Stat 431 Assignment 3 Spring 2021

Due by 4:00pm EDT on Friday July 23, 2021 via Crowdmark

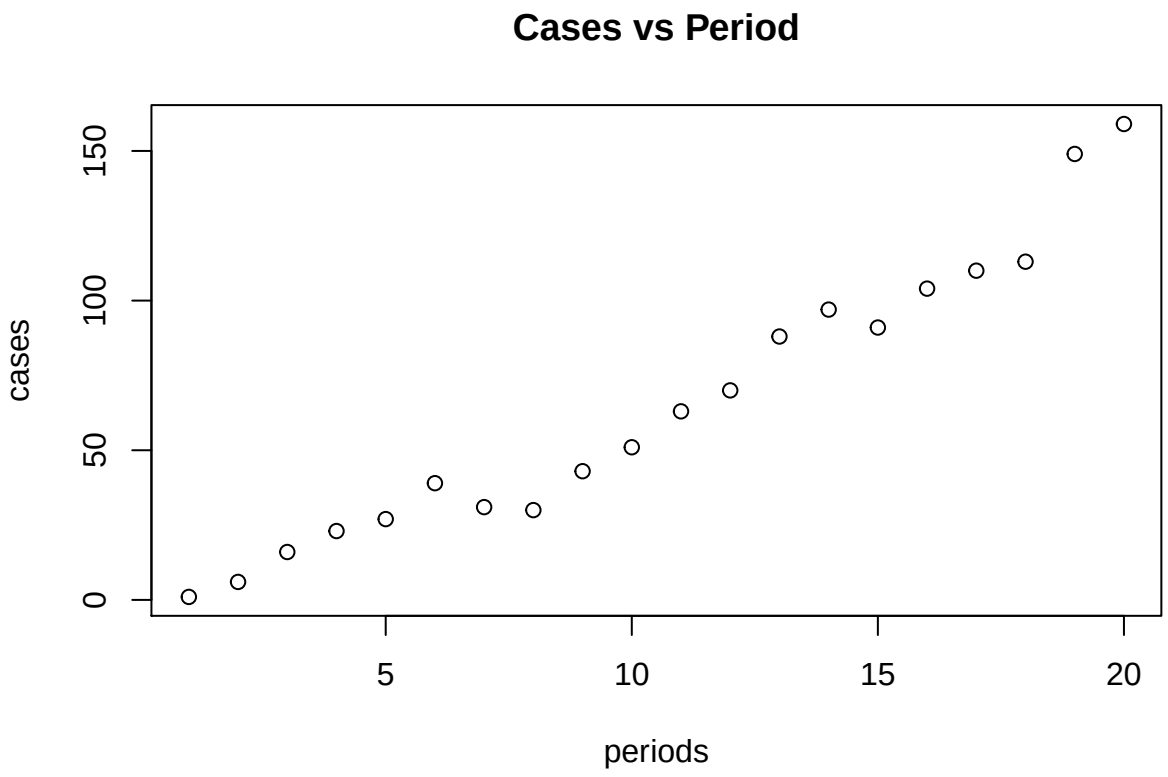
Question 1 [13 marks] Adapted from Problem 4.5 of Dobson & Barnett (2018)

The data below show the numbers of cases of AIDS in Australia by date of diagnosis for successive 3-month periods from 1984 to 1988. (Data from National Centre of HIV Epidemiology and Clinical Research, 1994)

	Quarter			
Year	1	2	3	4
1984	1	6	16	23
1985	27	39	31	30
1986	43	51	63	70
1987	88	97	91	104
1988	110	113	149	159

(a) **[1 mark]** Plot the number of cases y_i against time period i ($i = 1, \ldots, 20$). Comment on the relationship you see.

```
> period <- 1:20
> case <- c(1, 6, 16, 23, 27, 39, 31, 30, 43, 51, 63, 70, 88, 97, 91, 104, 110, 113, 149, 159)
> plot(period, case, main = "Cases vs Period", xlab = "periods", ylab = "cases")
```



Comment: They seem to form a linear relationship. With periods increase, the number of cases also increase.

(b) **[3 marks]** A possible model is the Poisson distribution with parameter $\lambda_i = i^\theta$, or equivalently

$$\log \lambda_i = \theta \log i$$

Find the MLE of θ. Plot $\log y_i$ against $\log i$. Comment on the relationship you see.

$$\text{Likelyhood function: } L(\lambda_i) = \prod_{i=1}^{n} \frac{\lambda_i^{y_i} exp(-\lambda_i)}{y_i!}$$

$$\text{So, } L(\theta) = \prod_{i=1}^{n} \frac{i^{\theta y_i} exp(-i^\theta)}{y_i!}$$

