McGill University

Faculty of Science

Department of Mathematics and Statistics

Part A Examination

Statistics: Methodology Paper

Date: 10th May 2019 Time: 1pm-5pm

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.**

Questions	Marks
L1	
L2	
L3	
G1	
G2	
G3	

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; $\mathbf{y}$ and $\mathbf{Y}$ are the $n \times 1$ vector versions of the responses; $\mathbf{x}_i$ is the row vector of predictor values, $\mathbf{X}$ is the matrix of predictor values; $\widehat{y}_i$, $\widehat{Y}_i$, $\widehat{\mathbf{y}}$ and $\widehat{\mathbf{Y}}$ are the fitted or predicted response values or vectors arising from a given model; β is the vector of regression coefficients; $\widehat{\beta}$ is the vector of estimates or estimators. Furthermore, $\mathbf{0}_n$ is the n-dimensional vector of zeros, and $\mathbf{I}_n$ is the n-dimensional identity matrix.

L1. 1. Suppose we wish to find the least-squares estimator of $\boldsymbol{\beta}$ in the model $= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ subject to a set of equality constraints on $\boldsymbol{\beta}$, say, $\mathbf{T}\boldsymbol{\beta} = \mathbf{c}$, where the matrix $\mathbf{T}$ and the vector $\mathbf{c}$ are specified in advance.

(a) Show that the new estimator is

$$\widetilde{\boldsymbol{\beta}} = \widehat{\boldsymbol{\beta}} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{T}^\top \left[\mathbf{T}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{T}^\top\right]^{-1} (\mathbf{c} - \mathbf{T}\widehat{\boldsymbol{\beta}}).$$

where $\widehat{\boldsymbol{\beta}}$ is the regular least-squares estimator of $\boldsymbol{\beta}$. 4 MARKS

(b) Briefly discuss the situation in which the new estimator might be appropriate. 1 MARKS

2. Consider the hat matrix $\mathbf{H} = \mathbf{X}(\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$, where $\mathbf{X}_{n\times p}$, with $p = k + 1$, is the design matrix.

(a) Consider the vector of fitted values $\widehat{\mathbf{Y}} = (\widehat{Y}_1, \widehat{Y}_2, \ldots, \widehat{Y}_n)^\top$. Show that

$$\text{Var}(\widehat{\mathbf{Y}}) = \sigma^2\mathbf{H} \qquad \text{and} \qquad \sum_{i=1}^{n} \text{Var}(\widehat{Y}_i) = p\sigma^2,$$

where $p = k + 1$. 2.5 MARKS

(b) Consider the least squares estimator $\widehat{\boldsymbol{\beta}}$ in a multiple linear regression model. Show that

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta} + \mathbf{R}\boldsymbol{\varepsilon} \qquad \text{where} \qquad \mathbf{R} = (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$$

2.5 MARKS

3. Suppose that we start from a way of creating confidence sets for the regression coefficient β_1, which we know have confidence level $1 - \alpha$

$$\text{CI}: \left(\widehat{\beta}_1 - k\widehat{\text{se}}\left(\widehat{\beta}_1\right), \widehat{\beta}_1 + k\widehat{\text{se}}\left(\widehat{\beta}_1\right)\right)$$

we can test the hypothesis

$$H_0: \beta_1 = \beta_1^*$$
$$H_1: \beta_1 \neq \beta_1^*$$

by rejecting $H_0 : \beta_1 = \beta_1^*$ when β_1^* is outside the confidence set and retaining $H_0 : \beta_1 = \beta_1^*$ when β_1^* is inside the confidence set. Show that this test is just the two sided t-test with size α. 5 MARKS

4. Consider the ridge regression estimator

$$\widehat{\boldsymbol{\beta}}_\lambda = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_2^2$$

(a) Derive the formula for computing $\widehat{\boldsymbol{\beta}}_\lambda$ using λ, $\mathbf{X}$ and $\mathbf{Y}$. 2 MARKS

(b) Show that $\widehat{\boldsymbol{\beta}}_\lambda \to 0$ as $\lambda \to \infty$. 1 MARKS

(c) Consider the following modified ridge estimator:

$$\widehat{\boldsymbol{\beta}}_\lambda = \lambda(\mathbf{X}^\top\mathbf{X} + \lambda\mathbf{I}_n)^{-1}\mathbf{X}^\top\mathbf{Y}.$$

what does this estimator converge to as $\lambda \to \infty$? 2 MARKS

L2. The Minnesota Twins professional baseball team plays its games in the Metrodome, an indoor stadium with a fabric roof. In addition to the large air fans required to keep to roof from collapsing, the baseball field is surrounded by ventilation fans that blow heated or cooled air into the stadium. Air is normally blown into the center of the field equally from all directions.

