Imperial College
London

BSc and MSci EXAMINATIONS (MATHEMATICS)

May-June 2012

This paper is also taken for the relevant examination for the Associateship of the Royal College of Science.

M3S2/M4S2/M5S2

STATISTICAL MODELLING II

Date: Tuesday, 8th May 2012 Time: 14:00 – 16:00

Credit will be given for all questions attempted but extra credit will be given for complete or nearly complete answers.

Calculators may not be used.

1. a. Write down the Poisson GLM with canonical link, its log-likelihood function, and state the typical assumptions made about the observations. Assuming a dataset $(y_i, \mathbf{X}_i)_{i=1:n}$, derive the formula for its deviance, D. Write down the formula for the Pearson statistic, X^2.

b. Recall the Taylor expansion of a function $f(y)$ around $y = a$:

$$f(y) = f(a) + f'(a)(x-a) + \frac{f''(a)(x-a)^2}{2!} + \frac{f'''(a)(x-a)^3}{3!} + ...$$

By considering the first three terms in the Taylor expansion of $y_i \log\left(\frac{y_i}{\hat{\mu}_i}\right)$ as a function of y_i around $y_i = \hat{\mu}_i$, show that $D \approx X^2$.

c. For the quasi-Poisson model the deviance is

$$D^{\mathsf{OD}} = \frac{1}{\tilde{\phi}} D$$

where D is the deviance you computed in the question above, and $\tilde{\phi}$ is an estimated dispersion parameter. State the estimator $\tilde{\phi}$ that was given in lectures. Hence establish that the deviance D^{OD} is approximately equal to $n-p$ in the over-dispersed case, where p is the dimension of $\mathbf{X}_i$. A rule of thumb for GLMs is that when the deviance is approximately equal to the degrees of freedom (in this case, $n-p$), then the model is a reasonably good fit. Comment in one sentence on the suitability of this rule in the over-dispersed case.

d. We wish to compare two models, one with and one without an intercept, but otherwise featuring the same set of covariates, using anova() in R. State whether an F-test or a χ^2-test is appropriate in each of the following cases: a) Gaussian where σ^2 is *known and fixed* to a certain value, b) inverse Gaussian, c) quasi-Poisson, d) Binomial, e) Gamma.

[END OF QUESTION 1.]

2. a. Write down the general form of the log-likelihood of the ith datapoint y_i under a GLM, in canonical parameterisation (in what we called 'Nelder' notation in lectures). Consider a set of observations recording the number y_i of successes in n_i trials of a certain experiment. A binomial model for y_i is given by:

$$f(Y_i = y_i; \mu_i) = \begin{pmatrix} n_i \\ y_i \end{pmatrix} \mu_i^{y_i} (1-\mu_i)^{n_i - y_i}$$

where $\mathbb{E}[y_i] = n_i \mu_i$. By letting $U_i = Y_i / n_i$ be the random variable instead, write the log-likelihood of this model in canonical form. State the formula for the variance function of the binomial *without proof*, and hence prove that $\mathsf{Var}[U_i] = \frac{1}{n_i}\mu_i(1-\mu_i)$.

b. The Fisher information for the binomial GLM in terms of U_i is given by

$$\mathcal{J}_{jk} = \sum_{i=1}^{n} \frac{x_{ij} x_{ik}}{\mathsf{Var}[U_i]} \left(\frac{\partial \mu_i}{\partial \eta_i} \right)^2$$

Show that in the special case where the number of trials is the same for all observations, $n_i = m$, $\mathcal{J}_{jk}$ is an *increasing* function of m. Write down the z-values for the regression coefficients (that R outputs via "summary()"), as a function of m, and hence determine whether the p-values increase, or decrease with m. Is this consistent with intuition?

c. Consider the data in Table 1, where $m = 10$ individuals of each age group were selected to answer a question in a survey, and the proportion thereof that answered "yes" was recorded. Is Age, treated as a continuous variable, significant at the 5% level? Answer this question on the basis of the R output in Figure 1 (overleaf), from a canonical binomial GLM. Also predict the proportion of 10-year-old respondents that would answer "Yes". You do not need to simplify your answer.

i	Percentage replied "Yes"	Age	Number of Respondents
1	0.1	25	10
2	0.1	30	10
3	0.2	35	10
4	0.3	40	10
5	0.3	45	10
6	0.6	50	10

Table 1: Dataset B

d. Would your answers in question 2(c) above have been different if $m = 100$ instead (all other numbers in Table 1 remaining equal)? Justify your answer mathematically.

e. If Age in Table 1 was treated as a (categorical) factor with 5 levels (one per age group), rather than as a continuous variable, what would be the deviance of the resulting model?

