Homework 6

ISTA 421/INFO 521

Student - Jung Mee Park

University of Arizona

Instructor - Cristian Roman-Palacios

School of Information, University of Arizona, Tucson, AZ

Objectives

This homework sheet will help reviewing the basic concepts associated with non-linear models. Please review the lectures, suggested readings, and additional resources *before* getting started on the HW.

Additional resources relevant to this HW

- **R Markdown**: Please review the basic R Markdown cheat sheet in case you have any questions regarding formatting the HW: https://www.rstudio.com/wp-content/uploads/2015/02/rmarkdown-cheatsheet.pdf.
- **R**: Please review the basic R` cheat sheet in case you have any questions regarding the programming language: https://www.soa.org/globalassets/assets/Files/Edu/2018/exam-pa-base-r.pdf.
- **RStudio**: Additional cheat sheets written by RStudio to help with specific `R` packages: https://www.rstudio.com/resources/cheatsheets/
- **Datasets**: The following website has access to the relevant datasets from the recommended textbook: https://book.huihoo.com/introduction-to-statistical-learning/data.html

The *Tidyverse*

I encourage students to check out functions from packages included in the `tidyverse` (https://www.tidyverse.org/) which greatly facilitates the productivity of novice coders. However, all instructions will be delivered using `Base R`. The main textbook also uses base `R` only. I will be happy to grade your code regardless whether it uses base `R` or functions in the `tidyverse.` For some steps, other packages (such as `data.table`) are an appropriate alternative. Most if not all the questions in this HW can be answered using the tidyverse. Please check out the "accompanying" book to our main textbook that uses packages from the tidyverse instead of base (https://emilhvitfeldt.github.io/ISLR-tidymodels-labs/index.html). For this chapter, please follow the instructions in the following site: https://emilhvitfeldt.github.io/ISLR-tidymodels-labs/linear-regression.html.

Scores

Please answer the questions from the section that you're enrolled in (labeled as either **421** or **521**). Below is a summary of the total scores associated with this HW (2 points per question).

- **ISTA 421**: 14 points (undergraduate)
- **INFO 521**: 16 points (graduate)

Submission:

Please follow the instructions outlined below to submit your assignment. **This HW is due at the end of the same week that is released (Sunday, 11:59 pm AZ time)**. Please get in touch with the instructor if you're (i) having issues opening the assignment, (2) not understanding the questions or (3) having issues submitting your assignment. Note that late submissions are subject to a penalty (see late work policies in the Syllabus).

-**Homework 6**: Please turn in a **single RMD file (this file) AND a rendered HTML**. Answers to each question should be in the relevant block of code (see below). Re-name your file to **lastname_Hw6.RMD** before submitting. **Make sure that you can correctly render (knit) your submission without errors before turning anything in**. If a given block of code is causing you issues and you didn't get to fix it on time, please use `eval=FALSE` in the relevant block. If you're using additional files to render your **RMD**, please include each of them in your submission.

Time commitment

Please do reach out if you're taking more than ~18h to complete (1) this HW, (2) reading the assigned book chapters, and (3) going over the lectures. I will be happy to provide accommodations if necessary. Do not wait until the last minute to start working on the HW. In most cases, working under pressure will certainly increase the time needed to answer each of these questions. The instructor might not be 100% available on Sundays to troubleshoot with you. Remember that you can sign up for office hours with the instructor 3 times a week.

Looking for help?

First, please go over the relevant readings for this week. Second, if you're still struggling with any of the questions, do some independent research (e.g. stackoverflow is a wonderful resource). **Don't forget that your classmates will also be working on the same questions - reach out for help (check the Discussion forum in D2L for advice from other students)**. Finally, the instructor will be happy to answer any questions during office hours. You can reach out to me by email (cromanpa94@arizona.edu) or simply schedule a 15 minute meeting through Calendly (https://calendly.com/cromanpa/15min)*

Please do not forget that the instructor holds office hours 3 times a week!!

Grading

Please note that grades are NOT exclusively based on your final answers. I will be grading the overall structure and logic of your code. Feel free to use as many lines as you need to answer each of the questions. I also highly recommend and strongly encourage adding comments (`#`) to your code. Comments will certainly improve the reproducibility and readability of your submission. Commenting your code is also good coding practice. Specifically for the course, you'll get better feedback if the instructor is able to understand your code in detail.

Questions

Conceptual

Question 1 (421/521)

Suppose that a curve $\hat{g}$ is computed to smoothly fit a set of `n` points using the following formula:

$$\hat{g} = \arg \min_{g} \sum_{(i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(m)}(x)]^2 dx)$$

where $g^{(m)}$ represents the m^{th} derivative of g (and $g^{(0)} = g$. Provide a description (or sketch) of $\hat{g}$ in each of the following scenarios. **Again, don't forget that m^{th} is related to the derivative of g.**

a. $\lambda = \infty$, $m = 0$.

> ***Answer:*** [BEGIN SOLUTION]. g hat would be 0 if lamda goes to infinity and m=0 because g^0 = 0. When lambda terms to infinity, the 2nd part is more important.

b. $\lambda = \infty$, $m = 1$.

> ***Answer:*** [BEGIN SOLUTION]. g would be 1. This would be the first first derivative of the horizontal line.

c. $\lambda = \infty$, $m = 2$.

> ***Answer:*** [BEGIN SOLUTION]. This will be the 2nd derivative of the horizontal line. This would be a linear line.

d. $\lambda = \infty$, $m = 3$.

> ***Answer:*** [BEGIN SOLUTION]. The 2nd derivative would be quadriatic.

e. $\lambda = 0$, $m = 3$.

> ***Answer:*** [BEGIN SOLUTION]. The equation becomes a standard least squares equation.

Question 2 (421/521)

For (1) step functions, (2) local regression, and (3) GAMs: provide one situation when using that particular model is advantageous. For instance, "GAMs is likely the preferred option when the question I'm interested in addressing involves…". #https://stat.ethz.ch/R-manual/R-devel/library/mgcv/html/smooth.terms.html > ***Answer:*** [BEGIN SOLUTION]. 1) *step functions:* You can use step functions when

you don't want to impose a non-linear structure for the entire model. A good example for using step functions comes when you want to model wage over age. 2) *local regression:* Maybe useful for fitting quadratic functions or sine waves. In R, loess() is used to smooth out the non-linear functions. 3) *GAMs:* generalized additive models are a blend of generalized linear and additive models. The benefit of GAMs is that we can use it for model categorical and numerical variables. So in the case of colleges, we can see if factors like whether the school is private/public can be examined along side numberical variables.

