THE UNIVERSITY OF AUCKLAND
Te Whare Wānanga o Tāmaki Makaurau
NEW ZEALAND

SCIENCE

STATS 330: Statistical Modelling

Assignment Tracking Sheet

Student Information

University ID:	190411106	Username:	hche737
Family Name:	Cheena	Given Names:	Hasnain

Assignment Information

Assignment Name:	Assignment 2	Due:	3:00 p.m. - 01 May, 2020 (NZ Time)
Department:			
Lab / Tutorial Day:		Time:	
Lab / Tutorial Group:		Tutor:	
Notes:		Word Count:	

Declaration: (please read and sign)

By submitting this assignment, I confirm that I am aware of The University expectation that all students complete coursework with integrity and honesty as stated in the Student Academic Conduct Statute.
http://www.auckland.ac.nz/uoa/home/about/teaching-learning/honesty/tl-uni-regs-statutes-guidelines

- I understand that the University of Auckland will not tolerate cheating or assisting other to cheat, and views cheating in coursework as a serious academic offence.
- I declare that where work from other sources (including sources on the world-wide web) has been used, it has been properly acknowledged and referenced.
- I confirm that this work represents my individual/ our team's effort and does not contain plagiarised material.
- I have checked the above details and verify them to be correct for the assignment I am submitting.
- I understand that the University of Auckland takes no responsibility for lost assignments and that I agree to provide a duplicate copy if requested.
- I understand that uncollected assignments will be retained in secure storage until the end of the examination period and thereafter destroyed.
- I agree that I will provide or submit an electronic version of my work for computerised review if requested.

Signed: Hasnain Cheena Date: 1/05/2020

Note:

1. Assignments are not accessible after they have been handed in. No additions/removals will be permitted.
2. Marks may be withheld for students who have not submitted their work to Turnitin.com if required in the course outline.
3. The University of Auckland views cheating in coursework as a serious academic offence. Accordingly it may require submitted work to be reviewed against electronic source material using computerised detection mechanisms.

Stats 330 Assignment 3

Hasnain Cheena, 190411106, hche737

18/04/2020

Question 1

```
#convert degrees fahrenheit to celsius
HighC = (nyBridges.df$High.Temp - 32) * 5/9
LowC = (nyBridges.df$Low.Temp - 32) * 5/9

#convert rain in inches to mm
#as a result of this conversion trace amounts (denoted with a T) are converte
d to NaN
Rainmm = as.numeric(as.character(nyBridges.df$Precipitation)) * 25.4

## Warning: NAs introduced by coercion

#relevel day variable
day.relevel = relevel(nyBridges.df$Day, "Monday")
```

Relevel Day

Monday was selected to be the baseline level as in New Zealand we consider Monday to be the first day of the week. Further by using Monday we can easily assess if there is a difference in bicycle activity between the first day of the week and the rest of the week.

