Research School of Finance Actuarial Studies & Statistics
ANU College of Business and Economics
Australian National University
Canberra ACT 2601
https://www.rsfas.anu.edu.au/

RSFAS ASSIGNMENT COVER SHEET

Submission and assessment is anonymous where appropriate and possible. Please do not write your name on this coversheet.

This coversheet must be attached to the front of your assessment when submitted in hard copy. If you have elected to submit in hard copy rather than Turnitin, you must provide copies of all references included in the assessment item.

Student ID: u7031432

Course Code and Name: ST7030 Generalized Linear Model

Assignment Number: 2

Assignment Due Date: Oct 26

Lecturer: Dr Andrew Wood

Tutor:

Tutorial number, day and time:

Word Count:

I declare that this work:

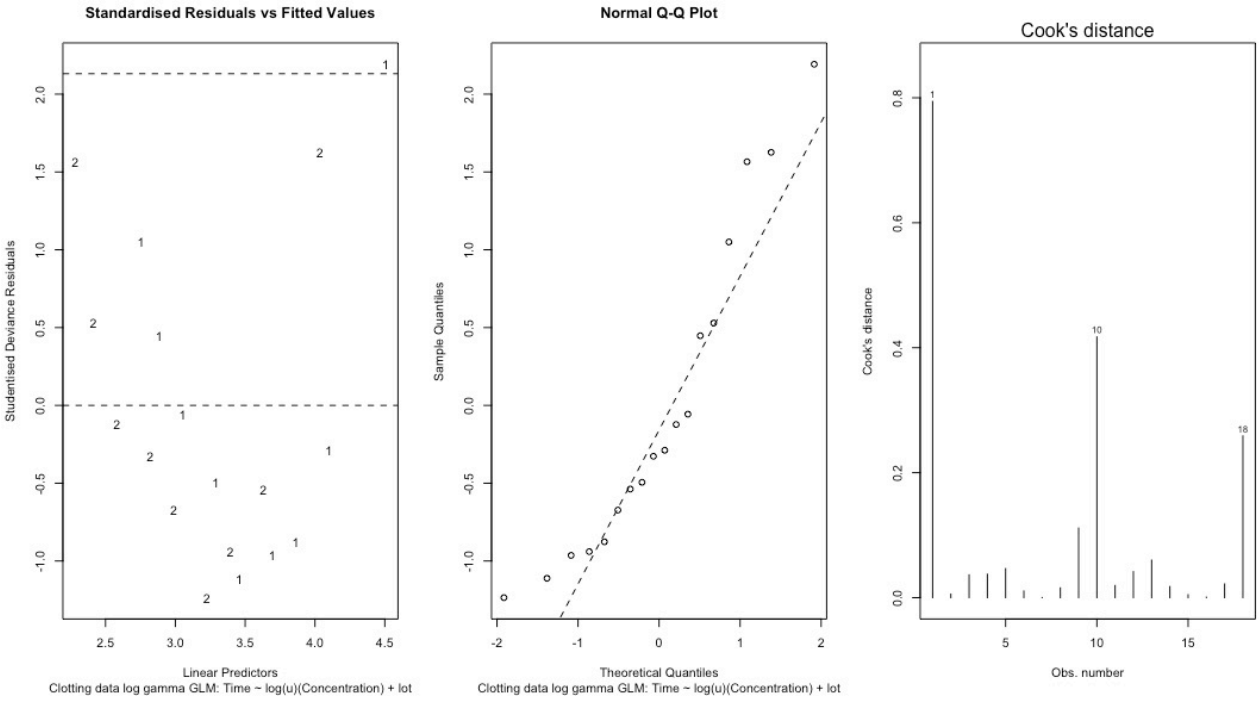
- upholds the principles of academic integrity, as defined in the ANU Policy: Code of Practice for Students University Academic Misconduct Rules;
- is original, except where collaboration (for example group work) has been authorised in writing by the course convener in the course outline and/or Wattle site;
- is produced for the purposes of this assessment task and has not been submitted for assessment in any other context, except where authorised in writing by the course convener;
- gives appropriate acknowledgement of the ideas, scholarship and intellectual property of others insofar as these have been used;
- in no part involves copying, cheating, collusion, fabrication, plagiarism or recycling.

