LOGISTIC REGRESSION

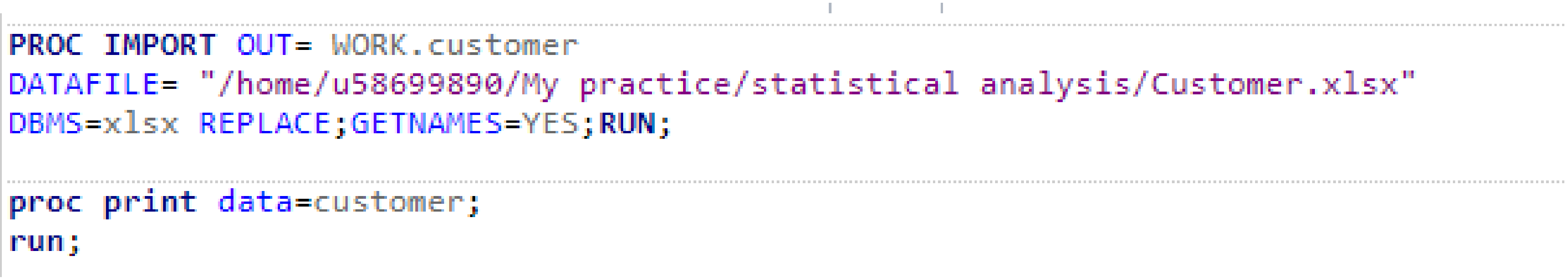
Maryam Najimigoshtasb

Problem 1 (12 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.

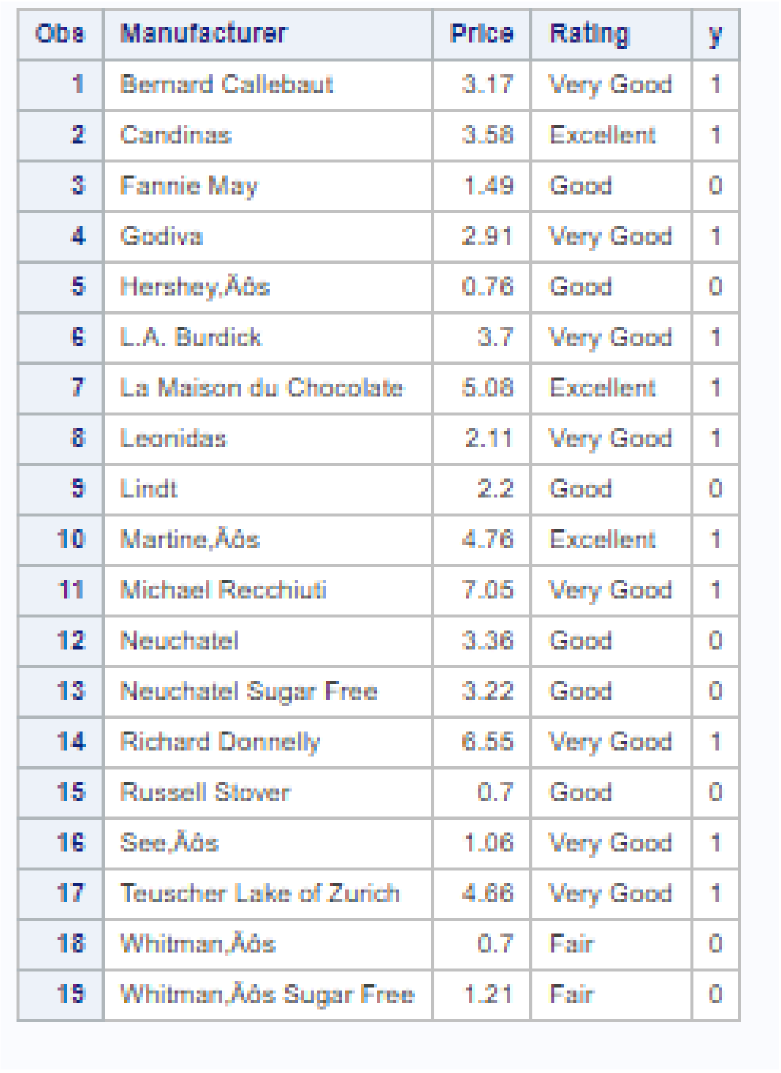
Suppose that you would like to determine whether products that cost more rate higher in quality. use the following binary dependent variable:

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



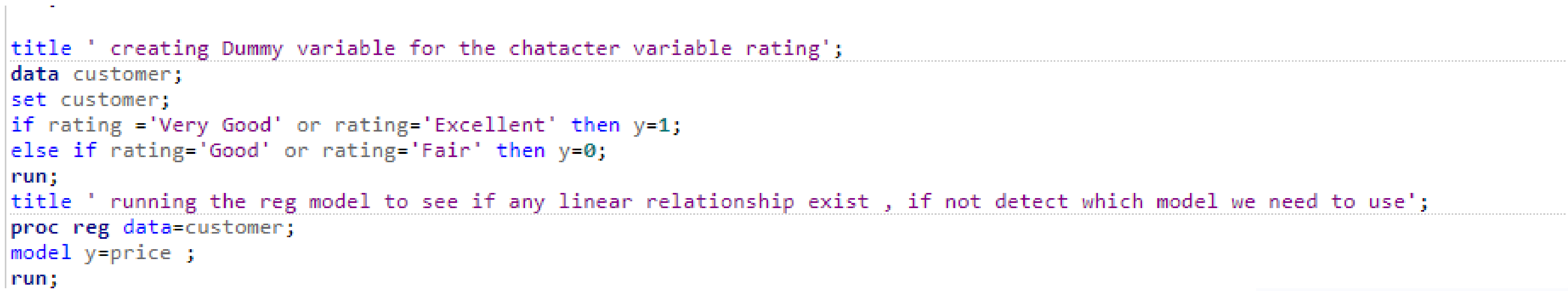
```
PROC IMPORT OUT= WORK.customer
DATAFILE= "/home/u58699890/My practice/statistical analysis/Customer.xlsx"
DBMS=xlsx REPLACE;GETNAMES=YES;RUN;

proc print data=customer;
run;
```



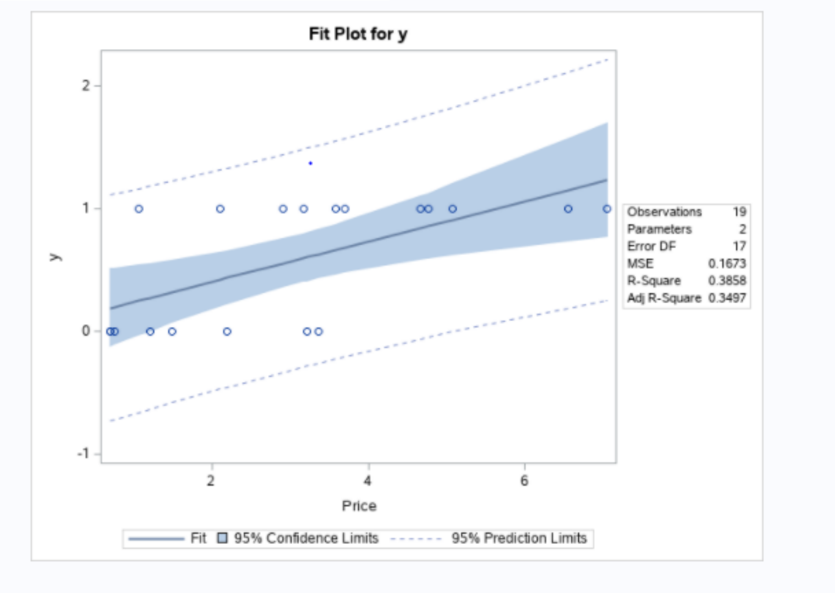
Obs	Manufacturer	Price	Rating	y
1	Bernard Callebaut	3.17	Very Good	1
2	Candinas	3.58	Excellent	1
3	Fannie May	1.49	Good	0
4	Godiva	2.91	Very Good	1
5	Hershey‚Äôs	0.76	Good	0
6	L.A. Burdick	3.7	Very Good	1
7	La Maison du Chocolate	5.08	Excellent	1
8	Leonidas	2.11	Very Good	1
9	Lindt	2.2	Good	0
10	Martine‚Äôs	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	0.7	Good	0
16	See‚Äôs	1.06	Very Good	1
17	Teuscher Lake of Zurich	4.66	Very Good	1
18	Whitman‚Äôs	0.7	Fair	0
19	Whitman‚Äôs Sugar Free	1.21	Fair	0

Creating Dummy variable and plotting the price vs dependent variable y which dummy of rating to explore the relationship



```
title ' creating Dummy variable for the chatacter variable rating';
data customer;
set customer;
if rating ='Very Good' or rating='Excellent' then y=1;
else if rating='Good' or rating='Fair' then y=0;
run;
title ' running the reg model to see if any linear relationship exist , if not detect which model we need to use';
proc reg data=customer;
model y=price ;
run;
```

Straight line is not a good fit for customer rating data, so run logistic regression for this data.