According to a retired supervisor in the Metrodome, in the late innings of some games the fans would be modified so that the ventilation air would blow out from home plate toward the outfield. The idea is that the air flow might increase the length of a fly ball. For example, if this were done in the middle of the eighth inning, then the air-flow advantage would be in favor of the home team for six outs, three in each of the eighth and ninth innings, and in favor of the visitor for three outs in the ninth inning, resulting in a slight advantage for the home team.

To see if manipulating the fans could possibly make any difference, a group of students at the University of Minnesota and their professor built a "cannon" that used compressed air to shoot baseballs. They then did the following experiment in the Metrodome in March, 2003:

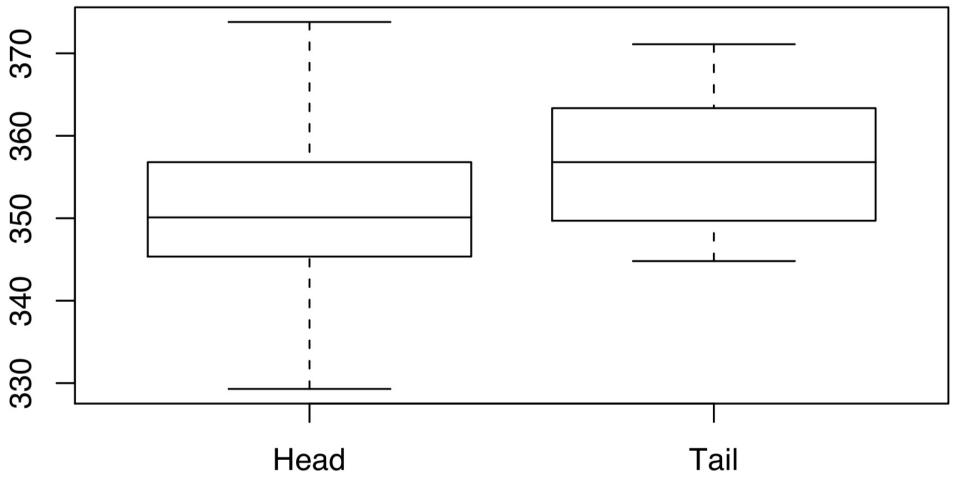
1. A fixed angle of 50 degrees and velocity of 150 feet per second was selected. In the actual experiment, neither the velocity nor the angle could be controlled exactly, so the actual angle and velocity varied from shot to shot.
2. The ventilation fans were set so that to the extent possible all the air was blowing in from the outfield towards home plate, providing a headwind. After waiting about 20 minutes for the air flows to stabilize, 20 balls were shot into the outfield, and their distances were recorded. Additional variables recorded on each shot include the weight (in grams) and diameter (in cm) of the ball used on that shot, and the actual velocity and angle.
3. The ventilation fans were then reversed, so as much as possible air was blowing out toward the outfield, giving a tailwind. After waiting 20 minutes for air currents to stabilize, 15 balls were shot into the outfield, again measuring the ball weight and diameter, and the actual velocity and angle on each shot.

In this data, the variable names are *Cond*, the condition, head or tail wind; *Velocity*, the actual velocity in feet per second; *Angle*, the actual angle; *BallWt*, the weight of the ball in grams used on that particular test; *BallDia*, the diameter in inches of the ball used on that test; *Dist*, distance in feet of the flight of the ball.

1. The following plot shows a boxplot of the response *Dist* for each value of *Cond*:

```
1 > boxplot(Dist ~ Cond, data=domedata)
```

