Regression Analysis of Esophageal Cancer Risk Factors

Anabel Costa, Mia Josephy-Zack, and Mila Lewis-Peinado
anabelcosta@umass.edu, mjosephyzack@umass.edu, mlewispeinad@umass.edu
STAT625 Final Project Report - Fall 2024
December 15th, 2024

Contribution Percentages: Anabel (0.33), Mia (0.33), Mila (0.33)

Abstract

The goal of this research project is to analyze the "esoph" data set in R to model the effects of age, alcohol consumption, and tobacco consumption on the development of esophageal cancer. This report begins by contextualizing the "esoph" data set, and then visually analyzing its variables in relation to the response. From there, our team modeled the outcome of the development of esophageal cancer based on its predictors through logistic regression, and then performed forward and backward stepwise regression for variable selection. Throughout this process, our team examined various measures of our models such as the Akaike Information Criterion (AIC) and Variance Inflation Indicator (VIF), as well as performed Analysis of Deviance testing and influential point detection to examine potential issues in our models such as poor fit, collinearity, and influential cases.

Introduction

In France, esophageal cancer ranks third among the most common forms of digestive cancer (Plessen, 2024). This disease arises when cancerous cell growth occurs within the esophageal tissue lining. For many patients, this type of cancer initially goes unnoticed and built-up cell growth eventually obstructs the opening to the esophagus, causing significant discomfort and long-term health consequences. Esophageal cancer is more likely to be seen in individuals who are both over the age of sixty and assigned male at birth (Cleveland Clinic, 2024).

In 1981, nearly 3,500 French men and women were surveyed across 65 geographical regions in the country regarding their smoking and drinking habits. Specifically looking into the results of the male respondents, more than eighty percent of them expressed that they had previously smoked cigarettes or currently smoked cigarettes (Wynder et al. 1981). Additionally, older male smokers at this time tended to use black or dark tobacco cigarettes, which are stronger than standard tobacco cigarettes with higher nicotine levels (Wynder et al. 1981). Researchers of this study highlighted that esophageal cancer is more commonly seen in males in France as opposed to lung cancer. General public health research has found that alcohol consumption and smoking tobacco are two prominent risk factors for esophageal cancer in adults. In our analysis, we aim to confirm the significance of alcohol and tobacco consumption in developing esophageal cancer.

Data Source

The data that we are using is a built-in R dataset called "esoph". This data originated from Chapter 4 and Chapter 6 of *Statistical Methods in Cancer Research: Volume 1 – The Analysis of Case-Control Studies* by Breslow and Day (1980). The data was collected in a case-control study conducted by Tuyns et al. (1977) in Ille-et-Vilaine, France. 200 men diagnosed with esophageal cancer at a regional hospital between January 1972 and April 1947 were interviewed along with 778 men without esophageal cancer drawn from electoral lists in the

same region. The subjects were interviewed at either the hospital or at home depending on whether they were in the case or control group. The researchers collected information about their lifestyle choices and recorded their daily alcohol consumption, daily tobacco consumption, and to what age group they belonged.

Figure 1.
A Glance at Our Data

	agegp <ord>	alcgp <ord>	tobgp <ord>	ncases <dbl>	ncontrols <dbl>
1	25-34	0-39g/day	0-9g/day	0	40
2	25-34	0-39g/day	10-19	0	10
3	25-34	0-39g/day	20-29	0	6
4	25-34	0-39g/day	30+	0	5
5	25-34	40-79	0-9g/day	0	27
6	25-34	40-79	10-19	0	7

6 rows

The dataset consists of 88 rows representing combinations of age group, alcohol consumption, and tobacco consumption. The first six rows of our data can be seen in Figure 1. Each row provides information on the number of cancer cases and controls for a specific combination of these variables. To get the total number of observations for each combination, we can sum all the cases and controls in that row. There are six different ordinal age groups: 25 - 35 years old, 35 - 44 years old, 45 - 54 years old, 65 - 74 years old, and 75 or older. There are four different ordinal groups for alcohol consumption: 0 - 34 grams a day, 40 - 79 grams a day, 80 - 119 grams a day, and 120+ grams a day. There are also four different ordinal groups for tobacco consumption: 0 - 9 grams a day, 10 - 19 grams a day, 20 - 29 grams a day, and 30+ grams a day. There are two more variables, “ncases” and “ncontrols”, which are the number of people with esophageal cancer and the number of people without esophageal cancer respectively for that specific combination of alcohol and tobacco consumption in that age group.

By analyzing this dataset, we aim to answer the central question, “How do age, alcohol consumption, and tobacco consumption influence the risk of esophageal cancer in French males?”

Proposed Methodology

We propose to use exploratory data analysis, logistic regression, and variable selection techniques to investigate further how these three variables contribute to the development of esophageal cancer in males. First, an exploratory data analysis will be performed to understand data distributions and initial patterns among the variables. Barplots will visualize the distribution of the relative frequency of cancer cases across predictor groups, and mosaic plots will explore positive and negative associations between variables and highlight significant relationships. We will also calculate odds ratios to identify combinations of alcohol and tobacco consumption that correspond to higher or lower esophageal cancer risk.

