University of St Andrews

DECEMBER 2018 SOME ANSWERS
EXAMINATION DIET

SCHOOL OF MATHEMATICS & STATISTICS

MODULE CODE:	**MT5761**
MODULE TITLE:	Statistical Modelling
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. Dr Econ O'Mist models the relationship between patents and R&D (Research & Development) expenditures, using U.S. data on 346 firms for each of the five years 1975 - 1979. This results in 1730 observations - 1 for each firm for each year. The dependent variable is successful patents, defined as the number of patents applied for during the year that were eventually granted. In particular, Dr O'Mist wants to model the log counts of patents (`logpat`) using linear regression with the following two explanatory variables:

- `logr` - the log R&D spendings in US dollars and
- `scisect` - a binary variable taking the value 1 if the firm is in the science sector and 0 otherwise.

An initial model `lmfit` is fitted using `logr` and `scisect` as main effects as well as an interaction term between the two main effects.

Use the output on page 3 to answer Question 1.

(a) What immediate problem does the log transformation of counts potentially cause? Suggest a quick fix. **[2]**

[redacted]

(b) Write out the equation for the `lmfit` model. Interpret the intercept and the coefficient of `logr` and explicitly state the error distribution. **[4]**

[redacted]

(c) Interpret the coefficient of the interaction term and thus explain how the interaction term affects the relationship between `logpat` and `logr`. Considering the output, was it justified to include the interaction term? Justify your answer. **[3]**

[redacted]

(d) Without using further diagnostic tests or plots, do you expect any of the assumptions of a linear regression to be violated? Justify your answer. **[2]**

[redacted]

The output for Question 1. appears on page 3.

```
> lmfit<-lm(logpat~logr+scisect+scisect:logr,data=pt)
> summary(lmfit)

Call:
lm(formula = logpat ~ logr + scisect + scisect:logr, data = pt)

Residuals:
    Min      1Q  Median      3Q     Max
-3.3930 -0.6422  0.0195  0.6792  3.2709

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     -0.45571    0.04192 -10.872  < 2e-16 ***
logr             0.85208    0.01751  48.655  < 2e-16 ***
scisectyes       0.33433    0.07158   4.671 3.24e-06 ***
logr:scisectyes -0.02373    0.02578  -0.920    0.358
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.032 on 1726 degrees of freedom
Multiple R-squared:  0.7298,Adjusted R-squared:  0.7293
F-statistic:  1554 on 3 and 1726 DF,  p-value: < 2.2e-16
```

2. Furthering his investigation Dr O'Mist re-analyses the data in question **1.** using a Generalised Least Squares (GLS) model.

[redacted]

(a) How does the `fitglsvp` model extend the linear model approach? Suggest a test to check whether this extension is necessary. Give the mean-variance relation that underpins the `fitglsvp` model and use the summary of this model to provide the estimate(s) for the parameter(s) required. **[4]**

[redacted]

(b) In addition to the GLS model above, further three models were fitted as shown on page 6. Table 1 below shows the AIC and BIC scores for each of these models. The autocorrelation function of the normalised residuals of models `fitglscorar1` and `fitglscorar3` are given in the figure on page 6.

Model	df	AIC	BIC
[redacted]glsvp	6	5010.45.6	5043.19
fitglsvpcorar1	7	1440.63	1473.37
fitglsvpcorar2	8	1440.96	1479.15
fitglsvpcorar3	9	1439.50	1483.14

Table 1: Table of GLS models fitted and corresponding AIC and BIC statistic.

Based on the information you have, which of the models would you choose as the 'best' model? Justify your answer. **[3]**

[redacted]

The output for Question 2. appears on pages 5-6.

```
> fitglsvp<-gls(logpat~logr+scisect+scisect:logr,data=pt, weights=varPower(),
               method="ML")
> summary(fitglsvp)
Generalized least squares fit by maximum likelihood
  Model: logpat ~ logr + scisect + scisect:logr
  Data: pt
       AIC      BIC    logLik
  5010.455 5043.191 -2499.228

Variance function:
 Structure: Power of variance covariate
 Formula: ~fitted(.)
 Parameter estimates:
      power
-0.08325532

Coefficients:
                     Value  Std.Error   t-value p-value
(Intercept)     -0.4546021 0.04322073 -10.51815  0.0000
logr             0.8577707 0.01678449  51.10497  0.0000
scisectyes       0.3054463 0.07476081   4.08565  0.0000
logr:scisectyes -0.0182418 0.02505615  -0.72804  0.4667

 Correlation:
                (Intr) logr   scscty
logr            -0.653
scisectyes      -0.578  0.378
logr:scisectyes  0.438 -0.670 -0.734

Standardized residuals:
        Min          Q1         Med          Q3         Max
-3.35048085 -0.63592440  0.01438334  0.65707434  3.27632128

Residual standard error: 1.039672
Degrees of freedom: 1730 total; 1726 residual
```

```
# cusip is the identifier for the company
#
> fitglsvpcorar1<-gls(logpat~logr+scisect+scisect:logr,data=pt,
                     correlation=corAR1(form= ~1|cusip), method="ML")
> fitglsvpcorar2<-gls(logpat~logr+scisect+scisect:logr,data=pt,
                    correlation=corARMA(p=2,q=0,form= ~1|cusip), method="ML")
> fitglsvpcorar3<-gls(logpat~logr+scisect+scisect:logr,data=pt,
                    correlation=corARMA(p=3,q=0,form= ~1|cusip), method="ML")
```

