-inspection[sample(nrow(inspection),50000),]
library(VGAM)
```

```
## Loading required package: stats4
```

```
## Loading required package: splines
```

```
L<-vglm(results~facility_type+risk+inspection_type,family=multinomial(refLevel="Fail"),data=inspect_sub
summary(L)
Lc<-vglm(results~facility_type+risk+inspection_type,family=multinomial(refLevel="Pass w/ Conditions"),da
summary(Lc)
```

```
1-pchisq(L@criterion$deviance,L@df.residual)
1-pchisq(Lc@criterion$deviance,Lc@df.residual)
```

```
##
## Call:
## vglm(formula = results ~ facility_type + risk + inspection_type,
##     family = multinomial(refLevel = "Fail"), data = inspect_sub)
##
## Coefficients:
##                                Estimate Std. Error z value Pr(>|z|)
## (Intercept):1                 0.7878789  0.0853917   9.227  < 2e-16 ***
## (Intercept):2                -0.3508796  0.1137184  -3.086  0.00203 **
## facility_typecatering:1       0.3252131  0.1246432   2.609  0.00908 **
## facility_typecatering:2       0.1244250  0.1813594   0.686  0.49267
## facility_typegas station:1   -0.2508790  0.3493778  -0.718  0.47271
## facility_typegas station:2   -0.0107595  0.5001939  -0.022  0.98284
## facility_typegrocery store:1  0.0898248  0.0906799   0.991  0.32190
## facility_typegrocery store:2  0.2440669  0.1206052   2.024  0.04300 *
## facility_typehospital:1       0.7111257  0.2346959   3.030  0.00245 **
## facility_typehospital:2       0.6025985  0.2897536   2.080  0.03755 *
## facility_typeliquor:1        -0.3779730  0.1837706  -2.057  0.03971 *
## facility_typeliquor:2        -0.0658997  0.2813647  -0.234  0.81482
## facility_typelong term care:1 0.0964457  0.1360491   0.709  0.47838
## facility_typelong term care:2 0.0335828  0.1798373   0.187  0.85186
## facility_typerestaurant:1     0.3607868  0.0855432   4.218 2.47e-05 ***
## facility_typerestaurant:2     0.4986027  0.1139365   4.376 1.21e-05 ***
## facility_typeschool:1         0.5830214  0.0905145   6.441 1.19e-10 ***
## facility_typeschool:2        -0.0320650  0.1225191  -0.262  0.79354
## riskRisk 2 (Medium):1         0.0266542  0.0311902   0.855  0.39279
## riskRisk 2 (Medium):2        -0.0009973  0.0394713  -0.025  0.97984
## riskRisk 3 (Low):1           -0.0450942  0.0559256  -0.806  0.42006
## riskRisk 3 (Low):2           -0.3760567  0.0837005  -4.493 7.03e-06 ***
## inspection_typecomplaint:1   -0.4349925  0.0301736 -14.416  < 2e-16 ***
## inspection_typecomplaint:2   -0.3218510  0.0369770  -8.704  < 2e-16 ***
## inspection_typelicense:1     -0.4846911  0.0291573 -16.623  < 2e-16 ***
## inspection_typelicense:2     -1.2627926  0.0472325 -26.736  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,2]/mu[,1]), log(mu[,3]/mu[,1])
##
## Residual deviance: 93871.89 on 99974 degrees of freedom
##
## Log-likelihood: -46935.94 on 99974 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Reference group is level  1  of the response
##
## Call:
## vglm(formula = results ~ facility_type + risk + inspection_type,
```

