Age and Esophageal Cancer

1. Calculate the OR and 95% CI for the association between esophageal cancer and age (agegrp).

Men aged 60 years and above were 2.7 times (95% CI 1.95-3.69) more likely to get esophageal cancer than those aged less than 60 years.

2. Consider whether age is linearly related to esophageal cancer.

a) Run a logistic regression model with a power (squared) term of ‘age’ (age*age) in addition to the independent variable (‘age’). Report the beta associated with age, OR, and 95% CI for the OR. Interpret the findings. Is the beta associated with ‘age*age’ significant? Is age linearly related to esophageal cancer?

Beta(age)= .4641
OR(age)=1.591, 95% CI: 1.375-1.839
Beta(age*age)= -.00359
OR(age*age)=.996, 95% CI: .995-.999

Variable ‘age*age’ is significant. Age is not linearly related to esophageal cancer.

To interpret the effect of age on esophageal cancer, you need to incorporate both the coefficient estimates associated with age and age*age. The negative significant square term (age*age) indicates that the odds of getting the cancer slow down as age increases. To see this, you can calculate the log odds of disease by substituting different age categories. For example,

<table>
<thead>
<tr>
<th>Age</th>
<th>.4641(age) + (-.00359)(age*age)</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>10.692</td>
</tr>
<tr>
<td>40</td>
<td>12.820</td>
</tr>
<tr>
<td>50</td>
<td>14.230</td>
</tr>
<tr>
<td>60</td>
<td>14.922</td>
</tr>
<tr>
<td>70</td>
<td>14.896</td>
</tr>
</tbody>
</table>

The odds of getting esophageal cancer tip down at 64.6 years of age (this is determined from taking the first derivative of the regression equation: Y=-15.5847+.4641(age)-.00359(age^2), setting it equal to zero and solving for age.)
This downturn of the cancer odds at 64.6 years of age is consistent with Figure 2 (see below).

One possible reason for this finding regarding age is the selective survival effect in which surviving older people are generally healthier (less healthy ones die earlier).

b) Set up a logistic regression to model the effects of indicator (or dummy) variables of 'age' using years 21 to 30 as the reference group.

\[
\begin{align*}
\text{age1} &= 1 \text{ if years 31 to 40, 0 otherwise} \\
\text{age2} &= 1 \text{ if years 41 to 50, 0 otherwise} \\
\text{age3} &= 1 \text{ if years 51 to 60, 0 otherwise} \\
\text{age4} &= 1 \text{ if years 61 to 70, 0 otherwise} \\
\text{age5} &= 1 \text{ if years \geq 71, 0 otherwise}
\end{align*}
\]

Plot a graph showing the relationship between the log odds of disease (betas) and the age categories (1, 2, 3, 4 & 5). Do you see a dose-response relationship between age (age1 age2 age3 age4 age5) and esophageal cancer?

See Figure 2. Age is not linearly related to esophageal cancer. Its relationship exhibits a convex shape.

![Figure 2. Relationships between the log odds of disease and age](image)

c) Based on findings in (a) & (b), which variable type for age (continuous, dichotomous or categorical) would you choose? Why?

Two pieces of evidence (significant square/power term and Figure 2) show that there is a non-linear relationship between age and esophageal cancer. A categorical (young, middle, older) age variable is more preferable than a binary age variable because it captures more information. Men, aged between years 21 and 60, will be categorized as the young group; men, aged between 61-70, will be categorized as the middle-aged group; men, aged 71 and above, will be categorized as the old group. However, for simplicity, a dichotomous age variable (21-60 years vs 61+ years) will be used.

Tobacco Consumption and Esophageal Cancer

1. Calculate the OR and 95% CI for the association between esophageal cancer and tobacco consumption (tobgrp).

Men who smoked 10 grams or more tobacco per day were 2.1 times (95% CI 1.54-2.92) more likely to get esophageal cancer than those smoking less than 10 grams per day.

2. Examine whether tobacco consumption is linearly related to esophageal cancer. Set up a logistic regression to model the effects of indicator variables of ‘tobacco’ using non-smokers as the reference. Plot a graph showing the relationship between the log odds of disease (betas) and smoking categories (1, 2, 3, 4, & 5). Do you see a dose-response relationship between cigarette consumption (tobamt1 tobamt2 tobamt3 tobamt4 tobamt5) and esophageal cancer?

\[
\begin{align*}
\text{tobamt1} &= 1 \text{ if 1-9 gms/day, 0 otherwise} \\
\text{tobamt2} &= 1 \text{ if 10-19 gms/day, 0 otherwise} \\
\text{tobamt3} &= 1 \text{ if 20-29 gms/day, 0 otherwise} \\
\text{tobamt4} &= 1 \text{ if 30-39 gms/day, 0 otherwise} \\
\text{tobamt5} &= 1 \text{ if 40+ gms/day, 0 otherwise}
\end{align*}
\]
See Figure 3. No dose-response relationship between smoking and esophageal cancer is observed.

