Seneca

SENECA COLLEGE OF APPLIED ARTS AND TECHNOLOGY

SENECA BUSINESS

BAN 100 - Statistics for Analytics

Other **Version NA**

DATE: 12/12/2021 **TIME ALLOWED: Two weeks**

PROFESSOR(S): Samaneh Gholami

Allowable Examination Aids: (check applicable boxes)

☒ Calculators (non-programmable only)	☒ Math Tables (normal distribution table)	☒ Periodic Tables
☐ Formula Sheets (attached)	☒ Textbooks	☒ Probability Tables
☒ Dictionary	☒ Notes	☒ Other

Answers to be completed on:

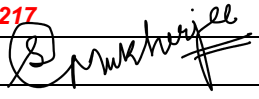
☐ Exam Booklet	☐ GradeMaster Card	☐ Exam Paper

TOTAL MARKS: 100 **WEIGHTED VALUE: 25**

INSTRUCTIONS:

Academic Integrity Policy. Seneca upholds a learning community that values academic integrity, honesty, fairness, trust, respect, responsibility and courage. These values enhance Seneca's commitment to students by delivering high-quality education and teaching excellence, while supporting a positive learning environment. The AI policy is always in effect. Note **Sections 2.3 and 2.4:**

*"**…2.3** Should there be a suspected violation of this policy (e.g.…cheating, falsification, impersonation or plagiarism), the academic integrity sanctions will be applied according to the severity of the offence committed. Refer to Appendix B for the academic integrity sanctions. **2.4** Should a suspected violation of this policy be a result of, or in combination with, a suspected violation of Seneca's Student Code of Conduct and/or another non-academic-related Seneca policy, the matter will be investigated and adjudicated through the process found in the Student Code of Conduct."*

TO BE COMPLETED BY STUDENT
SUBJECT SECTION NUMBER (e.g. QNM223 AA): ***BAN100ZBB***
STUDENT NAME: ***SUKANYA MUKHERJEE***
STUDENT NUMBER: ***128041217***
STUDENT SIGNATURE: 

APPROVED BY: _____

Cristina Italia, Interim Chair
School of Management and Entrepreneurship

DATE: ***12/12/2021***

ASSIGNMENT 6: LOGISTIC REGRESSION

DUE DATE IS DECEMBER 12, 2021

IT IS AN INDIVIDUAL ASSIGNMENT, WORTH 25% OF YOUR OVERALL GRADE. DOWNLOAD THE ASSIGNMENT QUESTIONS AND THE DATASETS FORM THE ASSIGNMENT SECTION ON BB.

Problem 1 (10 marks) File: Customer. xlsx

Consumer Reports conducted a taste test on some brands of boxed chocolates. The data show the price per serving, based on the FDA serving size of 1.4 ounces, and the quality rating for the chocolates tested.

```
proc import out = work.customer
datafile = "/home/u59406283/Assignment4PA/Customer.xlsx"
dbms = xlsx
replace;
getnames = yes;
datarow = 2 ; run;
proc print data= work.customer; format price dollar10.2; run;
```

H0: coefficient of all independent variables is zero

H1: at least one of the coefficients is non-zero

Suppose that you would like to determine whether products that cost more rate higher in

quality. use the following binary dependent variable:

y= 1 if the quality rating is very good or excellent and 0 if good or fair

```
data work.customer;
    set work.customer;
    if rating = "Very Good" or rating = "Excellent" then y = 1;
    else if rating = "Good" or rating = "Fair" then y = 0; run;
```

Obs	Manufacturer	Price	Rating	y
1	Bernard Callebaut	$3.17	Very Good	1
2	Candinas	$3.58	Excellent	1
3	Fannie May	$1.49	Good	0
4	Godiva	$2.91	Very Good	1
5	Hershey,Äôs	$0.76	Good	0
6	L.A. Burdick	$3.70	Very Good	1
7	La Maison du Chocolate	$5.08	Excellent	1
8	Leonidas	$2.11	Very Good	1
9	Lindt	$2.20	Good	0
10	Martine,Äôs	$4.76	Excellent	1
11	Michael Recchiuti	$7.05	Very Good	1
12	Neuchatel	$3.36	Good	0
13	Neuchatel Sugar Free	$3.22	Good	0
14	Richard Donnelly	$6.55	Very Good	1
15	Russell Stover	$0.70	Good	0
16	See,Äôs	$1.06	Very Good	1
17	Teuscher Lake of Zurich	$4.66	Very Good	1
18	Whitman,Äôs	$0.70	Fair	0
19	Whitman,Äôs Sugar Free	$1.21	Fair	0

a. Write the logistic regression equation relating x = price per serving to y. (3 marks)

```
proc reg data = work.customer;
model y = Price ; run;
```