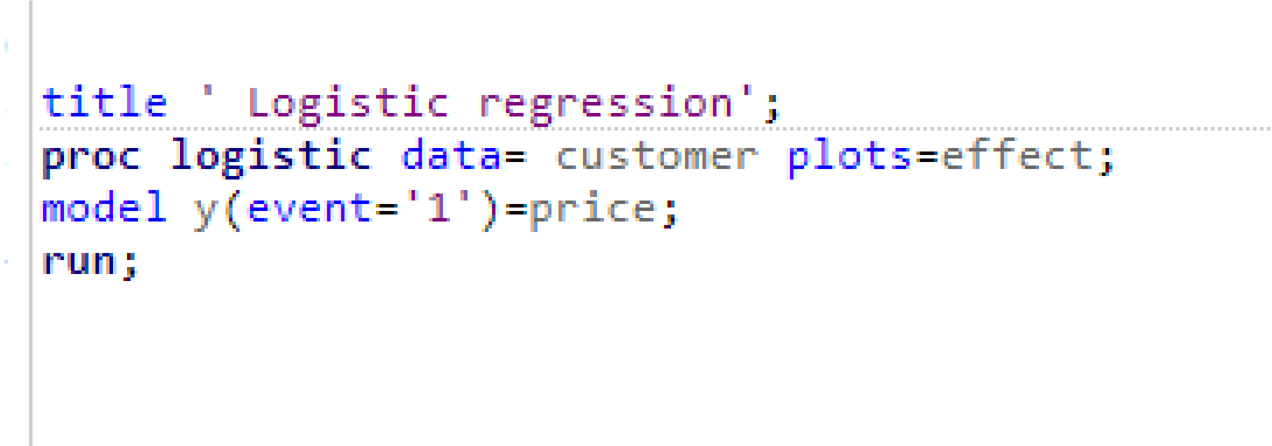


a. Write the logistic regression equation relating x = price per serving to y.

$$p(y\,|x) = \frac{e^{\beta 0+\ \beta 1\ x1}}{1+\ e^{\beta 0+\ \beta 1\ x1}}$$

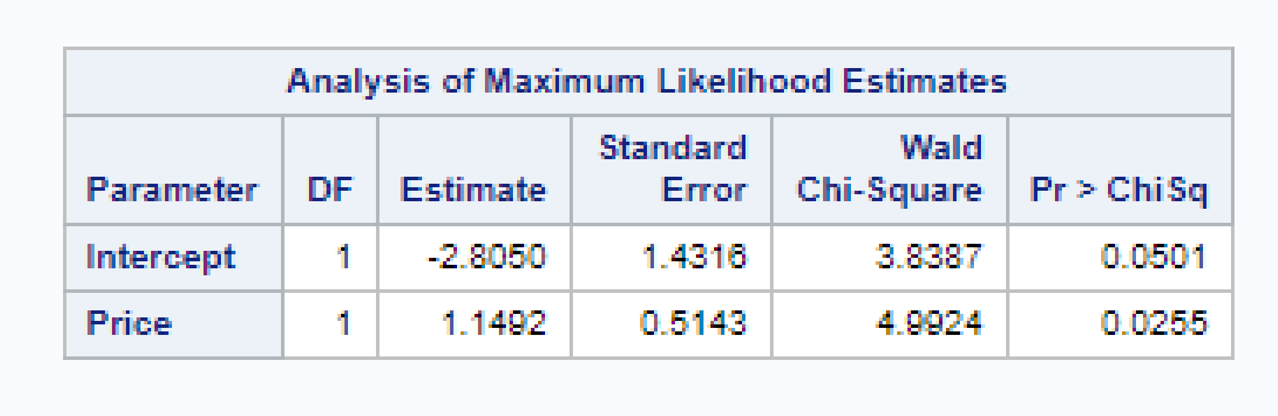
relating x = price per serving to y.

In this data there is just one independent variable price and one dependent variable y.
We run the logistic regression for y when the event is 1 which means when have a very good or excellent rating. So X1 here is price . Regression equation is the probability of having very good and excellent rate by given these set of prices in the data.



```
title ' Logistic regression';
proc logistic data= customer plots=effect;
model y(event='1')=price;
run;
```

b. Use SAS to compute the estimated logit.(



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

Estimate logit $\hat{g}$=−2.8050+1.1492*price

Estimated regression equation $\hat{Y}$=p(y=1|x)= $e^{-2.8050+1.1492*\text{price}}/(1+e^{-2.8050+1.1492*\text{price}})$

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.

$$\begin{aligned}
\hat{Y} = p(y=1 \mid x) &= e^{-2.8050+1.1492*Price} / (1+ e^{-2.8050+1.1492*Price}) \\
&= e^{-2.8050+1.1492*4.0} / (1+ e^{-2.8050+1.1492*4.0}) \\
&= e^{-2.8050+4.5968} / (1+ e^{-2.8050+4.5968}) \\
&= e^{1.7918} / (1+ e^{1.7918}) \\
&= 6.0002 / 7.0002 \\
&= 0.85715
\end{aligned}$$

85.72 percent(0.85715) is The probability of chocolate that has a price per serving of $4.00 having quality of rating very good or excellent.

d. What is the estimate of the odds ratio? What is its interpretation?

