Nonlinear models

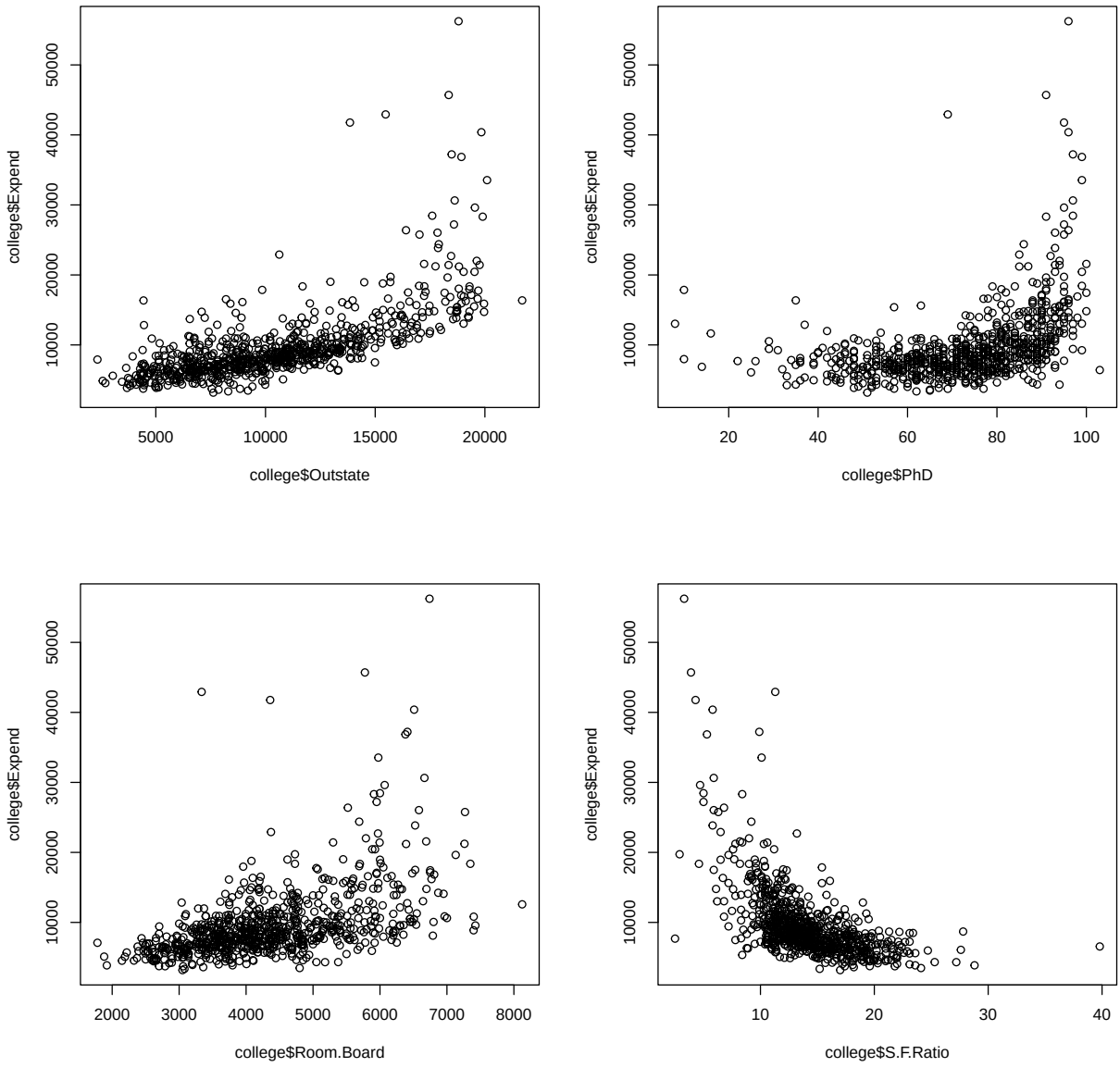
Justin Lo

2023-08-18

```
library(ISLR)
library(gam)
library(splines)
library(randomForest)
library(tidyverse)
```

