Final Project of MA4720 – Design and Analysis of Experiments

Brandon Coates

Analysis on Impact of Race and Geographical Location on COVID-19 Vaccination Rates in the United States

Section 1.0: Introduction

In a race to achieve herd immunity to end the Covid-19 pandemic that has been plaguing not only the United States, but the entire world, the United States government has been investigating ways to entice the citizens of this country to receive one of the multiple Covid-19 vaccinations approved by the government. However, with slowing vaccination rates across the country, there may be regional or racial concerns causing vaccine hesitancy. This analysis aims to address the effect of geographical region in the United States and race on the percentage of vaccinated individuals in the region. This information is useful in the sense that it could shed light on which races seem to have the most hesitancy to receive the Covid-19 vaccination, the regions where vaccination rates are lower, and if there is a correlation the two that would affect the chance that an individual receives the vaccination.

This study will examine these two main factors, the geographical regions of the United States in terms of the Midwest, Northeast, South and West, coded as 1, 2, 3, and 4, respectively, as seen in **Figure 1**. The races that will be examined in this analysis are White, African American, Hispanic, and Asian, which were coded as 1, 2, 3, and 4 respectively. The data used in this experiment was provided by the Kaiser Family Foundation (KFF)[1], in which vaccination rates for 41 states were listed in terms of percent vaccinated by race. To generate the necessary data, a list of all of the states in each level of the geographical region was made in alphabetical order, then randomization is applied by a random number generator which was used to select 7 of the states to be used in this analysis. This means that each level has 7 replications which all provide the response variable which is a percentage of vaccination rate. The data used in this experiment were collected on June 7, 2021, so the trend on these variables over time will not be considered in this analysis. The experimental units and the observational unit will be the citizens of the United States.

Section 2.0: Statistical Methods

The statistical programming language R was used within the RStudio suite for all of the statistical analyses that were conducted on the data set obtained for this analysis. For each analysis conducted in the subsequent sections, a significance level of 0.05 was used.

Section 2.1: Exploratory Data Analysis

Before any analysis of interest is conducted on the data set obtained for this experiment, an interaction plot for the factors of geographical region and race which will be used to investigate any interaction effects between these two factors. Box plots for each treatment combination will also be constructed in order to observe the means and variances across all of the possible combinations which may shed light onto the possible results of the ANOVA analyses that will be carried out.

Section 2.2: ANOVA

In order to analyze the data, a two-way complete model for geographical region and race will be constructed using the type I sum of squares calculations since the model is balanced. This model will be used to test if all treatment factors have the same effect on the response of vaccinated population, as well as the levels of the geographical region and the levels of race. The assumptions for the ANOVA model will be checked by examining residual plots of the data set along with conducting Levene's test to

[1] https://www.kff.org/coronavirus-covid-19/issue-brief/latest-data-on-covid-19-vaccinations-race-ethnicity/

determine if the assumption of equal variance is true for this data. Contrasts will also be used to examine the pairwise differences in the effect of geographical region as well as race. For these calculations, 95% confidence intervals will be constructed using the Tukey method for multiple comparisons.

Section 3.0: Results and Conclusion

During the initial exploration stage, the data was plotted in both an interaction plot and parallel box plot to gain an understanding of some of the results that will be found within the ANOVA test conducted. These plots can be seen in **Figure 2**, where the interaction plot is on the left and the parallel box plot is on the right.

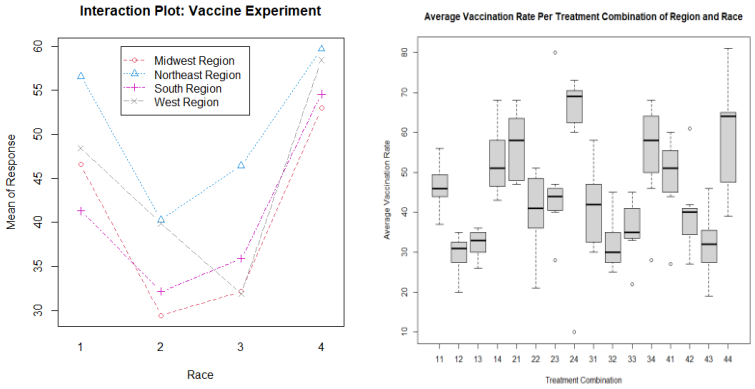

Figure 2: Exploratory Interaction Plot (Left) and Parallel Box Plot (Right)

The interaction plot suggests that there are several points in which interactions may occur, however, since the data is taken on a region to region basis, these intersections may not have an impact on each other. Similarly, the parallel box plot suggests that there is a high level of differences in means and variances across the treatment combinations. There also is evidence of several possible outliers that may cause violations in the assumptions needed for the ANOVA analysis.

