p8106_hw2_yg2625

Yue Gu

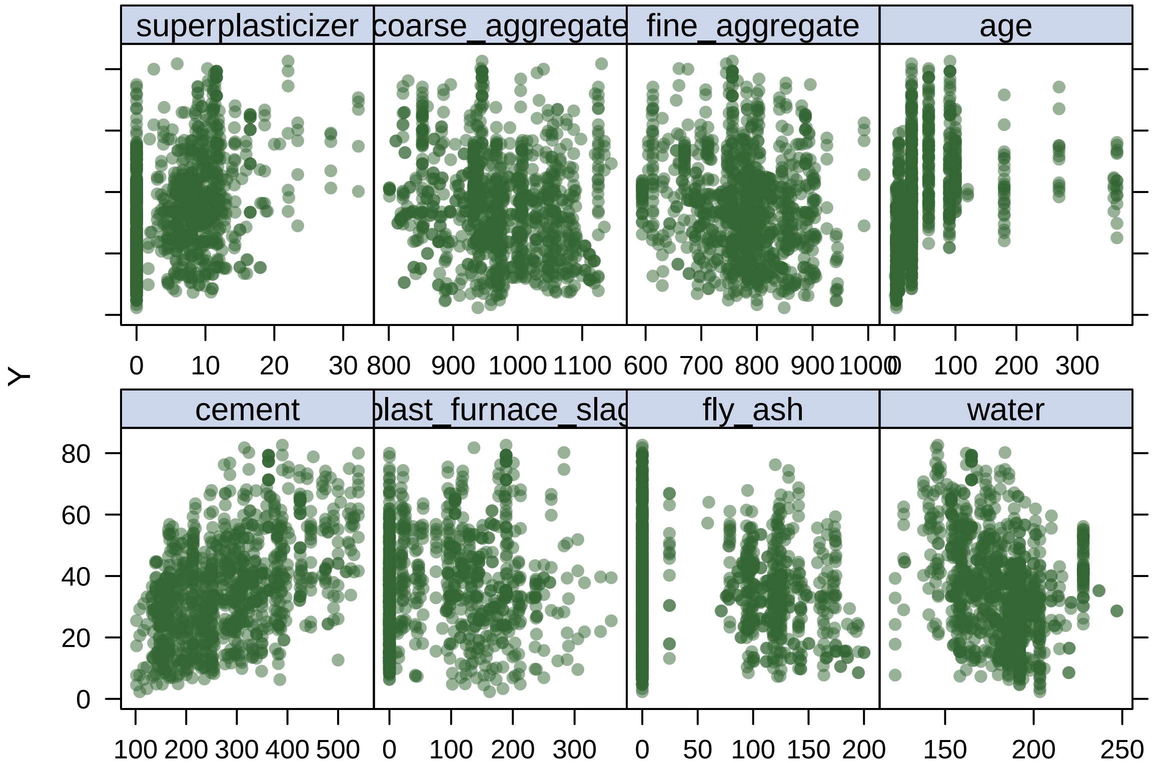
March 20, 2019

Load Data

```
concrete = read.csv("./data/concrete.csv") %>%
  janitor::clean_names()

# matrix of predictors
x <- model.matrix(compressive_strength~.,concrete)[,-1]
# vector of response
y <- concrete$compressive_strength
```

(a) Create scatter plots of response vs. predictors using the function featurePlot().

```
theme1 = trellis.par.get()
theme1$plot.symbol$col = rgb(.2, .4, .2, .5)
theme1$plot.symbol$pch = 16
theme1$plot.line$col = rgb(.8, .1, .1, 1)
theme1$plot.line$lwd = 2
theme1$strip.background$col = rgb(.0, .2, .6, .2)
trellis.par.set(theme1)
featurePlot(x, y, plot = "scatter", labels = c("","Y"),
            type = c("p"), layout = c(4, 2))
```

