Assignment3

Name : Sai Ram Oduri

CWID:A20522183

Chapter 6 –

Question 1

a) Best subset selection has the lowest training RSS. Because, it fits for every possible combination of predictors. Given a large p, it increases the chance of finding models that will give best fit to the data.
b) Less test RSS contains 2^p models and best subset selection will have high chances of choosing a model and the rest will consider only 1 + p(p+1)/2 models.
c) The following statements are true/false about variable selection methods in regression analysis:
   i) TRUE. Forward stepwise selection adds one predictor to the k-variable model.

   ii) TRUE, Backward stepwise selection obtains the k-variable model by removing one predictor from the (k+1)-variable model, which reduces the residual sum of squares (RSS).

   iii) FALSE, There is no relationship between the models obtained from forward and backward selection methods as they follow different criteria.
   iv) FALSE, The best subset method also follows different criteria and does not have a relationship between the models obtained from forward and backward selection methods.
   v)FALSE, The best subset method selects the best model among all possible models with k+1 predictors, so it does not necessarily choose the same predictors for the k-predictor model.

Question 2

a) Option iii) - Regularization with increased lambda leads to a less flexible fit, resulting in improved prediction accuracy if the increase in bias is outweighed by the decrease in variance. This is achieved by reducing the estimated coefficients, with some being set to zero, leading to a significant decrease in variance for a minor increase in bias.

(b) The same as (a) but all variables have non-zero coefficients.

(c) Option ii) is correct. Non-linear models tend to be more flexible and therefore have higher prediction variance and lower bias. To improve predictions, the variance must increase less than the decrease in bias, which is the trade-off between bias and variance.

Question 3

a)iv. Steadily decrease

The training RSS will steadily decrease as the value of s increases from 0 because this reduces the restriction on coefficients and allows the model to become more flexible.

b) ii.) The test RSS will initially decrease and then eventually start increasing in a U shape as the value of s increases from 0 because this reduces the restriction on coefficients and allows the model to become more flexible, resulting in less test RSS at first but then more as the model becomes too flexible.

c) iii).The variance of the model will steadily increase as the value of s increases from 0 because this reduces the restriction on coefficients and allows the model to become more flexible, leading to increased variance.

d) iv)The bias of the model will steadily decrease as the value of s increases from 0 because this reduces the restriction on coefficients and allows the model to become more flexible, leading to decreased bias.

e) v).The irreducible errors are independent of the model, so they will remain constant regardless of the value of s.

Question 4

a) The training RSS will steadily increase as the value of λ increases from 0 because this increases the restriction on coefficients (i.e. shrinking the coefficients) and makes the model less flexible, causing the coefficients to deviate from their least square estimates.

b) The test RSS will initially decrease and then eventually start increasing in a U shape as the value of λ increases from 0 because this increases the restriction on coefficients (i.e. shrinking the coefficients) and makes the model less flexible, resulting in less test RSS at first but then more as the model becomes too inflexible.

c) The variance of the model will steadily decrease as the value of λ increases from 0 because this increases the restriction on coefficients (i.e. shrinking the coefficients) and makes the model less flexible, leading to decreased variance.

d) The bias of the model will steadily increase as the value of λ increases from 0 because this increases the restriction on coefficients (i.e. shrinking the coefficients) and makes the model less flexible, leading to increased bias.

e) The irreducible error is independent of the model, so it will remain constant regardless of the value of λ.

Que5

(a) Ridge Regression is given by

$$\text{Minimize}: \sum_{i=1}^{n} \left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p} \hat{\beta}_j x_j\right)^2 + \lambda \sum_{j=1}^{p} \hat{\beta}_j^2$$

here $n = p = 2$ and $\hat{\beta}_0 = 0$

$$\rightarrow \min \left[(y_i - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \hat{\beta}_1 x_{21} - \hat{\beta}_{21} x_{22})^2 \right] + \lambda(\beta_1^2 + \beta_2^2) \quad — (1)$$

(b) expanding equation.

$$\rightarrow (y_1^2 + \hat{\beta}_1^2 x_{11}^2 + \hat{\beta}_2^2 x_{12}^2 - 2\hat{\beta}_1 x_{11} y_1 - 2\hat{\beta}_2 x_{12} y_1 + 2\hat{\beta}_1 \hat{\beta}_2 x_{11} x_{12}) + (y_2^2 + \hat{\beta}_1^2 x_{21}^2 + \hat{\beta}_2^2 x_{22}^2 - 2\hat{\beta}_1 y_2 - 2\hat{\beta}_2 x_{22} y_2 + 2\hat{\beta}_1 \hat{\beta}_2 x_{21} x_{22} + \lambda\beta_1^2 + \lambda\beta_2^2)$$