```
##     family = multinomial(refLevel = "Pass w/ Conditions"), data = inspect_sub)
##
## Coefficients:
##                                 Estimate Std. Error z value Pr(>|z|)
## (Intercept):1                  0.3508796  0.1137184   3.086 0.002032 **
## (Intercept):2                  1.1387585  0.1019021  11.175  < 2e-16 ***
## facility_typecatering:1       -0.1244250  0.1813594  -0.686 0.492670
## facility_typecatering:2        0.2007881  0.1637125   1.226 0.220023
## facility_typegas station:1     0.0107595  0.5001939   0.022 0.982838
## facility_typegas station:2    -0.2401195  0.4812884  -0.499 0.617843
## facility_typegrocery store:1  -0.2440669  0.1206052  -2.024 0.043003 *
## facility_typegrocery store:2  -0.1542420  0.1081732  -1.426 0.153903
## facility_typehospital:1       -0.6025985  0.2897536  -2.080 0.037554 *
## facility_typehospital:2        0.1085272  0.2277712   0.476 0.633736
## facility_typeliquor:1          0.0658997  0.2813647   0.234 0.814819
## facility_typeliquor:2         -0.3120733  0.2742830  -1.138 0.255213
## facility_typelong term care:1 -0.0335828  0.1798373  -0.187 0.851865
## facility_typelong term care:2  0.0628629  0.1589844   0.395 0.692545
## facility_typerestaurant:1     -0.4986027  0.1139365  -4.376 1.21e-05 ***
## facility_typerestaurant:2     -0.1378159  0.1020755  -1.350 0.176972
## facility_typeschool:1          0.0320650  0.1225191   0.262 0.793542
## facility_typeschool:2          0.6150864  0.1092695   5.629 1.81e-08 ***
## riskRisk 2 (Medium):1          0.0009973  0.0394713   0.025 0.979843
## riskRisk 2 (Medium):2          0.0276514  0.0335738   0.824 0.410166
## riskRisk 3 (Low):1             0.3760567  0.0837005   4.493 7.03e-06 ***
## riskRisk 3 (Low):2             0.3309625  0.0770540   4.295 1.75e-05 ***
## inspection_typecomplaint:1     0.3218510  0.0369770   8.704  < 2e-16 ***
## inspection_typecomplaint:2    -0.1131415  0.0321127  -3.523 0.000426 ***
## inspection_typelicense:1       1.2627926  0.0472325  26.736  < 2e-16 ***
## inspection_typelicense:2       0.7781015  0.0438237  17.755  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Names of linear predictors: log(mu[,1]/mu[,3]), log(mu[,2]/mu[,3])
##
## Residual deviance: 93871.89 on 99974 degrees of freedom
##
## Log-likelihood: -46935.94 on 99974 degrees of freedom
##
## Number of Fisher scoring iterations: 5
##
## No Hauck-Donner effect found in any of the estimates
##
##
## Reference group is level  3  of the response
## [1] 1
## [1] 1
```

Answer 3:

I ran a baseline-category logit model on the data because it seemed more appropriate with the outcome variable. The "Pass w/ Conditions" level has different specific conditions for each inspection, and it seems intuitive, but it's not entirely clear it should be classified between the "Pass" and "Fail" levels without knowing more about these conditions. Additional analysis on the conditions may warrant that a cumulative

logit model is more appropriate. The equations I obtained from the baseline-category logit model were approximately:

$\log(\frac{\hat{\pi}_P}{\hat{\pi}_F}) = 0.788 + 0.325x_{Fc} - 0.251x_{Fgas} + 0.090x_{Fg} + 0.711x_{Fh} - 0.378x_{Fl} + 0.096x_{Flt} + 0.361x_{Fr} + 0.583x_{Fs} + 0.027x_{R2} - 0.045x_{R3} - 0.435x_{Tc} - 0.485x_{Tl}$

$\log(\frac{\hat{\pi}_{Pw}}{\hat{\pi}_F}) = -0.351 + 0.124x_{Fc} - 0.011x_{Fgas} + 0.244x_{Fg} + 0.603x_{Fh} - 0.066x_{Fl} + 0.034x_{Flt} + 0.499x_{Fr} - 0.032x_{Fs} - 0.001x_{R2} - 0.376x_{R3} - 0.322x_{Tc} - 1.263x_{Tl}$

$\log(\frac{\hat{\pi}_P}{\hat{\pi}_{Pw}}) = \log(\frac{\hat{\pi}_P}{\hat{\pi}_F}) - \log(\frac{\hat{\pi}_{Pw}}{\hat{\pi}_F}) = 1.14 + 0.20x_{Fc} - 0.24x_{Fgas} - 0.15x_{Fg} + 0.11x_{Fh} - 0.31x_{Fl} + 0.06x_{Flt} - 0.14x_{Fr} + 0.62x_{Fs} + 0.03x_{R2} + 0.33x_{R3} - 0.11x_{Tc} + 0.78x_{Tl}$,

where

$\hat{\pi}_P$: estimated probability of passing inspection,

$\hat{\pi}_{Pw}$: estimated probability of passing inspection with conditions,