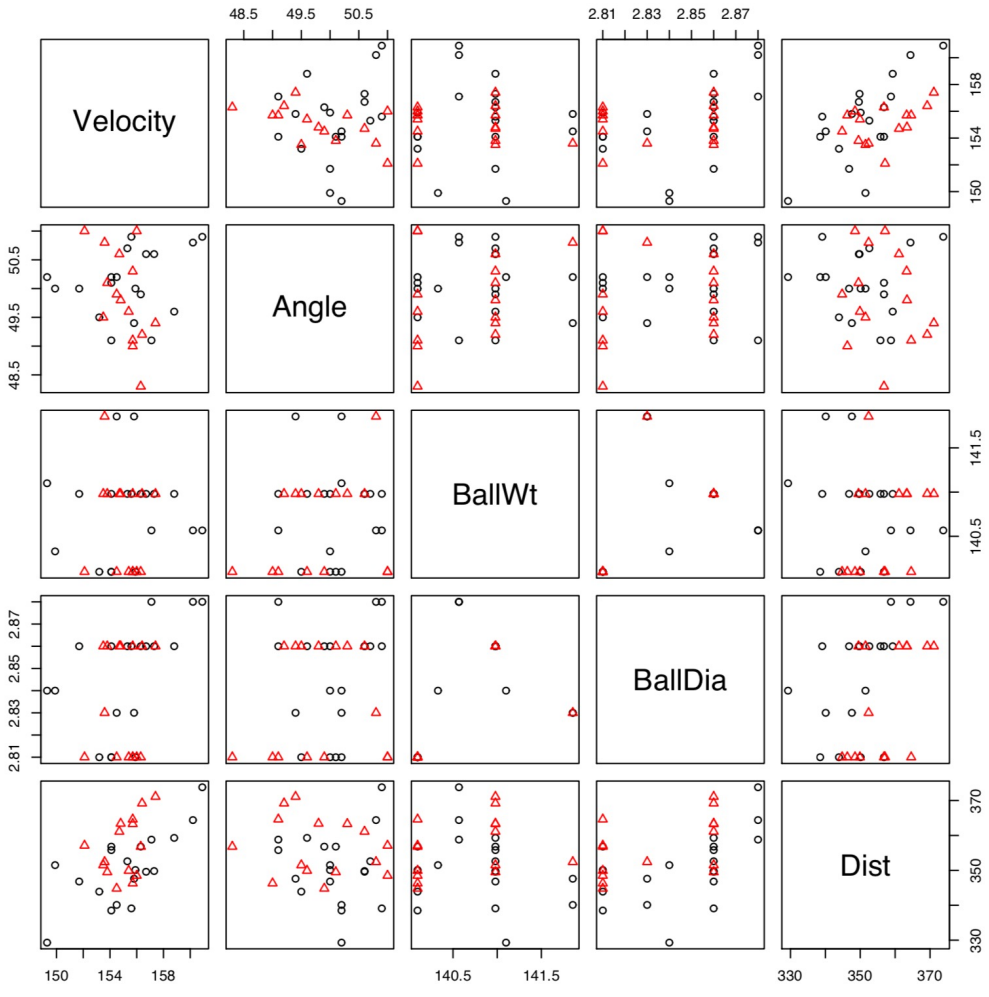
Question L2 continues on the next page.



Summarize the plot. Based on the boxplot, can we conclude that there is enough evidence that manipulating the fans can change the distance that a baseball travels? 2 MARKS

2. We next examine the scatterplot matrix of the response and the continuous predictors, using *Cond* to color and mark the points. Summarize the key features of the following graph. 3 MARKS

Question L2 continues on the next page.



3. Interpret the ANOVA output for `m1`, `m2` and `m3`. In particular, define appropriate hypotheses for these tests. Summarize the conclusions to be made from this output. 5 MARKS

```
> m1 <- lm( Dist ~ Velocity + Angle)
> m2 <- lm( Dist ~ Velocity + Angle + BallWt + BallDia)
> m3 <- lm( Dist ~ Velocity + Angle + BallWt + BallDia + Cond)
> anova(m1,m2,m3)
Analysis of Variance Table

Model 1: Dist ~ Velocity + Angle
Model 2: Dist ~ Velocity + Angle + BallWt + BallDia
Model 3: Dist ~ Velocity + Angle + BallWt + BallDia + Cond
  Res.Df    RSS Df Sum of Sq      F   Pr(>F)
1     31 2042.2
2     29 1747.0  2    295.15 3.1869 0.056627 .
3     28 1296.6  1    450.46 9.7279 0.004177 **
```

4. Test to see if the *Velocity* differential in *Dist* is the same in each *Cond*. Clearly state the null and alternative hypotheses of the tests. Comment on your findings. 5 MARKS

```
> ma <- lm(Dist ~ Velocity + Angle + BallWt + BallDia + Cond)
> mb <- update(ma, ~.+Velocity:Cond)
> anova(ma,mb)
Analysis of Variance Table

Model 1: Dist ~ Velocity + Angle + BallWt + BallDia + Cond
Model 2: Dist ~
Velocity + Angle + BallWt + BallDia + Cond + Velocity:Cond
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     28 1296.6
2     27 1296.5  1  0.078273 0.0016 0.9681
```

5. Using the this data, one fitted mean function is:

$$E(Dist|Velocity, Cond) = -24 + 2Velocity - 22Cond + 0.1Velocity \times Cond$$

   (a) Give the coefficients in the estimated mean function if *Cond* were coded so tailwind had the value 2 and headwind had the value 1 (the original coding given in the data file is 0 for headwind and 1 for tailwind). 2.5 MARKS

   (b) Give the coefficients if *Cond* is coded as -1 for headwind and +1 for tailwind. 2.5 MARKS

L3. The `twins` data give the IQ scores of identical twins, one raised in a foster home, *IQf*, and the other raised by birth parents, *IQb*. The data were published by Burt (1966), and their authenticity has been questioned. For purposes of this example, the twin pairs can be divided into three social classes C, *low*, *middle* or *high*, coded 1, 2, and 3, respectively, according to the social class of the birth parents. Treat *IQf* as the response and *IQb* as the predictor, with C as a factor. We perform an analysis of these data.