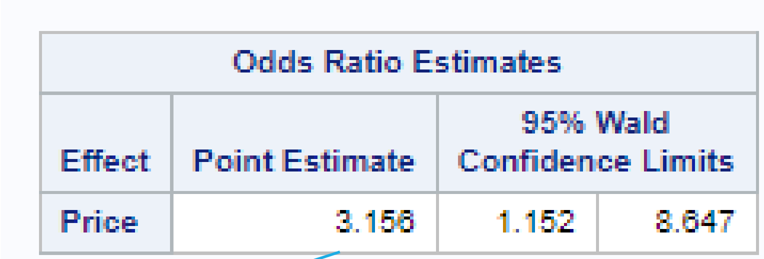
Odds ratio = e^{β_i}

Estimated odds ratio = $e^{\beta 1} = e^{1.1492} = \mathbf{3.156}$

Which gives us **3.156** the same number in the table of odds ratio

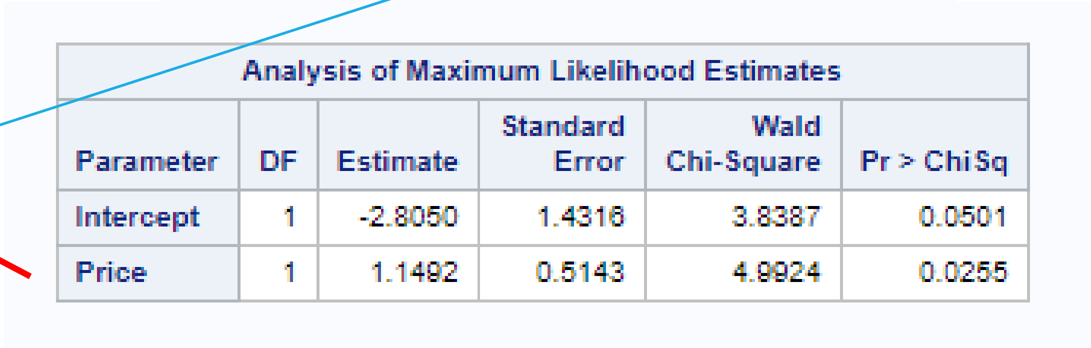
The odds ratio measures the impact on the odds of one-unite increase in only one of the independent variables. In fact, it is the probability of the excellent or very good rate over just one unite increase in price.

The odds ratio is the odds that y given that one of the independent variables has been increased by one unit (odds1) divided by the odds that y given no change in the values for the independent variables (odds0).



Odds Ratio Estimates

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



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

the estimated odds of rating “**Excellent and very good** ” are 3.156 greater than the estimated odds of rating “**Excellent and very good** ” when Price in increased by one unit. the 95% confidence interval is 1.152 to 8.647. An odds ratio greater than 1 implies there are greater odds of the event happening versus the non happening. Because the 95% confidence intervals does not include 1, they are all significant at the .05 level.

Extra Interpretations:

Data set: This is the data source.

Response variable: This is the dependent variable or the Y variable.

Number of response levels: This is the number of levels in the dependent variable (mostly Yes/No); it’s 1 or 0 in this example.

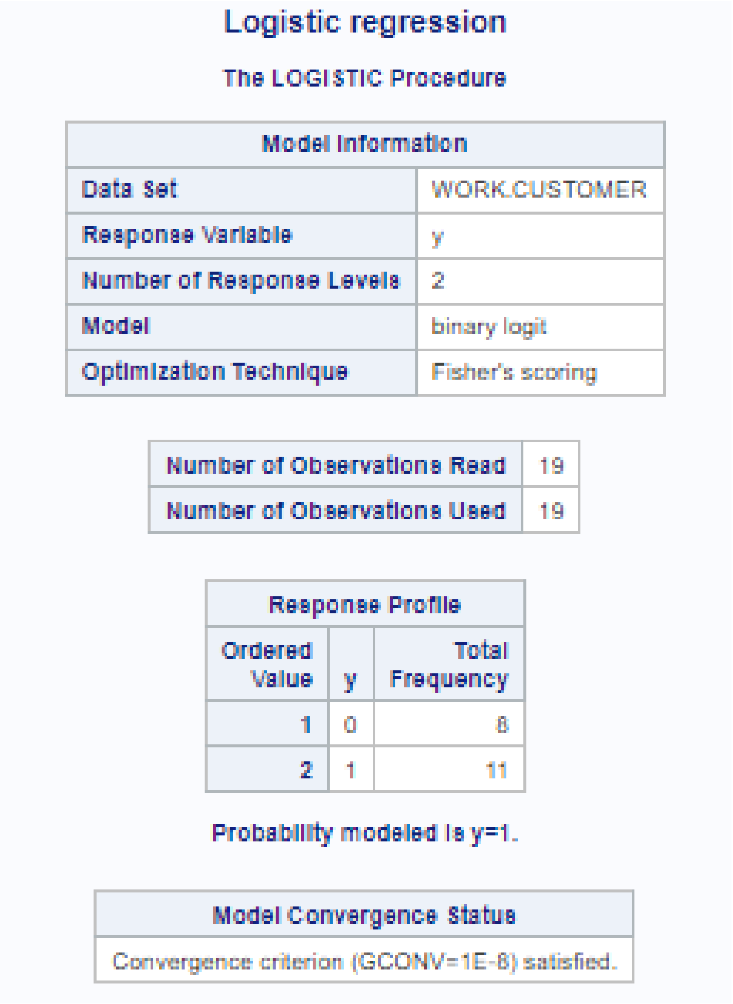
Model: This is the binary logistic regression. it’s the same as binary logit.

Optimization technique: Which optimization technique is used to find the regression coefficients? SAS chooses the most appropriate technique.