Question 3 (421/521)

Define the following types of smooth functions (one or two sentences per each; following the notation of `mgcv`) and list at least one relevant potential use: `s()`, `te()`, `ti()`.

a. `s()`

> ***Answer:*** [BEGIN SOLUTION]. s is a smooth specifier. x is the covariate. yi = f(xi) + Ei where Ei ~ N(0, sigma^2) gam(y ~ s(x))

b. `te()`

> ***Answer:*** [BEGIN SOLUTION]. te is the tensor product smooth term. With GAM models, te smooths = 1 penalty per marginal basis. te types of smooths are appropriate for location-time interactions. `te(X1,X2)` models `f1(X1)+f2(X2)+f3(X1,X2)`

c. `ti()`

> ***Answer:*** [BEGIN SOLUTION]. Similar to te() but ti() includes more smoothness parameters. ti() models have one penalty matrix per term compared to te() that have two per model.

Question 4 (421/521)

In the context of GAMs, if `te(X1,X2)` models `f1(X1)+f2(X2)+f3(X1,X2)`, what does `ti(X1,X2)` model? Explain.

> ***Answer:*** [BEGIN SOLUTION]. ti(X1) + ti(X2) + ti(X1, X2). The ti() has a penality per term plus one for the interaction.

Question 5 (421/521)

Briefly explain what spatial and temporal autocorrelation are? List one way in which GAMs could help to account for this (e.g. indicate the type of smooth function you would use or potential types of correlation structures).

> ***Answer:*** [BEGIN SOLUTION]. Temporal autocorrelation is to measure observations over time. Spatial autocorrelation extends from temporal correlation where we might be measuring distance. Spatial autocorrelation is used in ecological studies that measure migration patterns and spread. gam(y ~ s(lattitude, longitude) could be a potential equation studying spatial autocorrelation using GAM.

Question 6 (421/521)

Consider two curves $\hat{g}1$ and $\hat{g}2$, defined by

$$\hat{g}1 = \arg\min_{g} \sum_{(i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(3)}(x)]^2 dx)$$

$$\hat{g}2 = \arg\min_{g} \sum_{(i=1}^{n} (y_i - g(x_i))^2 + \lambda \int [g^{(4)}(x)]^2 dx)$$

where $g^{(m)}$ represents the m^{th} derivative of g. Now, answer the following questions (and include a brief explanation):

a. As $\lambda \to \infty$, will $\hat{g}1$ or $\hat{g}2$ have smaller training RSS?

> ***Answer:*** [BEGIN SOLUTION]. g2 will have the smaller RSS using the training dataset because as when the 4th derivative approaches 0, g2 will be more flexible than g1.

b. As $\lambda \to \infty$, will $\hat{g}_1$ or $\hat{g}_2$ have smaller test RSS?

> ***Answer:*** [BEGIN SOLUTION]. The question is similar to part a) in that if the relation is non-linear then the more flexible g2 model could have the smaller RSS than the g1 model using the testing data.

c. For $\lambda = 0$, will $\hat{g}_1$ or $\hat{g}_2$ have smaller training RSS?

> ***Answer:*** [BEGIN SOLUTION]. The penalty is both 0 for both equations, so the RSS would be the same for g1 and g2.

Applied

Please note that some of the questions outlined below make suggestions on potential functions to be used in the answers. Unless indicated otherwise, feel free to use any other function. For those interested in improving their R skills, I would recommend going over the information in the following book for *tidyverse* (https://emilhvitfeldt.github.io/ISLR-tidymodels-labs/linear-regression.html), using *caret* to answer some of the questions (https://topepo.github.io/caret/), or *mikropml* (https://cran.r-project.org/web/packages/mikropml/vignettes/introduction.html)

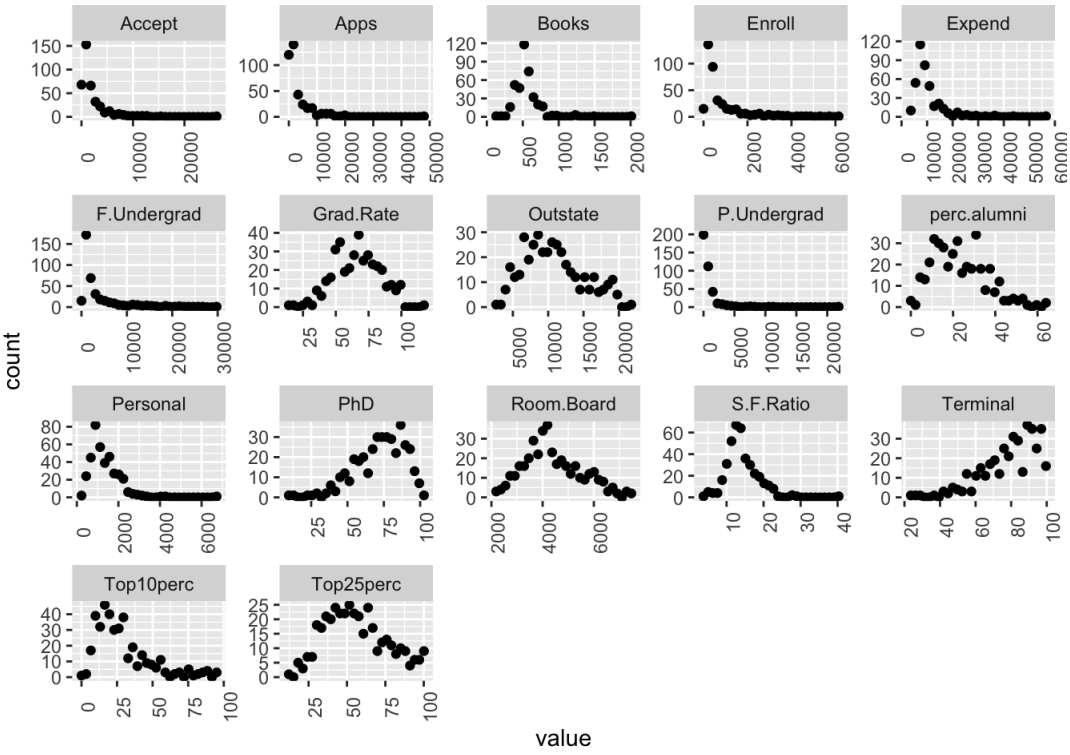
Question 7 (421/521)

```
# load libraries
library(ISLR2)
library(caret)
library(leaps)
library(tidyverse)
library(ggthemes)
library(purrr)
library(knitr)
library(gam)
library(maps)
library(mgcv)
library(broom)
library(modelr)
```

