May 2014

Ph.D. COMPREHENSIVE EXAMINATIONS
DEPARTMENT OF STATISTICAL SCIENCES
UNIVERSITY OF TORONTO

APPLIED STATISTICS COMPREHENSIVE EXAMINATION
May 9, 2014 12:30 p.m. – 4:30 p.m.
Sidney Smith Hall

- ***Attempt all questions*** (total # of questions = 5). (total # of pages = 14 including cover page)

- Please work neatly and legibly.

- ***Start each question in a new book, with your name and the number of the question on the front cover.*** If there is more than one book for a question, then also indicate which is the first book and which second, e.g., Jane Smith, Question 5, Book 1 of 2.

- The questions are not in any special order, nor are they all of equal difficulty.

- The problems may be improperly phrased or may contain a misprint. Should this happen, reflect it in your discussion. Faculty members are *not* available to answer questions during the exam.

- You are NOT permitted any aids (e.g., books, notes, etc.) **aside from a single non-programmable calculator**.

- Good luck!

1. Members of a Senior Kindergarten class (which we shall treat as a sample) try to zip their coats within one minute. We we observe whether each child succeeds or fails. A natural model is $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} B(1, \theta)$, where θ is the probability of success. But obviously, the probability of success is not the same for each child. Some are almost certain to succeed, and others have almost no chance. Thus a more reasonable model is that $Y_1, \ldots, Y_n$ are independent random variables, with $Y_i \sim B(1, \theta_i)$.

   This is a two-stage sampling model. First, sample from a population in which each child has a personal probability of success. Then for child i, use θ_i to generate success or failure. Note that in this formulation, $\theta_1, \ldots, \theta_n$ are random variables with some probability distribution supported on $[0, 1]$. For convenience, suppose this distribution is Beta with parameters $\alpha > 0$ and $\beta > 0$. The beta density is

$$f(\theta) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha - 1}(1 - \theta)^{\beta - 1}.$$

   Thus, the parameters of this problem are α and β.

   (a) The Bernoulli probability of success is a conditional probability. Find the unconditional probability $Pr\{Y_i = 1\}$. Show your work.

   (b) Using your answer above, write the likelihood. Again, it is a function of α and β.

   (c) Try to obtain the maximum likelihood estimates of α and β in the usual way. What happens? *Please comment.*

   (d) Geometrically, how does the likelihood function look? Make a rough sketch.

   (e) Is consistent estimation of the pair (α, β) possible? Clearly answer Yes or No and explain your answer. Provide a proof if you can.

   (f) Is there a meaningful *function* of α and β that can be estimated consistently? In your answer, make sure to state what the function is, why it's meaningful, and what estimator you would use.

2. The University of Toronto media room posted a report on a recently published study in the *Journal of Aggression, Maltreatment, and Trauma*: the headline of the media post was "Thirty per cent of adults with attention deficit disorder report childhood physical abuse".

   We first provide a summary of the study, taken from the published paper.[1] The data analyzed was a subset of the responses to the 2005 Canadian Community Health Survey (CCHS). This is a cross-sectional survey of a nationally representative sample of Canadians; in 2005 approximately 130,000 Canadians were interviewed. The current study is a secondary analysis of data from respondents 18 years or older, living in Manitoba or Saskatchewan. The variables included in the statistical analysis were age, gender, race, and the answers to two questions: (i): "(Remember, we are interested in conditions diagnosed by a health professional.) Do you have a learning disability?"; responses included attention deficit disorder (ADD), attention deficit hyperactive disorder (ADHD), dyslexia, and other; (ii): "Were you ever physically abused by someone close to you?" [before age 18]; responses were yes/no. All questions also included 'don't know', and 'refuse to answer'; there were 586 responses in this category, leaving a final sample size of 13,054.