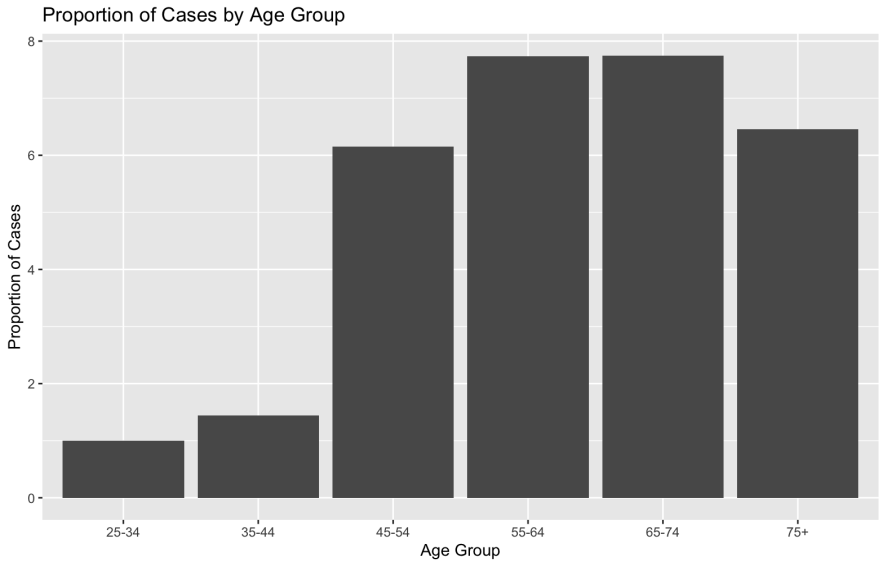
Following our exploratory data analysis, we will create multiple linear models and compare their effectiveness in addressing our central research question. For these different models, we will look at components such as their estimated coefficients and their significance, results from Analysis of Deviance testing, Variance Inflation Factor (VIF), and any notable outliers in the data.

We decided to use logistic regression for the "esoph" dataset as we are interested in the probability of an event occurring or not occurring. Instead of predicting a specific value like linear regression, logistic regression uses a sigmoid function to map predicted values to probabilities between 0 and 1. We also chose to use Analysis of Variance (ANOVA) testing to compare the fit of our models and Variance Inflation Factor (VIF) in order to look at the collinearity of our predictors in our models.

Analysis and Results

Exploratory Data Analysis

Figure 2.

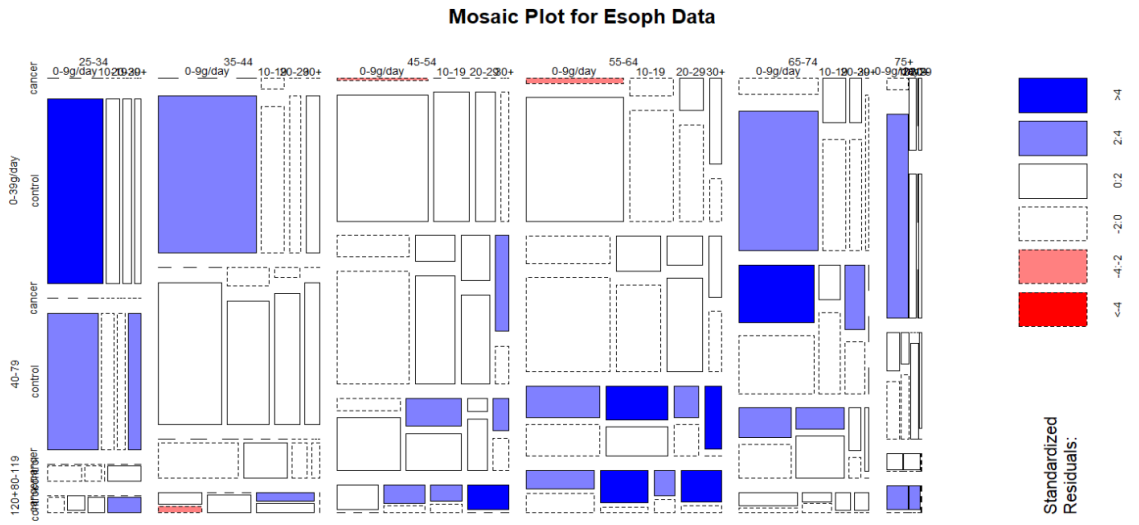


We started by conducting an exploratory data analysis to examine our dataset. We first looked at the distributions of our variables: age group, alcohol consumption, and tobacco consumption. Since the dataset is grouped by combinations instead of individual observations, the barplots visualize the frequency of observations for each variable. This tells us how each group within a variable is represented in the data. We can see that the distribution of this is roughly uniform (see Figures A1, A2, A3). This means that our dataset is well-balanced and doesn't seem to have any bias towards any group. After this, we were curious about the proportion of cases to the total number of cases and controls in each group for our variables. We calculated the proportion for each combination and visualized it in barplots for each variable.

This calculation gave us insight into the relative frequency of cases within each group for our variables. We can see in Figure 2, which shows the proportion of cases by age group, that the distribution is skewed left. The first two age groups have the lowest proportion of cases while the 55-64 and 65-74 groups have the highest proportion of cases. From this, we can predict that the likelihood of one getting esophageal cancer increases with age. The proportion of cases by tobacco consumption was uniformly distributed, except for the 20 to 29 grams a day grouping which was lower than all of the other groups (see Figure A4). Looking at the proportion of cases by alcohol consumption (see Figure A5), we can see that there is a linear increase in the barplot which tells us that there is a positive association between drinks per day and getting esophageal cancer.

We also decided to look into the odds ratio of our combinations to see what relationships we could uncover. We calculated the odds ratio of each combination of age group, tobacco consumption, and alcohol consumption, and then visualized them. The odds ratio plot by age group and tobacco consumption (Figure A6) shows the relationship between age group, alcohol consumption, and esophageal cancer risk. Higher odds ratios are observed in older age groups and for individuals consuming the highest levels of alcohol. In younger age groups and lower alcohol consumption categories, the odds ratios are notably smaller. This indicates that high alcohol intake, especially in older age groups, is strongly associated with increased odds of esophageal cancer. The other odds ratio plot was by age group and tobacco consumption (Figure A7). Higher odds ratios are observed in older age groups and among individuals with high tobacco consumption, suggesting a strong association between these factors and developing esophageal cancer. In younger age groups and lower tobacco levels, odds ratios are lower, with some values near zero or at zero.

Figure 3.