In this project, I am using the college dataset in ISLR package First, to load the dataset and do basic visualisation and to spot non-linear relationship visually

```
college<-College
par(mfrow=c(2,2))
plot(college$Outstate, college$Expend)
plot(college$PhD, college$Expend)
plot(college$Room.Board, college$Expend)
plot(college$S.F.Ratio, college$Expend)
```



I will continue this project with 'expend' as the response and 'PhD' as the predictor, here is the first non-linear regression model. It is a quadratic model.

```
regression_model<- college%>%
  lm(Expend ~ poly(PhD, 2), data = .)
summary(regression_model)
```

```
##
## Call:
## lm(formula = Expend ~ poly(PhD, 2), data = .)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12750  -2263   -357   1309  40415
```

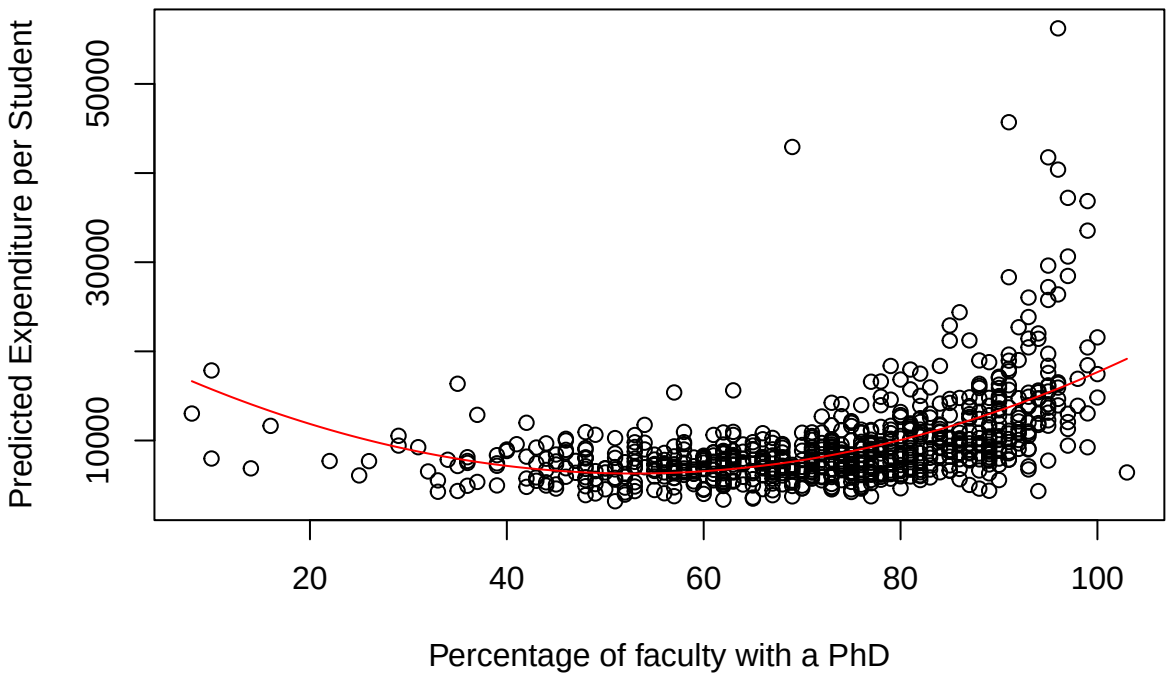
```
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)     9660.2      154.4   62.55   <2e-16 ***
## poly(PhD, 2)1  62950.2     4304.9   14.62   <2e-16 ***
## poly(PhD, 2)2  53405.8     4304.9   12.41   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4305 on 774 degrees of freedom
## Multiple R-squared:  0.3221, Adjusted R-squared:  0.3203
## F-statistic: 183.9 on 2 and 774 DF,  p-value: < 2.2e-16
```