   For the study, responses of interest were: (i) ADD or ADHD (yes, no) and (ii) physical abuse (yes, no). The results were:

		ADD/ADHD		
		Yes	No	
Childhood	Yes	19	935	954
Abuse	No	45	12,055	12,100
Total		64	12,990	13,054

   The authors used logistic regression, with ADD/ADHD as the response, and physical abuse as one of the explanatory variables. The other explanatory variables were age, gender and race. The results were reported as odds ratios.

   (a) Based on the table above, which ignores the other covariates, what is the estimated odds of ADD/ADHD in the abuse = "yes" group, relative to the odds in the abuse = "no" group? What is the estimated probability of ADD/ADHD in the full sample?

   (b) Writing y_i for the response of subject i, and x_i for the corresponding vector of explanatory variables (`abuse`, `age`, `gender`, `race`), the probability model used for fitting was

$$y_i \sim Bernoulli(p_i), \quad \log\{p_i/(1-p_i)\} = x_i^T\beta, i = 1, \ldots, n \tag{1}$$

[1] Esme Fuller-Thomson , Rukshan Mehta & Angela Valeo (2014) Establishing a Link Between Attention Deficit Disorder/Attention Deficit Hyperactivity Disorder and Childhood Physical Abuse, Journal of Aggression, Maltreatment & Trauma, 23:2, 188-198, DOI: 10.1080/10926771.2014.873510

with responses assumed independent across subjects. Age was treated as a continuous covariate; the others are factor variables with two levels each. The estimated coefficients and their estimated standard errors are given below:

variable	$\hat{\beta}_j$	$\widehat{s.e.}(\hat{\beta}_j)$	z-statistic
`age (years)`	-0.105	0.011	-9.54
`sex (male)`	0.727	0.234	3.11
`race (white)`	-0.041	0.275	-0.15
`abuse (yes)`	1.881	0.251	7.49

Use this information to provide an approximate 95% confidence interval for the odds of ADD/ADHD for those who had been physically abused relative to those who had not. Summarize the information in this whole table in one or two sentences, in non-technical language.

(c) Since both ADD/ADHD and abuse are responses to the survey, the authors might have carried out a logistic regression with abuse as the response, and ADD/ADHD as one of the explanatory variables. Why do you think they did their analysis instead?

(d) In the media report, one of the co-authors stated: "Our data do not allow us to know the direction of the association". Explain why this is the case, and suggest two interpretations of the result in light of this statement.

(e) For the model used in (1), write down the log-likelihood function based on the full data-set, and show that $\Sigma y_i x_i^T = \Sigma p_i(\hat{\beta}) x_i^T$ defines the maximum likelihood equations. How are the estimated standard errors of $\hat{\beta}_j$ determined?

(f) The residual deviance after fitting this model is defined by

$$D = 2\sum_{i=1}^{n}\{\ell(\tilde{p}; y) - \ell(\hat{p}; y)\}, \tag{2}$$

where $p = (p_1, \ldots, p_n), y = (y_1, \ldots, y_n)$, $\tilde{p}$ is determined by maximizing the log-likelihood function by ignoring the dependence of p_i on x_i, and $\hat{p} = (p_1(\hat{\beta}), \ldots, p_n(\hat{\beta}))$ is determined by the equation in Question (2e). Can this residual deviance be used to test the fit of model (1)? Why or why not? *Hint*: Show that $\tilde{p}_i = y_i$ and find an expression for D.

(g) Suppose we were interested in the *risk difference*, i.e. the difference between the probability of ADD/ADHD, in the `abuse = 1` group and the probability of ADD/ADHD in the `abuse = 0` group, adjusted for the other explanatory variables. How would you estimate this difference and its standard error? What is an advantage of reporting a risk difference rather than an odds ratio?

3. It is perfectly natural to assume that something like response to a drug might be approximately linear over some range of dosage values, but that each person in the population might have his or her own slope. Thus each time you select a random sample you'll get a different collection of slopes, and the regression coefficient corresponding to the slope would be a random variable. Here is a simple model illustrating this situation. Let

$$Y_i = S_i x_i + \epsilon_i,$$

where $x_1, \ldots, x_n$ are known constants, and independently for $i = 1, \ldots, n$,

S_i (S for slope) is a normal random variable with expected value β and variance σ_1^2,

ϵ_i is a normal random variable with expected value zero and variance σ_2^2, and

S_i and ϵ_i are independent.

