M3/M4
S2

Imperial College London

BSc, MSci and MSc EXAMINATIONS (MATHEMATICS)

May – June 2013

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

Statistical Modelling II

Date: Friday, 17 May 2013. Time: 10.00am. Time allowed: 2 hours.
This paper has FOUR questions.
Candidates should use ONE main answer book.
Supplementary books may only be used after the main book is full.
Statistical tables will not be provided.

- DO NOT OPEN THIS PAPER UNTIL THE INVIGILATOR TELLS YOU TO.
- Answer all the questions. Each question carries equal weight.
- Credit will be given for all questions attempted, but extra credit will be given for complete or nearly complete answers.
- Affix one of the labels provided to each answer book that you use, but DO NOT USE THE LABEL WITH YOUR NAME ON IT.
- Calculators may not be used.

1. Consider the dataset in Figure 1 on weight at birth for babies of different gender and different *gestational age* (i.e., duration of pregnancy). The gestational age is given in weeks ('Age'), as well as in *completed months* of gestation ('Age.Months'). Denote the birth weight of the ith individual by y_i, their age in weeks by a_i, their age in months by m_i, and their sex by g_i.

   a. The following model was fit in R:

   ```
   > M.a <- lm(Birth.Weight ~ Age + 0 + Age.Months)
   ```

   Let $\hat{\beta}_a$ be the regression coefficient of Age in this model, and $\hat{\beta}_m$ the coefficient of 'Age.Month'. Figure 2 contains two possible plots for the 95% confidence region of $\hat{\beta}_a$ and $\hat{\beta}_m$, one of which is correct. By considering the covariance between $\hat{\beta}_a$ and $\hat{\beta}_m$, determine which one is the correct plot. (Hint: you do not need to use the actual data values, other than to note that both Age and Age.Months are always positive).

   b. The model summary, as well as its analysis of variance results, are listed below:

   ```
   > summary(M.a)
   Call:
   lm(formula = Birth.Weight ~ Age + 0 + Age.Months)

   Residuals:
       Min      1Q  Median      3Q     Max
   -368.50 -157.88    4.09  132.21  372.99

   Coefficients:
              Estimate Std. Error t value Pr(>|t|)
   Age           56.10      36.33   1.544    0.137
   Age.Months    95.11     164.61   0.578    0.569

   Residual standard error: 203.9 on 22 degrees of freedom
   Multiple R-squared: 0.9957,       Adjusted R-squared: 0.9953
   F-statistic:  2553 on 2 and 22 DF,  p-value: < 2.2e-16
   > anova(M.a)
   Analysis of Variance Table

   Response: Birth.Weight
              Df    Sum Sq   Mean Sq   F value Pr(>F)
   Age         1 212270368 212270368 5105.3574 <2e-16 ***
   Age.Months  1     13880     13880    0.3338 0.5693
   Residuals  22    914715     41578
   ---
   Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
   ```

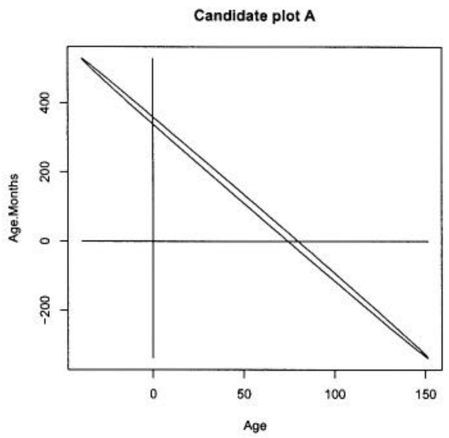
   Which covariates, if any, do you recommend should be removed from the model, on the basis of the output above? Is this recommendation consistent with the confidence region of Figure 2? Is it consistent with the meaning of the covariates?

[QUESTION 1 CONTINUED OVERLEAF.]