The mosaic plot above (Figure 3) reveals positive and negative associations between the variables in the dataset. For this plot, the cases have been separated into "cancer" and "control". Blue tiles with a solid outline indicate large, positive standardized residual values, and red tiles with a dashed outline indicate small, negative standardized residual values. Looking at the plot, we can see three red tiles or negative associations. One of the red tiles corresponds to observations falling under the following categories: "cancer", *tobgp* = 0-9 grams per day, *agegp* = 45-54, and *alcgp* = 0-39 grams per day. Another red tile corresponds to observations falling under these same categories, except the age group has increased to 55-64. Finally, the third red tile corresponds to the following categories: "control", *tobgp* = 0-9 grams per day, *agegp* = 35-44, and *alcgp* = greater than 120 grams per day. These red tiles all have standardized residual values between -4 and -2, and these patterns have a lower observed frequency.

There are several more positive associations reflected in this mosaic plot. For instance, one of the darker blue tiles corresponds to observations falling under the following categories: "control", *tobgp* = 0-9 grams per day, *agegp* = 25-34, and *alcgp* = 0-39 grams per day. Another darker blue tile corresponds to observations falling under the following categories: "cancer", *tobgp* = greater than 30 grams per day, *agegp* = 55-64, and *alcgp* = greater than 120 grams per day. These blue tiles all have standardized residual values greater than 4, and these patterns have a higher observed frequency. From this mosaic plot, we can see that old age, high alcohol consumption, and high tobacco consumption could potentially be high risk factors for esophageal cancer in males.

Regression Analysis: Main Effects Model

Applying logistic regression to this dataset, we first created a main-effects model, *model_main,* involving all three main ordinal variable groups: "*agegp"*, "*alcgp"*, and "*tobgp"*. The response for this model is the aggregated number of cases and controls, in columns of values for *ncases* and *ncontrols*. Our second model, *model_interactions*, includes the variable *agegp* as well as a singular interaction term between *agegp* and *tobgp*. In the model, this interaction is noted by the "*" operator. The response for this model is the same as that for our main-effects model. Both models were created using the *glm()* function in R. The output for our main-effects model is shown below in Figure 4.

Figure 4.

```
Call:
glm(formula = cbind(ncases, ncontrols) ~ agegp + alcgp + tobgp,
    family = binomial, data = esoph)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -1.19039    0.20737  -5.740 9.44e-09 ***
agegp.L      3.99663    0.69389   5.760 8.42e-09 ***
agegp.Q     -1.65741    0.62115  -2.668  0.00762 **
agegp.C      0.11094    0.46815   0.237  0.81267
agegp^4      0.07892    0.32463   0.243  0.80792
agegp^5     -0.26219    0.21337  -1.229  0.21915
alcgp.L      2.53899    0.26385   9.623  < 2e-16 ***
alcgp.Q      0.09376    0.22419   0.418  0.67578
alcgp.C      0.43930    0.18347   2.394  0.01665 *
tobgp.L      1.11749    0.24014   4.653 3.26e-06 ***
tobgp.Q      0.34516    0.22414   1.540  0.12358
tobgp.C      0.31692    0.21091   1.503  0.13294
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 367.953  on 87  degrees of freedom
Residual deviance:  82.337  on 76  degrees of freedom
AIC: 221.39

Number of Fisher Scoring iterations: 6
```

In this model, the coefficient for the intercept (-1.19039) represents the log-odds for the baseline group (individuals that are between 25-34 years of age, ingest an average of 0-39 grams of alcohol per day, and ingest an average of 0-9 grams of tobacco per day). This is equivalent to saying that on average, individuals that fit into these three categories of these groups have approximately a 23.32% probability of developing esophageal cancer. When looking at the effects of age on the log-odds outcome, it is shown in Figure 4 that "agegp.L" and "agegp.Q" are both statistically significant with p-values less than 0.01. This means that there is strong evidence that we can reject the null hypothesis that these coefficients are zero. Additionally, Figure 4 shows that the coefficient for "agegp.L" is 3.99663 which is both positive and relatively larger than other coefficients in this model, while the coefficient for "agegp.Q" is -1.65741, which is negative. This would imply that older age groups are more likely to develop esophageal cancer but the marginal effect of age on developing esophageal cancer decreases as age increases. The higher-ordered polynomial terms have much higher p-values indicating that there is very little support that these coefficients are not zero, meaning that there is very likely no higher-order polynomial trend between age and esophageal cancer.

When looking at the effects of alcohol consumption on the log-odds of developing esophageal cancer, it is shown in Figure 4 that "alcgp.L" and "alcgp.C" have p-values below 0.05, indicating that there is strong evidence to reject the null hypothesis that these coefficients are zero. The coefficient for "alcgp.L" is 2.53899 which is positive and relatively larger than other coefficients, while "alcgp.C" is 0.4393 which is still positive but is much smaller than the coefficient for "alcgp.L". This indicates that higher levels of alcohol increase the log odds of developing esophageal cancer with a strong linear trend and an additional smaller cubic trend.

When looking at the effects of tobacco consumption on the log-odds of developing esophageal cancer, it is shown in Figure 4 that "tobgp.L" has a p-value extremely close to zero,

indicating that there is strong evidence to reject the null hypothesis that this coefficient is zero. The coefficient for "tobgp.L" is 1.11749, indicating that higher levels of tobacco increase the log odds of developing esophageal cancer in a linear fashion. The higher-ordered polynomial terms have much higher p-values indicating that there is very little support that these coefficients are not zero, meaning that there is very likely no higher-order polynomial trend in relation to tobacco and esophageal cancer.