To minimize it we will take its derivative and equate it to zero.

$$\frac{d}{d\hat{\beta}_1} (2\hat{\beta}_1 x_{11}^2 - 2x_{11} y_1 + 2\hat{\beta}_2 x_{11} x_{12}) + (2\hat{\beta}_1 x_{21}^2 - 2x_{21} y_2 + 2\beta_2 x_{21} x_{22}) + 2\lambda\hat{\beta}_1 = 0$$

given $x_{11} = x_{12} = x_1$ & $x_{21} = x_{22} = x_2$ and divide throughout by 2

$(\hat{\beta}_1 x_1^2 - x_1 y_1 + \hat{\beta} x_2^2) + (\hat{\beta}_1 x_2^2 - x_2 y_2 + \hat{\beta}_2 x_2^2)$
$+ \lambda \beta_1 = 0$

$\Rightarrow \hat{\beta}_1 (x_1^2 + x_2^2) + \beta_2 (x_1^2 + x_2^2) + \lambda \hat{\beta}_1$

$= x_1 y_1 + x_2 y_2$

Add $2\hat{\beta}_1 x_1 x_2$ and $2\beta_2 x_1 x_2$ on both sides

$\Rightarrow \hat{\beta}_1 (x_1 + x_2)^2 + \beta_2 (x_1 + x_2)^2 + \lambda \hat{\beta}_1 = x_1 y_1 + x_2 y_2$

As $x_1 + x_2 = 0$

$\lambda \hat{\beta}_1 = x_1 y_1 + x_2 y_2 + 2\hat{\beta}_1 x_1 x_2 + 2\hat{\beta}_2 x_1 x_2$ (1)

$\lambda \hat{\beta}_2 = x_1 y_1 + x_2 y_2 + 2\hat{\beta}_1 x_1 x_2 + 2\hat{\beta}_2 x_1 x_2$

From eqn (1) & (2)

$\beta_1 = \beta_2$

(c)

$$\min \left[(y_1 - \hat{\beta}_1 x_{11} - \hat{\beta}_2 x_{12})^2 + (y_2 - \beta_1 x_{21} - \beta_2 x_{22})^2 \right] + \lambda \left(|\hat{\beta}_1| + |\hat{\beta}_2| \right)$$

(d) Replacing the penalty term from ridge regression the derivative term to β is

$$= \frac{\partial}{\partial \beta} (\lambda |\beta|) : \lambda \frac{|\beta|}{\beta}$$

Then, we have $\dfrac{\lambda |\beta_1|}{\beta_1} = \dfrac{\lambda |\beta_2|}{|\beta_2|}$

β_1 & β_2 are positive on both sids

Chapter 7

2,3,4,5

Question no 2

2 a)

a) In this case $\hat{g} = 0$, because a large smoothing parameter forces $g^{(0)}(x) \to 0$.

b) $\hat{g} = c$, because of large smoothing parameter $g^{(1)}(x) \to 0$.

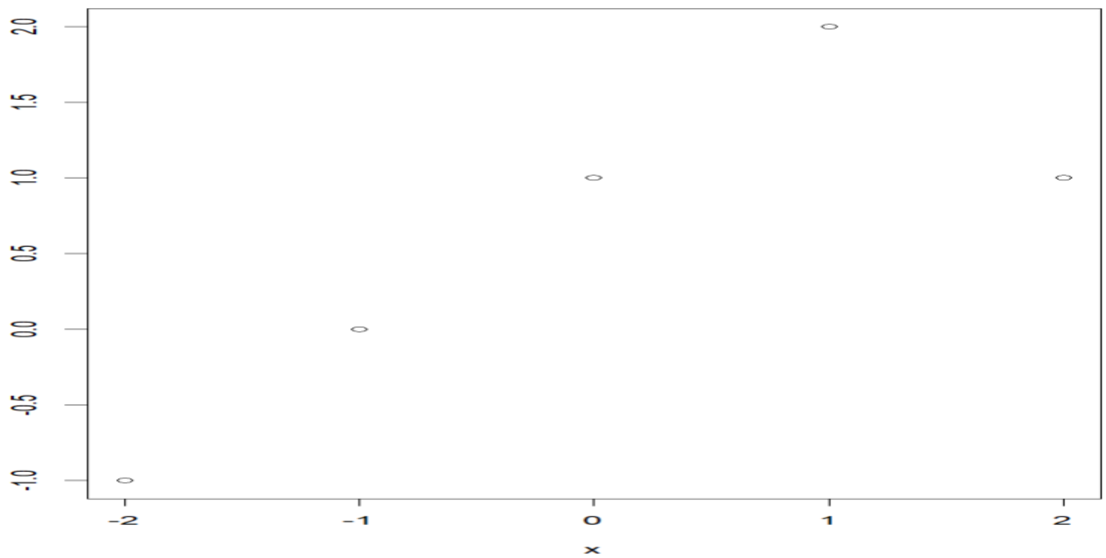
c) $\hat{g} = cx + d$, because of large smoothing parameter $g^{(2)}(x) \to 0$