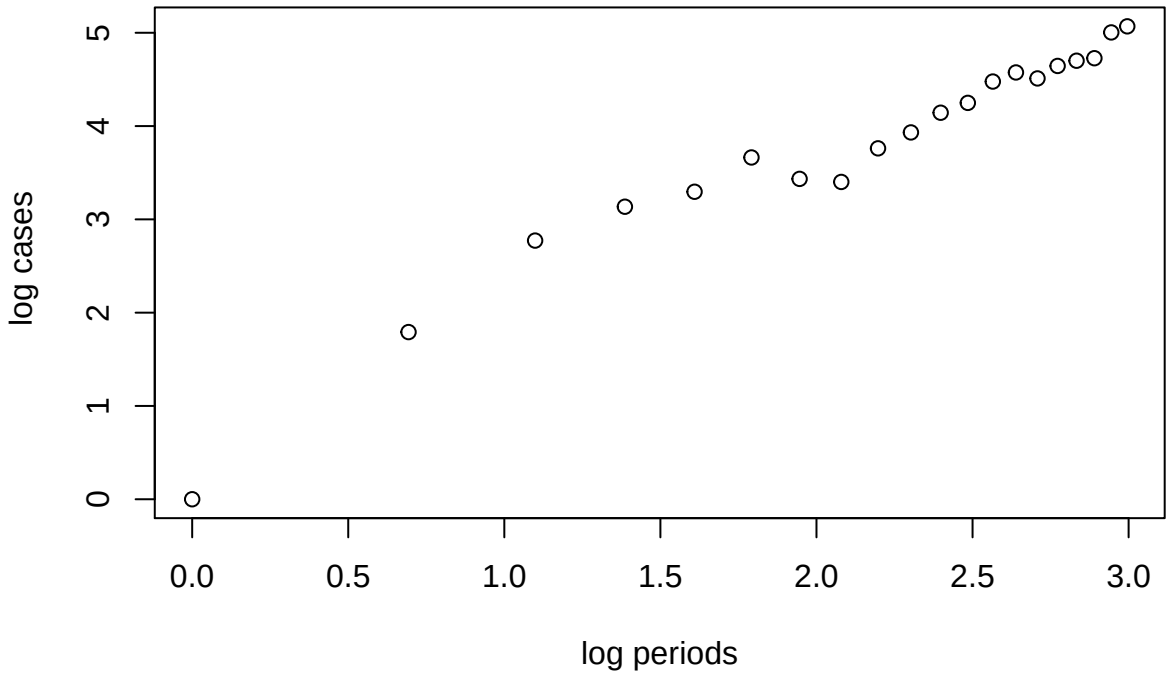
$$\text{Then, Log likely function: } l(\theta) = \sum_{i=1}^{n} [\theta y_i log(i) - i^\theta - log(y_i)]$$

$$\text{Score function: } S(\theta) = \sum_{i=1}^{n} [y_i log(i) - i^\theta log(i)]$$

Set the score function to 0, then we will have,

$$\sum_{i=1}^{n} [y_i log(i) - i^\theta log(i)] = 0$$

```
> score <- function(theta) {
+     sum(case * log(period) - period^theta * log(period))
+ }
> uniroot(score, lower = 0, upper = 2)$root
[1] 1.697999
>
> plot(log(period), log(case), xlab = "log periods", ylab = " log cases")
```



Comment: The MLE of θ in this question is 1.697999. From the plot, we can still see a linear relationship between period and cases

(c) **[3 marks]** Fit the following log-linear model to this data:

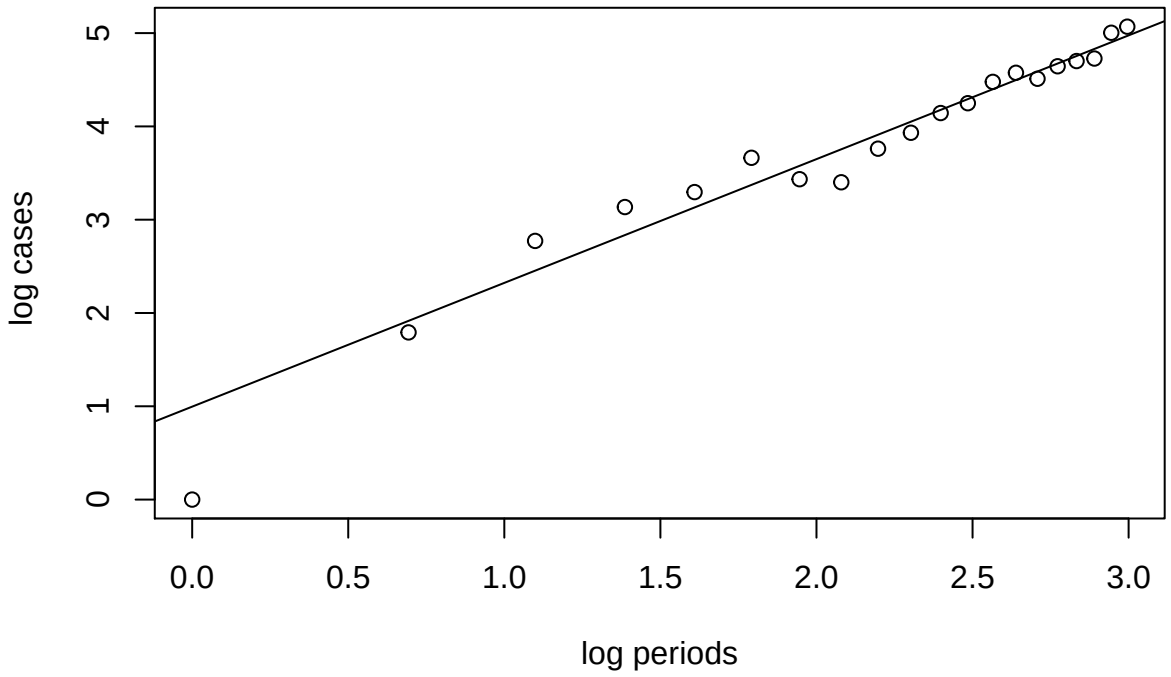
$$\log \lambda_i = \beta_1 + \beta_2 x_i$$

where $x_i = \log i$. Provide a precise written interpretation of the regression parameters. Add the fitted regression line to the plot from (b).

```
> m1 <- glm(case ~ log(period), family = poisson(link = "log"))
> summary(m1)

Call:
glm(formula = case ~ log(period), family = poisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.0568  -0.8302  -0.3072   0.9279   1.7310

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  0.99600    0.16971   5.869 4.39e-09 ***
log(period)  1.32661    0.06463  20.525  < 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: 677.264  on 19  degrees of freedom
Residual deviance:  21.755  on 18  degrees of freedom
AIC: 138.05

Number of Fisher Scoring iterations: 4
```