Figure 5.

```
Analysis of Deviance Table (Type II tests)

Response: cbind(ncases, ncontrols)
      LR Chisq Df Pr(>Chisq)
agegp  126.488  5  < 2.2e-16 ***
alcgp  127.933  3  < 2.2e-16 ***
tobgp   23.544  3   3.11e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

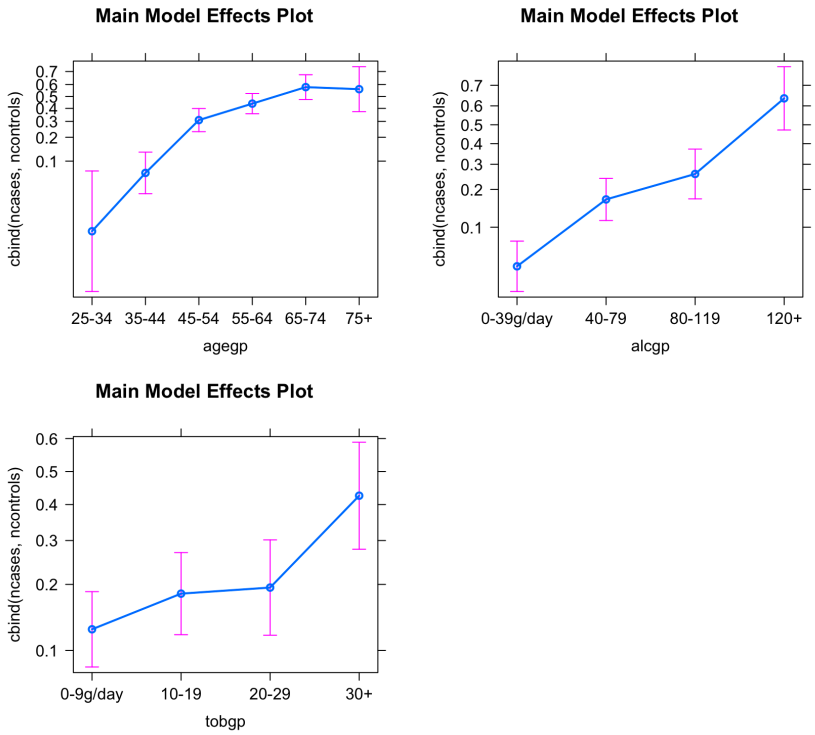
Shown above in Figure 5 is the Analysis of Deviance Table for our main effects model, where it is shown that the p-values for all of these variables are extremely close to zero. This means that after accounting for the other variables, each predictor contributes to the model's fit by increasing the likelihood of the model at a statistically significant level.

Figure 6.