Signed: LI ZIHAO

Dated Submitted: Oct 25 2020

Q1(a).

Given that we are required to use gamma errors with log link to explore different models. Therefore, I come up with 5 models, such as (1) time ~ u, (2) time ~ log (u), (3) time ~ u + lot, (4) time ~ log (u) + lot, (5) time ~ log (u) * lot. For every model, I firstly use summary statistics to get a brief view and then generate a series of diagnostic plots. Among all the models, the fourth one appears to be the optimal choice, as it has the lowest AIC value and the least residual deviance. Also, the diagnostic plots for the fourth model seems better comparing to all other models. (The diagnostic plots for the rest are not shown here, please find in appendices)



As we can see from the plots, the first point I noticed is the pattern in the Standardized Residuals vs. fitted value plot. Both of 1's and 2's are inverse bell-shaped, and 2's are intend to have lower standardized residuals than those for 1's. That might be related to the dataset we have, the 1's typically have bigger values of time than those for 2's. And both of the vertical scatters and horizontal scatters are equally randomly spread in the residuals vs. fitted value plot. The second point is no quantile deviation detected in the Normal Q-Q plot, there is only few points that are outside of the region, but that might be the result of limiting number of observations. And third point is the suspect outlier might

be the number 1 observation in the Cook's distance plot, as it is way too higher than the other observations. After that, I decide to check if there is over or under-dispersion in the model. The assumed dispersion used to calculate the model is given in the summary output, which is 0.02265. An alternative estimate can be found by dividing the residual deviance by the residual degrees of freedom, that is given in the below.

```
> summary(out3)$dispersion
[1] 0.02265072
> df<-out3$df.residual
> out3$deviance/df
[1] 0.02140642
```

We can see two estimates are fairly close, suggesting no over or under-dispersion, but we could confirm this by comparing the residual deviance with a X^2 distribution with the residual degrees of freedom.

```
> df
[1] 15
> c(qchisq(0.025,df),qchisq(0.975,df))
[1]  6.262138 27.488393
```

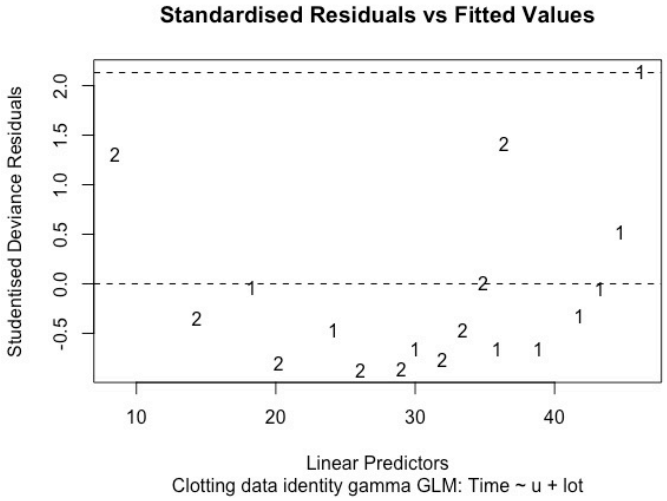
And we can see the residual deviance lies within the 95% interval, on the X^2 distribution with 15 degrees of freedom, so we do not reject the null hypothesis H_0: Φ = hat(Φ_{CV}) and conclude there is no evidence of significant over or under-dispersion. Then, I generate the table of coefficients for our model, and it suggests there is a negative relationship between u and time, also between lot and time.

```
> round(summary(out3)$coefficients,8)
             Estimate Std. Error    t value Pr(>|t|)
(Intercept)  5.4465951 0.13453229  40.485412 0.00e+00
log(u)      -0.5847614 0.03771584 -15.504398 0.00e+00
lot2        -0.4703448 0.07094711  -6.629513 8.02e-06
```