[QUESTION CONTINUED OVERLEAF]

```
Call:
glm(formula = Prob ~ Age, family = binomial, weights = Trials)

Deviance Residuals:
       1         2         3         4         5         6
 0.27745  -0.21467   0.08409   0.11769  -0.65844   0.44644

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.12304    1.68684  -3.037  0.00239 **
Age          0.10485    0.04064   2.580  0.00988 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 8.68365  on 5  degrees of freedom
Residual deviance: 0.77684  on 4  degrees of freedom
AIC: 19.015

Number of Fisher Scoring iterations: 4
```

Figure 1: R output on dataset B.

[END OF QUESTION 2.]

3. a. The pdf of the Gamma distribution is

$$f(y;k,\mu) = \frac{k^k}{\mu^k}\frac{1}{\Gamma(k)}y^{k-1}e^{-ky/\mu}, \quad k,\mu > 0, y \geq 0$$

where $\mathbb{E}[Y] = \mu$. Letting $\phi = \frac{1}{k}$, write the log-likelihood of the ith datapoint in canonical form, and hence derive the canonical link of the Gamma GLM. Let $r_D(i)$ indicate the ith deviance residual for a dataset $(y_i, \mathbf{X}_i)_{i=1:n}$. Give the formula for $r_D(i)$ in this case, recalling that $\sum_{i=1}^n r_D^2(i) = \phi D$ to help you check your answer.

b. State the variance function of the Gamma. State the variance function of the Gaussian. Hence write down the formulae for the Pearson residuals, in each case.

c. In the top plots of Figure 2, we observe the squared response residuals vs the fitted values for a canonical Gamma GLM (left), and a Gaussian GLM with inverse link (right), fitted on the *same dataset*. These plots look almost identical, but one of them indicates a poor fit. Which one, and why?

d. State four properties of residuals that indicate a good fit. Which of these can be assessed on the basis of the middle and bottom plots in Figure 2? Compare the two models.

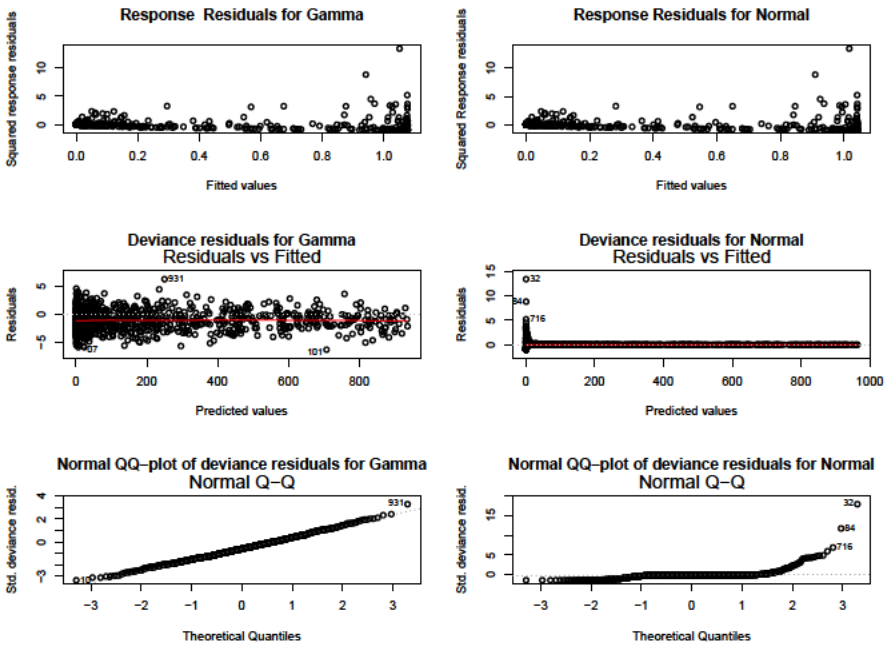

Figure 2: Top: squared response residuals vs fitted values for a Gamma model with canonical link, and a Normal model with inverse link. Middle: deviance residuals vs linear predictors for the two models. Bottom: QQ-plots for deviance residuals for the two models.

