Question 1 (20 marks)
Please tick the option(s) that is/are always TRUE.
2 marks are awarded for every correctly ticked or unticked option.

☐ (EXAMPLE) $2 + 1 = 0$

☑ (EXAMPLE) $1 + 1 = 2$

☐ (a) In an Exponential regression setup, let $\{Y_i\}_{i=1}^n$ denote independent observations with distribution $Y_i \sim \text{Exponential}(g^{-1}(\mathbf{x}_i^T\boldsymbol{\theta}))$, where g denotes the link function. Let $\mu_i = E(Y_i)$. Then, a suitable specification for g is $g(\mu_i) = \log(\mu_i)$, for $\mu_i > 0$.

☐ (b) Refer to part (a). Another suitable specification for g is $g(\mu_i) = \sin(\mu_i)$, for $\mu_i > 0$.

☐ (c) In a Binomial regression setup, let $\{Y_i\}_{i=1}^n$ denote independent observations with distribution $Y_i \sim \text{Binomial}(m_i, p_i)$, where $p_i = \Phi(\mathbf{x}_i^T\boldsymbol{\theta})$, Φ denotes the standard normal CDF, and $\boldsymbol{\theta} = (\theta_0, \theta_1, \theta_2)^T$ is a 3-dimensional column vector . We want to test $H_0 : \theta_2 = 0$ against $H_1 : \theta_2 \neq 0$. Then, the Wald's test and the likelihood ratio test are equivalent.

☐ (d) Consider the saturated model for independent observations $Y_i \sim \text{N}(\mu_i, \sigma^2)$ for $i = 1, \ldots, n$. Then, the MLE for μ_i is $\widetilde{\mu}_i = Y_i$ for $i = 1, \ldots, n$.

☐ (e) $\mathbf{X}$ is a n by p matrix such that $r = r(\mathbf{X}) < p$. $\mathbf{C}$ is a full rank p by r matrix. Then, $\mathbf{XC}$ is a full rank matrix.

☐ (f) $\mathbf{X}$ is a n by p matrix such that $n > p$ and $r = r(\mathbf{X}) < p$. $\mathbf{C}$ is a p by r matrix such that $\mathbf{XC}$ is a full rank matrix. Then, $\mathbf{C}$ is a matrix with entries equal to either 0 or 1.

☐ (g) $\mathbf{X}$ is the design matrix corresponding to the one-way ANOVA model with 3 levels: $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, where $i = 1, 2, 3$; $j = 1, \ldots, 7$, and $\epsilon_{ij} \sim N(0, \sigma^2)$. Let

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

Then, $\mathbf{XC}$ is a full rank matrix.

☐ (h) $\mathbf{X}$ is the design matrix corresponding to the one-way ANOVA model with 4 levels: $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, where $i = 1, 2, 3, 4$, $j = 1, \ldots, 7$, and $\epsilon_{ij} \sim N(0, \sigma^2)$. Let

$$\mathbf{C} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \end{pmatrix}.$$

Then, $\mathbf{XC}$ is a full rank matrix.

☐ (i) Consider the unit variance Gaussian density $f(x; \mu) = (2\pi)^{-1/2} \exp\{-\frac{1}{2}(x - \mu)^2\}$, for $x \in \mathbb{R}$, where the parameter is $\mu \in \mathbb{R}$. Then, f belongs to the exponential family of distributions.

☐ (j) Consider the Inverse-Gamma density $f(x; \alpha, \beta) = (\beta^\alpha/\Gamma(\alpha))(x)^{-\alpha-1} \exp\{-\beta x\}$, for $x \geq 0$, where the parameters are $\alpha > 0$ and $\beta > 0$. Then, f belongs to the exponential family of distributions.

Question 2 (18 marks)
A drug company wishes to compare the percentage change in tumour size induced by three drugs in a clinical trial. Patients are randomly assigned to three drug groups A, B and C. Assume that patients' responses are independent of each other.

	Change in tumour size (y)					
Drug A:	7	-8	-10	-7	-9	
Drug B:	3	8	4	-4	-7	-6
Drug C:	-9	3	-4	-3	-2	-5

We consider a one-factor classification model $y_{ij} = \mu + \tau_i + \epsilon_{ij}$, where $\epsilon_{ij} \sim N(0, \sigma^2)$ and $i = 1, 2, 3$; $j = 1, \ldots, n_i$

(a) (4 marks) Express the one-factor ANOVA model in terms of its less-than-full-rank matrix form: $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. Provide full details of the entries and dimension of $\mathbf{y}$, $\mathbf{X}$, $\boldsymbol{\beta}$, and $\boldsymbol{\epsilon}$.