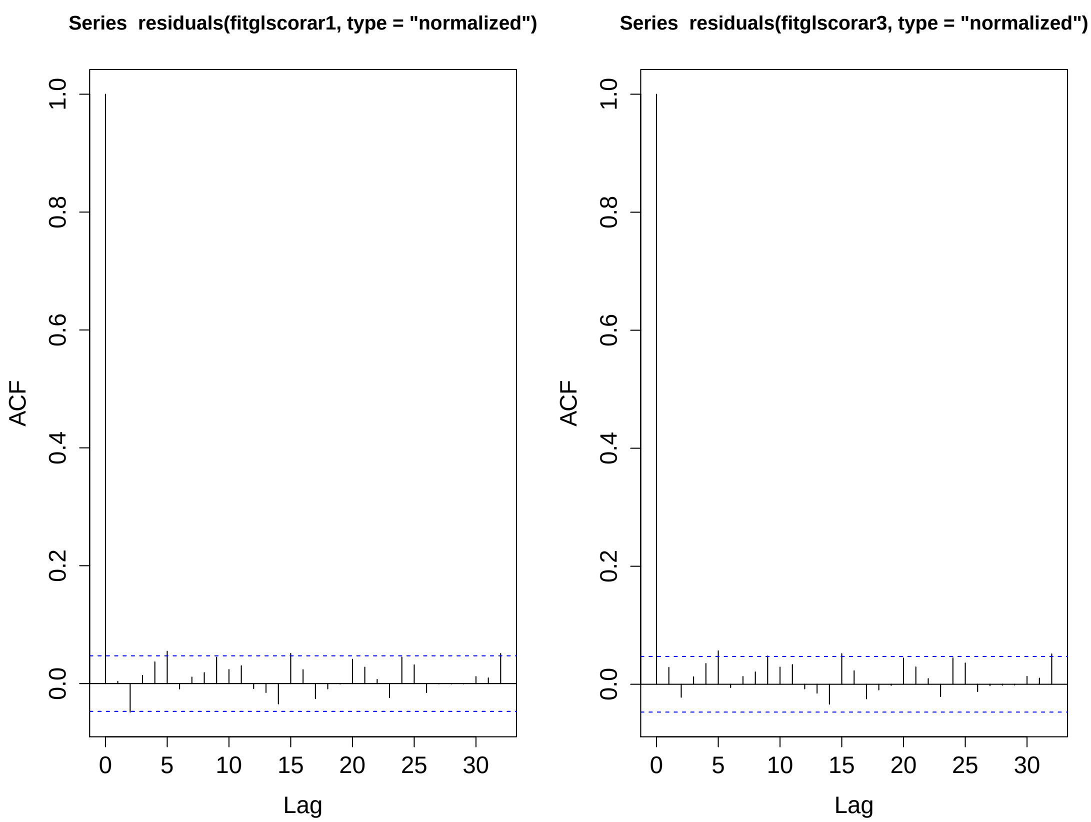

Figure 1: Autocorrelation function plots for the normalised residuals from the `fitglsvpcorar1` (left plot) and `fitglsvpcorar3` (right plot) models.

3. Dr Finn Ans, a colleague of Dr Econ O'Mist, is interested in modelling the patents of the firms only in 1979 so he considers the data of the 346 firms only for that year. As opposed to Dr O'Mist, Dr Ans decides to work with the counts of patents directly (instead of taking their logs). For this purpose he fits a Poisson model with `logr` and `scisect` (defined in Question **1.**) as explanatory variables but without an interaction term.

Use the output on pages 8 - 10 to answer Question 3.

(a) Give the formula for the link function used in the Poisson model. Specify the linear predictor without interpreting the parameters. [**1**]

[redacted]

(b) To account for possible overdisperssion, Dr Ans fits a Quasi-Poisson model with the same explanatory variables.