Probability modeled is y=‘1’: SAS is informing you that the model is built for y=1. in other words, the output probability will be given for the occurrence of Y being 1

Total frequency: This is the frequency of each category in a dependent variable.

The probability of model is for y=1 showing the frequency of 11



Logistic regression

The LOGISTIC Procedure

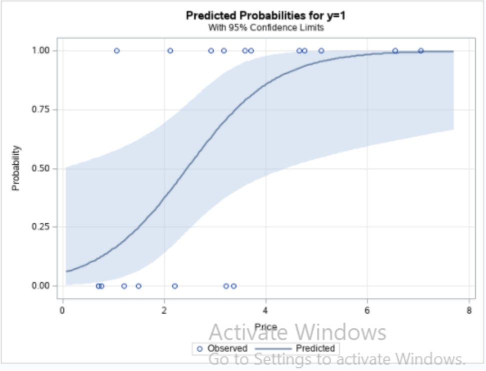
Model Information	
Data Set	WORK.CUSTOMER
Response Variable	y
Number of Response Levels	2
Model	binary logit
Optimization Technique	Fisher's scoring

Number of Observations Read	19
Number of Observations Used	19

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

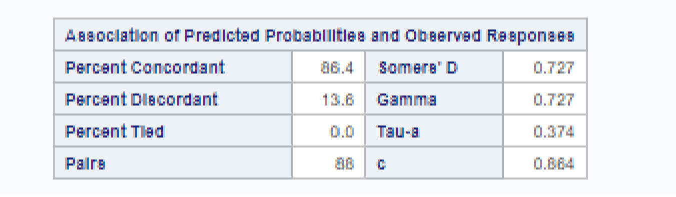
Probability modeled is y=1.

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



Predicted Probabilities for y=1

With 95% Confidence Limits



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

Making Hypothesis and Testing of significant
Null hypothesis : Coefficients of independent variable Price is equal to zero

Alternative hypothesis: Coefficients of independent variable Price is not equal to zero
If in any of the test p value is less than 0.05then Price variable having significant impact on the dependent variable. Here all test has probability less than 0.05, so we reject the Null hypothesis. This means "Price"'s coefficient is not equal to zero

$$H_0: \beta_1 = 0$$

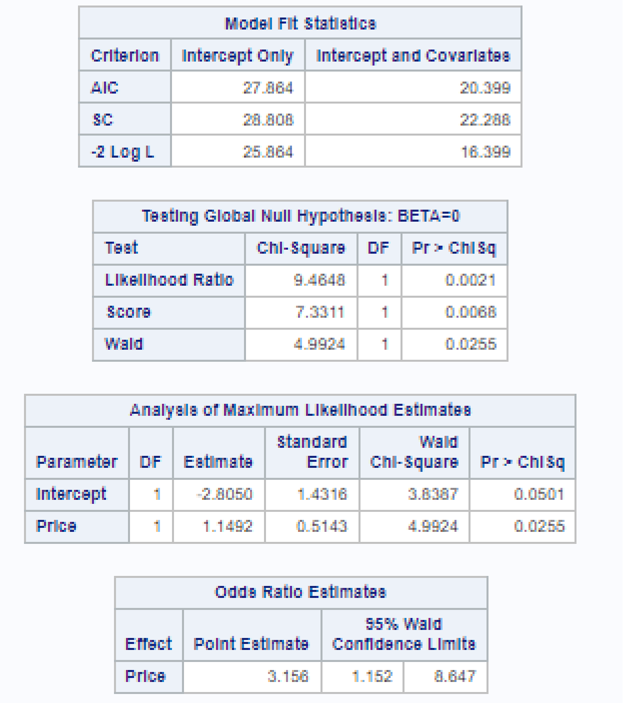
$$H_1 : \beta_1 \neq 0$$

Testing for Significance

AIC and SC are measures that is used to compare two models to pick the best one.

If AIC and SC have less value then we have better the model. As we don't have another model to compare. By just looking at these measure we can say they are low.

Concordant give estimate of accuracy or goodness of fit of logistic regression.so higher percent of Concordant is better.



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

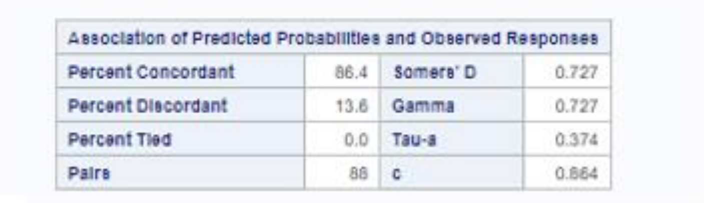
Analysis of Maximum Likelihood Estimates

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

Percent concordant = Percent of right classification = 86.4
Percent discordant = Percent of wrong classification = 13.6
Percent tied = 0.0
Looking at this model accuracy we can say the model is Good