(a) This is a special case of the general mixed linear model, in which $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Zb} + \boldsymbol{\epsilon}$, where

- $\mathbf{X}$ is an $n \times p$ matrix of known constants
- $\boldsymbol{\beta}$ is a $p \times 1$ vector of unknown constants.
- $\mathbf{Z}$ is an $n \times q$ matrix of known constants
- $\mathbf{b} \sim N_q(\mathbf{0}, \boldsymbol{\Sigma}_b)$ with $\boldsymbol{\Sigma}_b$ unknown
- $\boldsymbol{\epsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}_n)$, where $\sigma^2 > 0$ is an unknown constant.

At first, the model in this problem might not seem to fit the specifications above. Why? But actually it does. To see this, give the distribution of Y_i, and then write a general mixed model that yields this same probability distribution. Now you can answer these questions.

i. What is the matrix $\mathbf{X}$? What is p?

ii. What is the matrix $\boldsymbol{\beta}$?

iii. What is the matrix $\mathbf{Z}$? What is q?

iv. What is the matrix $\mathbf{b}$?

v. What is the matrix $\boldsymbol{\Sigma}_b$?

(b) In the model $Y_i = S_i x_i + \epsilon_i$, what would happen if you tried to estimate the scalar parameter β in the usual way with

$$\widehat{\beta} = \frac{\sum_{i=1}^n x_i Y_i}{\sum_{i=1}^n x_i^2}$$

Under what conditions on the x_i values is this estimator consistent?

(c) Find another estimator of β by calculating $E(\overline{Y}_n)$. Why does the Law of Large Numbers *not* apply here? Okay, anyway, propose an estimator, and give a set of conditions on the x_i values that will make it consistent for β.

(d) Which estimator do you like more? Why?

(e) Now suppose that all the x_i values are equal to one.

i. What is the distribution of Y_i in this situation?

ii. Propose an estimator of β that should satisfy anyone.

iii. Give an *exact* $(1-\alpha)100\%$ confidence interval for β; you don't have to show any work.

iv. Now suppose that you want to estimate σ_1^2 and σ_2^2. Is this possible when all the x_i values equal one? Briefly comment.

4. This question was adapted from Box, Hunter, and Hunter (2005, chapter 5). The following experiment studied two quantitative factors - temperature **T** and concentration **C** - and a single qualitative factor - type of catalyst **K** on yield.

T (°C)		C (%)		K	
-	+	-	+	-	+
160	180	20	40	A	B

The experiment was replicated twice and the response $y_{ij}, i = 1, ..., 8, j = 1, 2$ is the yield of the j^{th} replicate for the i^{th} experimental run. The data are displayed below.

i	T	C	K	y_{i1}^a	y_{i2}	$\bar{y}_{i\cdot}$
1	-	-	-	$59^{(6)}$	$61^{(13)}$	60
2	+	-	-	$74^{(2)}$	$70^{(4)}$	72
3	-	+	-	$50^{(1)}$	$58^{(16)}$	54
4	+	+	-	$69^{(5)}$	$67^{(10)}$	68
5	-	-	+	$50^{(8)}$	$54^{(12)}$	52
6	+	-	+	$81^{(9)}$	$85^{(14)}$	83
7	-	+	+	$46^{(3)}$	$44^{(11)}$	45
8	+	+	+	$79^{(7)}$	$81^{(15)}$	80

[a] Superscripts give the order in which the runs were made.

Answer the following questions.

(a) What type of experimental design was used to study the effects of the three factors on yield? State the linear model associated with this design.

(b) Let $\hat{T}, \hat{C}$ and $\hat{K}$ denote the differences between marginal means for **T**, **C** and **K** respectively. Calculate these three estimates using the data.

