Laboratory 6  
Case-Control Analysis (1)

Three weeks ago, you cleaned the esophageal cancer case-control study data and saved the updated program. Make sure that you corrected the miscoded values for the status and sex variables and also the two abnormally high values for alcohol (that is, change 868 to 268 and 916 to 216).

**Alcohol Consumption and Esophageal Cancer**

Part 1: Crude and Stratified Analysis

1. Calculate the OR and 95% CI for the association between alcohol consumption (alcgrp) and esophageal cancer by using the following SAS program.

   ```sas
   proc freq data=case_control978;
   tables alcgrp*status / measures chisq;
   run;
   ```

2. Examine whether age fulfills the basic criteria for potential confounding variable in the data.

   a). Examine whether age is associated with alcohol consumption (alcgrp) among controls.

   ```sas
   proc freq data=case_control978;
   where status=0;
   tables alcgrp*agegrp / measures;
   run;
   ```

   b). Examine whether age is associated with the disease among unexposed (alcgrp=0).

   ```sas
   proc freq data=case_control978;
   where alcgrp=0;
   tables agegrp*status / measures;
   run;
   ```

   c). Does age mediate the association between alcohol consumption and esophageal cancer?

   d). Based on your evaluations from a), b), c), is age a confounder in the association between alcohol consumption and esophageal cancer among men?
3. Examine whether tobacco consumption fulfills the basic criteria for potential confounding variable in the data.

4. Calculate the ORs and 95% CIs for the association between alcohol consumption and esophageal cancer within level of age and tobacco variables.

```
proc freq data=case_control978;
tables agegrp*alcgrp*status / measures cmh; /* CMH produces Mantel Haenszel Summary OR */
tables tobgrp*alcgrp*status / measures cmh;
tables agegrp*tobgrp*alcgrp*status / measures cmh;
run;
```

**Stratified by agegrp**

<table>
<thead>
<tr>
<th></th>
<th>Among agegrp=0</th>
<th>Among agegrp=1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Cancer</td>
<td>Cancer</td>
</tr>
<tr>
<td>Low alcohol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High alcohol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CI</td>
<td></td>
<td></td>
</tr>
<tr>
<td>MH OR</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Stratified by tobgrp**

<table>
<thead>
<tr>
<th></th>
<th>Among tobgrp=0</th>
<th>Among tobgrp=1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>No Cancer</td>
<td>Cancer</td>
</tr>
<tr>
<td>Low alcohol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High alcohol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MH OR =
95% CI =
Homogeneity Test Statistic= , df= , P =

Stratified by agegrp and tobgrp jointly

<table>
<thead>
<tr>
<th>Agegrp &amp; Tobacco</th>
<th>No Cancer</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low alcohol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High alcohol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Agegrp &amp; Tobacco</th>
<th>No Cancer</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low alcohol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>High alcohol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>OR</td>
<td></td>
<td></td>
</tr>
<tr>
<td>95% CI</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5. Does the pattern of stratum-specific effect estimates for alcohol consumption support your preliminary assessment of age and tobacco as confounders evaluated in Questions 2 and 3?

6. Based on information in Questions 4 and 5, evaluate the crude, stratum-specific and summary ORs as well as the P values for the homogeneity tests. Write your answer indicating what would be the most pertinent OR (and 95% CI) for describing the effect of alcohol consumption on esophageal cancer. Provide reasons for your answer.
Part 2: Logistic Regression

1. Run a logistic regression model to predict esophageal cancer with ‘alcgrp’ as the independent variable. Report the intercept, beta (i.e., estimate), OR and 95% CI for OR. Interpret the results. Compare this result with the one you did in part 1. Are they the same? Perform a likelihood ratio test to examine whether the model includes ‘alcgrp’ improves the predicting ability of the model. Interpret.

```latex
\texttt{proc logistic data=case_control978 descending;}
\texttt{model status=} ;
\texttt{run;}
\texttt{proc logistic data=case_control978 descending;}
\texttt{model status=}alcgrp ;
\texttt{run;}
```

\begin{tabular}{|l|l|l|}
\hline
Model & -2 Log Likelihood & DF \\
\hline
1. intercept only & 990.864 & -- \\
2. intercept & alcgrp & \\
\hline
\end{tabular}

2. Conduct an adjusted data analysis using a logistic regression model to examine the effects of alcohol on esophageal cancer, controlling for the potential confounding effect of ‘agegrp.’ Compare this result with the one you did in part 1 and answer the following questions.

```latex
\texttt{proc logistic data=case_control978 descending;}
\texttt{model status=}alcgrp agegrp ;
\texttt{run;}
```

a) Report the beta, OR and 95% CIs for OR. Which measure of association in a categorical data analysis is equivalent to the adjusted OR obtained in a logistic regression modeling?

b) Using the 10% change-in-estimate rule, is age a confounder?

c) Compare the result in logistic regression with the one you did before. Do you come to the same conclusion regarding age as a confounder?

d) Is the beta associated with ‘agegrp’ significant? Interpret.

e) Perform a likelihood ratio test to examine whether including ‘agegrp’ in the model (as compared to model with alcohol only) has improved the predicting ability of the model. Interpret. (Note: compare this result with d, do you come to the same conclusion?)

