**Multinomial Logistic Regression Analysis of Death and Recovery Probability Among Patients of Different Age Groups and Gender Who Were Hospitalized Under Different Lengths of Time**

Sean Cancino

Group 1

University of Delaware

Research Paper for STAT675 – Logistic Regression

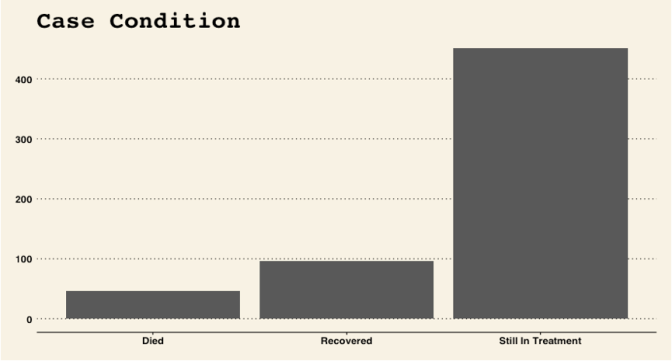
Spring 2020

**Abstract:**

This study aimed to analyze how gender and age influenced the probability of death or recovery under different time lengths in a hospital. 5 models corresponding to length of duration, which were 5 days, 10 days, 15 days, 20 days, and 25 days in a hospital were analyzed. It was found that age and gender play a significant role in death and recovery; the probability of death increased as age increased, the probability of recovery decreased as age increased, and males were more likely to die than females. Also, the probability of death or recovery increased as time duration increased.

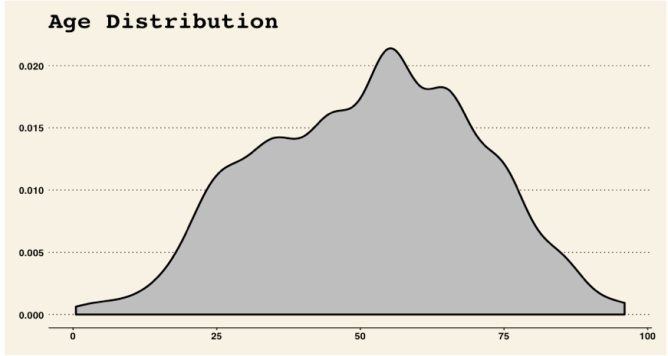
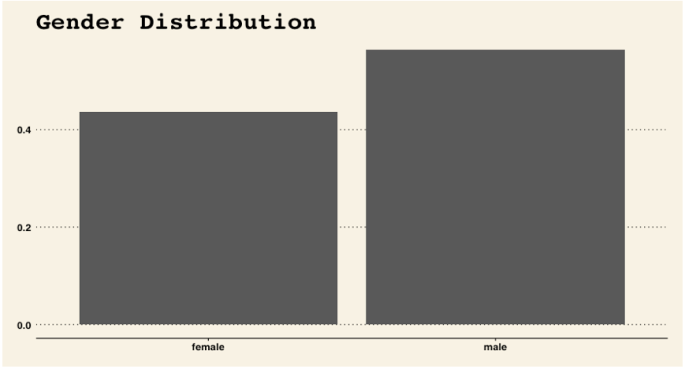
**Introduction and Exploratory Data Analysis:**

This project studied the effects of Covid-19 on the population of infected patients. The data was retrieved from DXY.cn, an online community for medical professional and news sharing of medical importance. The data set in total contained 3397 observations and 23 variables. Many of the cases had missing values in death, recovery, and other variables. The variables in the data set included death of a patient, recovery of a patient, age, gender, their name, the country they were from, whether they traveled or not, etc. Overall, there were 595 cases with non-missing death, recovery, and hospital visit date values. For the exploratory data analysis, several charts were made. The first chart, shown below, shows the number of cases who died, who recovered, or who were still in treatment. The chart is in the form of a bar graph and it shows around 600 observations.