![Figure 3. Relationships between the log odds of disease and tobacco categories](image)

3. Based on the above analyses, which variable type for tobacco (continuous, dichotomous, or categorical) would you choose? Why?

A categorical (3 levels: non-smokers, smoke 1-29 gms/day, smoke 30+ gms/day) or a dichotomous (0-29 vs 30+) variable of tobacco consumption may be the choice. The cutoff point for tobacco will be 30 grams/day.

Age, Tobacco, Alcohol and Esophageal Cancer

1. Having decided the variable types for alcohol, age and tobacco, we can now test the model using the preferred form of each variable. List the variables and their measurement types (continuous, dichotomous, or categorical) that you are going to put in a logistic regression model.

Alcohol consumption: use continuous measurement of alcohol: ‘alcohol’

Age: either use dichotomous measurement of age: ‘agegrp’ or 2 indicator variables (e.g., midage (61-70) and oldage (71+), using age less than 60 as the reference group)
Tobacco consumption: either use dichotomous measurement of tobacco: ‘tobgrp’ or 2 indicator variables (e.g., toblow (1-29 gms/day) and tohigh (30+ gms/day), using non-smokers as the reference group).

2. Perform a number of likelihood ratio tests to determine whether the additional variable(s) significantly improve(s) the model.

a) Run a logistic regression with alcohol consumption as the only explanatory variable for esophageal cancer. Report the likelihood ratio test (i.e., compare a model with your alcohol variable to a model with the intercept term only) and interpret.

A likelihood ratio test of 159.9 with df=1 is significant at a chi-square distribution of df=1. Alcohol is significantly associated with esophageal cancer.

b) Add age into the model. In addition to alcohol, is age significantly associated with esophageal cancer? Perform a likelihood ratio test and interpret the result.

(For dichotomous variables)
A likelihood ratio test of 36.7 (830.895-794.227) with df=1 is significant at a chi-square distribution of df=1. In addition to alcohol, ‘agegrp’ significantly improves the model.

(For indicator variables)
A likelihood ratio test of 37.7 (830.895-793.218) with df=2 is significant at a chi-square distribution of df=1. In addition to alcohol, ‘midage’ and ‘oldage’ significantly improve the model.

c) In addition to alcohol and age, add tobacco consumption into the model. Is tobacco consumption significantly associated with esophageal cancer? Perform a likelihood ratio test and interpret the result. Also report and interpret a goodness of fit test (use lackfit as the option).

(For dichotomous variables)
A likelihood ratio test of 18.08 (794.227-776.149) with df=1 is significant at a chi-square distribution of df=1. In addition to alcohol and ‘agegrp,’ ‘tobgrp’ significantly improves the model.

The chosen model shows a rather good fit between the data and model [goodness of fit test of 12.64 with df=8 (p-value=.125)]. It does not mean that it is the model. We are going to examine interaction terms later in the course.

(For indicator variables)
A likelihood ratio test of 36.92 (793.218-756.301) with df=2 is significant at a chi-square distribution of df=1. In addition to alcohol and age (‘midage’ and ‘oldage’), tobacco (‘toblow’ and ‘tobhigh’) significantly improve the model.

The chosen model shows a rather good fit between the data and model [goodness of fit test of 12.01 with df=8 (p-value=.15)]. It does not mean that it is the model. We are going to examine interaction terms later in the course.

Interaction

1. Run a logistic regression model to predict esophageal cancer with ‘alcgrp’ and ‘tobgrp’ as explanatory variables. Interpret the ORs for ‘alcgrp’ and ‘tobgrp.’ Cross-check these ORs with those in part 1 of homework 6. Which measure of effect in homework 6 corresponds to the ORs obtained via logistic regression?

This model assumes that the effect of alcohol is the same across different levels of tobacco consumption and vice versa. The OR of 5.6 indicates that the odds of disease is 5.6 times higher among men consuming alcohol >=40gm/day than those consuming <40gm/day, after controlling for the effect of tobacco. The OR for tobacco (1.8) can be interpreted similarly. These results are the same as the Mantel-Haenszel Summary ORs. Controlling for tobacco consumption (‘tobgrp’), the MHOR for alcohol consumption is 5.5. Controlling for alcohol consumption (‘alcgrp’), the MHOR for tobacco consumption is 1.8.

2. Run another logistic regression model by adding a cross-product term of ‘alcgrp’ and ‘tobgrp’ (alcgrp*tobgrp) to the above model.

   a) Perform a likelihood ratio test to examine whether the inclusion of the cross-product term of ‘alcgrp’ and ‘tobgrp’ significantly improves the predicting ability of the model? Does it confirm your findings in categorical data analysis?