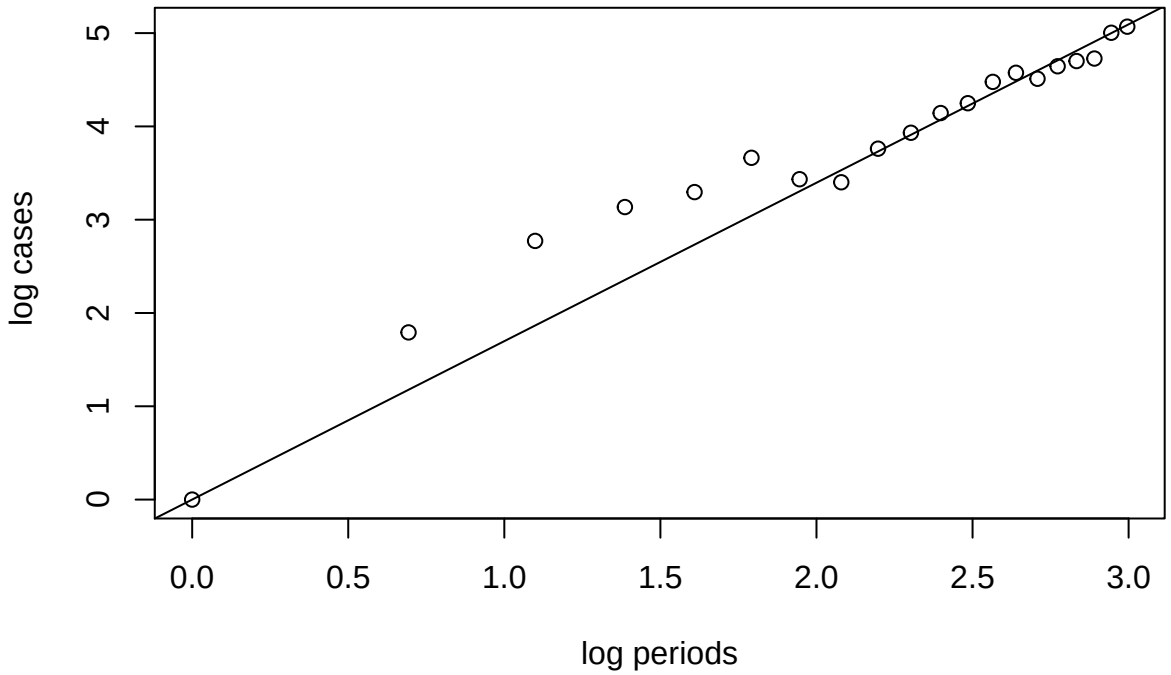
Comment: β_1 is the log of the number of AIDS cases in Australia in 1984 the first quarter. β_2 is the log relative rate of number of AIDS cases in Australia when log of period increases by one unit.

```
> plot(log(period), log(case), xlab = "log periods", ylab = " log cases")
> abline(a = 0.996, b = 1.32661)
```



(d) **[3 marks]** Fit the model implied by the relationship in (b) using a log-linear model. Add the fitted regression line to the plot from (b). Based on a visual assessment which model do you prefer and why?

```
> m2 <- glm(case ~ log(period) - 1, family = poisson(link = "log"))
> summary(m2)

Call:
glm(formula = case ~ log(period) - 1, family = poisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.9778  -0.3328   0.2202   1.1832   3.5159

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
log(period)  1.69800    0.01044   162.6   <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: 9062.600  on 20  degrees of freedom
Residual deviance:   52.381  on 19  degrees of freedom
AIC: 166.68

Number of Fisher Scoring iterations: 4
> plot(log(period), log(case), xlab = "log periods", ylab = " log cases")
> abline(a = 0, b = 1.698)
```



Comment: I prefer the model in part (c) because the distance between points in middle of the plot and the line are smaller than the second model.

(e) **[2 marks]** Test the null hypothesis that the model in (b) is adequate as compared to the model in (c). Be sure to state the null and alternative hypotheses, give the formula and distribution of the test statistic, calculate the test statistic, calculate the p-value and state the conclusion of the test. Based on this statistical assessment which model do you prefer?

$$H_0 : \beta_1 = 0 \text{ vs } H_A : \beta_1 \neq 0$$

we will use the wald based test to test the null hypothesis

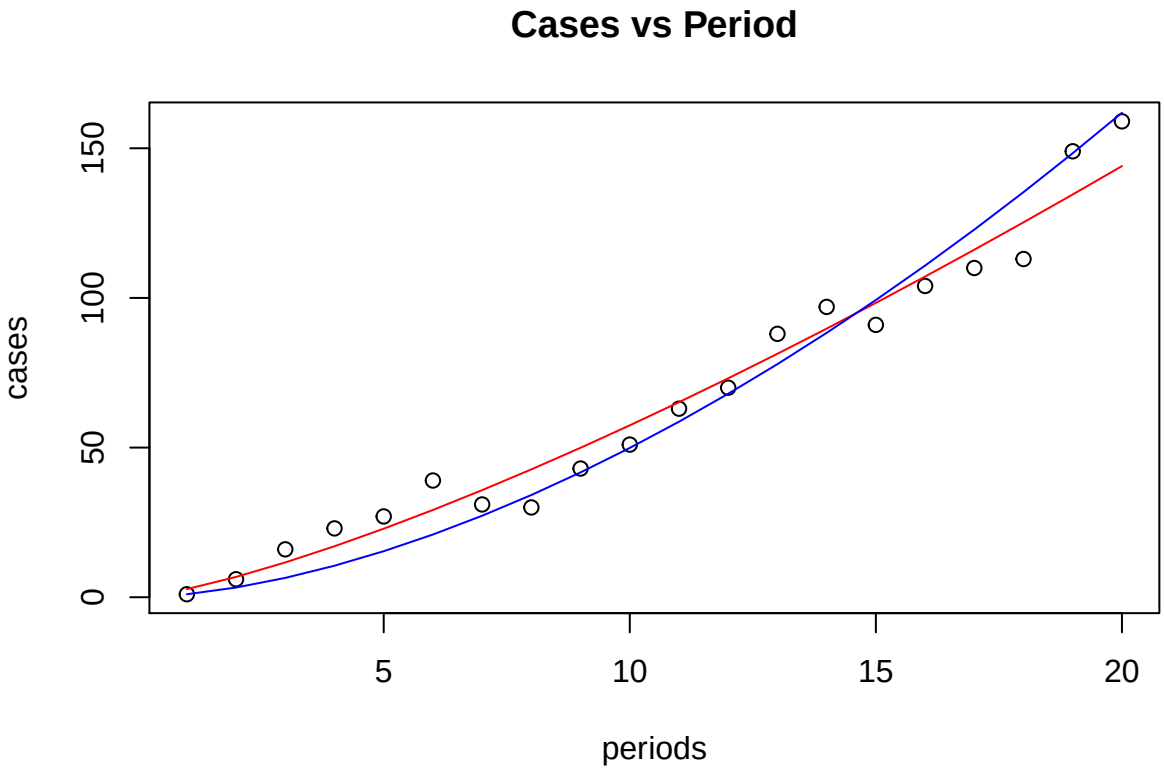
$$Z = \frac{\hat{\beta}_1 - 0}{se(\hat{\beta}_1)} = \frac{0.99600}{0.16971} = 5.868835 \sim N(0, 1)$$

```
> 2 * pnorm(-abs(5.868835))
[1] 4.388679e-09
```

