Final Project of MA4270 – Design and Analysis of Experiments

Kasia Krueger

Design and Analysis of Experiments for Esophageal Cancer Patients: Cancer in Different Lifestyle Groups

Section 1 Introduction

Esophageal cancer is one of the most common types of cancers in the United States (Esophageal, 2021), and the survival rate for esophageal cancer decreases with increasing age. The objective of this experiment is to study the relationship between case occurrence and age, alcohol and/or tobacco use. This data is retrieved from the R program (R Core Team, 2020) in the *esoph* dataset titled, "Esophageal Cancer in France". The data are obtained from 1175 patients with various ages and daily alcohol/tobacco use combinations. The data contain 200 cancer cases and 975 non-cases sampled from comparable hospital populations. The purpose of this experiment is to observe if different age groups, and different amount of tobacco use and alcohol use has a direct effect on the number of cases of esophageal cancer.

The data has 3 factors, daily alcohol use, daily tobacco use, and age group. There are 6 levels for Age Group: 25–34 years old, 35–44 years old, 45–54 years old, 55–64 years old, 65–74 years old, and 75+ years old. Alcohol Use has 4 levels: 0–39 gm/day, 40–79 gm/day, 80–119 gm/day, and 120+ gm/day. Tobacco Use has 4 levels: 0– 9 gm/day, 10–19 gm/day, 20–29 gm/day, and 30+ gm/day.

The response variable is the number of cases of esophageal cancer that occurred in these patients. By applying simple random sampling from 1175 patients in a hospital, the researcher applied randomization by sampling the entire hospital population, and not just those in a cancer ward. This allows for valid statistical analysis of the data. Replication is applied by surveying many different age groups in the hospital. Although the number of esophageal cancer cases can be affected by other factors, such as gender and other health complications, those factors were not considered. In addition, the blocking factors are not considered here. The experimental units are one of the 1175 patients surveyed, and the observational unit is one age group.

Section 2 Statistical Methods

All analyses have been conducted using RStudio Version 1.3.1093, using an overall significance level of 0.05 (95% confidence). The analysis in this report is conducted using an analysis of variance model using *aov* in the stats package (R Core Team, 2020) and statistical inference for pairwise contrasts using *lsmeans* (Russel 2016).

Section 2.1 Exploratory Data Analysis

Box plots are constructed for the number of esophageal cancer cases for each factor for exploring the data and making a hypothesis. A three-way interaction plot for the is constructed and used to investigate is there are any interaction effects between the three factors.

Section 2.2 ANOVA

The analysis of variance model with all two-way interactions is constructed to test if there is an effect on the response based on the three factors and their interactions. Simultaneous confidence intervals of pairwise contrasts using the Tukey method of pairwise comparisons will be constructed to determine which levels have a different effect on the number of esophageal cancer cases. The residuals will be obtained, and assumptions of equal variance, normality, independence, and outliers will be checked using appropriate residual plots.

Section 3 Results and Conclusion

The box plots in Figure 1 show the mean number of cancer cases for each factor. The box plots show the mean number of cases is close for alcohol use and tobacco use, but differs more for age group, although it's not clear if there is an actual difference. The box plots also show there are a few outliers in each factor. From this exploration, the first null hypothesis is that there is no difference in the number of esophageal cancer cases from the different factors and interactions. The alternative hypothesis is at least one factor or interaction has a different effect on the response. The three-way interaction plots shown in Figure 2 show Age as the third factor. There does not appear to be a strong effect in three-way interaction due to the frequent parallel lines between the two other factors, so the three-way interaction is determined to be negligible and will not be included in the linear model.