Be sure to draw and discuss a relevant graph. Are the within-class mean functions straight lines? Are there class differences? If there are differences, what are they?and Snyder, 1954).

1. For this question, use only the data from the low social class. 4 MARKS

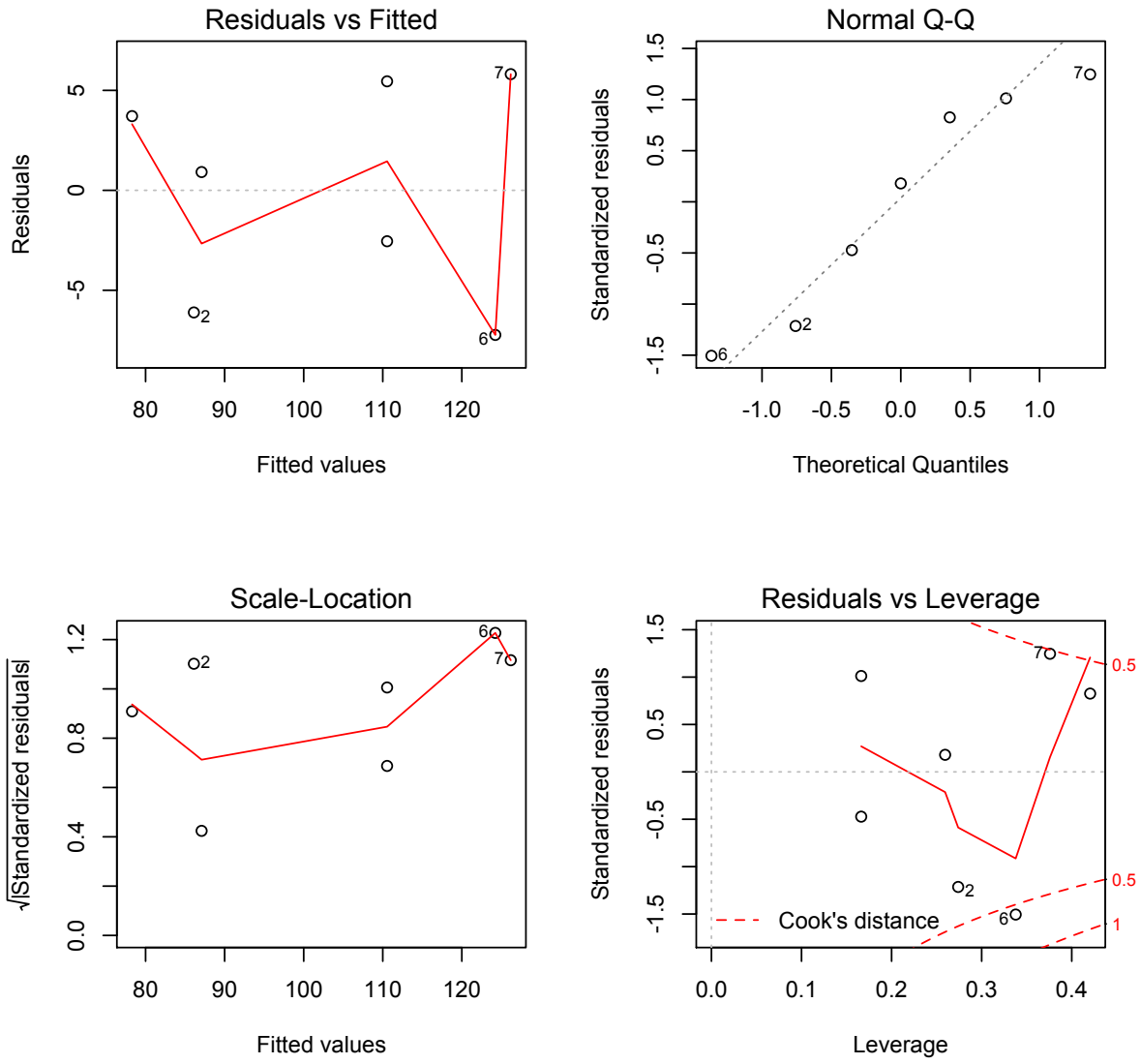
   (a) Fit the simple linear regression model with mean function

   $$E(IQf|IQb) = \beta_0 + \beta_1 IQb \tag{1}$$

   Using the R-output below, comment on the contribution of the covariate IQb in the fitted model. The significance and also performance of the fitted model. Explicitly state the hypotheses and test, and the conclusions.

```
> attach(twins)
> m0 <- lm(IQf ~ IQb, data=subset(twins,C==1))
> summary(m0)

Call:
lm(formula = IQf ~ IQb, data = subset(twins, C == 1))

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -1.8720    13.2725  -0.141 0.893339
IQb           0.9776     0.1216   8.037 0.000482 ***
---

Residual standard error: 5.903 on 5 degrees of freedom
Multiple R-squared:  0.9282,    Adjusted R-squared:  0.9138
F-statistic:  64.6 on 1 and 5 DF,  p-value: 0.0004823
```

   (b) Make comments on the following diagnostic plots.