d) $\hat{g} = cx^2 + dx + e$, because of large smoothing parameter $g^{(3)}(x) \to 0$

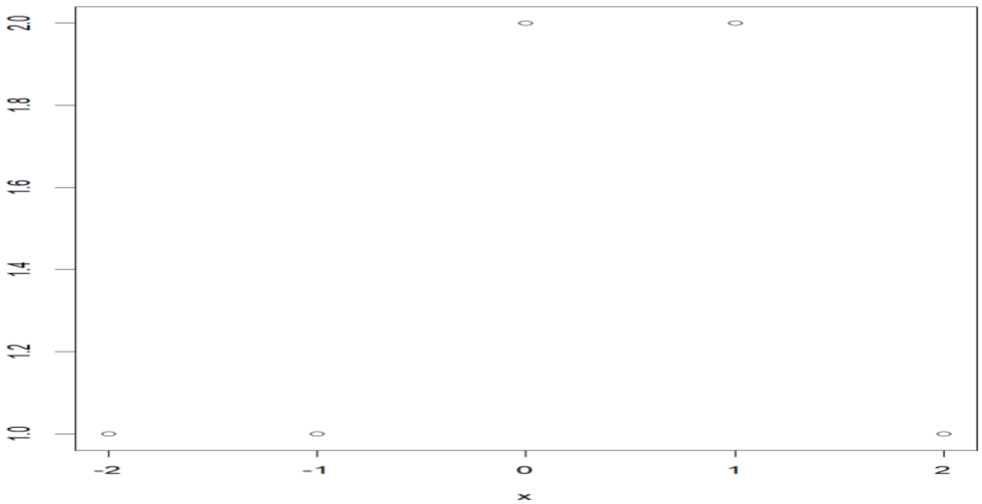
e) The penalty has no role, g is the interpolating spline

3.

```
x = -2:2

y = 1 + x + -2 * (x-1)^2 * I(x>1)
plot(x, y)
```



4.
```
x = -2:2

y = c(1 + 0 + 0, # x = -2
      1 + 0 + 0, # x = -1
      1 + 1 + 0, # x = 0
      1 + (1-0) + 0, # x = 1
      1 + (1-1) + 0 # x =2
      )
plot(x,y)
```



5Ans.

5.

spline.

5 (a) The smoothing spline $\hat{g}_2$ will probably have smaller ~~the~~ training RSS because it will be higher order polynomial.

(b) So, $\hat{g}_2$ has to flexible and it may overfit data. Probably $\hat{g}_1$ has smaller test RSS.

(c) if $\lambda = 0$, we have $\hat{g}_1 = \hat{g}_2$ both have same training and test RSS.