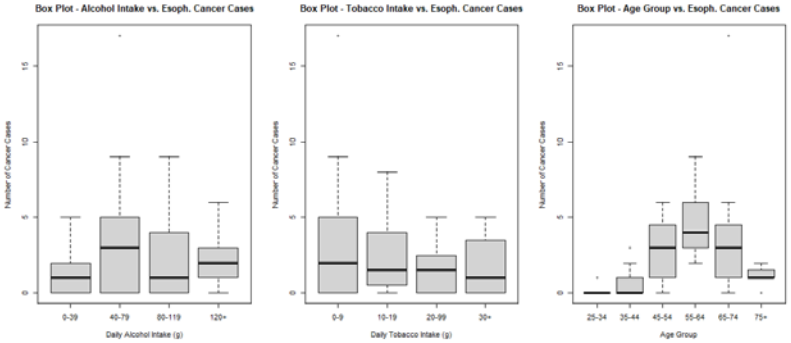

Figure 1, Box Plots for Main Factors

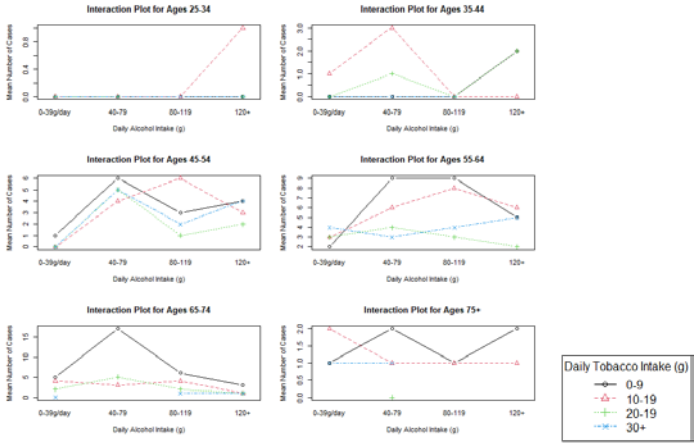

Figure 2, Three-way Interaction Plots for Factors

Table 1 shows the analysis of variance table for the Type III sum of squares that was obtained from the two-way complete model. From Table 1, an overall test was obtained for analyzing if the cancer cases differed across four ranges of daily alcohol use, four ranges of daily tobacco use, and six age groups. The corresponding F-statistics and p-values are 1.3714 and 0.236083, respectively, for the interaction of alcohol and tobacco use, and 1.8950 and 0.057106, respectively, for the interaction of tobacco use and age. The p-values for the main effects of alcohol use, tobacco use, age, and the interaction of age and alcohol are 0.001173, 0.004120, 1.13e-09, and 0.015714, respectively. Therefore, at a significance level of 0.05, we have sufficient evidence to support that the esophageal cancer case rate differed across four ranges of daily alcohol use, four ranges of daily tobacco use, and six age groups. The final model used for experiment includes the non-negligible effects alcohol use, tobacco use, age group, and the interaction between alcohol use and age.

Table 1, Two-way Analysis of Variance Table

Factor	Degrees of Freedom	F value	Pr(>F)
Alcohol Use (fA)	3	6.5317	0.001173
Tobacco Use (fB)	3	5.2316	0.004120
Age group (fC)	5	20.1972	1.13e-09
fA:fB	9	1.3714	0.236083
fA:fC	15	2.3900	0.015714
fB:fC	15	1.8950	0.057106
Residuals	37		

Since there is difference in response for each factor, the secondary null hypothesis is there is no difference between the levels of each factor on the effect on the number of cases of esophageal cancer. Simultaneous confidence levels of pairwise contrasts are constructed with the pairwise Tukey method with a 95% significance level. For space, only the statistically significant pairwise contrasts are presented for the age factor.

