Date: 8th May 2015

Instructions

- Answer only two questions from Section L. If you answer more than two questions, then only the FIRST TWO questions will be marked.

- Answer only two questions from Section G. If you answer more than two questions, then only the FIRST TWO questions will be marked.

<table>
<thead>
<tr>
<th>Questions</th>
<th>Marks</th>
</tr>
</thead>
<tbody>
<tr>
<td>L1</td>
<td></td>
</tr>
<tr>
<td>L2</td>
<td></td>
</tr>
<tr>
<td>L3</td>
<td></td>
</tr>
<tr>
<td>G1</td>
<td></td>
</tr>
<tr>
<td>G2</td>
<td></td>
</tr>
<tr>
<td>G3</td>
<td></td>
</tr>
</tbody>
</table>

This exam comprises the cover page and fourteen pages of questions.
Notation: In Section L, the following notation will be used: for \( i = 1, \ldots, n \), \( y_i \) is the observed response; \( Y_i \) is the random variable version of the response; \( y \) and \( Y \) are the \( n \times 1 \) vector versions of the responses; \( x_i \) is the row vector of predictor values, \( X \) is the matrix of predictor values; \( \hat{y}_i, \hat{Y}_i, \hat{y} \) and \( \hat{Y} \) are the fitted or predicted response values or vectors arising from a given model; \( \beta \) is the vector of regression coefficients; \( \hat{\beta} \) is the vector of estimates or estimators. Furthermore, \( 0_n \) is the \( n \)-dimensional vector of zeros, and \( I_n \) is the \( n \)-dimensional identity matrix.

L1. (a) Consider the linear regression model specified (in vector form) by the equation

\[
Y = X\beta + \varepsilon
\]

for random error vector \( \varepsilon \), where \( \beta \) is a \( p \times 1 \) vector of regression coefficients, and where

\[
E[\varepsilon|X] = 0_n, \quad \text{Var}[\varepsilon|X] = \sigma^2 I_n.
\]

(i) Derive the form of the ordinary least squares (OLS) estimator of \( \beta \), and show that it is unbiased.

4 MARKS

(ii) Show that for the linear regression model above, the vector of residuals

\[
e = y - \hat{y}
\]

is orthogonal to each of the columns of \( X \), and is also orthogonal to the vector of fitted values, \( \hat{y} \).

4 MARKS

(iii) Show that the (ANOVA) sums of squares decomposition for the model may be written

\[
Y^\top (I_n - H_I) Y = Y^\top (I_n - H) Y + Y^\top (H - H_I) Y
\]

for matrices \( H \) and \( H_I \) to be defined.

4 MARKS

(b) Suppose now that a random sample of pairs of random variables \( (X_i, Y_i), i = 1, \ldots, n \) is considered, and that a bivariate Normal distribution is presumed, that is

\[
\begin{bmatrix} X_i \\ Y_i \end{bmatrix} \sim \text{Normal}_2(\mu, \Sigma) \quad i = 1, \ldots, n
\]

where

\[
\mu = \begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \quad \Sigma = \begin{bmatrix} \sigma_X^2 & \sigma_{XY} \\ \sigma_{XY} & \sigma_Y^2 \end{bmatrix}.
\]

The conditional model for \( Y_i \) given \( X_i = x_i \) is to be considered.

Derive the maximum likelihood estimators of the parameters in the conditional model, and compare them with the estimators obtained using a regression approach and ordinary least squares.

Recall that, for the bivariate normal model, the conditional distribution of \( Y_i \) given \( X_i = x_i \) is Normal,

\[
\text{Normal} \left( \mu_Y + \frac{\sigma_Y}{\sigma_X} (x_i - \mu_X), \sigma_Y^2 (1 - \rho^2) \right)
\]

where

\[
\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}.
\]

8 MARKS
L2. The following data set relates to a small clinical study of three cholesterol lowering drugs (variable \( D \), with levels denoted 1, 2, 3); each patient in the study was classified into one of two subgroups (variable \( G \), with levels denoted 1, 2) according to a metabolic classification scheme, and then given a course of treatment using one allocated drug. The outcome of the study was recorded as the cholesterol reduction (in millimoles per litre, mmol/l) in standard blood samples taken before and after the course of treatment.

