ASSIGNMENT 4

(Chen Liang, Student ID 46275313)

Question 1

1. Introduction

Our earth is becoming warmer and warmer. As the country which is closest to the Antarctic region, the average temperature of New Zealand has maintained an upward trend in the past few years, and will rise by 6 degrees in the next century.

Due to the unique geographical environment, the rise in temperature of New Zealand will lead to a decrease in summer rainfall and increase in other seasons. This situation may cause agricultural activities to be affected by severe drought.

In addition, due to global warming, snow and glaciers are melting at an alarming rate. Research shows that in New Zealand, the most severely affected area is the Cook mountains. Since 1996, a glacier area which is 14 kilometres long, has gradually melted and formed a lake. In the basis of this development, the skiing industry that is popular with both locals and travellers will also face a crisis due to the global warming.

The increase in temperature also affects the species diversity of New Zealand. The New Zealand lizard is the only surviving reptile in the age of dinosaurs, with a very small number. However, because of global warming, the lizard, known as the living fossil of New Zealand, is at risk of extinction. Studies have shown that temperature increases the male ratio of the lizard, and may eventually become extinct due to the lack of female reproduction.

Therefore, the study of carbon dioxide emissions is of great significance not only for the preservation of ski resorts, but also for the beautiful Mount Cook and the world-famous lizards. The data set used in this study records carbon dioxide emissions and per capita emissions from 1960 to 2014.

2. Methods

The main tool for this analysis was R Studio, which was a compiler based on R, providing many embedded functions and packages for data analysis and model evaluation.

Data Description and Processing

The data set collected 55 observations for 4 variables, which were Year, CO2, CO2_per_capita and Population (using dim() and head() function to check). Then summary() function was conducted to figure out the quartiles of each variable and check if there were missing values as well.

In order to understand the relationship between variables, ggpairs () was conducted to plot the scatter matrix.

Model Fitting

In the model fitting, the generalised additive model was chosen. The tools used were from the gam package in R. Generalised additive model could use some smoothing functions to build models for

some or all of the independent variables. In this case, when fitting models to interpret the response variables, several different values of spar were set, which were 0.1, 0.3, 0.5 and 0.7. Spar, representing smooth parameters, which indicated the smoothness of the fitted model, the higher the value was set, the smoother the curve was.

After checking four different models based on varies smoothness, a model with well interpretation and appropriate smoothness would be obtained. However, when fitting some response variables, the models with the four smoothing parameters above were not ideal, so the parameter was adjusted again to fit new models.

After each model was fitted, use summary () to check the fitting result of the model and calculate the value of R- square.

3. Results

Data Description and Processing

As could be seen from Figure 1, CO2, CO2_per_capita and Population were all not linear with time (the independent variable Year). Moreover, the curves did not have good smoothness and all hold some jumping nodes in them. Therefore, it was appropriate to use GAM to fit these response variables based on Year.

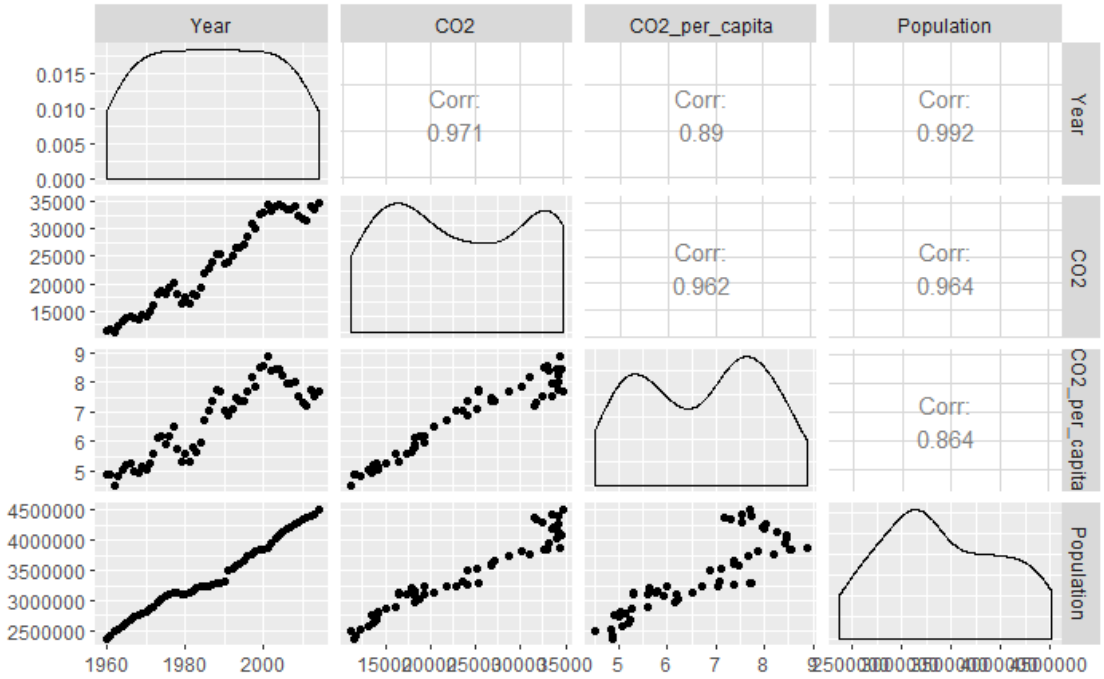
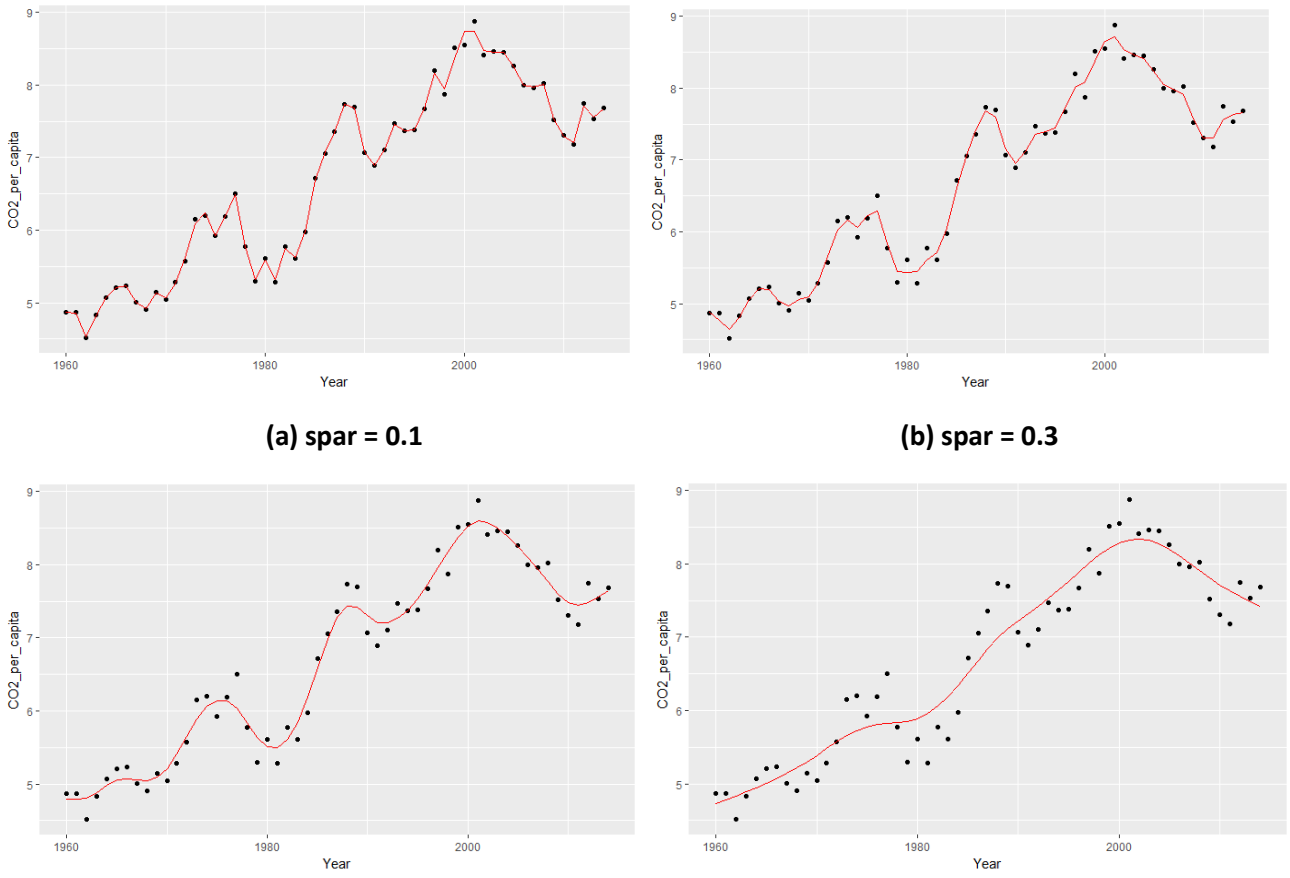


Figure 1 Scatter Plot Matrix for All Variables

Model Fitting

Figure 2 showed the fitting model when setting different smoothing parameters. It could be seen from Figure2 (a) that when the smoothing parameter was 0.1, the curve passed through almost all the observations in the sample except the response variable was closed to the peak value. After

setting the smoothing parameters up to 0.3, the smoothness of the curve (Figure 2 (b)) was improved, but there were still some unnecessary small fluctuations, such as those around year 1979 and 2002. When the smoothing parameter was set to 0.5, although the curve was better smoothing, the model did not miss any important inflection point as well, the estimation of the inflection point did not been described in place due to the pursuit of smoothness. While the smoothing parameter is 0.7, the curve was very smooth and contained rarely fluctuation.



(a) spar = 0.1 **(b) spar = 0.3**

(c) spar = 0.5 **(d) spar = 0.7**

Figure 2 Model Fitting for CO2_per_capita based on different Spar Values

Table 1 told the R square and AIC values of the models fitted above. It could be seen that when the smoothing coefficient increased, the R square value decreased and the AIC value grew up gradually. At the same time, when spar increased from 0.3 to 0.5, the value of R square decreased sharply compared to the situation when spar changed from 0.1 to 0.3.

Table 1 R Square and AIC Value of Models for CO2_per_capita

	spar = 0.1	spar = 0.3	spar = 0.5	spar = 0.7
R^2	0.9988	0.9937	0.9789	0.9307
AIC	-87.792	-30.775	1.575	51.277

The model obtained when the smoothing coefficient being set to 0.4 was displayed in Figure 3. It could be seen from the figure that all the features were captured by the model, and the peak fitting was appropriate and the smoothness was good. The R-Square of the model was 0.9869 and the value 0f AIC was -10.463.