This question relates to the `College` data set.

a. Split the data into a training set and a test set (50:50). Using out-of-state tuition as the response and the other variables as the predictors, perform *forward* stepwise selection on the training set in order to identify a satisfactory model that uses just a subset of the predictors.

```
# BEGIN SOLUTION
College <- read.csv("https://book.huihoo.com/introduction-to-statistical-learning/College.csv")

# replace first row X with college names
College <- College[,-1]

# splitting the dataset to training and testing
train_college <-
  College %>%
  as_tibble(rownames = "rowname") %>%
  sample_frac(size = 0.5)

test_college <-
  College %>%
  as_tibble(rownames = "rowname") %>%
  anti_join(train_college, by = "rowname")

# other visuals
college_gathered <- train_college %>%
  ## remove the variables that are factors
  select(-Private, -rowname) %>%
  gather()

ggplot(college_gathered, aes(x = value))+
  geom_point(stat = "bin", bins=30)+
  facet_wrap(.~key, scales = "free") +
  theme(axis.text.x = element_text(angle = 90))
```




```
# perform a forward selection
forward_selection <- regsubsets(Outstate ~ .,
                                data = select(train_college, -rowname),
                                method = "forward")

summary(forward_selection)
```

```
## Subset selection object
## Call: regsubsets.formula(Outstate ~ ., data = select(train_college,
##     -rowname), method = "forward")
## 17 Variables  (and intercept)
##              Forced in Forced out
## PrivateYes       FALSE      FALSE
## Apps             FALSE      FALSE
## Accept           FALSE      FALSE
## Enroll           FALSE      FALSE
## Top10perc        FALSE      FALSE
## Top25perc        FALSE      FALSE
## F.Undergrad      FALSE      FALSE
## P.Undergrad      FALSE      FALSE
## Room.Board       FALSE      FALSE
## Books            FALSE      FALSE
## Personal         FALSE      FALSE
## PhD              FALSE      FALSE
## Terminal         FALSE      FALSE
## S.F.Ratio        FALSE      FALSE
## perc.alumni      FALSE      FALSE
## Expend           FALSE      FALSE
## Grad.Rate        FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: forward
##          PrivateYes Apps Accept Enroll Top10perc Top25perc F.Undergrad
## 1  ( 1 ) " "        " "  " "    " "    " "        " "        " "
## 2  ( 1 ) "*"        " "  " "    " "    " "        " "        " "
## 3  ( 1 ) "*"        " "  " "    " "    " "        " "        " "
## 4  ( 1 ) "*"        " "  " "    " "    " "        " "        " "
## 5  ( 1 ) "*"        " "  " "    " "    " "        " "        " "
## 6  ( 1 ) "*"        " "  " "    " "    " "        " "        " "
## 7  ( 1 ) "*"        " "  " "    " "    " "        " "        " "
## 8  ( 1 ) "*"        " "  " "    " "    " "        " "        " "
##          P.Undergrad Room.Board Books Personal PhD Terminal S.F.Ratio
## 1  ( 1 ) " "          " "        " "   " "      " " " "      " "
## 2  ( 1 ) " "          " "        " "   " "      " " " "      " "
## 3  ( 1 ) " "          "*"        " "   " "      " " " "      " "
## 4  ( 1 ) " "          "*"        " "   " "      " " " "      " "
## 5  ( 1 ) " "          "*"        " "   " "      "*" " "      " "
## 6  ( 1 ) " "          "*"        " "   " "      "*" " "      " "
## 7  ( 1 ) " "          "*"        " "   "*"      "*" " "      " "
## 8  ( 1 ) " "          "*"        " "   "*"      "*" "*"      " "
##          perc.alumni Expend Grad.Rate
## 1  ( 1 ) " "         "*"    " "
## 2  ( 1 ) " "         "*"    " "
## 3  ( 1 ) " "         "*"    " "
## 4  ( 1 ) "*"         "*"    " "
## 5  ( 1 ) "*"         "*"    " "
## 6  ( 1 ) "*"         "*"    "*"
## 7  ( 1 ) "*"         "*"    "*"
## 8  ( 1 ) "*"         "*"    "*"
```

```
forward_models <-
  tibble(
    metric = c("adjr2", "cp", "bic"),
    best_model = c(
      summary(forward_selection)[["adjr2"]] %>% which.max(),
      summary(forward_selection)[["cp"]] %>% which.min(),
      summary(forward_selection)[["bic"]] %>% which.min()
    )
  )

forward_models
```

```
## # A tibble: 3 × 2
##   metric best_model
##   <chr>       <int>
## 1 adjr2           8
## 2 cp              8
## 3 bic             7
```

```
forward_selection %>% coef(id = 6)
```

```
##   (Intercept)    PrivateYes    Room.Board           PhD   perc.alumni
## -3947.6777825  2992.4649920     0.9233257    43.3233982    54.1924544
##        Expend     Grad.Rate
##     0.2048721    28.3833717
```