```
             GVIF Df GVIF^(1/(2*Df))
agegp 1.172205  5        1.016016
alcgp 1.129318  3        1.020476
tobgp 1.096993  3        1.015548
```

Shown above in Figure 6 is the output for the Variance Inflation Factor. As seen above, the GVIF and the adjusted GVIF are only slightly above 1, implying that collinearity among these variables is not likely an issue.

Figure 7.



We also created an effects plot (see Figure A8) to further look into the relationship of our predictors and our response variable. For the age group variable, the probability of esophageal cancer increases steadily with age which highlights older age as a strong risk factor. For alcohol consumption, there is a clear positive trend and alcohol seems to also be a significant risk factor. Similarly, for tobacco consumption, the probability of esophageal cancer increases with higher consumption. These findings reinforce that age, alcohol, and tobacco consumption are positively associated with esophageal cancer risk. The pink error bars reflect the uncertainty in the estimates, with narrower intervals indicating greater reliability of predictions. We can see that the estimates for the age group and alcohol consumption variables are more reliable than the tobacco consumption variable.

Regression Analysis: Forward and Backward Stepwise Selection Models

The next step in our model was to explore whether or not a more complex model could improve the model's goodness of fit. To investigate this, we ran both forward stepwise regression that began with a null model, and backward stepwise regression that began with two-way interactions in the model. Notably, the AIC for the null model was 485.01, but both the forward and backward stepwise selection models ended with the same model and respective coefficients as our main effects models. While backward stepwise and forward stepwise regression are not known to work particularly well with categorical variables, we were still surprised that they resulted in the exact same model, and wanted to explore whether collinearity was a possible cause of interaction terms not being included in the final models. We then decided