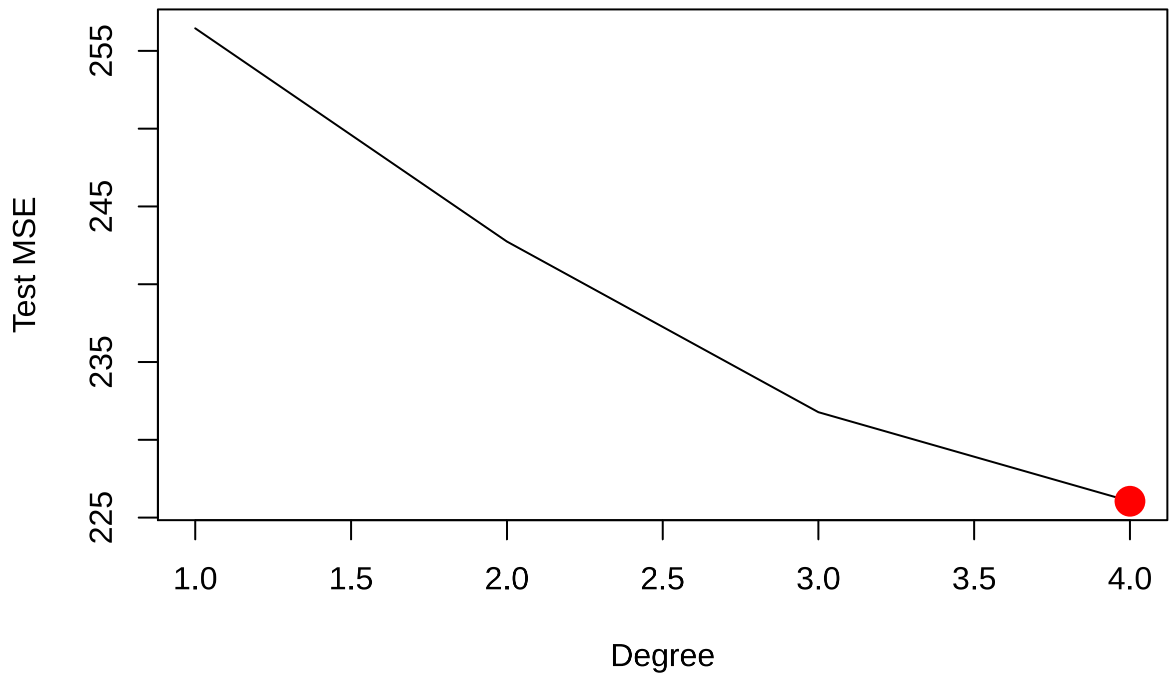


Based on the plots, we could observe that there is no linear relationship between the predictors and compressive strength except for cement.

(b) Perform polynomial regression to predict compressive strength using water as the predictor. For 1<=d<=4, use cross-validation to select the optimal degree d for the polynomial. What degree was chosen, and how does this compare to the results of hypothesis testing using ANOVA? Make a plot of different polynomial fits to the data.

CV selection of optimal degree d

```
set.seed(2)
delta = rep(NA, 4)
for (i in 1:4) {
  fit = glm(compressive_strength~poly(water,i), data = concrete)
  delta[i] = cv.glm(concrete, fit, K = 10)$delta[1]
}
plot(1:4, delta, xlab = "Degree", ylab = "Test MSE", type = "l")
points(which.min(delta), delta[which.min(delta)], col = "red", pch = 19, cex = 2)
```



Based on the CV plot, we choose degree of 4 for the smallest test MSE for $1 \leq d \leq 4$.