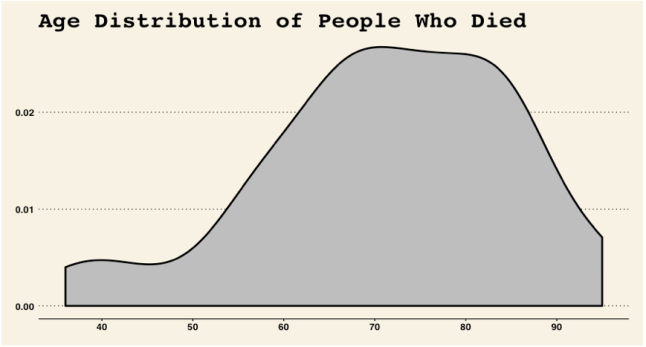
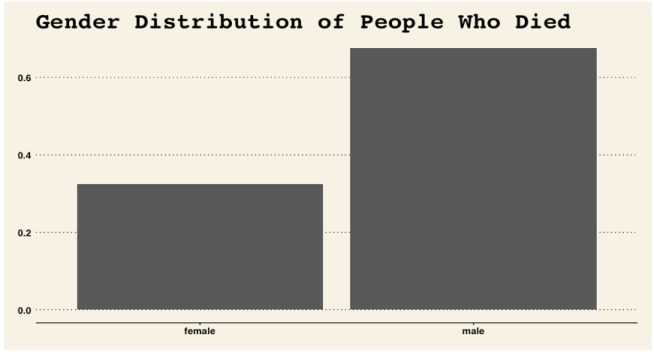


The above chart shows that there are less people who died than people who recovered, and there were approximately 50 deaths. There were approximately 100 people who recovered from the disease. The vast majority of people were still in treatment, as there are approximately more than 400 cases in treatment.

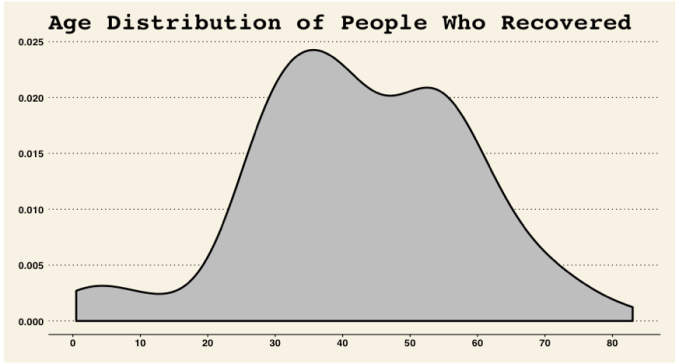
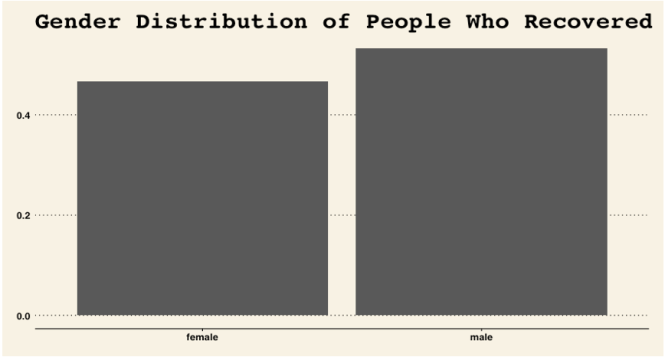
The next two charts show the age and gender distributions. The age distribution shows that most of the patients, around 80% or more, were around 25 to 75 years old. The maximum frequency of cases appears to be around the 50 to 55 year mark. As for the gender distribution, there were more male cases than female cases.

The next two charts show the age and gender distribution of people who died. For the age distribution graph, it appears that most of the cases who died were around 60 to 90 years old. For the gender distribution of people who died, the number of males who died appears to be double that of the number of males. From these two graphs, it can be predicted that the chance of death from COVID-19 is higher among males and among the older population (60+ years old).

The next two charts show the age and gender distribution of people who have recovered from COVID-19. The age distribution graph appears to be similar to the age distribution graph of people who got the disease. Most of the cases that recovered appear to fall within the 20 to 75 year range. As for the gender distribution graph, there were more males than females that recovered, but overall it appears to have a similar distribution as the gender distribution of cases of COVID-19. It might be predicted that gender has no significant effect on the chance of recovery. As for age, since the graph has a similar distribution as the age distribution of cases, one could predict that age has no significant effect on the chance of recovery.

**Methods:**

Considering the variables in that data set that could be used in the analysis, our team decided to examine how age and gender influence the chance of recovery and death among patients with COVID-19, and also how recovery and death rate were influenced by the duration of time spent in the hospital.