	Age	Birth.Weight	Sex	Age.Months
1	40	2968	boy	9
2	38	2795	boy	8
3	40	3163	boy	9
4	35	2925	boy	8
5	36	2625	boy	8
6	37	2847	boy	8
7	41	3292	boy	9
8	40	3473	boy	9
9	37	2628	boy	8
10	38	3176	boy	8
11	40	3421	boy	9
12	38	2975	boy	8

	Age	Birth.Weight	Sex	Age.Months
13	40	3317	girl	9
14	36	2729	girl	8
15	40	2935	girl	9
16	38	2754	girl	8
17	42	3210	girl	9
18	39	2817	girl	9
19	40	3126	girl	9
20	37	2539	girl	8
21	36	2412	girl	8
22	38	2991	girl	8
23	39	2875	girl	9
24	40	3231	girl	9

Figure 1: Data on birth weight from Question 1.



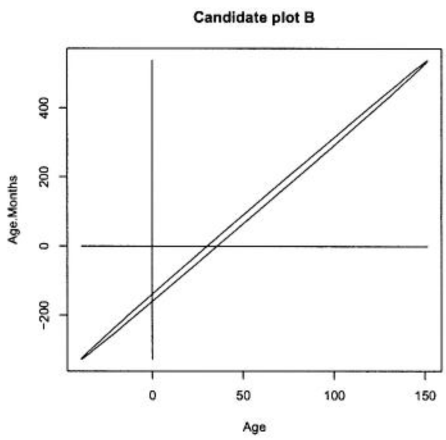


Figure 2: Two candidate plots for the confidence region of the regression coefficients of M.a.

[QUESTION 1 CONTINUED OVERLEAF.]

1. c. A second model is now fit, as follows:

```
> M.b <- lm(Birth.Weight ~ Age + Sex)
> summary(M.b)

Call:
lm(formula = Birth.Weight ~ Age + Sex)

Residuals:
    Min      1Q  Median      3Q     Max
-257.49 -125.28  -58.44  169.00  303.98

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -1610.28     786.08  -2.049   0.0532 .
Age           120.89      20.46   5.908 7.28e-06 ***
Sex girl     -163.04      72.81  -2.239   0.0361 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 177.1 on 21 degrees of freedom
Multiple R-squared:  0.64,     Adjusted R-squared: 0.6057
F-statistic: 18.67 on 2 and 21 DF,  p-value: 2.194e-05
```

Write down in a non-redundant form the design matrix, X, employed by this model, ensuring that the order of the columns agrees with the order of the estimated coefficients above. Using the output above, what is the expected birth weight of a female baby born after 40 weeks of gestation? You may use round numbers throughout, and do not need to perform simplifying arithmetic operations.

d. The variance-covariance matrix of the regression coefficient estimates is given by:

```
            (Intercept)    Age Sex girl
(Intercept)      617918 -16051     4074
Age              -16051    419     -174
Sex girl           4074   -174     5301
```

Define what a prediction interval (PI) is. Derive the standard error underlying the PI for the birth weight of a female baby born after 40 weeks of gestation under model M.b in terms of relevant parts of the above numerical output. Your answer must be a formula involving only numbers, but you do not need to perform any simplifying arithmetic operations.

[END OF QUESTION 1.]

	k	n	x
1	4	32	weekday
2	4	32	weekday
3	3	15	weekend
4	2	5	weekend
5	10	40	weekend

Table 1: Data for question 2.

Distribution	90% quantile	95% quantile	97.5% quantile
$N(0,1)$	1.28	1.65	1.96
t_4	1.53	2.13	2.78
t_3	1.64	2.35	3.18
t_2	1.89	2.92	4.30
t_1	3.10	6.31	12.70

Table 2: Quantiles of Gaussian and Student t

2. Consider the log-likelihood of an exponential family model, given in the form:

$$y\frac{\theta}{\alpha(\phi)} - \frac{d(\theta)}{\alpha(\phi)} + h(y,\phi),\ \alpha(\phi) = \frac{\phi}{w}$$

a. Prove that:

$$\mathbb{E}_\theta[Y] = d'(\theta),\ \mathsf{Var}_\theta[Y] = \alpha(\phi)d''(\theta) \tag{1}$$

where, letting $U_\theta = \frac{\partial \mathcal{L}(\theta;Y)}{\partial\theta}$, you may assume without proof the following:

