Introduction to Data Management and Analysis Using SAS

EPI 223 Final Project (70 points)
(Please check Canvas for assignment due date)
James Guo

1.

Q1. Using the data provided, complete Table 1.

TABLE 1. Demographic characteristics of subarachnoid (SAH) cases and controls, frequency matched on age, gender, King County, Washington, 1987-1989.

Characteristic	SAH Cases (n= 149)		Controls (n= 298)		p-value*
	No.	%	No.	%	
Age group (years)					
18 – 44	41	27.5	79	26.5	
45 – 54	29	19.5	61	20.5	
55 – 64	32	21.5	66	22.2	0.996
65 – 74	25	16.8	47	15.8	
ε 75	22	14.8	45	15.1	
Age (years, mean ± SD)	56.3 ± 16.7		56.2 ± 16.5		0.955
2. Sex (n,%)					
Male	46	30.9	92	30.9	1.000
Female	103	69.1	206	69.1	
3. Race (n, %)					
White	129	86.6	274	92.0	
Black	11	7.4	13	4.4	0.199
Other	9	6.0	11	3.7	
Education (years, mean ± SD)	12.8±2.6		13.5±2.6		0.011

*p-value based on Chi-square test for categorical variables and T-Test for continuous variables

Q2. Using the data provided, complete Table 2.

TABLE 2. Association between cigarette smoking and risk of subarachnoid (SAH), King County, Washington, 1987-1989.

Cigarette smoking	SAH cases		Controls		p-value	Crude Odds Ratio (95% CI)	Adjusted Odd ratio* (95% CI)
	n	%	n	%			
Never Smoker (ref. group)	40	26.9	154	52.0		1.0 (ref)	1.0 (ref)
Ever Smoker	109	73.2	142	48.0	<0.0001	2.96 [1.93, 4.53]	3.17 [2.04, 4.94]
Former	34	22.8	85	28.7		1.54 [0.91, 2.61]	1.58 [0.92, 2.73]
Current	75	50.3	57	19.3		5.07 [3.11, 8.27]	5.73 [3.43, 9.60]
Pack-yrs of smoking (tertiles)							
0 pk-yrs	40	27.0	154	55.2		1.0 (ref)	1.0 (ref)
0.025 - 16 pk-yrs	28	18.9	62	22.2	<0.0001	1.74 [0.99, 3.06]	1.91 [1.07, 3.42]
16.5 - 185.5 pk-yrs	80	54.1	63	22.6		4.89 [3.03, 7.90]	5.66 [3.40, 9.40]