Table 2, Simultaneous Pairwise Contrasts and Confidence Intervals for Each Factor

Daily Alcohol Use				Daily Tobacco Use				Age group		
contrasts	**95% CI**	**p-val**		**contrasts**	**95% CI**	**p-val**		**contrasts**	**95% CI**	**p-val**
(0-39g/day) - (40-79)	(-3.614, -0.839)	0.0004		(0-9g/day) - (10-19)	(-0.515, 2.18)	0.3683		(25-34) - (45-54)	(-4.7617, -1.0036)	0.0004
(0-39g/day) - (80-119)	(-2.272, 0.610)	0.4297		(0-9g/day) - (20-29)	(0.495, 3.38)	0.0041		(25-34) - (55-64)	(-6.6367, -2.8786)	<.0001
(0-39g/day) - (120+)	(-2.140, 0.747)	0.5821		(0-9g/day) - (30+)	(0.168, 3.05)	0.0228		(25-34) - (65-74)	(-5.8425, -2.0037)	<.0001
(40-79) - (80-119)	(-0.048, 2.839)	0.0619		(10-19) - (20-29)	(-0.34, 2.54)	0.1915		(35-44) - (45-54)	(-4.1510, -0.3929)	0.0091
(40-79) - (120+)	(0.088, 2.971)	0.0334		(10-19) - (30+)	(-0.666, 2.21)	0.4917		(35-44) - (55-64)	(-6.0260, -2.2679)	<.0001
(80-119) - (120+)	(-1.358, 1.627)	0.9952		(20-29) - (30+)	(-1.830, 1.17)	0.9389		(35-44) - (65-74)	(-5.2268, -1.3982)	0.0001
								(45-54) - (55-64)	(-3.7142, -0.0358)	0.0432
								(55-64) - (75+)	(1.8325, 6.0684)	<.0001
								(65-74) - (75+)	(0.9685, 5.2635)	0.0009

Table 2 shows the statistically significant confidence intervals for the pairwise contrasts of each factor. For contrasts in daily alcohol use, the number of cases decreases when comparing the lowest intake of alcohol per day to the higher

intake of alcohol per day. There is sufficient evidence to conclude there is statistically significant difference in response between 0-39 and 40-79 grams of alcohol per day and 40-79 and 120+ grams of alcohol per day. At a 95% confidence level, the number of cases decreases between 0-39g and 40-79 grams of alcohol per day by -3.614 to -0.839 cases and the number of cases increases between 40-79g/day and 120+ grams of alcohol per day by 0.088 to 2.971 cases. All other confidence intervals contain zero, so there is not sufficient evidence that there is a difference in response. For contrasts in daily tobacco use, the number of cases increases when the lowest amount of tobacco use per day is compared to the higher amounts of tobacco per day. There is sufficient evidence to conclude there is statistically significant difference in response between 0-9 and 20-29 grams of tobacco per day and 0-9 and 30+ grams of tobacco per day. At a 95% confidence level, the number of cancer cases increases when patients who use 0-9 grams of tobacco per day and 20-29 grams per day by 0.495 to 3.38 cases and the number of cases increases between 0-9 grams and 30+ grams of tobacco per day by 0.168 to 3.05 cases. For age group, the number of cases decreases when younger age groups are compared to older age groups, and the number of cases increases when older groups are compared to the oldest group, 75+ years old. For age group contrasts (25-34) and (45-54), (25-34) and (55-64), (25-34) and (65-74), (35-44) and (45-54), (35-44) and (55-64), (35-44) and (65-74), and (45-54) and (55-64), the number of cases of esophageal cancer decreases by -4.7617 to -1.0036 cases, -6.6367 to -2.8786 cases, -5.8425 to -2.0037 cases, -4.1510 to -0.3929 cases, -6.0260 to -2.2679 cases, -5.2268 to -1.3982 cases, and -3.7142 to -0.0358 cases, respectively. For age group contrasts (55-64) and (75+), and (65-74) and (75+), the number of esophageal cancer cases increases by 1.8325 to 6.0684 cases and 0.9685 to 5.2635 cases, respectively. The conclusion is that there is sufficient evidence to conclude there is a difference between the levels of each factor on the effect on the number of cases of esophageal cancer.

Figure 3 shows the model fit of residuals to check assumptions of the model. Using the residuals plots from Figure 3, three maximum standard residuals (6.406573, 3.391715, 3.391715) are removed as outliers from the 88 total residuals and a new data set is formed, *esoph.without.max.resid*.