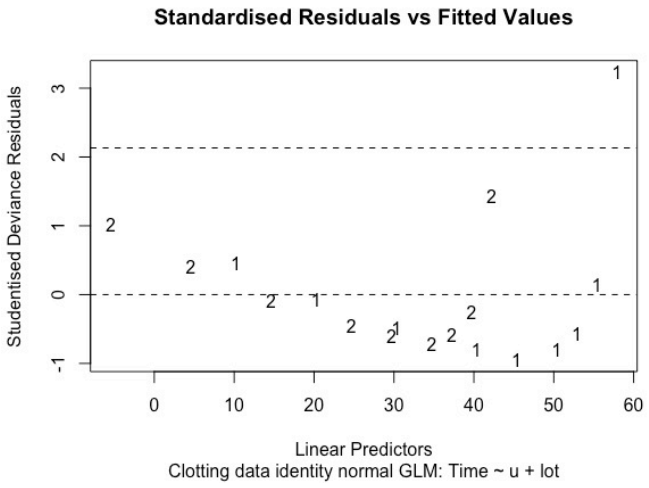
Suggesting that decreasing in plasma concentration and clotting agent type 1 can lead to the more clotting time.

Q1(b).

For this part, I use the similar procedures to create models based on two different ways, one is the gamma errors with identity link (the default) and the other one is the normal errors with identity link (using “family = Gaussian). The gamma errors with default link suggest that the model consists of time ~ u + lot performs better comparing to other models. That also indicates the model with each lot group has its own intercept fits better. The reasons for that are its summary statistics has the low AIC value and its residual plot has more randomly and equally spread than the others.



Similarly, the optimal model for normal errors with identity link has the same structure as the best model for gamma errors with identity link (which is likely no link).

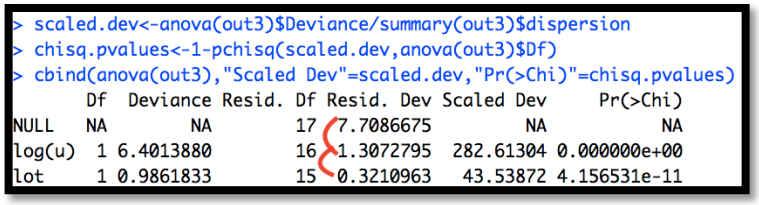


Then, I applied the same series of check for over or under-dispersion and the

relationship between explanatory variables and response variable. There seems no problem at all. Although, there is one thing against my choice of the optimal model is about the drop-in-deviance found in ANOVA table, that indicates that time ~ log (u) seems has the biggest gap for drop-in-deviance, and that might suggest time ~ log (u) tends to be better. Hence, for this part, there are arguably two options for the optimal model, one is u additive with lot and the other one is log transformation on u.

Q1(c).

Among all three approaches, none of them seems stands out more than the others. The first approach in the part(a) suggests that time ~ log (u) + lot has better performance not only on the residual plots, but also on the summary statistics. However, there is one problem about the drop-in-deviance from analysis of deviance table, it shows that additive explanatory variable lot is not that significant to the model (See the highlight in red).



```
> scaled.dev<-anova(out3)$Deviance/summary(out3)$dispersion
> chisq.pvalues<-1-pchisq(scaled.dev,anova(out3)$Df)
> cbind(anova(out3),"Scaled Dev"=scaled.dev,"Pr(>Chi)"=chisq.pvalues)
      Df  Deviance Resid. Df Resid. Dev Scaled Dev     Pr(>Chi)
NULL  NA        NA        17  7.7086675         NA           NA
log(u) 1 6.4013880        16  1.3072795  282.61304 0.000000e+00
lot    1 0.9861833        15  0.3210963   43.53872 4.156531e-11
```

Likewise, as I discussed in the bottom half of the part (b), there are arguably two models could be the optimal choice for the first and second error-link combinations, because they have the similar problem as the former talked above. Especially for the gamma errors with identity link, the time ~ log (u) seems has the biggest drop-in-deviance compared to other additive models. I think the problem might be the limiting size of observations and the characterization of the given dataset.

Q2(a).

