**330 Assignment 1 Answers**

**Question 1**

*Problem 1- height of 15.1m*

Looking at the height histogram in the pairs plot shows a single outlier (height = 15.1 metres).This height is not sensible and I propose the decimal point was inputted in the wrong place. However, because I do not know how the data entry occurred I have decided to change the outlier value to NA.

*Problem 2 – weight of 13.4kg*

From the pairs plot you can see that weight has a single outlier (weight = 13.4kg). The subject is a European male of height 1.75m, therefore a weight of 13.4kg is infeasible. I believe this value is an error and have replaced it with NA.

*Problem 3 – dbp of 163 mm Hg*

Looking at the histogram of dbp theres a clear outlier of 163 mm Hg. I believe the dbp value of the participant was incorrectly entered during the data collection. However, because I cannot confirm how the data was collected nor am I a healthcare expert (no knowledge about reasonable dbp values and the value could relate to medical condition the participant had) I have elected to retain the value.

**Question 2**

a) Firstly did some initial data exploration creating a box plot showing ethnicity versus number of eggs consumed per week. The plot showed a clear outlier where a Maori participant typically consumed 60 eggs a week. I have decided to keep this outlier as even though it is quite large it is definitely within the realm of possibility.

Furthermore a pairs plot was created for exploratory analysis. The pairs plot showed that the typical number of eggs consumed per week is right skewed. It also showed that many their were many more european participants than other ethnicities.

b) The response variable in the model fitted is the typical number of eggs consumed per week and the explanatory variable was the ethnicity of participants. The response variable is highly right skewed. Further is a count and therefore a Poisson regression model was fitted.

We were interested to answer the question whether the typical number of eggs consumed per week varied according to a participants ethnicity.

We have strong evidence to suggest that the typical number of eggs consumed per week does vary by ethnicity.

We have strong evidence that the expected number of eggs consumed per week by Maori participants is higher than European participants. We estimate that the expected number of eggs consumed per week by a Maori participant is 222.9% to 394.2% higher than an European participant.

We have strong evidence that the expected number of eggs consumed per week by a Polynesian participant is higher than an European participant. We estimate that the expected number of eggs consumed per week by a Polynesian participant is 12.3% to 151.2% higher than European participants.

We have no evidence for a difference in the expected number of eggs consumed per week between European participants and Other ethnicity participants.

**Question 3**

a) The response variable for this analysis is the maximum number of alcoholic drinks a person consumes in a day (within the last 3 months) and the explanatory variables are the person’s age and sex.

b) The response variable (*drinkmaxday*) is a count and thus cannot be negative. Furthermore, as can be seen from the pairs plot it is right skewed and thus I have choosen to use Poisson regression.

There is a outlier in the response variable of 50 drinks consumed in a day (within the last 3 months). However, I have decided to keep this outlier in my analysis as it could be feasible because we do not have any additional information on key factors such as size or alcohol type of the drinks consumed.

c) We were interested assessing whether there is evidence to suggest that and/or gender are related to the maximum number of alcoholic drinks consumed per day (within the last 3 months).

We have strong evidence to suggest that age and gender are related to the maximum number of drinks consumed per day (within the last 3 months). However, there is no evidence that the relationship between age and the maximum number of drinks consumed per day (within the last 3 months) depended on gender (and vice versa).

We estimate that for each additional year of age the expected number of the maximum number of alcoholic drinks consumed per day (within the last 3 months) decreases by 1.5% to 2.6%, regardless of gender.

We estimate that, for the same age, the expected number of of the maximum number of alcoholic drinks consumed per day (within the last 3 months) for male is 30.3% to 85.6% higher than a female.

d)

i)

ii) The plot show us the expected value of maximum number of alcoholic drinks consumed per day (within the last 3 months) is higher for males than females for the same age. Further, the plot shows that as a person gets older the maximum number of alcoholic drinks consumed per day (within the last 3 months) is decreases exponentially for both males and females. This decreasing relationship is the same for both males and females. The plot also shows us a clear outlier who is male, around 33 years old and has drunk a maximum of 50 alcoholic drinks per day (within the last 3 months) .

e) We estimate that the expected value of maximum number of alcoholic drinks consumed per day (within the last 3 months) for a 30 year old male is 8.4 drinks. In comparision we estimate that the expected value of maximum number of alcoholic drinks consumed per day (within the last 3 months) for a 50 year old male is 5.6 drinks.

We estimate the that the expected value of maximum number of alcoholic drinks consumed per day (within the last 3 months) for a 40 year old male is 6.9 drinks. In comparision we estimate that the expected value of maximum number of alcoholic drinks consumed per day (within the last 3 months) for a 40 year old female is 4.4 drinks.

**Question 4**

a) In this situation the explanatory variables are age of the participant and smoking history of the participant. The response variable is typical number of exercise hours per week.

This variable is a combination of *exermin*, *exerhour* and *exerday* variables. The method of calculation is shown below:

b) The response variable is numeric measure and thus linear regression has been fitted. One limitation of the modelling is that the response variable cannot be negative but the model doesn’t have this constraint.

c) We were interested in whether there is evidence that age and/or smoking history are related to the typical number of hours of exercise undertaken in a week.

We have evidence to suggest that age and smoking history are related to the typical number of hours of exercise undertaken in a week. However, we have no evidence that the relationship between age and the number of hours of exercise undertaken in a week depended on smoking history (and vice versa).

We estimate that for each additional year of age, the expected number of hours exercised per week decreases by between 0.04 hours and 0.25 hours, regardless of smoking history.

We estimate that the expected number of hours exercised per week for people who have smoked once a week or more in the past, is between 0.8 to 7.2 hours greater than people with no smoking history, for the same age.

d)

i) Plotted

ii) The plot show us the expected value of response variable is higher for participants with a smoking history than participants without a smoking history, for the same age. Further, the plot shows that as a person gets older, the expected number of hours exercised per week decreases for both smokers and non-smoking participants. The plot also shows the rate of decrease is the same for non-smokers and smokers.

The plot shows that there are a few large outliers, which are predominately participants with a smoking history.

The plot also shows that the model can predict negative hours exercised per week under certain conditions. For example, if a participant has no smoking history and is greater than 78 years old, our model will predict negative hours exercised per week.

**Question 5**

a) The response variable for this analysis is whether participants have had heart attacks and the explanatory variables are age and cholesterol level of the participant.

b) The response variable is a categorical variable with two levels, thus it is binary. As a result of having a binary response variable a logistic regression model has been fit.

c) We were interested in whether there is evidence to suggest that age and/or cholestrol level status have an effect on whether someone had a heart-attack or not.

We have strong evidence that age had an effect on whether a person had a heart-attack. However, we had no evidence that this relationship depended on cholestrol level status.

Likewise we had strong evidence that cholestrol level status had an effect on whether a person had a heart-attack. However, we have no evidence that this relationship depended on age.

We estimate that for every additional year of age, the odds of having a heart-attack increase between 4.4% to 19.7%, regardless of cholestrol level status.

We estimate that the odds of people who have low cholestrol having a heart-attack are 7.2% to 98.2% lower than people who have high cholestrol, for the same age.

d)

i) Plotted

ii) The plot shows that the probability of having a heart-attack for people with high cholestrol is higher than people with low cholestrol at the same age. Further, the plot shows that typicaly as a person gets older, the probability of having a heart-attack increases, regardless of cholestrol level.

The plot also highlights that there are very few participants in the study (6 observations out of 120 in the sample) that had heartattacks.