Hypothesis testing using ANOVA

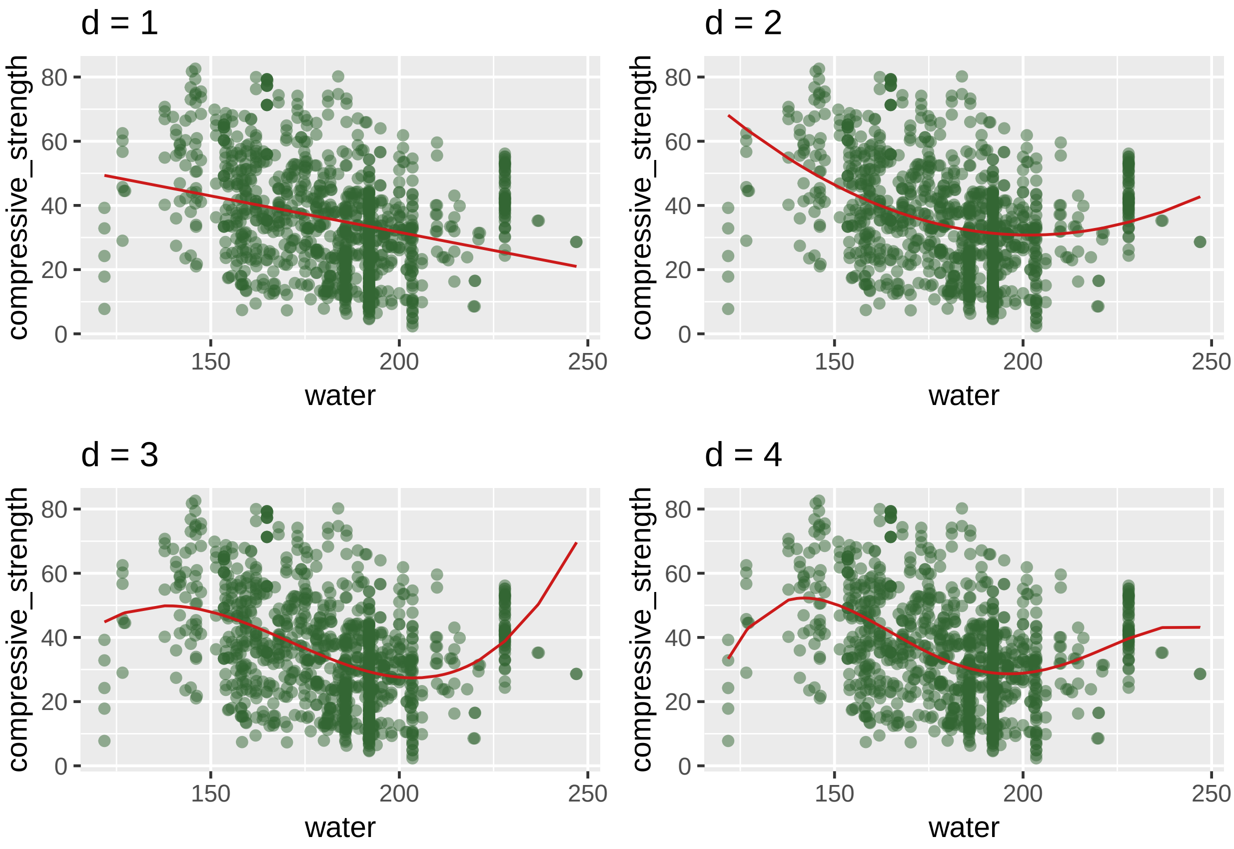
```
fit1 <- lm(compressive_strength~water, data = concrete)
fit2 <- lm(compressive_strength~poly(water,2), data = concrete)
fit3 <- lm(compressive_strength~poly(water,3), data = concrete)
fit4 <- lm(compressive_strength~poly(water,4), data = concrete)
anova(fit1, fit2, fit3, fit4)
```

```
## Analysis of Variance Table
##
## Model 1: compressive_strength ~ water
## Model 2: compressive_strength ~ poly(water, 2)
## Model 3: compressive_strength ~ poly(water, 3)
## Model 4: compressive_strength ~ poly(water, 4)
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1   1028 263085
## 2   1027 247712  1   15372.8 68.140 4.652e-16 ***
## 3   1026 235538  1   12174.0 53.962 4.166e-13 ***
## 4   1025 231246  1    4291.5 19.022 1.423e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Based on the result, model 4 with degree d = 4 provides the smallest MSE and it's the optimal model, which matches our choice from CV shown above.

Creating plots for different polynomial fits