The REG Procedure
Model: MODEL1
Dependent Variable: y

Number of Observations Read	19
Number of Observations Used	19

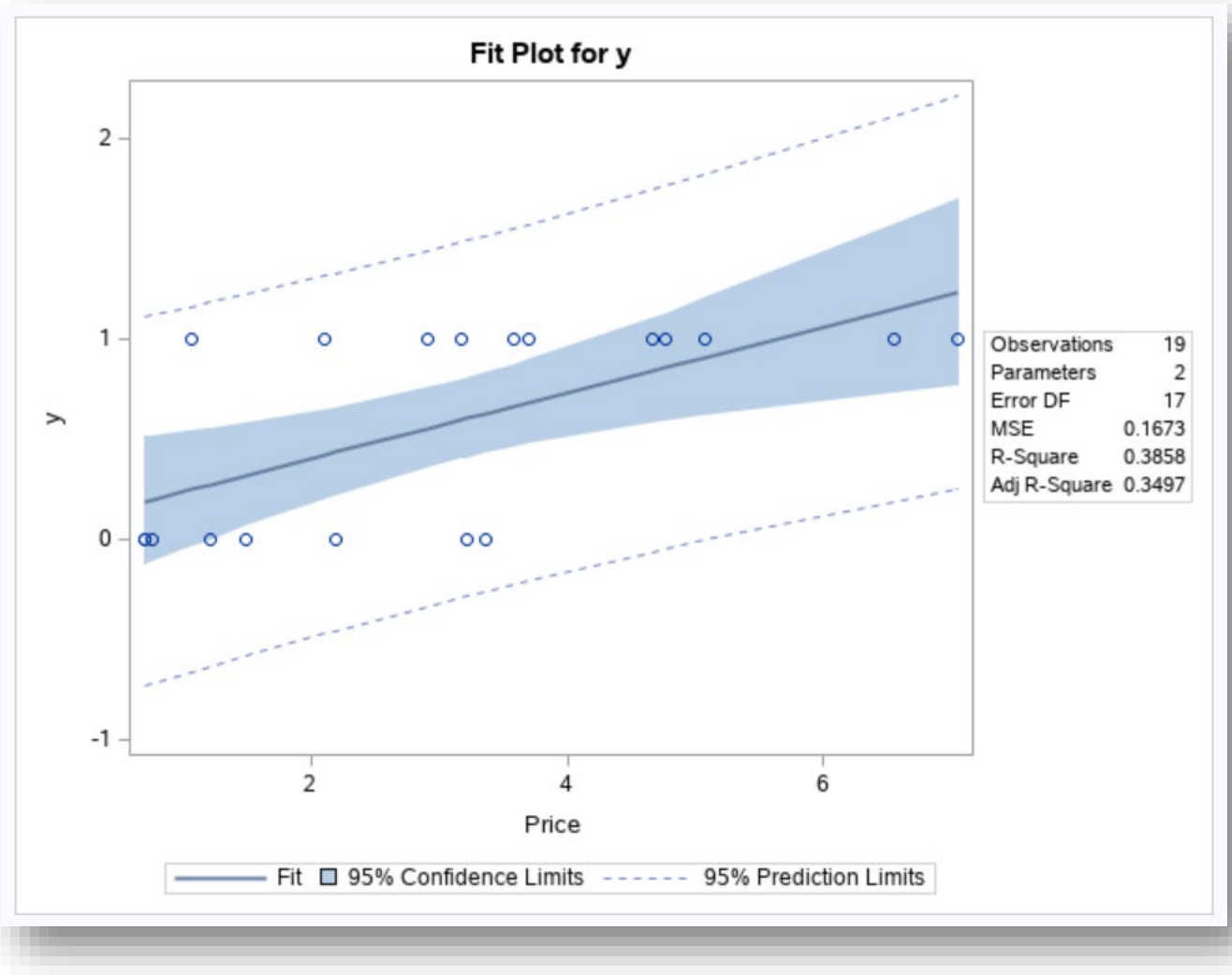
Analysis of Variance

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	1	1.78708	1.78708	10.68	0.0045
Error	17	2.84450	0.16732		
Corrected Total	18	4.63158			