<table>
<thead>
<tr>
<th>( D )</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>1</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>2</th>
<th>3</th>
<th>3</th>
<th>3</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>( G )</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>( Y )</td>
<td>0.63</td>
<td>0.99</td>
<td>1.31</td>
<td>3.10</td>
<td>1.99</td>
<td>1.55</td>
<td>1.70</td>
<td>-0.10</td>
<td>1.76</td>
<td>0.29</td>
<td>0.76</td>
<td>0.55</td>
<td>-0.16</td>
<td>0.09</td>
</tr>
</tbody>
</table>

(a) Derive an expression for the bias in the OLS estimators for the linear model parameters derived from the fit of an additive model \( D + G \) if in fact the true (data generating) model includes both main effects and an interaction term between \( D \) and \( G \).

(b) The following R output relates to a partial analysis of the data:

```r
> summary(lm(Y ~ D))
Coefficients:
            Estimate  Std. Error   t value     Pr(>|t|)
(Intercept) 1.5617     0.2277   6.858    1.75e-05 ***
D2          -0.9942    0.3600  -2.761    0.017241 *
D3          -1.4657    0.3378  -4.339    0.000963 ***
---
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5578 on 12 degrees of freedom
Multiple R-squared: 0.6232,  Adjusted R-squared: 0.5604
F-statistic: 9.922 on 2 and 12 DF,  p-value: 0.002863
```

```r
> anova(lm(Y ~ G+D))
Analysis of Variance Table

Response: Y

                      Df  Sum Sq Mean Sq F value    Pr(>F)
G                    1 2.1090 2.10901  7.9474 0.016694 *
D                    2 4.8796 2.43978  9.1938 0.004495 **
Residuals           11 2.9191 0.26537
```

Summarize the evidence available in this analysis to support the research hypothesis that the metabolic subgroups have a significantly different expected response to each other.

Note that the 0.95 quantile of the \( F_{1,11} \equiv \text{Fisher}(1,11) \) distribution is 4.844.

(c) Compute the \( R^2 \) statistic for the model \( G + D \).
L3. The following data relate a study of state by state crime statistics in the United States. The variables included in the data frame `cdata` are

- state id (`sid`),
- state name (`state`),
- violent crimes per 100,000 people (`crime`),
- murders per 1,000,000 (`murder`),
- the percent of the population living in metropolitan areas (`pctmetro`),
- the percent of the population that is white (`pctwhite`),
- percent of population with a high school education or above (`pcths`),
- percent of population living under poverty line (`poverty`),
- percent of population that are single parents (`single`).

There are 51 data points. The objective of the study was to understand whether the per capita rates of violent crime were related to the other measured predictors.

(a) Summarize the conclusions to be made from the following analysis in R:

```r
> fit.full<-lm(crime~(pctmetro+pctwhite+pcths+poverty+single)^2, data=cdata)
> anova(fit.full)
Analysis of Variance Table

  Response: crime

       Df Sum Sq Mean Sq F value    Pr(>F)
pctmetro  1 2879417 2879417 149.8809 3.313e-14 ***
pctwhite  1 2675621 2675621 139.2728 9.381e-14 ***
pcths     1  48679   48679  2.5339 0.12042
poverty   1 1725561 1725561  89.8198 3.389e-11 ***
single    1  938579  938579  48.8554 3.940e-08 ***
pctmetro:pctwhite  1  434360  434360  22.6096 3.356e-05 ***
pctmetro:pcths    1  48597   48597  2.5296 0.12072
pctmetro:poverty  1 12314   12314  0.6410 0.42876
pctwhite:pcths    1  4911    4911  0.2556 0.61631
pctwhite:poverty  1  1256    1256  0.0654 0.79968
pcths:poverty     1  8078    8078  0.4205 0.52094
poverty:single    1 103111  103111  5.3672 0.02649 *
Pcths:poverty     1  87040   87040  4.5307 0.04041 *
poverty:single    1  5.3672  5.3672  0.02649 *
Residuals    35 672398   19211
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

*Question L3 continues on the next page.*
Section L

Section L

(26) fit.1<-lm(crime~pctmetro+pctwhite+pcths+poverty+single,data=cdata)
27 > anova(fit.full,fit.1,test='F')
28 Analysis of Variance Table

29 Model 1: crime ~ (pctmetro + pctwhite + pcths + poverty + single)^2
30 Model 2: crime ~ pctmetro + pctwhite + pcths + poverty + single
31 Res.Df RSS Df Sum of Sq F Pr(>F)
32 1 35 672398
33 2 45 1460618 -10 -788221 4.1029 0.0008645 ***
34 ---
35 Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

4 MARKS

(b) A second phase of analysis was then considered:
37 > anova(fit1,fit2,fit3,fit4,fit5)
38 Analysis of Variance Table

39 Model 1: crime ~ pctmetro + pctwhite + pcths + poverty + single
40 Model 2: crime ~ pctmetro * pctwhite + pcths + poverty + single
41 Model 3: crime ~ pctmetro + pctwhite + pcths + poverty * single
42 Model 4: crime ~ pctmetro * pctwhite + pcths + poverty * single
43 Model 5: crime ~ pctmetro * pctwhite + poverty * single
44 Res.Df RSS Df Sum of Sq F Pr(>F)
45 1 45 1460618
46 2 44 1026258 1 434360 19.3773 6.978e-05 ***
47 3 44 1443119 0 -416861
48 4 43 963884 1 479235 21.3792 3.435e-05 ***
49 5 44 968243 -1 -4359 0.1945 0.6614
50 ---
51 Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
52 > drop1(fit5,test='F')
53 Single term deletions

54 Model:
crime ~ pctmetro * pctwhite + poverty * single
55 Df Sum of Sq RSS AIC F value Pr(>F)
56 <none> 968243 516.42
57 pctmetro:pctwhite 1 498431 1466674 535.60 22.6503 2.126e-05 ***
58 poverty:single 1 58529 1026772 517.42 2.6597 0.1101
59 ---
60 Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Summarize the conclusions to be made from this output. 5 MARKS

Question L3 continues on the next page.
A third analysis was carried out:

```r
fit6 <- lm(crime ~ murder + pctmetro * pctwhite + poverty * single, data = cdata)

> coef(summary(fit6))

Estimate  Std. Error     t value  Pr(>|t|)
(Intercept)  -3114.5252651  830.1133288  -3.751928  0.0005206568
murder          22.6502533    7.1452314    3.169982  0.0028083836
pctmetro       24.1719303    9.3092552    2.596548  0.0128367490
pctwhite       12.4397412    6.9541255    1.788829  0.0806885140
poverty        85.9493355  23.3236529    3.685072  0.0006358045
single         164.3391613  48.2100000    3.431642  0.0001181525
pctmetro:pctwhite  -0.2061182   0.1031445   -1.998344  0.0520279415
poverty:single    -6.5010010   2.0482300   -3.173661  0.0027797707

> anova(fit5, fit6)

Analysis of Variance Table

Model 1: crime ~ pctmetro * pctwhite + poverty * single
Model 2: crime ~ murder + pctmetro * pctwhite + poverty * single

Res.Df RSS Df Sum of Sq    F    Pr(>F)
1 44  968243
2 43  784833 1   183410  10.049  0.002808 **

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

> fit7 <- lm(murder ~ pctmetro * pctwhite + poverty * single, data = cdata)

> summary(fit7)

Coefficients:

Estimate  Std. Error     t value  Pr(>|t|)
(Intercept)  -49.082179    15.874574   -3.092  0.003447 **
pctmetro        0.860490    0.147479     5.835  5.90e-07 ***
pctwhite        0.558241    0.120188     4.645  3.09e-05 ***
poverty        -1.290914    0.451983    -2.856  0.006522 **
single         -0.187994    0.718661    -0.262  0.794859
pctmetro:pctwhite  -0.009216   0.001675   -5.502  1.81e-06 ***
poverty:single    0.148482    0.036971     4.016  0.000227 ***