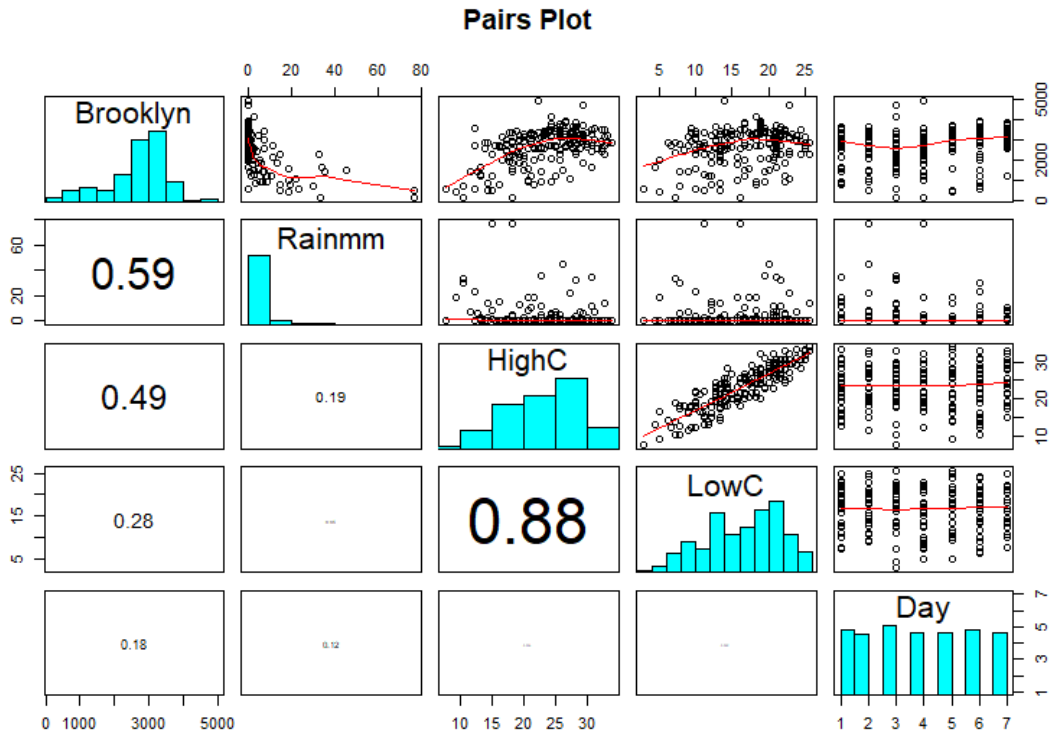
Trace Amounts

Trace characterizes an amount of precipitation that is greater than zero but too small to be measured reliably. The "T" used in the dataset denotes trace amounts of precipitation. As I am not a meterologist and therefore cannot confirm whether trace amounts of precipitation are equivalent to 0 I have chosen to recode trace amounts as NaN.

Reference for definition of trace:
https://www.wmo.int/pages/prog/www/IMOP/publications/CIMO-Guide/Prelim_2018_ed/8_I_6_en_MR_clean.pdf

Question 2

```
pairs20x(nyBridges.Q2.df, main="Pairs Plot")
```



a) The statistician did not include both high temperature ($HighC$) and low temperature ($LowC$) variables. This is because the high temperature variable and low temperature variable are highly correlated as shown on the pairs plot. Thus, both HighC and LowC explain the same pattern in the response variable and therefore the statistician only needed to include one of them. HighC in specific was chosen as HighC has a stronger correlation to the response than LowC (as shown on the pairs plot).

b) Firstly, the log transformation was required for Rainmm as the variable is very right skewed (as can be seen on the pairs plot). However, using the natural log transformation will not work as Rainmm contains the value zero. Therefore, log1p was used as log1p computes the natural logarithm of the given value plus one accounting for Rainmm values of zero.

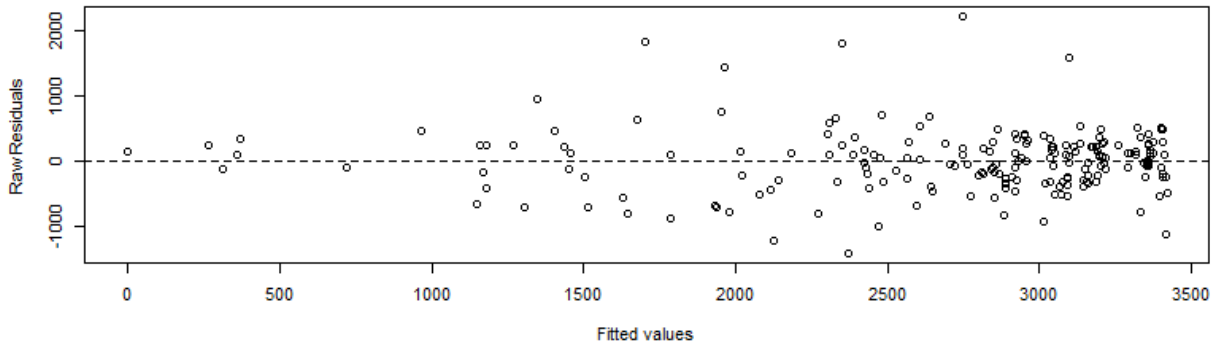
c) A quadratic effect for temperature was included in the models because the relationship between temperature ($HighC$) and the response variable showed curvature (as can be seen on the pairs plot). Therefore, a quadratic term for temperature was added to account for the non-linearity aspect of the relationship.