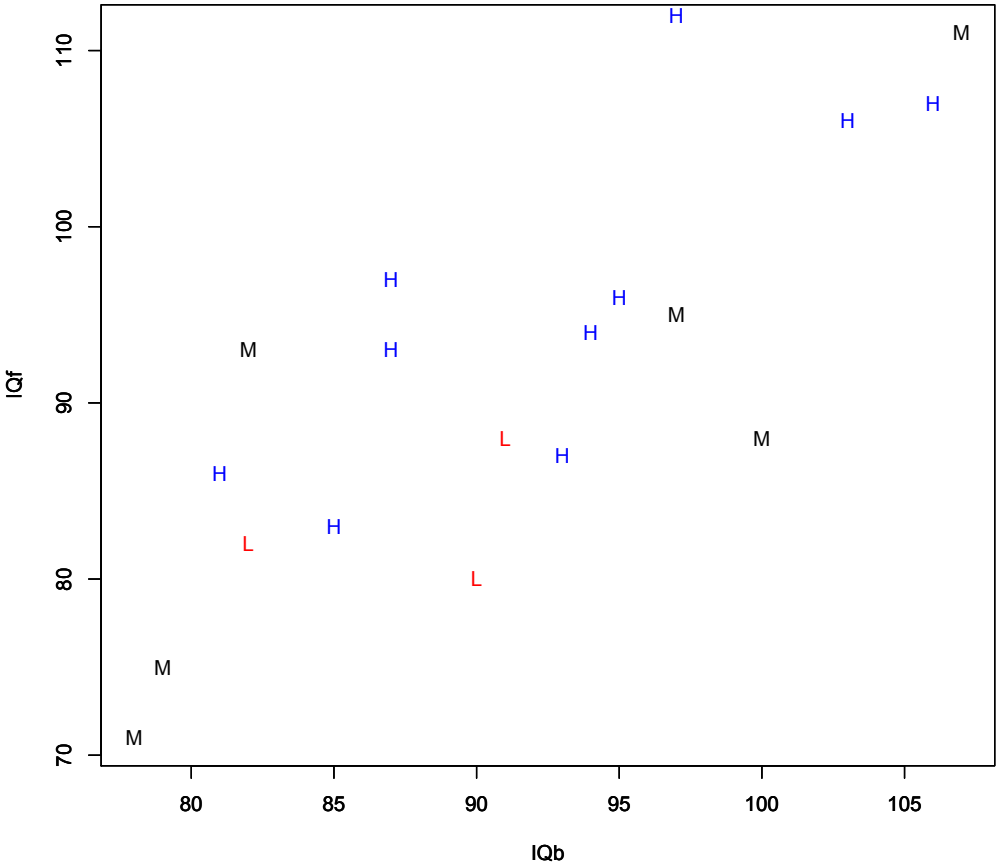


2. Now we fit the following model using the full dataset and obtain the sequential analysis of variance table for fitting the variables. State the hypotheses tested, and the conclusions for each of the tests. 4 MARKS

```
> m1 <- lm(IQf ~ IQb + factor(twins$C) + factor(twins$C):IQb, data=twins)
> anova(m1)
Analysis of Variance Table

Response: IQf
                     Df Sum Sq Mean Sq F value    Pr(>F)
IQb                   1 5231.1  5231.1 83.3823 9.279e-09 ***
factor(twins$C)       2  175.1    87.6  1.3958    0.2697
IQb:factor(twins$C)   2    0.9     0.5  0.0074    0.9926
Residuals            21 1317.5    62.7
---
```

Question L3 continues on the next page.

3. The scatterplot of IQf versus IQb is provided below, using a different symbol for social classes, *low*, *middle* or *high*. Comment on the information in the graph about an appropriate mean function for these data 2 MARKS