$\hat{\pi}_F$: estimated probability of failing inspection,

α : intercept term,

β_{Fc} : coefficient for catering facility,

x_{Fc} : dummy variable for catering facility,

β_{Fgas} : coefficient for gas station facility,

x_{Fgas} : dummy variable for gas station facility,

β_{Fg} : coefficient for grocery store facility,

x_{Fg} : dummy variable for grocery store facility,

β_{Fh} : coefficient for hospital facility,

x_{Fh} : dummy variable for hospital facility,

β_{Fl} : coefficient for liquor store facility,

x_{Fl} : dummy variable for liquor store facility,

β_{Flt} : coefficient for long-term care facility,

x_{Flt} : dummy variable for long-term care facility,

β_{Fr} : coefficient for restaurant facility,

x_{Fr} : dummy variable for restaurant facility,

β_{Fs} : coefficient for school facility,

x_{Fs} : dummy variable for school facility,

β_{R2} : coefficient for medium risk,

x_{R2} : dummy variable for medium risk,

β_{R3} : coefficient for low risk,

x_{R3} : dummy variable for low risk,

β_{Tc} : coefficient for complaint inspection,

x_{Tc} : dummy variable for complaint inspection,

β_{Tl} : coefficient for license inspection, and

x_{Tl} : dummy variable for license inspection.

We can see for the $\log(\frac{\hat{\pi}_P}{\hat{\pi}_F})$ model that inspection type (both license and complaint) are strongly significant ($p = 4.7232139 \times 10^{-62}$, $p = 4.0861261 \times 10^{-47}$), as well as the school, restaurant, hospital, catering, and liquor store facilities at the $\alpha = 0.05$ level. Among these, hospital facility had the greatest effect (0.7111257), with school facility and both inspection type variables also having large effects. The **estimated odds** of a hospital facility passing an inspection are approximately 2.0362823 times greater than the estimated odds of a bakery facility (null facility variable) passing an inspection.

For the $\log(\frac{\hat{\pi}_{Pw}}{\hat{\pi}_F})$ model, both inspection type variables were again strongly significant ($p = 1.8109933 \times 10^{-157}$, $p = 3.2017557 \times 10^{-18}$). Low risk and grocery store facility were significant at the $\alpha = 0.05$ level in addition to the same restaurant and hospital facilities that were significant in the previous model. Among these, license inspection had the greatest effect by far (-1.2627926), followed by hospital and restaurant facilities. The **estimated odds** of a license inspection passing are approximately 0.282863 times greater than the estimated odds of a canvass inspection (null inspection type variable) passing.

Lastly, for the $\log(\frac{\hat{\pi}_P}{\hat{\pi}_{Pw}})$ model, both inspection type variables were still significant ($p = 1.56909 \times 10^{-70}$, $p = 0.000426257$), but complaint inspection was not as significant as school facility and low risk. License inspection had the greatest effect (0.7781015), followed by school facility, low risk, and complaint inspection. The **estimated odds** of a license inspection passing are approximately 2.1773347 times greater than the estimated odds of a canvass inspection (null inspection type variable) passing. When performing multinomial model goodness-of-fit tests, we find the null hypotheses for the models soundly fail to be rejected ($\chi_{99974} = 93871.89$, $p = 1$), indicating the **model fits well**.

4. Perform a marginal model for a clustered binary response (`result` with only success or failure). You must clearly state

- What is the "cluster" here? What variable will you use from the data to express the clustering structure in the model? (Write out the model).
- What is the correlation structure you assumed? Why did you assume that?
- Interpret at least one β and compare the naive and robust standard errors from the model.
- Can you make likelihood based inferences on this model like you did in question 2 or 3?

