Problem 1 – 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

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

$$p(y = 1|Price) = \frac{e^{\beta_0+\beta_1 \times Price}}{1 + e^{\beta_0+\beta_1 \times Price}}$$

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

β_0 = -2.8050
β_1 = 1.1492

$$p(y = 0|Price) = \frac{e^{-2.8050+1.1492 \times Price}}{1 + e^{-2.8050+1.1492 \times Price}}$$

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

$$p(y = 1\,|4) = \frac{e^{-2.8050+1.1492\,\times 4}}{1 + e^{-2.8050+1.1492\,\times 4}}$$

$$p(y = 1\,|4) = 0.85714782002$$
$$p(y = 0\,|4) = 0.14285217998$$

$$odds_1 = \frac{p(y = 1\,|4)}{p(y = 0\,|4)}$$

$$odds_1 = 6$$

There is an 85.7% probability, and it is 6 times more likely for a chocolate at the price of $4.00 to have a quality rating of excellent or very good.

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

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

The estimate odds ratio is 3.156 and it means the odds in favor of chocolate having a quality rating of excellent or very good are 3.156 times higher if the price increase by one dollar.

Problem 2 - 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)

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

$$p(Survived = 1|Age) = \frac{e^{\beta_0 + \beta_{Age \times Age}}}{1 + e^{\beta_0 + \beta_{Age \times Age}}}$$

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

First, I ran the regression equation with all variables. From the results I decided to keep only Age, Sex and Class because the rest are not significant.

Type 3 Analysis of Effects			
Effect	DF	Wald Chi-Square	Pr > ChiSq
Age	1	21.2812	<.0001
Sex	1	135.7520	<.0001
Class	2	60.5585	<.0001
SiblingSpouse	5	11.0189	0.0510
ParentChild	6	2.3964	0.8799
Embarked	2	1.8571	0.3951

After running the code, the following estimates were given.

Analysis of Maximum Likelihood Estimates						
Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	3.7692	0.4011	88.3120	<.0001
Age		1	-0.0372	0.00767	23.5197	<.0001
Sex	male	1	-2.5139	0.2076	146.6147	<.0001
Class	2	1	-1.3012	0.2782	21.8811	<.0001
Class	3	1	-2.5724	0.2815	83.5005	<.0001

From that out regression model should be as the following

$$p(Survived = 1|Age) = \frac{e^{3.7692-0.0372\times Age-2.5139\times[male]-1.3014\times[class_2]-2.5724\times[class_2]}}{1+e^{3.7692-0.0372\times Age-2.5139\times[male]-1.3014\times[class_2]-2.5724\times[class_2]}}$$

Estimate the probability of surviving the passenger with the average Age 30.

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

$$p(y = 0|30) = \frac{e^{-0.0567-0.011\times 30}}{1+e^{-0.0567-0.011\times 30}}$$

$$p(y = 0|30) = 0.59548803971$$
$$p(y = 1|30) = 0.40451196029$$

$$odds_1 = \frac{p(y = 1\,|4)}{p(y = 0\,|4)}$$

$$odds_1 = 0.68$$

There is 40.5% probability, and odds are 0.68 times less likely a passenger will survive at the age of 30 than not survive.

c. 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?

$$0.5 = \frac{e^{-0.0567 - 0.011 \times Age}}{1 + e^{-0.0567 - 0.011 \times Age}}$$

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_{Age} \times Age$$

$$ln\left(\frac{0.5}{1-0.5}\right) = -0.0567 - 0.011 \times Age$$

$$Age = -5$$

The average age is equal to 5

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

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

The estimate odds ratio is 0.989 and it means the odds in favor of survival are 0.989 times lower if the age increased by one year.

Problem 3: Capital punishment

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

Model 1: whites 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

$$odds\ ratio = \frac{odds_1}{odds_0}$$

In Model 1 to calculate the odds ratio we use $odds_1$ is for blacks and $odds_0$ is for whites and therefore:

$$odds\ ratio = \frac{odds_{blacks}}{odds_{whites}} = \frac{0.838}{2.472} = 0.338$$

In Model 2 it is the other way around where $odds_1$ is for Whites and $odds_0$ is for blacks

$$odds\ ratio = \frac{odds_{whites}}{odds_{blacks}} = \frac{2.472}{0.838} = 2.949$$

b. Show the relation between the odd ratios and coefficient

The odds ratio can be calculated by the exponential of the coefficient and that is how the relation is defined.

$$odds\ ratio = \ e^{\beta_i}$$

To calculate the odds ratio for changes of more than one unit (for instance c units)

$$coeffecient\ odds\ ratio = \ e^{c\beta_1}$$