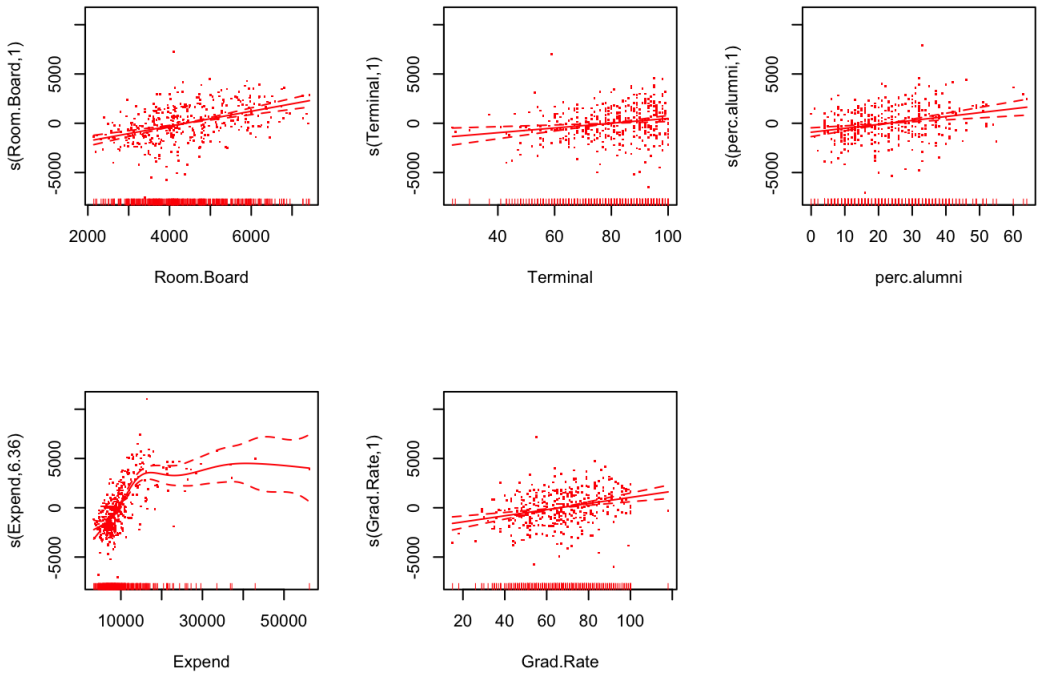
```
# END SOLUTION
```

b. Fit a GAM on the training data, using out-of-state tuition as the response and the features selected in the previous step as the predictors. Plot the results, and briefly explain your findings.

```
# BEGIN SOLUTION
# library(gam)

# fit GAM into training dataset
gam_college <-
  gam(Outstate ~ Private + s(Room.Board) + s(Terminal) + s(perc.alumni) + s(Expend) + s(Grad.Rate), data = t
rain_college)

summary(gam_college)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Outstate ~ Private + s(Room.Board) + s(Terminal) + s(perc.alumni) +
##     s(Expend) + s(Grad.Rate)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8629.5      210.3   41.03   <2e-16 ***
## PrivateYes    2702.4      257.5   10.49   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                   edf Ref.df      F  p-value
## s(Room.Board)   1.000  1.000 50.425  < 2e-16 ***
## s(Terminal)     1.000  1.000  9.698  0.00199 **
## s(perc.alumni) 1.000  1.000 16.412 6.22e-05 ***
## s(Expend)       6.359  7.501 29.738  < 2e-16 ***
## s(Grad.Rate)    1.000  1.000 22.760 2.84e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.821   Deviance explained = 82.6%
## GCV = 3.0837e+06  Scale est. = 2.9855e+06  n = 388
```

```
par(mfrow = c(2,3))
plot(gam_college, se = T, residuals = T, col = "red")

# END SOLUTION
```



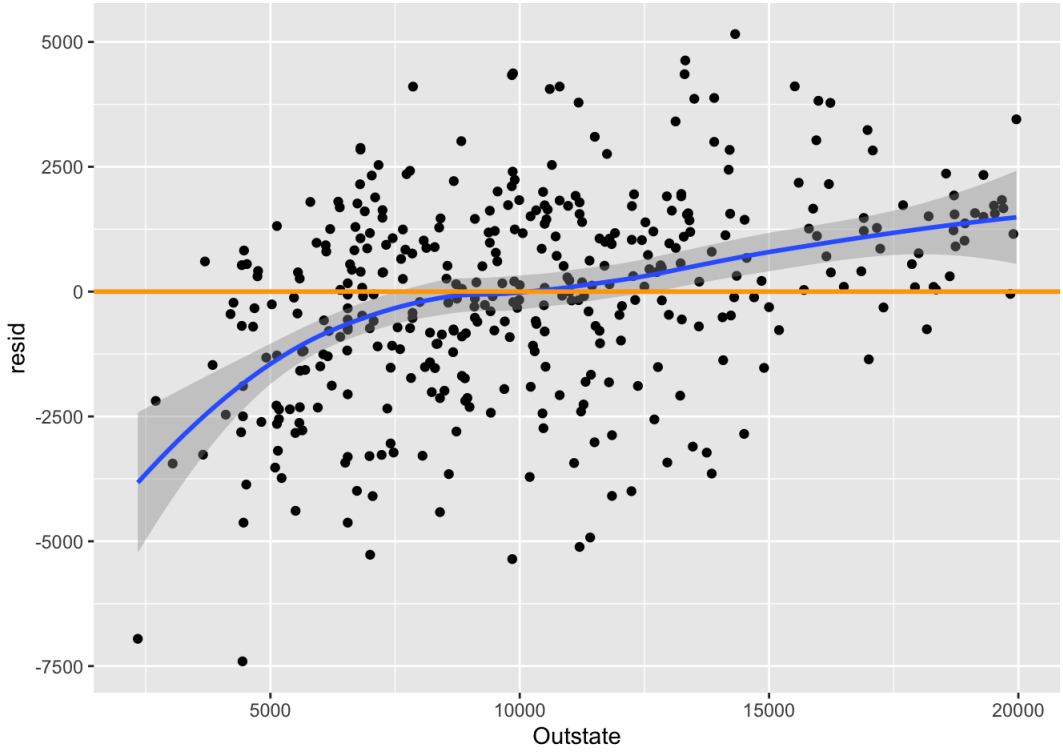
c. For which variables, if any, is there evidence of a non-linear relationship with the response?

```
# BEGIN SOLUTION
# compute RSS
(test_college$Outstate - predict(gam_college, newdata = test_college))^2 %>%
  sum()
```

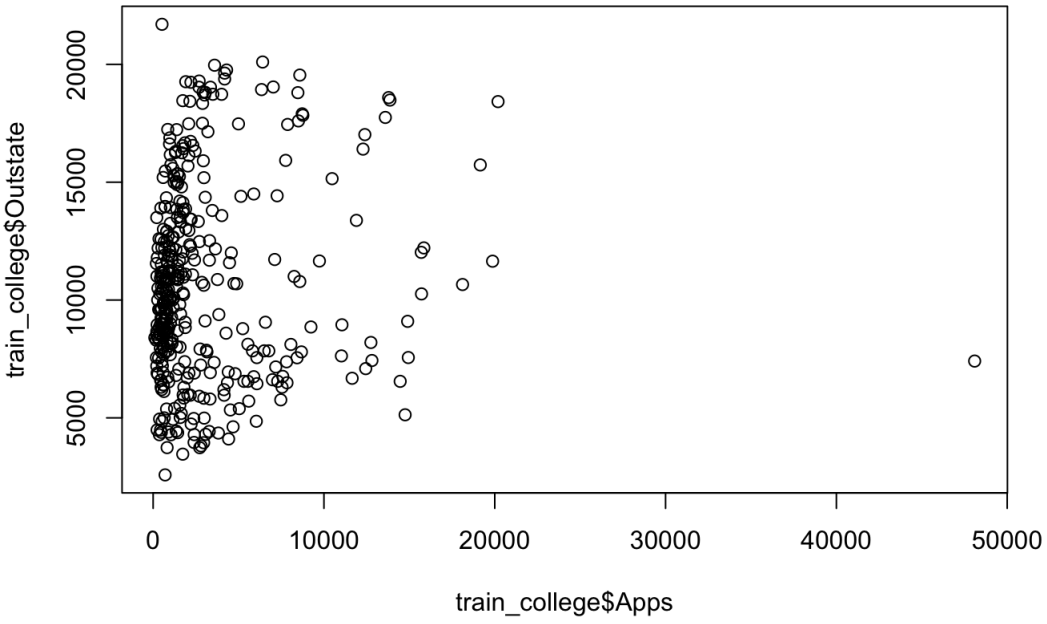
```
## [1] 1558366340
```

```
# plot residuals
test_college %>%
  add_predictions(gam_college) %>%
  mutate(resid = Outstate - pred) %>%
  ggplot(aes(Outstate, resid)) +
  geom_point() +
  geom_smooth() +
  geom_hline(yintercept = 0, color = "orange", size = 1)
```