$n = 17$

An n by 1 vector

$$\mathbf{y} = (7, -8, -10, -7, -9, 3, 8, 4, -4, -7, -6, -9, 3, -4, -3, -2, -5)^T$$

A n by 1 vector

$$\boldsymbol{\epsilon} = (\epsilon_{11}, \epsilon_{12}, \epsilon_{13}, \epsilon_{14}, \epsilon_{15}, \epsilon_{21}, \epsilon_{22}, \epsilon_{23}, \epsilon_{24}, \epsilon_{25}, \epsilon_{26}, \epsilon_{31}, \epsilon_{32}, \epsilon_{33}, \epsilon_{34}, \epsilon_{35}, \epsilon_{36})^T$$

A 4 by 1 vector

$$\boldsymbol{\beta} = (\mu, \tau_1, \tau_2, \tau_3)^T.$$

An n by 4 matrix

$$\mathbf{X} = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \\ 1 & 0 & 0 & 1 \end{pmatrix}$$

(b) (4 marks) Use the below matrix to reparameterise your model in part (a) as a full-rank linear model:

$$\mathbf{C} = \begin{pmatrix} 0 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

Provide full details of the full-rank design matrix and coefficient vector in your reparameterised model. Also, express the coefficients in your full-rank linear model as linear combinations of $\boldsymbol{\beta}$.

The reparameterised model is of the form:

$$\mathbf{y} = \widetilde{\mathbf{X}}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)^T$ is a 3 by 1 vector and

$$\widetilde{\mathbf{X}} = \mathbf{XC} = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}$$

and $\gamma_1 = \mu + \tau_1 = (1\,1\,0\,0)\boldsymbol{\beta}$, $\gamma_2 = \mu + \tau_2 = (1\,0\,1\,0)\boldsymbol{\beta}$, and $\gamma_3 = \mu + \tau_3 = (1\,0\,0\,1)\boldsymbol{\beta}$

(c) (4 marks) Compute the sample means of observations from each group: Drug A, Drug B, Drug C.

The samples means are, $\widehat{\mu}_1 = -5.4$, $\widehat{\mu}_2 = -1/3$ and $\widehat{\mu}_3 = -10/3$

(d) (6 marks) A clinician hypothesize that all drug induces the same amount of change in tumour size. Verify that the clinician's hypothesis is testable and compute the corresponding test statistic. What is your conclusion at 5% significance level?

```
> vy1 <- c(7,-8,-10,-7,-9); sum((vy1 - mean(vy1))^2)
[1] 197.2
> vy2 <- c(3,8,4,-4,-7,-6); sum((vy2 - mean(vy2))^2)
[1] 189.3333
> vy3 <- c(-9,3,-4,-3,-2,-5); sum((vy3 - mean(vy3))^2)
[1] 77.33333
> c(qf(0.95,2,14),qf(0.95,1,14),qf(0.95,1,13))
[1] 3.738892    4.600110    4.667193
```

The hypothesis can be written as $H_0 : \tau_1 = \tau_2 = \tau_3$. By some algebra, the null hypothesis can be re-expressed as $\mu + \tau_2 - \mu - \tau_1 = \gamma_2 - \gamma_1$ and $\mu + \tau_3 - \mu - \tau_2 = \gamma_3 - \gamma_2$. Note that the LHS of the original hypothesis can be written as a linear combination of the coefficients of the reparameterised model.

Note that

$$\widetilde{L} = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix}.$$

Note that $\widetilde{L}$ is a full rank matrix. Hence, H_0 is a testable hypothesis.
Moreover, $SS_{res} = \sum_{i=1}^{n}(y_{ij} - \widehat{\gamma}_i)^2 = 463.8667$ Also, note that $(\widetilde{X}^T\widetilde{X})^{-1}$ is diagonal matrix with entries 1/5, 1/6 and 1/6. Plug everything into the F-statistics formula

$$f = \frac{(\widetilde{L}\widehat{\gamma})^T\{\widetilde{L}(\widetilde{X}^T\widetilde{X})^{-1}\widetilde{L}^T\}^{-1}(\widetilde{L}\widehat{\gamma})/2}{SS_{res}/(n-3)} = \frac{71.89804/2}{463.8666/14} = 1.085 \sim F_{2,14}.$$

Since F-statistics is smaller than `qf(0.95,2,14)`, we don't reject H_0 and thus the clinician's claim is indeed plausible.