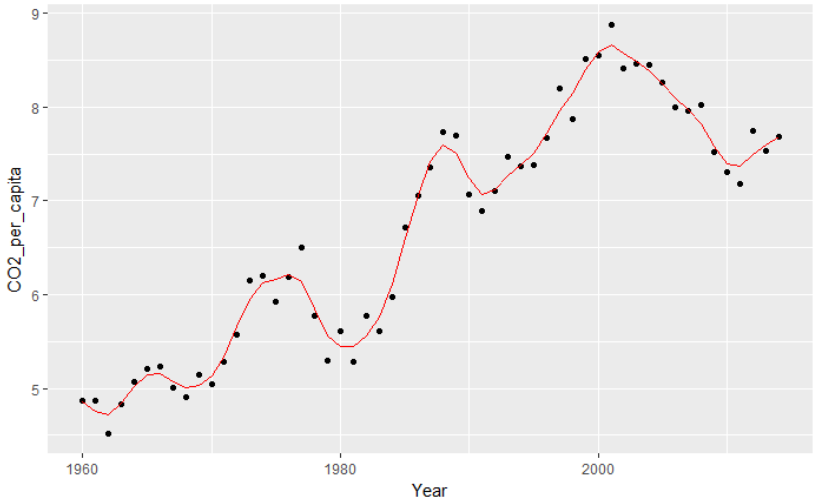
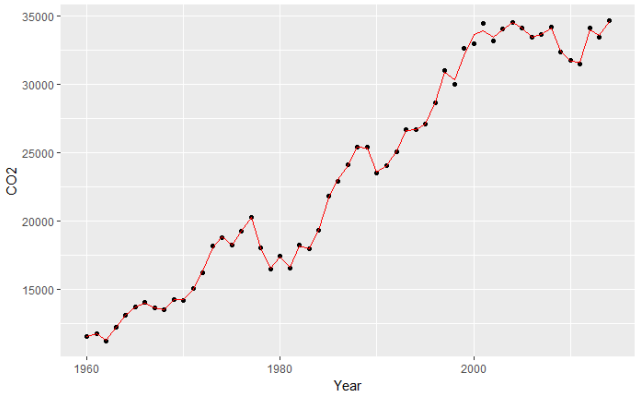
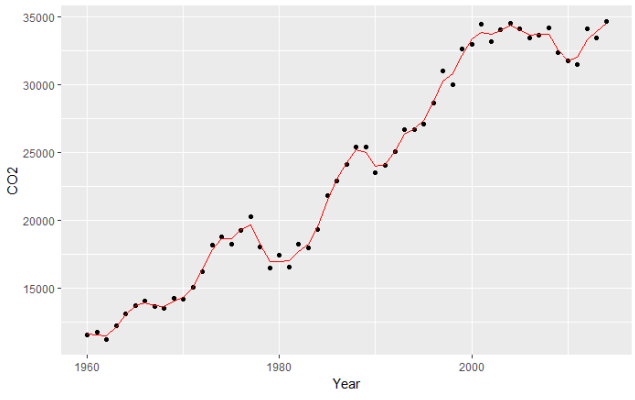


Figure 3 Model Fitting for CO2_per_capita based on Spar = 0.4

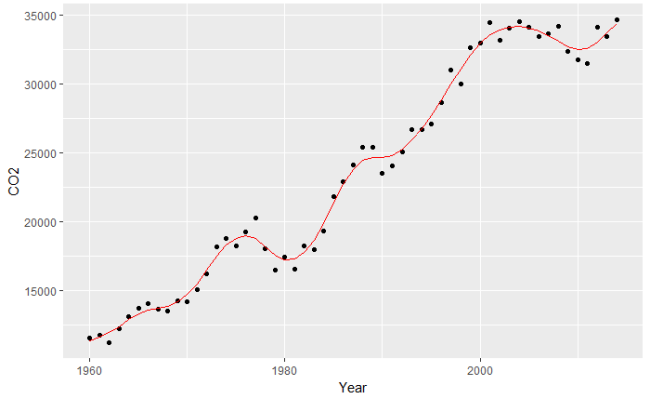
Figure 4 displayed the four models for CO2 with different smoothing coefficients. It could be seen from the Figure 4 (a) that when the spar value was chosen as 0.1, and the model captured all the sampling observations. However, some of the features were missing when spar was either 0.5 or 0.7. Therefore, in the four models, the model was most suitable when the spar value was 0.3 based on the Figure.



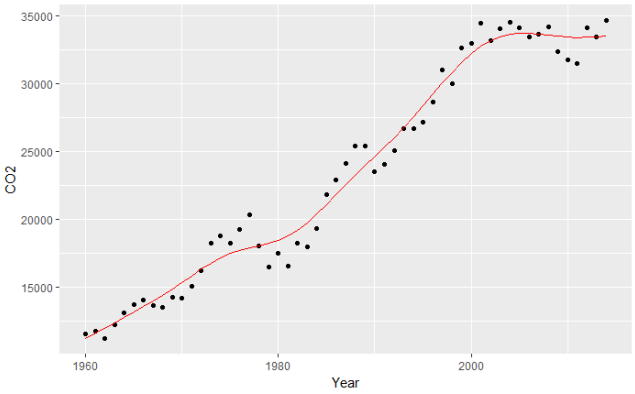
(a) spar = 0.1



(b) spar = 0.3



(c) spar = 0.5



(d) spar = 0.7

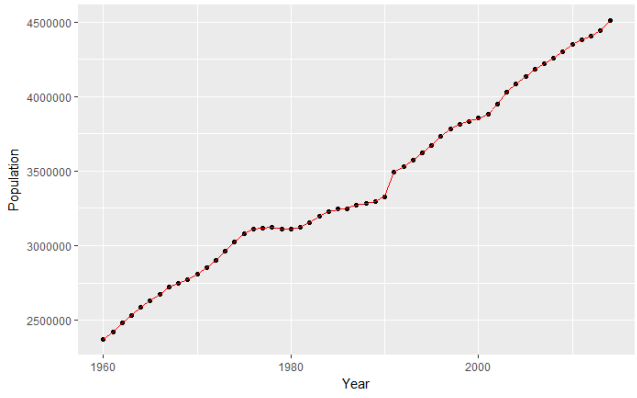
Figure 4 Model Fitting for CO2 based on different Spar Values

Table 2 showed that when the smoothing parameter increased, the R square value decreased gradually. However, when the value of spar was set as 0.1, 0.3 and 0.5, all the R-Square values were greater than 99%, and the value had a significant reduction when the value grew up to 0.7.

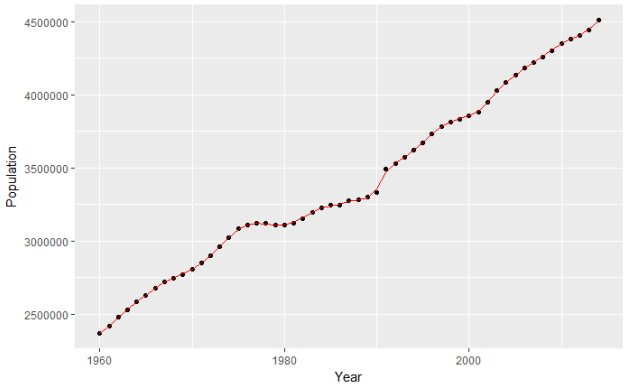
Table 2 R Square and AIC Value of Models for CO2

	spar = 0.1	spar = 0.3	spar = 0.5	spar = 0.7
R^2	0.9996	0.998	0.994	0.9808
AIC	815.377	867.78	894.895	942.908

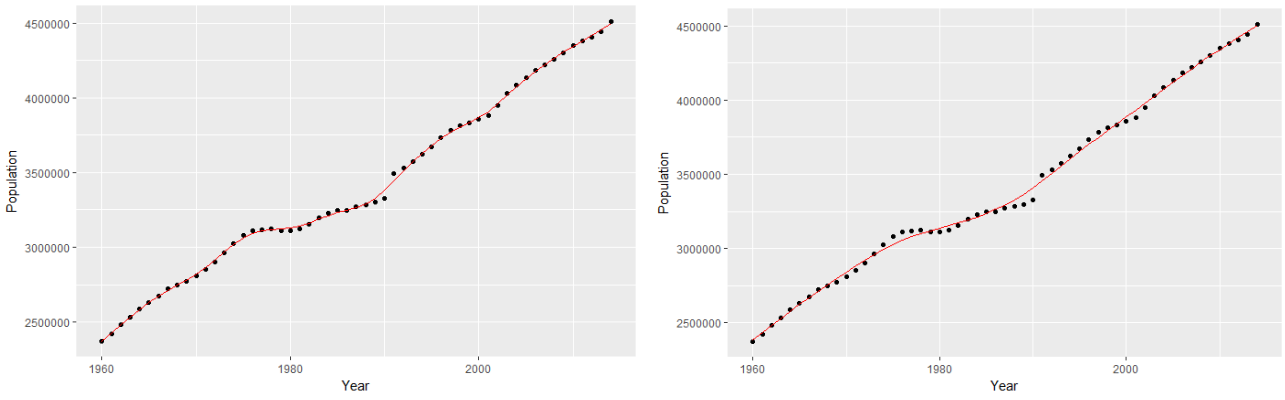
Figure 5 presented 4 models for different spar values in fitting Population changing with Year. The fluctuation of the trend was weak. Therefore, it could be seen from the Figure below that the all the obtained models could be successful to capture most of the features well.



(a) spar = 0.1



(b) spar = 0.3



(c) spar = 0.5 **(d) spar = 0.7**

Figure 5 Model Fitting for Population based on different Spar Values

As could be seen from Table 3, the R square for the four models all exceeded 99%. Moreover, with the increase of smoothing parameter, it only decreased slightly.

Table 3 R Square and AIC Value of Models for Population

	spar = 0.1	spar = 0.3	spar = 0.5	spar = 0.7
R^2	0.9999	0.9999	0.9994	0.998
AIC	1086.331	1178.276	1239.253	1294.239

4. Discussion

Model for CO2_per_capita

When spar was set as 0.3, the model was complex, and the description of some features of the sample was not accurate when taking 0.5. At the same time, the R square value decreased quickly when spar went down from 0.3 to 0.5. Therefore, the model was refitted with setting spar as 0.4, and a model with lower complexity and better interpretability was obtained. At this time, the R squared value was 0.9896, indicating that 98.96% of the deviance of CO2_per_capita can be explained by the model.