Now, I use the regression model to describe the association between the two variables, and then to calculate fitted values across the range of the 'PhD' variable. By recreating the plot between Expend and PhD and add a line representing these fitted values to illustrate the estimated relationship.

```
attach(college)
range(PhD)
```

```
## [1]   8 103
```

```
fitted_vals_quadratic <- predict(regression_model, newdata = data.frame(PhD= 8:103))
#It predicts the values for a new data frame with the PhD variable ranging from 8 to 103.
plot(PhD, Expend,
     xlab = "Percentage of faculty with a PhD",
     ylab = "Predicted Expenditure per Student")
lines(8:103, fitted_vals_quadratic, col = "red")
```



The graph suggests that predicted expenditure per student decreases when moving from low to moderate percentages of faculty with PhDs, and then increases when moving from moderate to high percentages of faculty with PhDs.

To further improve and re-estimate the model, this time including a cubic polynomial.

```
regression_model_2<- lm(Expend ~ poly(PhD,3), data=college)
summary(regression_model_2)
```

```
##
## Call:
## lm(formula = Expend ~ poly(PhD, 3), data = college)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -15884  -2266   -373   1330  39272
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)       9660        152  63.544  < 2e-16 ***
## poly(PhD, 3)1    62950       4238  14.855  < 2e-16 ***
## poly(PhD, 3)2    53406       4238  12.603  < 2e-16 ***
## poly(PhD, 3)3    21518       4238   5.078 4.79e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4238 on 773 degrees of freedom
```

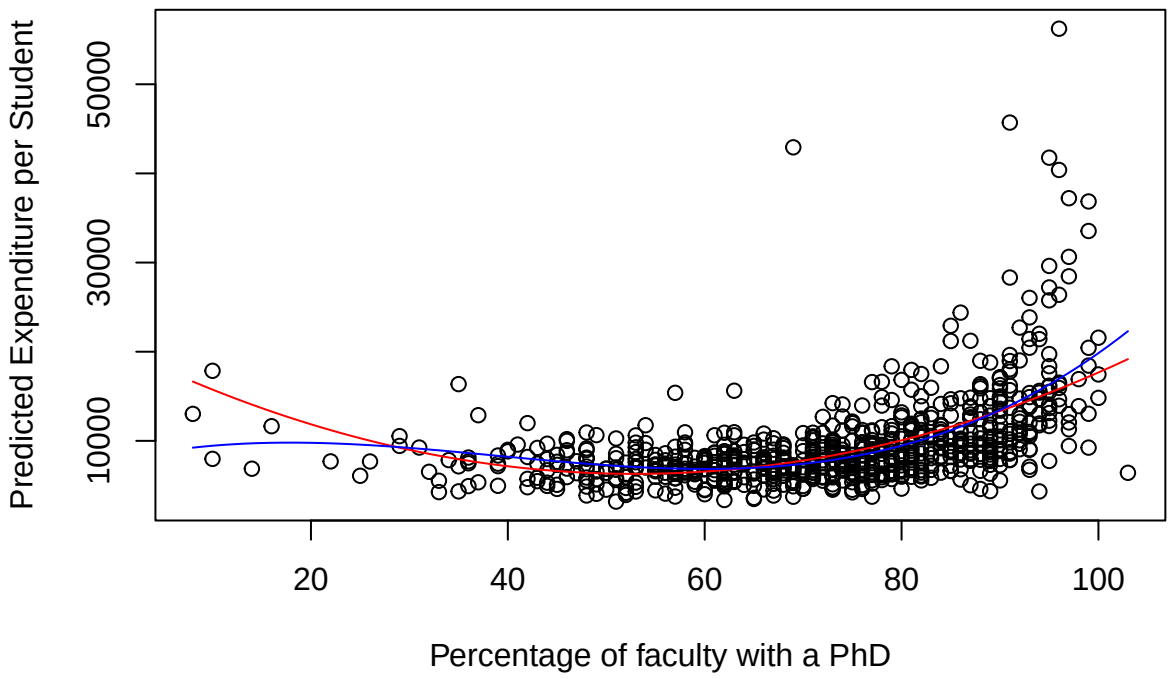
```
## Multiple R-squared:  0.344,  Adjusted R-squared:  0.3414
## F-statistic: 135.1 on 3 and 773 DF,  p-value: < 2.2e-16
```

```
fitted_vals_cubic<- predict(regression_model_2, newdata = data.frame(PhD=8:103))

fitted_vals_quadratic <- predict(regression_model, newdata = data.frame(PhD= 8:103))
attach(college)
```

```
## The following objects are masked from college (pos = 3):
##
##     Accept, Apps, Books, Enroll, Expend, F.Undergrad, Grad.Rate,
##     Outstate, P.Undergrad, perc.alumni, Personal, PhD, Private,
##     Room.Board, S.F.Ratio, Terminal, Top10perc, Top25perc
```

```
plot(PhD, Expend,
     xlab = "Percentage of faculty with a PhD",
     ylab = "Predicted Expenditure per Student")
lines(8:103, fitted_vals_quadratic, col = "red")
lines(8:103, fitted_vals_cubic, col = 'blue')
```