* odds ratio adjusted for age and gender

INSERT SAS LOG HERE USED FOR GENERATING RESULTS FOR TABLES 1 & 2 (in your log please include all data steps you used for cleaning the data and generating new variables and include comments so we can follow your steps)

```
 1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
 68
 69
/*************************************************************************
*********
 70         * Program: finalproject_code.sas
 71         * Author: Zhongyi (James) Guo
 72         * Date:   DEC. 10, 2023
 73         * Purpose: Code to answer questions on the final project of
EPI 223.
 74
*************************************************************************
********/
 75
 76         /* Data Pre-processing */
 77         /* include non-empty columns only */
 78         data epi223.sahdemographics;
 79         set epi223.sahdemographics;
 80         keep STUDYID2 CASECON AGE GENDER RACER EDUCATE RMARITAL;
 81         run;
```

```
 NOTE: There were 447 observations read from the data set
EPI223.SAHDEMOGRAPHICS.
 NOTE: The data set EPI223.SAHDEMOGRAPHICS has 447 observations and 7
variables.
 NOTE: DATA statement used (Total process time):
       real time           0.01 seconds
       user cpu time       0.00 seconds
       system cpu time     0.01 seconds
       memory              947.50k
       OS Memory           25256.00k
       Timestamp           12/08/2023 06:49:50 AM
       Step Count                        291  Switch Count  2
       Page Faults                       0
       Page Reclaims                     121
       Page Swaps                        0
       Voluntary Context Switches        42
       Involuntary Context Switches      0
       Block Input Operations            0
       Block Output Operations           264


 82
 83         /* sort two datasets by `STUDYID2` for the purpose of merging
*/
 84         proc sort data=epi223.sahdemographics;
 85         by STUDYID2;
 86         run;

 NOTE: There were 447 observations read from the data set
EPI223.SAHDEMOGRAPHICS.
 NOTE: The data set EPI223.SAHDEMOGRAPHICS has 447 observations and 7
variables.
 NOTE: PROCEDURE SORT used (Total process time):
       real time           0.01 seconds
       user cpu time       0.00 seconds
       system cpu time     0.00 seconds
       memory              929.28k
       OS Memory           25256.00k
       Timestamp           12/08/2023 06:49:50 AM
       Step Count                        292  Switch Count  2
       Page Faults                       0
       Page Reclaims                     116
       Page Swaps                        0
       Voluntary Context Switches        51
       Involuntary Context Switches      0
       Block Input Operations            288
```

```
      Block Output Operations           264


87
88          proc sort data=epi223.sahriskfactor;
89          by STUDYID2;
90          run;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              649.12k
      OS Memory           24996.00k
      Timestamp           12/08/2023 06:49:50 AM
      Step Count                        293  Switch Count  0
      Page Faults                       0
      Page Reclaims                     51
      Page Swaps                        0
      Voluntary Context Switches        6
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


91
92          /* merge two datasets by `STUDYID2` */
93          data combined;
94          merge epi223.sahdemographics epi223.sahriskfactor;
95          by STUDYID2;
96          run;

NOTE: There were 447 observations read from the data set
EPI223.SAHDEMOGRAPHICS.
NOTE: There were 450 observations read from the data set
EPI223.SAHRISKFACTOR.
NOTE: The data set WORK.COMBINED has 450 observations and 13 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1395.43k
      OS Memory           25516.00k
      Timestamp           12/08/2023 06:49:50 AM
      Step Count                        294  Switch Count  2
      Page Faults                       0
```

```
      Page Reclaims                     156
      Page Swaps                        0
      Voluntary Context Switches        26
      Involuntary Context Switches      0
      Block Input Operations            288
      Block Output Operations           264


 97
 98         /* detect duplicate IDs */
 99         proc sort data=combined nodupkey dupout=dup_IDs;
 100        by STUDYID2;
 101        run;

 NOTE: There were 450 observations read from the data set WORK.COMBINED.
 NOTE: 3 observations with duplicate key values were deleted.
 NOTE: The data set WORK.DUP_IDS has 3 observations and 13 variables.
 NOTE: The data set WORK.COMBINED has 447 observations and 13 variables.
 NOTE: PROCEDURE SORT used (Total process time):
       real time           0.00 seconds
       user cpu time       0.00 seconds
       system cpu time     0.01 seconds
       memory              1335.12k
       OS Memory           25516.00k
       Timestamp           12/08/2023 06:49:50 AM
       Step Count                        295  Switch Count  4
       Page Faults                       0
       Page Reclaims                     174
       Page Swaps                        0
       Voluntary Context Switches        24
       Involuntary Context Switches      0
       Block Input Operations            0
       Block Output Operations           528


 102
 103        /* The duplicate IDs are 11441, 50041, and 61412. */
 104        /* Each duplicate ID has two identical rows. One row was
removed for each ID. */
 105
 106        /* Confirm if duplicate rows were removed */
 107        /* proc print data=combined; */
 108        /* run; */
 109
 110
 111        /* EDUCATE, RMARITAL, BODYMASS, CIGSTAT, PACKYRS, ALCEVER,
HEAVDRIN, and FAMSAH
```

```
 112        all have missing values recorded as -1 */
 113        /* replace -1 with . */
 114        /* create a binary variable EVERSMOKER (1 = former and current
smokers, 0 = never smoked) */
 115        data combined;
 116        set combined;
 117        array vars {8} EDUCATE RMARITAL BODYMASS CIGSTAT PACKYRS
ALCEVER HEAVDRIN FAMSAH;
 118            do index=1 to 8;
 119            if vars{index} = -1 then vars{index} = .;
 120            end;
 121        /* create EVERSMOKER */
 122            if cigstat in (2,3) then EVERSMOKER = 1;
 123            else if cigstat=. then EVERSMOKER = .;
 124            else EVERSMOKER = 0;
 125            drop index;
 126        run;

 NOTE: There were 447 observations read from the data set WORK.COMBINED.
 NOTE: The data set WORK.COMBINED has 447 observations and 14 variables.
 NOTE: DATA statement used (Total process time):
       real time           0.00 seconds
       user cpu time       0.01 seconds
       system cpu time     0.00 seconds
       memory              1072.46k
       OS Memory           25256.00k