```
pred1 = predict(fit1)
pred2 = predict(fit2)
pred3 = predict(fit3)
pred4 = predict(fit4)

p = ggplot(concrete, aes(x = water, y = compressive_strength)) +
  geom_point(color = rgb(.2, .4, .2, .5))
p1 = p +
  geom_line(aes(x = water, y = pred1), concrete, color = rgb(.8, .1, .1, 1)) +
  ggtitle("d = 1")
p2 = p +
  geom_line(aes(x = water, y = pred2), concrete, color = rgb(.8, .1, .1, 1)) +
  ggtitle("d = 2")
p3 = p +
  geom_line(aes(x = water, y = pred3), concrete, color = rgb(.8, .1, .1, 1)) +
  ggtitle("d = 3")
p4 = p +
  geom_line(aes(x = water, y = pred4), concrete, color = rgb(.8, .1, .1, 1)) +
  ggtitle("d = 4")

p1 + p2 + p3 + p4
```



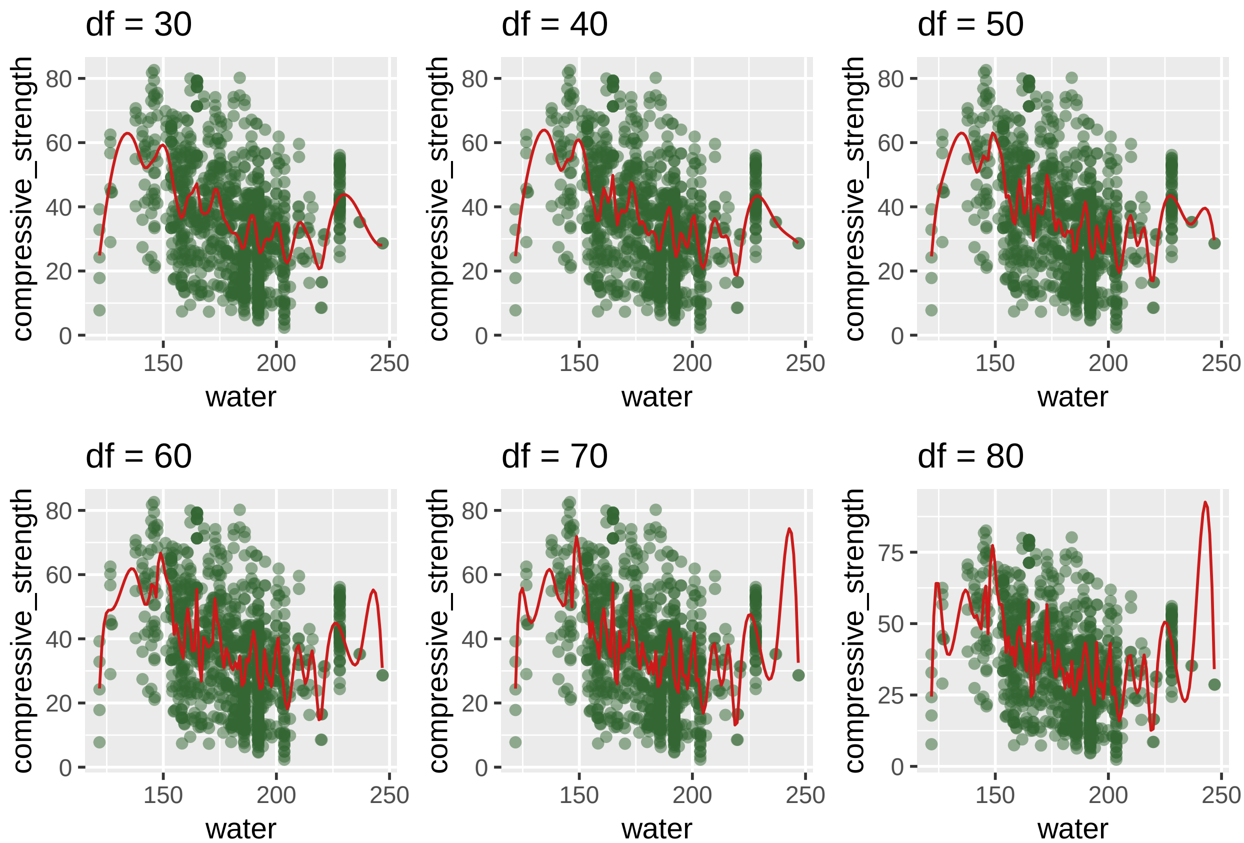
Based on the plots, d = 4 provides the optimal model fit for its closest trend for data points.

(c) Fit a smoothing spline using water as the predictor for a range of degrees of freedom, as well as the degree of freedom obtained by generalized cross-validation, and plot the resulting fits. Describe the results obtained.

fit smoothing spline using water for a range of df