[QUESTION CONTINUED OVERLEAF]

e. The ith jackknife residual for the GLM is defined as

$$\hat{\epsilon}_i^{(-i)} = y_i - \hat{y}_i^{(-i)}$$

where $\hat{y}_i^{(-i)}$ is the predicted value of y_i using the MLE $\hat{\beta}^{(-i)}$ of β obtained by fitting the GLM to the dataset after deleting the ith observation. Show that, for a general GLM,

$$\hat{\epsilon}_i^{(-i)} > \hat{\epsilon}_i \text{ if and only if } \mathbf{X}_i\hat{\beta}^{(-i)} < \mathbf{X}_i\hat{\beta}$$

recalling that the link function g is monotonically increasing. We are about to fit a normal linear regression of y to X without an intercept to the data in Figure 3. Can you tell whether $\hat{\epsilon}_{10}^{(-10)}$ will be larger, or smaller, than $\hat{\epsilon}_{10}$? Comment briefly on whether the jackknife or the simple residual is more suitable for outlier detection in this case.

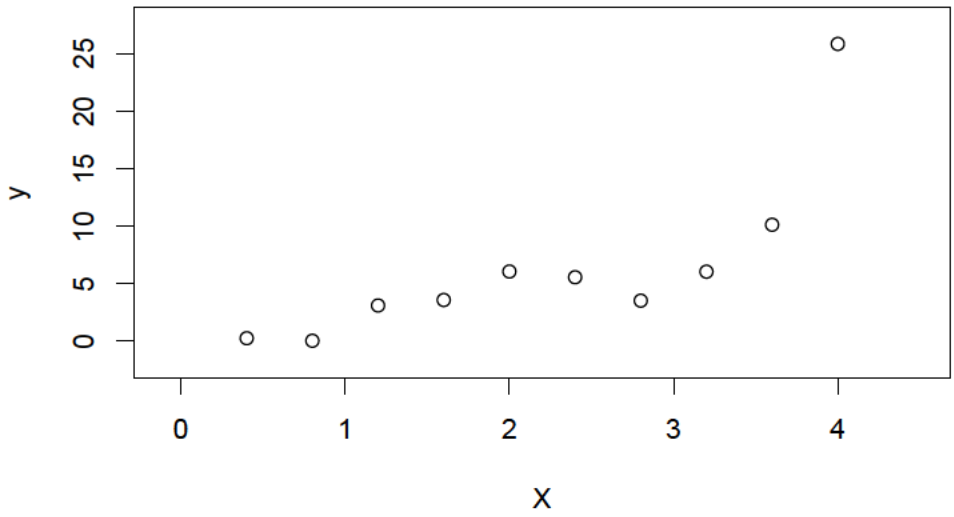

Figure 3: Data with an outlier.

[END OF QUESTION 3.]

4. a. Match each of the following four design matrices with one of the R commands listed further below, where $Z \in \{\text{'boy'}, \text{'girl'}\}$, and observations 1 and 2 refer to boys, whereas observations $3, 4$ and 5 to girls.

$$\mathbf{X}_A = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ 1 & x_3 \\ 1 & x_4 \\ 1 & x_5 \end{pmatrix}, \mathbf{X}_B = \begin{pmatrix} x_1 & 0 \\ x_2 & 0 \\ 0 & x_3 \\ 0 & x_4 \\ 0 & x_5 \end{pmatrix}, \mathbf{X}_C = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}, \mathbf{X}_D = \begin{pmatrix} 1 & x_1 & 0 \\ 1 & x_2 & 0 \\ 1 & x_3 & 0 \\ 1 & 0 & x_4 \\ 1 & 0 & x_5 \end{pmatrix}$$