       Timestamp           12/08/2023 06:49:50 AM
       Step Count                        296  Switch Count  2
       Page Faults                       0
       Page Reclaims                     120
       Page Swaps                        0
       Voluntary Context Switches        10
       Involuntary Context Switches      0
       Block Input Operations            0
       Block Output Operations           264


 127
 128        /* check the new variable `EVERSMOKER` vs `cigstat` */
 129        proc freq data=combined;
 130        tables EVERSMOKER*cigstat/list missing;
 131        run;

 NOTE: There were 447 observations read from the data set WORK.COMBINED.
 NOTE: PROCEDURE FREQ used (Total process time):
       real time           0.01 seconds
       user cpu time       0.01 seconds
```

```
      system cpu time     0.00 seconds
      memory              2001.34k
      OS Memory           26036.00k
      Timestamp           12/08/2023 06:49:50 AM
      Step Count                        297  Switch Count  8
      Page Faults                       0
      Page Reclaims                     321
      Page Swaps                        0
      Voluntary Context Switches        40
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           1064


132         /* successfully created */
133
134
135         /* categorical and continuous variable separation */
136         data cat_var;
137         set combined;
138         drop STUDYID2 AGE EDUCATE BODYMASS PACKYRS;
139         run;

NOTE: There were 447 observations read from the data set WORK.COMBINED.
NOTE: The data set WORK.CAT_VAR has 447 observations and 9 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              1067.40k
      OS Memory           25256.00k
      Timestamp           12/08/2023 06:49:50 AM
      Step Count                        298  Switch Count  2
      Page Faults                       0
      Page Reclaims                     118
      Page Swaps                        0
      Voluntary Context Switches        14
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


140
141         data cont_var;
142         set combined;
143         keep AGE EDUCATE BODYMASS PACKYRS;
144         run;
```

```
NOTE: There were 447 observations read from the data set WORK.COMBINED.
NOTE: The data set WORK.CONT_VAR has 447 observations and 4 variables.
NOTE: DATA statement used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              954.65k
      OS Memory           25256.00k
      Timestamp           12/08/2023 06:49:50 AM
      Step Count                        299  Switch Count  2
      Page Faults                       0
      Page Reclaims                     125
      Page Swaps                        0
      Voluntary Context Switches        10
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


145
146        /* confirm missing values were expressed as . for all
categorical variables */
147        proc freq data=cat_var;
148        tables _all_ /list missing;
149        run;

NOTE: There were 447 observations read from the data set WORK.CAT_VAR.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.05 seconds
      user cpu time       0.06 seconds
      system cpu time     0.00 seconds
      memory              1204.09k
      OS Memory           25256.00k
      Timestamp           12/08/2023 06:49:50 AM
      Step Count                        300  Switch Count  2
      Page Faults                       0
      Page Reclaims                     142
      Page Swaps                        0
      Voluntary Context Switches        12
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           280


150        /* confirmation completed */
151
```

```
 152        /* confirm missing values were expressed as . for all
continuous variables */
 153        proc univariate data=cont_var;
 154        var _all_;
 155        run;

 NOTE: PROCEDURE UNIVARIATE used (Total process time):
       real time           0.12 seconds
       user cpu time       0.13 seconds
       system cpu time     0.00 seconds
       memory              1075.43k
       OS Memory           24996.00k
       Timestamp           12/08/2023 06:49:50 AM
       Step Count                        301  Switch Count  0
       Page Faults                       0
       Page Reclaims                     61
       Page Swaps                        0
       Voluntary Context Switches        0
       Involuntary Context Switches      0
       Block Input Operations            0
       Block Output Operations           64


 156        /* confirmation completed */
 157
 158        /* Q1&2 */
 159        /* Table Establishment */
 160
 161        /* format age group according to Table 1 */
 162        proc format;
 163
 163      !  value agegroup
 164             18 - 45 = '18 – 44'
 165                 45 - 55 = '45 – 54'
 166                 55 - 65 = '55 – 64'
 167                 65 - 75 = '65 – 74'
 168                 75 - high = '≥ 75';
 NOTE: Format AGEGROUP is already on the library WORK.FORMATS.
 NOTE: Format AGEGROUP has been output.
 169             value gendergroup
 170             1 = 'Male'
 171             2 = 'Female';
 NOTE: Format GENDERGROUP is already on the library WORK.FORMATS.
 NOTE: Format GENDERGROUP has been output.
 172             value casecongroup
 173             1 = 'Case'
 174             0 = 'Control';
```

```
NOTE: Format CASECONGROUP is already on the library WORK.FORMATS.
NOTE: Format CASECONGROUP has been output.
175            value racergroup
176            1 = 'White'
177            2 = 'Black'
178            3 = 'Other';
NOTE: Format RACERGROUP is already on the library WORK.FORMATS.
NOTE: Format RACERGROUP has been output.
179            value cigstatgroup
180            1 = 'Never smoked'
181            2 = 'Former smoker'
182            3 = 'Current smoker';
NOTE: Format CIGSTATGROUP is already on the library WORK.FORMATS.
NOTE: Format CIGSTATGROUP has been output.
183            value alcevergroup
184            1 = 'No'
185            2 = 'Yes';
NOTE: Format ALCEVERGROUP is already on the library WORK.FORMATS.
NOTE: Format ALCEVERGROUP has been output.
186         /* create a format for `packyrstert` to be created */
187            value packyrstertgroup
188            0 = '0 pk-yrs'
189            1 = '0.025 - 16 pk-yrs'
190            2 = '16.5 - 185.5 pk-yrs';
NOTE: Format PACKYRSTERTGROUP is already on the library WORK.FORMATS.
NOTE: Format PACKYRSTERTGROUP has been output.
191            value eversmokergroup
192            0 = 'Never smoked'
193            1 = 'Ever smoked';
NOTE: Format EVERSMOKERGROUP is already on the library WORK.FORMATS.
NOTE: Format EVERSMOKERGROUP has been output.
194         run;

NOTE: PROCEDURE FORMAT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              248.90k
      OS Memory           24736.00k
