University of St Andrews

## DECEMBER 2019 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.

**PERMITTED MATERIALS:** Non-programmable calculator

**YOU MUST HAND IN THIS EXAM PAPER AT THE END OF THE EXAM.**

**PLEASE DO NOT TURN OVER THIS EXAM PAPER UNTIL YOU ARE INSTRUCTED TO DO SO.**

**1.** Figure 1 shows the speed at which 24 galaxies of the same kind are receding from us on the horizontal axis and the distance that they are from us on the vertical axis. A regression is performed in order to predict galaxy speed from distance. For the purposes of this question we assume that the distances are observed without error. Measurements of the speed of the galaxies were taken by three different laboratories (with no two laboratories measuring the same galaxy).

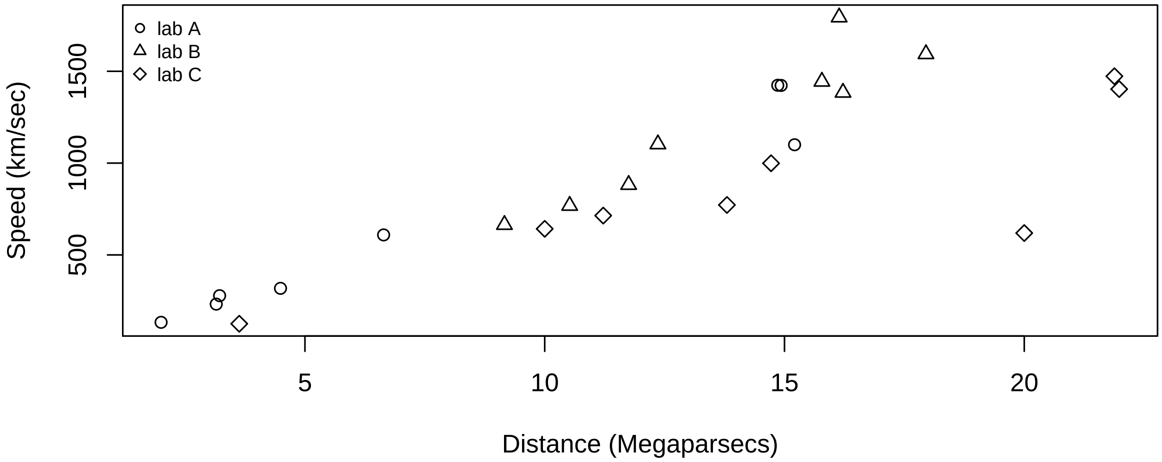

Figure 1: The speed at which 24 galaxies are receding from us, plotted against their distance from us. Measurements were taken by three different laboratories, with each laboratory indicated by a different symbol in the plot.

The following two models were fitted to these data

```
lm1 = lm(speed ~ distance*lab, data=hubble)
gls1 = gls(speed ~ distance*lab, data=hubble, weights=varFunc(~distance), method="ML")
```

Key parts of the R summary of each model is given below, together with their AIC values. Some diagnostic plots for model lm1 are shown in Figure 2 on page 4.

```
> summary(lm1)

Call:
lm(formula = speed ~ distance * lab, data = hubble)
```

```
Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)     -39.24     114.99  -0.341   0.7369
distance         90.40      11.77   7.681 4.36e-07 ***
labB           -430.09     327.25  -1.314   0.2053
labC             33.12     205.94   0.161   0.8740
distance:labB    31.44      24.77   1.269   0.2206
distance:labC   -32.43      15.96  -2.032   0.0572 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 183.7 on 18 degrees of freedom
Multiple R-squared:  0.8942,        Adjusted R-squared:  0.8649
F-statistic: 30.44 on 5 and 18 DF,  p-value: 3.514e-08


> AIC(lm1)
[1] 325.4419



> summary(gls1)
Generalized least squares fit by maximum likelihood
  Model: speed ~ distance * lab
  Data: hubble
       AIC      BIC    logLik
  313.2461 321.4925 -149.6231

Variance function:
 Structure: fixed weights
 Formula: ~distance

Coefficients:
                  Value Std.Error   t-value p-value
(Intercept)    -44.8335  52.92632 -0.847093  0.4081
distance        91.0956   8.59587 10.597604  0.0000
labB          -439.4843 260.71920 -1.685662  0.1091
labC           -16.3095 112.78084 -0.144612  0.8866
distance:labB   31.8343  20.91502  1.522079  0.1454
distance:labC  -29.3696  11.70565 -2.509009  0.0219

Residual standard error: 38.59752
Degrees of freedom: 24 total; 18 residual

> AIC(gls1)
[1] 313.2461
```