(c) Suppose that the variance is the same for all the treatment combinations. Show that an estimate of the variance based on just the data for treatment combination i is $s_i^2 = \frac{(y_{i1}-y_{i2})^2}{2}$.

(d) Show that the standard error of an individual difference between marginal means is $\frac{s}{2}$, where $s^2 = \sum_{i=1}^{8} s_i^2/8$, with s_i^2 defined above.

(e) Derive a test statistic and use it to evaluate if any individual difference between marginal means is readily explained by chance. Explicitly state any assumptions that you make in developing a test statistic. What is the distribution of this test statistic? Derive a $100(1-\alpha)\%$ confidence interval for an individual difference between marginal means.

(f) Which differences between marginal means are most likely explained by chance? Explain your reasoning.

(g) Suppose that the linear regression model $y = X\beta + \epsilon$, is fit to the data, where $y = (\bar{y}_{1\cdot}, \bar{y}_{2\cdot}, ..., \bar{y}_{8\cdot})'$,

$$X = \begin{pmatrix} 1 & -1 & -1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & 1 & -1 & 1 \\ 1 & -1 & 1 & 1 \\ 1 & 1 & 1 & 1 \end{pmatrix}$$

$\beta = (\beta_0, \beta_1, \beta_2, \beta_3)'$, and $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_8)'$, $\epsilon_i \sim N(0, \sigma^2)$.

Derive $\hat{\beta}$ the least squares estimator from first principles. Show that $\hat{\beta}_0 = \sum_{i=1}^{8} \bar{y}_{i\cdot}/8$, $\hat{\beta}_1 = \frac{1}{2}\hat{T}$, $\hat{\beta}_2 = \frac{1}{2}\hat{C}$, $\hat{\beta}_3 = \frac{1}{2}\hat{K}$, where $\hat{T}, \hat{C}, \hat{K}$ are the difference between marginal means for temperature, concentration, and catalyst respectively.

5. This question is adapted from Faraway (2006, Ch. 3.2). A data set from Purott and Reeder (1976) gives results from an experiment conducted to determine the effect of gamma radiation on the numbers of chromosomal abnormalities observed. The number of cells (in hundred) exposed in each run differs. The dose amount, and the rate at which the dose is applied, are the predictors of interest. The data is given in Table 1. In Figure 1 and Table 2 we show the rate of abnormalities, `ca/cells`, taken to be the response of interest. The questions for this problem appear after the graphs and R code on the next pages.

Table 1: Original data: chromosome abnormalities (ca) / cells ('000s) (`cells`.

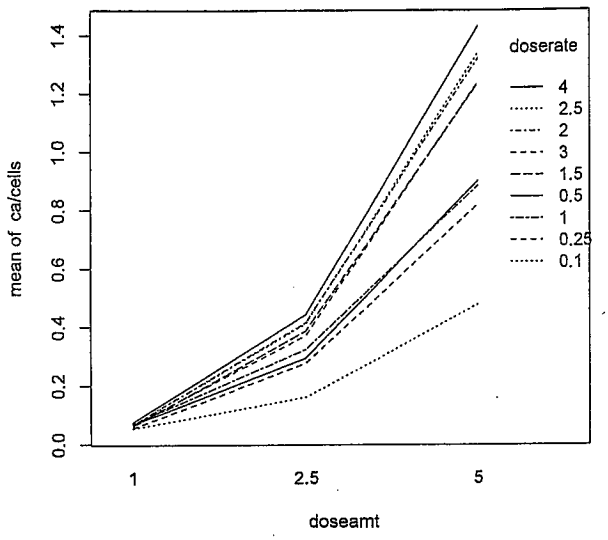
	dose amount		
doserate	1.0	2.5	5.0
0.10	25 / 478	52 / 328	100 / 210
0.25	102 / 1907	51 / 185	113 / 138
0.50	149 / 2258	100 / 342	144 / 160
1.00	160 / 2329	100 / 310	106 / 120
1.50	75 / 1238	107 / 278	111 / 90
2.00	100 / 1491	107 / 259	132 / 100
2.50	99 / 1518	102 / 249	419 / 313
3.00	50 / 764	110 / 298	225 / 182
4.00	100 / 1367	107 / 243	206 / 144