The likelihood ratio test indicates that the addition of the interaction term significantly improves the predictive ability of the model. The difference between the ‘-2 LL’ values for the models with and without the interaction term is 4.78 (879.522 - 874.745). This value is statistically significant (p=0.03) at 1 degree of freedom from the Chi-square distribution. It confirms the findings in categorical data analysis.

   b) Interpret the ORs for ‘alcgrp,’ ‘tobgrp’ and the cross-product term.
The model with the cross product term estimates the effects of alcohol and tobacco consumption at each level of the other variable. The ORs for main effects arising from the model reflect the magnitude of effect of a given variable among those who are at the lowest category of the other variable. So, the OR of 9.8 for ‘alcgrp’ indicates that among men who consume tobacco <15gm/day, consumption of alcohol >=40gm/day increases the odds of disease by 9.8 times as compared to those consuming <40gm/day. Similarly, the OR of 4.0 for ‘tobgrp’ indicates that among men who consume alcohol <40gm/day, consumption of tobacco >=15gm/day increases the odds of disease by 4 times as compared to those consuming <15gm/day.

The estimate for the interaction term reflects the amount by which the effect of a given variable differs (increases or decreases) across different levels of exposure for the other variable. The $\beta$ for the interaction term may be interpreted as the difference in the two $\beta$s (logarithm of odds ratios) for a given variable across different levels of the other variable. The OR ($e^{\beta}$) for the interaction term may be interpreted as the ratio of the two ORs for a given variable across different levels of the other variable.

So, the OR of 0.38 ($\beta = -.98$) for the interaction term (between ‘alcgrp’ and ‘tobgrp’) indicates that the odds ratio for alcohol consumption is 0.38 times lower among the high tobacco consumers than among the low tobacco consumers.

c) Estimate and interpret the odds ratio for the following subjects:

Among non-smokers, heavy drinkers versus non-drinkers:

$\exp(2.2786) = 9.8$

Among non-smokers, men who drink increase the odds of getting esophageal cancer by 9.8 times than those who do not drink.

Among non-drinkers, heavy smokers versus non-smokers:

$\exp(1.3818) = 4.0$

Among non-drinkers, men who smoke increase the odds of getting esophageal cancer by 4.0 times than those who do not smoke.

Subjects who are heavy smokers and heavy drinkers versus those who are light smokers and light drinkers:

$\exp(2.28+1.38-.98) = 14.6$

Men who are exposed to both tobacco and alcohol increase the odds of getting esophageal cancer by 14.6 [$\exp(2.28+1.38-.98)$] times than those who are exposed to neither.

d) What is the effect of alcohol consumption on cancer (i.e., OR) among heavy smokers?

Among heavy smokers, men who drink increase the odds of getting esophageal cancer by 3.7 times [$\exp((2.28+1.38-.98)-(1.38))$] than those who do not drink.
e) What is the effect of tobacco consumption on cancer (i.e., OR) among heavy drinkers?

Among heavy drinkers, men who smoke increase the odds of getting esophageal cancer by 1.5 times \(\exp((2.28+1.38-.98)-(2.28))\) than those who do not smoke.

f) What does the Hosmer-Lemeshow goodness of fit test tell you about the fit of the model? (use lackfit as the option)

We do not reject the Hosmer-Lemeshow goodness of fit test which suggests that the model with the interaction term is a good overall fit.

### Hosmer and Lemeshow Goodness-of-Fit Test

<table>
<thead>
<tr>
<th>Chi-Square</th>
<th>DF</th>
<th>Pr &gt; ChiSq</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.0000</td>
<td>2</td>
<td>1.0000</td>
</tr>
</tbody>
</table>

3. Now set up a model to estimate associations for the following 4 groups (you will need to create 3 indicator variables):

- Light drinkers, light smokers
- Light drinkers, heavy smokers
- Heavy drinkers, light smokers
- Heavy drinkers, heavy smokers

For this question, define light drinkers as those with alcgrp = 0, heavy drinkers as those with alcgrp = 1, light smokers as those with tobgrp = 0, and heavy smokers as those with tobgrp = 1.

Compare the results from this model to the model with the two main effects (alcgrp and tobgrp) and the cross-product term. Compare in terms of the overall log likelihood and the beta coefficients, as well as your overall interpretation of these beta coefficients. What does each model say about the presence or absence of interaction?

The overall likelihood is the same (874.745) whether we parameterize the model using the indicator variables (3 betas) or by using the cross product term (2 main effect + 1 cross product)

Here the odds ratios and 95% CI for ORs are:

- Light drinkers, light smokers 1.0 (referent)
- Light drinkers, heavy smokers 3.9 (1.75-9.04) * compares with main effect above
Heavy drinkers, light smokers 9.7 (4.76-20.03) * compares with main effect above

Heavy drinkers, heavy smokers 14.7 (7.2-29.7) * This is equivalent to adding all three beta from the cross product model and exponentiating: \( \exp(2.279 + 1.38 - 0.975) = 14.7 \)

Note: we can use the betas from either model to assess interaction because we can rearrange the betas from one model to get the results from the other. Both models are assessing interaction on a multiplicative scale.