Comment: The p-value is 4.388679e-09, which is smaller than 0.05. We reject the null hypothesis and conclude that the intercept can not be 0. Based on the statistic assessment, we still prefer to use the model from part (c), which has two parameters.

(f) **[1 mark]** Return to the plot from part (a) and add the fitted curves from the models in (c) and (d). Based on this visual assessment which model do you prefer and why?

```
> plot(period, case, main = "Cases vs Period", xlab = "periods", ylab = "cases")
> lines(period, m1$fitted.values, col = "red")
> lines(period, m2$fitted.values, col = "blue")
```



Comment: Based on the plot, I prefer the model from (c), the line of model from (c) goes through the points, while the line of model from (d) lies a little bit below the points, which provides a worse representation.

Question 2 [15 marks] Adapted from Problem 8.10 of Lachin (2000)

Fleming and Harrington (1991) present the results of a randomized clinical trial of the effects of gamma interferon versus placebo on the incidence of serious infections among children with chronic granulotomous disease (CGD). For each subject the number of infections experienced and the total duration of follow-up are presented ... the data set includes the patient `id`, number of severe infections experienced (`nevents`), the number of days of follow-up (`futime`) and the following covariates:

- `z1`: treatment group: interferon (1) versus placebo (2);
- `z2`: Inheritance pattern: X-linked (1) versus autosomal recessive (2);
- `z3`: Age (years);
- `z4`: Height (cm);
- `z5`: Weight (kg);
- `z6`: Corticosteroid use on entry: yes (1) versus no (2);
- `z7`: Antibiotic use on entry: yes (1) versus no (2);
- `z8`: Gender: male (1) versus female (2); and
- `z9`: Type of hospital: NIH (1), other US (2), Amsterdam (3), other European (4).

Use these data to conduct the following analyses.

```
> # Input the Fleming and Harrington Count Data (fhcnt) and name the covariates
> fhcnt = read.table("https://biostatcenter.gwu.edu/sites/biostatcenter.gwu.edu/files/Lachin%20Files/fhcnt.txt",
+     header = F)
> names(fhcnt) = c("id", "z1", "z2", "z3", "z4", "z5", "z6", "z7", "z8", "z9", "nevents", "futime")
```

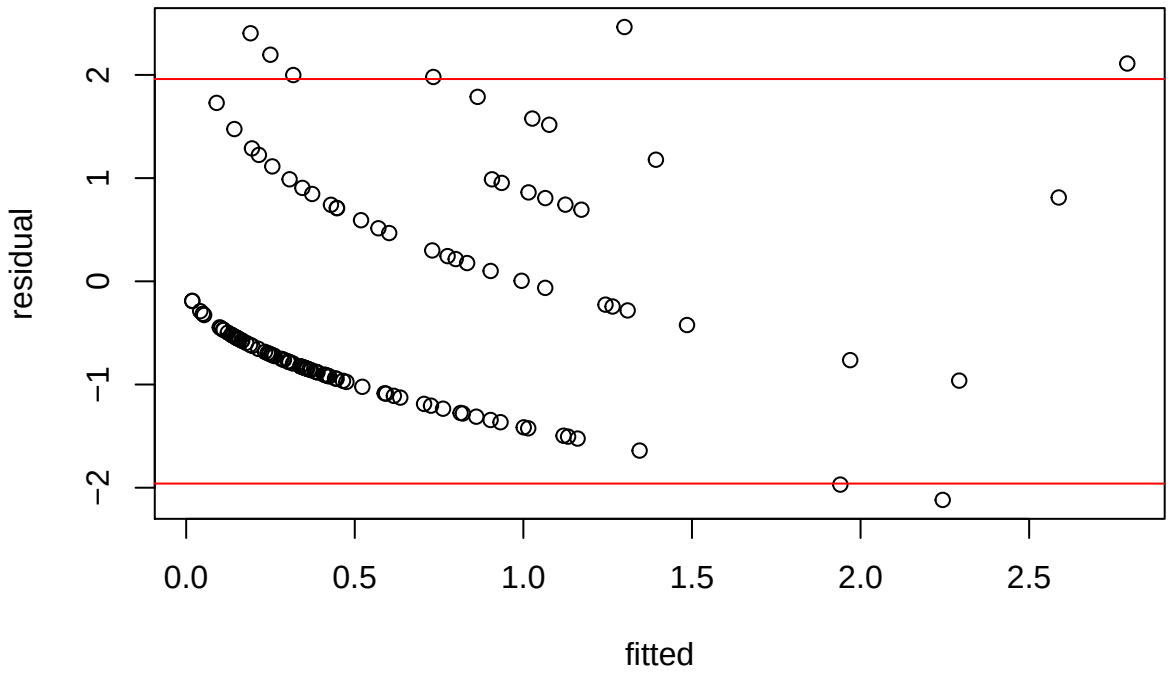
(a) **[2 marks]** Assume the infection counts follow a time homogeneous Poisson process. Fit the main effects Poisson GLM to the data. Be sure to declare covariates as factors, where necessary, and use an appropriate offset term. Print the R summary object for your fitted model.

```
> fhcnt$z1 <- as.factor(fhcnt$z1)
> fhcnt$z2 <- as.factor(fhcnt$z2)
> fhcnt$z6 <- as.factor(fhcnt$z6)
> fhcnt$z7 <- as.factor(fhcnt$z7)
> fhcnt$z8 <- as.factor(fhcnt$z8)
> fhcnt$z9 <- as.factor(fhcnt$z9)
> model <- glm(nevents ~ z1 + z2 + z3 + z4 + z5 + z6 + z7 + z8 + z9 + offset(log(futime)), family = pois
+     data = fhcnt)
> summary(model)

Call:
glm(formula = nevents ~ z1 + z2 + z3 + z4 + z5 + z6 + z7 + z8 +
    z9 + offset(log(futime)), family = poisson(link = "log"),
    data = fhcnt)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.1183  -0.8879  -0.5864   0.2579   2.4641

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -5.348270   1.058252  -5.054 4.33e-07 ***
z12          1.156140   0.277853   4.161 3.17e-05 ***
z22          0.754567   0.287943   2.621  0.00878 **
```