Table 2: `ca/cells`: Data on which Figure 1 is based

	0.1	0.25	0.5	1	1.5	2	2.5	3	4
1	0.052	0.053	0.066	0.069	0.061	0.67	0.065	0.065	0.073
2.5	0.159	0.276	0.292	0.322	0.384	0.413	0.410	0.369	0.440
5	0.476	0.818	0.900	0.883	1.233	1.320	1.339	1.2436	1.431

```
> library(faraway); data(dicentric)
> dim(dicentric)
[1] 27  4
> dicentric[1:11,]
   cells  ca doseamt doserate
1    478  25     1.0     0.10
2   1907 102     1.0     0.25
3   2258 149     1.0     0.50
4   2329 160     1.0     1.00
5   1238  75     1.0     1.50
6   1491 100     1.0     2.00
7   1518  99     1.0     2.50
8    764  50     1.0     3.00
9   1367 100     1.0     4.00
10   328  52     2.5     0.10
11   185  51     2.5     0.25
```

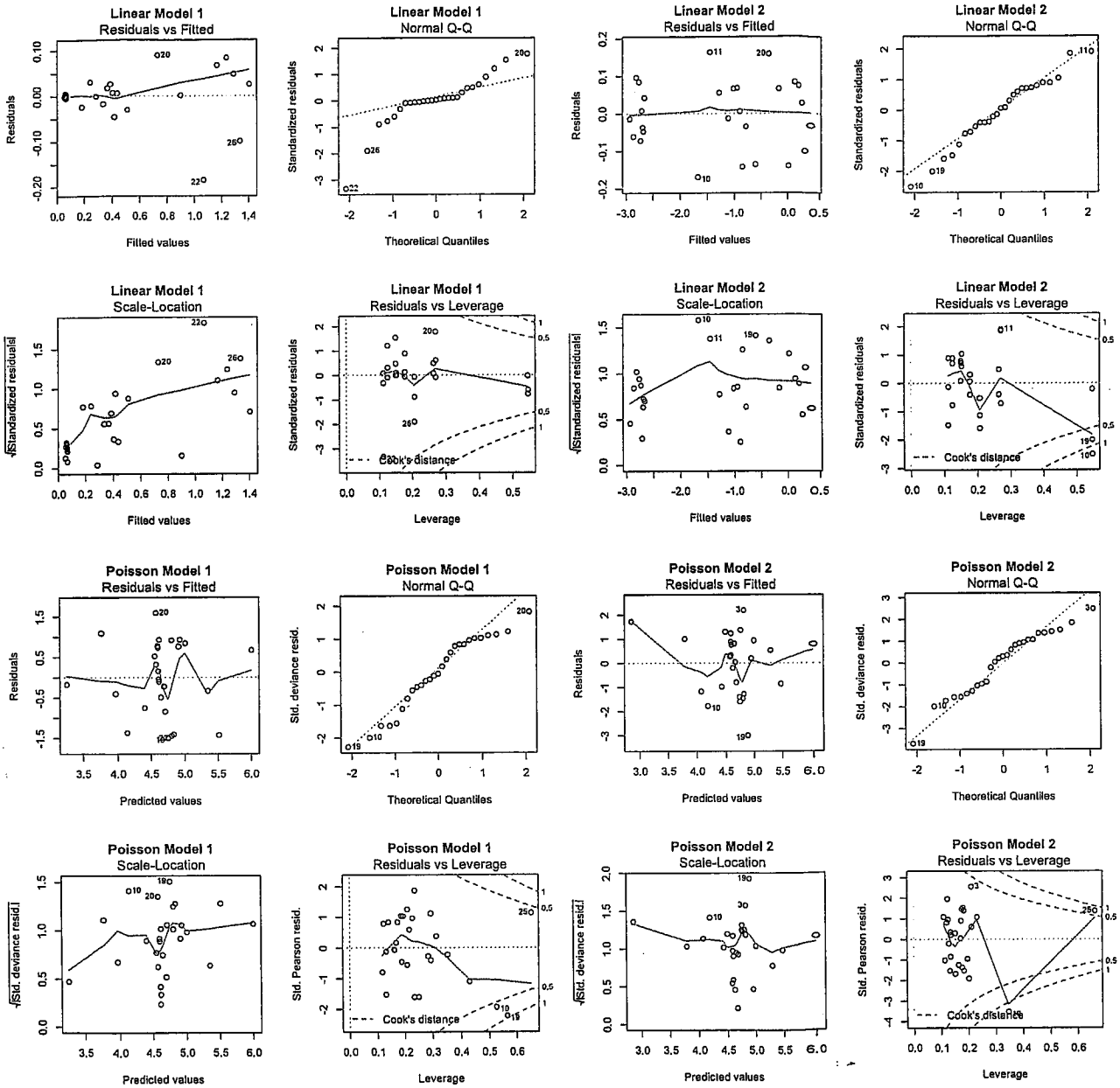
Figure 1: interaction plot of ca/cells against dose amount




```
> lm1 <- lm(cells/ca ~ log(doserate)*factor(doseamt), data = dicentric)
> summary(lm1); plot(lm1)
Call:
lm(formula = ca/cells ~ log(doserate) * factor(doseamt), data = dicentric)

Residuals:
     Min       1Q   Median       3Q      Max
-0.18427 -0.00421  0.00131  0.02121  0.08908

Coefficients:
                                  Estimate Std. Error t value Pr(>|t|)
(Intercept)                        0.49030    0.01127   43.49  < 2e-16 ***
log(doserate)                      0.10559    0.00964   10.96  3.8e-10 ***
factor(doseamt)1                  -0.42681    0.01594  -26.77  < 2e-16 ***
factor(doseamt)2                  -0.15050    0.01594   -9.44  5.3e-09 ***
log(doserate):factor(doseamt)1 -0.10102    0.01363   -7.41  2.7e-07 ***
log(doserate):factor(doseamt)2 -0.03709    0.01363   -2.72    0.013 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.0586 on 21 degrees of freedom
Multiple R-squared:  0.987,Adjusted R-squared:  0.984
F-statistic:  330 on 5 and 21 DF,  p-value: <2e-16
```

```
> lm2 <- lm(log(cells/ca) ~ log(dose rate)*factor(dose amt), data = dicentric)