$$\mathbb{E}_\theta[U_\theta] = 0, \text{ and } \mathbb{E}_\theta[U_\theta U_\theta^T] = -\mathbb{E}_\theta[\nabla_\theta U_\theta]$$

b. Write down the log-likelihood of the binomial model in the form above, letting $y = k/n$ be the random variable, where k is the number of successes in n trials. The formula for the ith Pearson residual in the presence of weights is given by:

$$r_i^{(P)} = \frac{y_i - \hat{\mu}_i}{\sqrt{V(\hat{\mu}_i)}}\sqrt{w_i}$$

where $V(\mu)$ is the variance function of the model. State the variance function for the binomial model, and hence derive the formula for the Pearson residuals in this case.

c. Consider the data in Table 1, consisting of successful attempts to contact households in a phone survey, as a function of whether the call was placed on a weekday, or over the weekend. Assuming a quasi-binomial model, prove that the standard estimate $\tilde{\phi}$ of the dispersion parameter ϕ for this dataset is $4/15$.

d. The coefficient estimate divided by the standard error for the weekend effect according to a canonical binomial model without an overdispersion parameter yields the value -1.76. Is this significant at the 5% level? Does the quasi-Poisson model agree? You may use Table 2 and the rough approximation $\sqrt{4/15} \approx 0.5$ to answer this question.

[END OF QUESTION 2.]

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	60.24	0.1458	413.2430	0.0000
operator2	-0.18	0.2062	-0.8731	0.3955
operator3	0.38	0.2062	1.8433	0.0839
operator4	0.44	0.2062	2.1343	0.0486

Figure 3: Model output for treatment contrasts for Question 3

3. Consider the pulp dataset, described in lectures, containing five observations of the brightness level of paper pulp, for each of four possible levels of the factor `operator` which describes the engineering process.

a. What is a balanced design? Is the design balanced in this dataset?

b. Let $\alpha_1, \alpha_2, \alpha_3, \alpha_4$ denote the group-specific expected values. Consider a linear Gaussian model of the brightness as a function of an intercept and of the unordered factor *operator* using treatment contrasts. The R output for the estimated coefficients is in Figure 3, each line performing a t-test for the hypothesis that the respective coefficient is non-zero.

- identify which of the four tests yields a significant result at the 5% level, and express the respective null hypothesis that was rejected as a statement about the α_js
- report the MLE estimate $\hat{\alpha}_2$ of α_2
- what does the intercept under sum contrasts represent? Show that its MLE is 60.4.

c. Consider the following random effects model:

$$y_{ij} = \mu_j + \epsilon_{ij},\ \epsilon_{ij} \sim N(0, \sigma_\epsilon^2),\ \mu_j \sim N(0, \sigma_\mu^2),\ i = 1, \ldots, n,\ j = 1, \ldots, K$$

and the following statistics:

$$\bar{y}_{\cdot,j} = \frac{1}{n}\sum_{i=1}^{n} y_{ij}, \quad \bar{y} = \frac{1}{nK}\sum_{i=1}^{n}\sum_{j=1}^{K} y_{ij}$$

$$\mathsf{SSE} = \sum_{j=1}^{K}\sum_{i=1}^{n}(y_{ij} - \bar{y}_{\cdot,j})^2, \quad \mathsf{SSA} = \sum_{j=1}^{K}\sum_{i=1}^{n}(\bar{y}_{\cdot,j} - \bar{y})^2, \quad \mathsf{SST} = \sum_{i=1}^{n}\sum_{j=1}^{K}(y_{i,j} - \bar{y})^2$$

Provide without proof unbiased estimators for σ_ϵ^2 and σ_μ^2 in terms of the above statistics. Hence prove the following formula:

$$\mathbb{E}[\mathsf{SST}] = (K-1)n\sigma_\mu^2 + (Kn-1)\sigma_\epsilon^2$$

[Hint: show that SST = SSE + SSA and take expectations.]

[END OF QUESTION 3.]