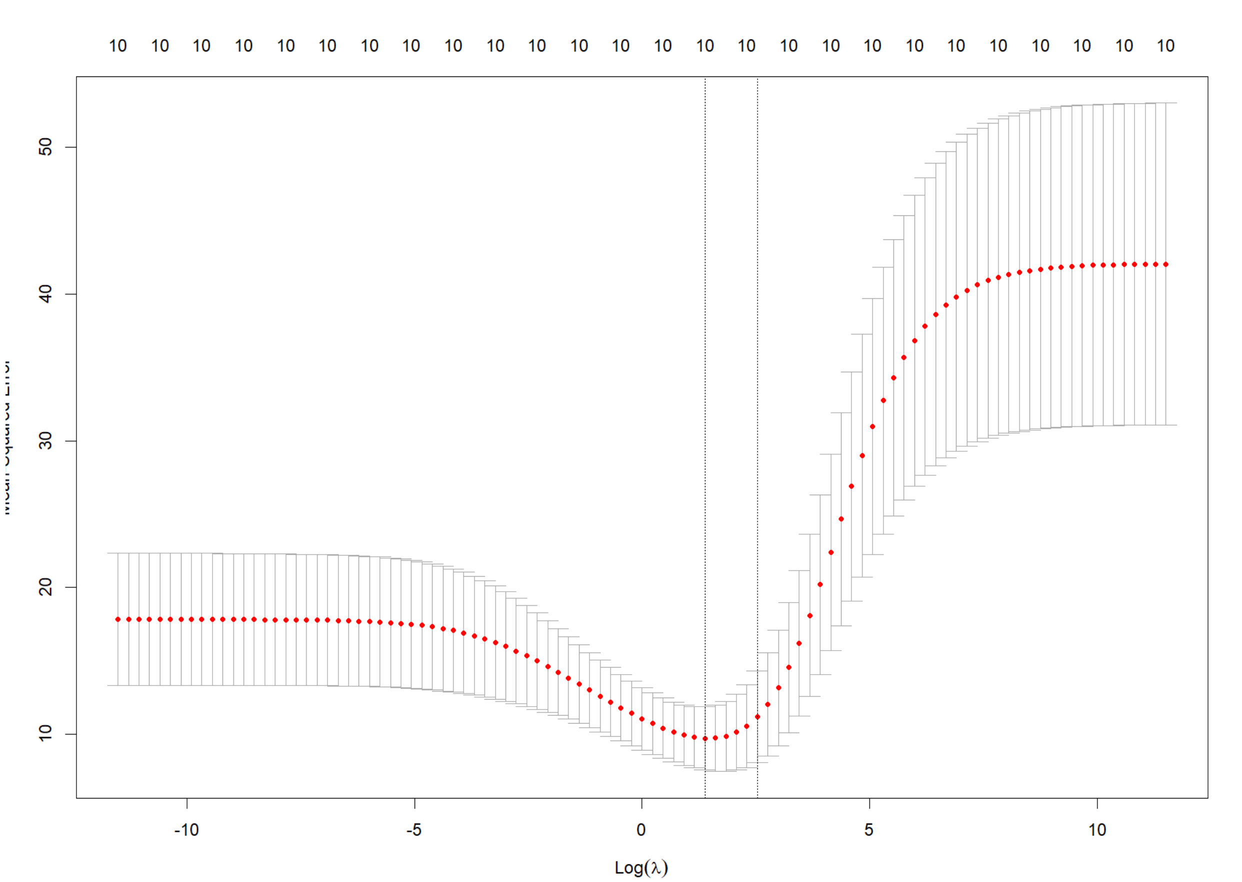
Programming Questions

Problem 2.1

```
>
> # Ridge Regression
> # Loading the library
> library(glmnet)
> # Getting the independent variable
> x <- model.matrix(mpg ~ ., train)[, -1]
>
> # Getting the dependent variable
> y <- train$mpg
>
> # Setting the range of lambda values
> lambda_seq <- 10^seq(5, -5, by = -0.1)
>
> # Using cross-validation glmnet
> ridge_cv <- cv.glmnet(x, y, alpha = 0, lambda = lambda_seq)
Warning message:
Option grouped=FALSE enforced in cv.glmnet, since < 3 observations per fold
> plot(ridge_cv)
>
> # Best lambda value
> best_lambda <- ridge_cv$lambda.min
> best_lambda
[1] 3.981072
>
> # Using glmnet function to build the ridge regression model
> fit <- glmnet(x, y, alpha = 0, lambda = best_lambda)
>
> # Checking the model
> summary(fit)
          Length Class     Mode
a0         1     -none-    numeric
beta      10     dgCMatrix S4
df         1     -none-    numeric
dim        2     -none-    numeric
lambda     1     -none-    numeric
dev.ratio  1     -none-    numeric
nulldev    1     -none-    numeric
npasses    1     -none-    numeric
jerr       1     -none-    numeric
offset     1     -none-    logical
call       5     -none-    call
nobs       1     -none-    numeric
>
> # Coefficients
```

