

Name Jerry

Student Number _____

⁴
STA 441s 2020 Quiz 6

1. (6 points) In a study of factors related to successful brain surgery in adults, the explanatory variables were age of patient in years (x_1), and experience of the surgeon, measured in number of times he or she had carried out the procedure (x_2). The response variable was $Y = 1$ for a successful surgery, or $Y = 0$ for unsuccessful. Denoting the probability of success by π , the logistic regression model is

$$\ln \left(\frac{\pi}{1-\pi} \right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$

- (a) You want to test whether, controlling for patient's age, the surgeon's experience is related to whether the surgery is successful. In symbols, what is the null hypothesis?

$$H_0: \beta_2 = 0$$

- (b) You want to test whether, controlling for surgeon's experience, the patient's age is related to whether the surgery is successful. In symbols, what is the null hypothesis?

$$H_0: \beta_1 = 0$$

- (c) You want to test patient's age and surgeon's experience simultaneously. In symbols, what is the null hypothesis?

$$H_0: \beta_1 = \beta_2 = 0$$

- (d) For fixed age of patient, each time the surgeon has carried out the procedure in the past multiplies the odds of success by ... what? Answer in terms of symbols from the logistic regression model.

$$e^{\beta_2}$$

- (e) For any fixed level of surgeon's experience, each each additional year of the patient's age multiplies the odds of success by ... what? Answer in terms of symbols from the logistic regression model.

$$e^{\beta_1}$$

- (f) Would you expect β_1 to be positive or negative? Why? I am asking for a common sense argument (a *short* one), not mathematical argument.

Negative. The older the patient, the lower the odds of success.

$$\ln\left(\frac{\pi}{1-\pi}\right) = \ln\left(\frac{P(Y=1|\mathbf{X}=\mathbf{x})}{P(Y=0|\mathbf{X}=\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}$$

$$P(Y = 1|\mathbf{X} = \mathbf{x}) = \frac{e^{\beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}}}{1 + e^{\beta_0 + \beta_1 x_1 + \dots + \beta_{p-1} x_{p-1}}}$$

2. This question is based on your analysis of the bird-keeping and cancer data.

(a) (2 points) You wish to test whether, controlling for the other variables, being a bird keeper is related to getting cancer.

i. Fill in the table below. On your printout, circle the chi-squared statistic and write "Question 2a" beside it.

All must
be correct
for one point

Chi-squared Statistic (a number)	p-value (a number)	Reject Null Hypothesis? (Yes or No)	Statistically Significant? (Yes or No)
10.9766	0.0009	Yes	Yes

Or LR G-squared = 11.67, p = 0.0006

Yes

Yes

ii. In plain, non-statistical language, what do you conclude? No marks for this without the first part right. Start your answer with "Allowing for other possible risk factors ..."

Allowing for other possible risk factors,
bird keepers are more likely to get cancer.

(b) (1 point) For a non-smoking, bird-keeping woman of average age and low socioeconomic status, what is the estimated probability of lung cancer? Write the answer in the space below. On your printout, circle the number (produced by `proc iml`) and write "Question 2b" beside it.

0.093

(c) (1 point) For a non-smoking, **non**-bird-keeping woman of average age and low socioeconomic status, what is the estimated probability of lung cancer? Write the answer in the space below. On your printout, circle the number (produced by `proc iml`) and write "Question 2c" beside it.

0.026 or 0.025

Attach your complete log file and your **COMPLETE** results file to the quiz paper. Make sure your name and student number are written clearly on both printouts.

STA441s24 Assignment 9 Check: Logistic Regression on Bird Lung Data

The FREQ Procedure

cancer	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	98	66.67	98	66.67
Yes	49	33.33	147	100.00

sex	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Male	111	75.51	111	75.51
Female	36	24.49	147	100.00

ses	Frequency	Percent	Cumulative Frequency	Cumulative Percent
Low	102	69.39	102	69.39
High	45	30.61	147	100.00

birdkeep	Frequency	Percent	Cumulative Frequency	Cumulative Percent
No	80	54.42	80	54.42
Yes	67	45.58	147	100.00

cigday	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	17	11.56	17	11.56
1	2	1.36	19	12.93
2	2	1.36	21	14.29
3	1	0.68	22	14.97
4	2	1.36	24	16.33
5	2	1.36	26	17.69
6	3	2.04	29	19.73
8	2	1.36	31	21.09
10	15	10.20	46	31.29
12	5	3.40	51	34.69
15	27	18.37	78	53.06
16	1	0.68	79	53.74
18	2	1.36	81	55.10
20	38	25.85	119	80.95
25	15	10.20	134	91.16
30	7	4.76	141	95.92
37	1	0.68	142	96.60
40	4	2.72	146	99.32
45	1	0.68	147	100.00