> summary(lm2); plot(lm2)

Call:
lm(formula = log(ca/cells) ~ log(doserate) * factor(doseamt),
    data = dicentric)

Residuals:
     Min       1Q   Median       3Q      Max
-0.16950 -0.05413  0.00585  0.06717  0.16343

Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                      -1.2883     0.0194  -66.56  < 2e-16 ***
log(doserate)                     0.1955     0.0165   11.82  9.7e-11 ***
factor(doseamt)1                 -1.4741     0.0274  -53.86  < 2e-16 ***
factor(doseamt)2                  0.1696     0.0274    6.20  3.8e-06 ***
log(doserate):factor(doseamt)1   -0.1199     0.0234   -5.12  4.5e-05 ***
log(doserate):factor(doseamt)2    0.0450     0.0234    1.92    0.068 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 0.101 on 21 degrees of freedom
Multiple R-squared:  0.994,Adjusted R-squared:  0.993
F-statistic:  729 on 5 and 21 DF,  p-value: <2e-16

> glm1 <- glm(ca ~ log(cells) + log(doserate)*factor(doseamt), family = poisson, data = dicentric)
> summary(glm1); plot(glm1)

Call:
glm(formula = ca ~ log(cells) + log(doserate) * factor(doseamt),
    family = poisson, data = dicentric)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.4990  -0.6223  -0.0502   0.7692   1.5953

Coefficients:
                                Estimate Std. Error z value Pr(>|z|)
(Intercept)                      -1.2998     0.3119   -4.17  3.1e-05 ***
log(cells)                        1.0025     0.0514   19.52  < 2e-16 ***
log(doserate)                     0.1901     0.0180   10.56  < 2e-16 ***
factor(doseamt)1                 -1.4655     0.0736  -19.92  < 2e-16 ***
factor(doseamt)2                  0.1643     0.0353    4.65  3.3e-06 ***
log(doserate):factor(doseamt)1   -0.1181     0.0273   -4.33  1.5e-05 ***
log(doserate):factor(doseamt)2    0.0430     0.0262    1.64      0.1
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