Question 3 (16 marks)
Pearlie conducted an experiment to determine the influence of temperature (`Temp`) and moisture (`Moisture`) on the health (`Health`) of Bonsai trees. Three different temperature settings (Low, Medium, and High) and two moisture settings (Low, High) were considered. Health scores range between 0 to 10 with higher scores indicating better overall health.

(a) (4 marks) In the initial phase of the experiment, she grew a total of $n = 6$ bonsai plants – one plant under each combination of temperature-moisture settings. She used the following codes to fit an ANOVA model.

```
> names(BonsaiData)
[1] "Temp"     "Moisture" "Health"
> BonsaiData$Health
[1] 4 6 3 2 9 7
> BonsaiData$Temp
[1] Low  Med  High Low  Med  High
Levels: High Low Med
> BonsaiData$Moisture
[1] "Low"  "Low"  "Low"  "High" "High" "High"
> summary(lm(Health~Temp))

Call:
lm(formula = Health ~ Temp)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    5.000      1.555   3.216   0.0487 *
TempLow       -2.000      2.198  -0.910   0.4300
TempMed        2.500      2.198   1.137   0.3381
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.198 on 3 degrees of freedom
Multiple R-squared:  0.5837,Adjusted R-squared:  0.3062
F-statistic: 2.103 on 2 and 3 DF,  p-value: 0.2686
```

Provide details of the fitted reparameterised model in matrix form: $\mathbf{y} = \widetilde{\mathbf{X}}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$, where $\widetilde{\mathbf{X}}$ is a full-rank design matrix of dimension 6 by 3.

$\mathbf{y} = (3, 7, 4, 2, 6, 9)^T$. $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \gamma_3)^T$. $\boldsymbol{\epsilon} = (\epsilon_{11}, \epsilon_{12}, \epsilon_{21}, \epsilon_{22}, \epsilon_{31}, \epsilon_{32})^T$. The first and second rows of $\widetilde{\mathbf{X}}$ is $(1, 0, 0)$. The third and fourth rows of $\widetilde{\mathbf{X}}$ is $(1, 1, 0)$. The fifth and sixth rows of $\widetilde{\mathbf{X}}$ is $(1, 0, 1)$. .

(b) (6 marks) In the second phase of the experiment, Pearlie grew an additional four bonsai plants under various combinations of temperature and moisture settings. Data from both phases of the experiment are combined and used to refit the model in part (a). The ANOVA output of the re-fitted model is as follows:

```
> str(AdditionalBonsaiData)
'data.frame': 4 obs. of  3 variables:
$ Temp    : chr  "Low" "Med" "High" "High"
$ Moisture: chr  "Low" "High" "Low" "High"
$ Health  : num  3 7 2 7
> CombineData = rbind(BonsaiData,AdditionalBonsaiData)
> anova(lm(Health~Temp*Moisture,data=CombineData))
Analysis of Variance Table

Response: Health
              Df Sum Sq Mean Sq F value   Pr(>F)
Temp           2 28.583 14.2917 19.0556 0.009022 **
Moisture       1 10.012 10.0119 13.3492 0.021701 *
Temp:Moisture  2 14.405  7.2024  9.6032 0.029710 *
Residuals      4  3.000  0.7500
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> c(qf(0.95,2,4),qf(0.98,2,4),qf(0.95,1,4),qf(0.98,1,4))
[1]  6.944272    12.142136     7.708647    14.039615
```

Use the R output to test H_0 : *no interaction effect between temperature and moisture* at 5% significance level. Include the null distribution of the test statistic, the value of your test statistic and conclusion in your answer.

Null distribution of test statistic: F distribution with 2 and 4 degrees of freedom Rejection region: Reject H_0 if test statistic is larger than 6.94427. OR Reject H_0 if pvalue < 0.05. F-statistic value $= 9.6032$. OR p-value $= 0.0297$ Conclusion: We reject H_0 and conclude that interaction effect is significant.