---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.85 on 44 degrees of freedom
Multiple R-squared: 0.9378, Adjusted R-squared: 0.9293
F-statistic: 110.5 on 6 and 44 DF,  p-value: < 2.2e-16
```

Comment on the utility of the predictor `murder` in predicting per capita crime rate.

4 MARKS

*Question L3 continues on the next page.*
(d) Define the *standardized* and *studentized* residuals that are typically examined when considering the fit of a regression model, and explain why each is preferable to the usual residual quantity $e_i = y_i - \hat{y}_i$.  

2 MARKS

(d) Define the *leverage* of a data point that may be computed for the fit of a regression model.  

2 MARKS

(f) The following residual plots were obtained from the model fit6:

Summarize the conclusions that should be drawn from these plots.  

3 MARKS
G1. The Bradley-Terry model is often used to rank individuals or objects subject to repeated comparisons amongst pairs. Take, for example, a competitive tennis league run by a local tennis club. For each match in league play, the two players’ names and the winner of the match are recorded. Assume that there are $K$ total players. Let $\pi_{ab}$ be the probability that player $a$ defeats player $b$ in a match, where $a \neq b$. Suppose that $\pi_{ab} + \pi_{ba} = 1$ (which in our context implies that there are no ties). The Bradley-Terry model assumes that:

$$\log \left( \frac{\pi_{ab}}{\pi_{ba}} \right) = \beta_a - \beta_b.$$

For $a < b$, let $N_{ab}$ denote the number of matches between players $a$ and $b$, with player $a$ winning $n_{ab}$ times and player $b$ winning $n_{ba}$ times.

(a) Treating the quantities $\{n_{ab} : a < b\}$ as realizations of independent binomial variates, show that the Bradley-Terry model can be formulated as a generalized linear model. Clearly define the design matrix, the link, and the response functions.

(b) Using your answer to part (a) or otherwise, find the minimal sufficient statistics for the model parameters in terms of the match data, $\{n_{ab} : a < b\}$.

(c) Describe how you would test, using the Bradley-Terry model, the hypothesis that player $a$ is more likely to win against player $b$ in a match than she is to lose. Clearly state the null and alternative hypotheses, how you would construct the test statistic, and the rejection region for your test.

(d) There is some concern that the players on the east side of the court (i.e. the side facing the sunset) are at a disadvantage because they face the setting sun. Assume that all courts are outdoors and that this potential disadvantage is the same across all matches. Describe how to modify your model to estimate the extent of the disadvantage.
G2. Cross-sectional data were collected on 428 women in order to model the association between women’s wages ($\text{WW}$, in dollars/10,000) and two covariates: the number of years of education ($\text{WE}$) and the number of children they have between the ages of 6 and 18 ($\text{K618}$). The figure below shows pairwise scatterplots of the three variable of interest.

(a) Researchers first fit the regression model contained in the output below. Summarize the conclusions from the following initial analysis in R. 5 MARKS

```r
> G2.M1 = glm(WW ~ WE+K618, data=wagedata)
> summary(G2.M1)

Coefficients:

    Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.84788   0.87958  -2.101  0.0362 *
    WE       0.48888   0.06622   7.382 8.28e-13 ***
    K618     -0.12074   0.11501  -1.050  0.2944

(Dispersion parameter for gaussian family taken to be 9.697015)

Null deviance: 4679.1 on 427 degrees of freedom
Residual deviance: 4121.2 on 425 degrees of freedom

> anova(G2.M1)

Analysis of Deviance Table

Model: gaussian, link: identity
Response: WW
Terms added sequentially (first to last)

    Df Deviance Resid. Df Dev
NULL          427 4679.1
WE            1   547.13    426 4131.9
K618          1    10.69    425 4121.2

> sum(resid(G2.M1,"pearson")^2)

[1] 4121.232
```

*Question G2 continues on the next page.*
(b) Based on the residual diagnostic plots below and the output above, do you think that the model in part (a) is appropriate? Explain your reasoning clearly.

Figure 1: Residual Diagnostic Plots for model G2.M1

Question G2 continues on the next page.
(c) A second regression model was fit using the code below. Summarize the conclusions drawn from this second analysis. Clearly define the model that is fit by the code below and interpret the parameter estimates obtained from the model.

```
27 > G2.M2 = glm(WW ~ WE+K618, data=wagedata, family=Gamma(link=log))
28 > summary(G2.M2)
29
30 Coefficients:
31 Estimate Std. Error t value Pr(>|t|)
32 (Intercept) 0.04125 0.21335 0.193 0.847
33 WE 0.10902 0.01606 6.787 3.87e-11 ***
34 K618 -0.01900 0.02790 -0.681 0.496
35 ---
36 Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 . ‘.’ 0.1 ‘ ’ 1
37
38 (Dispersion parameter for Gamma family taken to be 0.5705232)
39
40 Null deviance: 205.08 on 427 degrees of freedom
41 Residual deviance: 175.87 on 425 degrees of freedom
42 AIC: 1889.9
43 Number of Fisher Scoring iterations: 5
44
45 > anova(G2.M2)
46 Analysis of Deviance Table
47
48 Model: Gamma, link: log
49 Response: WW
50 Terms added sequentially (first to last)
51
52 Df Deviance Resid. Df Resid. Dev
53 NULL 427 205.08
54 WE 1 28.9431 426 176.14
55 K618 1 0.2701 425 175.87
56
57 > sum(resid(G2.M2, "pearson")^2)
58 [1] 242.4723
```

(d) Compute the estimated mean wages for a woman with 15 years of education and 3 children aged 6-18 under both models G1.M1 and G2.M2.