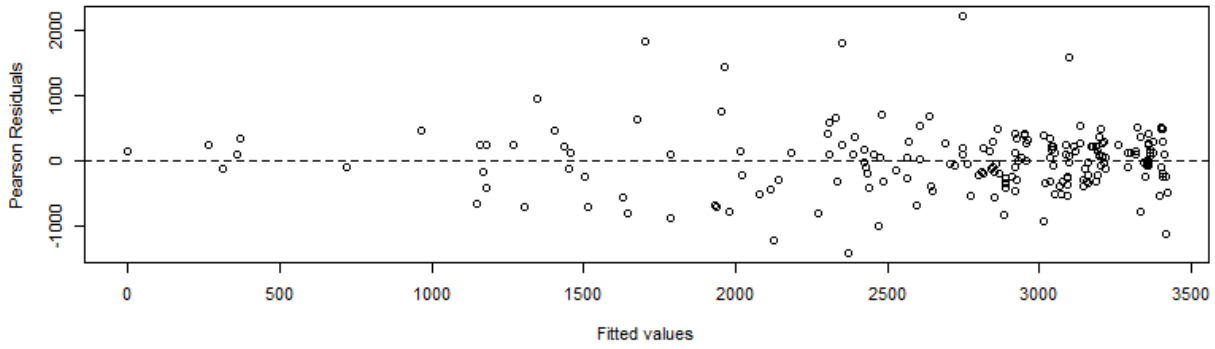
Question 3

Model A Exploration

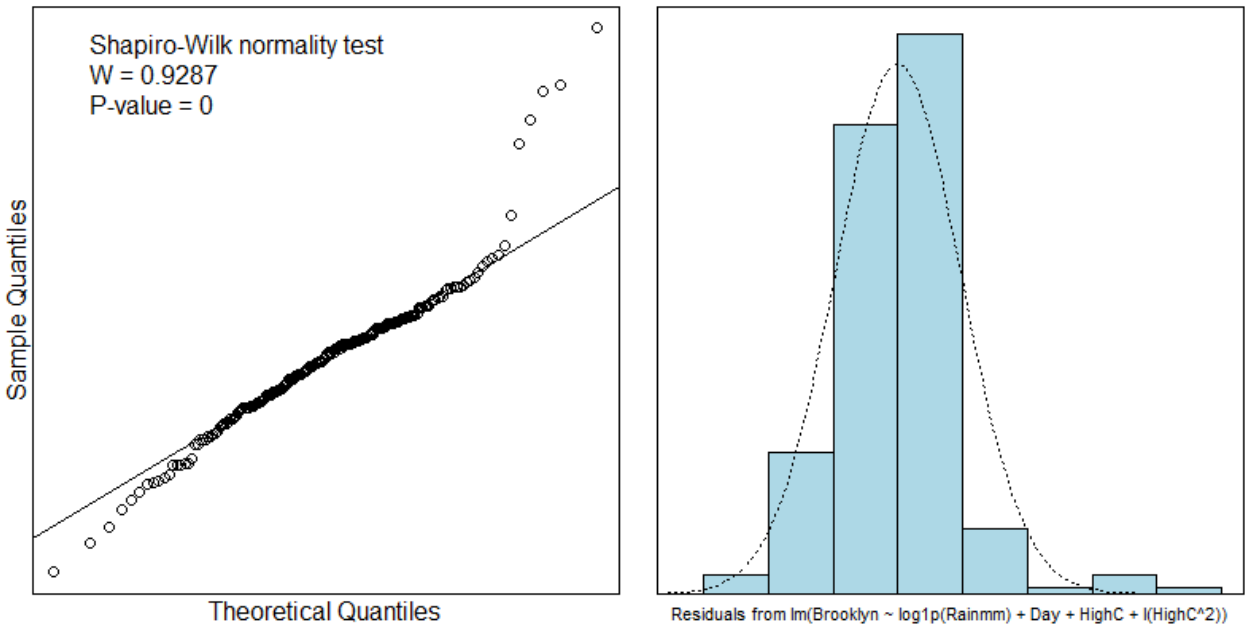
```
model.lin.a<-lm(Brooklyn~log1p(Rainmm)+Day+HighC+I(HighC^2),data=NYBridges.df
)
#group plots
layout(matrix(c(1,1,1,1,1,1,2,2,2,2,2,2), nrow = 4, ncol = 3, byrow = TRUE))
#raw residual
plot(predict(model.lin.a),residuals(model.lin.a), ylab="Raw Residuals",xlab="
Fitted values")
```

```r
abline(h=0, lty='dashed')
#pearson residual
plot(predict(model.lin.a),residuals(model.lin.a, type="pearson"), ylab="Pears
on Residuals", xlab="Fitted values")
abline(h=0, lty='dashed')
```






```r
normcheck(model.lin.a, shapiro.wilk = TRUE)
```