(a) `lm(y ~ X:Z + 1)`

(b) `lm(y ~ X:Z + Z)`

(c) `lm(y ~ X + Z + 1)`

(d) `lm(y ~ X + 1)`

(e) `lm(y ~ X:Z + 0)`

(f) `lm(y ~ X + Z + 0)`

(g) `lm(y ~ X:Z + Z + 0)`

(h) `lm(y ~ 1)`

(i) `lm(y ~ Z + 0)`

b. Consider the Carbs dataset from lectures, where a set of male individuals were instructed to eat fewer carbohydrates over a given time period, and the amount of carbohydrate consumed by each was recorded alongisde his/her weight (relative to ideal weight for his height), age, and protein intake. In Figure 4 (overleaf), we include R output of the "drop1()" and the "summary()" commands applied on a model including Weight, Age, Protein and an intercept. On the basis of this output, which variable should you delete, if any?

c. On the basis of the simpler model where Age is deleted, we wish to investigate whether it is possible for an individual of ideal weight for his height (i.e., weight $= 100$) whose protein intake is 20% of his diet to consume 40% in carbohydrates. The confidence interval around the prediction computed by the predict() function is:

```
      fit      lwr      upr
1 47.45124 40.85429 54.04819
```

and the prediction interval is

```
       fit      lwr      upr
1 47.45124 33.23146 61.67101
```

Explain in one to two lines why the prediction interval is larger than the confidence interval. Use the appropriate interval to answer the above question.

[QUESTION CONTINUED OVERLEAF]

```
> drop1(model.1,test="F")
Single term deletions

Model:
carbohydrate ~ age + weight + protein
        Df Sum of Sq    RSS    AIC F value   Pr(>F)
<none>               567.66 74.916
age      1     38.36 606.02 74.224  1.0812 0.313893
weight   1    265.91 833.57 80.600  7.4948 0.014599 *
protein  1    337.34 905.00 82.244  9.5082 0.007121 **

>summary(model.1)
Call:
lm(formula = carbohydrate ~ age + weight + protein, data = carbs)

Residuals:
     Min       1Q   Median       3Q      Max
-10.3424  -4.8203   0.9897   3.8553   7.9087

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 36.96006   13.07128   2.828  0.01213 *
age         -0.11368    0.10933  -1.040  0.31389
weight      -0.22802    0.08329  -2.738  0.01460 *
protein      1.95771    0.63489   3.084  0.00712 **
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

Figure 4: Output from R on the CARBS dataset.

[QUESTION CONTINUED OVERLEAF]

d. Consider the irrigation dataset from Lectures, where the yield of each of two varieties of crop is measured under one of four irrigation systems. The plot is split into 8 fields, as can be seen from the data in Table 2. The researcher is interested in the effect of irrigation on yield. A mixed model with a random field effect and fixed irrigation and variety effects has been fitted on this data:

$$y_{ij} = \mu_{ij} + \alpha_j + \epsilon_{ij}, \text{ where } \mu_{ij} = \mathbf{X}_{ij}\beta, \text{for } i = 1, 2, \text{ and } j = 1, \ldots, 8$$

The R output is in Figure 5. State the standard assumptions for the distributions of α_j and ϵ_{ij}, assuming the simplest possible covariance structures. Do you agree with the use of a random field effect in this setting? Derive the formula for the correlation between observations from the same field. On the basis of the R output, do you estimate this correlation to be greater or less than 0.5? If a fixed field effect had been employed instead, what would the correlation be?

	field	irrigation	variety	yield		field	irrigation	variety	yield
1	f1	i1	v1	35.4	9	f5	i1	v1	41.6
2	f1	i1	v2	37.9	10	f5	i1	v2	40.3
3	f2	i2	v1	36.7	11	f6	i2	v1	42.7
4	f2	i2	v2	38.2	12	f6	i2	v2	41.6
5	f3	i3	v1	34.8	13	f7	i3	v1	43.6
6	f3	i3	v2	36.4	14	f7	i3	v2	42.8
7	f4	i4	v1	39.5	15	f8	i4	v1	44.5
8	f4	i4	v2	40.0	16	f8	i4	v2	47.6

Table 2: Irrigation data.