```
inspect_4<-inspection[10100:10800,] # New subset per Campuswire discussion (https://campuswire.com/c/G1
inspect_4[inspect_4$results=="Pass w/ Conditions","results"]<-"Pass"
inspect_4$results<-factor(inspect_4$results,labels=0:1)
library(gee)
gee<-gee(results~facility_type+risk+inspection_type,id=license_number,corstr="exchangeable",family=bino
```

```
## Beginning Cgee S-function, @(#) geeformula.q 4.13 98/01/27

## running glm to get initial regression estimate
```

```
summary(gee)
```

```
##                 (Intercept)       facility_typecatering
##                1.091644925                  0.455208534
##  facility_typegrocery store facility_typelong term care
##                -0.305567092                  0.207638059
##    facility_typerestaurant         facility_typeschool
##                 0.271889867                  1.383623934
##         riskRisk 2 (Medium)            riskRisk 3 (Low)
##                 0.004676326                 -0.496478216
##    inspection_typecomplaint      inspection_typelicense
##                 0.030489195                 -0.540630190
##
##  GEE:  GENERALIZED LINEAR MODELS FOR DEPENDENT DATA
##  gee S-function, version 4.13 modified 98/01/27 (1998)
##
## Model:
```

```
##  Link:                      Logit
##  Variance to Mean Relation: Binomial
##  Correlation Structure:     Exchangeable
##
## Call:
## gee(formula = results ~ facility_type + risk + inspection_type,
##     id = license_number, data = inspect_4, family = binomial,
##     corstr = "exchangeable")
##
## Summary of Residuals:
##         Min          1Q      Median          3Q         Max
## -0.91609057  0.08390943  0.20967175  0.20967175  0.55540707
##
##
## Coefficients:
##                                Estimate Naive S.E.     Naive z Robust S.E.
## (Intercept)                1.094120822  0.3659718  2.98963154  0.02671576
## facility_typecatering      0.397549481  0.6642059  0.59853351  0.24275362
## facility_typegrocery store -0.324796510  0.4720554 -0.68804738  0.32181531
## facility_typelong term care 0.183968406  0.7083721  0.25970588  0.20773228
## facility_typerestaurant    0.232784354  0.3803942  0.61195550  0.11857432
## facility_typeschool        1.296256408  0.6739582  1.92334836  0.56145505
## riskRisk 2 (Medium)        0.018439879  0.2245480  0.08212000  0.19649216
## riskRisk 3 (Low)           -0.446422840  0.4078376 -1.09460926  0.37064896
## inspection_typecomplaint   0.008474979  0.2347072  0.03610873  0.21701913
## inspection_typelicense     -0.545443667  0.2329592 -2.34136981  0.21301520
##                              Robust z
## (Intercept)                40.95413265
## facility_typecatering       1.63766654
## facility_typegrocery store -1.00926371
## facility_typelong term care 0.88560334
## facility_typerestaurant     1.96319372
## facility_typeschool         2.30874478
## riskRisk 2 (Medium)         0.09384537
## riskRisk 3 (Low)           -1.20443571
## inspection_typecomplaint    0.03905176
## inspection_typelicense     -2.56058561
##
## Estimated Scale Parameter:  0.9992974
## Number of Iterations:  5
##
## Working Correlation
##              [,1]       [,2]       [,3]       [,4]       [,5]       [,6]
##  [1,]  1.0000000 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [2,] -0.0236558  1.0000000 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [3,] -0.0236558 -0.0236558  1.0000000 -0.0236558 -0.0236558 -0.0236558
##  [4,] -0.0236558 -0.0236558 -0.0236558  1.0000000 -0.0236558 -0.0236558
##  [5,] -0.0236558 -0.0236558 -0.0236558 -0.0236558  1.0000000 -0.0236558
##  [6,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558  1.0000000
##  [7,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [8,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [9,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [10,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [11,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
```