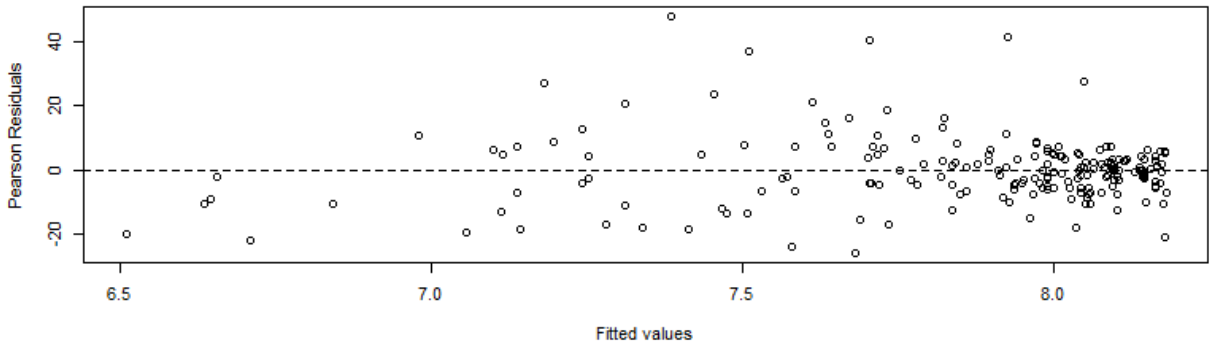


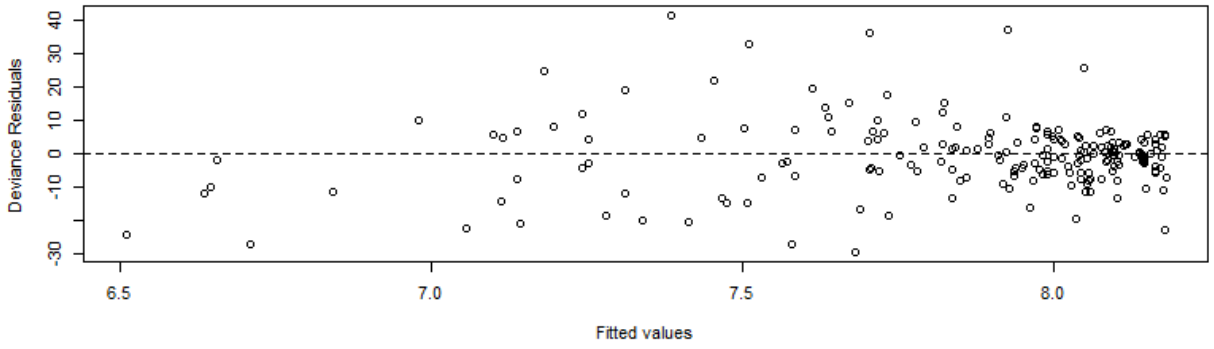
```r
#assess goodness of fit
deviance(model.lin.a)

## [1] 46551228
```

The raw residuals and Pearson residuals do not show a patternless band around 0. Both residual plots clearly display non-constant variance and majority of the Pearson residuals are not between -2 and 2. Furthermore, the histogram of residuals shows the residuals are very right skewed. All the evidence above coupled with the fact that the response variable is a count lead to the conclusion that the linear model is not appropriate.

Model B Exploration

```r
model.pois.b<-glm(Brooklyn~log1p(Rainmm)+Day+HighC+I(HighC^2),family=poisson,
data=NYBridges.df)


#group plots
layout(matrix(c(1,1,1,1,1,1,2,2,2,2,2,2), nrow = 4, ncol = 3, byrow = TRUE))
#pearson residual
plot(predict(model.pois.b),residuals(model.pois.b, type="pearson"),ylab="Pear
son Residuals", xlab="Fitted values")
abline(h=0, lty='dashed')
#deviance
plot(predict(model.pois.b),residuals(model.pois.b, type="deviance"),ylab="Dev
iance Residuals", xlab="Fitted values")
abline(h=0, lty='dashed')
```