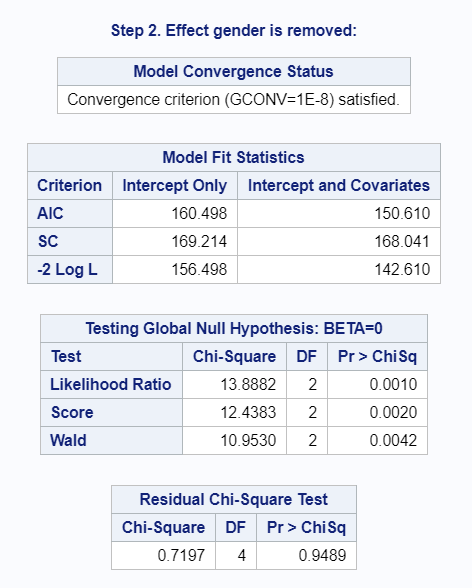
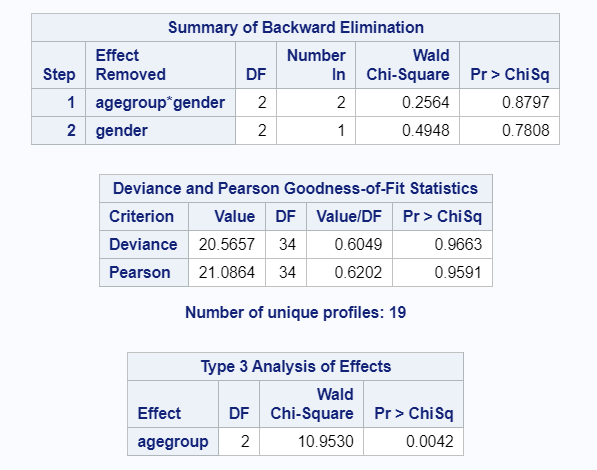
As mentioned above, age was used as a predictor variable, and it was turned into an ordinal categorical variable. People were classified into different age groups. The age groups were: 1 – 0 to 10 years old, 2 – 11 to 20 years old, 3 – 21 to 30 years old, 4 – 31 to 40 years old, 5 – 41 to 50 years old, 6 – 51 to 60 years old, 7 – 61 to 70 years old, 8 – 71 to 80 years old, 9 – 81 to 90 years old, 10 – 91 and older. As for gender, it is obviously a categorical variable. The response variable was a multinomial categorical variable with 3 levels – death, recovery, and still in treatment at a certain length of time. The “still in treatment” level was treated as a reference group for the other two levels. Data for the response variable was collected for 5 different time periods of patients who stayed in a hospital – 5 days in hospital, 10 days in hospital, 15 days in hospital, 20 days in hospital, and 25 days in hospital. 5 different models were created for this experiment corresponding to the different time periods of the response variable.

In order to select an optimal model, backwards selection was run with a cut-off probability of keeping a variable in the model at .15. Multinomial logistic regression of course was run for this model.

**Statistical Analysis:**

The first model that was run was the 5 days in hospital time period. As shown in the picture below, there were two variables that were removed from the final model – the age\*gender interaction term, and the gender variable. Overall, it appears that the model was a good fit for the data. The AIC statistic for the intercept and age variable was around 150, which is less than the intercept only model, which was around 160. The SC statistic for the intercept and age variable was around 168, which was only slightly lower than the SC statistic for the intercept only model. The -2\*Log L statistic for the intercept and age variable was around 142, which is much less than the intercept only model which was around 156.

For testing the global null hypothesis of the intercept only model, the Likelihood Ratio Test, Score Test, and Wald Test gave p-values of less than .05, which is an indicator that the age and intercept model is a better model for the data than the intercept model alone.

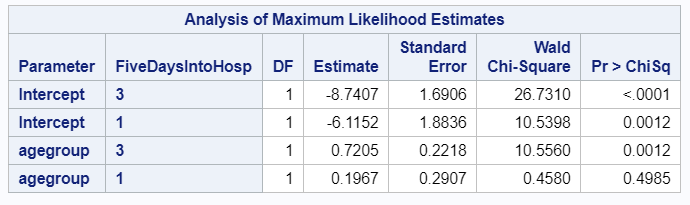
The Deviance and Pearson goodness of fit statistics also show that the model is a good fit. The p-values for Deviance and Pearson are clearly greater than .05, which are indicators that the model is a good fit.

Shown below are the model estimates. From the estimates, the model equations were found to be:

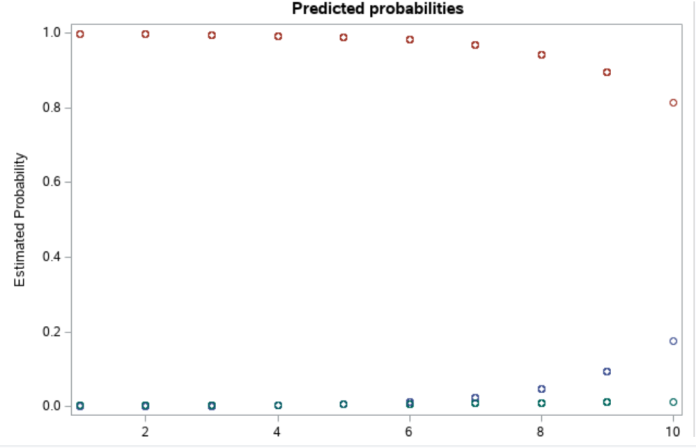
log(P(Y=1)/P(Y=2)) = -6.1152 + .1967\*I{agegroup=1}