(i) Explain what overdispersion in a Poisson model is and how it affects the model fit. Does the summary of the two model fits on pages 8 - 9 support your statement? [**4**]

[redacted]

(ii) Is the use of a Quasi-Poisson model justified? Base your answer on the summary of the model fit on page 9. [**1**]

[redacted]

(iii) As a part of the model assessment for the Quasi-Poisson model the scaled Pearson residuals were plotted against the fitted values. A plot of the autocorrrelation function of the same residuals is also provided (see the figure on page 10). Comment on the adequacy of the fit based on the two plots. [**3**]

[redacted]

The output for Question 3. appears on pages 8-10.

```
> # A Poisson GLM
> poisfit<-glm(pat~logr+scisect, data=pt1979, family=poisson)
> summary(poisfit)

Call:
glm(formula = pat ~ logr + scisect, family = poisson, data = pt1979)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-16.987   -1.867   -0.646    0.985   39.520

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -0.127570   0.044376  -2.875  0.00404 **
logr         1.020381   0.009213 110.756  < 2e-16 ***
scisectyes  -0.341524   0.021297 -16.036  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 30108.8  on 345  degrees of freedom
Residual deviance:  8203.4  on 343  degrees of freedom
AIC: 9311.7

Number of Fisher Scoring iterations: 6
```

```
> poisfit_OD<-glm(pat~logr+scisect, data=pt1979, family=quasipoisson)
> summary(poisfit_OD)

Call:
glm(formula = pat ~ logr + scisect, family = quasipoisson, data = pt1979)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-16.987   -1.867   -0.646    0.985   39.520

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -0.12757    0.28110  -0.454   0.6502
logr         1.02038    0.05836  17.485   <2e-16 ***
scisectyes  -0.34152    0.13491  -2.532   0.0118 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for quasipoisson family taken to be 40.12521)

    Null deviance: 30108.8  on 345  degrees of freedom
Residual deviance:  8203.4  on 343  degrees of freedom
AIC: NA

Number of Fisher Scoring iterations: 6
```

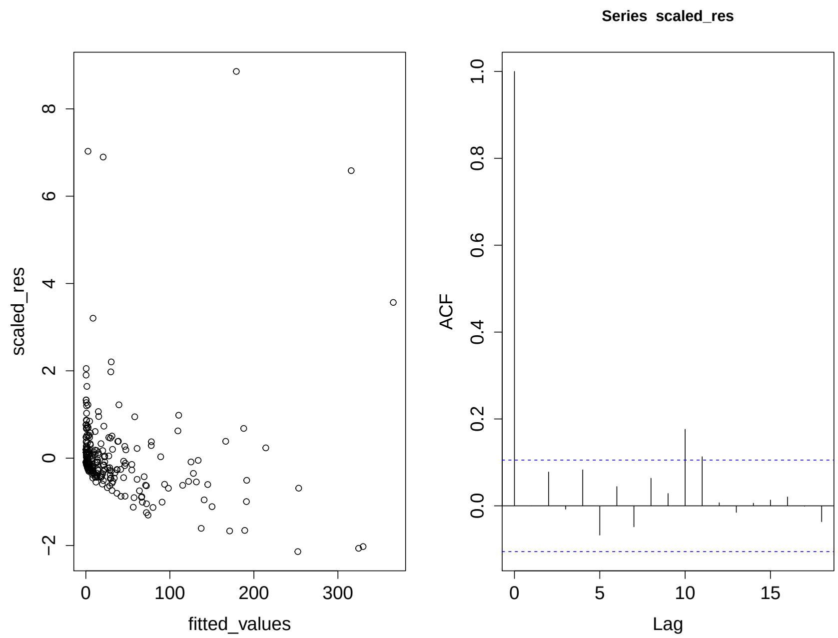

Figure 2: Model assessment for the Quasi-Poisson model. The left plot gives the scaled Pearson residuals against the fitted values. The right plot gives the autocorrelation function of the scaled Pearson residuals.

4. In this question, we consider some data on life expectancy from 142 countries across the world, for 12 time periods (approximately every five years from 1952 to 2007). For each country and time period, we have recorded whether or not the life expectancy at birth is equal to or greater than 70 years old (1) or not (0). The data are then grouped by country to give the average probability of life expectancy ≥ 70 in each country over the study period (55 years). We will investigate if the probability of life expectancy at birth ≥ 70 years can be estimated for each country using the following covariates:

- `slife70` - number of successes per country (i.e. number of time periods per country where life expectancy at birth ≥ 70)
- `n` - total number of trials per country (i.e the number of time period records for each country)
- `continent` - one of five continents:
  - Africa
  - Americas
  - Asia
  - Europe
  - Oceania
- `mgdp` - mean GDP per capita (Gross domestic product (GDP) is a monetary measure of the market value of all final goods and services produced in a period of time).

Use the output on page 13 to answer Question 4.