```
## [12,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [13,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [14,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [15,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [16,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [17,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [18,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [19,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [20,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##             [,7]       [,8]       [,9]      [,10]      [,11]      [,12]
##  [1,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [2,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [3,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [4,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [5,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [6,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [7,]  1.0000000 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [8,] -0.0236558  1.0000000 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [9,] -0.0236558 -0.0236558  1.0000000 -0.0236558 -0.0236558 -0.0236558
## [10,] -0.0236558 -0.0236558 -0.0236558  1.0000000 -0.0236558 -0.0236558
## [11,] -0.0236558 -0.0236558 -0.0236558 -0.0236558  1.0000000 -0.0236558
## [12,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558  1.0000000
## [13,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [14,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [15,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [16,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [17,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [18,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [19,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [20,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##             [,13]      [,14]      [,15]      [,16]      [,17]      [,18]
##  [1,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [2,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [3,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [4,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [5,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [6,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [7,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [8,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##  [9,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [10,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [11,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [12,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [13,]  1.0000000 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [14,] -0.0236558  1.0000000 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [15,] -0.0236558 -0.0236558  1.0000000 -0.0236558 -0.0236558 -0.0236558
## [16,] -0.0236558 -0.0236558 -0.0236558  1.0000000 -0.0236558 -0.0236558
## [17,] -0.0236558 -0.0236558 -0.0236558 -0.0236558  1.0000000 -0.0236558
## [18,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558  1.0000000
## [19,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
## [20,] -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558 -0.0236558
##             [,19]      [,20]
##  [1,] -0.0236558 -0.0236558
##  [2,] -0.0236558 -0.0236558
```