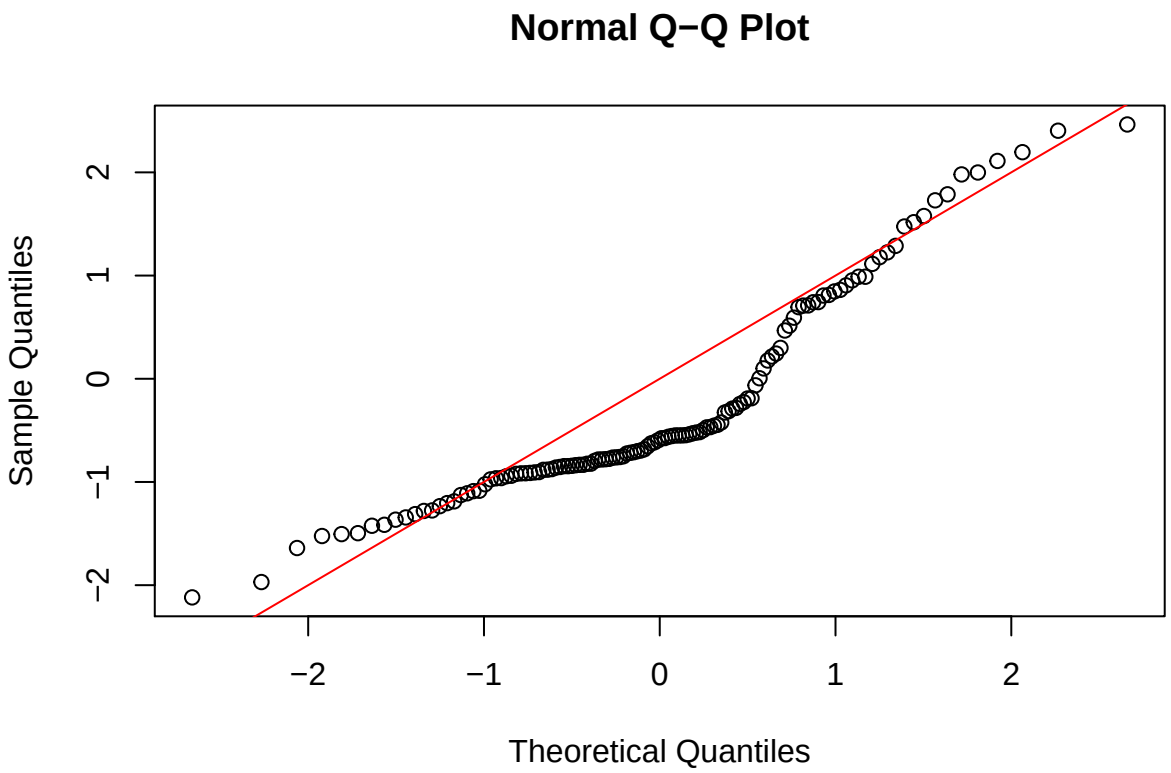
```
z3          -0.083938   0.035200  -2.385  0.01710 *
z4           0.007649   0.010556   0.725  0.46869
z5           0.010345   0.016127   0.641  0.52122
z62         -1.958265   0.597411  -3.278  0.00105 **
z72          0.668546   0.343399   1.947  0.05155 .
z82         -0.856499   0.393411  -2.177  0.02947 *
z92         -0.110733   0.326942  -0.339  0.73484
z93         -0.955742   0.487365  -1.961  0.04987 *
z94         -0.788223   0.494094  -1.595  0.11065
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 179.79  on 127  degrees of freedom
Residual deviance: 133.27  on 116  degrees of freedom
AIC: 260.3

Number of Fisher Scoring iterations: 6
```

(b) **[3 marks]** Conduct a residual analysis of the model from part (a). Include at least one scatterplot and one quantile-quantile plot. Investigate and explain any patterns you see in these plots. Are you satisfied with the fit of the model?

```
> residual <- residuals.glm(model, "deviance")
> fitted <- model$fitted.values
> plot(fitted, residual)
> abline(h = -1.96, col = "red")
> abline(h = 1.96, col = "red")
```



```
>
> qqnorm(residual)
> abline(a = 0, b = 1, col = "red")
```



Comment: We see from the residual plot that the points form several curves and going down with fitted value. They are not distributed randomly. Most points are in the interval. From the qq plots, we see that points in the middle are shifting away from the line. Both plots indicate that the fitted model does not follow a normal distribution. So, I'm not satisfied with the model.