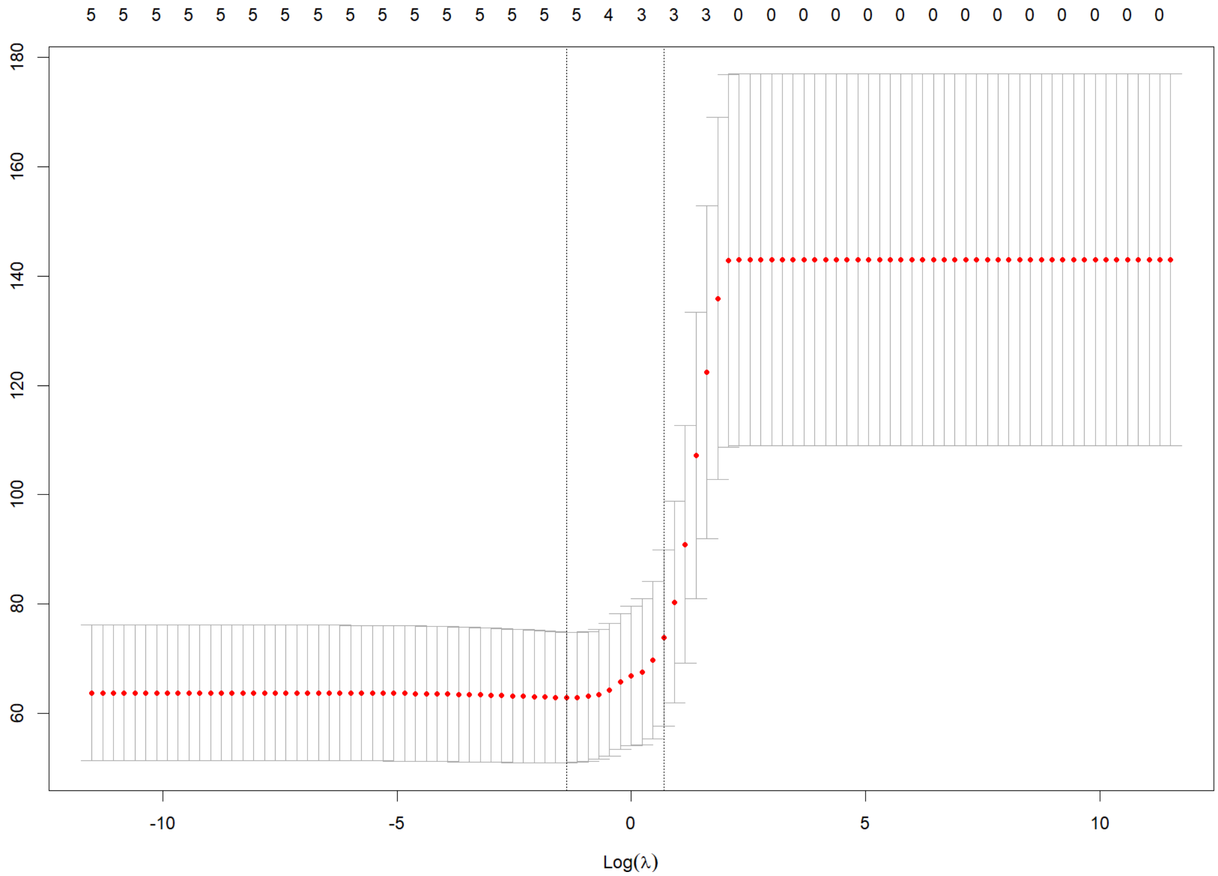
```
> coef(ridge_cv, s = "lambda.min")
11 x 1 sparse Matrix of class "dgCMatrix"
                      s1
(Intercept) 19.533705869
cyl         -0.368008786
disp        -0.005720897
hp          -0.011099008
drat         1.156418468
wt          -1.109528763
qsec         0.203566030
vs           0.804978288
am           1.520934064
gear         0.588710051
carb        -0.497348516
>
> # For test dataset
> xx <- model.matrix(mpg ~ ., test)[, -1]
> model_predict <- predict(fit, s = best_lambda, newx = xx, type = "response"
)
>
> # MSE on test data
> mean((model_predict - test$mpg)^2)
[1] 1.184656
```