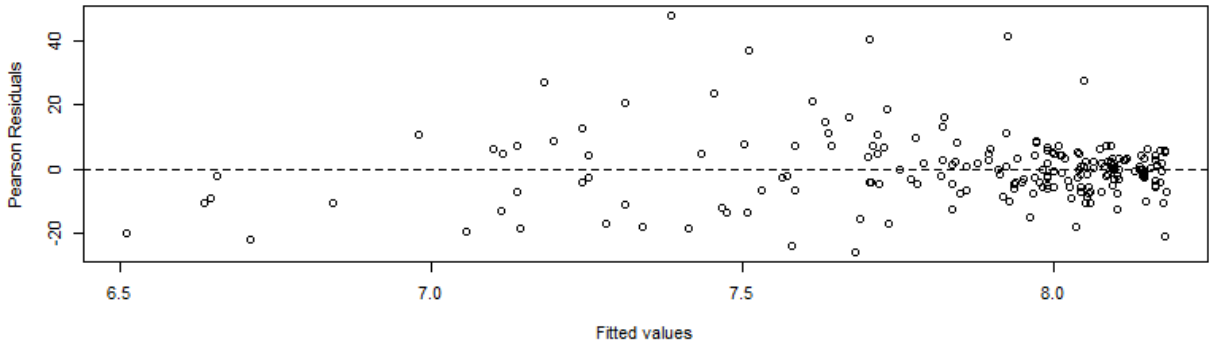


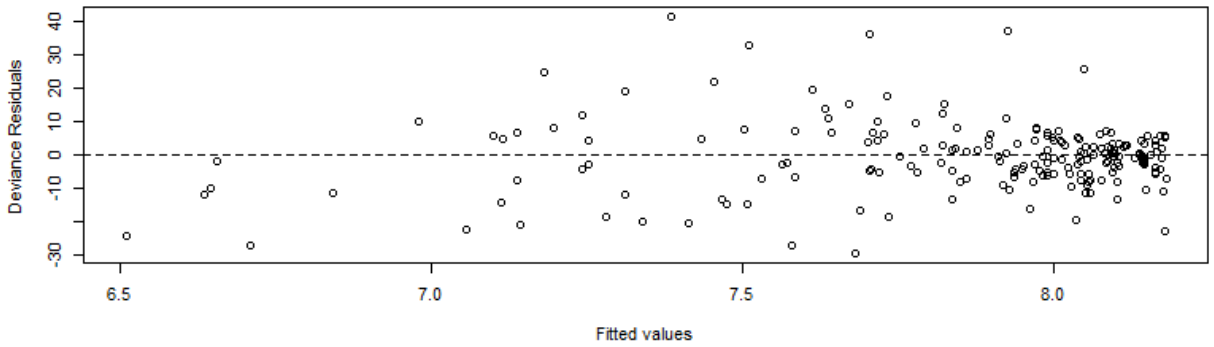
```r
#assess goodness of fit
1-pchisq(model.pois.b$deviance, model.pois.b$df.residual)

## [1] 0
```

The Pearson and deviance residual plots do not show a patternless band around 0. Therefore, the residual plots show non-constant variance. Furthermore, many of the residuals are not within -2 and 2. Moreover, we have very strong evidence (p-value = 0) that the model is not a good fit. Therefore, the Poisson model is not appropriate to capture the relationship between the bicycle activity on the Brooklyn bridge and explanatory variables.

Model C Exploration

```r
model.qpois.c<-glm(Brooklyn~log1p(Rainmm)+Day+HighC+I(HighC^2),family=quasipoisson,data=NYBridges.df)

layout(matrix(c(1,1,1,1,1,1,2,2,2,2,2,2), nrow = 4, ncol = 3, byrow = TRUE))
#pearson residual
plot(predict(model.qpois.c),residuals(model.qpois.c, type="pearson"),ylab="Pearson Residuals", xlab="Fitted values")
abline(h=0, lty='dashed')
#deviance
plot(predict(model.qpois.c),residuals(model.qpois.c, type="deviance"),ylab="Deviance Residuals", xlab="Fitted values")
abline(h=0, lty='dashed')
```