Regardless of your conclusion in (b), use the model from (a) to answer the following parts.

(c) **[3 marks]** Is treatment with interferon effective in reducing serious infections among children with CGD? Justify your response with an appropriate estimate, its precise interpretation, and 95% confidence interval.

Comments: Yes, it is effective. Since the question is asking for the effectiveness of treatment, we will look at the treatment group, which is z1. From the fitted model, we see that the estimate of log of relative rate of placebo vs interferon is 1.156140. We can take the exponential and get the relative rate of taking placebo vs interferon is 3.177644. This means that group taking placebo is more likely to cause serious infections. Therefore, treatment with interferon is considered effective.

```
> c(exp(1.15614 - 0.277853 * 1.96), exp(1.15614 + 0.277853 * 1.96))
[1] 1.843283 5.477955
```

Therefore, the 95% confidence interval is (1.843283, 5.477955)

(d) **[3 marks]** Estimate the relative rate of serious infections for children treated at a hospital in Amsterdam versus those treated at other European hospitals. Include a 95% confidence interval with your estimate.

```
> # relative risk
> exp(-0.955742 - (-0.788223))
[1] 0.8457605
>
> # 95% confidence interval
> x <- as.matrix(c(0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, -1), ncol = 1)
> v <- summary(model)$cov.unscaled
> se <- sqrt(t(x) %*% v %*% x)
> dif <- -0.955742 - (-0.788223)
> c(exp(dif - 1.96 * se), exp(dif + 1.96 * se))
[1] 0.2761451 2.5903447
```

(e) **[2 marks]** Estimate the number of infections that a 12 year old male child, 142 cm tall, 34 kg in weight, with X-linked inheritance who was on corticosteroids but not antibiotics at entry, and was randomized the the treatment group at a non-NIH US hospital would expect to experience over one year.

```
> coef <- model$coefficients
> x <- as.matrix(c(1, 0, 0, 12, 142, 34, 0, 1, 0, 1, 0, 0), ncol = 1)
> exp(sum(coef * x) + log(365.25))
[1] 4.667999
```

Therefore, the number of infections that this particular child will expect to experience over one year is 4.667999.

(f) **[2 marks]** Write one paragraph summarizing the results and highlighting the important associations identified by this log linear model.

Comment: From p value, we see that z1 (treatment group), z2 (inheritance pattern), z6 (Corticosteriod use on entry) do affect the number of infections as they have really small p values comparing to 0.05, while z4 (height), z5 (weight), z7 (antibiotic?), z9 (type of hospital) do not really affect the number of infections. On top of that, we see that using interferon will make a huge difference as it significantly reduces the risk of causing infection and the relative rate is 3.177644. On the other hand, young age do not have an obvious advantage comparing to older ages. There might be an advatage of taking treatment in Amsterdam, but we need to details to support that.

Question 3 [12 marks] Adapted from Problem 7.6 of Agresti (2007) and 7.4 of Agresti (2018)

At the website www.stat.ufl.edu/~aa/intro-cda/data for the second edition of this book, the MBTI data file cross-classifies the MBTI Step II National Sample on four binary scales of the Myers–Briggs personality test: Extroversion/Introversion (E/I), Sensing/iNtuitive (S/N), Thinking/Feeling (T/F), and Judging/Perceiving (J/P). The 16 cells in this table correspond to the 16 personality types: ESTJ, ESTP, ESFJ, ESFP, ENTJ, ENTP, ENFJ, ENFP, ISTJ, ISTP, ISFJ, ISFP, INTJ, INTP, INFJ, INFP. Also collected was data on whether individuals report smoking or drinking alcohol frequently which we will ignore for the time being. The code below inputs the data set and prints out the 4-way $(2 \times 2 \times 2 \times 2)$ contingency table.

```
> # Input the Agresti Myers-Briggs data set
> MBTI = read.table("http://users.stat.ufl.edu/~aa/intro-cda/data/MBTI.dat", header = T)
> kable(cbind(MBTI[1:8, c(1, 2, 3, 4, 7)], MBTI[9:16, c(1, 2, 3, 4, 7)]))
```

EI	SN	TF	JP	n	EI	SN	TF	JP	n
e	s	t	j	77	i	s	t	j	140
e	s	t	p	42	i	s	t	p	52
e	s	f	j	106	i	s	f	j	138
e	s	f	p	79	i	s	f	p	106
e	n	t	j	23	i	n	t	j	13
e	n	t	p	18	i	n	t	p	35
e	n	f	j	31	i	n	f	j	31
e	n	f	p	80	i	n	f	p	79

(a) **[3 marks]** Consider first just the scales of Extroversion/Introversion (E/I) and Judging/Perceiving (J/P). Produce the appropriate 2-way contingency table and fit the main effects log linear model to the data. Perform a formal test of the null hypothesis of independence between the two scales. Be sure to carefully state the null and alternative hypotheses in terms of the regression coefficients (be explicit about which model you are referring to) and give the formula of the test statistic and its asymptotic distribution under the null hypothesis. What is the conclusion of the test?

```
> # E and J
> 77 + 106 + 23 + 31
[1] 237
>
> # E and P
> 42 + 79 + 18 + 80
[1] 219
>
> # I and J
> 140 + 138 + 13 + 31
[1] 322
>
> # I and P
> 52 + 106 + 35 + 79
[1] 272
```

```
> t <- matrix(c(237, 322, 219, 272), ncol = 2, byrow = TRUE)
> colnames(t) <- c("E", "I")
> rownames(t) <- c("J", "P")
> as.table(t)
   E   I
J 237 322
P 219 272
```

```
> ei <- c(1, 1, 2, 2)
> jp <- c(1, 2, 1, 2)
> num <- c(237, 322, 219, 272)
> model <- glm(num ~ factor(ei) + factor(jp), family = poisson(link = "log"))
> summary(model)

Call:
glm(formula = num ~ factor(ei) + factor(jp), family = poisson(link = "log"))

Deviance Residuals:
      1       2       3       4
-0.3715  0.3232  0.3931 -0.3472

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  5.49210    0.05504  99.782  < 2e-16 ***
factor(ei)2 -0.12971    0.06185  -2.097    0.036 *
factor(jp)2  0.26439    0.06226   4.246 2.17e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 23.11417  on 3  degrees of freedom
Residual deviance:  0.51756  on 1  degrees of freedom
AIC: 36.109

Number of Fisher Scoring iterations: 3
```