(a) Write out the equations for the link and inverse link functions used in the `fit.bi`[redacted]model. Be sure to explain any parameters in your formulae. **[3]**

(b) Calculate the odds of a life expectancy at birth ≥ 70 for Germany using the `fit.bin` model (`mgdp` = 20557, `continent` = 'Europe') and interpret your answer. Calculate the estimated expected probability of having a life expectancy at birth of at least 70 years in Germany. **[3]**

[redacted]

(c) Interpret the estimated value of the coefficient for Europe. **[2]**

[redacted]

(d) Explain what a confusion matrix for binary data is. Can you use it for the fitted model? **[2]**

[redacted]

```
Call:
glm(formula = cbind(slife70, n - slife70) ~ mgdp + continent,
    family = binomial, data = newdat)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-6.6524  -0.8311  -0.7704   0.7385   4.0864

Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -3.741e+00  2.361e-01 -15.846  < 2e-16 ***
mgdp               1.095e-04  1.173e-05   9.341  < 2e-16 ***
continentAmericas  2.277e+00  2.703e-01   8.424  < 2e-16 ***
continentAsia      1.646e+00  2.748e-01   5.991 2.09e-09 ***
continentEurope    3.176e+00  2.867e-01  11.078  < 2e-16 ***
continentOceania   4.108e+00  7.961e-01   5.161 2.46e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 1046.92  on 141  degrees of freedom
Residual deviance:  301.56  on 136  degrees of freedom
AIC: 518.67

Number of Fisher Scoring iterations: 5
```

5. As part of a study on the factors determining happiness, 39 students in Chicago independently completed a questionnaire. The variable of interest is `happy`, which measures the happiness of the surveyed students on a 5-point scale, from 1 (extremely unhappy) to 5 (extremely happy). The explanatory variables are:

- `money` - annual family income, in thousands of dollars.
- `love` - 3-point scale, from 1 (lonely) to 3 (deep falling in love)
- `sex` - binary, taking the values 0 (unsatisfactory) and 1 (satisfactory)
- `work` - 5-point scale, from 1 (unemployed) to 5 (in enjoyable employment).

A proportional odds model was fitted to these data.

Use the output on page 16 to answer Question 5.

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

[redacted]

(ii) State procedures you would use to check each assumption, and what you would look for if the assumption was met. (Give the name of each procedure or a brief description; do not give the name of a computer function that implements the procedure.) **[3]**

[redacted]

(b) (i) What is the estimated probability of a student having a happy score of 4 and above if they have \$0 annual income and love, sex and work scores of 3, 0 and 1, respectively?
(Hint - the `vglm` *function uses the* $Pr(Y \geq j)$ *specification of the linear predictors.)* **[1]**

[redacted]

(ii) What is the estimated probability of a student having a happy score of 1 if they have \$0 annual income and love, sex and work scores of 1, 1 and 1, respectively (essentially an unemployed student engaging in loveless sex)? **[2]**

[redacted]

(c) Interpret in as plain English as possible, the value of the parameter estimate for the variable love, i.e. what exactly does the value mean for the effect of love on happiness? Considering the given output, is the inclusion of this variable justified?

[redacted]

(d) The same data were fit using a multinomial logit model. Explain how this model differs from proportional odds. Why is it less appropriate in this case?

[2]

[redacted]

```
> fit<- vglm(happy~money+sex+love+work,propodds,data=happy,reverse=TRUE)
> summary(fit)

Call:
vglm(formula = happy ~ money + sex + love + work, family = propodds,
    data = happy, reverse = TRUE)

Pearson residuals:
                   Min        1Q   Median       3Q    Max
logit(P[Y>=2]) -3.498  0.008281  0.01694  0.06913 0.9005
logit(P[Y>=3]) -2.006  0.026533  0.04717  0.19866 1.0434
logit(P[Y>=4]) -4.602 -0.141476  0.11672  0.28956 1.6573
logit(P[Y>=5]) -1.549 -0.315831 -0.07641 -0.02890 3.3381

Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept):1  -5.96780    2.32371  -2.568 0.010222 *
(Intercept):2  -9.23359    2.53797  -3.638 0.000275 ***
(Intercept):3 -11.43621    2.78936  -4.100 4.13e-05 ***
(Intercept):4 -16.96118    3.71628  -4.564 5.02e-06 ***
money           0.02284    0.01183   1.932 0.053421 .
sex             0.10493    0.85329   0.123 0.902127
love            3.23005    0.89547   3.607 0.000310 ***
work            1.00916    0.44963   2.244 0.024805 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Number of linear predictors:  4

Names of linear predictors: logit(P[Y>=2]), logit(P[Y>=3]),
                            logit(P[Y>=4]), logit(P[Y>=5])

Residual deviance: 59.2774 on 148 degrees of freedom

Log-likelihood: -29.6387 on 148 degrees of freedom

Number of iterations: 7

Warning: Hauck-Donner effect detected in the following estimate(s):
'(Intercept):4', 'love'

Exponentiated coefficients:
    money       sex      love      work
 1.023107  1.110637 25.280834  2.743290
```

END OF PAPER