Figure 4 shows the assumptions for the model do not appear to be violated. The normality assumption for the new data set holds as the QQ-plot shows close to a 45-degree angle line and fairly linear distribution of residual and theoretical quantiles. The independence assumption does not exhibit any patterns or outliers and overall equal variances appear to be normally distributed with mean zero and equal variance among the residuals..

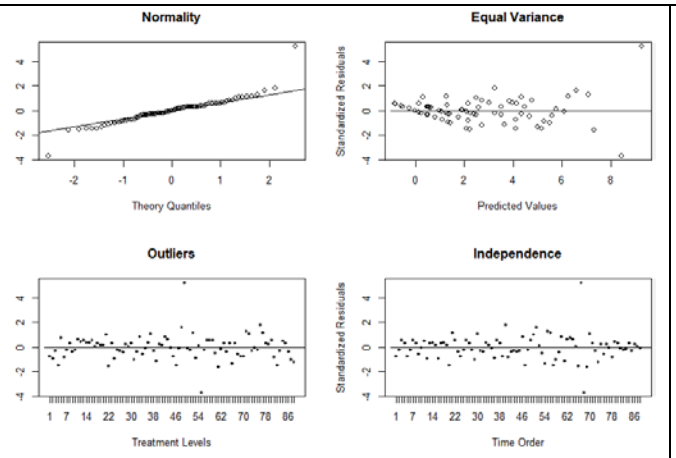

Figure 3, Checking Model Fit of Residuals: (1) normal Q-Q plot for standardized residuals; (2) scatter plot for standardized residuals and predicted values (3) scatter plot for standardized residuals and treatment combinations; (4) scatter plot for standardized residuals and order.

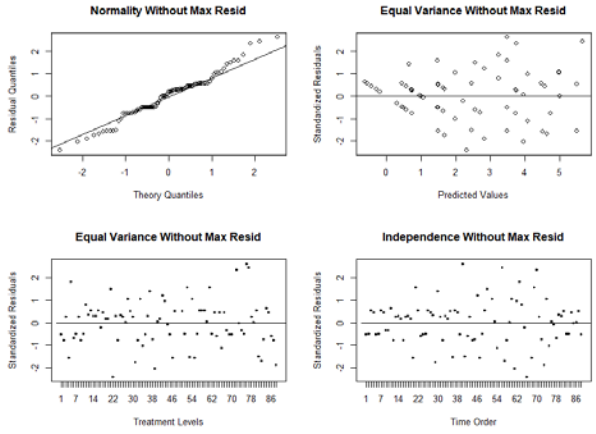

Figure 4, Checking Model Fit of Residuals without Maximum Residuals (1) normal Q-Q plot for standardized residuals; (2) scatter plot for standardized residuals and predicted values (3) scatter plot for standardized residuals and treatment combinations; (4) scatter plot for standardized residuals and order**.**

Section 4 Discussion

The analysis of this experiment found there is sufficient evidence to support there is a statistically significant difference on the response from different age groups and different amounts of tobacco use and alcohol use, and that there is sufficient evidence to support there is a statistically significant difference on the response from the different levels of each factor. The motivations for the experiment are met and show that any amount of alcohol and tobacco may affect one's risk of esophageal cancer, and age differences increases this risk as well.

References

Esophageal Cancer - Statistics. Cancer.Net. (2021, May 27). https://www.cancer.net/cancer-types/esophageal-cancer/statistics.

Oesophageal cancer survival statistics. Cancer Research UK. (2020, April 7). https://www.cancerresearchuk.org/health-professional/cancer-statistics/statistics-by-cancer-type/oesophageal-cancer/survival.

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Russell V. Lenth (2016). Least-Squares Means: The R Package lsmeans. Journal of Statistical Software, 69(1), 1-33. doi:10.18637/jss.v069.i01