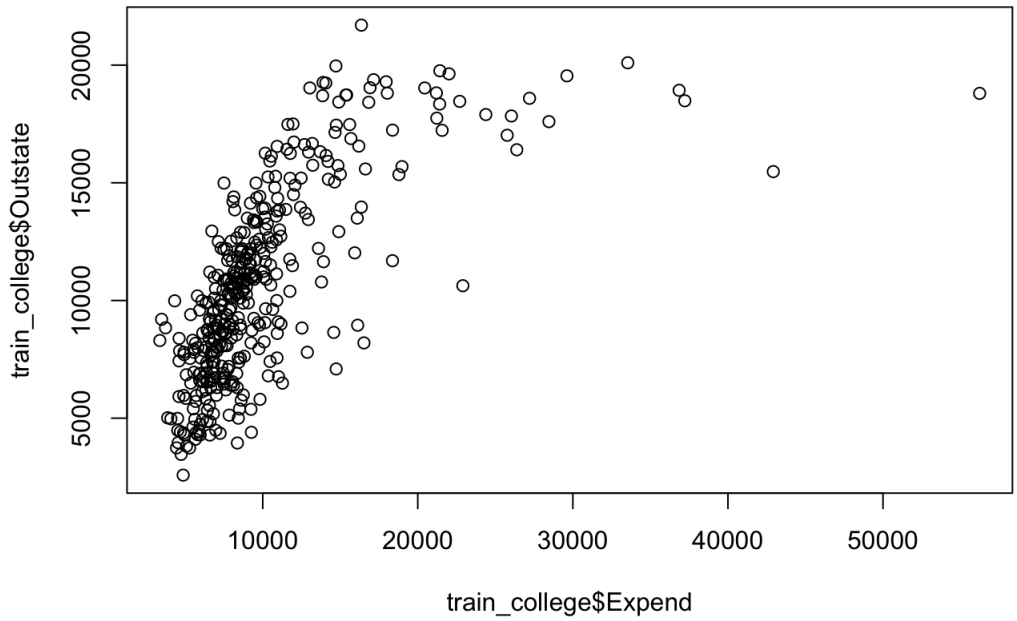
```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```



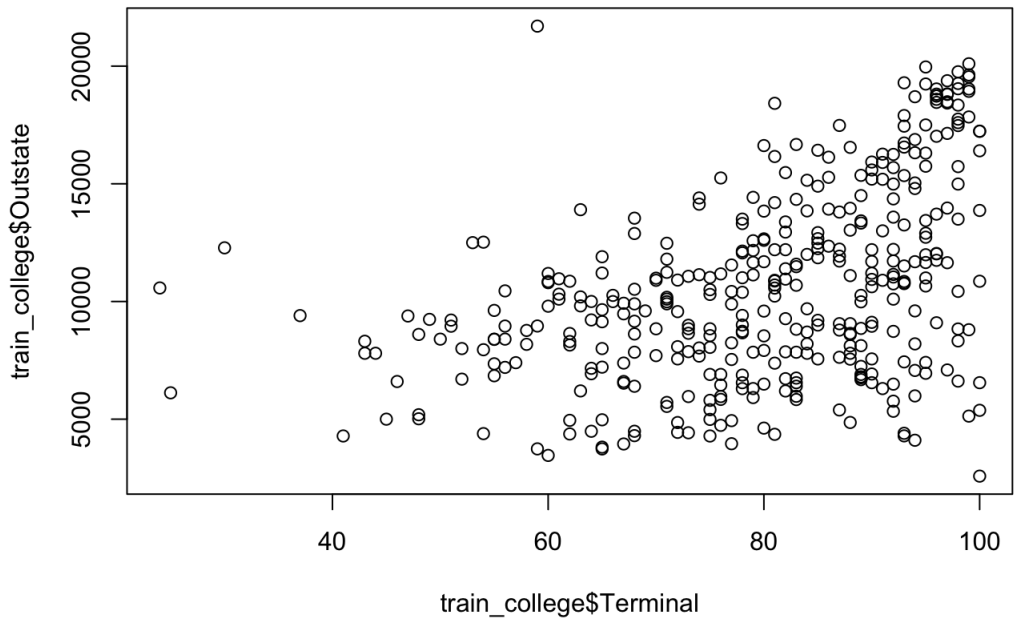

```
# keyvar <- data.matrix(train_college[, c("Apps", "perc.alumni", "Personal","PhD","S.F.Ratio", "Top25perc", "Outstate", "Room.Board", "Expend", "Grad.Rate")])
plot(x=train_college$Apps, y=train_college$Outstate)
```




```
plot(x=train_college$Expend, y=train_college$Outstate)
```




```
plot(x=train_college$Terminal, y=train_college$Outstate)
```




```
# END SOLUTION
```

From part B, I saw that expenditure had a strong non-linear relationship.

Question 8 (521)

Answer the questions below using the following dataset:

```
# Change eval=FALSE to TRUE once HW.RData is in the same folder as this script.
load("HW.RData")
data <- data.frame(data)
map <- data.frame(map)
```

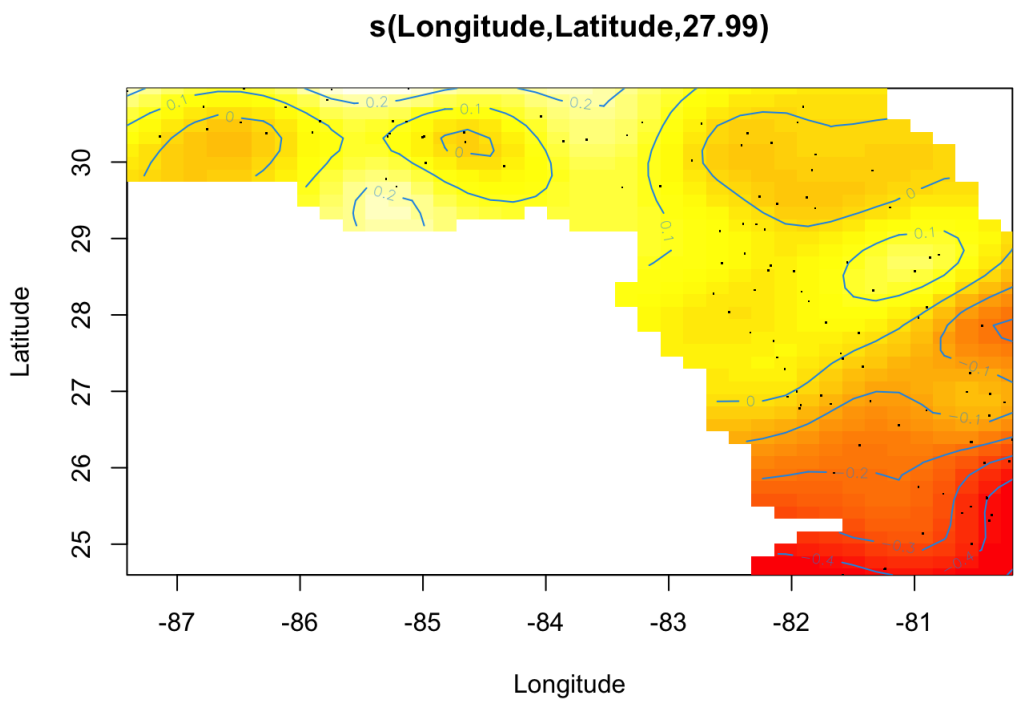
a. Only by inspecting the plots shown below: Is there any temporal or spatial trend in the response variable `richness`?

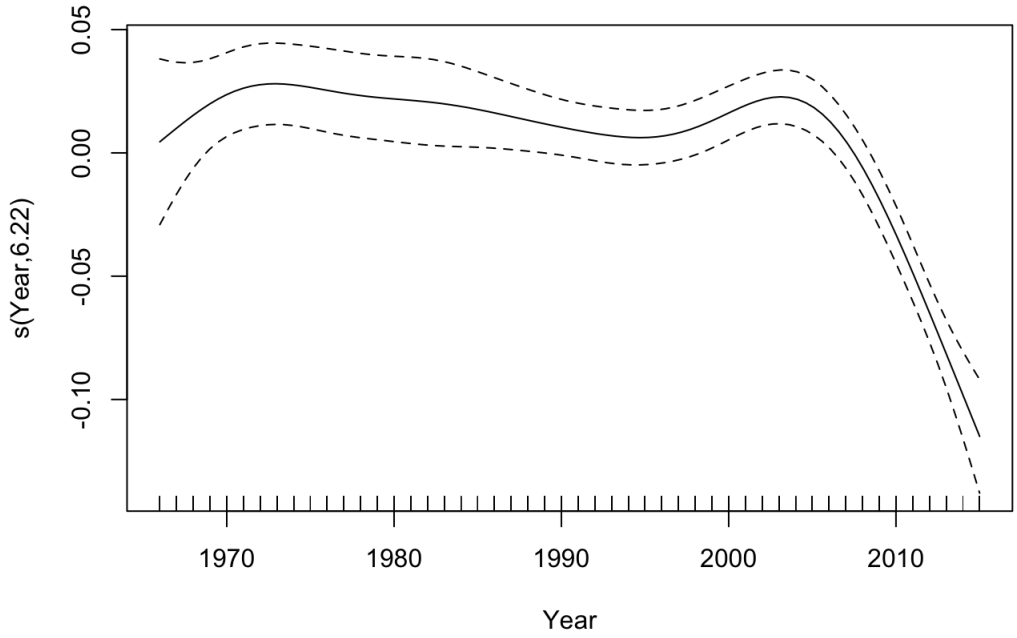
```
# library(maps) #Install if needed
# library(tidyverse) #Iuploaded earlier
head(data)
ggplot(aes(x=Year,y=richness),data=data)+geom_point()+
  theme_bw()
ggplot(aes(x=Longitude,y=Latitude,col=richness), data = data) +
  geom_polygon(data=map,fill=NA, col="black")+
  geom_point()+theme_bw()
```

Richness is more dispersed over the years. Spatially, there is less richness below the 27th latitude and between -82 to -80 longitude.

b. Now, fit a Poisson GAM with the following structure `E(richness) = exp(f(year))*exp(g(location))`. Note that location is actually two features: `Longitude` and `Latitude.` Name this model `M1.` Plot the model (`plot(M1, scheme=2)`), get the `summary` of `M1` , and answer: Is there evidence for effects of location and year?

```
# BEGIN SOLUTION
# Poisson GAM
# E(richness) = exp(f(year))*exp(g(location))
# MGCV gam, gam(richness~ s(Longitude, Latitude) + s(Year), data=data, family=poisson, method = REML)

M1 <- mgcv::gam(richness ~ s(Longitude, Latitude) + s(Year), data = data, family = poisson(), method = "REML
")