Model for CO2

Although the model using spar as 0.3 seemed to be more reasonable being based on Figure 4, the decrease of R-Square was very small when spar was set to 0.5 compared to 0.3. This meant that when spar was 0.5, the model could already capture the features of the sample well enough. Therefore, a simpler model with spar as 0.5 was chose for interpreting CO2. The R square was 0.994 which meant that 99.4% of the deviance of CO2 could be explained by the model.

Model for Population

Figure 5 showed that the four models all fitted the data well, and hold a high R square (Table 3), so the simplest model was chosen here, that was, the model fitted by spar at 0.7. In addition, with the increase of smoothness, the fitting performance did not go down, which indicated that the

relationship between population and year was quite close to linearity. What is more, Figure 1 also evidenced that the distribution of Population was approximately normal which reached the assumption of linear regression, so this model could also be fitted by linear regression. The R square was 0.998 which meant that 99.8% of the deviance of Population could be explained by the model.

Question 2

1. Introduction

When driving on the roads in New Zealand, cows can be seen walking through the mountains anywhere, which are the foundation of the dairy and meat industry in New Zealand.

However, cows need to digest about 68 kilograms of grass, hay, silage and 9 kilograms of concentrated feed every day. When cows digest food, they release hydrogen and carbon dioxide, while Archaea transform some of the gas they release into methane. Therefore, each cow can produce enormous amounts of carbon dioxide and methane every single day, mainly by burping and breathing. In New Zealand, methane and carbon dioxide produced by burping and breathing of cows account for nearly half of the total greenhouse gas emissions in the country.

To reduce greenhouse gas emissions and effectively curb the greenhouse effect, it is necessary to study greenhouse gas released by cows.

There were 429 pieces of data of 27 cows in two months were recorded in this study. The purpose was to study whether the carbon dioxide and methane emitted by cows were related to other variables and fit model for carbon dioxide and methane emission changing with time.

2. Methods

The main tool for this analysis was R Studio, which was a compiler based on R, providing many embedded functions and packages for data analysis and model evaluation.

Data Description and Processing

First, use summary () to check the statistics of the data, and conduct na.omit () to remove the observations with missing values. In addition, the variable representing the cow numbers was integer, so factor () was implemented to turn it into factor. Next, use ggpairs () to figure out the relationship of variables except cow and time (date and day).

In order to know the trend of CO2 and CH4 changing with TotDM and Pasture, scatter diagrams (using geom_point() in ggplot2 package)and conditioning plots (using coplot ()) were selected. In the conditioning plots, the cow was kept as given condition, and the relationship of CO2 and CH4 with TotDM and Pasture for each cow were described respectively.

Then, use boxplot to check the differences of CO2 and CH4 emissions among different cows. The box plot was plotted using the geom_boxplot () function, and could display the overall distribution and quantile clearly for each cow.

At last in this part, coplot () was used to figure out the change of CO2 and CH4 emissions per cow

over time. The cow was given in the figure, each plot showed the amount of CO2 (or CH4) emissions varied with time for each cow.

Model Fitting

In the model fitting, the generalised additive model was firstly chosen. The tools used were from the gam package in R. Generalised additive model could use smoothing function to build models for some or all the independent variables. In this case, nonparametric smooth fitting for some variables were needed. For variables being added into the model, testing their significance when using smooth functions or not, and remove the insignificant ones (P value > 0.05).

Next, polynomial linear regression was used to fit how CO2 and CH4 changed with time. The fitting process started with the full model, and the variables were removed by stepwise (realised by step ()). In this case, because of the nonlinear relationship between the greenhouse gas emission and some predictors could be found out in the figure, the polym() (degree=2) function was implemented in the fitting process.

Finally, a linear mixed model was used, base on the data in which 27 cows in two months were measured in this experiment. Since the difference among cows would produce random errors, cow was then added to the model fitting process as a random effect to fit the linear fitting model.

In the regression diagnosis, the QQ plot was used to diagnose the normality of residual of the above three models. In the QQ plot, if all the points were approximately in a straight line, the regression assumption was true.

Model Comparison

In order to evaluate the effect of the model, the Adjusted R square (in gam and lm) and AIC (a criterion for evaluating the regression model) value of the models were calculated.

Adjusted R square indicates the goodness of fit, and the obtained value is the proportion of the variance of the response variable that can be described by the model.

The AIC value reaches the minimum Kullback–Leibler distance between the estimated value and the true value, while controlling the over fitting. AIC encourages the goodness of data fitting, but avoids overfitting as far as possible, so the best model is the one with the lowest AIC value. By taking the model with lowest AIC value, the best model for interpreting data but with the least free parameters could be got.

3. Results

Data Description and Processing

Figure 1 displayed the relationship of the variables except cow and variables representing time (day and date). As could be seen from the figure, the correlation between Pasture and TotDM was as high as 0.994 and that between Pellets and Ndropperday was 1. In addition, the correlation between the two response variables was strong as well. Figure 1 also told that the distributions of CO2 and CH4 were approximately normal distribution and they were not need to be transformed when fitting the models.

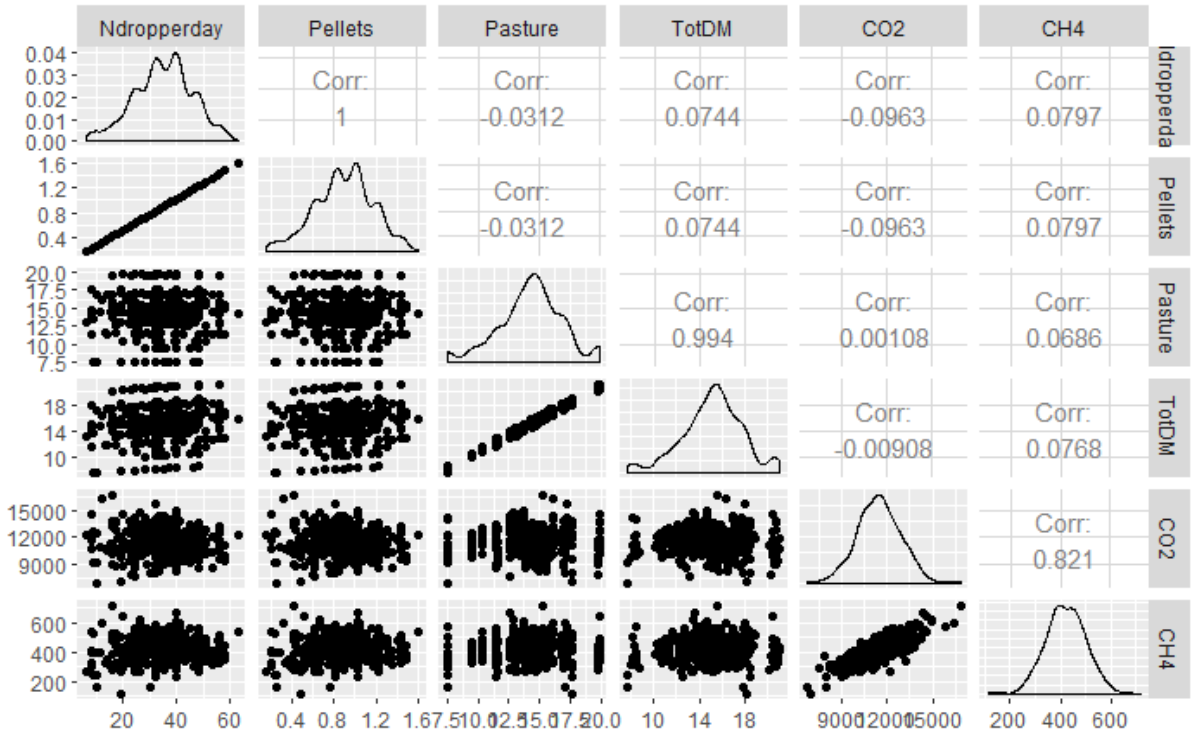
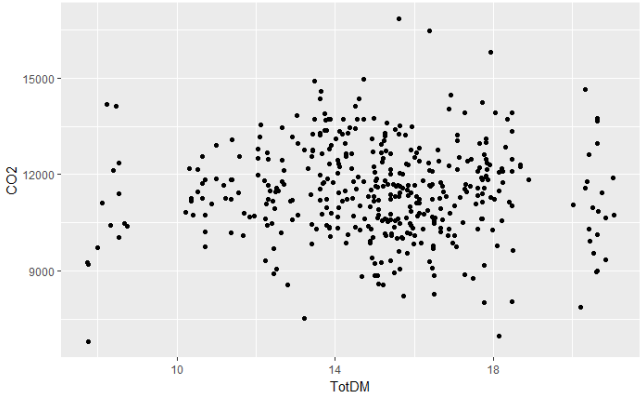
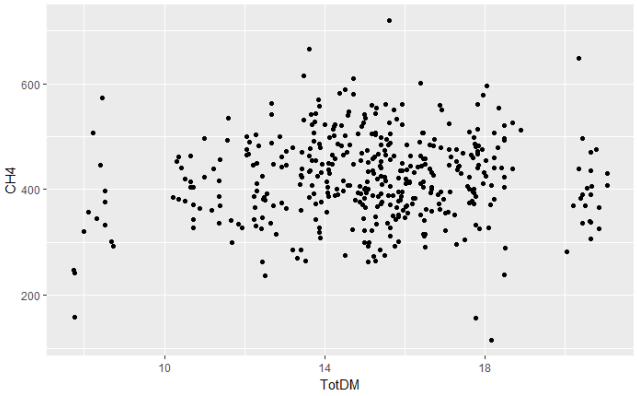


Figure 1 Relationship between Variables

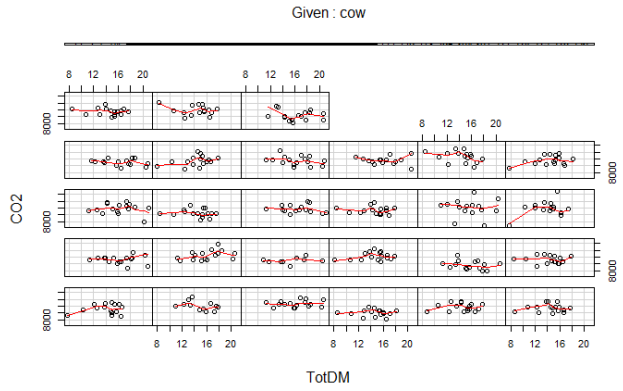
Figure 2 was the change of CO2 and CH4 emission with TotDM, which displayed that the intake for most cows were between 10kgDM and 19kgDM per day. From (a) (c), it could be seen that with the growth of TotDM, CO2 performed no strong evidence for an obvious uniform trend. However, (d) told that for most cows, the emission of CH4 would perform a slight hump and reach a peak within the range when TotDM was between 13 and 18.