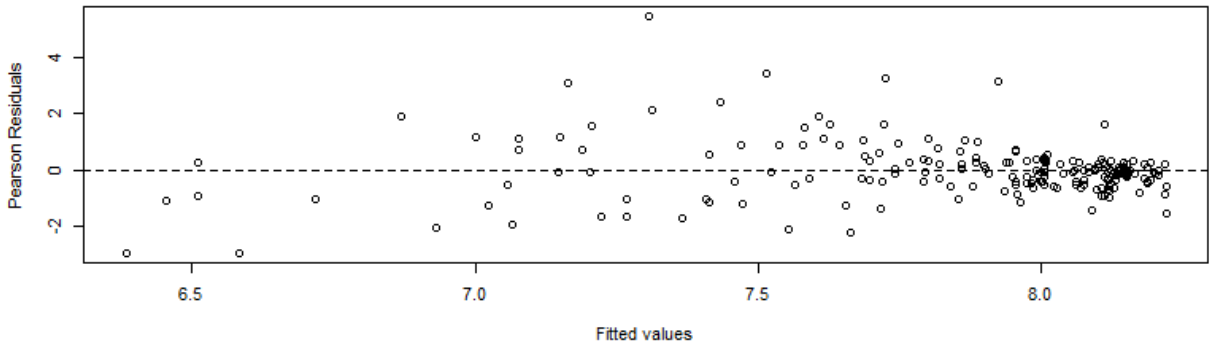


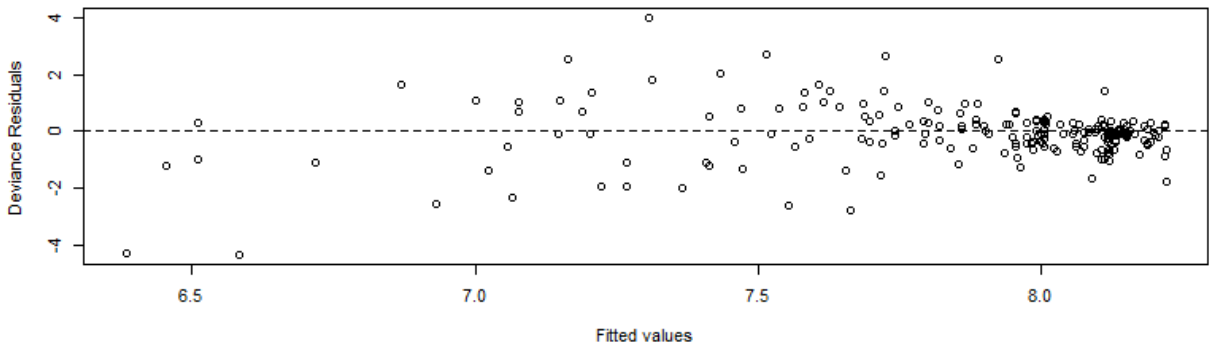


The Pearson and deviance residual plots are still showing non-constant variance as they are not a patternless band around 0. This suggests that the quasi-Poisson model is not suitable to model the relationship between the bicycle activity on the Brooklyn bridge and explanatory variables.

Model D Exploration

```
model.nb.d<-glm.nb(Brooklyn~log1p(Rainmm)+Day+HighC+I(HighC^2),data=NYBridges
.df)

layout(matrix(c(1,1,1,1,1,1,2,2,2,2,2,2), nrow = 4, ncol = 3, byrow = TRUE))
#pearson residual
plot(predict(model.nb.d),residuals(model.nb.d, type="pearson"),ylab="Pearson
Residuals", xlab="Fitted values")
abline(h=0, lty='dashed')
#deviance
plot(predict(model.nb.d),residuals(model.nb.d, type="deviance"),ylab="Devianc
e Residuals", xlab="Fitted values")
abline(h=0, lty='dashed')
```






```
#summary
summary(model.nb.d)

##
## Call:
## glm.nb(formula = Brooklyn ~ log1p(Rainmm) + Day + HighC + I(HighC^2),
##     data = NYBridges.df, init.theta = 15.71142767, link = log)
##
```

```
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3575  -0.4805  -0.0738   0.2983   3.9887
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    6.1101329  0.2556884  23.897  < 2e-16 ***
## log1p(Rainmm) -0.2715243  0.0176485 -15.385  < 2e-16 ***
## DayFriday     -0.0125467  0.0677290  -0.185   0.8530
## DaySaturday   -0.0177899  0.0654899  -0.272   0.7859
## DaySunday     -0.1311051  0.0673415  -1.947   0.0516 .
## DayThursday    0.0148485  0.0675741   0.220   0.8261
## DayTuesday     0.0831982  0.0666101   1.249   0.2117
## DayWednesday   0.0572638  0.0673593   0.850   0.3953
## HighC          0.1455586  0.0227150   6.408 1.47e-10 ***
## I(HighC^2)    -0.0026132  0.0005004  -5.222 1.77e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(15.7114) family taken to be 1)
##
##     Null deviance: 560.21  on 199  degrees of freedom
## Residual deviance: 203.62  on 190  degrees of freedom
##   (14 observations deleted due to missingness)
## AIC: 3167.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  15.71
##           Std. Err.:  1.58
##
##  2 x log-likelihood:  -3145.446

#assess goodness of fit
1-pchisq(model.nb.d$deviance, model.nb.d$df.residual)

## [1] 0.2367973
```