log(P(Y=3)/P(Y=2)) = -8.7407 + .7205\*I{agegroup=3}

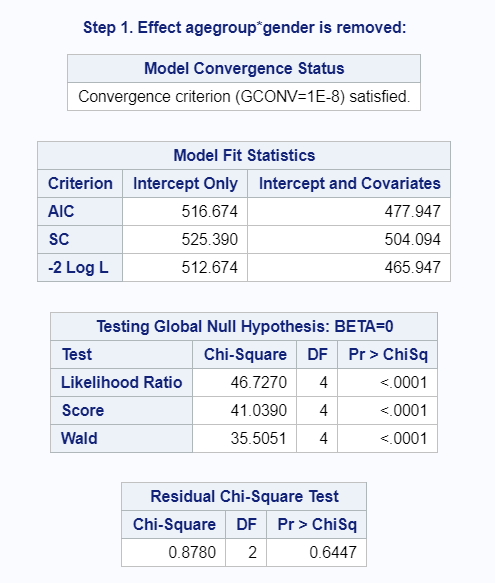
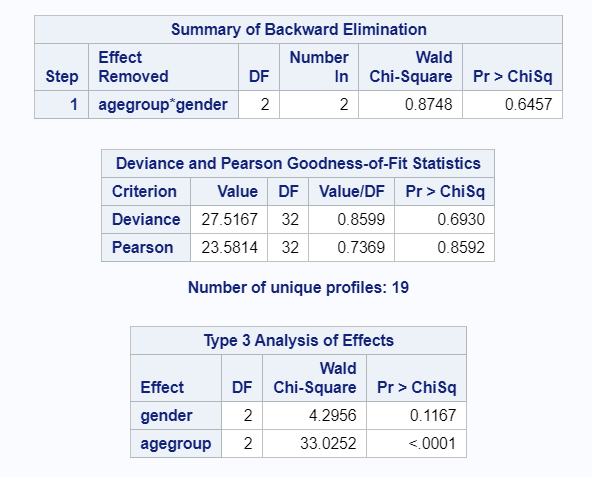
Where log(P(Y=1)/P(Y=2)) are the log odds of recovery over the still in treatment reference group, and log(P(Y=3)/P(Y=2)) are the log odds of death over the still in treatment reference group.



Shown below is the predicted probability graph for the model. It appears that recovery rate is flat among all age groups, and that death rate starts increasing from around 5 to 20% at age 70. From this graph, it can be concluded that the chance of recovery from COVID is very small for all age groups for patients who stay in a hospital for 5 or less days, and the chance of death from COVID is around 5 to 20% for people greater than 70 years old and who stay in a hospital for 5 or less days.



The next model that was investigated was the 15 days in hospital model. As can be seen in the pictures below, the only variable that was removed from the model was the age\*gender interaction term. The AIC, SC, and -2\*LogL terms for the final model are all less than the intercept only model, which indicate that the final model is a better fit than the intercept only model. The testing global null hypothesis p-values for likelihood, score, and wald test are all less than .0001, which indicate the final model is a better fit than the intercept only model. The Deviance and Pearson goodness of fit statistics have p-values greater than .05, which are also indications that the final model is a good fit for the data.

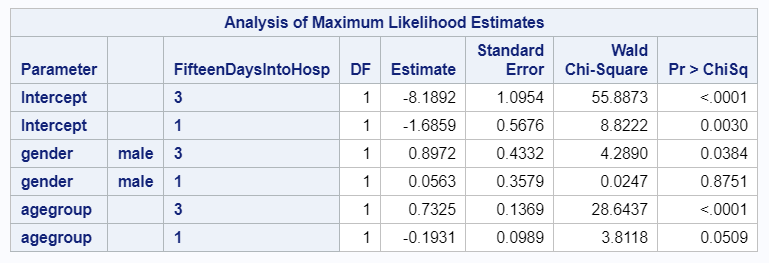
 

Shown below are the model estimates. From the estimates, the model equations were found to be:

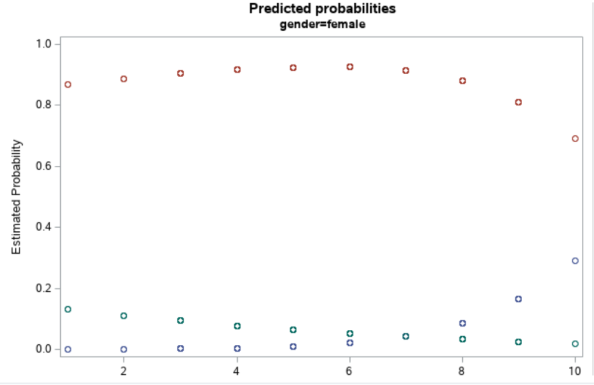
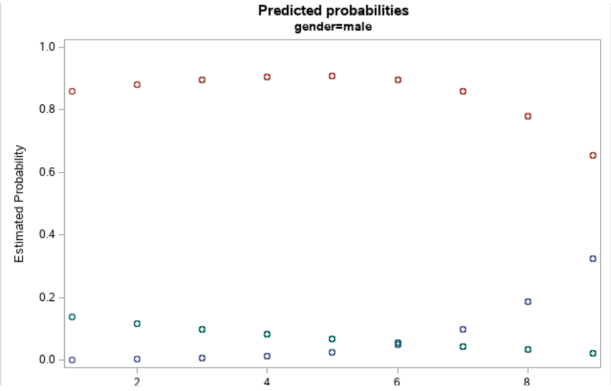
log(P(Y=1)/P(Y=2)) = -6.1152 + .1967\*I{agegroup=1}

log(P(Y=3)/P(Y=2)) = -8.7407 + .7205\*I{agegroup=3}

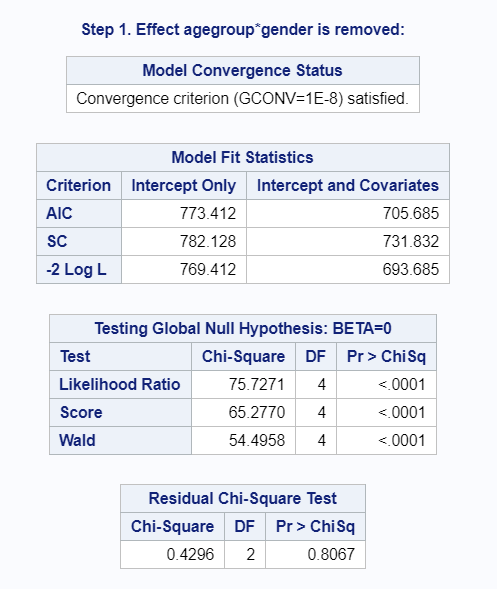
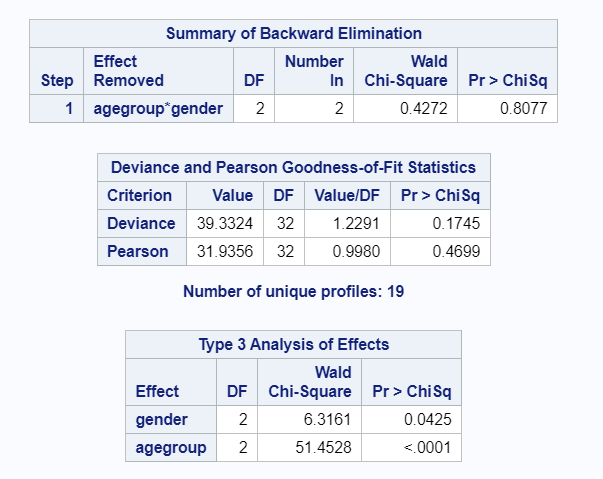
Where log(P(Y=1)/P(Y=2)) are the log odds of recovery over the still in treatment reference group, and log(P(Y=3)/P(Y=2)) are the log odds of death over the still in treatment reference group.



Shown below are the predicted probability plots for females and males. The recovery probability was highest among younger age groups, age 20 or less, and the recovery probability decreased as age increased. The death probability was close to zero for most age groups besides people age 60 or higher, where the death probability ranged from 5% up to 35%. The predicted probability of males dying at age 81-90 or higher is around 35%, which is much higher than females of the same age range, which have a predicted probability of death around 20%.

The last model that was investigated was the 25 days in hospital model. As shown below, only the interaction term (agegroup\*gender) was removed from the final model. The AIC, SC, and -2 Log L criterions for the final model were all lower than the intercept only model. The likelihood ratio, score, and wald tests that tested the global null hypothesis of the intercept only model all had p-values of less than .05, which indicates that the final model is a better fit for the data than the intercept only model. The Deviance and Pearson goodness of fit tests show p-values of greater than .05, which are also indicators that this final model is a good model. Overall, the final model included age and gender predictor variables.