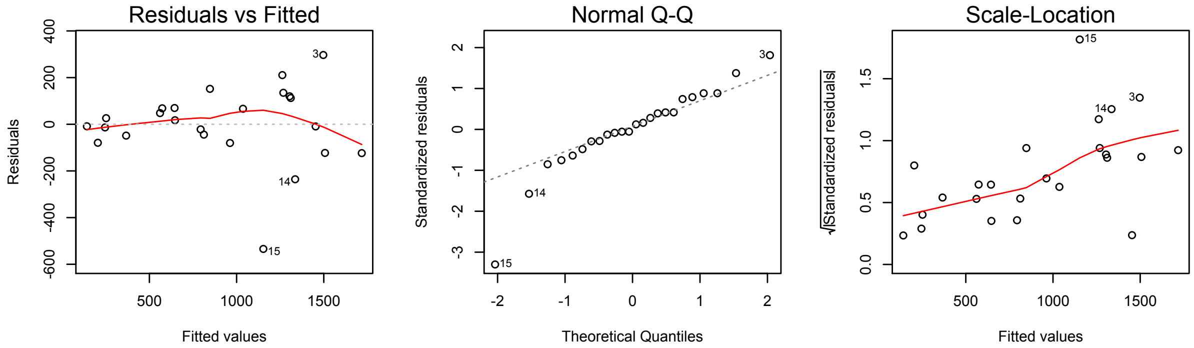

Figure 2: Some diagnostic plots for model `lm1`.

(a) Comment on the adequacy of model `lm1`. **[3]**

**Solution**:
First plot suggests linearity assumption is fine.
Q-Q plot suggest some lack of normality of errors (too heavy right tail), but this mainly due to 1 observation.
Third plot suggests that assumption of constant variance may be violated.

(b) What is the estimated standard error of the error distribution at `distance=16` for the second model (model `gls1`)? Explain your answer and show your working. **[1]**

**Solution**:
$\approx 154.3901$

(c) Which of these two models is preferable? Explain your answer. **[2]**

**Solution**:
*Difficulty: Medium.*
Model gls1 is preferable because (i) it allows increasing variance with distance and there is evidence of this in Figure 1, (ii) it has much lower AIC (313.2 *vs* 325.4).

(d) Using the results from model `gls1`, calculate the speed at which lab C predicts that a galaxy that is at 15 Megaparsecs from us, is moving away from us. **[3]**

**Solution**:
864.747 km/sec

**2.** The number of lizards of each of two species (*grahami* or *opalinus*) found at 24 locations was recorded in order to draw conclusions about the type of perches that each species prefers. In addition to counts, the data contain information on whether each perch was in the sun or in the shade. A generalised linear model was fitted to these data, as shown below.

```
> summary(lizardata)
    sun          species       count      
 shady:24   grahami :24   Min.   : 0.00  
 sunny:24   opalinus:24   1st Qu.: 2.75  
                          Median : 5.50  
                          Mean   :11.75  
                          3rd Qu.:13.50  
                          Max.   :69.00  
> GLM1 = glm(count~species*sun,family=poisson,data=lizardata)
> summary(GLM1)

Call:
glm(formula = count ~ species * sun, family = poisson, data = lizardata)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-5.7698  -1.8057  -0.5713   1.0806   6.4047  

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)    
(Intercept)               3.34990    0.05407  61.951   <2e-16 ***
speciesopalinus          -1.06413    0.10676  -9.967   <2e-16 ***
sunsunny                 -1.34617    0.11900 -11.313   <2e-16 ***
speciesopalinus:sunsunny -0.71646    0.29883  -2.398   0.0165 *  
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 737.56  on 47  degrees of freedom
Residual deviance: 322.65  on 44  degrees of freedom
AIC: 493.89

Number of Fisher Scoring iterations: 5
```

(a) Write down the equation for the expected number of counts in terms of the model coefficients and say what each of the coefficients in the equation represents. **[4]**

**Solution**:

The interpretation requires understanding of the data.