```
> lmod <- lmer(yield~irrigation+variety+0+(1|field),data=irrigation)
Linear mixed model fit by REML
Random effects:
 Groups   Name        Variance Std.Dev.
 field    (Intercept) 16.5409  4.067
 Residual              1.4257  1.194
Number of obs: 16, groups: field, 8
Fixed effects:
             Estimate Std. Error t value
irrigationi1   38.425      2.952  13.015
irrigationi2   39.425      2.952  13.354
irrigationi3   39.025      2.952  13.219
irrigationi4   42.525      2.952  14.404
varietyv2       0.750      0.597   1.256
```

Figure 5: R output for a mixed effects model of irrigation data.

[END OF QUESTION 4.]

	EXAMINATION SOLUTIONS 2011-12	Course
Question **1**		Marks & seen/unseen
Part a)	Poisson GLM: for $Y_i \in \mathbb{N}$, we have $Y_i \sim Poi(\mu_i)$, where $\mu_i = \mathbb{E}[Y_i]$ and $\log(\mu_i) = \mathbf{X}_i\beta$, and Y_i is independent from Y_j for all $i \neq j$. The student can either state the Poisson density, or the log-likelihood: $$\mathcal{L}(\mu_i; y_i) = y_i \log \mu_i - \mu_i - \log(y_i!)$$ The saturated model sets $\hat{\mu}_{i,\text{max}} = y_i$, so $$D = 2\left(\mathcal{L}(\hat{\beta}_{\text{max}}; \mathbf{y}) - \mathcal{L}(\hat{\beta}; \mathbf{y})\right) = 2\sum_{i=1}^{n}\left(y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right)$$ The Pearson X^2 is $X^2 = \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i}$, since for the Poisson the variance function is $V(\hat{\mu}_i) = \hat{\mu}_i$ (no proof required).	2 [SEEN] 2 [SEEN] 2 [SEEN]
Part b)	For the first three terms of the Taylor expansion, we need the first two derivatives of $f(y_i) = y_i \log \frac{y_i}{\hat{\mu}_i}$ evaluated at $y_i = \hat{\mu}_i$. This yields: $$f'(y_i) = \log \frac{y_i}{\hat{\mu}_i} + y_i\left(\frac{1}{y_i/\hat{\mu}_i}\right)\frac{1}{\hat{\mu}_i} = 1 + \log \frac{y_i}{\hat{\mu}_i},\ f''(y_i) = \frac{1}{y_i}$$ $$\therefore\ f'(\hat{\mu}_i) = 1 + \log \frac{\hat{\mu}_i}{\hat{\mu}_i} = 1 + 0 = 1,\ f''(\hat{\mu}_i) = \frac{1}{\hat{\mu}_i}$$ $$\therefore\ y_i \log \frac{y_i}{\hat{\mu}_i} \approx (y_i - \hat{\mu}_i) + \frac{1}{2\hat{\mu}_i}(y_i - \hat{\mu}_i)^2$$ $$\therefore\ D = \sum_{i=1}^{n} 2\left(y_i \log \frac{y_i}{\hat{\mu}_i} - (y_i - \hat{\mu}_i)\right) \approx \sum_i \frac{(y_i - \hat{\mu}_i)^2}{\hat{\mu}_i} = X^2$$	4 [UNSEEN]
Part c)	Our standard estimate for ϕ is $\tilde{\phi} = \frac{X^2}{n-p}$. So $$D^{\text{OD}} = \frac{D}{\tilde{\phi}} = \frac{D}{X^2/(n-p)} = (n-p)\frac{D}{X^2} \approx 1,\ \text{using (b).}$$ Therefore the deviance in the quasi-Poisson case is approximately equal to $n - p$ regardless of the fit, and hence not a good indication of fit.	2 [SEEN] 2 [UNSEEN] 1 [UNSEEN]
Part d)	(b) and (e) naturally feature a dispersion parameter, therefore need an F-test. (c) has an additional (over)-dispersion parameter, and hence also needs an F-test. (a) and (d) have no nuisance parameter, and therefore require a χ^2-test.	5 [SEEN] (1 EACH)
	Setter's initials Checker's initials	Page number

	EXAMINATION SOLUTIONS 2011-12	Course
Question **2**		Marks & seen/unseen