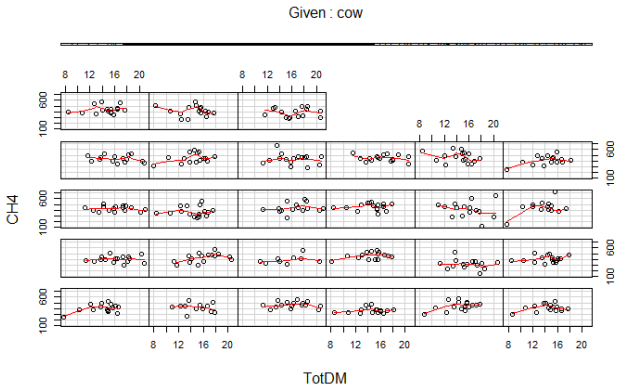


(a) Scatterplot for CO2 changing by TotDM



(b) Scatterplot for CH4 changing by TotDM

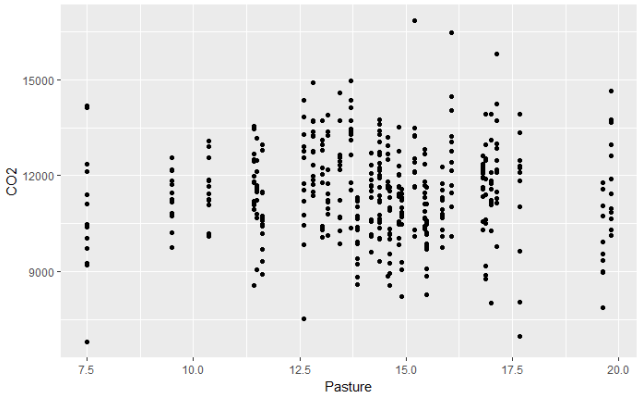


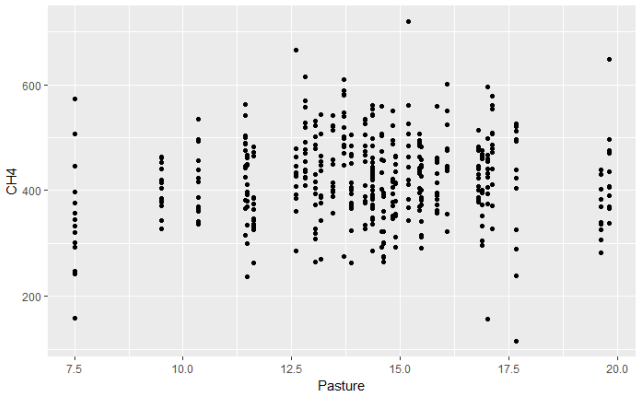


(c) CO2 changing by TotDM for different cow **(d) CH4 changing by TotDM for different cow**

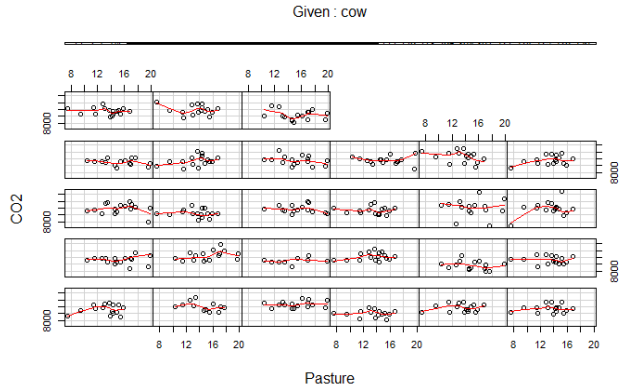
Figure 2 Plots for CO2 and CH4 and TotDM

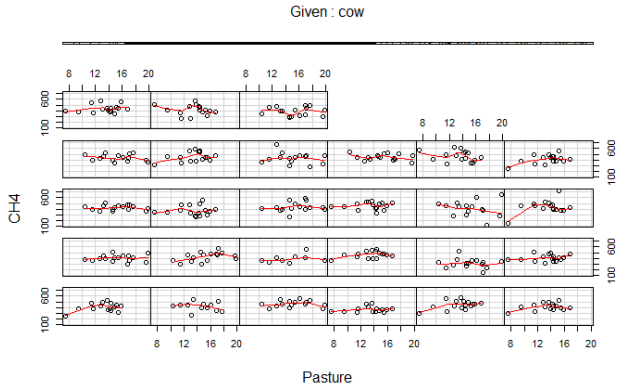
Figure 3 showed the change of CO2 and CH4 emission with Pasture. The information provided by the plots was similar to figure 2, which was that the trend of CO2 was not obvious. At the same time, the change of CH4 displayed a slight hump. This was also consistent with the highly collinearity between Pasture and TotDM in Figure 1.





(a) Scatterplot for CO2 changing by Pasture **(b) Scatterplot for CH4 changing by Pasture**





(c) CO2 changing by Pasture for different cow **(d) CH4 changing by Pasture for different cow**

Figure 3 Plots for CO2 and CH4 changing by Pasture

It could be seen in Figure 4 which was the CO2 emission for each cow that there were some differences. Firstly, from the location of the box, it could be figured out that some cows emitted more CO2, such as No. 321, while some emit relatively less, such as No. 151. In addition, the distribution of CO2 values emitted by cow was different as well. Some of them like No. 222 were more dispersed, some of which were more concentrated and varied less, which could be like No. 133.

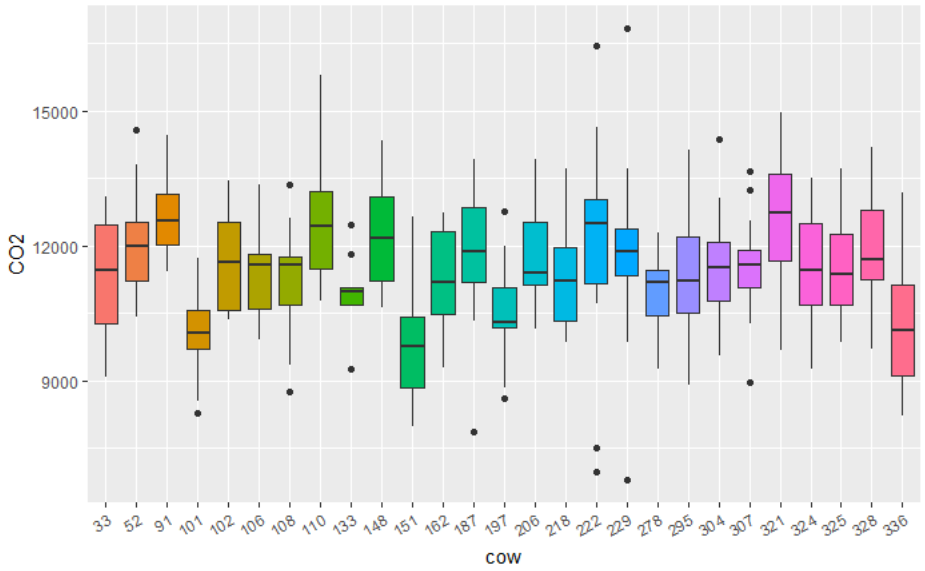


Figure 4 CO2 Emission for Each Cow

Compared with CO2 emission, the difference of CH4 emission for cows which was shown in Figure 5 was smaller, with median distributing between 300~500 g/d. No. 148 had the greatest median while No. 151 was the smallest. For the distribution of the value, No. 222 varied most.