4. Consider the following GLM of the probability μ of a patient surviving after being given a dose X of a certain poison, where $X \in \mathbb{R}^{+}$:

$$g(\mu) = \beta_0 + \beta_1 X$$

a. State two invertible functions $g()$ for which $g^{-1} : \mathbb{R} \to [0, 1]$

b. Let $\hat{\beta}_1$ and $\hat{\beta}_0$ be estimates of β_1 and β_0 given a certain dataset $(y_i, x_i)_{i=1:n}$. Write down the formula for the estimated effective dose for 20% probability of survival, in terms of $\hat{\beta}$ and $g()$.

c. Assuming that $g()$ is the logistic function, reformulate the model in terms of odds of survival, rather than probability of survival.

d. Assume that g is the logistic, that the effective dose for 20% probability of survival is 2.34, and that $\hat{\beta}_1 = ln(2)$. Calculate the probability of survival for a dose of 4.34.

[END OF QUESTION 4.]

	EXAMINATION SOLUTIONS 2012-13	Course S2
Question **1**		Marks & seen/unseen
Part a)	The design matrix of this model is $\mathbf{X} = (\mathbf{X}_a \mid \mathbf{X}_b)$, where X_a is an $n \times 1$ vector containing the values of 'Age', and X_m the respective values of 'Age.Months'. Whether the confidence region agrees with plot A or plot B will be determined by the covariance of $\hat{\beta}_a$ and $\hat{\beta}_m$. We first write down the Fisher information matrix: $$J = X^T X = \begin{pmatrix} X_a^T X_a & X_a^T X_m \\ X_m^T X_a & X_m^T X_m \end{pmatrix}$$ $$J_{12}^{-1} = \frac{-X_m^T X_a}{\lvert J \rvert} < 0, \text{ since } X_m, X_a > 0$$ Therefore, plot A is correct, since the two regression coefficient estimates are negatively correlated.	seen 1 part seen 2 2 1 =6
Part b)	The t-tests reported in `summary()` suggest that either covariate may be removed, if the other one remains in the model. The confidence region agrees with this conclusion, since the region intersects both axes, but does not contain $\hat{\beta}_a = \hat{\beta}_m = 0$. Consequently, deleting both covariates is not advised, but deleting either one of them is recommended. The output of `anova()` agrees with the results of the confidence region. This also agrees with our intuitive understanding of the covariates' meaning: they both represent the same piece of information, so including both introduces duplication, whereas deleting both removes useful information.	seen 1 seen 1 seen 1 unseen 1 =4
Part c)	Let $\mathbf{1}$ be an $n \times 1$ vector of 1s, and X_g be an $n \times 1$ vector containing $n/2$ 0s followed by $n/2$ 1s (i.e,. the indicator function of being a girl). The design matrix in question is then: $$X = (\mathbf{1} \mid X_a \mid X_g)$$ Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3)$ correspond to the estimated coefficient values of the R output. The datapoint we are required to predict is given by $X_{n+1} = (1, 40, 1)$, so: $$\hat{y}_{n+1} = X_{n+1}\hat{\beta} = \hat{\beta}_1 + 40\hat{\beta}_2 + \hat{\beta}_3 = -1610 + 40 \cdot 121 - 163 = 3067$$ Arithmetic errors are not penalised.	seen 1 seen 1 seen 1 1 =4
	Setter's initials Checker's initials	Page number

	EXAMINATION SOLUTIONS 2012-13	Course
Question 1		Marks & seen/unseen
Part d)	A prediction interval is an estimated interval in which future observations will fall, with a certain probability. It is based on the following prediction standard error:	seen 1
	$$\sigma_{PI} = \sqrt{\sigma_{Y\|X} + \mathsf{Var}[\hat{\mu}_{n+1}]}$$	seen 1
	We may estimate the residual variance by the standard formula:	
	$$\tilde{\sigma}_{Y\|X} = \sqrt{\mathsf{RSS}/(n-p)} = 177.1$$	seen 1
	We may also estimate $\mathsf{Var}[\hat{\mu}_{n+1}]$ as follows:	
	$$\tilde{\mathsf{Var}}[\hat{\mu}_i] = \tilde{\mathsf{Var}}[X_{n+1}\hat{\beta}] = X_{n+1}\tilde{\mathsf{Var}}[\hat{\beta}]X_{n+1}^T$$	seen 1
	We then have: $$se_{\mathsf{PI}} = \sqrt{\tilde{\mathsf{Var}}[\hat{\mu}_i] + \tilde{\sigma}_{Y\|X}}$$	seen 1
	and $$\tilde{\mathsf{Var}}[\hat{\mu}_i] = (1,40,1)\begin{pmatrix} 617918 & -16051 & 4074 \\ -16051 & 419 & -174 \\ 4074 & -174 & 5301 \end{pmatrix}(1,40,1)^T$$	1 =6
	Setter's initials Checker's initials	Page number