(c) (6 marks) Using the combined data from the two phases, Pearlie fitted a two-way ANOVA without interaction. The ANOVA output of the model is as follows:

```
> anova(lm(Health~Temp+Moisture,data=CombineData))
Analysis of Variance Table

Response: Health
          Df Sum Sq Mean Sq F value  Pr(>F)
Temp       2 28.583 14.2917  4.9268 0.05421 .
Moisture   1 10.012 10.0119  3.4514 0.11257
Residuals  6 17.405  2.9008
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
> c(qf(0.95,2,6),qf(0.98,2,6),qf(0.95,1,6),qf(0.98,1,6))
[1] 5.143253    8.052094    5.987378    9.876365
```

Use the R output to test H_0 : *no Moisture effect* at 5% significance level. Include the null distribution of the test statistic, the value of your test statistic and conclusion in your answer.

Null distribution of test statistic: F distribution with 1 and 6 degrees of freedom Rejection region: Reject H_0 if test statistic is larger than 5.987378. OR Reject H_0 if pvalue < 0.05. F-statistic value $= 3.4514$. OR p-value $= 0.11257$ Conclusion: We don't reject H_0 and conclude that there is plausibly no moisture effect.

Question 4 (14 marks)
The data frame `beetle` contains data on the concentration, `x`, of insecticides given to groups of beetles of size, `m`, and the subsequent number that died, `y`. Here is a table of the `beetle` data frame:

x	1.6907	1.7242	1.7552	1.7842	1.8113	1.8369	1.8610	1.8839
m	59	60	62	56	63	59	62	60
y	6	13	18	28	52	53	61	60

A binomial model is to be used to model the log odds of a beetle dying. At concentration x_i, the log odds are assumed to be $\eta_i = \beta_1 + \beta_2 x_i$, with parameters $\boldsymbol{\beta} = (\beta_1, \beta_2)$.

(a) (6 marks) Write down the canonical link of a Binomial regression model. Write down the domain and range of the link function, and verify that it is a monotonic function.

$Y \sim \text{Binomial}(m, p)$, where $\mu = mp$ and $g(\mu) = \eta$. Canonical link:

$$g(\mu) = \log(\mu) - \log(m - \mu)$$

Domain: $(0, m)$. Range: $\mathbb{R}$. Take derivative $g'(\mu) = 1/\mu + 1/(m - \mu) > 0$ for all $\mu \in (0, m)$. Hence, g is monotonic.

OR

$$g(p) = \log(p) - \log(1 - p)$$

Domain: $(0, 1)$. Range: $\mathbb{R}$
Take derivative $g'(p) = 1/p + 1/(1 - p) > 0$ for all $p \in (0, 1)$. Hence, g is monotonic.

(b) (3 marks) We found that the maximum likelihood estimates of $\boldsymbol{\beta}$ is $\widehat{\boldsymbol{\beta}} = \begin{bmatrix} -60.71746 & 34.27033 \end{bmatrix}^T$ and the inverse of the corresponding Fisher Information matrix is:

$$\boldsymbol{\mathcal{I}}(\widehat{\boldsymbol{\beta}})^{-1} = \begin{bmatrix} 26.839771 & -15.0821509 \\ -15.082151 & 8.4805597 \end{bmatrix}$$

Find the 90% confidence interval for β_1. Is -60 a plausible value for β_1 at this confidence level?

```
> c(qnorm(0.95), qnorm(0.975))
[1] 1.644854    1.959964
```

$-60.71746 \pm 1.645\sqrt{26.839771} = (-69.23973, -52.19519)$
-60 is in the 90% CI, so it is plausible that $\beta_1 = -60$

(c) (5 marks) Find the 95% confidence interval for the probability that a beetle dies at insecticide concentration 1.7.

Let $\mathbf{x}_i = \begin{bmatrix} 1 & 1.7 \end{bmatrix}^T$, the CI for η_i is $\mathbf{x}_i^T\hat{\beta} \pm 1.96\sqrt{\mathbf{x}_i^T(\boldsymbol{\mathcal{I}}(\hat{\beta})^{-1})\mathbf{x}_i}$
We compute $\mathbf{x}_i^T(\boldsymbol{\mathcal{I}}(\hat{\beta})^{-1})\mathbf{x}_i = 0.069275$
So the 95% CI for η is $(-2.97377, -1.94202)$
Note that $p_i = 1/(1 + e^{-\eta_i})$, so the CI for p_i is $(0.0486, 0.1254)$