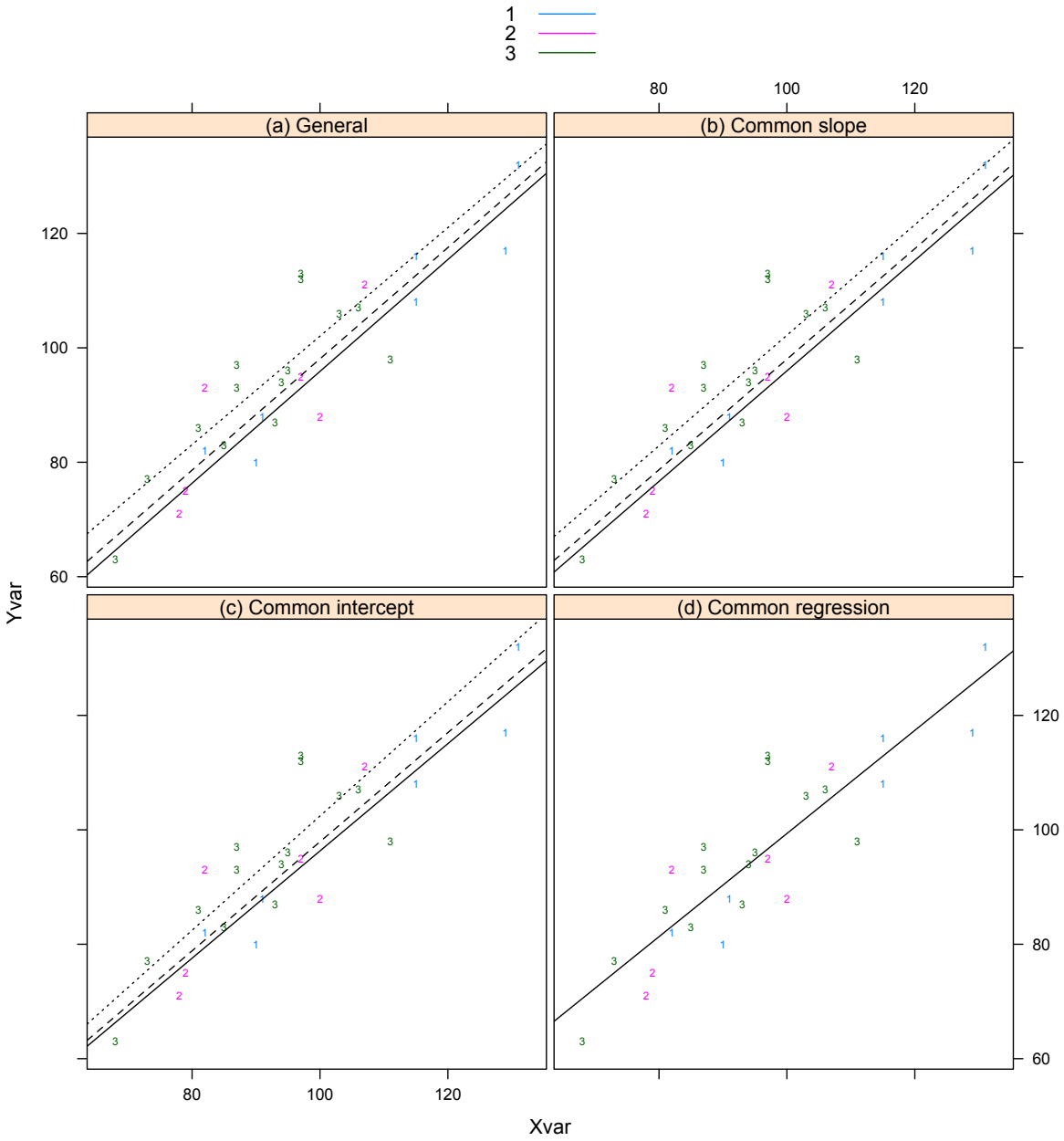
Question L3 continues on the next page.

4. Now we fit four models with different mean functions. Summarize the conclusions to be made from the following analysis in R: 4 MARKS

```
> model1 <- lm(IQf ~ IQb + factor(twins$C) + Fvar:IQb)
> model2 <- lm(IQf ~ IQb + factor(twins$C))
> model3 <- lm(IQf ~ IQb + factor(twins$C):IQb)
> model4 <- lm(IQf ~ IQb)
> anova(model4,model2,model1)
Analysis of Variance Table

Model 1: IQf ~ IQb
Model 2: IQf ~ IQb + factor(twins$C)
Model 3: IQf ~ IQb + factor(twins$C) + Fvar:IQb
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     25 1493.5
2     23 1318.4  2   175.131 1.3958 0.2697
3     21 1317.5  2     0.932 0.0074 0.9926
> anova(model4,model3,model1)
Analysis of Variance Table

Model 1: IQf ~ IQb
Model 2: IQf ~ IQb + factor(twins$C):IQb
Model 3: IQf ~ IQb + factor(twins$C) + Fvar:IQb
  Res.Df    RSS Df Sum of Sq      F Pr(>F)
1     25 1493.5
2     23 1326.5  2   167.074 1.3315 0.2855
3     21 1317.5  2     8.989 0.0716 0.9311
> drop1(model1,test="F")
Single term deletions

Model:
IQf ~ IQb + factor(twins$C) + Fvar:IQb
                Df Sum of Sq    RSS    AIC F value Pr(>F)
<none>                       1317.5 116.97
factor(twins$C)  2    8.9888 1326.5 113.15  0.0716 0.9311
IQb:Fvar         2    0.9318 1318.4 112.98  0.0074 0.9926
```

Question L3 continues on the next page.