# plot
plot(M1, scheme = 2)
```






```
# END SOLUTION
```

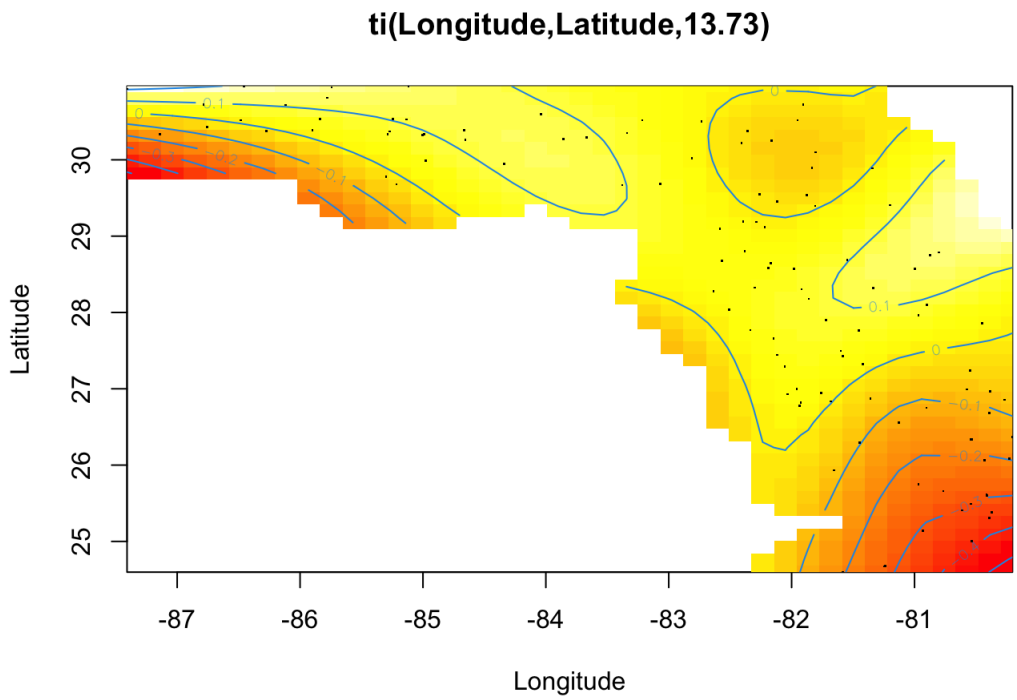
Yes, there is evidence of temporal and spatial affects on the outcome. The spread occurs after 2005 and below 26 latitude.

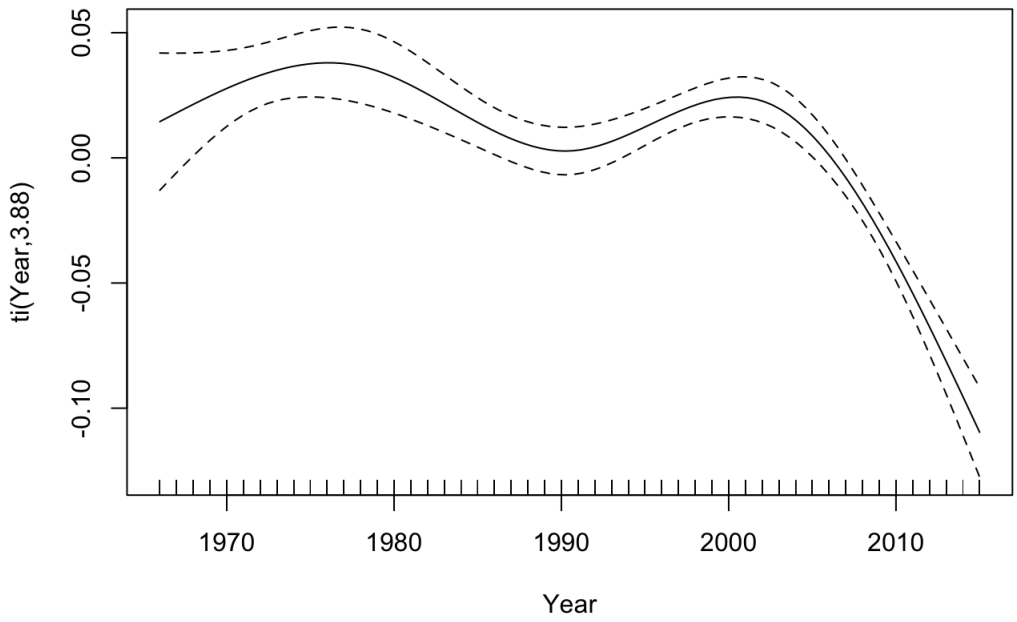
c. Now, fit another Poisson GAM model with the same structure as in `M1` but using `ti()` to account for the interactions between `longitude`, `latitude`, and `year`. Ideally, use `d=c(2,1)` as argument to `ti()`.

```
# BEGIN SOLUTION
M2 <- gam(richness ~ ti(Longitude,Latitude) + ti(Year), data = data, family = poisson(), method = "REML")
# M2 <- gam(richness ~ s(Longitude, Latitude) + s(Year, d=c(2,1)), data = data, family = poisson(), method =
"REML")

summary(M2)
```

```
##
## Family: poisson
## Link function: log
##
## Formula:
## richness ~ ti(Longitude, Latitude) + ti(Year)
##
## Parametric coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) 3.835761   0.005567     689   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##                           edf Ref.df Chi.sq p-value
## ti(Longitude,Latitude) 13.73 14.683 1436.5  <2e-16 ***
## ti(Year)                3.88  3.991  158.6  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.323   Deviance explained = 33.6%
## -REML = 9739.9  Scale est. = 1         n = 2824
```

```
plot(M2, scheme = 2)
```






```
# END SOLUTION
```

d. Use an ANOVA to compare `M1` and `M2.` Which model should we select based on this test?