STA441s24 Assignment 9 Check: Logistic Regression on Bird Lung Data

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
age	147	56.9659864	7.3488560	37.0000000	67.0000000
yrsmoke	147	27.8503401	13.9756901	0	50.0000000
cigday	147	15.7482993	9.7037230	0	45.0000000

STA441s24 Assignment 9 Check: Logistic Regression on Bird Lung Data
Q11: Full model

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TWEET
Response Variable	cancer
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	147
Number of Observations Used	147

Response Profile		
Ordered Value	cancer	Total Frequency
1	No	98
2	Yes	49

Probability modeled is cancer='Yes'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	189.135	168.198
SC	192.126	189.131
-2 Log L	187.135	154.198

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	32.9367	6	<.0001
Score	27.9620	6	<.0001
Wald	22.5697	6	0.0010

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-1.9375	1.8042	1.1531	0.2829
sex	1	0.5613	0.5312	1.1166	0.2907
ses	1	0.1054	0.4688	0.0506	0.8221
birdkeep	1	1.3626	0.4113	10.9766	0.0009
age	1	-0.0398	0.0355	1.2552	0.2626
yrsmoke	1	0.0729	0.0265	7.5677	0.0059
cigday	1	0.0260	0.0255	1.0390	0.3081

Question 2a

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
sex	1.753	0.619	4.964
ses	1.111	0.443	2.785
birdkeep	3.906	1.745	8.747
age	0.961	0.896	1.030
yrsmoke	1.076	1.021	1.133
cigday	1.026	0.976	1.079

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	77.4	Somers' D	0.549
Percent Discordant	22.5	Gamma	0.550
Percent Tied	0.0	Tau-a	0.246
Pairs	4802	c	0.775

STA441s24 Assignment 9 Check: Logistic Regression on Bird Lung Data Estimated Probability of Cancer

	pihat1
Non-smoking, bird-keeping woman of average age and low socioeconomic status	0.0929601

	Pihat2
Non-smoking, NON-bird-keeping woman of average age and low socioeconomic status	0.0255657

STA441s24 Assignment 9 Check: Logistic Regression on Bird Lung Data Estimated Probability of Cancer copying numbers from the printout

	pihat1
Non-smoking, bird-keeping woman of average age and low socioeconomic status	0.0927217

Question 2b

	Pihat2
Non-smoking, NON-bird-keeping woman of average age and low socioeconomic status	0.025495

Question 2c

STA441s24 Assignment 9 Check: Logistic Regression on Bird Lung Data

Q12: Reduced model for LR test

The LOGISTIC Procedure

Model Information	
Data Set	WORK.TWEET
Response Variable	cancer
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	147
Number of Observations Used	147

Response Profile		
Ordered Value	cancer	Total Frequency
1	No	98
2	Yes	49

Probability modeled is cancer='Yes'.

Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	189.135	177.868
SC	192.126	195.811
-2 Log L	187.135	165.868

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	21.2668	5	0.0007
Score	17.0706	5	0.0044
Wald	13.6467	5	0.0180

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.2973	1.6600	0.0321	0.8579
sex	1	0.7492	0.5050	2.2010	0.1379

ses	1	-0.0730	0.4389	0.0277	0.8679
age	1	-0.0580	0.0343	2.8542	0.0911
yrsmoke	1	0.0796	0.0264	9.1056	0.0025
cigday	1	0.0198	0.0242	0.6667	0.4142

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
sex	2.115	0.786	5.692
ses	0.930	0.393	2.197
age	0.944	0.882	1.009
yrsmoke	1.083	1.028	1.140
cigday	1.020	0.973	1.070

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	72.2	Somers' D	0.446
Percent Discordant	27.7	Gamma	0.446
Percent Tied	0.1	Tau-a	0.199
Pairs	4802	c	0.723

STA441s24 Assignment 9 Check: Logistic Regression on Bird Lung Data LR Test for Bird Keeping

Gsq	df	pvalue
11.67	1	0.0006352

This is okay for Q2a

```

1      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
NOTE: ODS statements in the SAS Studio environment may disable some output features.
73
74      /***** 441s24HW9Birdlung.sas *****/
75      title 'STA441s24 Assignment 9 Check: Logistic Regression on Bird Lung Data';
76
77      proc format;
78          value ynfmt 0 = "No" 1 = "Yes";
NOTE: Format YNFMT is already on the library WORK.FORMATS.
NOTE: Format YNFMT has been output.
79          value sexfmt 0 = "Male" 1 = "Female";
NOTE: Format SEXFMT is already on the library WORK.FORMATS.
NOTE: Format SEXFMT has been output.
80          value sesfmt 0 = "Low" 1 = "High";
NOTE: Format SESFMT is already on the library WORK.FORMATS.
NOTE: Format SESFMT has been output.
81