Part a)	Canonical form of the log-likelihood for an EF distribution is: $$y_i\frac{\theta}{\alpha_i(\phi)} - \frac{d(\theta)}{\alpha_i(\phi)} + h(y,\phi), \alpha_i(\phi) = \frac{\phi}{w_i}$$ where w_i is a known fixed constant, θ is the canonical parameter, and ϕ a nuisance parameter representing dispersion. We may write $$\theta = \theta(\mu_i),\ \mu_i = \mathbb{E}[Y_i]$$ For the GLM, we additionally require that $\mu_i = \mathbf{X}_i\beta$. For the binomial in terms of u_i, we have $$\mathcal{L}(\theta; u_i) = n_i u_i \log\frac{\mu_i}{1-\mu_i} + n_i\log(1-\mu_i) + \text{constant}$$ where $\mathbb{E}[U_i] = \mu_i$, $\theta(\mu_i) = \log\frac{\mu_i}{1-\mu_i}$, $d(\theta(\mu_i)) = -\log(1-\mu_i)$. The variance function of the binomial is $V(\mu_i) = \mu_i(1-\mu_i)$ (no proof required), and since Var$[U_i] = V(\mu_i)\alpha_i(\phi)$ from lectures (no proof needed), we obtain Var$[U_i] = \frac{1}{n_i}\mu_i(1-\mu_i)$.	2 [SEEN] 2 [SEEN] 2 [SEEN]
Part b)	Fisher information is, for $n_i = m$, $$\mathcal{J}_{jk} = \sum_{i=1}^{n}\frac{x_{ij}x_{jk}}{\frac{1}{m}\mu_i(1-\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 = m\sum_{i=1}^{n}\frac{x_{ij}x_{jk}}{\mu_i(1-\mu_i)}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2$$ So $\mathcal{J}_{jk}$ is <u>increasing with m.</u> The z-values are given by $z_j = \frac{\hat\beta_j}{s.e.(\hat\beta_j)}$, where $s.e.(\hat\beta_j) = \sqrt{\mathcal{J}_{jj}^{-1}}$ (because $\hat\beta \sim N(\beta, \mathcal{J}^{-1})$). So $s.e.(\hat\beta_j) = \frac{1}{\sqrt{m}}(...)$, i.e., <u>decreasing with m</u>, so $\lvert z_j\rvert = \frac{\lvert\hat\beta_j\rvert}{s.e.(\hat\beta_j)}$, hence $\lvert z_j\rvert$ is <u>decreasing with m</u> (since $s.e.(\hat\beta_j) > 0$). Since <u>$P(>\lvert z_j\rvert)$ (probability of observing something as extreme as z_j)</u> is decreasing with $\lvert z_j\rvert$, it is also <u>decreasing with m</u>. This agrees with intuition (more data = small differences are more trustworthy).	1 [UNSEEN] ↑ or ↓: 3 [UNSEEN] formulae for z and p-values: 3 [SEEN] intuition: 1 [SEEN]
	Setter's initials Checker's initials	Page number

	EXAMINATION SOLUTIONS 2011-12	Course
Question **2**		Marks & seen/unseen
Part c)	Table 2 only has a z-test, so we use that, and Age has a p-value of $0.03554 < 0.05$, so Age is significant at 5%. To predict the proportion, $$\hat{\mu}_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}, \text{ where } \hat{\eta}_i = -4.29 + 10 \cdot 0.082$$	2 [SEEN]
Part d)	Using (b), it is clear that Age will remain significant for $m = 100$ if it was significant already for $m = 10$, since p-values are $\downarrow$ with m. As for the prediction, it depends only on $\hat{\beta}$, which is itself independent of m in this case. This can be seen in many ways. Probably the easiest is to show that a factor of m can be extracted from the log-likelihood - so maximising with respect to $\hat{\beta}$ is independent of m. Alternatively, the students may show that $\mathcal{J}^{-1}\mathbf{U}$ is independent of m, where $\mathbf{U}$ is the score vector $\nabla_\beta \mathcal{L}(\beta; \mathbf{y})$ – this implies that the Fisher scoring algorithm does not involve m. This is more tedious, though.	1 [UNSEEN] 1 [SEEN]
Part e)	If Age had been employed as a factor with 5 levels, this would have introduced 4 additional degrees of freedom, so the model would be saturated, hence $$D = 2(\mathcal{L}(\hat{\beta}_{\text{max}}; \mathbf{y}) - \mathcal{L}(\hat{\beta}_{\text{max}}; \mathbf{y})) = 0$$	2 [SEEN]