(b) Assuming that model `GLM1` is adequate, use the output above to decide which, if any, of the two species has greater preference for perching in the sun compared to perching in the shade. Explain your choice and show your working. **[2]**

**Solution**:
*Difficulty: Moderate difficulty (because of interaction).*
GLM1.
*grahami* has greater preference for the sun than does *opalinus.*

(c) A second generalised linear model is fitted to these data, using a quasi-Poisson error model, as shown below

```
> GLM2 = glm(count~species*sun,family=quasipoisson,data=lizardata)
> summary(GLM2)

Call:
glm(formula = count ~ species * sun, family = quasipoisson, data = lizardata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.7698  -1.8057  -0.5713   1.0806   6.4047

Coefficients:
                         Estimate Std. Error t value Pr(>|t|)
(Intercept)                3.3499     0.1490  22.482  < 2e-16 ***
speciesopalinus           -1.0641     0.2942  -3.617 0.000764 ***
sunsunny                  -1.3462     0.3279  -4.105 0.000172 ***
speciesopalinus:sunsunny  -0.7165     0.8234  -0.870 0.388977
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for quasipoisson family taken to be 7.59301)

    Null deviance: 737.56  on 47  degrees of freedom
```

```
Residual deviance: 322.65  on 44  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 5
```

(i) Explain why the estimated standard errors of the coefficients are greater for the model that assumes a quasi-Poisson error distribution than the model that assumes a Poisson error distribution, and why the interaction parameter is not significant in the case of the quasi-Poisson model, while it is significant at the 5% level in the case of the Poisson model. **[4]**

**Solution**:
*Difficulty: Medium.*
Thsi question is about overdispersion and its consequences.

(ii) Which of the two models is to be preferred? Justify your answer. **[1]**

**Solution**:
*Difficulty: Easy.*
The model with overdispersion.

(iii) Calculate the expected numbers of each species that are to be found on sunny perches. **[2]**

**Solution**:
*grahami*: 7.4
*opalinus*: 2.56 (neglecting non-significant interaction)

(iv) Obtain a 95% confidence interval for the number of species *grahami* to be found on shady perches. State any assumptions you use in doing this and show your working. **[3]**

**Solution**:
A 95% CI for expected count is $(21.3, 38.2)$.

**3.** For this question we consider a study on the effectiveness of a new teaching method in economics. Data on 32 students were collected. The response variable is `grade`, which takes the value 1 if the exam grades of the student were improved and 0 otherwise. We model the probability of an improvement using the following covariates:

- `exposure` - a binary variable taking the value 1 if the student was exposed to PSI (the new teaching method) and 0 otherwise.
- `ability` - a measure of ability when entering the class
- `gpa` - grade point average of the student

A binary model with a logit link is fitted to the data. The corresponding R output is given on page 10.

(a) Write out the link function for the `fit.tm` model. Explain any parameters in your formula. **[2]**

Bookwork

(b) Give the effect of `exposure` on the **odds** for improvement of the grades and use your result to interpret the respective parameter. Is this parameter significant? **[2]**

$e^{2.37869} \approx 10.79.$

(c) Provide a 95% confidence interval for the `exposure` parameter and a 95% confidence interval for the effect of `exposure` on the odds of improvement. **[3]**

A 95 % confidence intreval for `exposure` is given by

$$\approx [0.2922, 4.4653]$$

Respective confidence interval for the effect:

$$\approx [1.3393, 86.9408]$$

(d) Explain what a saturated model is and what `Null deviance` and `Residual deviance` in the R output stand for. Do a likelihood ratio test with Null hypothesis "The probability of improvement does not depend on any of the covariates." and clearly state your conclusion. You may use the following result:

```
> qchisq(0.95,df=3)
[1] 7.814728
```

**[4]**

There is compelling evidence that the model has some explanatory power.

(e) To further assess the quality of the model fit, the following confusion matrix was produced using the mean fitted values as a threshold:

Table 1: Table depicting the output of the confusion matrix. Each cell gives a count.

| | | **Observed** | **Values** |
|---|---|---|---|
| | | 0 | 1 |
| **Predicted** | 0 | 17 | 2 |
| | 1 | 4 | 9 |

(i) Give the **accuracy** of the fitted model. **[1]**

81.25 %

(ii) An optimal threshold could be chosen using a Receiver Operating Characteristic (ROC) curve. Figure 3 on page 10 gives the ROC curve for `fit.tm`. Explain how the best threshold is chosen. Indicate (using simple words) the location on the graph of a model with a perfect predictive power. **[2]**

Discussed in the lectures.

```
> fit.tm<-glm(grade~exposure+ability+gpa, family = binomial, data=spector)
> summary(fit.tm)

Call:
glm(formula = grade ~ exposure + ability + gpa, family = binomial,
    data = spector)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9551  -0.6453  -0.2570   0.5888   2.0966

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -13.02135    4.93127  -2.641  0.00828 **
exposure      2.37869    1.06456   2.234  0.02545 *
ability       0.09516    0.14155   0.672  0.50143
gpa           2.82611    1.26293   2.238  0.02524 *
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 41.183  on 31  degrees of freedom
Residual deviance: 25.779  on 28  degrees of freedom
AIC: 33.779

Number of Fisher Scoring iterations: 5
```