      Timestamp           12/08/2023 06:49:50 AM
      Step Count                        302  Switch Count  0
      Page Faults                       0
      Page Reclaims                     14
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
```

```
      Block Output Operations           40


195
196        /* sort `combined` by CASECON to display 'Case' first */
197        proc sort data=combined;
198        by descending casecon age;
199        run;

NOTE: There were 447 observations read from the data set WORK.COMBINED.
NOTE: The data set WORK.COMBINED has 447 observations and 14 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              926.34k
      OS Memory           25256.00k
      Timestamp           12/08/2023 06:49:50 AM
      Step Count                        303  Switch Count  2
      Page Faults                       0
      Page Reclaims                     104
      Page Swaps                        0
      Voluntary Context Switches        12
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           264


200
201        /* no missing values for `age` */
202        /* create the 2x2 table for age and casecon & assign formats
created to the corresponding variables*/
203        proc freq data=combined order=data;
204            tables age*casecon / chisq nocum norow nopercent;
205            format casecon casecongroup. age agegroup.;
206        run;

NOTE: There were 447 observations read from the data set WORK.COMBINED.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              1316.21k
      OS Memory           25516.00k
      Timestamp           12/08/2023 06:49:50 AM
      Step Count                        304  Switch Count  4
      Page Faults                       0
```

```
      Page Reclaims                     192
      Page Swaps                        0
      Voluntary Context Switches        26
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           544


 207
 208        /* obtain mean and SD for each group in `casecon` by `age` and
p-value using t test */
 209        proc ttest data=combined order=data;
 210            class casecon;
 211            var age;
 212            format casecon casecongroup.;
 213        run;

 NOTE: PROCEDURE TTEST used (Total process time):
       real time           0.29 seconds
       user cpu time       0.14 seconds
       system cpu time     0.06 seconds
       memory              16766.90k
       OS Memory           39120.00k
       Timestamp           12/08/2023 06:49:51 AM
       Step Count                        305  Switch Count  48
       Page Faults                       0
       Page Reclaims                     27741
       Page Swaps                        0
       Voluntary Context Switches        1034
       Involuntary Context Switches      0
       Block Input Operations            0
       Block Output Operations           1808


 214
 215        /* create the 2x2 table for gender and casecon & assign
formats created to the corresponding variables*/
 216        proc freq data=combined order=data;
 217            tables gender*casecon / chisq nocum norow nopercent;
 218            format gender gendergroup. casecon casecongroup.;
 219        run;

 NOTE: There were 447 observations read from the data set WORK.COMBINED.
 NOTE: PROCEDURE FREQ used (Total process time):
       real time           0.02 seconds
       user cpu time       0.03 seconds
       system cpu time     0.00 seconds
```

```
      memory                  1159.37k
      OS Memory               33716.00k
      Timestamp               12/08/2023 06:49:51 AM
      Step Count                        306  Switch Count  4
      Page Faults                       0
      Page Reclaims                     195
      Page Swaps                        0
      Voluntary Context Switches        24
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           528


220
221        /* sort the data by `racer` */
222        proc sort data=combined;
223        by racer;
224        run;

NOTE: There were 447 observations read from the data set WORK.COMBINED.
NOTE: The data set WORK.COMBINED has 447 observations and 14 variables.
NOTE: PROCEDURE SORT used (Total process time):
      real time               0.00 seconds
      user cpu time           0.00 seconds
      system cpu time         0.00 seconds
      memory                  1040.62k
      OS Memory               33456.00k
      Timestamp               12/08/2023 06:49:51 AM
      Step Count                        307  Switch Count  2
      Page Faults                       0
      Page Reclaims                     104
      Page Swaps                        0
      Voluntary Context Switches        13
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           272


225
226        /* create the 2x2 table for racer and casecon & assign formats
created to the corresponding variables*/
227        proc freq data=combined order=data;
228            tables racer*casecon / chisq nocum norow nopercent;
229            format racer racergroup. casecon casecongroup.;
230        run;

NOTE: There were 447 observations read from the data set WORK.COMBINED.
```

```
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              1014.53k
      OS Memory           33716.00k
      Timestamp           12/08/2023 06:49:51 AM
      Step Count                        308  Switch Count  4
      Page Faults                       0
      Page Reclaims                     191
      Page Swaps                        0
      Voluntary Context Switches        25
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           544


231
232        /* obtain mean and SD for each group in `casecon` by `educate`
and p-value using t test */
233        proc ttest data=combined order=data;
234        where educate^=.;
235        class casecon;
236        var educate;
237        format casecon casecongroup.;
238        run;

NOTE: PROCEDURE TTEST used (Total process time):
      real time           0.28 seconds
      user cpu time       0.13 seconds
      system cpu time     0.05 seconds
      memory              10090.28k
      OS Memory           39636.00k
      Timestamp           12/08/2023 06:49:51 AM
      Step Count                        309  Switch Count  56
      Page Faults                       0
      Page Reclaims                     25751
      Page Swaps                        0
      Voluntary Context Switches        1177
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           1264


239
240
241        /* Q2 */
```

```
 242        /* sort the data first to display `cigstat` in an assending
order */
 243        proc sort data=combined;
 244        by descending casecon cigstat;
 245        run;

 NOTE: There were 447 observations read from the data set WORK.COMBINED.