\begin{tabular}{|l|l|l|}
\hline
Model = intercept + & -2 Log Likelihood & DF \\
\hline
1. alcgrp & 901.036 & 1 \\
2. alcgrp + agegrp & & \\
\hline
\end{tabular}
3. Conduct an adjusted data analysis using a logistic regression model to examine the effect of alcohol on esophageal cancer, controlling for the potential confounding effect of ‘tobgrp.’

```
proc logistic data=case_control1978 descending;
model status=alcgrp tobgrp;
run;
```

a) Report the beta, OR and 95% CIs for OR.

b) Using the 10% change-in-estimate rule, is tobacco a confounder?

c) Compare this result with the one we did before. Do you come to the same conclusion regarding tobacco as a confounder?

d) Is the beta associated with ‘tobgrp’ significant? Interpret.

e) Perform a likelihood ratio test to examine whether including ‘tobgrp’ in the model (as compared to model with alcohol only) has improved the predicting ability of the model. Explain. (Note: compare this result with d, do you come to the same conclusion?)

<table>
<thead>
<tr>
<th>Model</th>
<th>-2 Log Likelihood</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. alcgrp</td>
<td>901.036</td>
<td>1</td>
</tr>
<tr>
<td>2. alcgrp + tobgrp</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

4. Control for the joint confounding effects of age ‘agegrp’ and tobacco ‘tobgrp’ on the effect of alcohol consumption on esophageal cancer.

```
proc logistic data=case_control1978 descending;
model status=alcgrp agegrp tobgrp;
run;
```

a) Perform a likelihood ratio test to examine whether including ‘tobgrp’ in addition to ‘agegrp’ and ‘alcgrp’ has improved the predicting ability of the model. Explain.

<table>
<thead>
<tr>
<th>Model</th>
<th>-2 Log Likelihood</th>
<th>DF</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. alcgrp</td>
<td>901.036</td>
<td>1</td>
</tr>
<tr>
<td>2. alcgrp + agegrp</td>
<td>860.149</td>
<td>2</td>
</tr>
<tr>
<td>3. alcgrp + agegrp + tobgrp</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

b) Do ‘agegrp’ and ‘tobgrp’ jointly confound the association between alcohol consumption and esophageal cancer (based on 10% change in beta)?
5. Examine the linear association between alcohol consumption and esophageal cancer.

a) Run a logistic regression model with a power (squared) term of ‘alcohol’ (alcohol*alcohol) in addition to the independent variable (‘alcohol’). Is the beta associated with ‘alcohol*alcohol’ significant? Is alcohol consumption linearly related to esophageal cancer?

```plaintext
proc logistic data=case_control978 descending;
  model status=alcohol alcohol*alcohol;
run;
```

b) Set up a logistic regression to model the effects of indicator variables (i.e., dummy variables) of ‘alcohol’ using non-drinkers as the reference group.

- alc1 = 1 if 1-39 gms/day, 0 otherwise
- alc2 = 1 if 40-79 gms/day, 0 otherwise
- alc3 = 1 if 80-119 gms/day, 0 otherwise
- alc4 = 1 if 120-159 gms/day, 0 otherwise
- alc5 >= 1 if 160 gms/day, 0 otherwise

```plaintext
data case_control978;
set case_control978;
  alcohol1=0;
  if (1<=alcohol<=39) then alcohol1=1;
  alcohol2=0;
  if (40<=alcohol<=79) then alcohol2=1;
  alcohol3=0;
  if (80<=alcohol<=119) then alcohol3=1;
  alcohol4=0;
  if (120<=alcohol<=159) then alcohol4=1;
  alcohol5=0;
  if (alcohol>=160) then alcohol5=1;
proc logistic data=case_control978 descending;
  model status=alcohol1 alcohol2 alcohol3 alcohol4 alcohol5;
run;
```

Plot a graph showing the relationship between the log odds of disease (i.e., betas) and the alcohol categories (1, 2, 3, 4, & 5). Do you see a dose-response relationship between alcohol consumption and esophageal cancer?

```plaintext
data dose_response;
input beta alcohol;
cards;
;
run;
proc gplot data=dose_response;
plot beta*alcohol; symbol line=1 width=5 interpol=join;
title 'Fig 1. Relationship between the log odds of esophageal cancer and alcohol consumption';
run;
```
c) Based on the above analyses, which variable type for alcohol (continuous, dichotomous or categorical) would you choose? Why?

6. Examine the effect of the continuous measurement of alcohol consumption (i.e., ‘alcohol’) on esophageal cancer via a logistic regression.

a) Report the estimate ($\beta$), OR and 95% CI for the crude association between ‘alcohol’ and esophageal cancer? Interpret the results.

```proc logistic data=case_control978 descending;
model status=alcohol;
run;
```

b) What is the crude OR for esophageal cancer for a 40-grams increase in alcohol consumption per day?

```proc logistic data=case_control978 descending;
model status=alcohol/waldrl;
units alcohol=40;
run;
```

or you can calculate by hand

c) Compare the odds of getting esophageal cancer for a man who drinks 120 grams per day with a man who is similar in other respects but drinks 80 grams per day.

d) Compare the odds of getting esophageal cancer for a man who drinks 160 grams per day with a man who is similar in other respects but drinks 40 grams per day.

e) Examine results in (b), (c), & (d). What is the assumption in the interpretation of the log odds if we choose to keep alcohol as a continuous variable?