5. A visual display of the four models, `model1`, `model2`, `model3` and `model4`, considered in Question 4 is provided in the figure below. Identify the corresponding model in each sub-figure. 3 MARKS



6. Now consider the problem with two continuous predictors rather than one, i.e. the regression of IQf on IQb and EQb (another continuous variable) and the grouping factor C. Assuming **no interaction** between IQb and EQb, suppose that we want to test for the null hypothesis that the regression planes are parallel for different C versus the alternative that separate planes are required for each C. Write down the psudo `R` code for this test using the `lm` and `anova` commands. 3 MARKS

G1. (a) Consider a family of distributions with parameters $\mu > 0$ and $\theta_Z > 0$ and the probability mass function given by

$$f(y;\mu,\theta_z) = \frac{\Gamma(y+\theta_z)}{\Gamma(y+1)\Gamma(\theta_z)} \left(\frac{\theta_z}{\mu+\theta_z}\right)^{\theta_z} \left(\frac{\mu}{\mu+\theta_z}\right)^{y}, \quad y = 0, 1, \ldots.$$

Show whether or not this family is an exponential dispersion family. 3 MARKS

(b) Formulate the Poisson rate model, explain in which situations it may be useful, and how it relates to a binomial GLM. 4 MARKS

Consider the following study on non-melanoma skin cancer among Caucasians in two areas of the United States. The variables collected were the number of `cases`, the population size `n`, the `age.range` with values `15--24`, `25-34`, `35-44`, `45-54`, `55-64`, `65-74`, `75-84`, `85+`, and `city` which is either `Dallas` or `Minneapolis`. Also available is `age.score`, assigning a score from 1 to 8 to each of the age.range categories.

(c) Based on the output lines 1–11, decide which one of the Poisson GLMs fitted to the data is most appropriate. Clearly formulate the test you use and test at the 5% level. Is there evidence that age has a different impact on the rate of non-melanoma skin cancer in Dallas then in Minneapolis? 4 MARKS

(d) Consider the model `mod4` (output lines 15–43). Calculate the parameter estimates and their standard errors when the quasipoisson model is used, and decide which predictors remain significant at the 5% level. 4 MARKS

(e) Using `mod4`, calculate the expected rate of non-melanoma skin cancer for someone aged between 45 and 55 living in Minneapolis. 3 MARKS

(f) Which other model would you consider fitting to these data and why? 2 MARKS

R Code and Output for Question G1

```
Analysis of Deviance Table

Model 1: cases ~ offset(log(n)) + city
Model 2: cases ~ offset(log(n)) + age.score
Model 3: cases ~ offset(log(n)) + city + age.score
Model 4: cases ~ offset(log(n)) + city * age.score

  Resid. Df Resid. Dev Df Deviance
1        14    2569.15
2        14     459.63  0  2109.51
3        13     190.56  1   269.07
4        12     187.64  1     2.92


> mod4 <- glm(formula = cases ~ offset(log(n)) + city * age.score, family = "poisson",
    data = nonmel)

> summary(mod4)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-8.1029  -3.4510  -0.0648   1.9230   4.0486

Coefficients:
                     Estimate Std. Error z value Pr(>|z|)
(Intercept)          -9.94374    0.14598 -68.116  < 2e-16 ***
cityDallas            1.09243    0.16993   6.429 1.29e-10 ***
age.score             0.63555    0.02471  25.721  < 2e-16 ***
cityDallas:age.score -0.04962    0.02918  -1.700   0.0891 .
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 2789.68  on 15  degrees of freedom
Residual deviance:  187.64  on 12  degrees of freedom
AIC: 289.89

Number of Fisher Scoring iterations: 5

> sum((residuals(mod4,type="pearson"))^2)

[1] 145.206
```

G2. (a) Calculate the deviance residuals for a binomial GLM with responses $m_i y_i \sim \mathcal{B}(\pi, m_i)$, $i = 1, \ldots, n$. Prove that the residuals simplify to

$$r_{D_i} = \text{sign}(y_i - \hat{\pi}_i)\sqrt{-2\log(1 - |y_i - \hat{\pi}_i|)}$$

when $m_i = 1$ for $i = 1, \ldots, n$. Sketch how the residuals would look like in this case if they are plotted against the linear predictor, and explain the consequences of this for residual model diagnostics. 7 MARKS

Consider the following data on home-well contamination in 3020 households in Araihazar Upazila, Bangladesh. The response variable is `switch` (binary variable whether or not the household switched to another well from an unsafe well). Other variables collected for each household were `arsenic` (the level of arsenic contamination in the household's original well, in hundreds of micrograms per liter), `dist100` (distance in 100-meter units to the closest known safe well), `educ` (years of education of the head of the household) and `assoc` (whether or not any members of the household participated in any community organizations: no or yes).