 NOTE: The data set WORK.COMBINED has 447 observations and 14 variables.
 NOTE: PROCEDURE SORT used (Total process time):
       real time           0.00 seconds
       user cpu time       0.00 seconds
       system cpu time     0.01 seconds
       memory              1039.62k
       OS Memory           33456.00k
       Timestamp           12/08/2023 06:49:51 AM
       Step Count                        310  Switch Count  2
       Page Faults                       0
       Page Reclaims                     104
       Page Swaps                        0
       Voluntary Context Switches        17
       Involuntary Context Switches      0
       Block Input Operations            0
       Block Output Operations           264


 246
 247        /* create a 2x2 table between `eversmoker` and `casecon` */
 248        proc freq data=combined order=data;
 249        where eversmoker^=.;
 250        tables eversmoker*casecon/ chisq nocum norow nopercent;
 251        format eversmoker eversmokergroup. casecon casecongroup.;
 252        run;

 NOTE: There were 445 observations read from the data set WORK.COMBINED.
       WHERE eversmoker not = .;
 NOTE: PROCEDURE FREQ used (Total process time):
       real time           0.02 seconds
       user cpu time       0.03 seconds
       system cpu time     0.00 seconds
       memory              1151.78k
       OS Memory           33716.00k
       Timestamp           12/08/2023 06:49:51 AM
       Step Count                        311  Switch Count  7
       Page Faults                       0
       Page Reclaims                     191
       Page Swaps                        0
       Voluntary Context Switches        31
```

```
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           528


253
254
255         /* create a 3x2 table between `cigstat` and `casecon` */
256         proc freq data=combined order=data;
257         where cigstat^=.;
258         tables cigstat*casecon/ chisq nocum norow nopercent;
259         format cigstat cigstatgroup. casecon casecongroup.;
260         run;

NOTE: There were 445 observations read from the data set WORK.COMBINED.
      WHERE cigstat not = .;
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      user cpu time       0.02 seconds
      system cpu time     0.00 seconds
      memory              1127.75k
      OS Memory           33716.00k
      Timestamp           12/08/2023 06:49:51 AM
      Step Count                        312  Switch Count  7
      Page Faults                       0
      Page Reclaims                     190
      Page Swaps                        0
      Voluntary Context Switches        37
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           544


261
262         /* use logistic regression to calculate odds ratio between
`eversmoker` and `casecon`*/
263         proc logistic data=combined;
264         where eversmoker^=.;
265         format eversmoker eversmokergroup. casecon casecongroup.;
266         class eversmoker(ref='Never smoked');
267         model casecon(event='Case') = eversmoker;
268         oddsratio eversmoker;
269         run;

NOTE: PROC LOGISTIC is modeling the probability that CASECON='Case'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 445 observations read from the data set WORK.COMBINED.
```

```
      WHERE eversmoker not = .;
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.17 seconds
      user cpu time       0.11 seconds
      system cpu time     0.01 seconds
      memory              5314.50k
      OS Memory           35400.00k
      Timestamp           12/08/2023 06:49:51 AM
      Step Count                        313  Switch Count  13
      Page Faults                       0
      Page Reclaims                     701
      Page Swaps                        0
      Voluntary Context Switches        369
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           504


270
271
272        /* use logistic regression to calculate odds ratio between
`cigstat` and `casecon`*/
273        proc logistic data=combined;
274        where cigstat^=.;
275        format cigstat cigstatgroup. casecon casecongroup.;
276        class cigstat(ref='Never smoked');
277        model casecon(event='Case') = cigstat;
278        oddsratio cigstat;
279        run;

NOTE: PROC LOGISTIC is modeling the probability that CASECON='Case'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 445 observations read from the data set WORK.COMBINED.
      WHERE cigstat not = .;
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.17 seconds
      user cpu time       0.11 seconds
      system cpu time     0.01 seconds
      memory              5087.78k
      OS Memory           35400.00k
      Timestamp           12/08/2023 06:49:51 AM
      Step Count                        314  Switch Count  13
      Page Faults                       0
      Page Reclaims                     699
      Page Swaps                        0
      Voluntary Context Switches        387
      Involuntary Context Switches      0
```

```
       Block Input Operations            0
       Block Output Operations           504


 280
 281       /* use logistic regression to calculate odds ratio for
`eversmoker`, adjusted for age and gender */
 282       proc logistic data=combined;
 283       where eversmoker^=.;
 284       format eversmoker eversmokergroup. casecon casecongroup.
gender gendergroup.;
 285       class eversmoker(ref='Never smoked');
 286       model casecon(event='Case') = eversmoker age gender;
 287       oddsratio eversmoker;
 288       run;

 NOTE: PROC LOGISTIC is modeling the probability that CASECON='Case'.
 NOTE: Convergence criterion (GCONV=1E-8) satisfied.
 NOTE: There were 445 observations read from the data set WORK.COMBINED.
       WHERE eversmoker not = .;
 NOTE: PROCEDURE LOGISTIC used (Total process time):
       real time           0.18 seconds
       user cpu time       0.11 seconds
       system cpu time     0.00 seconds
       memory              5095.37k
       OS Memory           35400.00k
       Timestamp           12/08/2023 06:49:52 AM
       Step Count                        315  Switch Count  13
       Page Faults                       0
       Page Reclaims                     689
       Page Swaps                        0
       Voluntary Context Switches        380
       Involuntary Context Switches      1
       Block Input Operations            0
       Block Output Operations           496


 289
 290       /* use logistic regression to calculate odds ratio for
`cigstat`, adjusted for age and gender */
 291       proc logistic data=combined;
 292       where cigstat^=.;
 293       format cigstat cigstatgroup. casecon casecongroup. gender
gendergroup.;
 294       class cigstat(ref='Never smoked');
 295       model casecon(event='Case') = cigstat age gender;
 296       oddsratio cigstat;
```