An initial ANOVA analysis was conducted on the data and the results can be seen in **Table 1** below. This two-way complete model considers the two main factors of the data set, region and race each with 4 levels, as well as the interaction effects between the two. From this table, it was found that when examining the regional factor, there was sufficient evidence to support that the vaccination rates differed across the four regions of the United States since the F-test statistic and p-value was 5.009 and 0.0029, respectively. When examining the effects of race, there was enough evidence to support that the vaccination rate differed across the four races examined since the F-test statistic was 21.874 and the p-value was 7.049e-11. Lastly it was found that there was not enough evidence to support that the interaction effect between the two differed across all of the races and regions since the p-value was 0.7657.

Table 1: ANOVA on Two-Way Complete Model of Vaccination Data

	Degrees of Freedom	Sum of Squares	Mean Sum of Squares	F-test Statistic	p-value
Region	3	1928.7	642.92	5.009	0.0029
Race	3	8421.8	2807.27	21.874	7.049e-11
Region:Race	9	732.0	81.33	0.6337	0.7657
Error	96	12320.6	128.34		
Total	111	23403.1			

The assumptions for the above ANOVA analysis were checked in **Figure 3**. The first assumption that was examined was the assumption of model fit was checked. In order to check this assumption, a plot of the standardized residuals and treatment level was constructed in R (**Figure 3a)**. From this plot it was

observed that the majority of the residuals exhibited a nonrandom pattern about the zero line of this plot, however there were some residual points that were considered outliers since their values are above 3 and below -3 on the y-axis. Since the majority of points are within the standard residual ranges, it can be concluded that the assumption of model fit is retained for this data set.

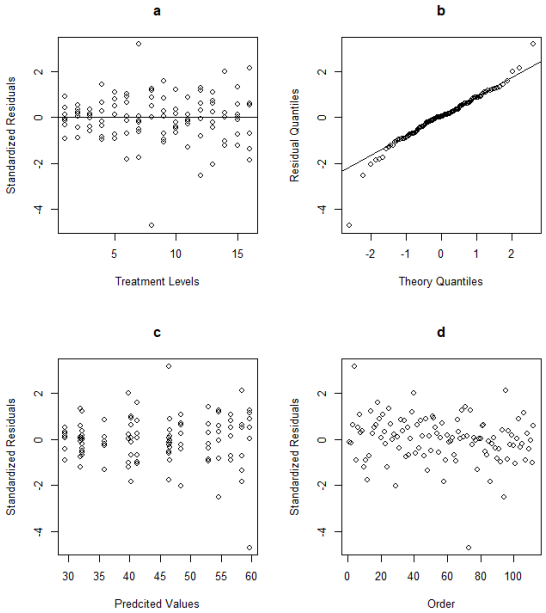

Figure 3: Residual Plots to Check Model Assumptions

The assumption of normality was the next also checked using a quantile-quantile plot (**Figure 3b**). It was observed that the residual points followed along the linear trendline except for at the right and left sides of the plot, where outliers were present. The assumption of normality is violated since the residuals no longer follow a linear trend. The assumption of equal variance was tested using a standardized residual versus predicted values plot (**Figure 3c**). It was observed in this plot that there was a slight fan effect in the residuals, meaning that as the predicted value increased, the variance in the points also increased. To properly test the assumption, Levene's test was conducted on the data, which found that the assumption was not violated.

The pairwise contrasts were used to create 95% confidence intervals for each of the main factors. For the region factor, it was found that only two contrasts had a significant difference, which were Midwest-Northeast (-10.46, 95% CI (-19.18, -1.75)) and Northeast-South (9.79, 95% CI (1.07, 18.50)). For the race factor, four pairwise contrasts had a significant difference which were: White-African American (12.79, 95% CI (4.07, 21.50)), White-Hispanic (11.64, 95% CI (2.93, 20.36)), African American-Asian (-21, 95% CI (-29.71, -12.29)), and Hispanic-Asian (-19.86, 95% CI (-28.57, -11.14)).

Section 4.0: Discussion

From the analysis conducted in this experiment, there are several conclusions that can be made. It was found that there is enough evidence to suggest that there are differences in average vaccination rates within the various regions of the United States. However, when examining regions to each other, it was found that there were only two regions that showcased statistical differences, which were Midwest-Northeast and Northeast-South. When examining the racial factors, there was enough evidence to suggest that there are differences in average vaccination rates across races. When comparing races to each other, there were four instances in which there were statistical differences between vaccination rates for

different races. So, it was found that the race of a person seems to have a larger effect on if the person will receive one of the vaccines for Covid-19, not region. However, this study should be refined further with more data and higher significance levels to increase the overall power of the experiment to further refine the understanding of where and which groups of the United States population are refusing the vaccine.