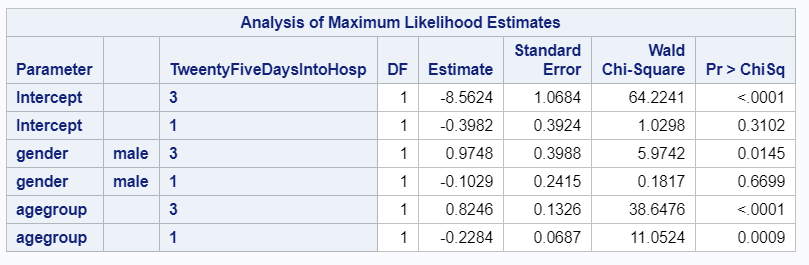
 

Shown below are the model estimates for the data. From the estimates, the model equations were found to be:

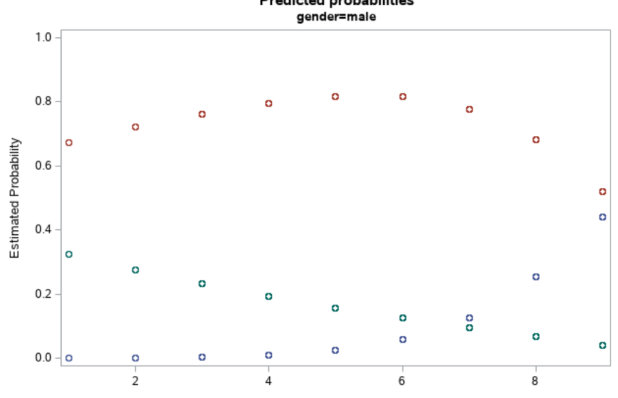
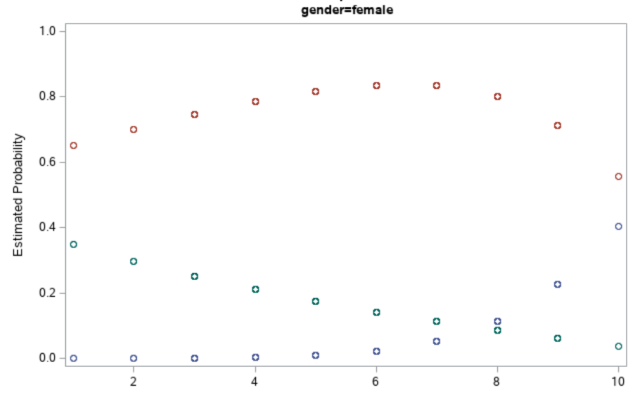
log((Y=1)/(Y=2)) = -.3982+-.2284\*I{agegroup=1}+-.1029\*I{gender|male=1}

log((Y=3)/(Y=2)) = -8.5624+.8246\*I{agegroup=3}+.9748\*I{gender|male=3}

Where log(P(Y=1)/P(Y=2)) are the log odds of recovery over the still in treatment reference group, and log(P(Y=3)/P(Y=2)) are the log odds of death over the still in treatment reference group.

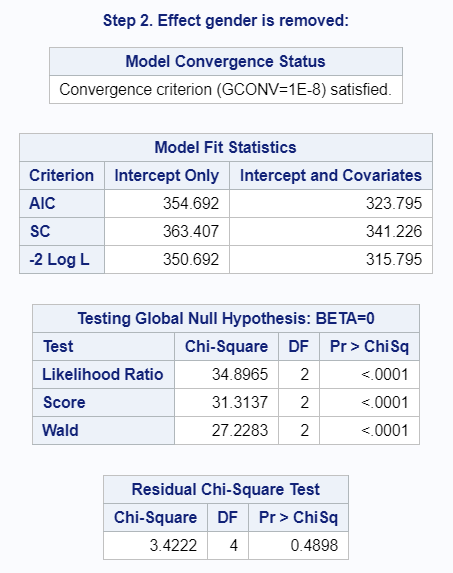
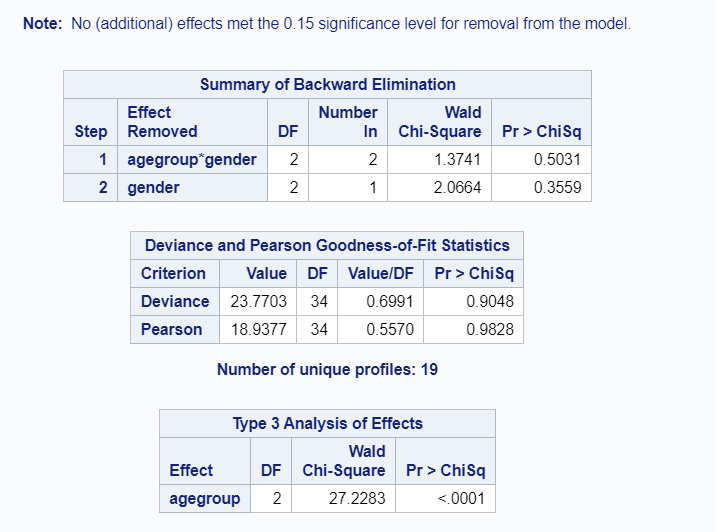


