University of St Andrews

# DECEMBER 2020 EXAMINATION DIET

# SCHOOL OF MATHEMATICS & STATISTICS

**MODULE CODE:** MT5761 SOME ANSWERS

**MODULE TITLE:** Applied Statistical Modelling using GLMs

**EXAM DURATION:** 2 hours

**EXAM INSTRUCTIONS:** Attempt ALL questions.

The number in square brackets shows the maximum marks obtainable for that question or part-question.

Your answers should contain the full working required to justify your solutions.

## INSTRUCTIONS FOR ONLINE EXAMS:

Each page of your solution must have the page number, module code, and your student ID number at the top of the page. You must make sure all pages of your solutions are clearly legible.

**1.** A real estate agent is interested in predicting the sale price of houses in the city of Windsor, Canada. They have access to a sample of 546 house prices (variable `price`, in Canadian dollars), and the following explanatory variables:

- `lotsize` - the lot size of a property in square feet
- `bedrooms` - the number of bedrooms and
- `prefarea` - a binary variable taking the value 1 if the property is located in a preferred neighbourhood area and 0 otherwise.

An initial model `model1` is fitted using `lotsize`, `bedrooms` and `prefarea` as main effects as well as an interaction term between `bedrooms` and `prefarea`.

**Use the output on pages 3 - 4 to answer Question 1.**

(a) Write out the equation for `model1` and explicitly state the error distribution. **[1]**

Bookwork.

(b) Interpret the coefficient of the interaction term and thus explain how the interaction term affects the relationship between `price` and `bedrooms`. **[3]**

Requires understanding but seen similar.

(c) Additionally, a Generalised Least Squares (GLS) model `model2` was fitted to the data. Which model (`model1` or `model2`) would you prefer? Justify your answer using the provided output. **[2]**

Model 2 (the GLS model) should be preferred.

(d) For the GLS model give the estimated standard error of the error distribution for a fitted value $\hat{y} = 100,000$ (a hundred thousand). Show your working. **[3]**

The estimated standard error of the error term is 33502.32

(e) State one formal test and one graphical tool to check the independence assumption of the observations in a linear model. **[1]**

Bookwork.

Linear Regression output:

```
> model1<-lm(price ~ lotsize + bedrooms+ prefarea + bedrooms:prefarea,
+                        data=Housing)
> summary(model1)

Call:
lm(formula = price ~ lotsize + bedrooms + prefarea + bedrooms:prefarea,
    data = Housing)

Residuals:
   Min     1Q Median     3Q    Max
-58068 -13202  -2444   9375  83261

Coefficients:
                       Estimate Std. Error t value Pr(>|t|)
(Intercept)           9655.5395  4228.6758   2.283   0.0228 *
lotsize                  5.4722     0.4211  12.995  < 2e-16 ***
bedrooms              9193.4876  1302.0455   7.061 5.11e-12 ***
prefareayes          -9612.8808 10604.7071  -0.906   0.3651
bedrooms:prefareayes  7329.6064  3407.4240   2.151   0.0319 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20520 on 541 degrees of freedom
Multiple R-squared:  0.4138,        Adjusted R-squared:  0.4095
F-statistic: 95.48 on 4 and 541 DF,  p-value: < 2.2e-16

> AIC(model1)
[1] 12399.07
> ncvTest(model1)
Non-constant Variance Score Test
Variance formula: ~ fitted.values
Chisquare = 99.83502, Df = 1, p = < 2.22e-16
```

GLS model output:

```
> model2<-gls(price ~ lotsize + bedrooms+ prefarea + bedrooms:prefarea,
+                                           data=Housing, weights=varExp())
> summary(model2)
Generalized least squares fit by REML
  Model: price ~ lotsize + bedrooms + prefarea + bedrooms:prefarea
  Data: Housing
       AIC      BIC    logLik
  12219.72 12249.78 -6102.861

Variance function:
 Structure: Exponential of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
       expon
1.897348e-05

Coefficients:
                        Value Std.Error   t-value p-value
(Intercept)          3701.784  3427.020  1.080176  0.2805
lotsize                 6.918     0.480 14.420775  0.0000
bedrooms             9047.399  1071.357  8.444806  0.0000
prefareayes          4251.227 10757.265  0.395196  0.6929
bedrooms:prefareayes 2024.573  3678.954  0.550312  0.5823

 Correlation:
                     (Intr) lotsiz bedrms prfrys
lotsize              -0.494
bedrooms             -0.791 -0.093
prefareayes          -0.254  0.027  0.264
bedrooms:prefareayes  0.269 -0.052 -0.284 -0.980

Standardized residuals:
       Min         Q1        Med         Q3        Max
-2.2389134 -0.6810988 -0.1380312  0.5641746  4.0411725

Residual standard error: 5024.202
Degrees of freedom: 546 total; 541 residual
> AIC(model2)
[1] 12219.72
```

**2.** An ecologist compared the count of birds at a series of sites in two areas either side of a stockproof fence. One side had limited grazing (mainly from native herbivores), and the other was heavily grazed by feral herbivores, mostly horses. Bird counts were recorded at the sites either side of the fence (the "before" measurements). Then both the feral and the native herbivores were removed, and bird counts recorded again (the "after" measurements).