Question 5 (22 marks)
The data `moons` contains the diameter, mass, distance from the sun, and number of moons for 13 planets, gas giants, and dwarf planets in our solar system. The variables are

- `Name`: a character variable with the name of the planet, gas giant, or dwarf planet
- `Distance`: distance from sun, relative to earth's
- `Diameter`: diameter of the planet, relative to earth's
- `Mass`: mass, relative to earth's
- `Moons`: number of moons

The first 3 rows of the data are shown below:

Name	Distance	Diameter	Mass	Moons
Mercury	0.39	0.382	0.0600	0
Venus	0.72	0.949	0.8200	0
Earth	1.00	1.000	1.0000	1

We want to see if the number of moons of a planet is related to its size. We will use Poisson regression (with log link) to model the number of moons of a planet.

(a) (6 marks) We first fit a model that assumes the number of moons does not depend on any other variable in the data (Model 0). Next, we fit a model for the number of moons with predictors `Diameter` and `Mass`. We call this Model 1. The following R output is provided.

```
model0 <- glm(Moons~1,family = poisson(link = 'log'),data = moons)
model1 <- glm(Moons~ Mass + Diameter,family = poisson(link = 'log'),data = moons)
summary(model1)

# Call:
# glm(formula = Moons ~ Mass + Diameter, family = poisson(link = "log"),
#       data = moons)
#
# Deviance Residuals:
#      Min        1Q    Median        3Q       Max
# -2.45936  -2.05425  -0.91278   0.27366   4.32665
#
# Coefficients:
#                Estimate  Std. Error z value  Pr(>|z|)
# (Intercept)  0.71322809  0.19912109  3.5819 0.0003411 ***
# Mass        -0.00401033  0.00094609 -4.2388 2.247e-05 ***
# Diameter     0.41803308  0.03270228 12.7830 < 2.2e-16 ***
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for poisson family taken to be 1)
#
#     Null deviance: 388.2529  on 12  degrees of freedom
# Residual deviance:  44.8173  on 10  degrees of freedom
# AIC: 84.8331
> c(qchisq(0.95,10), qchisq(0.95,2),qnorm(0.975))
[1] 18.307038  5.991465  1.959964
```

Compared to Model 0, is Model 1 a better model for predicting the number of moons? Conduct an appropriate hypothesis test at 10% significance level and state clearly the test you use, the test statistic value, its null distribution, and the conclusion. *Hint: the Null deviance in the R output corresponds to the deviance of Model 0. The residual deviance corresponds to the deviance of Model 1.*

We can use the log-likelihood ratio test .
The test statistics is 388.2529 - 44.8173 = 343.4356 , follow a χ^2 distribution with degree of freedom 2 .
We reject H_0 if test statistic is larger than `qchisq(0.9,2)` which is not provided in the R output.
Since 343.4356 is greater than `qchisq(0.95,2)` and hence is greater than `qchisq(0.95,2)`, we reject the null hypothesis that Model 0 is correct. Model 1 is better/ correct.

(b) (5 marks) Use the R output in part (a) to answer this part: Test $H_0 : \beta_{Diameter} = 0$ at 5% signifiance level. State the test statistic, null distribution of your test statistic, p-value, and the conclusion.

Wald's test statistic: $\widehat{Z} = \widehat{\beta}_{Diameter}/\text{SE}(\widehat{\beta}_{Diameter}) = 0.41803308/0.03270228 = 12.7830$.
Under H_0, $\widehat{Z} \sim N(0, 1)$.
p-value from R output $< 2.2 \times 10^{-16}$.
Since p-value < 0.05, we reject H_0 and conclude that the Diameter coefficient is non-zero.

(c) (7 marks) We suspect that distance from sun is also useful in predicting the number of moons of a planet. We then add `Distance` to our model, and call this new model Model 2. The model's summary is shown below (with `MISSING` output):