to calculate the VIF for a model that included all interaction terms as well as main effects, and below is the result.

Figure 8.

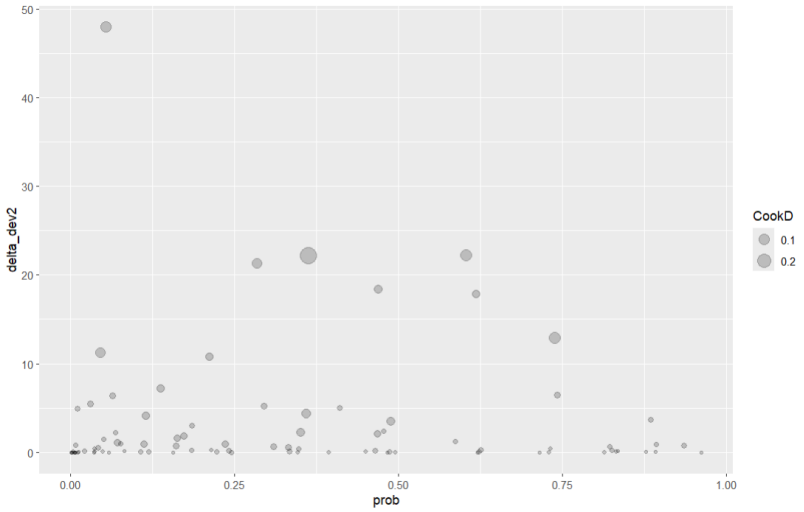
	GVIF	Df	GVIF^(1/(2*Df))
agegp	1.275990e+18	5	64.652409
alcgp	1.819080e+24	3	11048.635368
tobgp	6.364219e+23	3	9274.507499
agegp:alcgp	4.536790e+40	15	22.658176
agegp:tobgp	3.156048e+39	15	20.731853
alcgp:tobgp	2.002681e+01	9	1.181168

As shown above, many of the adjusted values for the VIF were far above 1, indicating extreme levels of collinearity among the regressors, and would explain how including these terms would not substantially improve the model's AIC.

Regression Analysis- Influential Point Analysis:

One aspect of the data that we believed was worth examining was whether or not there were any influential cases in our data. Shown below is a graph of our data in terms of the square of an observation's delta deviance, the fitted probability, and the associated Cook's distance. The delta deviance is the change in the deviance model when we include the case, so the overall delta deviance for a particular case gives an idea of the overall contribution of a particular case to a model's goodness of fit.

Figure 9.



We see in the graph below there is one case (case 13 which corresponds to (25-24 year olds, 120+ grams of alcohol per day, 10-19 grams of tobacco per day with ncase = 1, ncontrol = 0), that has a particularly high delta deviance squared and a low fitted probability. Given this, we

wanted to see the effects of removing this variable from our model, and then running the same main effects model. The coefficients to the model are shown below.

Figure 10.

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -3.87958  317.05554  -0.012   0.9902
agegp.L       13.63493 1136.86154   0.012   0.9904
agegp.Q      -10.46003 1037.80784  -0.010   0.9920
agegp.C        6.12427  708.95775   0.009   0.9931
agegp^4       -2.97293  359.50726  -0.008   0.9934
agegp^5        0.75650  119.83590   0.006   0.9950
alcgp.L        2.50302    0.26494   9.448  < 2e-16 ***
alcgp.Q        0.06037    0.22495   0.268   0.7884
alcgp.C        0.42592    0.18377   2.318   0.0205 *