```
(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 916.127  on 26  degrees of freedom
Residual deviance:  21.748  on 20  degrees of freedom

> glm2 <- glm(ca ~ log(cells) + log(doserate) + factor(doseamt), family = poisson, data = dicentri
> summary(glm2); plot(glm2)

Call:
glm(formula = ca ~ log(cells) + log(doserate) + factor(doseamt),
    family = poisson, data = dicentric)

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.996  -1.076   0.272   0.913   2.197

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       -1.5438     0.3132   -4.93  8.3e-07 ***
log(cells)         1.0427     0.0516   20.22  < 2e-16 ***
log(doserate)      0.2151     0.0167   12.85  < 2e-16 ***
factor(doseamt)1  -1.5408     0.0729  -21.13  < 2e-16 ***
factor(doseamt)2   0.1867     0.0350    5.33  9.7e-08 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 916.127  on 26  degrees of freedom
Residual deviance:  42.089  on 22  degrees of freedom
AIC: 227.5

Number of Fisher Scoring iterations: 4

> anova(glm1,glm2)
Analysis of Deviance Table

Model 1: ca ~ log(cells) + log(doserate) * factor(doseamt)
Model 2: ca ~ log(cells) + log(doserate) + factor(doseamt)
  Resid. Df Resid. Dev Df Deviance
1        20       21.7
2        22       42.1 -2    -20.3
```


Linear Model 1
Residuals vs Fitted
Residuals
Fitted values
Linear Model 1
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Linear Model 2
Residuals vs Fitted
Residuals
Fitted values
Linear Model 2
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Linear Model 1
Scale-Location
√|Standardized residuals|
Fitted values
Linear Model 1
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Linear Model 2
Scale-Location
√|Standardized residuals|
Fitted values
Linear Model 2
Residuals vs Leverage
Standardized residuals
Cook's distance
Leverage
Poisson Model 1
Residuals vs Fitted
Residuals
Predicted values
Poisson Model 1
Normal Q-Q
Std. deviance resid.
Theoretical Quantiles
Poisson Model 2
Residuals vs Fitted
Residuals
Predicted values
Poisson Model 2
Normal Q-Q
Std. deviance resid.
Theoretical Quantiles
Poisson Model 1
Scale-Location
√|Std. deviance resid.|
Predicted values
Poisson Model 1
Residuals vs Leverage
Std. Pearson resid.
Cook's distance
Leverage
Poisson Model 2
Scale-Location
√|Std. deviance resid.|
Predicted values
Poisson Model 2
Residuals vs Leverage
Std. Pearson resid.
Cook's distance
Leverage

(a) The R code shows four different model fits, `lm1`, `lm2`, `glm1` and `glm2`. Write each of these three models in mathematical form, something like " $y_i \sim N(\mu_i, \sigma^2)$; $\mu_i = \beta_0 + \beta_1 x_i$, where $y =$, $x =$ ". (This quotation is not correct, of course.)

(b) For the model `lm2`, use the R code to give estimates of each of the parameters in the model in part (a). Interpret each of these estimates in terms of their effect on the appropriate response.

(c) Explain in non-technical terms the differences between models `lm2` and `glm1`.

(d) Why do you think the explanatory variable dose rate was transformed to the log-scale?

(e) Does there seem to be an interaction between dose rate and dose amount? Explain.

(f) Which of the four models do you think is best for assessing the effect of gamma radiation on chromosomal abnormalities?