Association of Predicted Probabilities and Observed Responses

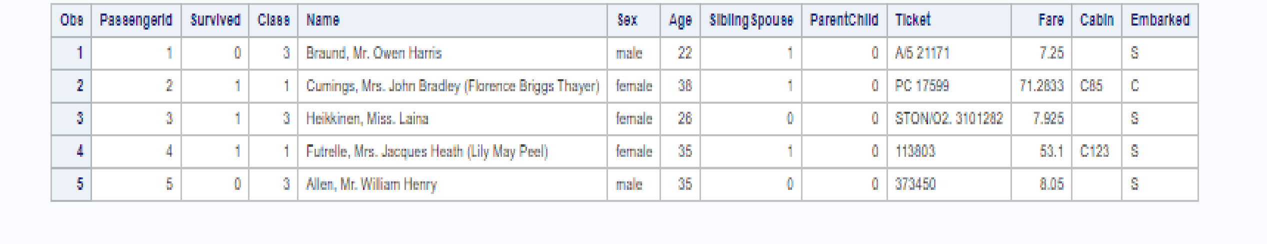
Percent Concordant	86.4	Somers' D	0.727
Percent Discordant	13.6	Gamma	0.727
Percent Tied	0.0	Tau-a	0.374
Pairs	88	c	0.864

$$\hat{Y} = p(y=1 \mid x) = e^{-2.8050+1.1492*Price} / (1+ e^{-2.8050+1.1492*Price})$$

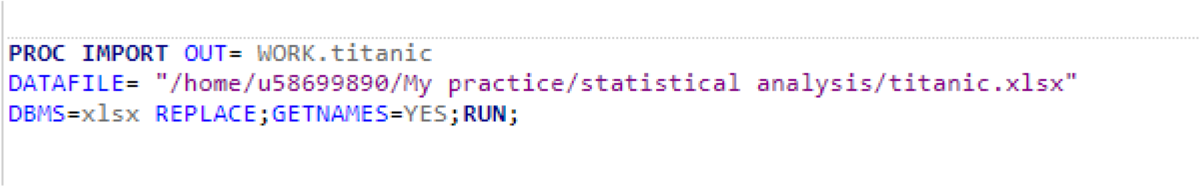
Problem 2 (13 marks) File: Titanic. xlsx

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 d variables)

y=1 if the passenger was survived and y=0 if not



Obs	PassengerId	Survived	Class	Name	Sex	Age	SiblingSpouse	ParentChild	Ticket	Fare	Cabin	Embarked
1	1	0	3	Braund, Mr. Owen Harris	male	22	1	0	A/5 21171	7.25		S
2	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0	PC 17599	71.2833	C85	C
3	3	1	3	Heikkinen, Miss. Laina	female	26	0	0	STON/O2. 3101282	7.925		S
4	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0	113803	53.1	C123	S
5	5	0	3	Allen, Mr. William Henry	male	35	0	0	373450	8.05		S

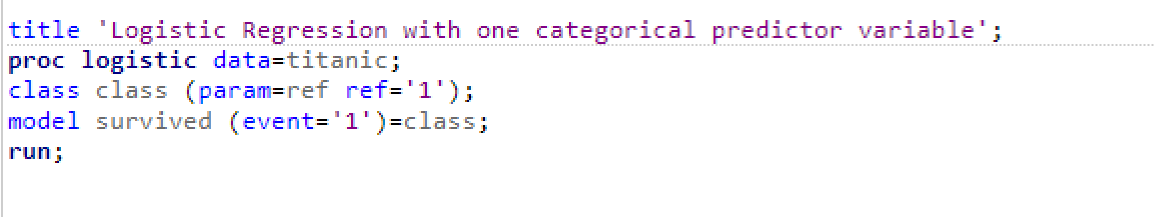


```
PROC IMPORT OUT= WORK.titanic
DATAFILE= "/home/u58699890/My practice/statistical analysis/titanic.xlsx"
DBMS=xlsx REPLACE;GETNAMES=YES;RUN;
```

a. Write the logistic regression equation relating Class and Survived.

$$p(y \mid x) = \frac{e^{\beta 0+\ \beta 1\ x1}}{1+\ e^{\beta 0+\ \beta 1\ x1}}$$

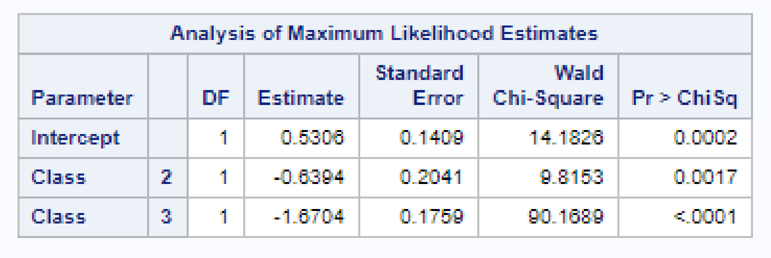
x1 = Class



```
title 'Logistic Regression with one categorical predictor variable';
proc logistic data=titanic;
class class (param=ref ref='1');
model survived (event='1')=class;
run;
```

y=1 is Survived
y=0 Is Not survived

b. For the Titanic data, use SAS to compute the estimated logistic regression equation.