The Pearson and deviance residuals are scattered in a reasonably patternless band around 0. Further, a majority of the residuals are between -2 and 2. Therefore, the residuals do not show a pattern in the mean and variance. Further, the deviance does not provide evidence (p-value ≈ 0.24) for a lack of fit. The negative binomial model is therefore an appropriate model to capture the relationship between the bicycle activity on the Brooklyn bridge and explanatory variables.

Question 4

```
#weekend vs weekday
#relevel with baseline as sunday
NYBridges.df$Day <- relevel(NYBridges.df$Day, "Saturday")
model.nb.d<-glm.nb(Brooklyn~log1p(Rainmm)+Day+HighC+I(HighC^2),data=NYBridges
```

```
.df)
summary(model.nb.d)

##
## Call:
## glm.nb(formula = Brooklyn ~ log1p(Rainmm) + Day + HighC + I(HighC^2),
##     data = NYBridges.df, init.theta = 15.71142767, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3575  -0.4805  -0.0738   0.2983   3.9887
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    6.0923430  0.2534232  24.040  < 2e-16 ***
## log1p(Rainmm) -0.2715243  0.0176485 -15.385  < 2e-16 ***
## DayMonday      0.0177899  0.0654899   0.272   0.7859
## DayFriday      0.0052432  0.0668158   0.078   0.9375
## DaySunday     -0.1133152  0.0665772  -1.702   0.0888 .
## DayThursday    0.0326384  0.0666821   0.489   0.6245
## DayTuesday     0.1009881  0.0655065   1.542   0.1232
## DayWednesday   0.0750537  0.0666616   1.126   0.2602
## HighC          0.1455586  0.0227150   6.408 1.47e-10 ***
## I(HighC^2)    -0.0026132  0.0005004  -5.222 1.77e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(15.7114) family taken to be 1)
##
##     Null deviance: 560.21  on 199  degrees of freedom
## Residual deviance: 203.62  on 190  degrees of freedom
##   (14 observations deleted due to missingness)
## AIC: 3167.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  15.71
##           Std. Err.:  1.58
##
##  2 x log-likelihood:  -3145.446

#relevel with baseline as saturday
NYBridges.df$Day <- relevel(NYBridges.df$Day, "Sunday")
model.nb.d<-glm.nb(Brooklyn~log1p(Rainmm)+Day+HighC+I(HighC^2),data=NYBridges
.df)
summary(model.nb.d)

##
## Call:
```

```
## glm.nb(formula = Brooklyn ~ log1p(Rainmm) + Day + HighC + I(HighC^2),
##     data = NYBridges.df, init.theta = 15.71142767, link = log)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -4.3575  -0.4805  -0.0738   0.2983   3.9887
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    5.9790278  0.2574954  23.220  < 2e-16 ***
## log1p(Rainmm) -0.2715243  0.0176485 -15.385  < 2e-16 ***
## DaySaturday    0.1133152  0.0665772   1.702  0.08875 .
## DayMonday      0.1311051  0.0673415   1.947  0.05155 .
## DayFriday      0.1185584  0.0684876   1.731  0.08344 .
## DayThursday    0.1459536  0.0679912   2.147  0.03182 *
## DayTuesday     0.2143033  0.0674407   3.178  0.00148 **
## DayWednesday   0.1883689  0.0677742   2.779  0.00545 **
## HighC          0.1455586  0.0227150   6.408 1.47e-10 ***
## I(HighC^2)    -0.0026132  0.0005004  -5.222 1.77e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for Negative Binomial(15.7114) family taken to be 1)
##
##     Null deviance: 560.21  on 199  degrees of freedom
## Residual deviance: 203.62  on 190  degrees of freedom
##   (14 observations deleted due to missingness)
## AIC: 3167.4
##
## Number of Fisher Scoring iterations: 1
##
##
##               Theta:  15.71
##           Std. Err.:  1.58
##
##  2 x log-likelihood:  -3145.446
```

```
100*(exp(confint(model.nb.d))-1)
```

```
## Waiting for profiling to be done...
```

```
##                       2.5 %        97.5 %
## (Intercept)   23655.2192952 66322.1334187
## log1p(Rainmm)   -26.4002232   -21.0172861
## DaySaturday      -1.9179299    27.8613408
## DayMonday        -0.1535242    30.1692216
## DayFriday        -1.5960193    28.8301055
## DayThursday       1.3179836    32.1561966
## DayTuesday        8.5174119    41.4508963
## DayWednesday      5.6734862    37.9264816
```

```
## HighC            10.4271304    21.0577286
## I(HighC^2)       -0.3610522    -0.1591307
```