Plots



Problem 2.2

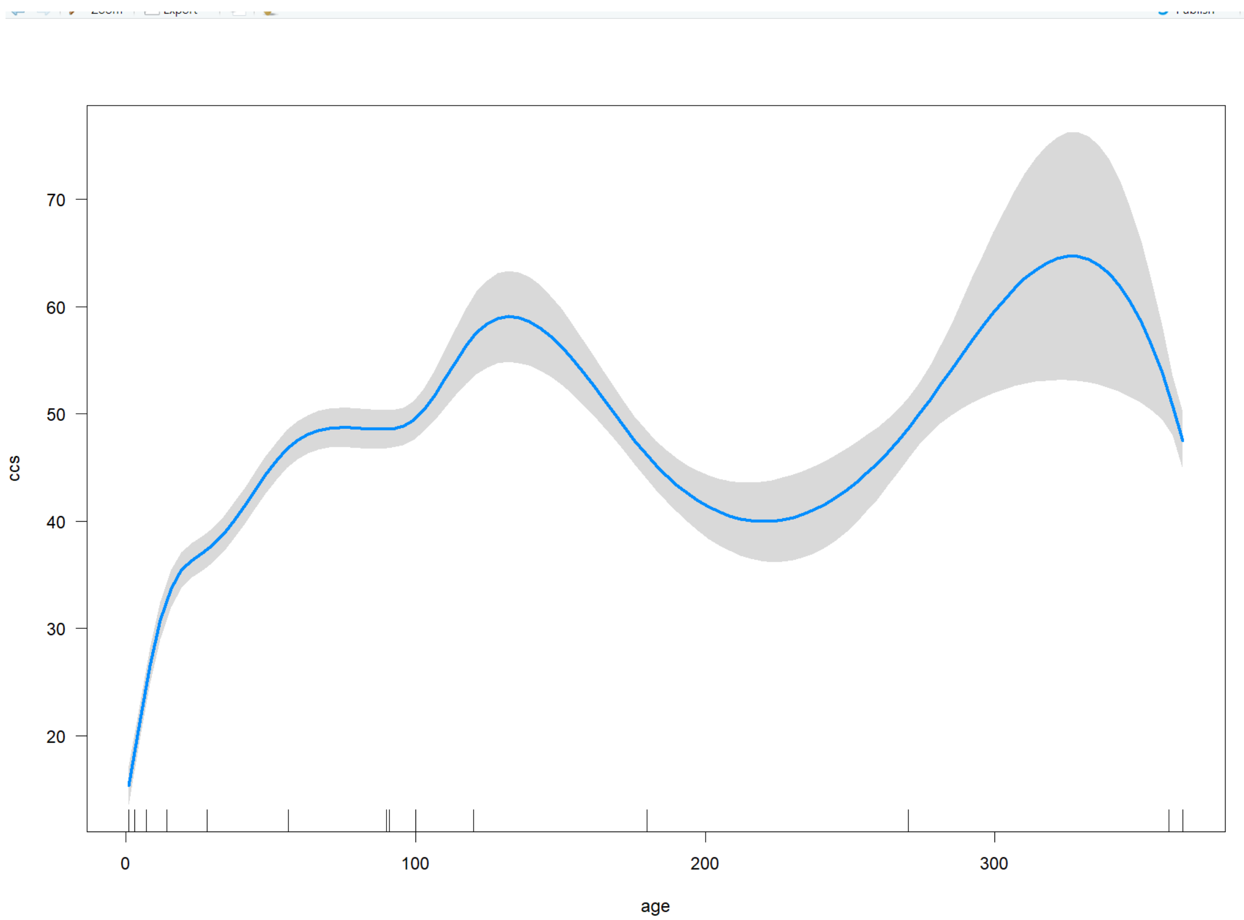
```
> # Calculate test mean squared error for the LASSO model
> lasso_predictions <- predict(lasso_model, newx=model.matrix(Fertility ~ .,
data=test_data)[,-1], type="response")
> lasso_mse <- mean((test_data$Fertility - lasso_predictions)^2)
> cat("LASSO Model Test MSE:", lasso_mse, "\n")
LASSO Model Test MSE: 57.83554
>
> # Print the coefficients of both models
> cat("\nLinear Model Coefficients:\n")

Linear Model Coefficients:
> print(coef(lm_model))
     (Intercept)      Agriculture      Examination        Education         C
atholic Infant.Mortality
     66.16965921      -0.17497395      -0.05176448      -1.06932048       0.1
1713319       1.03247401
> cat("\nLASSO Model Coefficients:\n")

LASSO Model Coefficients:
> print(coef(lasso_model))
6 x 1 sparse Matrix of class "dgCMatrix"
                          s0
(Intercept)      63.52172468
Agriculture      -0.13304558
Examination      -0.02199682
Education        -0.99598167
Catholic          0.10961142
Infant.Mortality  1.00951187
```

