11

REGRESSION SPECIAL TOPICS

QUESTION 1.

In the box below write down what you think dummy variables are and why they are used.

Dummy are variables are:
They are used when:

QUESTION 2.

Your Marketing Manager wants to predict which markets for the company's products should be targeted for investment. She has asked you to look at existing markets and provide a yes/no answer for ones that should be developed.

Currently there are 3 markets:

UK:	55% market share
Japan:	15% market share
North America:	30% market share

As part of the team, your job is to decide on which statistical analysis technique would be appropriate, and decide upon the type of variables and give them a unique identifier.

You have decided logistic regression is the best technique to use because the outcome is _____.

You have decided to use variables _____ using the market share percentage for the UK as the model _____. The reason for choosing the UK is _____.

Complete the following table of dummy variables:

Country	x_1	x_2
UK		
Japan		
North America		

QUESTION 3.

A fast food company is proposing to set up two new restaurants in your area. One will be in Lowerbunwich, which has a population of 8000 people and already has two other fast food restaurants, the other in Higherbunwich which has a population of 3000 people and one existing fast food restaurant. The sales department uses the following regression model to predict sales:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

where:

x_1 = number of competitors within the town.

x_2 = population of the town (1000s).

$$x_3 = \begin{cases} 1 & \text{if drive-through window available} \\ 0 & \text{otherwise} \end{cases}$$

The fast food company already has 20 restaurants and based on data from these outlets the following estimated regression equation was developed:

$$\hat{y} = 10.1 - 4.2x_1 + 6.8x_2 + 15.3x_3$$

The Sales Manager wants to know:

1. Should a new restaurant be built in either Lowerbunwich, Higherbunwich or both?
2. The expected sales if a drive-through is added to the restaurants.

Complete the table below:

Restaurant name	Predicted sales no drive-through	Predicted sales with drive through
Lowerbunwhich		
Higherbunwich		

QUESTION 4.

The following equation describes _____.

$$\text{where } \_\_\_\_\_ = \Sigma_{i-1}^{N}(y_i \ln(P(y_i)) + (1 - y_i)\ln(1 - P(y_i)))$$

Complete the following sentence:

The Wald statistic is used to test the _____ of each predictor. In order to assess how much individual predictors contribute to the overall model the following (partially completed) equation is used:

$$R^2 = \pm\sqrt{\left(\frac{?-(2\times ?)}{?}\right)}$$

Write the completed equation in the box below:

QUESTION 5.

You have been given the job by a credit card company to organise a campaign to increase the number of premium card holders. Standard credit card holders have been targeted. At the moment these customers do not pay an annual fee, whereas if they switch to the premium card they will have to pay an annual fee.

The data from 30 standard credit card holders has been used to develop a logistic regression model. The following output is obtained from running the model on a computer:

Predictor	Coefft	SE Coefft	*Z*	*P*	Odds ratio	95% CI lower	95% CI upper
Constant	-6.93984	2.94712	–2.35	0.019			
Purchases	0.139469	0.0680641	2.05	0.040	1.15	1.01	1.31
Extra cards (1)	2.77434	1.19267	2.33	0.020	16.03	1.55	165.99

LL = –10.038

Method	χ^2	*df*	*p*
Pearson	18.5186	27	0.887
Hosmer-Lemeshow	6.5174	8	0.589

Write down an explanation for each of the coefficients:

b_0: _____

_____

b_1: _____

_____

b_2: _____

_____

The estimated logistic regression equation is: _____

Using the goodness-of-fit tests results, is the model a good fit? (yes/no)

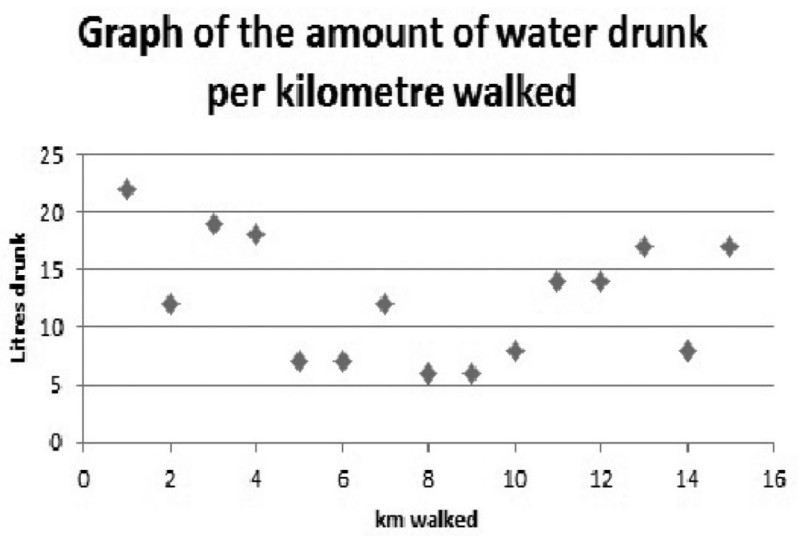
Give your explanation for your decision in the box below.

Roger Spenditall has been a standard credit card holder for 7 years. He, on average, charges £36,000 to his card annually and other family members also have credit cards from the same company.

The estimated probability that Roger will upgrade is: _____.

QUESTION 6.

The graph below shows how much water a hiker drank depending on how far he walked.



How would you classify the relationship between kilometres walked and the amount of water drunk?

linear/non-linear (delete as appropriate).

The equation $\hat{y} = 0.2x^2 - 3.42x + 23.4$ is is linear/quadratic/cubic (delete as appropriate).

MINI PROJECT

You have been asked by the Director of Studies for a university business school to develop a statistical model which will enable her to predict the success of students based on two independent variables. The students enter the programme after completing an undergraduate degree or equivalent programme. The grades from this qualification are scored from 1 to 5 with 5 being the highest grade. They also sit an entrance exam which is scored from 0 to 700.

The table below shows the results of a random sample of 30 students who have been enrolled on the programme. A 0 indicates the student did not successfully complete the programme and a 1 indicates a student who did.

Successful completion	Entry grade	Entrance exam	Successful completion	Entry grade	Entrance exam
0	2.93	617	1	3.17	639
0	3.05	557	1	3.24	632
0	3.11	599	1	3.41	639
0	3.24	616	1	3.37	619
0	3.36	594	1	3.46	665
0	3.41	567	1	3.57	694
0	3.45	542	1	3.62	641
0	3.60	551	1	3.66	594
0	3.64	573	1	3.69	678
0	3.57	536	1	3.70	624
1	2.75	688	1	3.78	654
1	2.81	647	1	3.84	718
1	3.03	652	1	3.77	692
1	3.10	608	1	3.79	632
1	3.06	680	1	3.97	784

The Director of Studies wants you to produce a report which gives the details and an explanation of your model. You need to explain the meaning of the regression coefficients and the level of significance used. You will need to justify the use of two predictors by developing similar models using single predictors. Your recommendations should compare and discuss the predictive power of each of the models.

And finally...

'To be or not to be', Shakespeare's version of logistic regression.