```
##  [3,] -0.0236558 -0.0236558
##  [4,] -0.0236558 -0.0236558
##  [5,] -0.0236558 -0.0236558
##  [6,] -0.0236558 -0.0236558
##  [7,] -0.0236558 -0.0236558
##  [8,] -0.0236558 -0.0236558
##  [9,] -0.0236558 -0.0236558
## [10,] -0.0236558 -0.0236558
## [11,] -0.0236558 -0.0236558
## [12,] -0.0236558 -0.0236558
## [13,] -0.0236558 -0.0236558
## [14,] -0.0236558 -0.0236558
## [15,] -0.0236558 -0.0236558
## [16,] -0.0236558 -0.0236558
## [17,] -0.0236558 -0.0236558
## [18,] -0.0236558 -0.0236558
## [19,]  1.0000000 -0.0236558
## [20,] -0.0236558  1.0000000
```

Answer 4:

I believe the most appropriate variable for a cluster is **license number**. The inspection ID would not be the most appropriate as it is unique for each inspection, and by clustering on license number we can look at all inspections for a single license (or facility). Other candidates like name DBA and address would also not be completely accurate as there could be multiple facilities with the same name or at the same address. The corresponding generalized estimating equation (GEE) is:

$logit[P(Y_t = 1)] = 1.094 + 0.398z_{Fc} - 0.325z_{Fg} + 0.184z_{Flt} + 0.233z_{Fr} + 1.296z_{Fs} + 0.018z_{R2} - 0.446z_{R3} + 0.008z_{Tc} - 0.545z_{Tl} + \gamma x$

where

$\alpha = 1.094$: intercept term,

$\beta_{Fc} = 0.398$: coefficient for catering facility,

z_{Fc} : dummy variable for catering facility,

$\beta_{Fg} = -0.325$: coefficient for grocery store facility,

z_{Fg} : dummy variable for grocery store facility,

$\beta_{Flt} = 0.184$: coefficient for long-term care facility,

z_{Flt} : dummy variable for long-term care facility,

$\beta_{Fr} = 0.233$: coefficient for restaurant facility,

z_{Fr} : dummy variable for restaurant facility,

$\beta_{Fs} = 1.296$: coefficient for school facility,

z_{Fs} : dummy variable for school facility,

$\beta_{R2} = 0.018$: coefficient for medium risk,

z_{R2} : dummy variable for medium risk,

$\beta_{R3} = -0.446$: coefficient for low risk,

z_{R3} : dummy variable for low risk,

$\beta_{Tc} = 0.008$: coefficient for complaint inspection,

z_{Tc} : dummy variable for complaint inspection,

$\beta_{Tl} = -0.545$: coefficient for license inspection, and

z_{Tl} : dummy variable for license inspection.

I decided to use the **exchangeable** correlation structure as the facility type, risk, and inspection type do not appear to be completely independent, as seen from the statistical significance of the interaction terms in the generalized linear models (GLM), but the correlations between variables were also not entirely different from one another. The textbook writes in Section 9.2.2 (page 257): "In practice, usually little if any *a priori* information is available about the correlation structure. ... Unless you expect dramatic differences among the correlations, we recommend using the exchangeable working correlation structure." We can see from the output that school type has by far the greatest effect (1.2962564), followed by license inspection and restaurant type. The **estimated odds** of a school facility passing an inspection are approximately 3.655586 times the estimated odds of a bakery facility (null facility variable) passing an inspection. We can see there are **significant differences** between the naive and robust standard errors for all variables, leading to different z-statistics. This may be because of the large number of variables in the model. The robust standard errors were all smaller than the naive standard errors which actually would have led to different conclusions at the $\alpha = 0.05$ level for the school type variable.

The equation does not have a known distribution, causing a loss of generality, so likelihood-based inferences **cannot** be made on this model. The textbook explains these limitations in Section 9.2.4 (page 259) and Section 10.2.5 (page 283), writing: "Because the GEE method does not specify the complete multivariate distribution, it does not have a likelihood function. In this sense, the GEE method is a multivariate type of quasi-likelihood method. Therefore, its estimates are not ML estimates. ... With the GEE approach to fitting marginal models, a drawback is that likelihood-based inferences are not available. The GEE approach does not specify a joint distribution of the responses, so it does not have a likelihood function."

5. Perform a GLMM on the response `result` as a binary response. You must clearly state

- What would be the logical choice of a random effect here? How will we signify that in the model? (Write out the model).
- Interpret at least one β and the standard deviation of the random effect. How is the model fitting?
- Discuss a comparison of the marginal vs GLMM model. Not only in terms of parameters but also discuss how this changes the interpretation and assumptions being made about each model. Why would you use one over the other? Which one do you think is correct for this dataset?

```
inspect_sub[inspect_sub$results=="Pass w/ Conditions","results"]<-"Pass"
inspect_sub$results<-factor(inspect_sub$results,labels=0:1)
library(lme4)
```

```
## Loading required package: Matrix
```

```
##
## Attaching package: 'Matrix'
```

```
## The following objects are masked from 'package:tidyr':
##
##     expand, pack, unpack
```

```
glmm<-glmer(results~(1|license_number)+facility_type+risk+inspection_type,family=binomial,nAGQ=20,data=
summary(glmm)
```

```
##
## Correlation matrix not shown by default, as p = 13 > 12.
## Use print(x, correlation=TRUE)  or
##     vcov(x)        if you need it
```

```
as.numeric(1-pchisq(summary(glmm)$AICtab["deviance"],summary(glmm)$AICtab["df.resid"]))
library(geepack)
```

```r
inspect_sub$results<-as.numeric(inspect_sub$results)-1
gw<-geeglm(results~facility_type+risk+inspection_type,id=license_number,corstr="exchangeable",family="bi
anova(gw)
```