```
(d) Compute the estimated mean wages for a woman with 15 years of education and 3 children aged 6-18 under both models G1.M1 and G2.M2.
```

*Question G2 continues on the next page.*
(e) Based on the residual diagnostic plots below and the output above, which of the two models (G1.M1 and G2.M2) do you believe is most appropriate? Explain your reasoning clearly.

4 MARKS

Figure 2: Residual Diagnostic Plots for model G2.M2
G3. The following four questions were asked in a social science survey of a random sample of adults.

1. How would you classify yourself on the political spectrum: Conservative, Moderate, or Liberal? (coded by variable `pol` in the R code that follows)
2. Do you believe that teenagers should have access to birth control (yes/no)? (`bc`)
3. Do you frequently attend religious services (yes/no)? (`re`)
4. Do you believe that it is appropriate for unmarried people to have sexual relations before marriage (yes/no)? (`pr`)

The data are summarized in the contingency table below:

<table>
<thead>
<tr>
<th></th>
<th>Prem. Sex</th>
<th>No</th>
<th>Yes</th>
<th>Att. Relig.</th>
<th>No</th>
<th>Yes</th>
<th>No</th>
<th>Yes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Birth Control</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Pol: Cons</td>
<td>73</td>
<td>25</td>
<td>99</td>
<td>15</td>
<td>24</td>
<td>22</td>
<td>8</td>
<td>4</td>
</tr>
<tr>
<td>Pol: Mod</td>
<td>87</td>
<td>37</td>
<td>73</td>
<td>20</td>
<td>50</td>
<td>60</td>
<td>20</td>
<td>13</td>
</tr>
<tr>
<td>Pol: Lib</td>
<td>51</td>
<td>36</td>
<td>51</td>
<td>19</td>
<td>33</td>
<td>88</td>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>

(a) For the following questions, consider the analyses in the R code on page 14 involving only the questions relating to birth control, religious services attendance, and premarital sex.

(i) For each of the four models (G3.M1, G3.M2, and G3.M3, describe what each of the models assume in terms of the dependence structure amongst the three variables. 4 MARKS

(ii) For model G3.M4, interpret the values of the parameter estimates for the coefficient `bcYes:reYes` with respect to conditional odds ratios. 2 MARKS

(iii) Choose which of the four models presented is most appropriate for inference. Explain your answers using the output. 4 MARKS

(b) For the following questions, consider the analyses in the R code on page 15 (two pages following) involving all four survey questions.

(i) What restriction is placed upon all pairwise associations between variables by the model G3.M5? 2 MARKS

(ii) Is there evidence that a more complex model is required than G3.M5? Explain your answer clearly. 3 MARKS

(iii) Is there evidence that any of the terms in model G3.M5 can be dropped? If so, indicate which term can be dropped and describe how that changes the interpretation of the conditional associations between the variables. 5 MARKS

Question G3 continues on the next page.
Question G3 continues on the next page.
> G3.M5 = glm(counts ~ bc*re+bc*pol+bc*pr +
       re*pol + re*pr + pol*pr, family="poisson")
> summary(G3.M5)

Coefficients:

(Intercept) 4.33602 0.10469 41.419 < 2e-16 ***
bcYes -1.18446 0.16976 -6.977 3.01e-12 ***
reYes 0.22899 0.13268 1.726 0.084379 .
polLib -0.39337 0.15248 -2.580 0.009885 **
polMod 0.08541 0.13825 0.618 0.536694
prYes -1.22124 0.17507 -6.976 3.04e-12 ***
bcYes:reYes -0.59793 0.16255 -3.678 0.000235 ***
bcYes:polLib 0.92882 0.19362 4.797 1.61e-06 ***
bcYes:polMod 0.30484 0.18933 1.610 0.107377
bcYes:prYes 1.14683 0.15315 7.488 6.99e-14 ***
reYes:polLib -0.34408 0.18830 -1.827 0.067658 .
reYes:polMod -0.25827 0.17291 -1.494 0.135261
reYes:prYes -1.14593 0.16980 -6.749 1.49e-11 ***
polLib:prYes 0.80176 0.20306 3.948 7.87e-05 ***
polMod:prYes 0.71995 0.19521 3.688 0.000226 ***

---

(Dispersion parameter for poisson family taken to be 1)

Null deviance: 477.7846 on 23 degrees of freedom
Residual deviance: 6.9631 on 9 degrees of freedom
AIC: 161.4

> min(fitted(G3.M5))
[1] 4.769702

> drop1(G3.M5)

Single term deletions

Model:
counts ~ bc * re + bc * pol + bc * pr + re * pol + re * pr +
       pol * pr

Df Deviance   AIC
<none> 6.963 161.40
bc:re 1 20.725 173.16
bc:pol 2 32.950 183.39
bc:pr 1 64.064 216.50
re:pol 2 10.701 161.14
re:pr 1 56.124 208.56
pol:pr 2 25.866 176.31