<table>
<tr><td></td><td>EXAMINATION SOLUTIONS 2012-13</td><td>Course</td></tr>
<tr><td>Question 2</td><td></td><td>Marks & seen/unseen</td></tr>
<tr><td>Part a)</td><td>We have:

$$U_\theta = \frac{\partial \mathcal{L}(\theta; Y)}{\partial \theta} = \frac{Y}{\alpha(\phi)} - \frac{d'(\theta)}{\alpha(\phi)}$$
$$\therefore\ \mathbb{E}_\theta[U_\theta] = \frac{\mathbb{E}_\theta[Y]}{\alpha(\phi)} - \frac{d'(\theta)}{\alpha(\phi)} = 0$$
$$\therefore\ \mathbb{E}_\theta[Y] = d'(\theta)$$

We now replace $d'(\theta) = \mathbb{E}_\theta[Y]$ in the expression for U_θ in $\mathcal{J}_\theta$:

$$\mathbb{E}_\theta[U_\theta U_\theta^T] = \mathbb{E}_\theta\left[\frac{1}{\alpha(\phi)^2}\left(Y - d'(\theta)\right)^2\right] =$$
$$= \mathbb{E}_\theta\left[\frac{1}{\alpha(\phi)^2}\left(Y - \mathbb{E}_\theta[Y]\right)^2\right] = \frac{1}{\alpha(\phi)^2}\mathsf{Var}_\theta[Y]$$

and we also have:

$$-\mathbb{E}_\theta\left[\nabla_\theta U_\theta\right] = -\mathbb{E}\left[-\frac{d''(\theta)}{\alpha(\phi)}\right] = \frac{d''(\theta)}{\alpha(\phi)}$$

Equating the two given expressions yields:

$$\mathsf{Var}_\theta[Y] = \alpha(\phi) d''(\theta)$$</td><td>all seen

2

2

2

1 =7</td></tr>
<tr><td>Part b)</td><td>The log-likelihood is given by:

$$\mathcal{L} = ny\log\left(\frac{\mu}{1-\mu}\right) + n\log(1-\mu) + \text{ constant.}$$

The variance function for the binomial is given by $V(\mu) = \mu(1-\mu)$. Therefore, the formula for the ith Pearson residual is:

$$r_i^{(P)} = \frac{y_i - \hat{\mu}_i}{\hat{\mu}_i(1-\hat{\mu}_i)}\sqrt{n_i}$$</td><td>1

1

1 =3</td></tr>
<tr><td></td><td>Setter's initials   Checker's initials</td><td>Page number</td></tr>
</table>