```
## Generalized linear mixed model fit by maximum likelihood (Adaptive
##   Gauss-Hermite Quadrature, nAGQ = 20) [glmerMod]
##  Family: binomial  ( logit )
## Formula:
## results ~ (1 | license_number) + facility_type + risk + inspection_type
##    Data: inspect_sub
##
##      AIC      BIC   logLik deviance df.resid
##  52630.2  52753.7 -26301.1  52602.2    49986
##
## Scaled residuals:
##     Min      1Q  Median      3Q     Max
## -2.4147  0.4535  0.4770  0.5703  1.0006
##
## Random effects:
##  Groups         Name        Variance Std.Dev.
##  license_number (Intercept) 0.03099  0.176
## Number of obs: 50000, groups:  license_number, 15782
##
## Fixed effects:
##                              Estimate Std. Error z value Pr(>|z|)
## (Intercept)                   1.07699    0.08385  12.845  < 2e-16 ***
## facility_typecatering         0.29871    0.12294   2.430  0.01511 *
## facility_typegas station     -0.20549    0.33677  -0.610  0.54173
## facility_typegrocery store    0.11958    0.08874   1.348  0.17781
## facility_typehospital         0.68541    0.23185   2.956  0.00311 **
## facility_typeliquor          -0.32697    0.17762  -1.841  0.06565 .
## facility_typelong term care   0.08006    0.13325   0.601  0.54794
## facility_typerestaurant       0.39217    0.08374   4.683 2.83e-06 ***
## facility_typeschool           0.48332    0.08887   5.438 5.37e-08 ***
## riskRisk 2 (Medium)           0.01850    0.03056   0.605  0.54500
## riskRisk 3 (Low)             -0.10967    0.05492  -1.997  0.04586 *
## inspection_typecomplaint     -0.40053    0.02905 -13.786  < 2e-16 ***
## inspection_typelicense       -0.62598    0.02874 -21.778  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## [1] 2.220446e-16
## Analysis of 'Wald statistic' Table
## Model: binomial, link: logit
## Response: results
## Terms added sequentially (first to last)
##
##                  Df     X2 P(>|Chi|)
## facility_type     8 225.51 < 2.2e-16 ***
## risk              2  30.15 2.832e-07 ***
## inspection_type   2 560.96 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Answer 5:

The logical choice of a random effect would again be the **license number** for the same reasons as in the GEE. This would be signified in the model as $\{u_i\}$, with the full model being:

$logit[P(Y_{it} = 1)] = u_i + 1.07698973 + \hat{\beta}_f + \hat{\beta}_r + \hat{\beta}_t + \gamma x$

where

$\{u_i\}$: intercept term for license number i,

$\hat{\alpha} = 1.07698973$: estimated general intercept term,

$\hat{\beta}_f$: estimated coefficient for facility;

where
$f = b$ for bakery ($\hat{\beta}_b = 0$),
$f = c$ for catering ($\hat{\beta}_c = 0.29870541$),
$f = gas$ for gas station ($\hat{\beta}_{gas} = -0.20549148$),
$f = g$ for grocery store ($\hat{\beta}_g = 0.11958310$),
$f = h$ for hospital ($\hat{\beta}_h = 0.68540966$),
$f = q$ for liquor store ($\hat{\beta}_q = -0.32696931$),
$f = lt$ for long-term care ($\hat{\beta}_{lt} = 0.08006439$),
$f = r$ for restaurant ($\hat{\beta}_r = 0.39217254$), and
$f = s$ for school ($\hat{\beta}_s = 0.48331578$);

$\hat{\beta}_r$: estimated coefficient for risk;

where
$r = 1$ for low risk ($\hat{\beta}_1 = 0$)
$r = 2$ for medium risk ($\hat{\beta}_2 = 0.01849662$), and
$r = 3$ for high risk ($\hat{\beta}_3 = -0.10966511$); and

$\hat{\beta}_t$: estimated coefficient for type of inspection;

where
$t = cv$ for canvass ($\hat{\beta}_{cv} = 0$)
$t = ci$ for complaint inspection ($\hat{\beta}_{ci} = -0.40053479$), and
$t = l$ for license ($\hat{\beta}_l = -0.62597678$).