Root MSE	0.40905	R-Square	0.3858
Dependent Mean	0.57895	Adj R-Sq	0.3497
Coeff Var	70.65444		

Parameter Estimates

Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|
Intercept	Intercept	1	0.07589	0.18028	0.42	0.6791
Price	Price	1	0.16403	0.05019	3.27	0.0045



b. Use SAS to compute the estimated logit.(2 mark)

```
Proc logistic data=work.customer;
model y = Price; run;
```

The LOGISTIC Procedure

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

Number of Observations Read	19
Number of Observations Used	19

Response Profile		
Ordered Value	y	Total Frequency
1	0	8
2	1	11

Probability modeled is y=0.

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	27.864	20.399
SC	28.808	22.288
-2 Log L	25.864	16.399

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.4648	1	0.0021
Score	7.3311	1	0.0068
Wald	4.9924	1	0.0255

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	2.8050	1.4316	3.8387	0.0501
Price	1	-1.1492	0.5143	4.9924	0.0255

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Price	0.317	0.116	0.868

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	86.4	Somers' D	0.727
Percent Discordant	13.6	Gamma	0.727
Percent Tied	0.0	Tau-a	0.374
Pairs	88	c	0.864

Concordant is 86.4 = accurate and good fit

Price (pvalue = 0.02) < alpha (0.05) = at least one of the coefficients is non-zero.

c. Use the estimated logit computed in part (b) to compute an estimate of the probability a chocolate that has a price per serving of $4.00 will have a quality rating of very good or excellent. (3 marks)

We have no exact observation for $4.00 and y=1 (very good or excellent),

As seen in the below output:

Obs	Manufacturer	Price	Rating	y
1	Bernard Callebaut	$3.17	Very Good	1
2	Candinas	$3.58	Excellent	1
3	Fannie May	$1.49	Good	0
4	Godiva	$2.91	Very Good	1
5	Hershey,Äôs	$0.76	Good	0
6	L.A. Burdick	$3.70	Very Good	1
7	La Maison du Chocolate	$5.08	Excellent	1
8	Leonidas	$2.11	Very Good	1
9	Lindt	$2.20	Good	0
10	Martine,Äôs	$4.76	Excellent	1
11	Michael Recchiuti	$7.05	Very Good	1
12	Neuchatel	$3.36	Good	0
13	Neuchatel Sugar Free	$3.22	Good	0
14	Richard Donnelly	$6.55	Very Good	1
15	Russell Stover	$0.70	Good	0
16	See,Äôs	$1.06	Very Good	1
17	Teuscher Lake of Zurich	$4.66	Very Good	1
18	Whitman,Äôs	$0.70	Fair	0
19	Whitman,Äôs Sugar Free	$1.21	Fair	0

d. What is the estimate of the odds ratio? What is its interpretation? (2 marks)

```
proc logistic data=work.customer;
class y (ref='1') / param=ref;
model y (event='1') = Price ;
```

The LOGISTIC Procedure

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

Number of Observations Read	19
Number of Observations Used	19

Response Profile		
Ordered Value	y	Total Frequency
1	0	8
2	1	11

Probability modeled is y=1.

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	27.864	20.399
SC	28.808	22.288
-2 Log L	25.864	16.399

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	9.4648	1	0.0021
Score	7.3311	1	0.0068
Wald	4.9924	1	0.0255

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-2.8050	1.4316	3.8387	0.0501
Price	1	1.1492	0.5143	4.9924	0.0255

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Price	3.156	1.152	8.647

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	86.4	Somers' D	0.727
Percent Discordant	13.6	Gamma	0.727
Percent Tied	0.0	Tau-a	0.374
Pairs	88	c	0.864

In the odds ratio estimates: odds ratio is greater than 1 (3.156) when y=1 means that there is greater odd of cost being high when rating is high than the cost being low when rating is high.