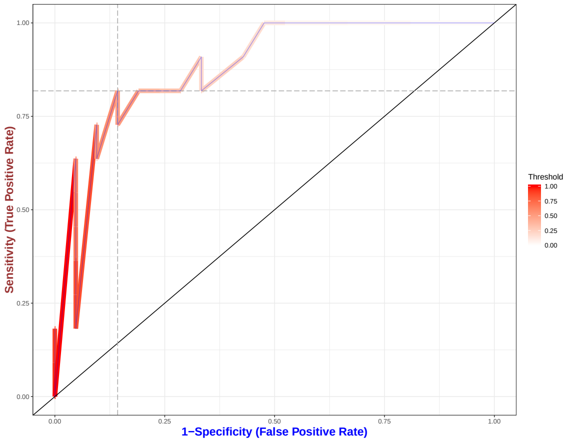

Figure 3: The ROC curve for the `fit.tm` model. The location of best threshold for the model fit is indicated by the intersection of the dashed lines.

**4.** Data were collected in a large postal survey on the psychology of debt. The response variable `ccarduse` is the frequency of credit card use. It is a three-level factor taking the values 1=never, 2=occasionally and 3=regularly. The explanatory variables are:

- `incomegp` - income group taking integers from 1 (lowest) to 5 (highest)
- `children` - number of children in the household
- `cigbuy` - a binary variable taking the value 1 if the person buys cigarettes and 0 otherwise
- `prodebt` - a continuous variable giving the score on a scale of attitudes to debt (high values = favourable to debt)

A proportional odds model was fitted to these data. The R output is given on page 13.

(a) Describe the difference between nominal and ordinal data. Use an example to illustrate your explanation. **[1]**

Bookwork

(b) State the assumptions of the proportional odds model. (You may state them using words alone or words supplemented by mathematical equations.) **[3]**

Bookwork

(c) Write down an equation for the probabilities $Pr(ccarduse \geq j)$ for $j = 1, 2, 3$ and explain the meaning of each of the parameters in the equation. **[2]**

$$Pr(Y \geq j) = \frac{\exp(\beta_{0j} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}{1 + \exp(\beta_{0j} + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)},$$

for $j = 2, 3$ with $Pr(Y \geq 1) = 1$.

(d) What is the estimated probability of never using credit cards (i.e. `ccarduse` taking the value 1) for a person who is in income group 1, has 1 child, doesn't buy cigarettes and has a score of attitude to debt equal to 2? **[2]**

0.8049352

(e) Explain what the parameter estimate for the variable `incomegp` implies about the effect of `incomegp` on the response variable. Comment whether the inclusion of this variable is justified and suggest an explanation for the estimated effect of `incomegp` on the response variable. **[3]**

The larger the positive value of the coefficient for `incomegp`, the more likely a respondent is to use credit cards more frequently.

```
Call:
vglm(formula = ccarduse ~ incomegp + children + cigbuy + prodebt,
    family = propodds, data = debt2, reverse = TRUE)


Pearson residuals:
                 Min      1Q  Median      3Q   Max
logit(P[Y>=2]) -2.24 -0.7093 -0.3457  0.7668 2.611
logit(P[Y>=3]) -1.72 -0.5023 -0.2315 -0.1234 4.183

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept):1 -2.71888    0.53757  -5.058 4.24e-07 ***
(Intercept):2 -4.06567    0.56350  -7.215 5.39e-13 ***
incomegp       0.49555    0.08192   6.049 1.46e-09 ***
children      -0.13066    0.09656  -1.353 0.176009
cigbuy        -0.86715    0.24599  -3.525 0.000423 ***
prodebt        0.46828    0.15421   3.037 0.002392 **
---
Signif. codes:  0 ?***? 0.001 ?**? 0.01 ?*? 0.05 ?.? 0.1 ? ? 1

Number of linear predictors:  2

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

Residual deviance: 659.7429 on 720 degrees of freedom

Log-likelihood: -329.8714 on 720 degrees of freedom

Number of iterations: 5

No Hauck-Donner effect found in any of the estimates

Exponentiated coefficients:
 incomegp  children    cigbuy   prodebt
1.6414026 0.8775194 0.4201466 1.5972430
```

**END OF PAPER**