```

```

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             243.56k
      OS Memory          27812.00k
      Timestamp          03/14/2024 04:31:30 PM
      Step Count          91  Switch Count  0
      Page Faults         0
      Page Reclaims       25
      Page Swaps          0
      Voluntary Context Switches 0
      Involuntary Context Switches 0
      Block Input Operations 0
      Block Output Operations 32

```

```

82      data tweet;
83          infile '/home/u1407221/441s24/data/birdlung.data.txt';
84          input cancer sex ses birdkeep age yrs smoke cigday;
85          format sex sexfmt.;
86          format ses sesfmt.;
87          format cancer birdkeep ynfmt.;
88

```

```

NOTE: The infile '/home/u1407221/441s24/data/birdlung.data.txt' is:
      Filename=/home/u1407221/441s24/data/birdlung.data.txt,
      Owner Name=u1407221, Group Name=oda,
      Access Permission=-rw-r--r--,
      Last Modified=14Mar2024:11:54:19,
      File Size (bytes)=3526

```

```

NOTE: 147 records were read from the infile '/home/u1407221/441s24/data/birdlung.data.txt'.
      The minimum record length was 21.
      The maximum record length was 22.

```

```

NOTE: The data set WORK.TWEET has 147 observations and 7 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             760.34k
      OS Memory          28328.00k
      Timestamp          03/14/2024 04:31:30 PM
      Step Count          92  Switch Count  2
      Page Faults         0
      Page Reclaims      155
      Page Swaps          0
      Voluntary Context Switches 15
      Involuntary Context Switches 0
      Block Input Operations 0

```


Block Output Operations 264

```

89      proc freq data=tweet;
90          tables cancer sex ses birdkeep cigday;
91

```

NOTE: There were 147 observations read from the data set WORK.TWEET.

NOTE: PROCEDURE FREQ used (Total process time):

```

real time          0.06 seconds
user cpu time      0.07 seconds
system cpu time    0.00 seconds
memory            3403.50k
OS Memory          29612.00k
Timestamp          03/14/2024 04:31:30 PM
Step Count         93  Switch Count  2
Page Faults        0
Page Reclaims      363
Page Swaps         0
Voluntary Context Switches  11
Involuntary Context Switches 1
Block Input Operations  0
Block Output Operations 296

```

```

92      proc means data=tweet;
93          var age yrs smoke cigday;
94      run;

```

NOTE: There were 147 observations read from the data set WORK.TWEET.

NOTE: PROCEDURE MEANS used (Total process time):

```

real time          0.02 seconds
user cpu time      0.02 seconds
system cpu time    0.01 seconds
memory            6604.78k
OS Memory          35516.00k
Timestamp          03/14/2024 04:31:30 PM
Step Count         94  Switch Count  1
Page Faults        0
Page Reclaims      1542
Page Swaps         0
Voluntary Context Switches  19
Involuntary Context Switches 0
Block Input Operations  0
Block Output Operations  0

```

```

94      !      /* Without this run statement, ods output gives a misleading warning. */
95
96      ods output ParameterEstimates = estimout;
97      /* The ParameterEstimates table will be written to a SAS data
98      set called estimout. */
99      proc logistic data=tweet;
100          title2 'Q11: Full model';
101          model cancer (event='Yes') = sex ses birdkeep age yrs smoke cigday;
102

```

NOTE: PROC LOGISTIC is modeling the probability that cancer='Yes'.

NOTE: Convergence criterion (GCONV=1E-8) satisfied.

NOTE: The data set WORK.ESTIMOUT has 7 observations and 7 variables.

NOTE: There were 147 observations read from the data set WORK.TWEET.

NOTE: PROCEDURE LOGISTIC used (Total process time):

```

real time          0.06 seconds
user cpu time      0.06 seconds
system cpu time    0.01 seconds
memory            2750.46k
OS Memory          31932.00k
Timestamp          03/14/2024 04:31:30 PM
Step Count         95  Switch Count  3
Page Faults        0

```

```

Page Reclaims          350
Page Swaps              0
Voluntary Context Switches 24
Involuntary Context Switches 33
Block Input Operations   0
Block Output Operations  320