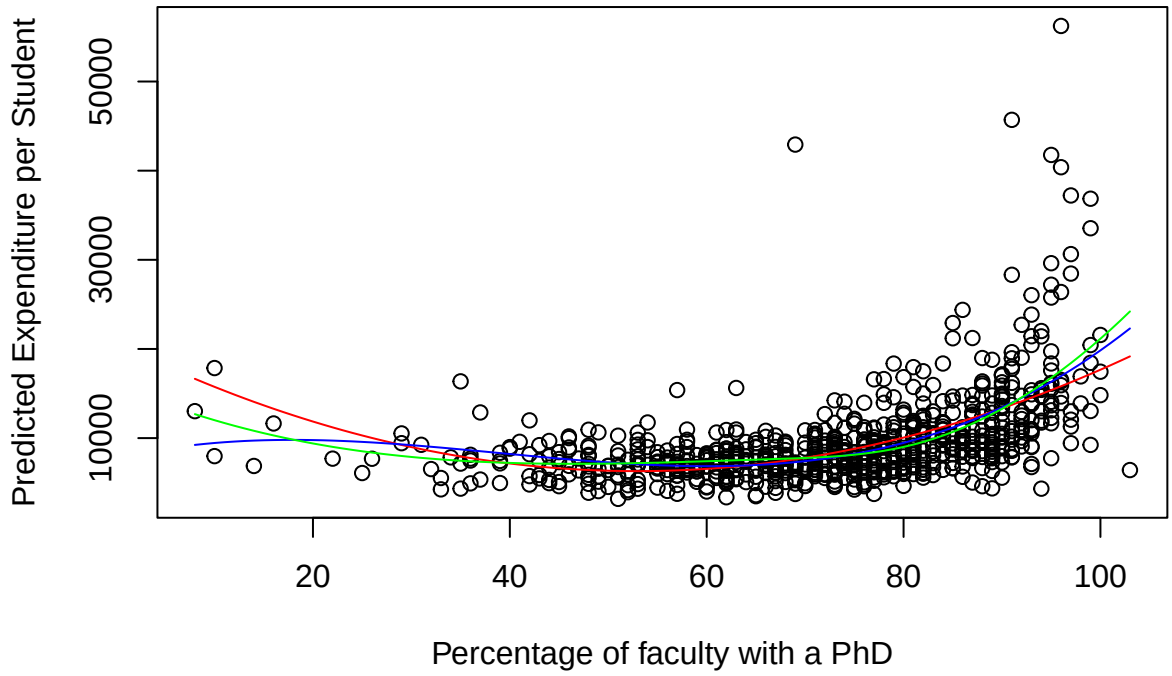


Now, we re-estimate with a cubic spline model

```
phd_mod_spline <- lm(Expend ~ bs(PhD, df = 5, degree = 3), data = College)

fitted_vals_spline <- predict(phd_mod_spline, newdata = data.frame(PhD = 8:103))
```

```
plot(College$PhD, College$Expend,
     xlab = "Percentage of faculty with a PhD",
     ylab = "Predicted Expenditure per Student")
lines(8:103, fitted_vals_quadratic, col = "red")
lines(8:103, fitted_vals_cubic, col = "blue")
lines(8:103, fitted_vals_spline, col = "green")
```