The dependent variable `Birds` is the total number of birds observed in three 20-min surveys of 2 hectares quadrats. Explanatory variables are:

- `When` - a factor variable with two levels: `Before` for the before measurements and `After` for the after measurements
- `Grazed` - another factor variable with two levels: `Feral` for the side with feral herbivores and `Reference` for the side with the native herbivores

A generalised linear model was fitted to the data as shown below:

```
> fit1<-glm(Birds~When*Grazed, data=grazing, family=poisson)
> summary(fit1)
Call:
glm(formula = Birds ~ When * Grazed, family = poisson, data = grazing)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.8053  -2.4524  -1.0198   0.2336   8.9360

Coefficients:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                 1.85630    0.08839  21.002  < 2e-16 ***
WhenBefore                 -0.67764    0.15231  -4.449 8.62e-06 ***
GrazedReference             0.44629    0.13001   3.433 0.000598 ***
WhenBefore:GrazedReference  0.82135    0.20040   4.098 4.16e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 528.07  on 61  degrees of freedom
Residual deviance: 437.23  on 58  degrees of freedom
AIC: 624.66
Number of Fisher Scoring iterations: 6
```

(a) Which are the reference levels in the fitted model `fit1` for the two factor variables? **[1]**

For `When` it is `After`, for `Grazed` it is `Feral`.

(b) Write down the equation for the expected number of birds in terms of the model coefficients and state what each of the coefficients in the equation represents. **[4]**

$$E(\text{count}) = \exp\left(\beta_0 + \beta_B x_B + \beta_R x_R + \beta_{B*R}(x_B * x_R)\right)$$

- $\beta_0$ gives the log expected counts for birds after measurement on the side with feral herbivores.
- $\beta_B$ gives the change to the log expected count of birds for the "before" measurements compared to the "after" measurements.
- $\beta_R$
- $\beta_{B*R}$

(c) Calculate the expected counts of birds on the side with native herbivores before measurements. **[1]**

The expected count is 11.54, which is around 12.

(d) A model with overdispersion is fitted to these data. The resultant estimate for the dispersion parameter $\phi$ is 9.75, i.e. $\hat{\phi} = 9.75$, indicating overdispersion. Under the **overdispersed model** calculate a 95% confidence interval for the number of birds on the side with feral herbivores after measurements. State any assumptions you make and show your working.
*Hint 1: Accounting for overdispersion does not affect the estimates of the coefficients in the model equation.*
*Hint 2: The z-multiplier is 1.96.* **[4]**

The confidence interval for the counts is:

$$[3.726018; 10.99301]$$

**3.** (a) Let $y_1, y_2, ..., y_n$ be $n$ realisations of independent and identically distributed binary random variables with probability of "success" $p$. The probability mass function of a binary random variable $Y_i$, $i = 1, .., n$, is given as

$$P(Y_i = y_i) = p^{y_i}(1-p)^{1-y_i}, \qquad y_i \in \{0, 1\}$$

(i) Give the likelihood $L(p)$ and the log likelihood $\ln L(p)$ for this sample.

Bookwork.

(ii) Show that the Maximum Likelihood estimator for $p$ is given by

$$\hat{p} = \frac{\sum_{i=1}^n y_i}{n}$$

Taking the first derivative and setting it to zero:

$$\frac{\partial \ln L(p)}{\partial p} = \sum_{i=1}^n \left( \frac{y_i}{p} - \frac{1-y_i}{1-p_i} \right) = \frac{1}{p}\sum_{i=1}^n y_i - \frac{n}{1-p} + \frac{1}{1-p}\sum_{i=1}^n y_i = 0$$

Rearranging:

$$(1-p)\sum_{i=1}^n y_i - np + p\sum_{i=1}^n y_i = 0$$

Thus

$$\sum_{i=1}^n y_i - np = 0, \qquad \Leftrightarrow \qquad \hat{p} = \frac{\sum_{i=1}^n y_i}{n}$$