(b) Based on output lines 1–11, find the most appropriate model for these data, among the models that appear on lines 3–6. 4 MARKS

(c) In Bangladesh, primary education is 1–8 years, and secondary and above more than 9 years. Consider the predictor `feduc` whose value is 0 if `educ` is at most 8, and 1 otherwise. Interpret the model whose summary is on output lines 13–37. 5 MARKS

(d) Describe in detail one graphical technique that could be used to choose between the models `switch~arsenic+dist100+feduc+dist100:feduc` and `switch~arsenic+dist100+educ+dist100:educ`. 4 MARKS

R Code and Output for Question 2

```
Analysis of Deviance Table

Model 1: switch ~ arsenic * dist100 * assoc * educ
Model 2: switch ~ arsenic + dist100 + assoc + educ + dist100:educ
Model 3: switch ~ arsenic + dist100 + assoc + educ
Model 4: switch ~ arsenic + dist100 + educ + dist100:educ
  Resid. Df Resid. Dev  Df Deviance
1      3004     3876.0
2      3014     3893.1 -10  -17.076
3      3015     3907.8  -1  -14.703
4      3015     3896.2   0   11.676

Call:
glm(formula = switch ~ arsenic + dist100 + feduc + dist100:feduc,
    family = "binomial")

Deviance Residuals:
   Min      1Q  Median      3Q     Max
-2.663  -1.190   0.740   1.044   1.882

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   -0.01489    0.08606  -0.173    0.863
arsenic        0.47840    0.04205  11.378  < 2e-16 ***
dist100       -1.15433    0.12112  -9.530  < 2e-16 ***
feduc          0.04607    0.15625   0.295    0.768
dist100:feduc  1.20655    0.26380   4.574 4.79e-06 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1  1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 4118.1  on 3019  degrees of freedom
Residual deviance: 3869.0  on 3015  degrees of freedom
AIC: 3879

Number of Fisher Scoring iterations: 4
```

G3. (a) Suppose that $Y_1, \ldots, Y_K$ are independent Poisson variables with means $\mu_1, \ldots, \mu_K$. Calculate the conditional distribution of $(Y_1, \ldots, Y_K)$ given $Y_1 + \ldots + Y_K = m_K$.

4 MARKS

(b) By inspection of the log-likelihoods, and using suitable reparametrization, demonstrate the equivalence of parameter estimates obtained using Poisson log-linear modeling and multinomial modeling. 4 MARKS

Consider the following study on diabetes and retinopathy status reported by Bender and Grouven (1998). The question of interest is how retinopathy status is associated with risk factors. The variables collected are `RET` with values 1 (no retinopathy), 2 (nonproliferative retinopathy) and 3 (advanced retinopathy or blind), as well as `SM` (1: smoker, 0: nonsmoker), `DIAB` (diabetes duration in years), `GH` (glycosylated hemoglobin measured in percent), and `BP` (diastolic blood pressure in mmHg).

(c) A baseline category logit model `Model 1` was fitted to these data; its summary appears in output lines 1–16. Formulate this model. Furthermore, using `Model 1`, calculate the probability that a smoker who has had diabetes for 10 years, whose level of glycosylated hemoglobin is 10%, and whose diastolic blood pressure is 80 mmHg suffers from an advanced retinopathy. How would this probability change if the same patient were not smoking? 5 MARKS

(d) Based on the output provided, is smoking a significant risk factor for retinopathy?

4 MARKS

(e) Describe in detail an alternative to the baseline category logit model that could have been used to analyze these data. 3 MARKS

R Code and Output for Question Q3

```
Call:
multinom(formula = RET ~ SM + DIAB + GH + BP, data = retinopathy)

Coefficients:
  (Intercept)         SM       DIAB        GH         BP
1   -8.801117 0.59433108 0.09231948 0.3035673 0.04451891
2  -18.407243 0.08424936 0.18992340 0.6645750 0.10445957

Std. Errors:
  (Intercept)        SM       DIAB        GH         BP
1    1.577684 0.2382201 0.01659192 0.0917103 0.01704098
2    2.008648 0.2795210 0.02041861 0.1105393 0.01961002

Residual Deviance: 895.2518
AIC: 915.2518
Likelihood ratio tests of Multinomial Models

Response: RET
                Model Resid. df Resid. Dev
1      DIAB + GH + BP      1218   901.9623
2 SM + DIAB + GH + BP      1216   895.2518
```