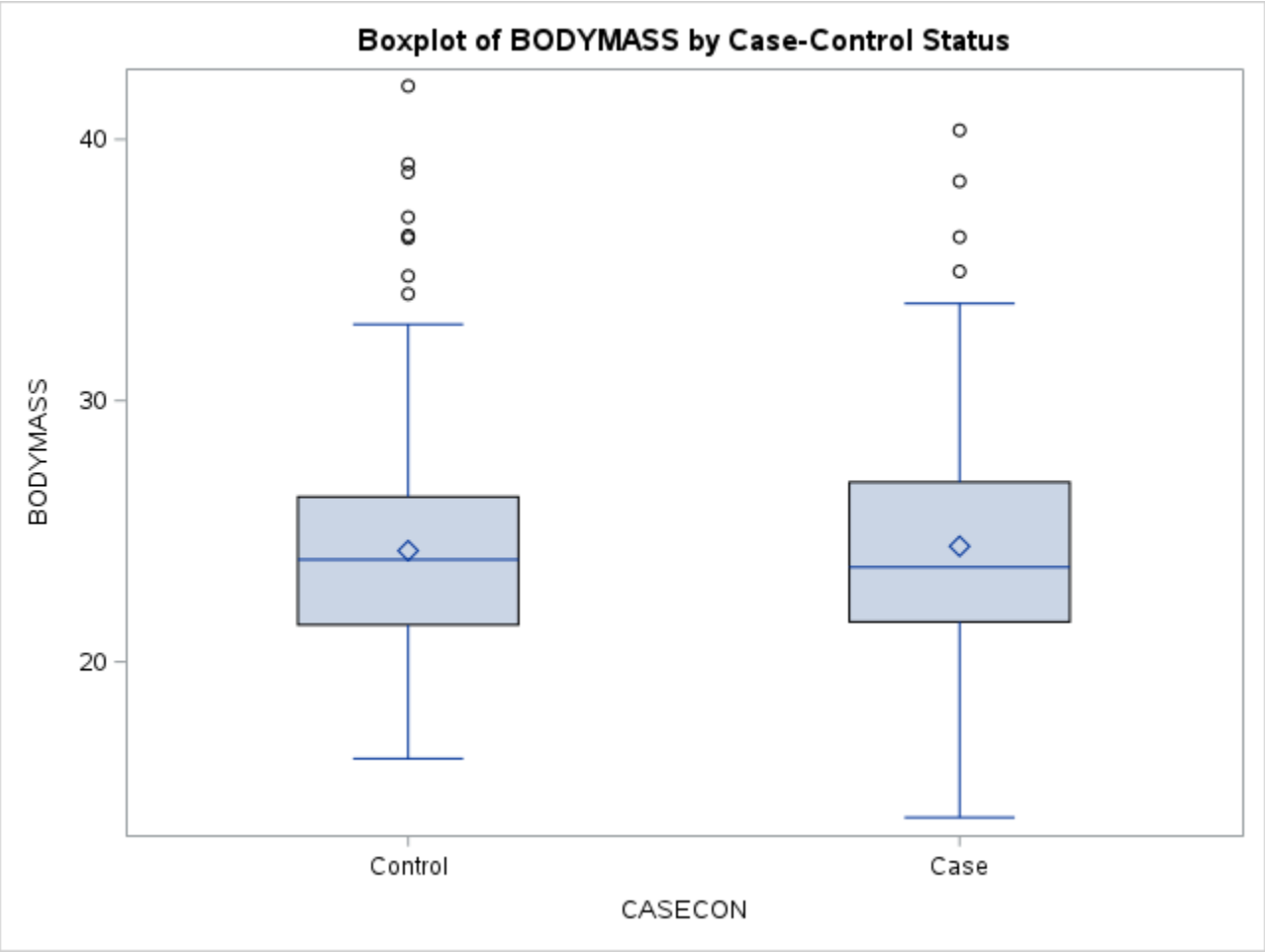
```
297         run;

NOTE: PROC LOGISTIC is modeling the probability that CASECON='Case'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 445 observations read from the data set WORK.COMBINED.
      WHERE cigstat not = .;
NOTE: PROCEDURE LOGISTIC used (Total process time):
      real time           0.17 seconds
      user cpu time       0.11 seconds
      system cpu time     0.01 seconds
      memory              5117.81k
      OS Memory           35400.00k
      Timestamp           12/08/2023 06:49:52 AM
      Step Count                        316  Switch Count  13
      Page Faults                       0
      Page Reclaims                     691
      Page Swaps                        0
      Voluntary Context Switches        398
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           520


298
299         /* create a tertile variable for `PACKYRS` */
300         proc rank data=combined out=combined groups=3;
301         where PACKYRS^=.;
302         var PACKYRS;
303         ranks PACKYRSTERT;
304         run;

NOTE: The data set WORK.COMBINED has 427 observations and 15 variables.
NOTE: PROCEDURE RANK used (Total process time):
      real time           0.00 seconds
      user cpu time       0.00 seconds
      system cpu time     0.00 seconds
      memory              3110.65k
      OS Memory           35768.00k
      Timestamp           12/08/2023 06:49:52 AM
      Step Count                        317  Switch Count  11
      Page Faults                       0
      Page Reclaims                     204
      Page Swaps                        0
      Voluntary Context Switches        48
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           536
```

```
305
306         /* check the new variable PACKYRSTERT */
307         proc freq data=combined;
308         tables PACKYRSTERT*PACKYRS/list missing;
309         run;

NOTE: There were 427 observations read from the data set WORK.COMBINED.
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.07 seconds
      user cpu time       0.07 seconds
      system cpu time     0.00 seconds
      memory              1641.37k
      OS Memory           34236.00k
      Timestamp           12/08/2023 06:49:52 AM
      Step Count                        318  Switch Count  8
      Page Faults                       0
      Page Reclaims                     322
      Page Swaps                        0
      Voluntary Context Switches        49
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           1104


310
311         /* From the frequency table, we can observe that: */
312         /* first tertile: 0 pk-yrs */
313         /* second tertle: 0.025 - 16 pk-yrs */
314         /* third tertle: 16.5 - 185.5 pk-yrs */
315
316         /* create a 3x2 table for `packyrstert` and `casecon` */
317         proc freq data=combined order=data;
318         where packyrstert^=.;
319         tables packyrstert*casecon/ chisq nocum norow nopercent;
320         format packyrstert packyrstertgroup. casecon casecongroup.;
321         run;

NOTE: There were 427 observations read from the data set WORK.COMBINED.
      WHERE packyrstert not = .;
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.02 seconds