**[3]**

(b) A study of the habitats of noisy miners (a small but aggressive native Australian bird) recorded whether noisy miners were detected in $n = 31$ transects in woodland patches. The response variable `Miners` takes the value 1 if noisy miners were present and 0 otherwise. An ecologist wants to study whether the presence of noisy miners is impacted by whether or not the number of eucalypts for each studied transect exceeds 15. For that purpose the researcher considers the explanatory variable `Eucs15` taking the value 1 if there were more than 15 eucalypts and 0 otherwise. A GLM model was fitted to the data leading to the following (truncated) output:

```
> binGLM<-glm(Miners~Eucs15, data=nminer, family = binomial)
> summary(binGLM)
Call:
glm(formula = Miners ~ Eucs15, family = binomial, data = nminer)

Deviance Residuals:
     Min         1Q     Median         3Q        Max
-0.84460   -0.84460   0.00008    0.00008    1.55176

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.8473     0.4880  -1.736   0.0825 .
Eucs15TRUE    20.4134  3242.4569   0.006   0.9950
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(i) Write out the link function for the `binGLM` model providing as much detail as possible. **[1]**

Bookwork.

(ii) Interpret the coefficient for `Eucs15TRUE`. **[2]**

The change will lead to an increase in the odds for detecting the noisy miners.

(iii) Comment on the statistical significance of the explanatory variable using the provided output. **[1]**

Bookwork.

(iv) The following code is used to calculate a confusion matrix for the `binGLM` model. Explain what the first two lines of the code do and calculate the predictive accuracy of the fitted model. **[3]**

```
> val  <- mean(fitted(binGLM))
> resp <- ifelse(fitted(binGLM)>val,1,0)
> table(resp,nminer$Miners)

resp  0  1
```

```
0 14  6
1  0 11
```

The accuracy of the fitted model is

$$0.8064516$$

(v) Take a critical view of your results from parts (iii) and (iv) in light of the goodness-of-fit of `binGLM`. **[2]**

The point of this question is for the students to notice that something went wrong. In that particular case the binary explanatory variable taking the value 1 leads to perfect prediction - the Hauck–Donner effect. This has not been discussed in the lecture but I expect comments on the mixed messages one obtains for the model fit and the prediction.

**4.** A data set contains the results of chemical analysis on 178 wines grown in a specific area in Italy. The dependent variable, `Type`, is a nominal categorical variable, representing three types of wine based on three different types of grapes (Type 1, Type 2 and Type 3). Three explanatory variables are considered:

- `Alcohol` - a continuous variable giving the % of alcohol for each wine
- `Color` - a continuous variable giving the intensity of the colour of each wine
- `Magnesium` - a continuous variable giving the quantity of magnesium

The following multinomial logit model is fitted in R:

```
mult<-multinom(Type ~ Alcohol + Color + Magnesium, data=wine)
```

(a) Are the data aggregated or disaggregated? Justify your answer. **[1]**

The data are disaggregated.

(b) State the mathematical formulation for the model being fitted, assuming that the first category (Type 1) is the baseline. Include the distribution of the dependent variable, the link function(s) and the total number of parameters to estimate. **[3]**

The total number of parameters to estimate is 8.

(c) State the model assumptions and **briefly** discuss ideas for checking their validity. **[4]**

Bookwork.

**5.** A survey on marijuana usage was conducted on a group of 11-17 year olds (116 male, 120 female), over a period of 5 consecutive years. The response `potuse` has three levels: 1 = never used, 2 = used no more than once a month and 3 = used more than once a month. There are two explanatory variables:

- `time` - an integer taking the values from 1 to 5 for each year the individuals were asked about their marijuana usage.
- `sex` - a binary variable taking the value 1 if the individual is female and 0 if the individual is male.

A proportional odds model was fitted to this data.

**The R output for Question 5. is given on pages 13 - 14.**

(a) Using the `fitpropodds` model, calculate the probabilities for each marijuana usage type for a male individual in the first year of the survey. **[3]**

The probabilities for each category are:

$$\hat{P}(Y = 1) = 0.8581063$$
$$\hat{P}(Y = 2) = 0.08928148$$
$$\hat{P}(Y = 3) = 0.05261222$$

(b) If the variable `time` is turned into a factor and everything else is kept the same, how many parameters will the modified proportional odds model have? Justify your answer. **[2]**

The number of parameters in the modified model will be 7.

(c) Considering only the survey design, i.e. without running any diagnostics, which model assumption is unlikely to be met? Justify your answer. **[2]**

The independence assumption is likely to be violated.

(d) Under the condition that the model assumptions are valid, describe the trend over time and the difference between sexes using the plots on page 14. Suggest possible explanations. **[3]**

I would accept any sensible interpretation. For full points the main trends should be identified and an attempt for their explanation should be made.