```
waterlims = range(concrete$water)
water.grid = seq(from = waterlims[1],to = waterlims[2])


fit.ss1 = smooth.spline(concrete$water, concrete$compressive_strength,df = 30)
fit.ss2 = smooth.spline(concrete$water, concrete$compressive_strength,df = 40)
fit.ss3 = smooth.spline(concrete$water, concrete$compressive_strength,df = 50)
fit.ss4 = smooth.spline(concrete$water, concrete$compressive_strength,df = 60)
fit.ss5 = smooth.spline(concrete$water, concrete$compressive_strength,df = 70)
fit.ss6 = smooth.spline(concrete$water, concrete$compressive_strength,df = 80)
fit.ss7 = smooth.spline(concrete$water, concrete$compressive_strength,df = 90)
fit.ss8 = smooth.spline(concrete$water, concrete$compressive_strength,df = 100)
fit.ss9 = smooth.spline(concrete$water, concrete$compressive_strength,df = 110)

pred1 = as.data.frame(predict(fit.ss1, x = water.grid))
pred2 = as.data.frame(predict(fit.ss2, x = water.grid))
pred3 = as.data.frame(predict(fit.ss3, x = water.grid))
pred4 = as.data.frame(predict(fit.ss4, x = water.grid))
pred5 = as.data.frame(predict(fit.ss5, x = water.grid))
pred6 = as.data.frame(predict(fit.ss6, x = water.grid))
pred7 = as.data.frame(predict(fit.ss7, x = water.grid))
pred8 = as.data.frame(predict(fit.ss8, x = water.grid))
pred9 = as.data.frame(predict(fit.ss9, x = water.grid))

q1 = p +
  geom_line(aes(x = x, y = y), data = pred1,
            color = rgb(.8, .1, .1, 1)) +
  ggtitle("df = 30")
q2 = p +
  geom_line(aes(x = x, y = y), data = pred2,
            color = rgb(.8, .1, .1, 1)) +
  ggtitle("df = 40")
q3 = p +
  geom_line(aes(x = x, y = y), data = pred3,
            color = rgb(.8, .1, .1, 1)) +
  ggtitle("df = 50")
q4 = p +
  geom_line(aes(x = x, y = y), data = pred4,
            color = rgb(.8, .1, .1, 1)) +
  ggtitle("df = 60")
q5 = p +
  geom_line(aes(x = x, y = y), data = pred5,
            color = rgb(.8, .1, .1, 1)) +
  ggtitle("df = 70")
q6 = p +
```

```r
  geom_line(aes(x = x, y = y), data = pred6,
            color = rgb(.8, .1, .1, 1)) +
  ggtitle("df = 80")