$$H_0 : \pi_{ij} = \pi_i * \pi_j \text{ for all } i \in (\text{"E","I"}) \text{ and } j \in (\text{"J", "P"})$$
$$H_A : \pi_{ij} \neq \pi_i * \pi_j \text{ for some } i \in (\text{"E","I"}) \text{ and } j \in (\text{"J", "P"})$$
$$\text{The test statistic is Residual deviance } D = 0.51756 \sim X^2_{(1)}$$

```
> 1 - pchisq(0.51756, 1)
[1] 0.4718844
```

Comment: 0.4718844>0.05, there is no evidence against the null hypothesis. We conclude that EI and JP are independent coefficients.

(b) **[2 marks]** For the fitted model from (a) calculate (by hand, using the relevant formula) the deviance residual for the count of individuals who are Introverted and Judging.

$$\text{The expected value } \hat{\mu} = \frac{(237+322)*(322+272)}{(237+322+322+272)} = 287.984$$
$$d = 2[y_{IJ} * log(\frac{y_{IJ}}{\hat{\mu}}) - (y_{IJ} - \hat{\mu})] = 2[287.984 * log(\frac{322}{287.984}) - (322 - 287.984)] = -3.727$$
$$\text{The deviance residual is } r = sign(y_{IJ} - \hat{\mu})\sqrt{|d|} = \sqrt{|d|} = \sqrt{3.727} = 1.9305$$

(c) **[2 marks]** Working with the 4-way table, fit the homogeneous association model (includes all main effects and 2-way interactions). Use a deviance test to compare the fit of this model to the saturated

model. Be sure to carefully state the null and alternative hypotheses in terms of the regression coefficients (be explicit about which model you are referring to) and give the formula of the test statistic and its asymptotic distribution under the null hypothesis. What is the conclusion of the test?

$H_0 : u_{ijk}^{EST} = u_{ikl}^{ETJ} = u_{ijl}^{ESJ} = u_{jkl}^{STJ} = u_{ijkl}^{ESTJ} = 0$ for all i,j,k,l H_A : Some of them are not 0 for some i,j,k,l

```
> E <- c(rep(1, 8), rep(0, 8))
> E <- factor(E)
> S <- c(1, 1, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 0, 0, 0, 0)
> S <- factor(S)
> T <- c(1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0, 1, 1, 0, 0)
> T <- factor(T)
> J <- c(1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0, 1, 0)
> J <- factor(J)
> num <- c()
> model2 <- glm(MBTI$n ~ E * S + E * T + E * J + S * J + S * T + J * T, family = poisson(link = "log"))
> summary(model2)

Call:
glm(formula = MBTI$n ~ E * S + E * T + E * J + S * J + S * T +
    J * T, family = poisson(link = "log"))

Deviance Residuals:
       1         2         3         4         5         6         7         8
-0.72826   1.00215   0.05168  -0.01429   1.49947  -1.29325  -0.07596   0.00231
       9        10        11        12        13        14        15        16
 0.56850  -0.82975  -0.04948   0.01728  -1.57051   1.09960   0.08587  -0.00804

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.37035    0.09913  44.087  < 2e-16 ***
E1           0.01142    0.12516   0.091  0.92732
S1           0.29141    0.12138   2.401  0.01636 *
T1          -1.00681    0.14898  -6.758 1.40e-11 ***
J1          -0.95183    0.14661  -6.492 8.45e-11 ***
E1:S1       -0.30212    0.14233  -2.123  0.03378 *
E1:T1       -0.19449    0.13121  -1.482  0.13826
E1:J1        0.01766    0.13160   0.134  0.89326
S1:J1        1.22153    0.14547   8.397  < 2e-16 ***
S1:T1        0.40920    0.15243   2.684  0.00727 **
T1:J1        0.55936    0.13512   4.140 3.48e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for poisson family taken to be 1)

    Null deviance: 399.944  on 15  degrees of freedom
Residual deviance:  10.162  on  5  degrees of freedom
AIC: 125

Number of Fisher Scoring iterations: 4
```

$$\text{The Deviance } D = 2 * \sum_{i=1}^{2}\sum_{j=1}^{2}\sum_{k=1}^{2}\sum_{l=1}^{2} O_{ijkl} * log(\frac{O_{ijkl}}{E_{ijkl}}) = 10.162 \sim X^2_{(5)}$$

```
> 1 - pchisq(10.162, 5)
[1] 0.07077304
```

Comment: Since the p-value 0.07077304 $>$ 0.05, we can not reject the null hypothesis. Therefore, we conclude that the homogeneous association model is adequate.