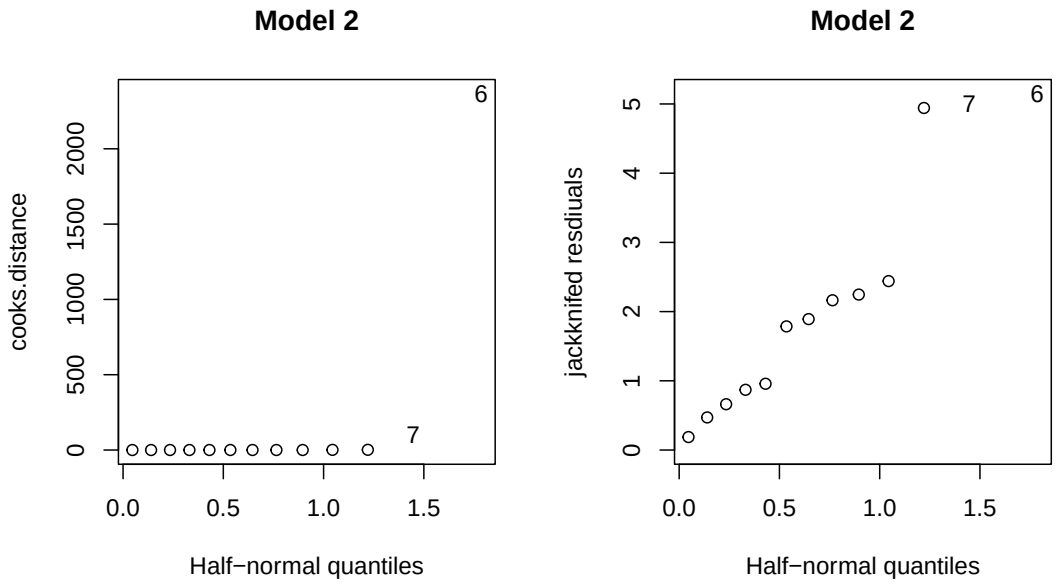
```
model2 <- glm(Moons~ Mass + Diameter + Distance, family = poisson(link = 'log'),
data = moons)
summary(model2)

# Call:
# glm(formula = Moons ~ Mass + Diameter + Distance, family = poisson(link = "log"),
#     data = moons)
#
# Deviance Residuals:
#      Min        1Q    Median        3Q       Max
# -2.34540  -1.73936  -0.87962   0.27045   4.40580
#
# Coefficients:
#               Estimate   Std. Error   z value    Pr(>|z|)
# (Intercept)   MISSING1   0.33261090   0.9390     0.34776
# Mass        -0.00392976  0.00094607  MISSING2   3.271e-05 ***
# Diameter     0.44500885  MISSING3     11.6350   < 2.2e-16 ***
# Distance     0.01410618  MISSING4     1.7067     0.08788 .
# ---
# Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
#
# (Dispersion parameter for poisson family taken to be MISSING5)
#
#     Null deviance: MISSING6  on 12  degrees of freedom
# Residual deviance:  41.8974  on  MISSING7  degrees of freedom
```

Compute all `MISSING` output

Missing1 $= 0.9390 \times 0.33261090 = 0.3123216$
Missing2 $= -0.00392976/0.00094607 = -4.153773$
Missing3 $= 0.44500885/11.6350 = 0.03824743$
Missing4 $= 0.01410618/1.7067 = 0.008265178$
Missing5 $= 1$ (dispersion parameter of Poisson distribution)
Missing6 $= 388.2529$ (null deviance is the same as part (a) since we are using same response variable and distributional assumption)
Missing7 $= n - 4 = 13 - 4 = 9$

(d) (4 marks) The figure below shows the half-normal plots of Cook's distance (left panel) and jackknife residuals (right panel) from Model 2. Can you identify any influential observations or outliers from the two plots?



The jackknife residuals and Cook's distance of 6th and 7th observations are large in magitude and deviate from the rest , collectively suggesting that they might be influential outliers .

Question 6 (10 marks)
A large postal survey on the psychology of debt was conduct with 381 participants. The data is imported in R under the name `debt`. In this data, the frequency of credit card use is a three-level factor ranging from never, through occasionally to regularly. We are interested in building a model for predicting credit card use as a function of the other variables. The variables in the data are:

- `ccarduse`: how often did s/he use credit cards (a factor of 3 levels: never, occasionally, regularly)
- `children`: number of children in household
- `manage`: self-rating of money management skill (high values=high skill)
- `locintrn`: score on a locus of control scale (high values=internal)
- `prodebt`: score on a scale of attitudes to debt (high values=favourable to debt)
- `singpar`: is the respondent a single parent ? (a factor with levels: 0 = No, 1 = Yes)

We will fit multinomial logit models to this data.