Problem 2 (11 marks) File: Titanic. xlsx

The data set contains personal information for 891 passengers, including an indicator variable for their survival, and the objective is to predict survival, or probability thereof, from the other characteristics. The survival data for all passengers is stored in the binary variable called Survived. The predictors include Sex (modeled with male/female dummy variables), Age (and additional dummy variables for ranges), Class (first, second, or third, modeled with dummy variables), SiblingSpouse (number of siblings and spouses accompanying the passenger, and corresponding dummy variables), ParentChild (number of parents and children accompanying the passenger, and corresponding dummy variables), and Embarked (ports of Cherbourg, QueensTown, and Southampton, modeled by dummy variables)

```
proc import out = work.titanic
datafile = "/home/u59406283/Assignment4PA/titanic.csv"
dbms = csv
replace;
getnames = yes;
datarow = 2 ; run;
proc print data= work.titanic;  run;
```

a. **Write the logistic regression equation relating Age and Survived. (2 mark)**

```
proc logistic data=work.titanic;
class Survived (ref='1') / param=ref;
model Survived (event='1') = Age ; run;
```

The LOGISTIC Procedure

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

Number of Observations Read	891
Number of Observations Used	714

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	424
2	1	290

Probability modeled is Survived='1'.

Note: 177 observations were deleted due to missing values for the response or explanatory variables.

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	966.516	964.228
SC	971.087	973.370
-2 Log L	964.516	960.228

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	4.2876	1	0.0384
Score	4.2577	1	0.0391
Wald	4.2310	1	0.0397

Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-0.0567	0.1736	0.1068	0.7438
Age	1	-0.0110	0.00533	4.2310	0.0397

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age	0.989	0.979	0.999

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	52.1	Somers' D	0.062
Percent Discordant	45.9	Gamma	0.063
Percent Tied	2.0	Tau-a	0.030
Pairs	122960	c	0.531

b. **For the Titanic data, use SAS to compute the estimated logistic regression equation. (2 marks)**

```
proc logistic data=work.titanic;
class Survived (ref='1')
Sex (ref='female') / param=ref;
model Survived (event='1') = Class Sex Age SiblingSpouse ParentChild; run;
```

The LOGISTIC Procedure

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

Number of Observations Read	891
Number of Observations Used	714

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	424
2	1	290

Probability modeled is Survived='1'.

Note: 177 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information		
Class	Value	Design Variables
Sex	female	0
	male	1

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	966.516	648.622
SC	971.087	676.048
-2 Log L	964.516	636.622

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	327.8937	5	<.0001
Score	285.5391	5	<.0001
Wald	192.9260	5	<.0001

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Class	1	87.2741	<.0001
Sex	1	144.4986	<.0001
Age	1	29.6795	<.0001
SiblingSpouse	1	8.3073	0.0039
ParentChild	1	0.0964	0.7561

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	5.6196	0.5467	105.6584	<.0001
Class		1	-1.3160	0.1409	87.2741	<.0001
Sex	male	1	-2.6374	0.2194	144.4986	<.0001
Age		1	-0.0445	0.00816	29.6795	<.0001
SiblingSpouse		1	-0.3646	0.1265	8.3073	0.0039
ParentChild		1	-0.0371	0.1196	0.0964	0.7561

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Class	0.268	0.204	0.353
Sex male vs female	0.072	0.047	0.110
Age	0.957	0.941	0.972
SiblingSpouse	0.694	0.542	0.890
ParentChild	0.964	0.762	1.218

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	85.8	Somers' D	0.718
Percent Discordant	14.0	Gamma	0.719
Percent Tied	0.1	Tau-a	0.347
Pairs	122960	c	0.859

Concordant is 85.8 = accurate and good fit

Class (pvalue < 0.0001) < alpha (0.05) = at least one of the coefficients is non-zero.

Sex (pvalue < 0.0001) < alpha (0.05) = at least one of the coefficients is non-zero.

Age (pvalue < 0.0001) < alpha (0.05) = at least one of the coefficients is non-zero.

Sibling Spouse (pvalue = 0.0039) < alpha (0.05) = at least one of the coefficients is non-zero.