For this problem, I created two models at first, one is the model without interaction term and the other one is the model including the interaction term between the factor tumour type and factor tumour site. Then I use anova (.) function to apply with two models above using the likelihood-ratio test, and we get the result shown below.

```
> anova(out4,out3,test="LRT")
Analysis of Deviance Table

Model 1: Count ~ 0 + Tumour + Site
Model 2: Count ~ 0 + Tumour * Site
  Resid. Df Resid. Dev Df Deviance Pr(>Chi)
1         6     51.795
2         0      0.000  6   51.795 2.05e-09 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

As we can see from above, the p-value we get is $2.05 * 10^{-9}$, which is statistically significant. Based on that, we can safely reject the null hypothesis that the response distributions are independent. So we can conclude that not only the model with interaction term fits better than the one without interaction term, but also the relationship of two response factors are dependent.

Q2(b).

After treated tumour site as the explanatory variable, I found out the new model does not fit quite well as the former model does. And we can see the difference between the predictions for both two models shown below.

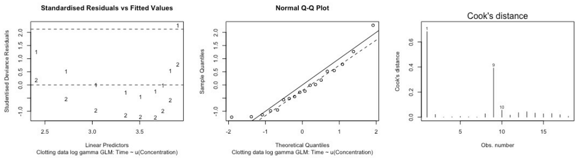
```
> xtabs(h0 ~ Tumour + Site, M)
      Site
Tumour       E      HN       T
     H  19.210   5.780   9.010
     I  31.640   9.520  14.840
     N  70.625  21.250  33.125
     S 104.525  31.450  49.025
> M$h1 <- predict(out3, M, type="response")
> xtabs(h1 ~ Tumour + Site, M)
      Site
Tumour   E  HN   T
     H  10  22   2
     I  28  11  17
     N  73  19  33
     S 115  16  54
```

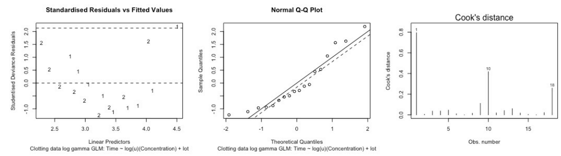
In the screenshot given here, we can see that the new model seems not handle our original data properly. So that might lead to conclude that the new model with treating tumour site as explanatory variable seems not a good choice to improve the model.

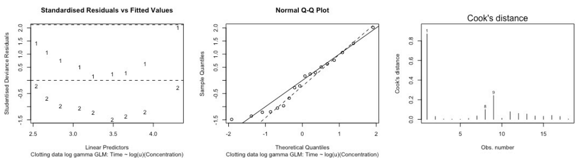
Appendix:

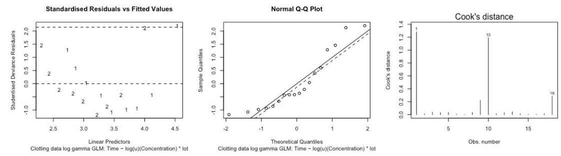
Q1.

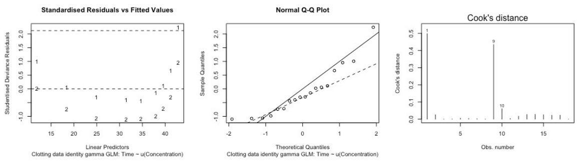
```
C=read.table(file.choose())
names(C)
View(C)
lot=factor(lot)
out1<-glm(time~u,family = Gamma(link = "log"))
