# STATS 330 Mid-semester Test

Hasnain Cheena, 190411106

Due Date: 5pm, Thursday 7th May 2020

## Question 1

For the 24 hour duration of this test, I confirm that I will not discuss the content of the test with anyone else. I will not give any assistance to another student taking this test. I will not receive any assistance from any person or tutoring service.

Hasnain Cheena

## Question 2

a)

$$log(\mu_i) = \beta_0 + \beta_1 \times age_i + \beta_2 \times (age)^2 + \beta_3 \times smoke.Yes_i + log(years_i)$$

$$Y_i \sim Poisson(\mu_i)$$

Where $smoke.Yes_i = 1$ if the ith doctor replied yes, he/she smokes and 0 if the ith doctor replied no he/she does not smoke. Furthermore, $age_i$ is the mid-point of a given age group in years and $log(years_i)$ is the log transformed total number of person-years for a given group. Moreover, $Y_i$ follows a Poisson distribution with mean of $\mu_i$ and represents the number of deaths in the ith age group.

b)

$$logit(p_i) = \beta_0 + \beta_1 \times age_i + \beta_2 \times (age)^2 + \beta_3 \times smoke.Yes_i$$

$$Y_i \sim Binomial(n_i, p_i)$$

Where $smoke.Yes_i = 1$ if the ith doctor replied yes, he/she smokes and 0 if the ith doctor replied no he/she does not smoke. Furthermore, $age_i$ is the mid-point of a given age group in years. Moreover, $Y_i$ follows a Binomial distribution with number of trials $n_i$ and probability of success of $p_i$.

c) In $poisson.fit$ the variable $years$ is being used as an offset. This is because we know the effect of years on the response variable, deaths and as such do not need to estimate it. Generally, as someone gets older, their chance of dying increases. This effect is displayed as years decreases and the number of deaths increases. The offset is incorporated into the model by fitting the log of years.

## Question 3

a)

$$logit(p_i) = \beta_0 + \beta_1 \times weight_i + \beta_2 \times weight_i + \beta_3 \times sex.male_i$$
$$Y_i \sim Binomial(n_i, p_i)$$

where $sex.male_i = 1$ if the insect is a male and 0 if the insect is a female. Moreover, $age_i$ is the age in days of the ith insect and $weight_i$ is the weight in grams of the ith insect. Furthermore, $Y_i$ follows a Binomial distribution with number of trials $n_i$ and probability of success of $p_i$.

b) We have evidence ($p - value \approx 0.02$) to suggest that the sex of an insect had an effect on whether it became infected.
We estimate that the odds of insects that are male becoming infected is 21.6% to 94.9% less than insects that are female, for the same age and weight.

c) Looking purely at the GAM plots I agree with her decision to fit age with a quadratic effect but not weight. This is because looking at the GAM plot for weight, a straight line can easily be fitted in between the dashed lines. Therefore it was sensible not to fit a quadratic term for weight at this stage. In comparison, looking at the GAM plot for age, while a straight line can be fit between the dashed lines on the GAM plot, there is a reasonable amount of curvature in the plot hinting at a non-linear (quadratic) relationship between age and the response.

d) I would recommend $insects.quad$ over $insects1.fit$. This is because even though both models are appropriate (randomised quantile residual plots are patternless bands around 0) and good fits (the deviance provides evidence suggesting the models are good fits) the quadratic term in $insects.quad$ is significant and cannot be ignored. This implies a non-linear relationship between the response and age which is captured by $insects.quad$ and not captured by $insects1.fit$ . Furthermore, the deviance of $insects.quad$ is lower than $insects1.fit$ implying it is a better fit.

e) The null hypothesis being tested is that there is no interaction between any of the explanatory variables (age, weight or sex).

f) We have no evidence (p-value ≈ 0.39) to suggest that there are any interactions between age, weight and/or sex.

## Question 4

a)

i)

$$\hat{V}ar(Y_{10}) = n_i\hat{p}_i(1 - \hat{p}_i)$$

$$\hat{V}ar(Y_{10}) = 10 \times 0.7555038 \times (1 - 0.0.7555038)$$

$$\hat{V}ar(Y_{10}) = 1.85$$

ii)

$$\hat{V}ar(Y_{10}) = \hat{k} \times n_i\hat{p}_i(1 - \hat{p}_i)$$

$$\hat{V}ar(Y_{10}) = 2.991942 \times 10 \times 0.7555038 \times (1 - 0.0.7555038)$$

$$\hat{V}ar(Y_{10}) = 5.53$$

iii)

$$\hat{V}ar(Y_{10}) = n_i\hat{p}_i(1 - \hat{p}_i) \times (1 + \hat{\rho}(n_i - 1))$$

$$\hat{V}ar(Y_{10}) = (10 \times 0.7555038 \times (1 - 0.0.7555038)) \times (1 + 0.2365657 \times (10 - 1))$$

$$\hat{V}ar(Y_{10}) = 5.78$$

b)

i) Residuals for $binom.fit$

$$r_{raw} = y_{10} - \widehat{y_{10}}$$
$$r_{raw} = 0.7 - 0.7555038$$
$$r_{raw} = -0.06$$

$$r_{pearson} = \frac{y_{10} - \widehat{y_{10}}}{\sqrt{\hat{V}ar(Y_{10})}}$$

$$r_{pearson} = \frac{0.7 - 0.7555038}{\sqrt{1.85}}$$

$$r_{pearson} = -0.04$$

ii) Residuals for $bbinom.fit$

$$r_{raw} = y_{10} - \widehat{y_{10}}$$

$$r_{raw} = 0.7 - 0.7886691$$

$$r_{raw} = -0.09$$

$$r_{pearson} = \frac{y_{10} - \widehat{y_{10}}}{\sqrt{\hat{V}ar(Y_{10})}}$$

$$r_{pearson} = \frac{0.7 - 0.7886691}{\sqrt{5.78}}$$

$$r_{pearson} = -0.04$$

c) I would plot and examine the pearson, deviance and randomised quantile residual plots of $binom.fit$ to assess if the model is appropriate. If the residual plots are fine, I would then use the deviance to check the goodness of fit. If $binom.fit$ was a good fit ($p - value \geq 0.05$) I would then use information-theory approaches (AIC, AICc and/or BIC) to compare both models. If $binom.fit$ showed lower AIC, AICc and/or BIC scores (and all other conditions above were met) I would then consider $binom.fit$ a suitable replacement for $bbinom.fit$.