Question 2		Marks & seen/unseen
Part c)	First, we compute the fitted values by grouping by covariate pattern: $$\hat{\mu}_1 = \hat{\mu}_2 = \frac{4+4}{32+32} = 0.125,\ \hat{\mu}_3 = \hat{\mu}_4 = \hat{\mu}_5 = \frac{3+2+10}{15+5+40} = 0.25$$ We will use the following formula to estimate the dispersion parameter: $$\tilde{\phi} = \frac{1}{n-p}\sum_{i=1}^{5}\frac{n_i(y_i-\hat{\mu}_i)^2}{\mu_i(1-\mu_i)}$$ The Pearson residuals for the first two obseravtions, as well as the last observation, are exactly 0. The others yield: $$\tilde{\phi} = \frac{1}{5-2}\left(\frac{15\,(3/15-1/4)^2+5\,(2/5-1/4)^2}{1/4(1-1/4)}\right) = \frac{1}{3}\left(\frac{15\frac{1}{400}+5\frac{9}{400}}{3/16}\right) = \frac{1}{3}\left(\frac{\frac{6}{40}}{3/16}\right) = \frac{1/20}{3/16} = \frac{16}{60} = \frac{4}{15},$$ as required.	part unseen 1 seen 1 1 1 =4
Part d)	Let $\hat{\beta}_W$ be the coefficient representing the weekend effect. For the Binomial model, the test forms the z-value $\hat{\beta}_W/\text{se}(\hat{\beta}_W)$ first, and then performs a two-tailed normal test: $$2\cdot(1-p_N(z)) < 0.05 \Leftrightarrow z > q_N(0.975)$$ where p_N is the Normal $N(0,1)$ probability distribution, and q_N is its quantile function. By referring to Table 2, we can see that $\lvert z\rvert = \lvert -1.76\rvert < q_N(0.975) = 1.96$, so the intercept is not significant at the 5% level. For the quasi-Poisson model, note that the estimated coefficient remains identical, whereas the s.e. is inflated by $\sqrt{\tilde{\phi}}$: $$\text{se}_Q(\hat{\beta}) = \text{se}(\hat{\beta} \mid \phi = 1)\sqrt{\tilde{\phi}}$$ Consquently $\hat{\beta}_W/\text{se}_Q(\hat{\beta}_W)$ is given by multiplying the previous standard error by $1/\sqrt{\phi} \approx 2$. Moreover, we will need to compare to a t distribution with 3 degrees of freedom. So $\lvert t\rvert \approx \lvert -1.76\cdot 2\rvert = 3.52 > q_{t,3}(0.975) = 3.18$, and the intercept becomes significant.	seen 1 1 1 (coefficients identical) 1 1 1 =6
	Setter's initials Checker's initials	Page number

	EXAMINATION SOLUTIONS 2012-13	Course
Question **3**		Marks & seen/unseen
Part a)	A balanced design is one where each combination of factor levels has been observed the same number of times. The design in question is balanced because we have 5 observations per factor level.	2 2 =4
Part b)	The first and last lines are significant, and they represent, respectively, the following null hypotheses: • $\alpha_1 = 0$ • $\alpha_4 = \alpha_1$ The value $\hat{\alpha}_2$ is given by $\hat{\beta}_1 + \hat{\beta}_2 = 60.06$. The intercept under sum contrasts represents the grand mean $\frac{1}{4}(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4)$, which is given by: $\begin{aligned} &\frac{1}{4}(\alpha_1 + \alpha_2 + \alpha_3 + \alpha_4) = \\ &= \frac{1}{4}\hat{\beta}_1 + \frac{1}{4}\left(\hat{\beta}_1 + \hat{\beta}_2\right) + +\frac{1}{4}\left(\hat{\beta}_1 + \hat{\beta}_3\right) + +\frac{1}{4}\left(\hat{\beta}_1 + \hat{\beta}_4\right) \\ &= \hat{\beta}_1 + \frac{1}{4}\hat{\beta}_2 + \frac{1}{4}\hat{\beta}_3 + \frac{1}{4}\hat{\beta}_4 = \\ &= 60.24 + \frac{-0.18 + 0.38 + 0.44}{4} = 60.24 + \frac{0.64}{4} = 60.4 \end{aligned}$	part seen 1 1 (unseen) 1 1 2 =6
	Setter's initials Checker's initials	Page number

Part c)

The estimators given in lectures are: seen

$$\hat{\sigma}^2_\mu = \frac{\text{MSA} - \text{MSE}}{n} = \frac{(K-1)\text{SSA} + K(n-1)\text{SSE}}{n},$$
$$\hat{\sigma}^2_\epsilon = \text{MSE} = \frac{\text{SSE}}{K(n-1)}$$ 2

We now show that:

$$SST = SSA + SSE$$

This follows from the following derivation:

$$\sum_{i=1}^{n}\sum_{j=1}^{K}(y_{i,j} - \bar{y})^2 = \sum_{i=1}^{n}\sum_{j=1}^{K}(y_{i,j} - \bar{y}_{.,j} + \bar{y}_{.,j} - \bar{y})^2 =$$
$$= SSA + SSE + 2\sum_{i=1}^{n}\sum_{j=1}^{K}(y_{i,j} - \bar{y}_{.,j})(\bar{y}_{.,j} - \bar{y})$$ 2