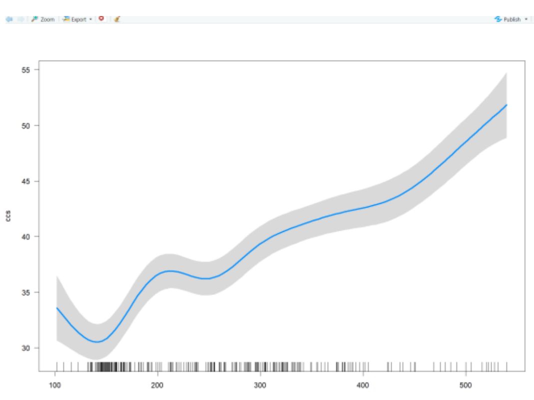


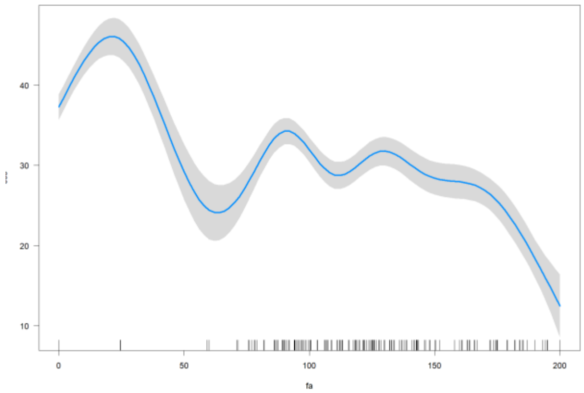
Question 2.3

```
> library(mgcv)
> library(visreg)
> library(glmnet)
> library(readxl)
>
> concrete_data <- read_excel('Z://IIT/DPA/Assignment3/Concrete_Data.xls')
>
> concrete <- data.frame(concrete_data)
> colnames(concrete) <- c("cem", "bfs", "fa", "water", "sp", "cagg", "fagg",
"age", "ccs")
>
> gam.linear <- gam(ccs~cem+fa+water+sp+cagg+fagg+age, data=concrete)
> # r2
> summary(gam.linear)$r.sq
[1] 0.5730282
>
> gam.smooth<- gam(ccs~s(cem)+s(fa)+s(water)+s(sp)+s(cagg)+s(fagg)+s(age), da
ta=concrete)
> # r2
> summary(gam.smooth)$r.sq
[1] 0.8758309
>
> visreg(gam.smooth, scale = "response", alpha = 0.2, rug = TRUE)
Hit <Return> to see next plot:
Hit <Return> to see next plot:
```

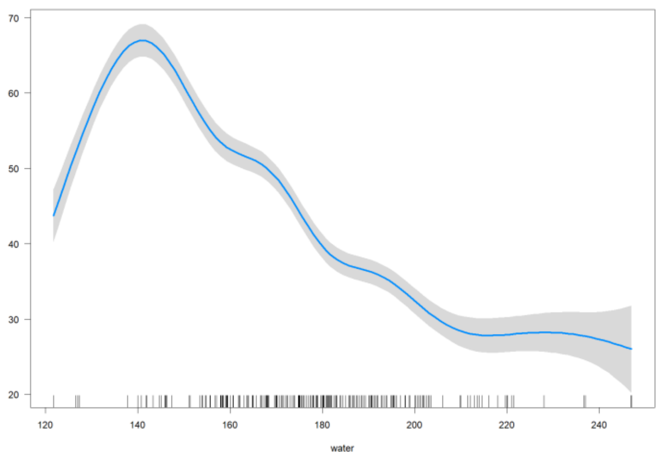
```
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
Hit <Return> to see next plot:
```