(a) (6 marks) We first assume that frequency of credit card use depends on all other variables except `manage`. We call this Model A. The summary of Model A is shown below.

```
# Call:
# multinom(formula = ccarduse ~ children + locintrn + prodebt +
#              singpar, data = debt)
#
# Coefficients:
#              (Intercept)  children  locintrn  prodebt  singpar1
# occasionally    -2.06177 -0.114668 0.0211954 0.407830 -0.432893
# regularly       -5.77064 -0.210404 0.4371748 0.955684 -0.463336
#
# Std. Errors:
#              (Intercept) children locintrn  prodebt singpar1
# occasionally    0.926473 0.116407 0.142538 0.185453 0.531529
# regularly       1.082938 0.127205 0.158297 0.199827 0.591830
```

According to model A, what is the probability of a person to "never", "occasionally" and "regularly" use their credit card, given that this person has 2 children, locus of control score 2, attitudes to debt score 3.0 and a single parent ?

For this model, $\eta_{ij} = \mathbf{x}_i^T \boldsymbol{\beta}_j$ for $j = 1, 2, 3$. $\mathbf{x}_i^T = \begin{bmatrix} 2 & 2 & 3.0 & 1 \end{bmatrix}$ We have $\beta_1 = 0$ so $\eta_{i1} = 0$

$$\eta_{i2} = -2.06177 + (-0.114668) \times 2 + 0.0211954 \times 2 + 0.407830 \times 3.0 - 0.432893 \times 1 = -1.458118$$

and

$$\eta_{i3} = -5.77064 + (-0.210404) \times 2 + (0.4371748) \times 2 + 0.955684 \times 3.0 - 0.463336 \times 1 = -2.913382$$

We have

$$p_{ij} = \frac{\exp(\eta_{ij})}{\sum_k \exp(\eta_{ik})},$$

so $p(\text{never}) = 0.7770216$, $p(\text{occasionally}) = 0.1807925$ and $p(\text{regularly}) = 0.0421859$ (round to any number of decimal places)

(b) (4 marks) We now add variable `manage` to Model A and call this Model B. The new model's summary is shown below. At significant level 5%, test the adequacy of model B.
The deviance of Model A and Model B are 734.3599 and 726.31306 respectively.

```
modelB <- multinom(ccarduse ~ ., debt)
summary(modelB,digits = 6)
# Call:
# multinom(formula = ccarduse ~ ., data = debt)
#
# Coefficients:
#              (Intercept)   children   manage    locintrn  prodebt  singpar1
# occasionally    -3.61581 -0.0930887 0.316138 -0.00812246 0.516055 -0.372842
# regularly       -7.55512 -0.1875308 0.373647  0.39712055 1.073635 -0.402110
#
# Std. Errors:
#              (Intercept) children   manage locintrn  prodebt singpar1
# occasionally     1.18869 0.118150 0.150050 0.143516 0.193305 0.535481
# regularly        1.34353 0.128873 0.160573 0.159870 0.208181 0.600022
> c(qchisq(0.95,2),qchisq(0.95,12),qchisq(0.95,369))
[1]   5.991465  21.026070 414.792164
```

The scaled deviance equals to deviance in multinomial regression. Hence, test statistics equals to deviance of model B $= 726.31306$, follows a χ^2 distribution with df $381 - 12 = 369$. The test statistic is greater than the critical value 414.7922 so we reject H_0 and conclude that model B is inadequate.