      user cpu time       0.03 seconds
      system cpu time     0.00 seconds
      memory              1014.75k
      OS Memory           33716.00k
```

```
       Timestamp           12/08/2023 06:49:52 AM
       Step Count                        319  Switch Count  5
       Page Faults                       0
       Page Reclaims                     190
       Page Swaps                        0
       Voluntary Context Switches        28
       Involuntary Context Switches      0
       Block Input Operations            0
       Block Output Operations           528


 322
 323        /* use logistic regression to calculate odds ratio for
packyrstert*/
 324        proc logistic data=combined;
 325        where packyrstert^=.;
 326        format packyrstert packyrstertgroup. casecon casecongroup.;
 327        class packyrstert(ref='0 pk-yrs');
 328        model casecon(event='Case') = packyrstert;
 329        oddsratio packyrstert;
 330        run;

 NOTE: PROC LOGISTIC is modeling the probability that CASECON='Case'.
 NOTE: Convergence criterion (GCONV=1E-8) satisfied.
 NOTE: There were 427 observations read from the data set WORK.COMBINED.
       WHERE packyrstert not = .;
 NOTE: PROCEDURE LOGISTIC used (Total process time):
       real time           0.17 seconds
       user cpu time       0.11 seconds
       system cpu time     0.01 seconds
       memory              5104.34k
       OS Memory           35400.00k
       Timestamp           12/08/2023 06:49:52 AM
       Step Count                        320  Switch Count  12
       Page Faults                       0
       Page Reclaims                     690
       Page Swaps                        0
       Voluntary Context Switches        382
       Involuntary Context Switches      0
       Block Input Operations            0
       Block Output Operations           528


 331
 332        /* use logistic regression to calculate odds ratio for
packyrstert, adjusted for age and gender*/
 333        proc logistic data=combined;
```

```
 334        format packyrstert packyrstertgroup. casecon casecongroup. age
agegroup. gender gendergroup.;
 335        class packyrstert(ref='0 pk-yrs');
 336        model casecon(event='Case') = packyrstert age gender;
 337        oddsratio packyrstert;
 338        run;

 NOTE: PROC LOGISTIC is modeling the probability that CASECON='Case'.
 NOTE: Convergence criterion (GCONV=1E-8) satisfied.
 NOTE: There were 427 observations read from the data set WORK.COMBINED.
 NOTE: PROCEDURE LOGISTIC used (Total process time):
       real time           0.18 seconds
       user cpu time       0.10 seconds
       system cpu time     0.01 seconds
       memory              5007.37k
       OS Memory           35140.00k
       Timestamp           12/08/2023 06:49:52 AM
       Step Count                        321  Switch Count  0
       Page Faults                       0
       Page Reclaims                     657
       Page Swaps                        0
       Voluntary Context Switches        357
       Involuntary Context Switches      0
       Block Input Operations            0
       Block Output Operations           528


 339
 340        OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
 350
```