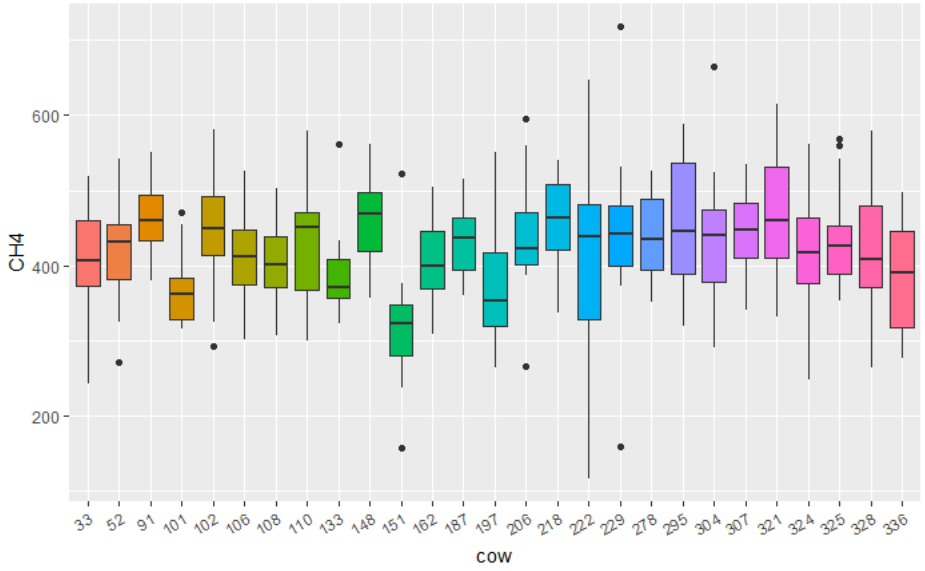
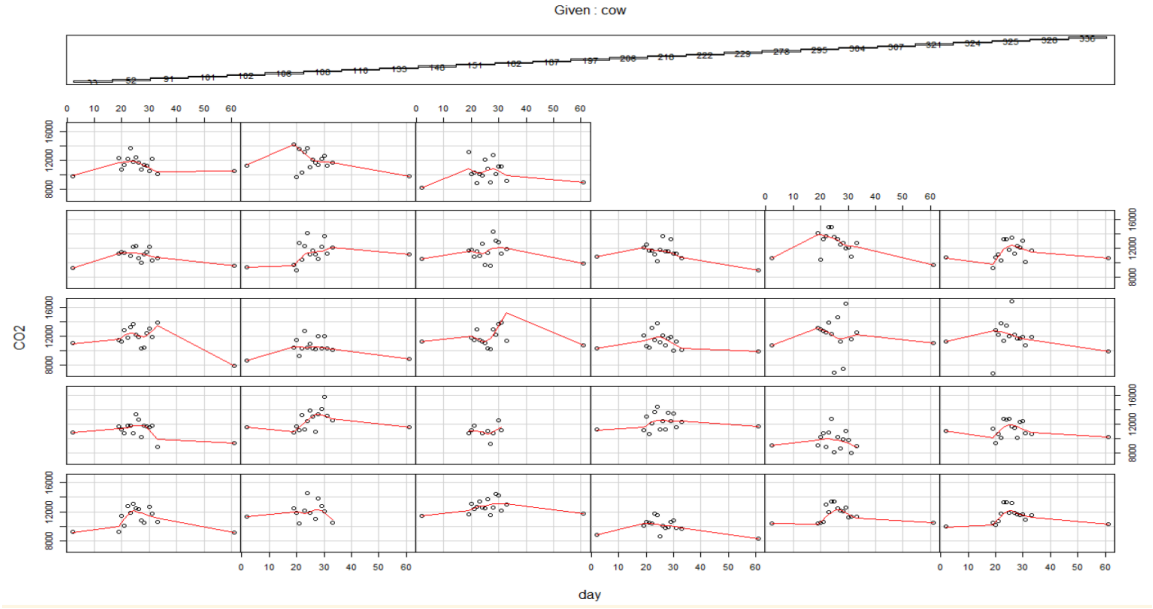
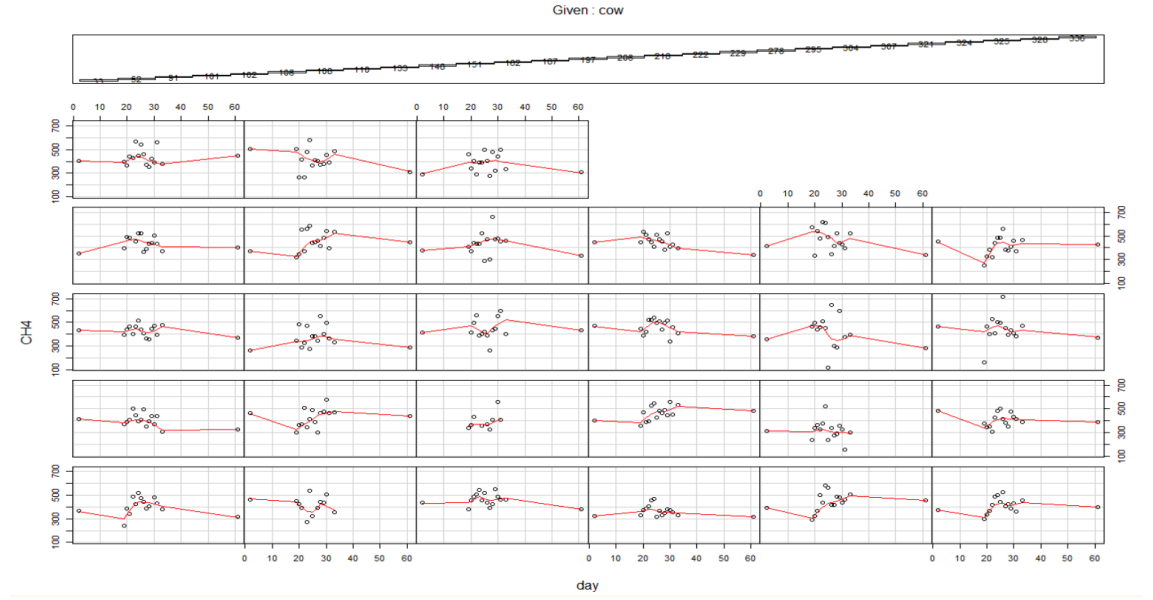


Figure 5 CH4 Emission for Each Cow

Figure 6 showed the change in emissions of CO2 and CH4 over time. It could be seen from the figure that most cows, excepting cows holding missing data, tend to remain stable (slightly rising or descending) at first, and reached a peak between 20 and 30 days and then slowed down. Some of them also had more complex fluctuations, such as cattle No. 321 in (b). In addition, the scatter plots in Appendix Figure 1 and 2 could also show a higher value in the middle part of the whole period.



(a) CO2 Emission for Each Cow changing by day



(b) CH4 Emission for Each Cow changing by day

Figure 6 Greenhouse gas Emission Changing with Time

Model Fitting

After removing the variables with p-value which was greater than 0.05, the final generalised additive model for CO2 was shown as model 1, which took cow as Parametric effects and day as nonparametric effects. The final generalised additive model for CH4 was shown as Model 2. In this model, cow was included into the parametric effects, day and Pasture were taken into nonparametric effects. The QQ plots for regression diagnose were shown in Appendix Figure 5 and Figure 6.

CO2 ~ s(day) + cow **(Model 1)**

CH4 ~ s(day) + s(Pasture) + cow **(Model 2)**

Table 1 Significance of Parameters of Model 1

Parameter Effects	P value	assessment by R	Nonparameter Effects	P value	assessment by R
cow	2.425e-16	***	s(day)	<2.2e-16	***

Table 2 Significance of Parameters of Model 2

Parameter Effects	P value	assessment by R	Nonparameter Effects	P value	assessment by R
cow	2.449e-0.8	***	s(day)	1.884e-05	***
			s(Pasture)	0.01203	*

The final models obtained by linear regression were model 3 and model 4. For the response variable CO2, the predicting variables were cow and the two-degree polym (day). For the response variable CH4, the predictors were cow, two-degree polym (day) and polym (Pasture). The QQ plots for regression diagnose were shown in Appendix Figure 7 and Figure 8.

CO2 ~ cow + Polym(day, degree=2) **(Model 3)**

CH4 ~ cow + Polym(day, degree=2) + Polym(Pasture, degree=2) **(Model 4)**

Table 3 Significance of Parameters of Model 3

	p value	assessment by R
cow	4.031e-15	***
Polym(day)	1.316e-15	***

Table 4 Significance of Parameters of Model 4

	p value	assessment by R
Polym(Pasture)	0.00019	***
cow	9.99e-08	***
Polym(day)	9.427e-05	***

The obtained final linear mixed models were model 5 and Model 6. In model 5, the fixed effect contained the two-degree polym (day). In the fixed effect of model 6, the variables being kept were cow, two-degree polym (day) and polym (Pasture). The QQ plots for regression diagnose were shown in Appendix Figure 9 and Figure 10.

CO2 ~ Polym(day, degree=2)+ (1 | cow) **(Model 5)**

CH4 ~ Polym(Pasture, degree=2) + Polym(day, degree=2) + (1 | cow) **(Model 6)**

Table 5 Fixed Effects and Random Effects of Model 5

Fixed Effects	Estimate	p value	Random Effects	Variance	Std.Eev
intercept	11465.9	<2e-16	cow	449243	670.3
Polym(day)1	-2456.3	0.0436	Residual	1467163	1211.3
Polym(day)2	-10151.3	1.07E-15			

Table 6 Fixed Effects and Random Effects of Model 6

Fixed effects	Estimate	p value	Random Effects	Variance	Std.Eev
intercept	421.003	<2e-16	cow	891.5	29.86
Polym(Pasture)1	204.77	0.0168	Residual	5376.2	73.32
Polym(Pasture)2	-248.647	0.00117			
Polym(day)1	-116.634	0.15873			
Polym(day)2	-298.942	6.22e-0.5			

Model Comparison

Table 7 and table 8 listed the adjusted R square values that needed to be used for model comparison. It could be seen from the tables that the adjusted R square of GAM was higher than polynomial LM for the model of both CO2 and CH4 emissions, so fit of GAM was closer. However, the lowest AIC values in the two tables were all produced by linear mixed models.

Table 7 Adjusted R square and AIC of Models for CO2

CO2	Adjusted R square	AIC
GAM	0.386	7105.109
LM	0.315	7117.235
LMM		7097.413

Table 8 Adjusted R square and AIC of Models for CH4

CH4	Adjusted R square	AIC
GAM	0.266	4782.142
LM	0.186	4785.711
LMM		4749.312

4. Discussion

When fitting the CO2 and CH4 emission, the variables Pasture, Pellets, cow and day were selected. This was due to firstly, the strong correlation between Pasture and TotDM, Pellets and Ndropsperday in Figure 1. Therefore, one variable of each group was kept for fitting the model. From the Figure 4 and Figure 5 , it could be seen that there were obvious differences of CO2 and CH4 emitted by

different cows, and Figure 6 showed evidence that CO2 and CH4 would change with day. Therefore, when fitting the data, the original models of several types of models all included Pasture, Pellets, cow and day.

In the fitting process, the generalized additive model, multivariate linear model and linear mixed model were selected for comparison.

The reason for the use of the generalized addition model was that in the Figure 6 CO2 and CH4 had a slightly humped function of day, and there were some discontinuous points in the data. Therefore, the generalized additive model should to be used to conduct nonparametric smoothing function for day. In the process of fitting, CO2 and CH4 did not hold a unified trend for each cow with the change of time. Therefore, in order to prevent over fitting, the smoothing parameters was not defined as a smaller value to fit the data, but using the default value instead, which could help get a low complexity model for all cows to use.

At the same time, the polynomial linear model was conducted to fit the non-linear relationship shown by Figure 6 by adding higher order of day. Finally, the linear mixed model was used, because the difference in the cows in the data would bring random error into the model, which could be estimated by linear mixed model by adding cow as a random effect.

In the process of model fitting, in addition to day, Pasture was considered as a nonlinear term, that was, the smooth function was used on it in GAM, and quadratic term was added into both LM and LMER. This was because, according to the information provided in Figure 3, the effect of Pasture on CO2 and CH4 emission was highly likely to require higher order fitting.

Models Selection for Response variable CO2

Model 1, model 3 and model 5 showed that the predictors which were significant for final models obtained by the three methods were all day and cow, which were consistent with the information conveyed by Figure4 and Figure6, and other variables had little effect on the emission of CO2.

Table 7 showed that compared with model 3, the adjusted R square value of model 1 was larger than that of model 3, indicating that model 1 fitted the data better. However, the difference of the AIC values of the three models was small, and the model 5 owed the smallest one, which evidenced that fitting day and Pasture using higher order was reasonable in polynomial linear regression and linear mixed models. However, because cow was added to the model as factors, models 1 and 3 cannot work if models were used to predict the CO2 emission changing by day for new cows. At this point, model 5 should be chosen. When predicting, model 5 would add random effects to the prediction of new cows according to the average level of the existing cows.