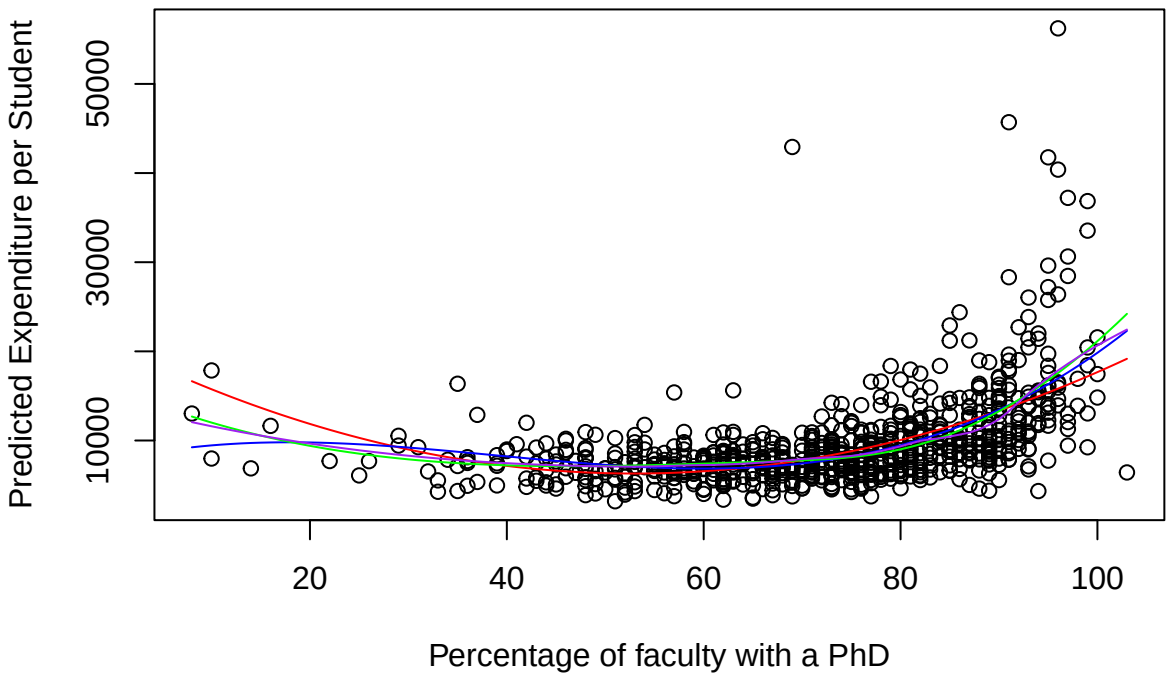


Now, with a loess model

```
phd_mod_loess <- loess(Expend ~ PhD, data = College, span = .4)

fitted_vals_loess <- predict(phd_mod_loess, newdata = data.frame(PhD = 8:103))

plot(College$PhD, College$Expend,
     xlab = "Percentage of faculty with a PhD",
     ylab = "Predicted Expenditure per Student")
lines(8:103, fitted_vals_quadratic, col = "red")
lines(8:103, fitted_vals_cubic, col = "blue")
lines(8:103, fitted_vals_spline, col = "green")
lines(8:103, fitted_vals_loess, col = "purple")
```


Predicted Expenditure per Student
10000
30000
50000
20
40
60
80
100
Percentage of faculty with a PhD