lp1<-3.9796-0.0157*u
res1<-residuals(out1,type=c("deviance"))
plot(lp1,res1)
plot(lp1,res1,col=c("red","blue")[lot])
plot(out1,1,col=c("red","blue")[lot])
#Test with second case of time as response and log(u) as covariate
out2<-glm(time~log(u),family = Gamma(link = "log"))
summary(out2)
lp2=5.25197-0.58874*log(u)
res2=residuals(out2,type = c("deviance"))
plot(lp2,res2)
plot(lp2,res2,col=c("red","blue")[lot])
#Test with third case of time as response and log(u) and factor lot as covariates
out3<-glm(time~log(u)+lot,family = Gamma(link = "log"))
lot1<-as.numeric(lot)
summary(out3)
lp3<-5.44660-0.58476*log(u)-0.47034*(lot1-1)
res3<-residuals(out3,type = c("deviance"))
plot(lp3,res3)
plot(lp3,res3,col=c("red","blue")[lot])
#Test with fourth case of time as response and log(u) and lot with its own intercept as covariate
out4<-glm(time~log(u)*lot,family = Gamma(link = "log"))
summary(out4)
lp4=5.50323-0.60192*log(u)-(lot1-1)*(0.58447-0.03448*log(u))
res4=residuals(out4,type=c("deviance"))
plot(lp4,res4)
plot(lp4,res4,col=c("red","blue")[lot])
std.residuals <- function(model, type="deviance"){
  # Function to standardise residuals from a GLM model object
  # Produces standardised deviance residuals, unless type="pearson" requested
  std.error <- sqrt(summary(model)$dispersion * (1 - influence(model)$hat))
  std.res <- residuals(model)/std.error
  if (type=="pearson") std.res <- residuals(model, "pearson")/std.error
  std.res
}
```

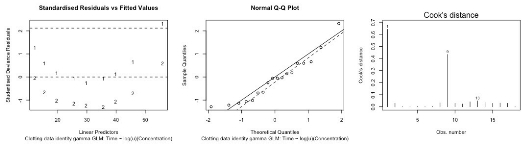


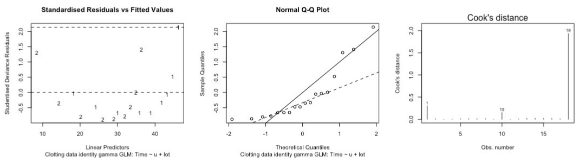


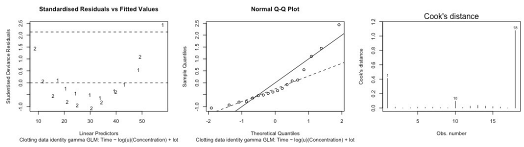


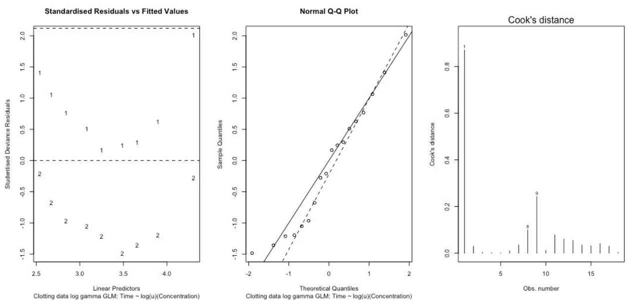


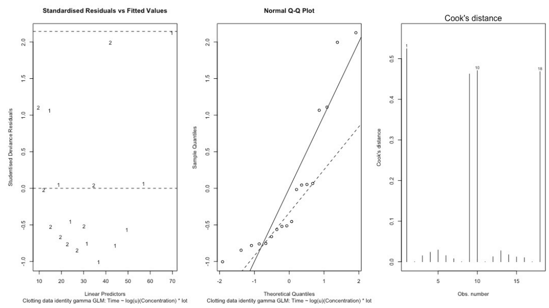












Q2.

```
View(melanoma_data)
M<-melanoma_data
names(M)
attach(M)
levels(Tumour)
levels(Site)
t=factor(Tumour)
s=factor(Site)
out3<-glm(Count~0+Tumour*Site,family = poisson)
out4<-glm(Count~0+Tumour+Site,family = poisson)
summary(out3)
summary(out4)
anova(out4,out3,test="LRT")
library(car)
Anova(out3,test.statistic = "LR",type=2)
X2<-sum(residuals(out3,type=c("pearson"))**2)
anova(out1)  anova(out2)
M$h0 <- predict(out4, M, type="response")   xtabs(h0 ~ Tumour + Site, M)   M$h1 <- predict(out3, M, type="response")  xtabs(h1 ~ Tumour + Site, M)
```