Models Selection for Response variable CH4

Model 2, model 4 and Model 6 were the three final models for CH4. The predictors were day, Pasture and cow, which was consistent with the information provided by figures. It could be seen from table8 that the minimum AIC value was still given by the linear mixture model, which showed that it was suitable to fit quadratic to Pasture and day based on polym (). Like model 5, the model 6 had predictive function, so it was selected here.

Conclusion

Through data analysis and modelling, it could be known,

(a) in this sample, in addition to each cow's own differences, the amount of CO2 emissions is only affected by time, and it increases first and then decreases with time for a certain cow. Model 5 obtained from linear mixture model is the best model for predicting CO2 emissions based on time.

(b) for this sample, the emission of CH4 is related to the cow itself, time and the pasture eaten per day. If pasture eaten per day is fixed, the emission of CH4 will slightly increase and then decrease with time for a certain cow. The predicting model for CH4 emission is recommended the use of model6.

Reference

[1] R in Action. Robert I, Kabacoff, 2013

[2] Yi X, Liping C, etc. <Statistical model and R>

Appendix: Figures

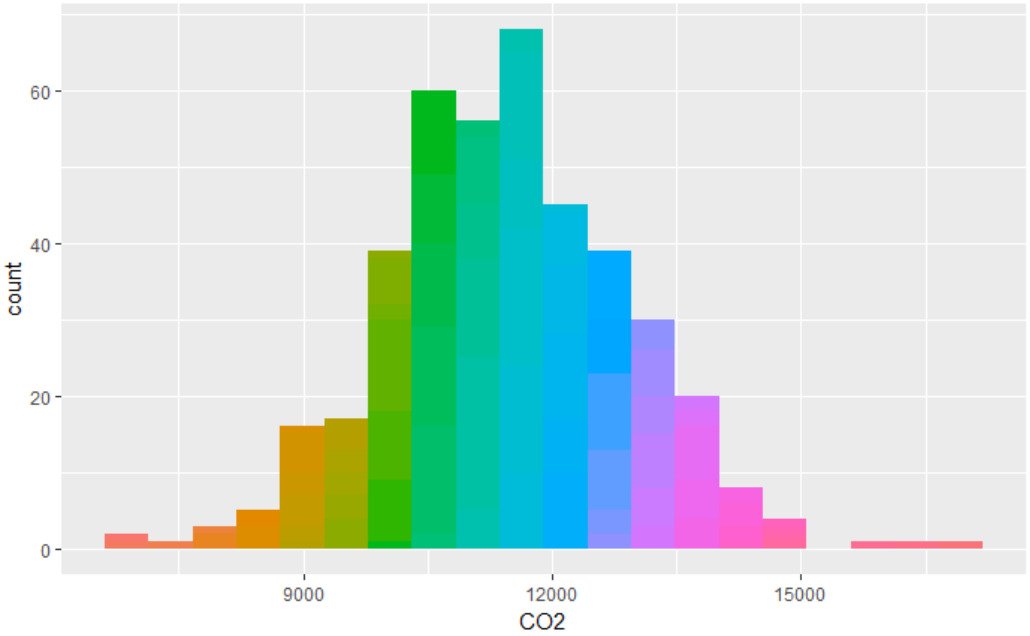


Figure 1 distribution of CO2

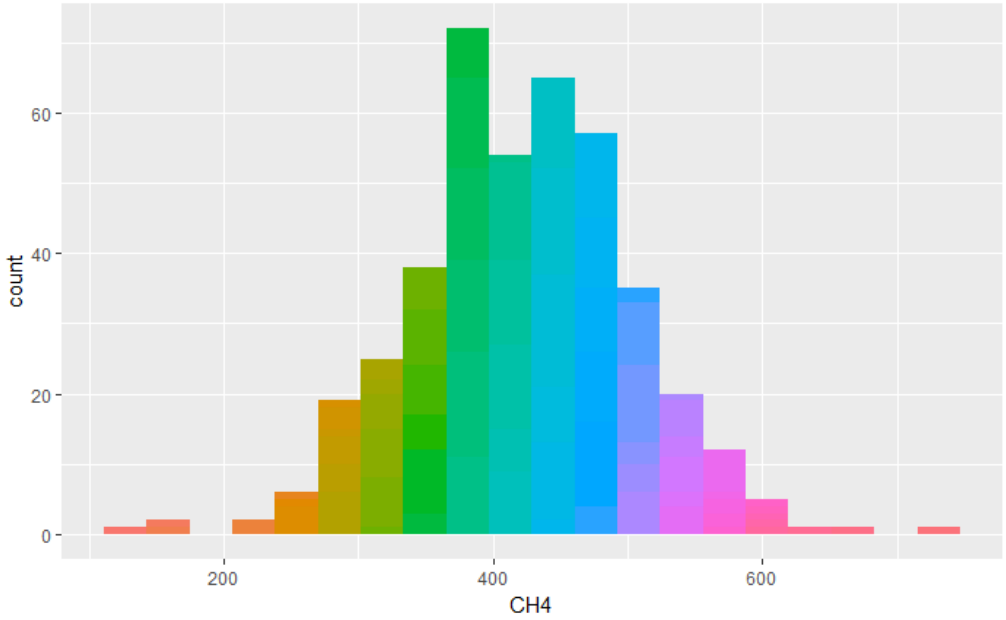


Figure 2 distribution of CH4

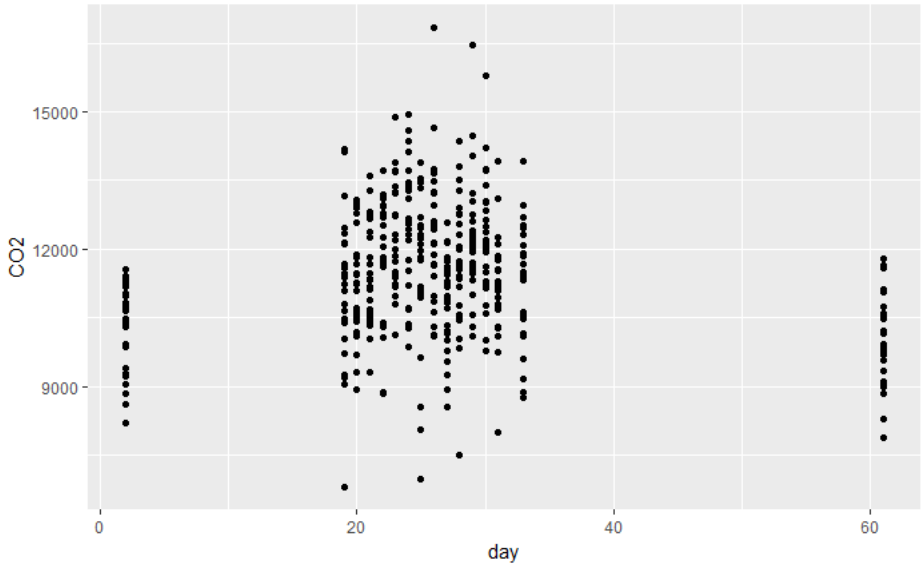


Figure 3 CO2 Emission changing with day

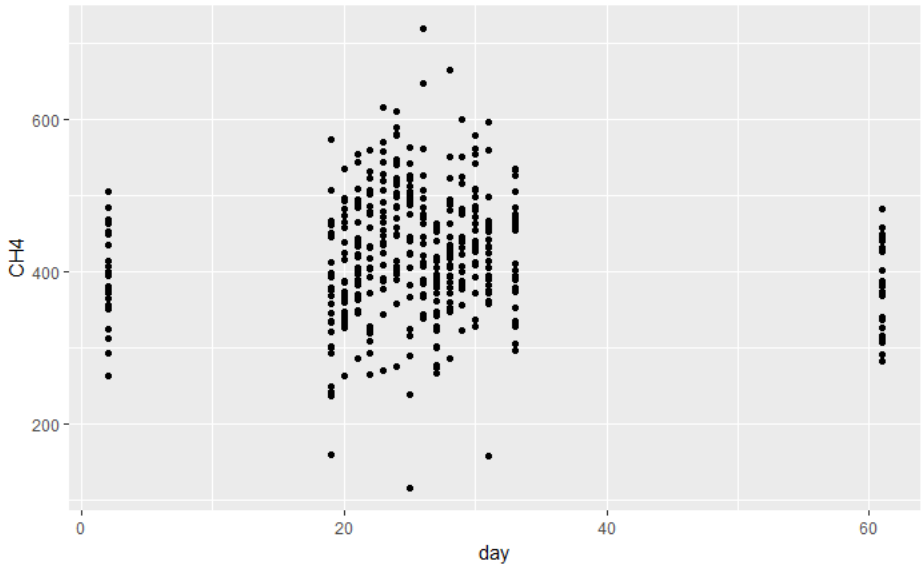


Figure 4 CH4 Emission changing with day

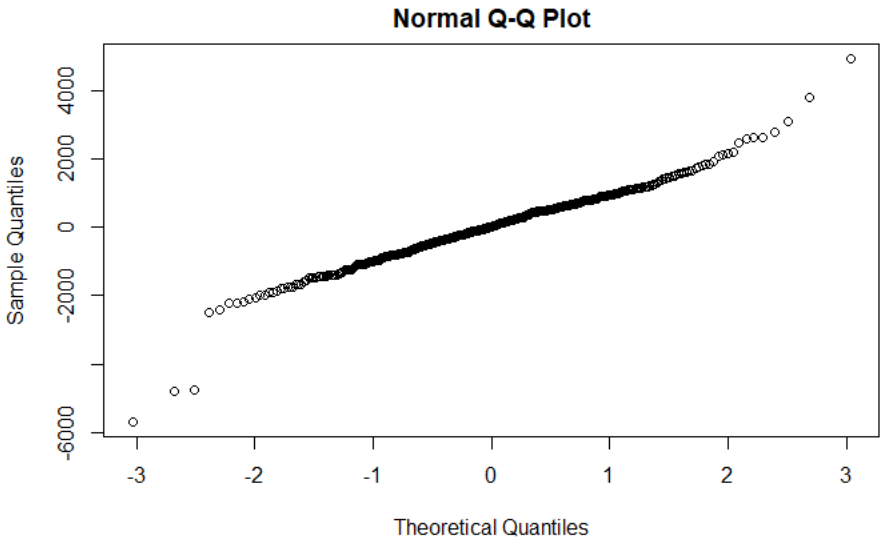


Figure 5 QQ Plot for Model1

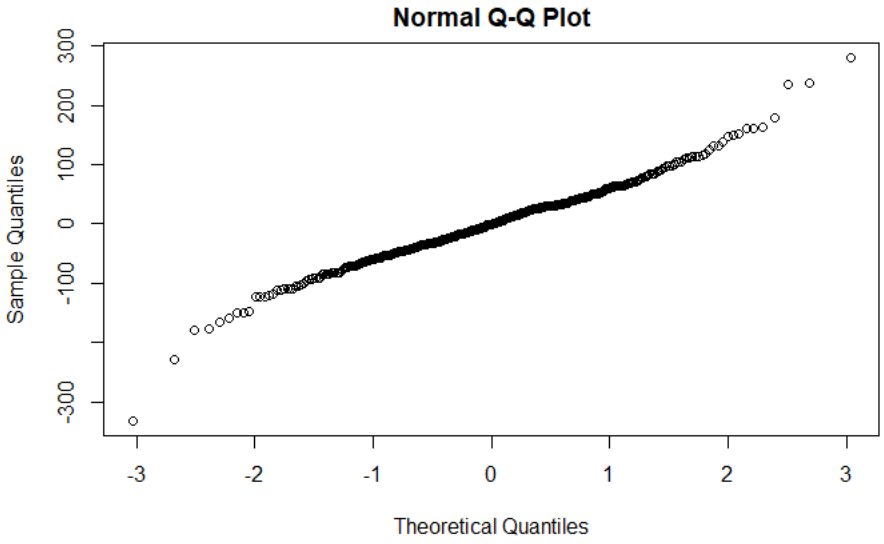


Figure 6 QQ Plot for Model2

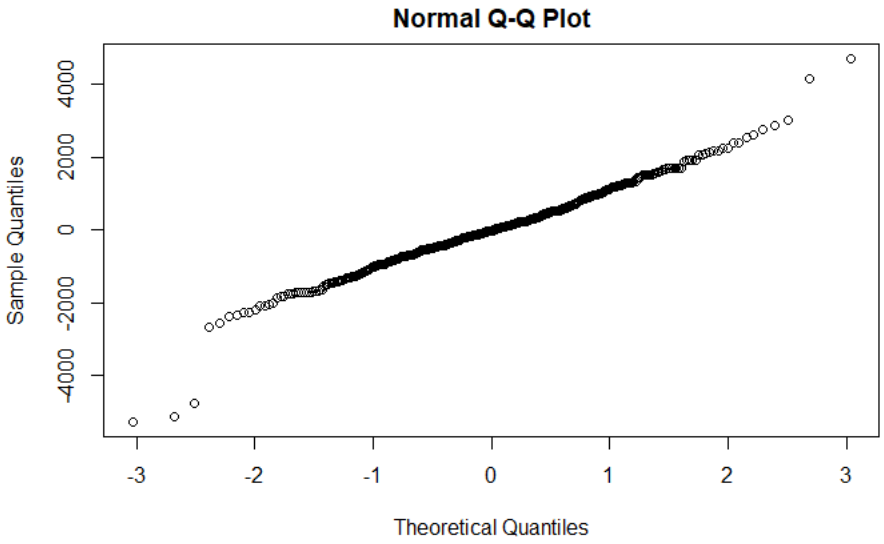


Figure 7 QQ Plot for Model3

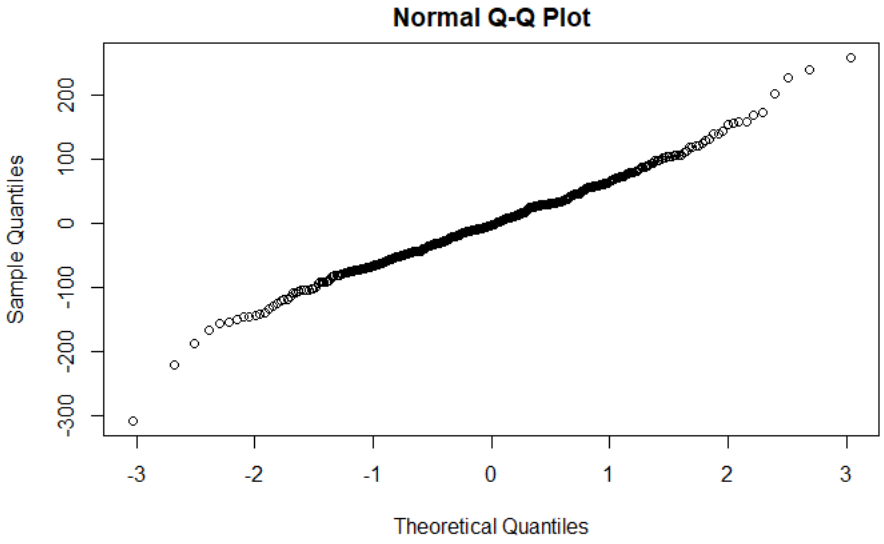


Figure 8 QQ Plot for Model4

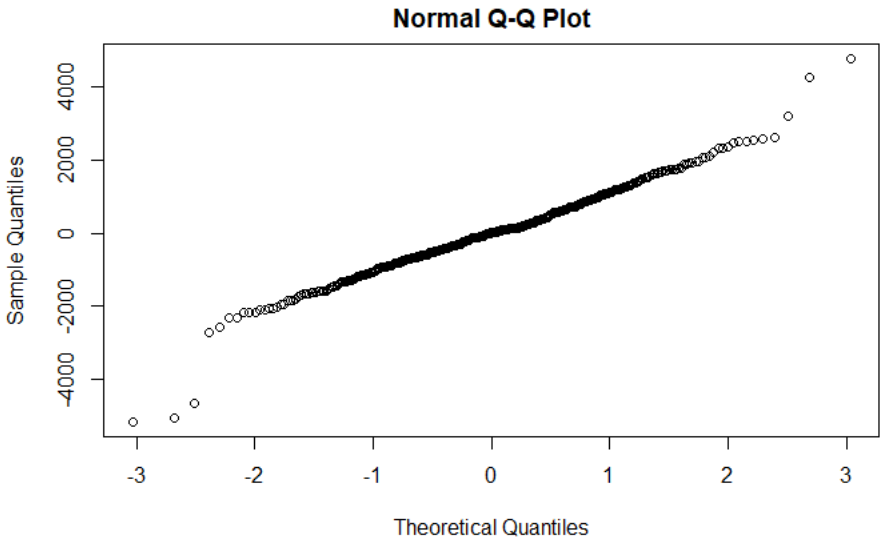


Figure 9 QQ Plot for Model5

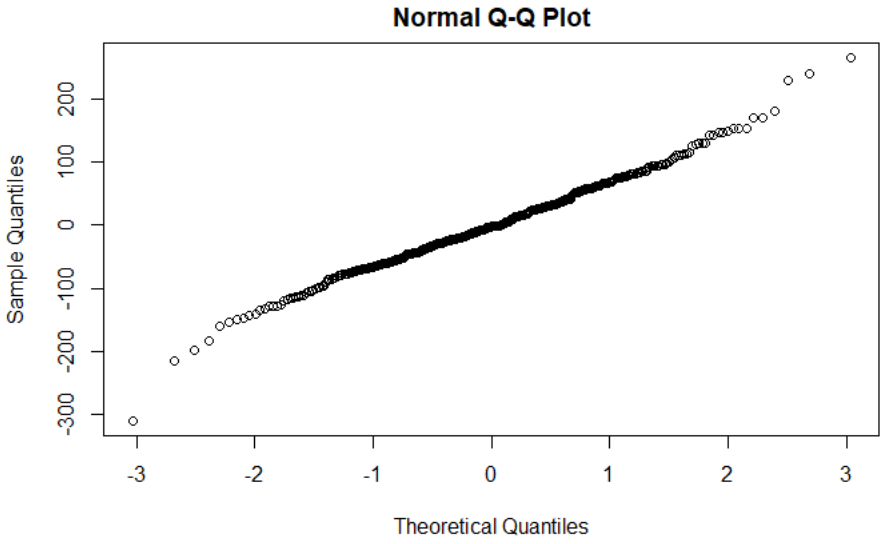


Figure 10 QQ Plot for Model6

Appendix: R CODE

QUESTION1

```
library(GGally)
co2<-read.csv('NZCO2.csv',header = TRUE)
head(co2)
dim(co2)
summary(co2)
ggpairs(co2)
library(ggplot2)
plot1 <- ggplot(co2, aes(x=Year,y=CO2)) + geom_point()