q1 + q2 + q3 + q4 + q5 + q6
```



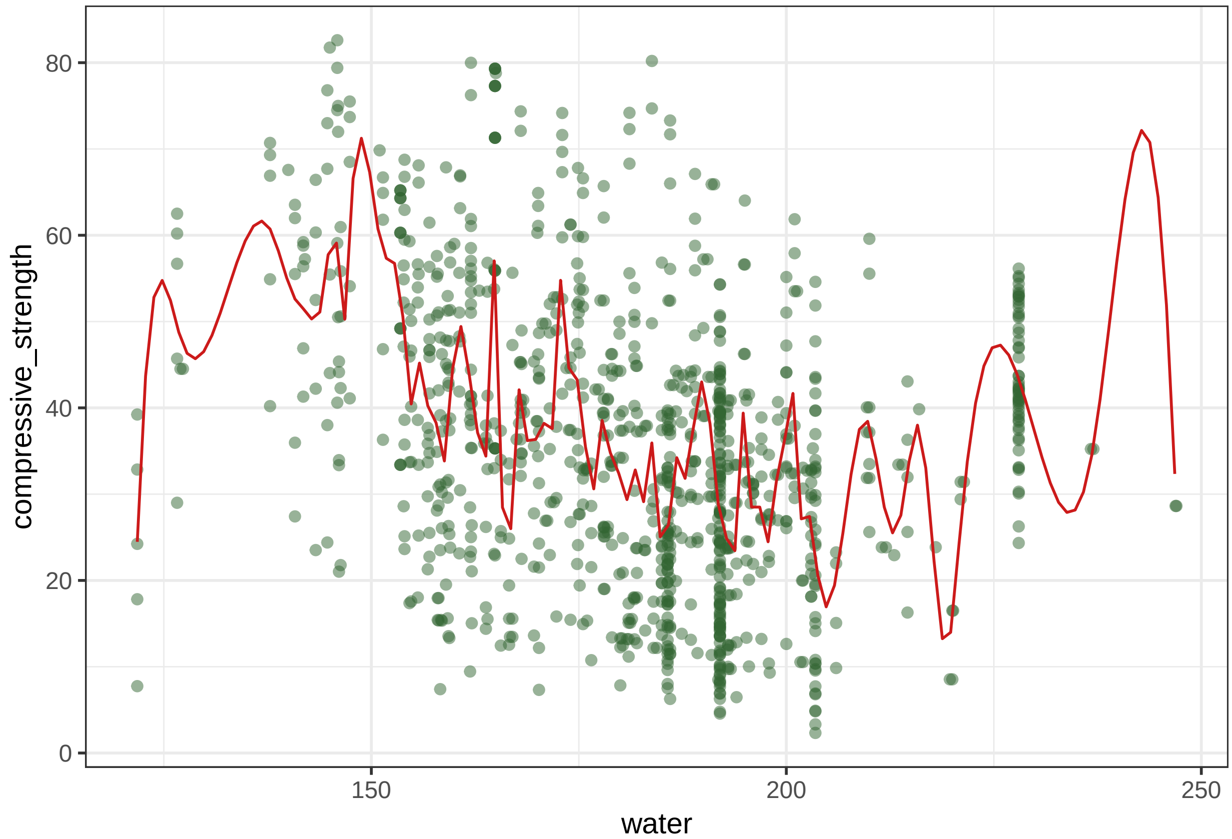
Based on the plots, with increasing degree of freedom from 30 to 80, the flexibility also increases for each model.

fit smoothing spline using df obtained by generalized CV

```r
fit.ss <- smooth.spline(concrete$water, concrete$compressive_strength)
fit.ss$df
```

```
## [1] 68.88205
```

```r
pred.ss <- predict(fit.ss,
                   x = water.grid)

pred.ss.df <- data.frame(pred = pred.ss$y,
                         water = water.grid)