ParentChild (pvalue = 0.7561) > alpha (0.05) = coefficient of independent variable is zero

c. **Estimate the probability of surviving the passenger with the average Age 30. (2 marks)**

```
data work.titanic;
    set work.titanic;
    if missing(Age) then Age_Group = ' ';
    else if Age lt 25 then Age_Group = '1:< 25';
    else if Age le 35 then Age_Group = '2:25-35';
    else Age_Group = '3:35+'; run;
proc logistic data=work.titanic;
class Age_Group (ref='2')
Survived (ref='1') / param=ref;
model Survived (event='1') = Age_Group ; run;
```

The LOGISTIC Procedure

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

Number of Observations Read	891
Number of Observations Used	714

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	424
2	1	290

Probability modeled is Survived='1'.

Note: 177 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information			
Class	Value	Design Variables	
Age_Group	1	1	0
	2	0	0
	3	0	1

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	966.516	969.624
SC	971.087	983.337
-2 Log L	964.516	963.624

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.8918	2	0.6403
Score	0.8902	2	0.6408
Wald	0.8896	2	0.6410

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age_Group	2	0.8896	0.6410

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.3789	0.1376	7.5846	0.0059
Age_Group	1	1	0.0744	0.1834	0.1645	0.6850
Age_Group	3	1	-0.1001	0.1961	0.2607	0.6097

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age_Group 1 vs 2	1.077	0.752	1.543
Age_Group 3 vs 2	0.905	0.616	1.329

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	35.0	Somers' D	0.039
Percent Discordant	31.2	Gamma	0.059
Percent Tied	33.8	Tau-a	0.019
Pairs	122960	c	0.519

Probability (Survived =1, Age_Group=2) = e (-0.3789) /1+ e (-0.3789) = 0.68/1.68 = 0.40

d. **Suppose we want to check who have a 0.50 or higher probability of surviving. What is the average age to achieve this level of probability? (3 marks)**

Survival = 0 or 1 in the dataset.

e. **What is the estimated odds ratio? What is the interpretation? (2 marks)**

```
proc logistic data=work.titanic;
class Age_Group (ref='3')
Survived (ref='1') / param=ref;
model Survived (event='1') = Age_Group ; run;
```

The LOGISTIC Procedure

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

Number of Observations Read	891
Number of Observations Used	714

Response Profile		
Ordered Value	Survived	Total Frequency
1	0	424
2	1	290

Probability modeled is Survived='1'.

Note: 177 observations were deleted due to missing values for the response or explanatory variables.

Class Level Information			
Class	Value	Design Variables	
Age_Group	1	1	0
	2	0	1
	3	0	0

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

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	966.516	969.624
SC	971.087	983.337
-2 Log L	964.516	963.624

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	0.8918	2	0.6403
Score	0.8902	2	0.6408
Wald	0.8896	2	0.6410

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age_Group	2	0.8896	0.6410

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-0.4790	0.1397	11.7596	0.0006
Age_Group	1	1	0.1745	0.1850	0.8895	0.3456
Age_Group	2	1	0.1001	0.1961	0.2607	0.6097

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Age_Group 1 vs 3	1.191	0.828	1.711
Age_Group 2 vs 3	1.105	0.753	1.623

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	35.0	Somers' D	0.039
Percent Discordant	31.2	Gamma	0.059
Percent Tied	33.8	Tau-a	0.019
Pairs	122960	c	0.519

Odds ratio tells us that:

Age_Group (2: 25-35) and Age_Group (1:<25) have similar approx. 1.1 more odds of survival than Age_Group (3: 35+)

Problem 3 (4 marks): Capital punishment

I ran two models between race and capital punishment. However, I recoded race two different ways:

Model 1: white coded as 0, blacks coded as 1

Model 2: blacks coded as 0, whites coded as 1

This gave me the following results:

	Model 1	Model 2
Coefficient	-1.081	1.081
Odds for whites	2.472	2.472
Odds for blacks	0.838	0.838
Odds ratio	0.34	2.95

a. **Why the odds ratios are different? Explain it (2 marks)**

Model 1 (Coefficient) = -1.081 so the exp (-1.081) = 0.34 (Odds ratio)

Model 2 (Coefficient) = 1.081 so the exp (1.081) = 2.95 (Odds ratio)

b. **Show the relation between the odd ratios and coefficient (2 marks)**

If a = coefficient then odds ratio = exp (a)

Model 1 (Coefficient) = -1.081 so the exp (-1.081) = 0.34 (Odds ratio)

Model 2 (Coefficient) = 1.081 so the exp (1.081) = 2.95 (Odds ratio)