plot1 <- plot1 + theme_bw()
plot1
plot2 <- ggplot(co2, aes(x=Year,y=CO2_per_capita)) + geom_point()
plot2 <- plot2 + theme_bw()
plot2
plot3 <- ggplot(co2, aes(x=Year,y=Population)) + geom_point()
plot3 <- plot3 + theme_bw()
plot3
```

#fit CO2_per_capita

```
library(gam)
yearCO2_per_capita_1<- gam(CO2_per_capita ~ s(Year, spar=0.1), data=co2)
scapital1<-summary(yearCO2_per_capita_1)
scapital1
R1_1 <- (scapital1$null.deviance-scapital1$deviance)/scapital1$null.deviance
R1_1
co2$fits1 <- fitted(yearCO2_per_capita_1)
Plot101 <- ggplot(co2, aes(x=Year, y=CO2_per_capita)) + geom_point()
Plot101 <- Plot101 + geom_line(aes(x=Year, y=fits1), color="red")
Plot101

yearCO2_per_capita_3<- gam(CO2_per_capita ~ s(Year, spar=0.3), data=co2)
scapital3<-summary(yearCO2_per_capita_3)
scapital3
R1_3 <- (scapital3$null.deviance-scapital3$deviance)/scapital3$null.deviance
R1_3
co2$fits3 <- fitted(yearCO2_per_capita_3)
Plot103 <- ggplot(co2, aes(x=Year, y=CO2_per_capita)) + geom_point()
Plot103 <- Plot103 + geom_line(aes(x=Year, y=fits3), color="red")
Plot103

yearCO2_per_capita_5<- gam(CO2_per_capita ~ s(Year, spar=0.5), data=co2)
scapital5<-summary(yearCO2_per_capita_5)
```

```
scapital5
R1_5 <- (scapital5$null.deviance-scapital5$deviance)/scapital5$null.deviance
R1_5
co2$fits5 <- fitted(yearCO2_per_capita_5)
Plot105 <- ggplot(co2, aes(x=Year, y=CO2_per_capita)) + geom_point()
Plot105 <- Plot105 + geom_line(aes(x=Year, y=fits5), color="red")
Plot105

yearCO2_per_capita_7<- gam(CO2_per_capita ~ s(Year, spar=0.7), data=co2)
scapital7<-summary(yearCO2_per_capita_7)
scapital7
R1_7 <- (scapital7$null.deviance-scapital7$deviance)/scapital7$null.deviance
R1_7
co2$fits7 <- fitted(yearCO2_per_capita_7)
Plot107 <- ggplot(co2, aes(x=Year, y=CO2_per_capita)) + geom_point()
Plot107 <- Plot107 + geom_line(aes(x=Year, y=fits7), color="red")
Plot107

yearCO2_per_capita_4<- gam(CO2_per_capita ~ s(Year, spar=0.4), data=co2)
scapital4<-summary(yearCO2_per_capita_4)
scapital4
R1_4 <- (scapital4$null.deviance-scapital4$deviance)/scapital4$null.deviance
R1_4
co2$fits4 <- fitted(yearCO2_per_capita_4)
Plot104 <- ggplot(co2, aes(x=Year, y=CO2_per_capita)) + geom_point()
Plot104 <- Plot104 + geom_line(aes(x=Year, y=fits4), color="red")
Plot104
```