Analysis of Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.5306	0.1409	14.1826	0.0002
Class	2	1	-0.6394	0.2041	9.8153	0.0017
Class	3	1	-1.6704	0.1759	90.1689	<.0001

estimated logistic regression equation.

Ŷ = p(y =1 |x) = $e0.5306-0.6394*ifClass=2-1.6704(ifClass=2)$/ (1+ $e0.5306-0.6394*ifClass=2-1.6704(ifClass=3)$)

Estimated Logit ĝ =$0.5306-0.6394*ifClass=2-1.6704(ifClass=3)$

C. What is the interpretation of E(y) when$X2$= 2? (2 marks)
So, Class =2.

ĝ = $0.5306-0.6394*ifClass=2-1.6704(ifClass=3)$

ĝ =$0.5306-0.6394*1-1.6704(0)$

ĝ =$0.5306-0.6394$

ĝ = -0.1088

Ŷ = p(y =1 |x) = $e0.5306-0.6394*ifClass=2-1.6704(ifClass=2)$/ (1+ $e0.5306-0.6394*ifClass=2-1.6704(ifClass=3)$)

Ŷ = p(y =1 |x) = $e-0.1088$/ (1+ $e-0.1088$)

Ŷ = p(y =1 |x) = 0.8969/ (1+ 0.8969)

Ŷ = p(y =1 |x) = 0.8969/ 1.8969

Ŷ = p(y =1 |x) = 0.4728

E(Y=1 that is survived) = 0.4728 That is 47.28%

E(Y=0 that is not survived) = 100%-47.28% = 0.5272 that is 52.72%

d. Estimate the probability of surviving the 2nd class passengers and the 3rd class passengers. (3 marks)

Class 2

ĝ = $0.5306-0.6394*ifClass=2-1.6704(ifClass=3)$

ĝ =$0.5306-0.6394*1-1.6704(0)$

Ŷ = p(y =1 |x) = $e0.5306-0.6394*ifClass=2-1.6704(ifClass=3)$/ (1+ $e0.5306-0.6394*ifClass=2-1.6704(ifClass=3)$)

Ŷ = p(y =1 |x) = $e-0.1088$/ (1+ $e-0.1088$)

Ŷ = p(y =1 |x) = 0.8969/ 1.8969

Ŷ = p(y =1 |x) = 0.4728

probability of surviving for 2nd class passengers E(Y=1 that is survived) = 0.4728 that is 47.28%

probability of Not surviving for 2nd class passengers E(Y=0 that is not survived) is (100%-47.28% = 52.72%) 0.5272

Class 3

ĝ = $0.5306-0.6394*ifClass=2-1.6704(ifClass=3)$

ĝ =$0.5306-0.6394*0-1.6704(1)$=ĝ = -1.1398

Ŷ = p(y =1 |x) = $e0.5306-0.6394*ifClass=2-1.6704(ifClass=3)$/ (1+ $e0.5306-0.6394*ifClass=2-1.6704(ifClass=3)$)

Ŷ = p(y =1 |x) = $e-1.1398$/ (1+ $e-1.1398$.)

Ŷ = p(y =1 |x) = 0.31988/ 1.31988

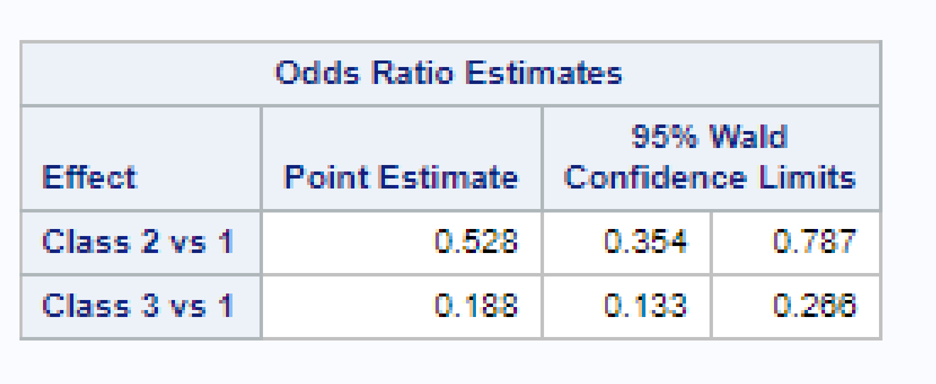
Ŷ = p(y =1 |x) = 0.24245

probability of surviving for 3rd class passengers E(Y=1 that is survived) = 0.24245 that is 24.24%

probability of not surviving for 3rd class passengers E(Y=0 that is not survived) is (100%-24.24% = 75.76%) = 0.7576

E. What is the estimated odds ratio? What is the interpretation?

- Odds Ratio estimate for **Class 2 vs 1 = 0.528**
- People in the Class 2 have 0.528 times the odds of surviving compared to the people of class 1.
- Odds Ratio estimate for **Class 3 vs 1 = 0.188**
- People in the Class 3 have 0.188 times the odds of surviving compared to the people of class 1.
- Because none of the 95% confidence intervals includes 1, they are all significant at the .05 level.