Inspection type is again strongly significant ($p = 3.743867 \times 10^{-105}$, $p = 3.0780941 \times 10^{-43}$), followed by school and restaurant facilities. Hospital and catering facilities are also significant along with low risk. Among these, hospital facility has the greatest effect (0.6854097), followed by license inspection and school facility. The **estimated odds** of a hospital facility passing an inspection are approximately 1.9845847 times the estimated odds of a bakery facility (null facility variable) passing an inspection. We can also see that the **standard deviation of the random effect** $\hat{\sigma} = 0.1760465$. A one-standard deviation change in the individual intercept term between license numbers is approximately a 17.6 percent change in probability of passing an inspection, holding all other variables constant.

We reject H_0 at the $\alpha = 0.05$ level for the χ^2 test. There is strong statistical evidence ($\chi_{49986} = 52602.18$, $p = 2.220446 \times 10^{-16}$) that the **model fits well**. Additionally, we reject H_0 at the $\alpha = 0.05$ level for all three variables (facility type, risk level, and inspection type). There is sufficient evidence ($p = 0$, $p = 0.0000002832244$, $p = 0$) that at least one β_f, β_r, and β_t is different.

The results from the GEE and the generalized linear mixed model (GLMM) are *relatively* similar. It is difficult to compare them directly since they were performed on different and different-sized subsets, but the methods of interpreting variables remain the same. The **main assumption** for the GEE is on the correlation structure of the variables (independence, exchangeable, unstructured, etc.), while the GLMM requires an approximate normality assumption on all variables. For these data, using a GEE would be to compare results

between different license numbers with the marginal probabilities, while using a GLMM would be more to compare results within each license number. Both models are available and the more appropriate model would **depend on the specific analysis** being conducted. For predicting the result of an inspection, I believe the GLMM would be more appropriate as prior results for a given license number would be relevant. Alternatively, if comparing the results of multiple facilities with one another (e.g., the proportion of passed inspections for each facility), I believe the GEE would be more appropriate.

6. Graduate Student questions: You do not have to model in R and this can be written out and attached to the upload.

a) Describe what a transitional model would look like for this data. What is the point of adding a transitional term to the model and what assumption are be making when we use it? What would the β parameter be describing/interpreted as? What would the difference between first order and second order term be? What is the max order term we could use in this data (how far back in time can we go?)?
b) If you could make this dataset be a multilevel model what would your first and second level models be? Hypothesize other variables you might have about the business/establishment (full dataset have year of creation, etc.). Would you have random effects and if so at which level?

Answer 6:

(**a**) A transitional model would incorporate the inspection date variable in an attempt to predict the outcome of future inspections. The main assumption being made would be that all past inspections recorded except the last (most recent) one are **conditionally independent**, which may not necessarily be true (follow-up or duplicate inspections, violation-specific inspections, etc.). In the model, β would represent the coefficient for the result of the previous inspection, with a greater effect ($|\beta|$) indicating a greater influence of the previous inspection result on the model. The first-order Markov model would only account for the **result of the last inspection**, while the second-order Markov model would account for the results of the previous two inspections. Accordingly, the nth-order model would account for the results of the previous n inspections. The maximum order term we would be able to use would be the number of inspections in the data (50,000) and we would be able to go back to the earliest inspection listed (January 05, 2010). However, this would be less useful as the previous methods used in this exam, as it would be impractical to interpret that many terms in the model and it seems that accounting for every previous inspection would defeat the purpose of using a Markov model to begin with.

(**b**) Looking at the variables in the data, it seems like license number and facility type would be logical choices for the first and second levels respectively. Variability between licenses would be captured by the first level and variability between facility types would be captured by the second level. Alternate choices for first and second levels could be inspection ID and license number to capture variability between inspections and types of licenses, or maybe address and ZIP code to capture variability between locations and ZIP codes. There are countless **other variables** that could be added to a multilevel model, but I think square area, number of entrances/exits, and floor number(s) of the facility, expired business (or other applicable) permit (binary), Americans with Disabilities Act compliance (binary), alcohol sale (binary), public restrooms (binary), and customer rating (Yelp, Facebook, etc.) could be suitable variables. These variables could all be random effects on the **first level**.