##fit CO2

```
yearco2_01<- gam(CO2 ~ s(Year, spar=0.1), data=co2)
sco21<-summary(yearco2_01)
sco21
R2_1 <- (sco21$null.deviance-sco21$deviance)/sco21$null.deviance
R2_1
co2$fits01 <- fitted(yearco2_01)
Plot201 <- ggplot(co2, aes(x=Year, y=CO2)) + geom_point()
Plot201 <- Plot201 + geom_line(aes(x=Year, y=fits01), color='red')
Plot201

yearco2_03<- gam(CO2 ~ s(Year, spar=0.3), data=co2)
sco23<-summary(yearco2_03)
sco23
R2_3 <- (sco23$null.deviance-sco23$deviance)/sco23$null.deviance
R2_3
co2$fits03 <- fitted(yearco2_03)
Plot203 <- ggplot(co2, aes(x=Year, y=CO2)) + geom_point()
Plot203 <- Plot203 + geom_line(aes(x=Year, y=fits03), color='red')
```

```
Plot203

yearco2_05<- gam(CO2 ~ s(Year, spar=0.5), data=co2)
sco25<-summary(yearco2_05)
sco25
R2_5 <- (sco25$null.deviance-sco25$deviance)/sco25$null.deviance
R2_5
co2$fits05 <- fitted(yearco2_05)
Plot205 <- ggplot(co2, aes(x=Year, y=CO2)) + geom_point()
Plot205 <- Plot205 + geom_line(aes(x=Year, y=fits05), color='red')
Plot205

yearco2_07<- gam(CO2 ~ s(Year, spar=0.7), data=co2)
sco27<-summary(yearco2_07)
sco27
R2_7 <- (sco27$null.deviance-sco27$deviance)/sco27$null.deviance
R2_7
co2$fits07 <- fitted(yearco2_07)
Plot207 <- ggplot(co2, aes(x=Year, y=CO2)) + geom_point()
Plot207 <- Plot207 + geom_line(aes(x=Year, y=fits07), color='red')
Plot207
#fit Population
yearpop01<- gam(Population ~ s(Year, spar=0.1), data=co2)
spop1<-summary(yearpop01)
spop1
R3_1 <- (spop1$null.deviance-spop1$deviance)/spop1$null.deviance
R3_1
co2$fits001 <- fitted(yearpop01)
Plot301 <- ggplot(co2, aes(x=Year, y=Population)) + geom_point()
Plot301 <- Plot301 + geom_line(aes(x=Year, y=fits001), color='red')
Plot301

yearpop03<- gam(Population ~ s(Year, spar=0.3), data=co2)
spop3<-summary(yearpop03)
spop3
R3_3 <- (spop3$null.deviance-spop3$deviance)/spop3$null.deviance
R3_3
co2$fits003 <- fitted(yearpop03)
Plot303 <- ggplot(co2, aes(x=Year, y=Population)) + geom_point()
Plot303 <- Plot303 + geom_line(aes(x=Year, y=fits003), color='red')
Plot303

yearpop05<- gam(Population ~ s(Year, spar=0.5), data=co2)
spop5<-summary(yearpop05)
spop5
R3_5 <- (spop5$null.deviance-spop5$deviance)/spop5$null.deviance
```

```
R3_5
co2$fits005 <- fitted(yearpop05)
Plot305 <- ggplot(co2, aes(x=Year, y=Population)) + geom_point()
Plot305 <- Plot305 + geom_line(aes(x=Year, y=fits005), color='red')
Plot305

yearpop07<- gam(Population ~ s(Year, spar=0.7), data=co2)
spop7<-summary(yearpop07)
spop7
R3_7 <- (spop7$null.deviance-spop7$deviance)/spop7$null.deviance
R3_7
co2$fits007 <- fitted(yearpop07)
Plot307 <- ggplot(co2, aes(x=Year, y=Population)) + geom_point()
Plot307 <- Plot307 + geom_line(aes(x=Year, y=fits007), color='red')
Plot307
lmfit<-lm(Population ~ Year, data=co2)
summary(lmfit)
```

QUESTION2

```
library(GGally)
gf<-read.csv('Ass4GF_data.csv',header = TRUE)
head(gf)
dim(gf)
gf<-na.omit(gf)
dim(gf)
gf$cow<-factor(gf$cow)
summary(gf)
ggpairs(gf[,4:9])
PlotTC <- ggplot(gf, aes(x=TotDM, y=CO2)) + geom_point()
PlotTC
coplot(CO2 ~TotDM|cow, data=gf, panel = panel.smooth)
PlotTC <- ggplot(gf, aes(x=Pasture, y=CO2)) + geom_point()
PlotTC
coplot(CO2 ~Pasture|cow, data=gf, panel = panel.smooth)
PlotTH <- ggplot(gf, aes(x=TotDM, y=CH4)) + geom_point()
PlotTH
coplot(CH4 ~TotDM|cow, data=gf, panel = panel.smooth)
PlotTH <- ggplot(gf, aes(x=Pasture, y=CH4)) + geom_point()
PlotTH
coplot(CH4 ~Pasture|cow, data=gf, panel = panel.smooth)
plotCCO2<-ggplot(data=gf, aes(x=cow,y=CO2, fill = cow))+geom_boxplot(show.legend =
FALSE)+theme(axis.text.x = element_text(angle=30, hjust=1, vjust=1))
plotCCO2
plotCCH4<-ggplot(data=gf, aes(x=cow,y=CH4, fill = cow))+geom_boxplot(show.legend =
FALSE)+theme(axis.text.x = element_text(angle=30, hjust=1, vjust=1))
plotCCH4
```

```
ggplot(gf, aes(x=day, y=CO2)) + geom_point()
coplot(CO2 ~day|cow, data=gf, panel = panel.smooth)
ggplot(gf, aes(x=day, y=CH4)) + geom_point()
coplot(CH4 ~day|cow, data=gf, panel = panel.smooth)
#check distribution of CO2 and CH4
ggplot(gf, aes(CO2, fill = cut(CO2,100))) + geom_histogram(bins = 20,show.legend = FALSE)
ggplot(gf, aes(CH4, fill = cut(CH4,100))) + geom_histogram(bins = 20,show.legend = FALSE)
#fit gam
gamCO2<- gam(CO2 ~ s(day) + s(Pellets) + s(Pasture) + cow, data=gf, family = gaussian)
sgamCO2<-summary(gamCO2)
sgamCO2
gamCO2_1<- gam(CO2 ~ s(day) + cow + Pellets + Pasture, data=gf, family = gaussian)
sgamCO2_1<-summary(gamCO2_1)
sgamCO2_1
gamCO2_2<- gam(CO2 ~ s(day) + Pasture + cow, data=gf, family = gaussian)
sgamCO2_2<-summary(gamCO2_2)
sgamCO2_2
gamCO2_3<- gam(CO2 ~ s(day) + cow, data=gf, family = gaussian)
sgamCO2_3<-summary(gamCO2_3)
sgamCO2_3
RCO2 <- (sgamCO2_3$null.deviance-sgamCO2_3$deviance)/sgamCO2_3$null.deviance
RCO2
qqnorm(resid(gamCO2_3))
gamCH4<- gam(CH4 ~ s(day) + s(Pellets) + s(Pasture) + cow, data=gf, family = gaussian)
sgamCH4<-summary(gamCH4)
sgamCH4
gamCH4_1<- gam(CH4 ~ s(day) + s(Pasture) + cow, data=gf, family = gaussian)
sgamCH4_1<-summary(gamCH4_1)
sgamCH4_1
RCH4 <- (sgamCH4_1$null.deviance-sgamCH4_1$deviance)/sgamCH4_1$null.deviance
RCH4
qqnorm(resid(gamCH4_1))
#fit lm(polym)
CO2full<-lm(CO2~Pellets + polym(Pasture,degree = 2) + cow + polym(day,degree = 2), gf)
CO2step<-step(CO2full)
summary(CO2step)
anova(CO2step)
CO2step1<-lm(CO2~cow + polym(day,degree = 2), gf)
summary(CO2step1)
anova(CO2step1)
AIC(CO2step1)
qqnorm(resid(CO2step1))
library(DAAG)
cv.lm(gf, CO2step1, m=10)
CH4full<-lm(CH4~Pellets + polym(Pasture,degree=2) + cow + polym(day,degree=2), gf)
CH4step<-step(CH4full)
```

```
summary(CH4step)
anova(CH4step)
AIC(CH4step)
qqnorm(resid(CH4step))
#fit lmm
library(lmerTest)
lmeCO2 <- lmer(CO2 ~ polym(Pasture,degree=2) + polym(day,degree = 2) + Pellets + (1|cow), data = gf)
summary(lmeCO2)
lmeCO2_1 <- lmer(CO2 ~ polym(day,degree = 2) + (1|cow), data = gf)
summary(lmeCO2_1)
AIC(lmeCO2_1)
qqnorm(resid(lmeCO2_1))
lmeCH4 <- lmer(CH4 ~ polym(Pasture,degree = 2) + polym(day,degree = 2) + Pellets + (1|cow), data = gf)
summary(lmeCH4)
lmeCH4_1 <- lmer(CH4 ~ polym(Pasture,degree = 2) + polym(day,degree = 2) + (1|cow), data = gf)
summary(lmeCH4_1)
anova(lmeCH4_1,lmeCH4, test = 'F')
AIC(lmeCH4_1)
qqnorm(resid(lmeCH4_1))
```