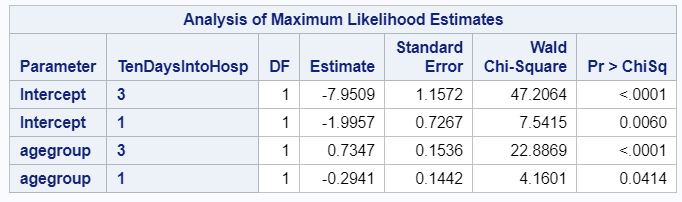
Shown below are the predicted probability plots for males and females. Both plots show that the predicted probability of recovery was highest among younger age groups, having a max of around .35 to .4 for people age 20 or younger. The probability of recovery steadily declined as age group increased. For the predicted probability of death, people who were age 60 or older had a relatively high chance of dying, with predicted probabilities between .2 and .4. Both graphs appear to have higher maximum predicted probabilities compared to the other models, which show that as time spent in hospital increases, the predicted probability of death or recovery also increases. Finally, it appears that there is no estimated probability for males in the age 10 group. For age group 9, which corresponds to an age range between 81 to 90 years old, males have around a 40% probability of death, compared to females age 81 to 90 years old, which have around a 30% probability of death. It can be concluded that male probability of death is significantly higher than females.



There were two other models that were investigated, 10 days in hospital and 20 days in hospital. Since the limit for this report is 10 pages, detailed paragraph summaries of the output won’t be covered for these models, because these models are very similar to the other models presented earlier. Instead, only the output will be presented.

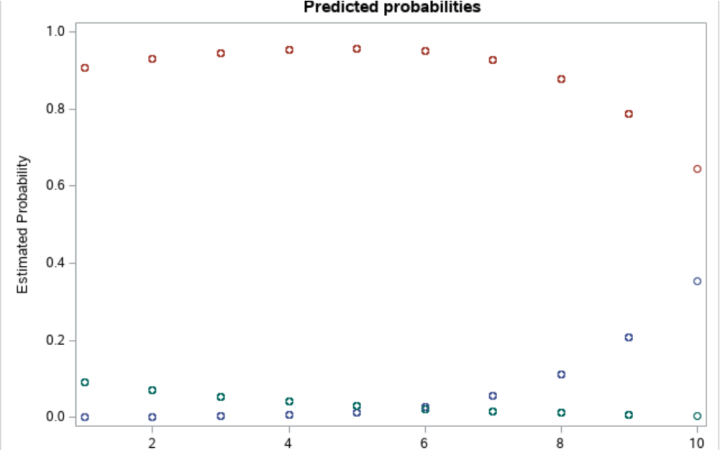
10 days in hospital model:

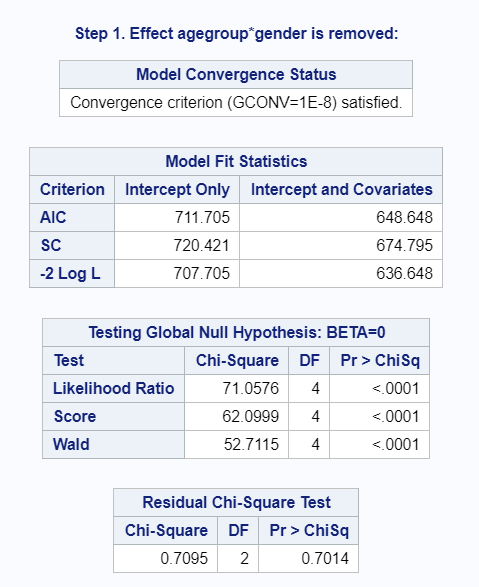
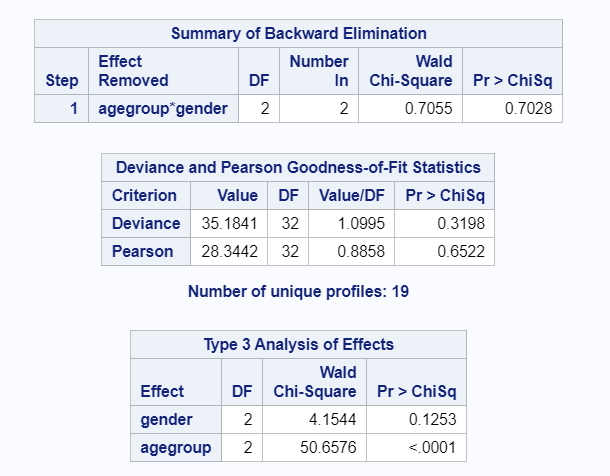