But the rightmost summand is in fact zero, because:

$$\sum_{i=1}^{n}\sum_{j=1}^{K}(y_{i,j} - \bar{y}_{.,j})(\bar{y}_{.,j} - \bar{y}) = \sum_{j=1}^{K}(\bar{y}_{.,j} - \bar{y})\sum_{i=1}^{n}(y_{i,j} - \bar{y}_{.,j})$$ 2

and

$$\sum_{i=1}^{n}(y_{i,j} - \bar{y}_{.,j}) = \sum_{i=1}^{n} y_{i,j} - \sum_{i=1}^{n}\bar{y}_{.,j} = n\bar{y}_{.,j} - n\bar{y}_{.,j} = 0$$

Taking expectations and using unbiasedness, we have that $\mathbb{E}[\text{MSE}] = \sigma^2_\epsilon$, and that $\mathbb{E}[\text{MSA}] = n\sigma^2_\mu + \sigma^2_\epsilon$. This completes the proof, since 2

$$\mathbb{E}[SST] = \mathbb{E}[SSA] + \mathbb{E}[SSE] = (K-1)\mathbb{E}[MSA] + K(n-1)\mathbb{E}[MSE] =$$
$$= (K-1)(n\sigma^2_\mu + \sigma^2_\epsilon) + K(n-1)\sigma^2_\epsilon$$
$$= (K-1)n\sigma^2_\mu + (Kn-1)\sigma^2_\epsilon$$ 2

=10

	EXAMINATION SOLUTIONS 2012-13	Course
Question **4**		Marks & seen/unseen
Part a)	Two such choices include: • the logistic transformation, $g(x) = \log \frac{x}{1-x}$ whose inverse is $g^{-1}(y) = \frac{e^y}{1+e^y}$, satisfies the property in question since $0 < e^y < 1 + e^y$, and $g^{-1}(y)$ is defined for all y. • the complementary log-log function $g(x) = \log(-\log(1-x))$ has the inverse $g^{-1}(y) = 1 - e^{-e^y}$, which satisfies the property as it's defined for all y, and $-e^y < 0$, which implies $0 < e^{-e^y} < 1$ (since $e^z < 1$ whenever $z < 0$).	seen 3 part seen 4 =7
Part b)	The estimated effective dose $X^\star$ for 20% probability of survival is: $$X^\star = \frac{g(0.2) - \hat{\beta}_0}{\hat{\beta}_1}$$	seen 2 =2
Part c)	The odds of survival o are given by $$o = \frac{\mu}{1-\mu}$$ Consequently, if $g(\mu) = \log \frac{\mu}{1-\mu}$, we may write: $$\log o = \beta_0 + \beta_1 X \ \therefore\ o = e_0^\beta (e_1^\beta)^X$$	seen 2 2 =4
Part d)	Using our equation of the model in terms of odds and the given effective dose, we have: $$\hat{o} = \frac{0.2}{0.8} = e^{\hat{\beta}_0}(e^{\hat{\beta}_1})^2.34$$ Then, $\hat{o}^\star$ for $X^\star = 4.34$ will be given by: $$\hat{o}^\star = e^{\hat{\beta}_0}(e^{\hat{\beta}_1})^{(2.34+2)} = e^{\hat{\beta}_0}(e^{\hat{\beta}_1})^{2.34}(e^{\hat{\beta}_1})^2$$ Since $\hat{\beta}_1 = \ln(2)$, $(e^{\hat{\beta}_1})^2 = 2^2 = 4$, so: $$\hat{o}^\star = e^{\hat{\beta}_0}(e^{\hat{\beta}_1})^{2.34}4 = \hat{o} \cdot 4 = \frac{0.2}{0.8} \cdot 4 = 1$$ So the probability of survival is 0.5.	part seen 2 2 1 1 1 =7
	Setter's initials Checker's initials	Page number