Odds Ratio Estimates

Effect	Point Estimate	95% Wald Confidence Limits	
Class 2 vs 1	0.528	0.354	0.787
Class 3 vs 1	0.188	0.133	0.266

Extra Interpretations:

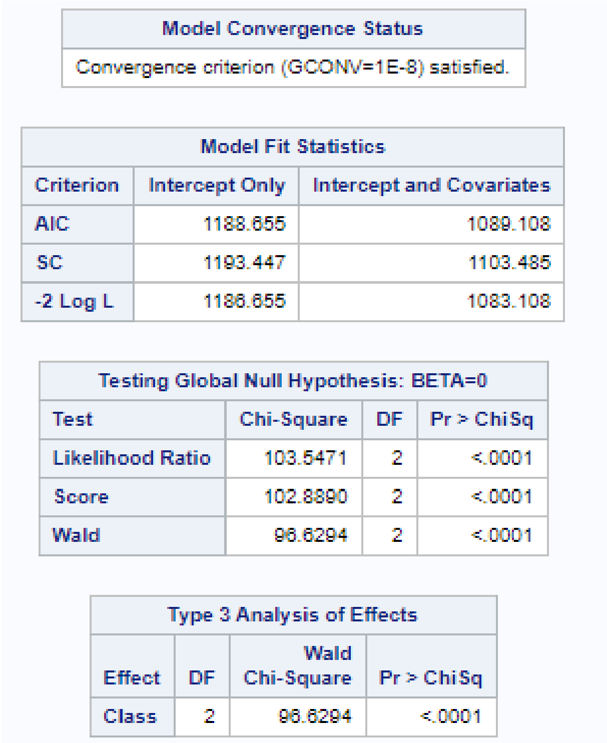
Null hypothesis : Coefficients of all independent variables are equal to zero

Alternative hypothesis: At least one of the confidents is nonzero
If in any of the test p value is less than 0.05 then there is at least one variable having significant impact on the dependent variable.
Here all test has P less than 0.05 we reject the Null hypothesis.
This means there is at least one independent variable whose coefficient is not equal to zero.
Overall, The model is significant.

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

$$H_1 : \beta_1 \neq 0 \text{ or } \beta_2 \neq 0 \text{ or } \beta_3 \neq 0 \text{ or } \ldots \beta_p \neq 0$$

Percent concordant : Percent of right classification = 51.2
•**Percent discordant** : Percent of wrong classification = 14.9
•**Percent tied** :33.9
•model accuracy is not very good, we could add other variable to the logistics regression to improve model Accuracy. As, The questions in the assignment are dependent on only “Class” variable we will continue with this model

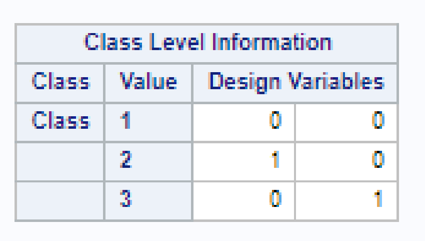


Model Convergence Status
Convergence criterion (GCONV=1E-8) satisfied.

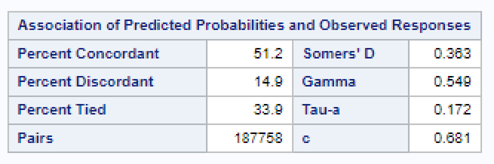
Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	1188.655	1089.108
SC	1193.447	1103.485
-2 Log L	1186.655	1083.108

Testing Global Null Hypothesis: BETA=0			
Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	103.5471	2	<.0001
Score	102.8890	2	<.0001
Wald	96.6294	2	<.0001

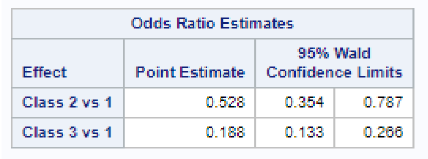
Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Class	2	96.6294	<.0001



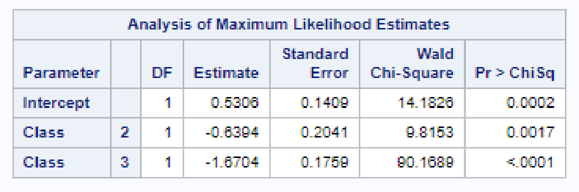
Class Level Information			
Class	Value	Design Variables	
Class	1	0	0
	2	1	0
	3	0	1



Association of Predicted Probabilities and Observed Responses			
Percent Concordant	51.2	Somers' D	0.363
Percent Discordant	14.9	Gamma	0.549
Percent Tied	33.9	Tau-a	0.172
Pairs	187758	c	0.681



Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
Class 2 vs 1	0.528	0.354	0.787
Class 3 vs 1	0.188	0.133	0.266



Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	0.5306	0.1409	14.1826	0.0002
Class	2	1	-0.6394	0.2041	9.8153	0.0017
Class	3	1	-1.6704	0.1759	90.1689	<.0001

Thank You