Q3a. Generate a box plot for BMI by case-control status.



Q3b. Does the mean BMI differ among cases and controls?
Answer:

1. From eyeballing the plot: they are equal.
2. Using the result from t test:
The results from the t-test indicate a p-value of 0.083 in the Equality of Variances test. This suggests that the variances of the two groups are not significantly different, warranting the use of the Pooled method for further analysis. Applying the Pooled method yields a p-value of 0.693. Given that this p-value is higher than 0.050, we fail to reject the null hypothesis. Therefore, based on this analysis, the means of BMI between the case group and the control group are equal.

SAS log for code related to Q3:

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         /* generate a box plot for BMI by case-control status */
70         proc sgplot data=combined;
71         where BODYMASS^=.;
72         format casecon casecongroup.;
73         vbox BODYMASS / category=casecon;
74         title "Boxplot of BODYMASS by Case-Control Status";
75         run;

NOTE: PROCEDURE SGPLOT used (Total process time):
      real time           0.08 seconds
      user cpu time       0.04 seconds
      system cpu time     0.01 seconds
      memory              8971.78k
      OS Memory           31536.00k
      Timestamp           12/08/2023 06:53:10 AM
      Step Count                        339  Switch Count  5
      Page Faults                       0
      Page Reclaims                     1593
      Page Swaps                        0
      Voluntary Context Switches        261
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           744

NOTE: There were 425 observations read from the data set WORK.COMBINED.
      WHERE BODYMASS not = .;

76
77         /* perform a t test to compare means of BODYMASS between Cases
and Controls */
78         proc ttest data=combined order=data;
79         format casecon casecongroup.;
80         class casecon;
81         var bodymass;
82         run;

NOTE: PROCEDURE TTEST used (Total process time):
      real time           0.30 seconds
      user cpu time       0.15 seconds
      system cpu time     0.06 seconds
      memory              11551.40k
      OS Memory           38864.00k
```

```
      Timestamp           12/08/2023 06:53:11 AM
      Step Count                        340  Switch Count  44
      Page Faults                       0
      Page Reclaims                     26200
      Page Swaps                        0
      Voluntary Context Switches        1122
      Involuntary Context Switches      2
      Block Input Operations            0
      Block Output Operations           1592


83
84         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
94
```

Q4. In Q2 you calculated the crude odds ratio for the cigarette smoking (ever/never) association with SAH. Calculate the (a) odds ratios for smoking (ever vs. never)-SAH association, stratified by alcohol status (ALCEVER), and the (b) Mantel-Haenszel odds ratio (i.e., alcohol adjusted cigarette smoking-SAH association)

(1) ALCEVER=No, odds ratio = 2.24
ALCEVER=Yes, odds ratio = 4.02
(2) MH-OR = 4.02

Q4. Extra credit:
Is alcohol an effect modifier of the cigarette smoking - SAH association? (Support your answer quantitatively)

There is a noticeable difference in the odds ratios between the two strata (2.24 vs. 4.02). This difference suggests that the effect of cigarette smoking on the outcome is different depending on the alcohol status, which is indicative of effect modification.

SAS log for code related to Q4:

```
1          OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
68
69         /* Q4 */
70         /* sort the data first by alcever */
71         proc sort data=combined;
72         by alcever;
73         run;

NOTE: Input data set is already sorted, no sorting done.
NOTE: PROCEDURE SORT used (Total process time):
      real time           0.00 seconds
      user cpu time       0.01 seconds
      system cpu time     0.00 seconds
      memory              649.53k
      OS Memory           25764.00k
      Timestamp           12/08/2023 07:03:29 AM
      Step Count                        476  Switch Count  0
      Page Faults                       0
      Page Reclaims                     50
      Page Swaps                        0
      Voluntary Context Switches        0
      Involuntary Context Switches      0
      Block Input Operations            0
      Block Output Operations           0


74
75         /* obtain odds ratio for `eversmoker` stratified by `ALCEVER`
*/
76         proc freq data=combined;
77         where eversmoker^=. and ALCEVER^=.;
78             tables eversmoker*casecon / chisq nocum norow nopercent
relrisk;
79             by ALCEVER;
80             format eversmoker eversmokergroup. casecon casecongroup.
alcever alcevergroup.;
81         run;

NOTE: There were 426 observations read from the data set WORK.COMBINED.
      WHERE (eversmoker not = .) and (ALCEVER not = .);
NOTE: PROCEDURE FREQ used (Total process time):
      real time           0.05 seconds
      user cpu time       0.05 seconds
      system cpu time     0.00 seconds
      memory              2677.46k
      OS Memory           26284.00k
```

```
       Timestamp           12/08/2023 07:03:29 AM
       Step Count                        477  Switch Count  5
       Page Faults                       0
       Page Reclaims                     196
       Page Swaps                        0
       Voluntary Context Switches        23
       Involuntary Context Switches      0
       Block Input Operations            0
       Block Output Operations           552


 82
 83         /* obtain MH odds ratio for `eversmoker` stratified by
`ALCEVER` */
 84         proc freq data=combined;
 85         where eversmoker^=. and ALCEVER^=.;
 86             tables eversmoker*casecon / chisq nocum cmh norow
nopercent;
 87             by ALCEVER;
 88             format eversmoker eversmokergroup. casecon casecongroup.
alcever alcevergroup.;
 89         run;

 NOTE: There were 426 observations read from the data set WORK.COMBINED.
       WHERE (eversmoker not = .) and (ALCEVER not = .);
 NOTE: PROCEDURE FREQ used (Total process time):
       real time           0.07 seconds
       user cpu time       0.08 seconds
       system cpu time     0.00 seconds
       memory              1439.21k
       OS Memory           26284.00k
       Timestamp           12/08/2023 07:03:30 AM
       Step Count                        478  Switch Count  7
       Page Faults                       0
       Page Reclaims                     197
       Page Swaps                        0
       Voluntary Context Switches        29
       Involuntary Context Switches      1
       Block Input Operations            0
       Block Output Operations           560


 90
 91         OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
 101
```