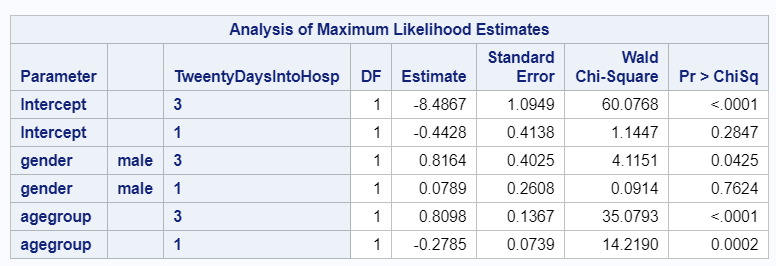
log(P(Y=1)/P(Y=2)) = -1.9957-.2941\*I{agegroup=1}

log(P(Y=3)/P(Y=2)) = -7.9509+.7347\*I{agegroup=3}



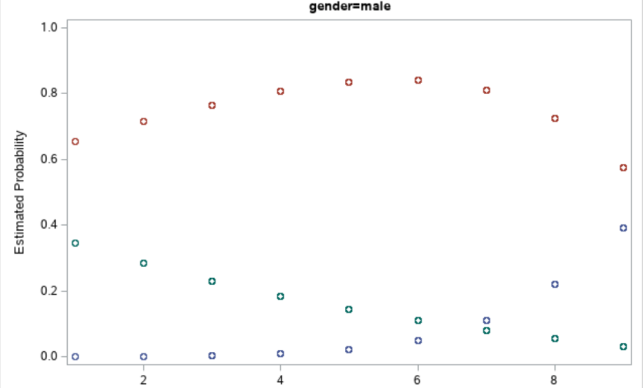
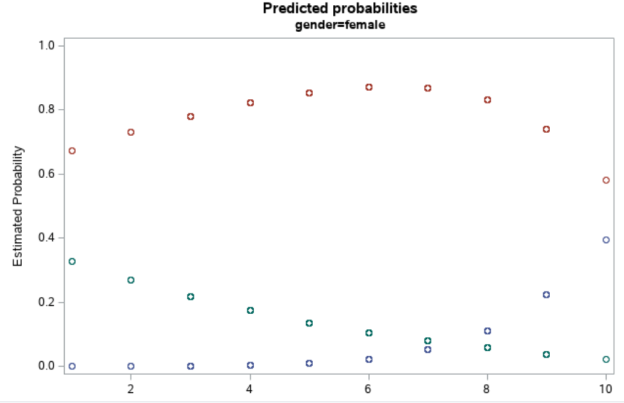
20 days in hospital model:



log(P(Y=1)/P(Y=2)) = -.4428+-.2785\*I{agegroup=1}+.0789\*I{gender|male=1}

log(P(Y=3)/P(Y=2)) = -8.4867+.8098\*I{agegroup=3}+.8164\*I{gender|male=3}

**Conclusion:**

Overall, based on all of the models presented in this project, it appears that the probability of death and recovery increase as time spent in hospital increases. It also appears that there is a significant probability for all age groups to either die or recovery within 20 to 25 days of being in a hospital, compared to smaller time durations. Males are more likely than females to die from COVID-19. Younger age groups (age 0 to 30) have a higher chance of recovery than older age groups (age 70 or older). Older age groups (age 70 or older) have a higher chance of death than younger age groups (age 0 to 30).

One flaw of this analysis is that probability of death seems a little too high for older age groups (around 30% probability of death for those age higher than 90). There are only two observations that are higher than 90 years old. Collecting more data on people age 90 or higher might portray a more accurate probability for that age group.

Continuing on the point made in the previous paragraph, in all only 595 observations were analyzed. Some age groups were very underrepresented in this analysis. Getting more observations evenly distributed in age would make for a more accurate prediction model than the current one proposed.

Finally, although this model captured how the death and recovery probability increased as time duration increased, with having death and recovery probabilities around 30 to 40% for different age groups, it would be useful to have analyses that would describe the average time for patients to recover or die from the disease, instead of using logistic models of different time periods to capture which time intervals are when patients are likely to recover or die from the disease. This kind of analysis is probably outside the scope of this course, and is probably covered in a survival analysis course. Nonetheless, a multinomial logistic regression model has proved to be adequate at analyzing certain aspects of death and recovery from COVID, as was described in earlier paragraphs.