```
# BEGIN SOLUTION
anova(M1, M2)
```

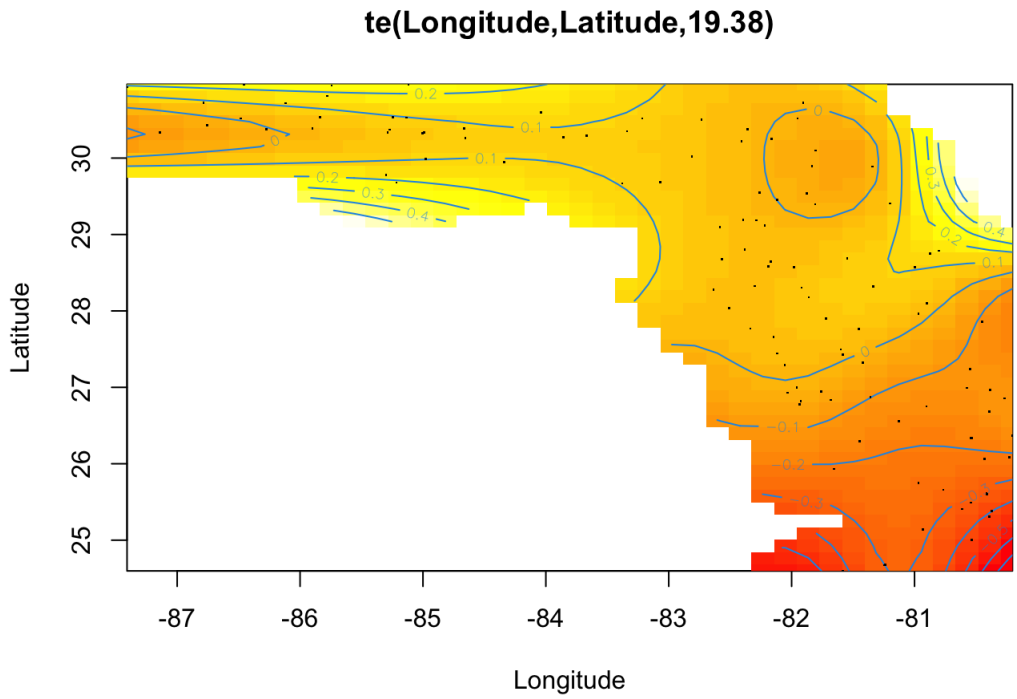
```
## Analysis of Deviance Table
##
## Model 1: richness ~ s(Longitude, Latitude) + s(Year)
## Model 2: richness ~ ti(Longitude, Latitude) + ti(Year)
##   Resid. Df Resid. Dev      Df Deviance
## 1    2786.2       2747
## 2    2803.3       3346 -17.168  -599.02
```

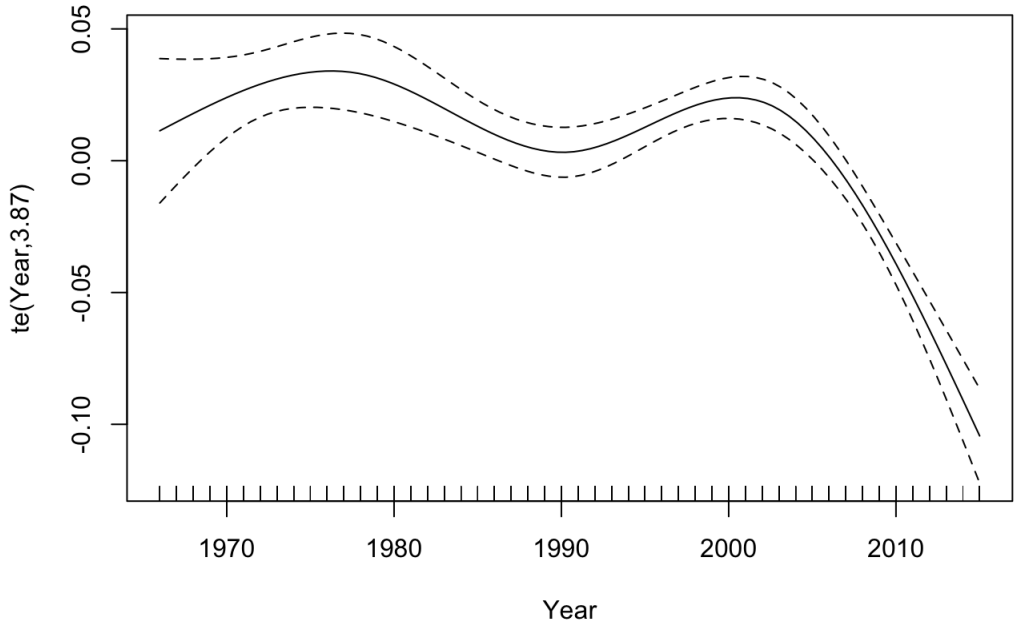
```
# END SOLUTION
```

I would select M1 because of the smaller residual difference.

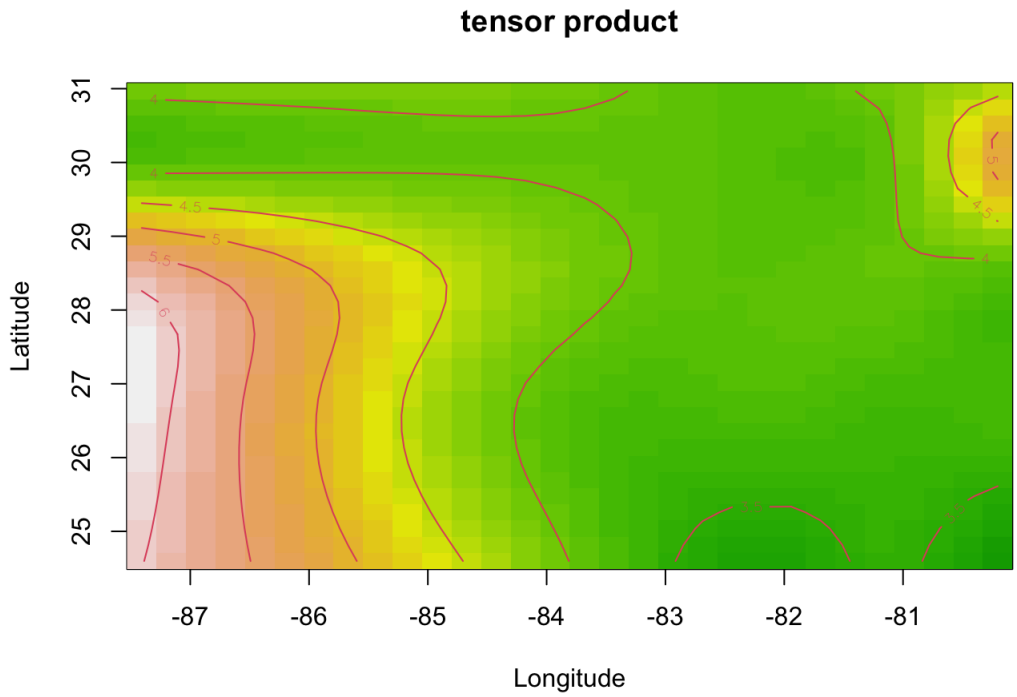
e. Finally, can you come up with a simple approach to examine which of the features (`latitude` , `longitude` , or `year`) have a larger impact on the response (`richness`) using GAMs?

```
# BEGIN SOLUTION
M3 <- gam(richness ~ te(Longitude, Latitude) + te(Year), data=data, family = poisson(), method="REML")
plot(M3, scheme = 2)
```






```
vis.gam(M3, plot.type = "contour", color = "terrain", main = "tensor product")
```




```
anova(M1, M2, M3)
```

```
## Analysis of Deviance Table
##
## Model 1: richness ~ s(Longitude, Latitude) + s(Year)
## Model 2: richness ~ ti(Longitude, Latitude) + ti(Year)
## Model 3: richness ~ te(Longitude, Latitude) + te(Year)
##   Resid. Df Resid. Dev       Df Deviance
## 1    2786.2     2747.0
## 2    2803.3     3346.0 -17.1681  -599.02
## 3    2798.1     2979.5   5.2381   366.46
```

```
AIC(M1, M2, M3)
```

```
##          df      AIC
## M1 35.74125 18842.74
## M2 19.59453 19409.47
## M3 25.06911 19053.96
```

```
# END SOLUTION
```

Processing math: 100%

ce that was one we had not tried yet. But M1 might be the best model due to the low AIC.