(d) **[3 marks]** Fit a series of log-linear models and find the model that is most appropriate for characterizing the association between the E/I, S/N, T/F, and J/P factors. Do not use automatic model selection functions/procedures. Print the R summary object of your final fitted model. Provide a precise interpretation of one interaction term from your final model.

```
> # saturated
> model1 <- glm(MBTI$n ~ E * S * T * J, family = poisson(link = "log"))
> model2 <- glm(MBTI$n ~ E * S + E * T + E * J + S * J + S * T + J * T, family = poisson(link = "log"))
> 1 - pchisq(model2$deviance - model1$deviance, model2$df.residual - model1$df.residual)
[1] 0.0707809
> model3 <- glm(MBTI$n ~ E * S + E * T + E * J + S * J + S * T, family = poisson(link = "log"))
> 1 - pchisq(model3$deviance - model2$deviance, model3$df.residual - model2$df.residual)
[1] 3.133661e-05
> model4 <- glm(MBTI$n ~ E * S + E * T + E * J + S * J + J * T, family = poisson(link = "log"))
> 1 - pchisq(model4$deviance - model2$deviance, model4$df.residual - model2$df.residual)
[1] 0.006766581
> model5 <- glm(MBTI$n ~ E * S + E * T + E * J + S * T + J * T, family = poisson(link = "log"))
> 1 - pchisq(model5$deviance - model2$deviance, model5$df.residual - model2$df.residual)
[1] 0
> model6 <- glm(MBTI$n ~ E * S + E * T + S * J + S * T + J * T, family = poisson(link = "log"))
> 1 - pchisq(model6$deviance - model2$deviance, model6$df.residual - model2$df.residual)
[1] 0.8932506
> model7 <- glm(MBTI$n ~ E * S + E * J + S * J + S * T + J * T, family = poisson(link = "log"))
> 1 - pchisq(model7$deviance - model2$deviance, model7$df.residual - model2$df.residual)
[1] 0.1376742
> model8 <- glm(MBTI$n ~ E * T + E * J + S * J + S * T + J * T, family = poisson(link = "log"))
> 1 - pchisq(model8$deviance - model2$deviance, model8$df.residual - model2$df.residual)
[1] 0.03382609
>
> # We pick model5 and continue
> model9 <- glm(MBTI$n ~ E * S + E * T + E * J + S * T, family = poisson(link = "log"))
> 1 - pchisq(model9$deviance - model5$deviance, model9$df.residual - model5$df.residual)
[1] 3.543691e-07
> model10 <- glm(MBTI$n ~ E * S + E * T + E * J + J * T, family = poisson(link = "log"))
> 1 - pchisq(model10$deviance - model5$deviance, model10$df.residual - model5$df.residual)
[1] 6.583101e-05
> model11 <- glm(MBTI$n ~ E * S + E * T + S * T + J * T, family = poisson(link = "log"))
> 1 - pchisq(model11$deviance - model5$deviance, model11$df.residual - model5$df.residual)
[1] 0.6535965
> model12 <- glm(MBTI$n ~ E * S + E * J + S * T + J * T, family = poisson(link = "log"))
> 1 - pchisq(model12$deviance - model5$deviance, model12$df.residual - model5$df.residual)
[1] 0.1639346
> model13 <- glm(MBTI$n ~ E * T + E * J + S * T + J * T, family = poisson(link = "log"))
```

```
> 1 - pchisq(model13$deviance - model5$deviance, model13$df.residual - model5$df.residual)
[1] 0.03039413
>
> # We pick model11 and continue
> model14 <- glm(MBTI$n ~ E * S + E * T + S * T, family = poisson(link = "log"))
> 1 - pchisq(model14$deviance - model11$deviance, model14$df.residual - model11$df.residual)
[1] 2.208931e-07
> model15 <- glm(MBTI$n ~ E * S + E * T + J * T, family = poisson(link = "log"))
> 1 - pchisq(model15$deviance - model11$deviance, model15$df.residual - model11$df.residual)
[1] 6.583101e-05
> model16 <- glm(MBTI$n ~ E * S + S * T + J * T, family = poisson(link = "log"))
> 1 - pchisq(model16$deviance - model11$deviance, model16$df.residual - model11$df.residual)
[1] 0.1390014
> model17 <- glm(MBTI$n ~ E * T + S * T + J * T, family = poisson(link = "log"))
> 1 - pchisq(model17$deviance - model11$deviance, model17$df.residual - model11$df.residual)
[1] 0.03039413
>
> # We pick model 16 and continue
> model18 <- glm(MBTI$n ~ E * S + E * T, family = poisson(link = "log"))
> 1 - pchisq(model18$deviance - model16$deviance, model18$df.residual - model16$df.residual)
[1] 2.296134e-10
> model19 <- glm(MBTI$n ~ E * S + S * T, family = poisson(link = "log"))
> 1 - pchisq(model19$deviance - model16$deviance, model19$df.residual - model16$df.residual)
[1] 2.208931e-07
> model20 <- glm(MBTI$n ~ E * T + S * T, family = poisson(link = "log"))
> 1 - pchisq(model20$deviance - model16$deviance, model20$df.residual - model16$df.residual)
[1] 6.3354e-08
>
> # We can not further reduce the model
```

```
> summary(model16)

Call:
glm(formula = MBTI$n ~ E * S + S * T + J * T, family = poisson(link = "log"))

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.2863  -2.0841  -0.5816   2.2231   3.9065

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  4.08785    0.09485  43.098  < 2e-16 ***
E1          -0.03871    0.11361  -0.341   0.7333
S1           0.80826    0.10441   7.741 9.84e-15 ***
T1          -1.27423    0.14639  -8.704  < 2e-16 ***
J1          -0.11706    0.07858  -1.490   0.1363
E1:S1       -0.32190    0.13598  -2.367   0.0179 *
S1:T1        0.58786    0.14597   4.027 5.64e-05 ***
T1:J1        0.66001    0.13012   5.073 3.93e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

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

```
    Null deviance: 399.944  on 15  degrees of freedom
Residual deviance:  87.185  on  8  degrees of freedom
AIC: 196.02

Number of Fisher Scoring iterations: 4
```

Comments: Therefore, model 16 is the final fitted model. E*S represents the association between E and S.

(e) **[2 marks]** For introverts, what does your final model from (d) tell you about the relative probabilities for the other three scales of their Myers–Briggs personality type? Include estimates, where appropriate.

Comments: From the fitted model, we see that the association E1:S1 is -0.32190. This means that introvert person are less likely to be intuitive. Since the association S1:T1 is 0.58786. This means that intuitive person will more likely to be feeling. Since the association T1:J1 is 0.66001. This means that the feeling person are more likely to be perceiving. Therefore, combining them together, an introvert person are liekly to be sensing, thinking and judging.