```

```

103      proc iml;
NOTE: IML Ready
104      title2 'Estimated Probabilty of Cancer';
105      use estimout;
106      read all var {Estimate} into b;
107      x1 = {1, 1, 0, 1, 56.9659864, 0, 0};
107      !                               /* Rows are separated by commas */
108      pihat1 = exp(x1`*b)/(1+exp(x1`*b));
109      print "Non-smoking, bird-keeping woman of average age and low socioeconomic status" pihat1;
110      x2 = {1, 1, 0, 0, 56.9659864, 0, 0};
110      !                               /* Rows are separated by commas */
111      Pihat2 = exp(x2`*b)/(1+exp(x2`*b));
112      print "Non-smoking, NON-bird-keeping woman of average age and low socioeconomic status" pihat2;
113
114      /* This was a nasty lesson in proc iml for me. When I tried to manually
115      assign b0 = -1.9375; b1 = 0.5613; and so on and then b = {b0, b1 etc,
116      the vector b was character valued and I could not find a way around
117      it. */
118
119      /* The crude way, taking numbers from the printout for comparison */
NOTE: Exiting IML.
NOTE: PROCEDURE IML used (Total process time):
      real time          0.01 seconds
      user cpu time      0.01 seconds
      system cpu time    0.00 seconds
      memory             811.90k
      OS Memory          30632.00k
      Timestamp          03/14/2024 04:31:30 PM
      Step Count         96      Switch Count  1
      Page Faults        0
      Page Reclaims      95
      Page Swaps         0
      Voluntary Context Switches 8
      Involuntary Context Switches 1
      Block Input Operations 0
      Block Output Operations 16

120      proc iml;
NOTE: IML Ready
121      title2 'Estimated Probabilty of Cancer copying numbers from the printout';
122      b = { -1.9375, 0.5613, 0.1054, 1.3626, -0.0398, 0.0729, 0.0260};
123      x1 = {1, 1, 0, 1, 56.9659864, 0, 0};
123      !                               /* Rows are separated by commas */
124      pihat1 = exp(x1`*b)/(1+exp(x1`*b));
125      print "Non-smoking, bird-keeping woman of average age and low socioeconomic status" pihat1;
126      x2 = {1, 1, 0, 0, 56.9659864, 0, 0};
126      !                               /* Rows are separated by commas */
127      Pihat2 = exp(x2`*b)/(1+exp(x2`*b));
128      print "Non-smoking, NON-bird-keeping woman of average age and low socioeconomic status" pihat2;
129
130
NOTE: Exiting IML.
NOTE: PROCEDURE IML used (Total process time):
      real time          0.01 seconds
      user cpu time      0.01 seconds
      system cpu time    0.00 seconds
      memory             527.00k
      OS Memory          30372.00k
      Timestamp          03/14/2024 04:31:30 PM
      Step Count         97      Switch Count  1
      Page Faults        0

```

```

Page Reclaims          61
Page Swaps              0
Voluntary Context Switches 10
Involuntary Context Switches 10
Block Input Operations   0
Block Output Operations   0

131      proc logistic data=tweet;
132          title2 'Q12: Reduced model for LR test';
133          model cancer (event='Yes') = sex ses age yrs smoke cigday; /* Deleting birdkeep */
134      run;

NOTE: PROC LOGISTIC is modeling the probability that cancer='Yes'.
NOTE: Convergence criterion (GCONV=1E-8) satisfied.
NOTE: There were 147 observations read from the data set WORK.TWEET.
NOTE: PROCEDURE LOGISTIC used (Total process time):
    real time          0.06 seconds
    user cpu time       0.06 seconds
    system cpu time     0.00 seconds
    memory              2132.87k
    OS Memory           31672.00k
    Timestamp           03/14/2024 04:31:30 PM
    Step Count          98   Switch Count   1
    Page Faults         0
    Page Reclaims       198
    Page Swaps          0
    Voluntary Context Switches 10
    Involuntary Context Switches 0
    Block Input Operations 0
    Block Output Operations 72

135
136      proc iml;
NOTE: IML Ready
137          title2 'LR Test for Bird Keeping';
138          Gsq = 165.868-154.198;
139      !                df=1;
139          pvalue = 1-probchi(Gsq,df);
140          print Gsq df pvalue;
141
142      quit;
NOTE: Exiting IML.
NOTE: PROCEDURE IML used (Total process time):
    real time          0.00 seconds
    user cpu time       0.01 seconds
    system cpu time     0.00 seconds
    memory              531.40k
    OS Memory           30372.00k
    Timestamp           03/14/2024 04:31:30 PM
    Step Count          99   Switch Count   1
    Page Faults         0
    Page Reclaims       57
    Page Swaps          0
    Voluntary Context Switches 10
    Involuntary Context Switches 0
    Block Input Operations 0
    Block Output Operations 8

143
144
145
146
147
148      OPTIONS NONOTES NOSTIMER NOSOURCE NOSYNTAXCHECK;
160

```