```
> fitpropodds<-vglm(potuse~sex+time, data=data2, family = propodds)
> summary(fitpropodds)

Call:
vglm(formula = potuse ~ sex + time, family = propodds, data = data2)

Pearson residuals:
                      Min      1Q  Median      3Q   Max
logitlink(P[Y>=2]) -1.013 -0.5748 -0.3778  0.4644 4.090
logitlink(P[Y>=3]) -1.146 -0.2837 -0.1878 -0.1410 5.639

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -2.28302    0.19550 -11.678  < 2e-16 ***
(Intercept):2 -3.37413    0.21112 -15.982  < 2e-16 ***
sex           -0.62303    0.13707  -4.545 5.49e-06 ***
time           0.48337    0.05131   9.420  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Names of linear predictors: logitlink(P[Y>=2]), logitlink(P[Y>=3])

Residual deviance: 1667.795 on 2356 degrees of freedom

Log-likelihood: -833.8973 on 2356 degrees of freedom

Number of Fisher scoring iterations: 4

No Hauck-Donner effect found in any of the estimates


Exponentiated coefficients:
      sex      time
0.5363146 1.6215348
```

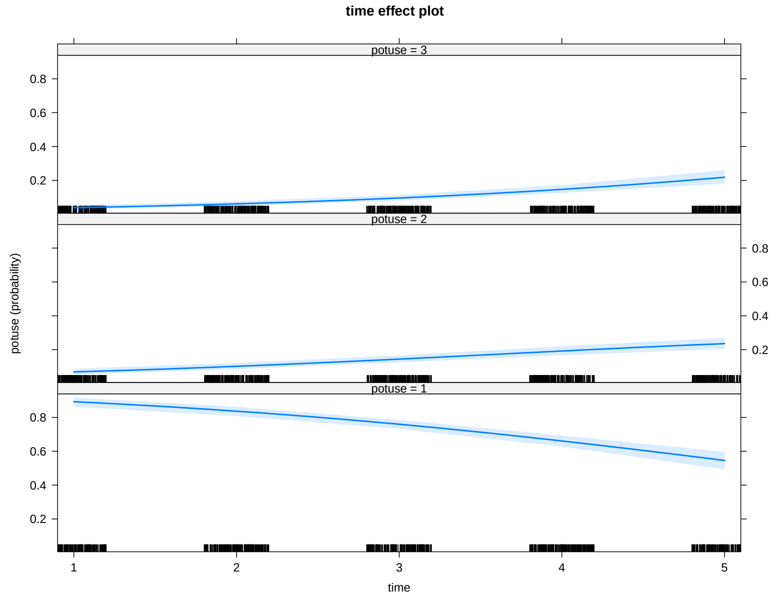


Figure 1: Effects for the variable `week` in the proportional odds model

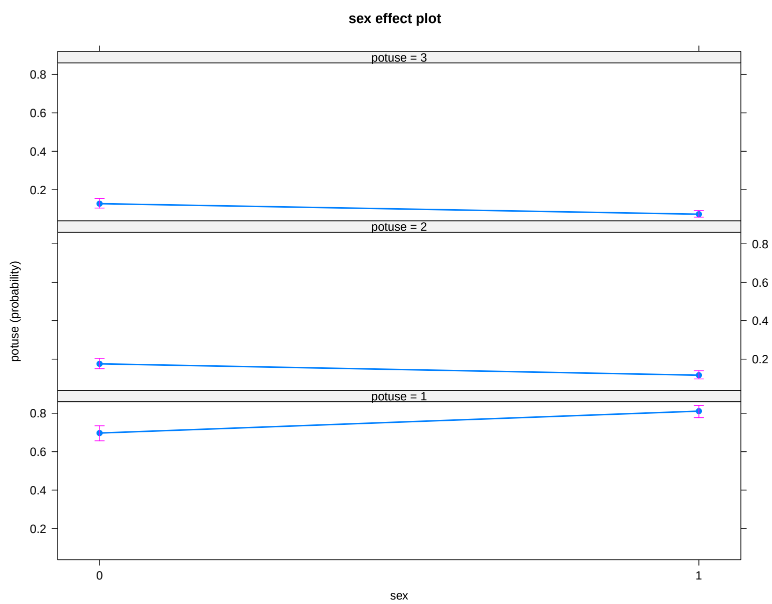


Figure 2: Effects for the variable `sex` in the proportional odds model

**END OF PAPER**