p +
geom_line(aes(x = water, y = pred), data = pred.ss.df,
          color = rgb(.8, .1, .1, 1)) + theme_bw()
```



The plot above shows the optimal df calculated by generalized CV.

(d) Fit a GAM using all the predictors. Plot the results and explain your findings.

Fit GAM

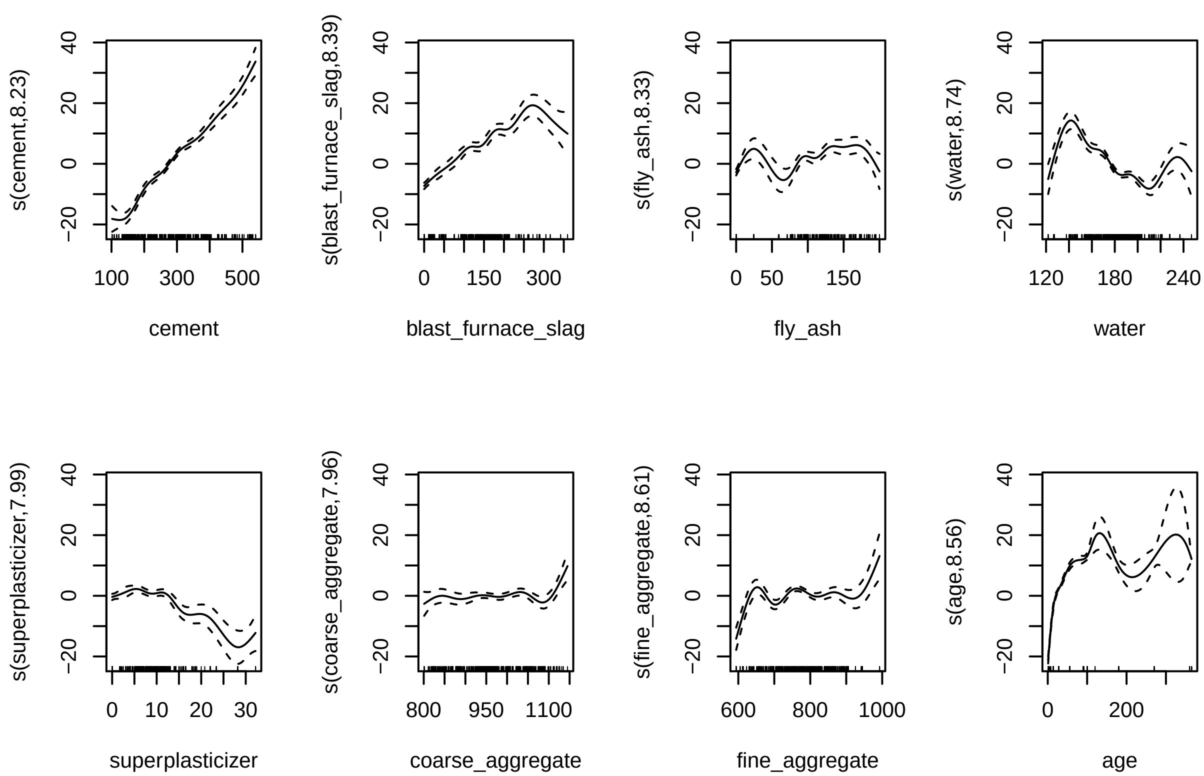
```
gam_fit = gam(compressive_strength ~ s(cement) + s(blast_furnace_slag) + s(fly_ash) + s(water) + s(supe

summary(gam_fit)
```

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## compressive_strength ~ s(cement) + s(blast_furnace_slag) + s(fly_ash) +
##     s(water) + s(superplasticizer) + s(coarse_aggregate) + s(fine_aggregate) +
##     s(age)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  35.8180     0.1671   214.4   <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(cement)             8.228  8.833  48.285  < 2e-16 ***
## s(blast_furnace_slag) 8.388  8.874  24.855  < 2e-16 ***
## s(fly_ash)            8.331  8.851   9.742 3.04e-14 ***
## s(water)              8.742  8.974  26.469  < 2e-16 ***
## s(superplasticizer)   7.989  8.714  10.871 7.77e-16 ***
## s(coarse_aggregate)   7.956  8.702   3.595 0.000305 ***
## s(fine_aggregate)     8.614  8.950  18.405  < 2e-16 ***
## s(age)                8.561  8.901 366.698  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## R-sq.(adj) =  0.897   Deviance explained = 90.4%
## GCV = 30.786  Scale est. = 28.759    n = 1030
```

Plot the result

```r
par(mfrow = c(2, 4))
plot(gam_fit)
```



Based on the plots, we observe that there is a linear relationship between compressive strength and cement, where we could furtherly build linear regression model. However, there is no clear linear relationship between compressive strength and other predictors.