	Setter's initials Checker's initials	Page number

	EXAMINATION SOLUTIONS 2011-12	Course
Question **3**		Marks & seen/unseen
Part a)	The log-likelihood is $\mathcal{L}(y_i;\mu_i,k) = -k\log(\mu_i) + (k-1)\log(y_i) + k\log k - \frac{ky_i}{\mu_i} - \log\Gamma(k)$ So $\theta(\mu_i) = -\frac{1}{\mu_i}$, $\alpha_i(\phi) = \frac{1}{k}$, where $\phi = \frac{1}{k}$, $w_i = 1$. Therefore the canonical link is $g(\mu) = -\frac{1}{\mu}$. For the saturated model, $\hat{\mu}_{i,\text{max}} = y_i$, so: $\mathcal{L}(\hat{\beta}_{\text{max}};\mathbf{y}) =$ $= \sum_i \left\{ -k\log y_i + (k-1)\log(y_i) + k\log(k) - \frac{ky_i}{y_i} - \log\Gamma(k) \right\}$ So $\phi D = \frac{2}{k}\left(\mathcal{L}(\hat{\beta}_{\text{max}};\mathbf{y}) - \mathcal{L}(\hat{\beta};\mathbf{y})\right)$ $= 2\sum_i \left(-\log y_i - 1 + \log\hat{\mu}_i + \frac{y_i}{\hat{\mu}_i} \right)$ $= 2\sum_i \left(-\log\frac{y_i}{\hat{\mu}_i} + \left(\frac{y_i - \hat{\mu}_i}{\hat{\mu}_i}\right) \right)$ So $r_D(i) = sign(d_i)\sqrt{d_i}$, where $d_i = 2\left(-\log\frac{y_i}{\hat{\mu}_i} + \left(\frac{y_i-\hat{\mu}_i}{\hat{\mu}_i}\right)\right)$.	2 [SEEN] 2 [SEEN] 2 [UNSEEN]
Part b)	The variance function for the Gaussian is constant $V_{\text{Gaussian}}(\mu) = 1$, and for the Gamma it is quadratic $V_{\text{Gamma}} = \mu^2$. Hence: $r_P^{\text{Gamma}}(i) = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i^2},\ r_P^{\text{Gaussian}}(i) = y_i - \hat{\mu}_i.$	2 [SEEN]
Part c)	The squared response residuals should roughly follow the variance function. On this basis, the Gaussian is a poor fit (RHS), since the spread of the residuals around their mean should be constant as we move from left to right, whereas the Gamma (LHS) is a good fit, as the spread shows a roughly quadratic increase, as expected from the distribution's variance function.	2 [SEEN]
	Setter's initials Checker's initials	Page number

	EXAMINATION SOLUTIONS 2011-12	Course
Question **3** Cont'd		Marks & seen/unseen
Part d)	Residuals should be: • zero mean • constant variance • (roughly) independent • approximately Gaussian, or, at least, lying on a smooth line in a QQ-plot Middle plot shows: • zero-mean constant variance for Gamma • non-constant variance, and probably non-zero mean for Gaussian QQ-plot (bottom) shows: • approximate normality for Gamma • departure from normality for Gaussian	2 [SEEN] 2 [SEEN] 2 [SEEN]
Part e)	$\hat{\epsilon}_i^{(-i)} > \hat{\epsilon}_i \Leftrightarrow \hat{\epsilon}_i^{(-i)} - \hat{\epsilon}_i > 0 \Leftrightarrow (y_i - \hat{y}_i^{(-i)}) - (y_i - \hat{y}_i) > 0$ $\Leftrightarrow \hat{y}_i - \hat{y}_i^{(-i)} > 0 \Leftrightarrow g^{-1}(X_i\hat{\beta}) > g^{-1}(X_i\hat{\beta}^{(-i)})$ $\Leftrightarrow X_i\hat{\beta} > X_i\hat{\beta}^{(-i)}$ as required, since g is monotonically increasing. For Figure 3, clearly $\hat{\beta}^{(-10)} < \hat{\beta}$, since taking (y_{10}, X_{10}) into account would increase the slope. So $\hat{\epsilon}_{10}^{(-10)} > \hat{\epsilon}_{10}$. Therefore the jackknife residual is in this case more sensitive to this outlier.	2 [UNSEEN] 2 [UNSEEN]
	Setter's initials Checker's initials	Page number

	EXAMINATION SOLUTIONS 2011-12	Course
Question **4**		Marks & seen/unseen