tobgp.L        1.15872    0.24246   4.779 1.76e-06 ***
tobgp.Q        0.38536    0.22572   1.707   0.0878 .
tobgp.C        0.30569    0.21154   1.445   0.1484
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the above summary, we can see that removing the 13th case resulted in age no longer being considered statistically significant in the outcome of developing esophageal cancer.

Conclusions

Based on our findings, we determined that our main-effects model, *model_main*, is our best model for the data as there was a lot of collinearity in our models with interaction terms. Future directions for this research could involve creating predictive models to determine an individual's risk of developing esophageal cancer based on factors such as age, alcohol consumption, and tobacco consumption. This could be achieved by building a training and test dataset and utilizing lasso or ridge regression in our model. Applying either lasso and ridge regression would not be suitable for further analyzing any trends in the data, but these techniques would aid in predicting future values. Additionally, we found that removing (25-24 year olds, 120+ grams of alcohol per day, 10-19 grams of tobacco per day with ncase = 1, ncontrol = 0) changed our main effects model. However, given the limited nature of the data and surveying method, and specifically the limited number of observations within our data, it feels inappropriate to conclude that this reduced model is more appropriate. This supports the natural next step which would be finding newer and more comprehensive data about esophageal cancer, as this dataset is from the 1970s and may not reflect modern alcohol and tobacco consumption in males, as well as current esophageal cancer prevalence.

Appendix A

Figure A1

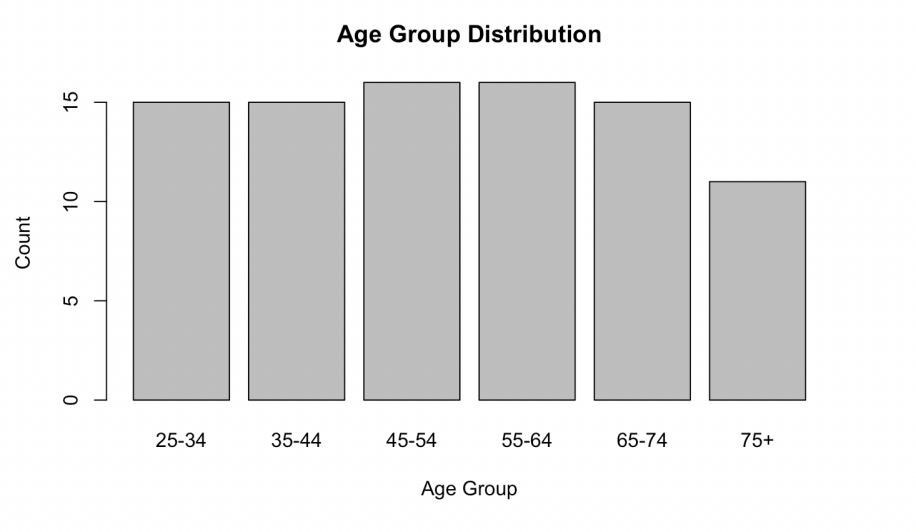


Figure A2

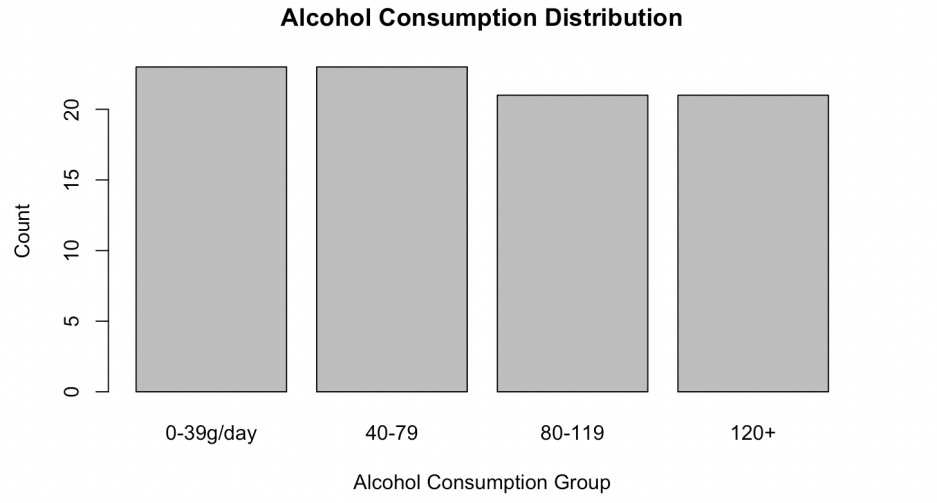


Figure A3

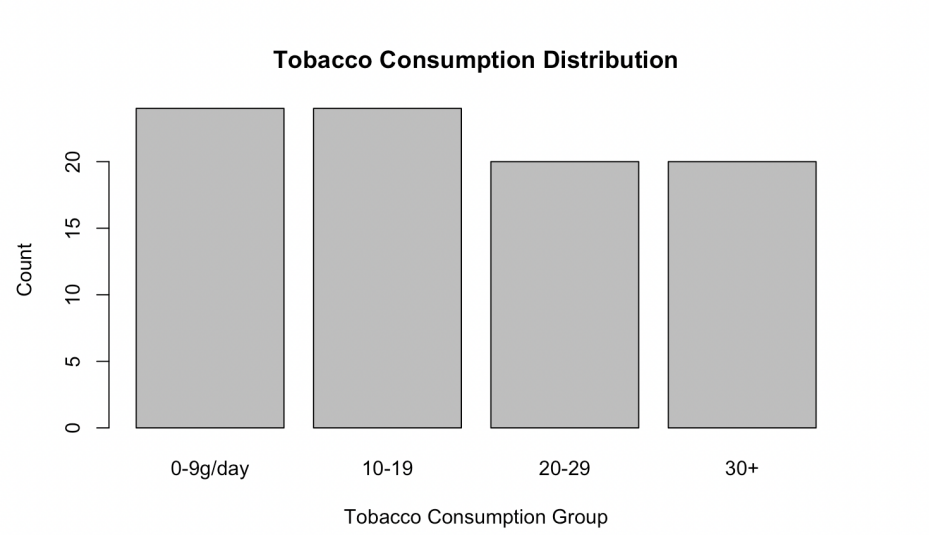


Figure A4

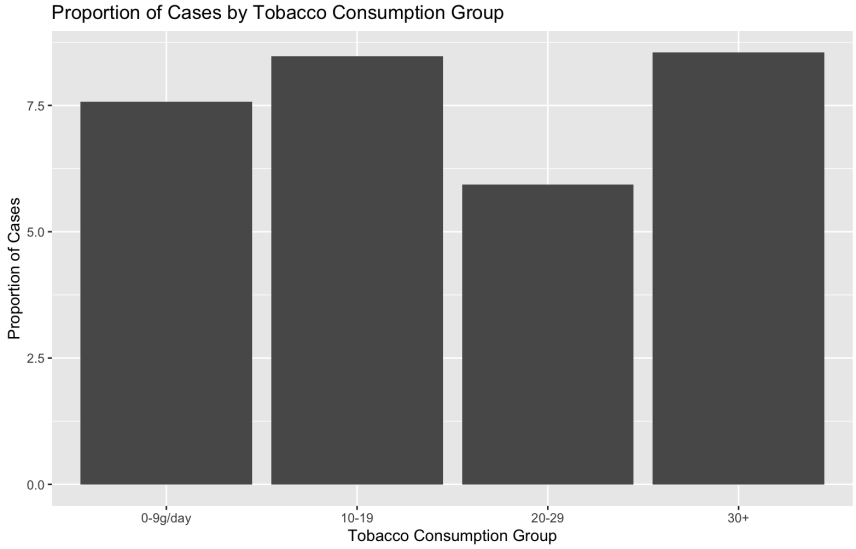


Figure A5

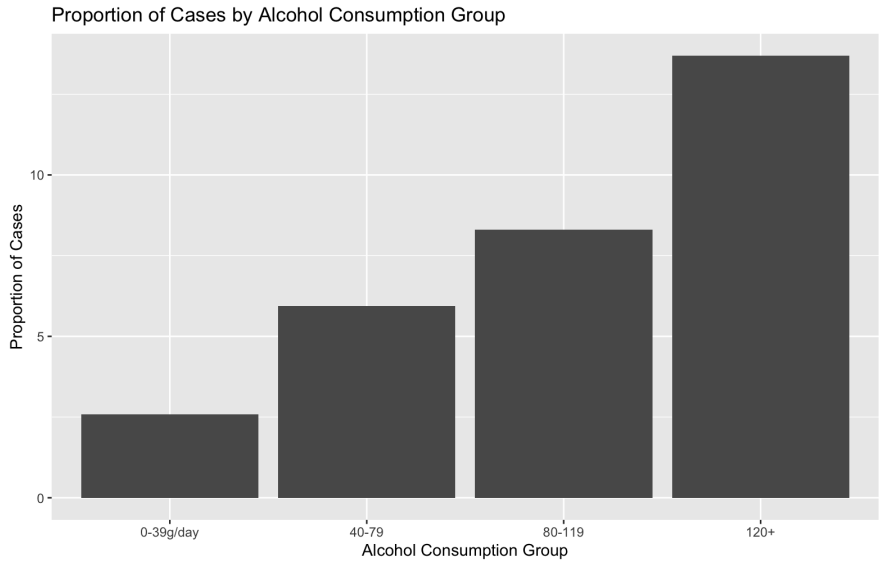


Figure A6

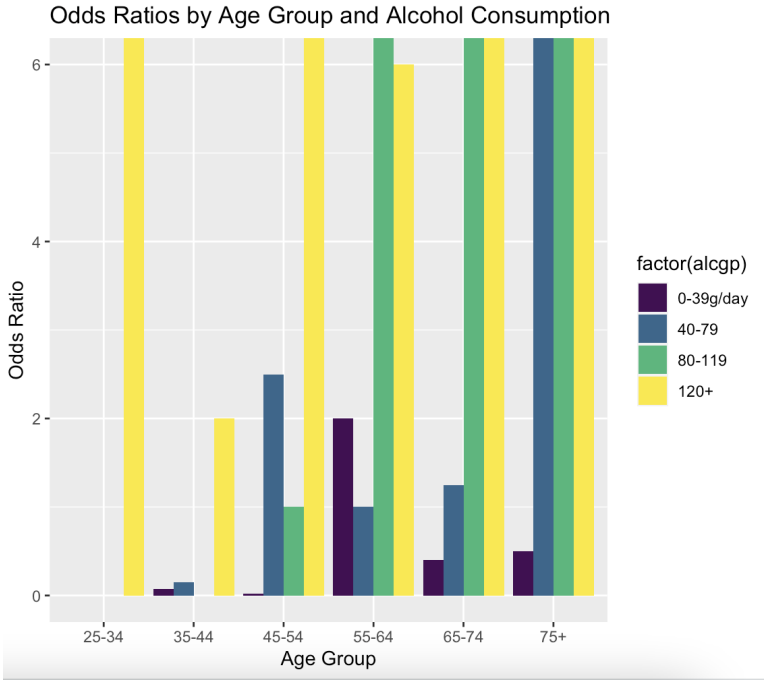


Figure A7

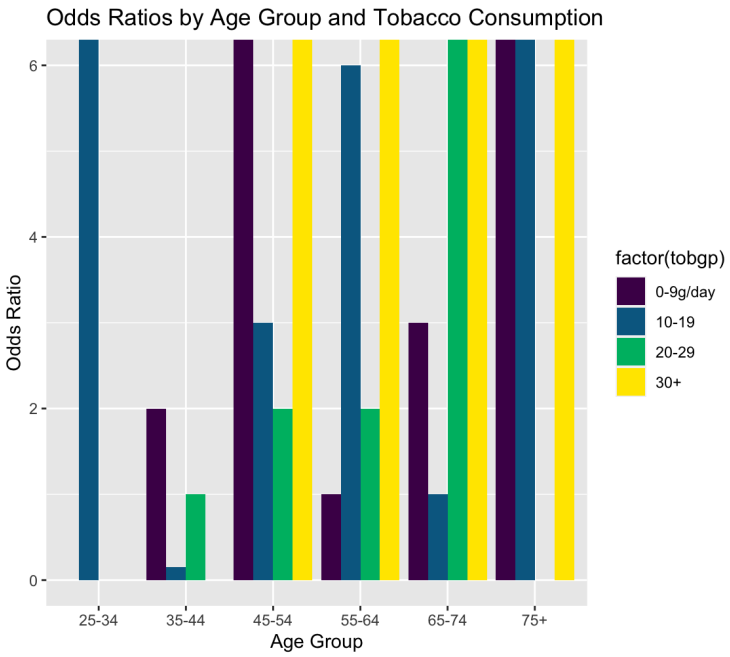


Figure A8