Additional writing space for any question commences on the next page

Additional answer space for any question—submit this page even if blank

Additional answer space for any question—submit this page even if blank

Additional answer space for any question—submit this page even if blank

Additional answer space for any question—submit this page even if blank

End of Exam — Total Available Marks = 100

Turn the page for appended material

You must tick this box if you have used extra booklets ☐

Table XII: Discrete Distributions

Probability Distribution and Parameter Values	Probability Mass Function	Moment-Generating Function	Mean $E(X)$	Variance $\mathrm{Var}(X)$	Examples
Bernoulli $0 < p < 1$ $q = 1 - p$	$p^x q^{1-x}$, $x = 0, 1$	$q + pe^t$	p	pq	Experiment with two possible outcomes, say success and failure, $p = P(\text{success})$
Binomial $n = 1, 2, 3, \ldots$ $0 < p < 1$	$\binom{n}{x} p^x q^{n-x}$, $x = 0, 1, \ldots, n$	$(q + pe^t)^n$	np	npq	Number of successes in a sequence of n Bernoulli trials, $p = P(\text{success})$
Geometric $0 < p < 1$ $q = 1 - p$	$q^{x-1}p$, $x = 1, 2, \ldots$	$\dfrac{pe^t}{1 - qe^t}$	$\dfrac{1}{p}$	$\dfrac{q}{p^2}$	The number of trials to obtain the first success in a sequence of Bernoulli trials
Hypergeometric $x \le n, x \le N_1$ $n - x \le N_2$ $N = N_1 + N_2$ $N_1 > 0, \ N_2 > 0$	$\dfrac{\binom{N_1}{x}\binom{N_2}{n-x}}{\binom{N}{n}}$		$n\left(\dfrac{N_1}{N}\right)$	$n\left(\dfrac{N_1}{N}\right)\left(\dfrac{N_2}{N}\right)\left(\dfrac{N-n}{N-1}\right)$	Selecting r objects at random without replacement from a set composed of two types of objects
Negative Binomial $r = 1, 2, 3, \ldots$ $0 < p < 1$	$\binom{x-1}{r-1} p^r q^{x-r}$, $x = r, r+1, \ldots$	$\dfrac{(pe^t)^r}{(1 - qe^t)^r}$	$\dfrac{r}{p}$	$\dfrac{rq}{p^2}$	The number of trials to obtain the rth success in a sequence of Bernoulli trials
Poisson $0 < \lambda$	$\dfrac{\lambda^x e^{-\lambda}}{x!}$, $x = 0, 1, \ldots$	$e^{\lambda(e^t - 1)}$	λ	λ	Number of events occurring in a unit interval, events are occurring randomly at a mean rate of λ per unit interval
Uniform $m > 0$	$\dfrac{1}{m}$, $x = 1, 2, \ldots, m$		$\dfrac{m+1}{2}$	$\dfrac{m^2 - 1}{12}$	Select an integer randomly from $1, 2, \ldots, m$

Table XIII: Continuous Distributions

Probability Distribution and Parameter Values	Probability Density Function	Moment-Generating Function	Mean $E(X)$	Variance $\mathrm{Var}(X)$	Examples
Beta $0 < \alpha$ $0 < \beta$	$\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$, $0 < x < 1$		$\dfrac{\alpha}{\alpha+\beta}$	$\dfrac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$	$X = X_1/(X_1 + X_2)$, where X_1 and X_2 have independent gamma distributions with same θ
Chi-square $r = 1, 2, \ldots$	$\dfrac{x^{r/2-1}e^{-x/2}}{\Gamma(r/2)2^{r/2}}$, $0 < x < \infty$	$\dfrac{1}{(1-2t)^{r/2}}$, $t < \dfrac{1}{2}$	r	$2r$	Gamma distribution, $\theta = 2$, $\alpha = r/2$; sum of squares of r independent $N(0, 1)$ random variables
Exponential $0 < \theta$	$\dfrac{1}{\theta} e^{-x/\theta}$, $0 \le x < \infty$	$\dfrac{1}{1-\theta t}$, $t < \dfrac{1}{\theta}$	θ	θ^2	Waiting time to first arrival when observing a Poisson process with a mean rate of arrivals equal to $\lambda = 1/\theta$
Gamma $0 < \alpha$ $0 < \theta$	$\dfrac{x^{\alpha-1}e^{-x/\theta}}{\Gamma(\alpha)\theta^\alpha}$, $0 < x < \infty$	$\dfrac{1}{(1-\theta t)^\alpha}$, $t < \dfrac{1}{\theta}$	$\alpha\theta$	$\alpha\theta^2$	Waiting time to αth arrival when observing a Poisson process with a mean rate of arrivals equal to $\lambda = 1/\theta$
Normal $-\infty < \mu < \infty$ $0 < \sigma$	$\dfrac{e^{-(x-\mu)^2/2\sigma^2}}{\sigma\sqrt{2\pi}}$, $-\infty < x < \infty$	$e^{\mu t + \sigma^2 t^2/2}$	μ	σ^2	Errors in measurements; heights of children; breaking strengths
Uniform $-\infty < a < b < \infty$	$\dfrac{1}{b-a}$, $a \le x \le b$	$\dfrac{e^{tb} - e^{ta}}{t(b-a)}$, $t \ne 0$ 1, $t = 0$	$\dfrac{a+b}{2}$	$\dfrac{(b-a)^2}{12}$	Select a point at random from the interval $[a, b]$

Appendix C Tables