Brief Model Interpretation
There is no evidence of a difference between the number of cyclists in Brooklyn on Saturday compared to all the weekdays.

There is evidence of a difference between the number of cyclists in Brooklyn on Sunday compared to Tuesday, Wednesday and Thursday. However, there is no evidence of a difference between the number of cyclists in Brooklyn on Sunday compared to Monday and Friday.
We estimate that, for the same level of precipitation and temperature, the expected number of cyclists on the Brooklyn bridge on Tuesday is 8.5% to 41.5% higher than Sunday.
We estimate that, for the same level of precipitation and temperature, the expected number of cyclists on the Brooklyn bridge on Wednesday is 5.7% to 37.9% higher than Sunday.
We estimate that, for the same level of precipitation and temperature, the expected number of cyclists on the Brooklyn bridge on Thursday is 1.3% to 32.2% higher than Sunday.

Which of the four models do you think is best?
From the evidence presented above the negative binomial model ($model.nb.d$) is the best. Relative to the other models the Pearson and deviance residual plots of the negative binomial model show reasonably constant variance and no pattern in the mean. Furthermore, most of the residuals are within -2 and 2. Thus the modelling assumptions have been met unlike the other models. Further, for the negative binomial model the deviance does not provide evidence for a lack of fit. In contrast, all the other models specified provide evidence against the hypothesis that the model is a good fit.

```
AIC(model.lin.a,model.pois.b,model.nb.d)

##              df       AIC
## model.lin.a  11  3061.125
## model.pois.b 10 23972.325
## model.nb.d   11  3167.446

AICc(model.lin.a,model.pois.b,model.nb.d)

##              df      AICc
## model.lin.a  11  3062.529
## model.pois.b 10 23973.489
## model.nb.d   11  3168.851

BIC(model.lin.a,model.pois.b,model.nb.d)

##              df       BIC
## model.lin.a  11  3097.406
## model.pois.b 10 24005.308
## model.nb.d   11  3203.728
```

Do you think the model fits well?
I believe the negative binomial model ($model.nb.d$) is a good fit but it does not fit the data well. This is because of the AIC, AICc and BIC scores. The AICc, AIC and BIC scores show that the linear model ($model.lin.a$) is considerably better supported than the the negative binomial model ($model.nb.d$). However, we know the linear model is not appropriate and does not fit the data well.

Thus, since it is least-worst out of the candidate models, all the candidate models also must not fit the data well.

Other variables

Additional meteorological variables I believe this dataset could benefit from are:

- Cloud cover: Days with high cloud cover may indicate chance of rain and thus deter cyclists.
- Average Wind speed: Days with higher wind speed makes bicycle riding harder increasing the possibilty of less cyclists on the Brooklyn bridge.
- Hours of sunshine: Hours of daily sunshine could give an indication on the season. For example, lower hours of sunshine could mean winter and thus less cyclists on the Brooklyn bridge due to the cold conditions.
- Relative Humidity: Quite humid days are uncomfortable and may deter cyclists.