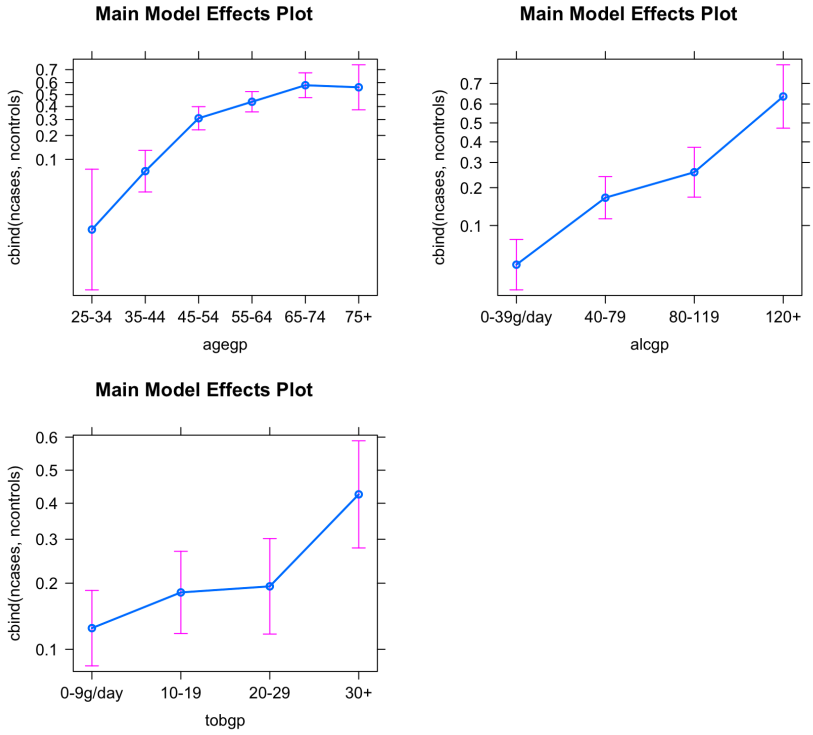


Figure A9

	GVIF	Df	GVIF^(1/(2*Df))
agegp	1.275990e+18	5	64.652409
alcgp	1.819080e+24	3	11048.635368
tobgp	6.364219e+23	3	9274.507499
agegp:alcgp	4.536790e+40	15	22.658176
agegp:tobgp	3.156048e+39	15	20.731853
alcgp:tobgp	2.002681e+01	9	1.181168

Bibliography

Breslow, N. E. and Day, N. E. (1980) *Statistical Methods in Cancer Research. Volume 1: The Analysis of Case-Control Studies.*

Cleveland Clinic. *Esophageal cancer: Symptoms, causes & treatment*. (2024, November 19). https://my.clevelandclinic.org/health/diseases/6137-esophageal-cancer

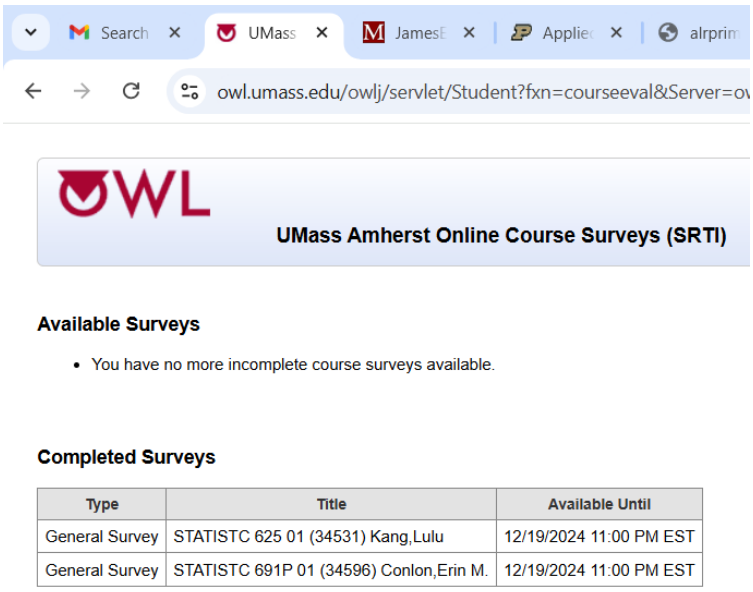
LoConte, Noelle K., et al. "Alcohol and cancer: A statement of the American Society of Clinical Oncology." *Journal of Clinical Oncology*, vol. 36, no. 1, 1 Jan. 2018, pp. 83–93, https://doi.org/10.1200/jco.2017.76.1155.

Plessen, G. (2022). P-152 Real-world data in patients with localized esophageal and gastro-esophageal junction cancer undergoing surgery in France: Results from the FREGAT database. *Annals of Oncology*, *33*(4).

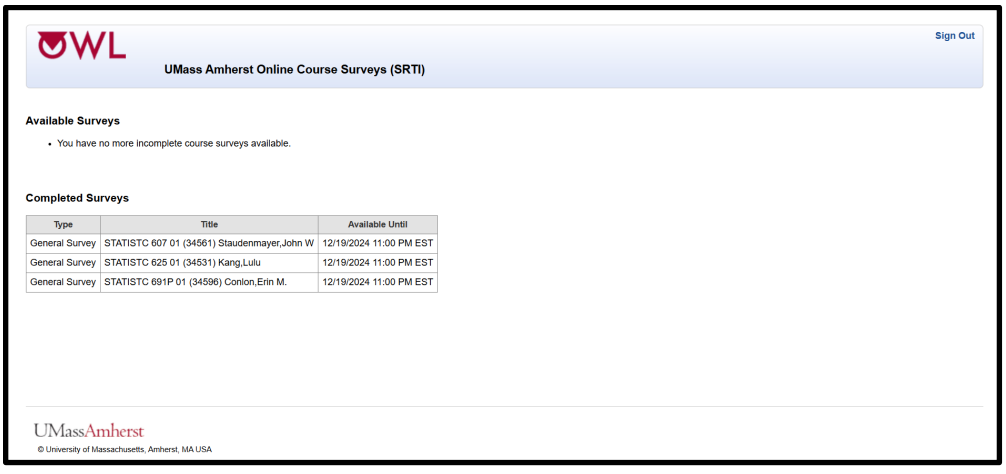
Wynder, E. L., Mushinski, M. H., Stellman, S. D., & Choay, P. (1981). Tobacco usage in France: An epidemiological study. *Preventive Medicine*, *10*(3), 301–315.

Individual Confirmation of Course Survey Completion:

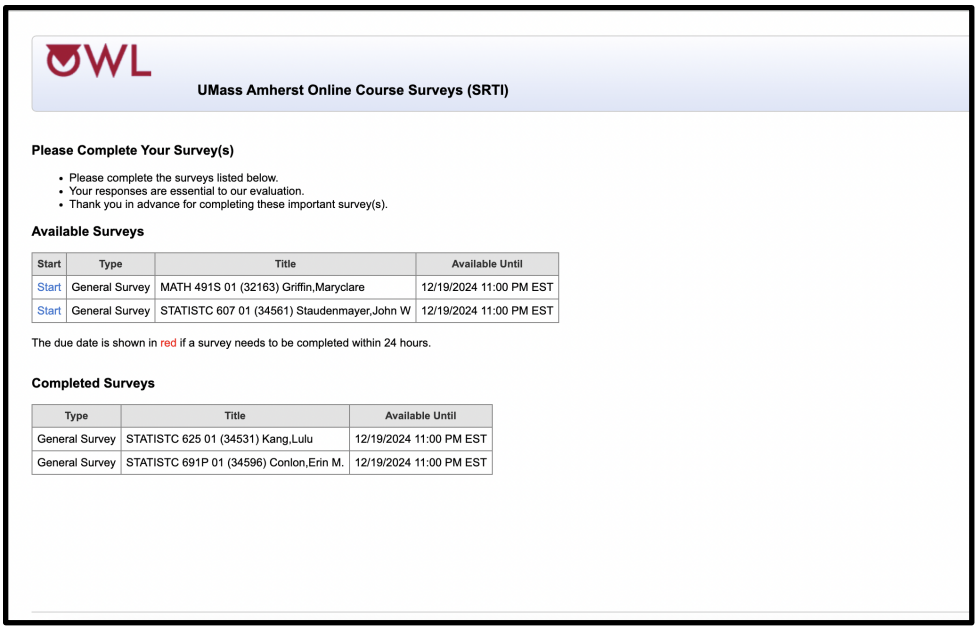
Anabel Costa:



Mia Josephy-Zack:



Mila Lewis-Peinado:



UMass Amherst Online Course Surveys (SRTI)

Please Complete Your Survey(s)

- Please complete the surveys listed below.
- Your responses are essential to our evaluation.
- Thank you in advance for completing these important survey(s).

Available Surveys

Start	Type	Title	Available Until
Start	General Survey	MATH 491S 01 (32163) Griffin,Maryclare	12/19/2024 11:00 PM EST
Start	General Survey	STATISTC 607 01 (34561) Staudenmayer,John W	12/19/2024 11:00 PM EST

The due date is shown in red if a survey needs to be completed within 24 hours.

Completed Surveys

Type	Title	Available Until
General Survey	STATISTC 625 01 (34531) Kang,Lulu	12/19/2024 11:00 PM EST
General Survey	STATISTC 691P 01 (34596) Conlon,Erin M.	12/19/2024 11:00 PM EST