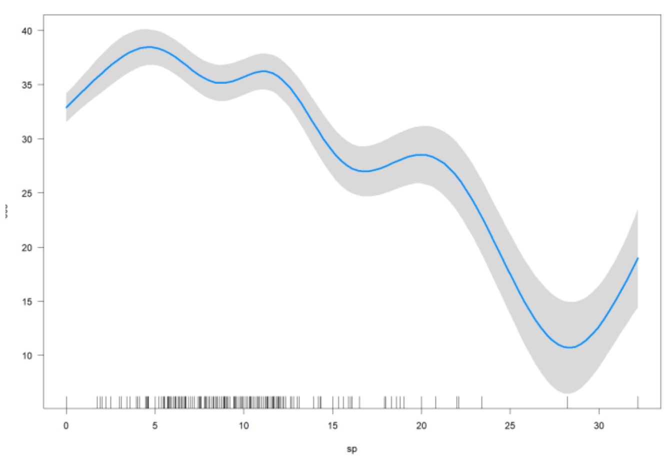




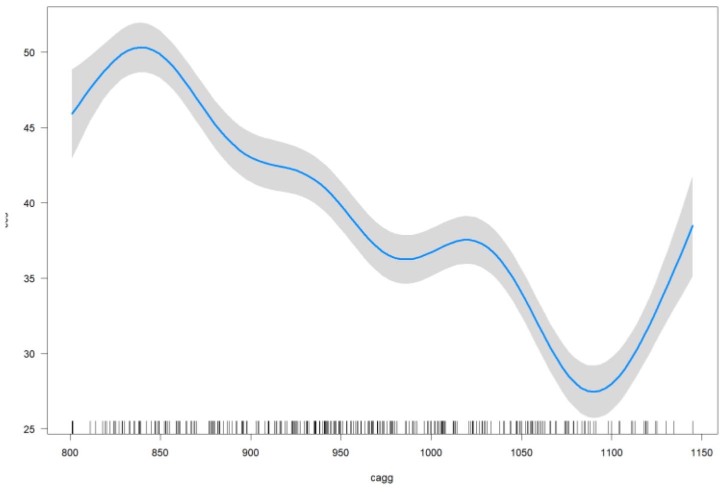
40
30
20
10
0
50
100
150
200
fa

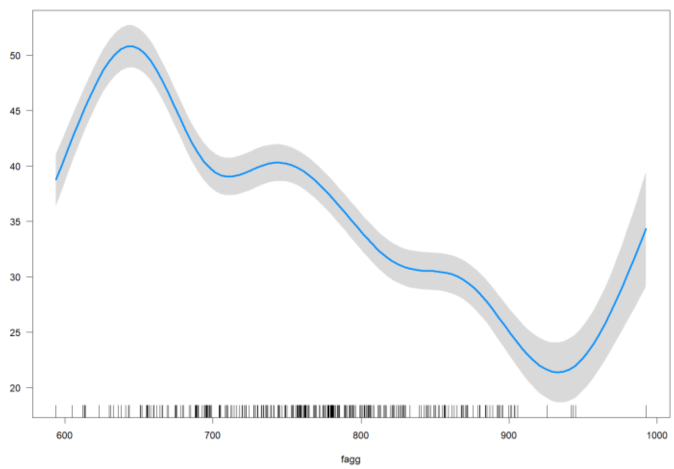


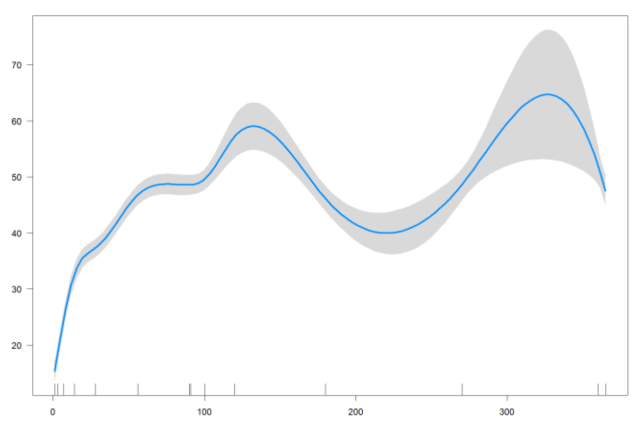
70
60
50
40
30
20
120
140
160
180
200
220
240
water



40
35
30
25
20
15
10
0
5
10
15
20
25
30
sp


50
45
40
35
30
25
800
850
900
950
1000
1050
1100
1150
cagg


50
45
40
35
30
25
20
600
700
800
900
1000
fagg


70
60
50
40
30
20
0
100
200
300
age