Part a)	$X_A = (d), X_B = (e), X_C = (i), X_D = (a)$	4 [SEEN]
Part b)	The F-test suggests Age should be deleted. Instead the students may choose to report the t-test from the summary() function, which shows $\hat{\beta}_{\text{Age}}$ is insignificant. Further evidence is that AIC is lowered by removing Age, indicating that the simpler model is preferable.	F/t-test: 1 [SEEN] AIC: 1 [SEEN]
Part c)	40 is <u>not</u> in the confidence interval, but it <u>is</u> in the prediction interval, so we have no evidence to suggest that the proposition is impossible. The prediction interval is always wider because $$s.e._{\text{prediction}}(\hat{y}(\mathbf{X}_{n+1})) = \sqrt{\hat{\sigma}^2_{y\|X_{n+1}} + s.e.(\hat{\mu}(x_{n+1}))}$$ for a new datum $\mathbf{X}_{n+1}$, where $\sigma^2_{y\|\mathbf{X}_{n+1}} > 0$: the prediction interval takes into account both estimation uncertainty and data scatter.	1 [UNSEEN] 2 [SEEN]
Part d)	The (two-way anova) simplest mixed linear model assumes $$\alpha_j \sim N(0, \sigma^2_\alpha),\ \epsilon_{ij} = N(0, \sigma^2_\epsilon)$$ where all of the α_js and ϵ_{ij}s are independent. A random effect is appropriate: • to reflect the constraint on randomisation imposed by the fact that we can use one irrigation system per field • since field effects are of no interest *per se* • because the mixed model is more efficient (only two observations per covariate pattern otherwise) • because we can interpret the observed field levels to arise out of a 'population' of fields (nearby pieces of land) The student need not give all four reasons - but efficiency must be mentioned. Note also that a large proportion of the variance is explained by the random effect, so that is an additional reason to use random effects here, as opposed to disregarding the field factor. CONTINUED OVERLEAF	2 [SEEN] 3 [SEEN]
	Setter's initials Checker's initials	Page number

Question **4** cont'd | Marks & seen/unseen

Part d) cont'd

First we state the definition of correlation:

$$Corr(y_{1j}, y_{2j}) = \frac{\mathbb{E}\left[(y_{1j} - \mathbb{E}[y_{1j}])(y_{2j} - \mathbb{E}[y_{2j}])\right]}{\sqrt{Var[y_{1j}]Var[y_{2j}]}}$$

$Corr(y_{1j}, y_{2j})$: 3 [SEEN]

Now clearly $Var[y_{1j}] = Var[y_{2j}] = \sigma_\alpha^2 + \sigma_\epsilon^2$. So the denominator is given by $\sigma_\alpha^2 + \sigma_\epsilon^2$. Since $y_{ij} - \mathbb{E}[y_{ij}] = \alpha_j + \epsilon_{ij}$, the nominator is

$$\begin{aligned}&\mathbb{E}\left[(y_{1j} - \mathbb{E}[y_{1j}])(y_{2j} - \mathbb{E}[y_{2j}])\right] = \mathbb{E}[(\alpha_j + \epsilon_{1j})(\alpha_j + \epsilon_{2j})] \\ &= \mathbb{E}[\alpha_j^2] + \mathbb{E}[\alpha_j(\epsilon_{1j} + \epsilon_{2j})] + \mathbb{E}[(\epsilon_{1j}\epsilon_{2j})], \text{ but } \epsilon_{ij} \text{ is indnt of } \alpha_j, \text{ so} \\ &= \mathbb{E}[\alpha_j^2] = \sigma_\alpha^2\end{aligned}$$

Hence $Corr(y_{1j}, y_{2j}) = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}$.

If a fixed effect had been employed, $\sigma_\alpha^2 = 0$, so the correlation itself is 0 (alternative explanations are possible: e.g., with fixed effects only, observations are independent). — 1 [SEEN]

Given the estimated variances by R, $\hat{\rho} = \frac{16.5409}{16.5409 + 1.4257} > 0.5$. — 2 [SEEN]