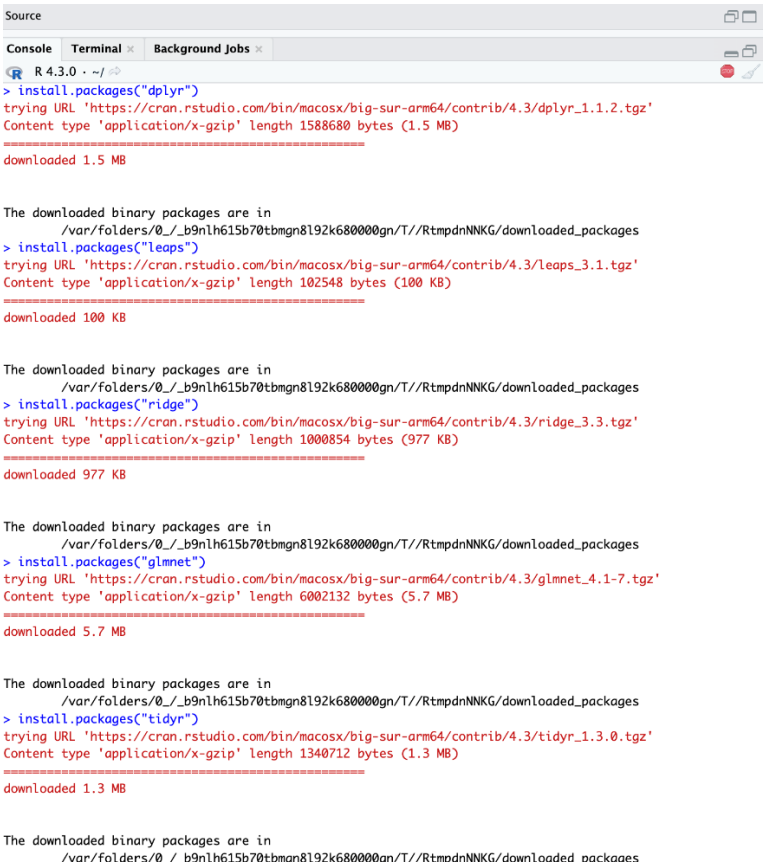
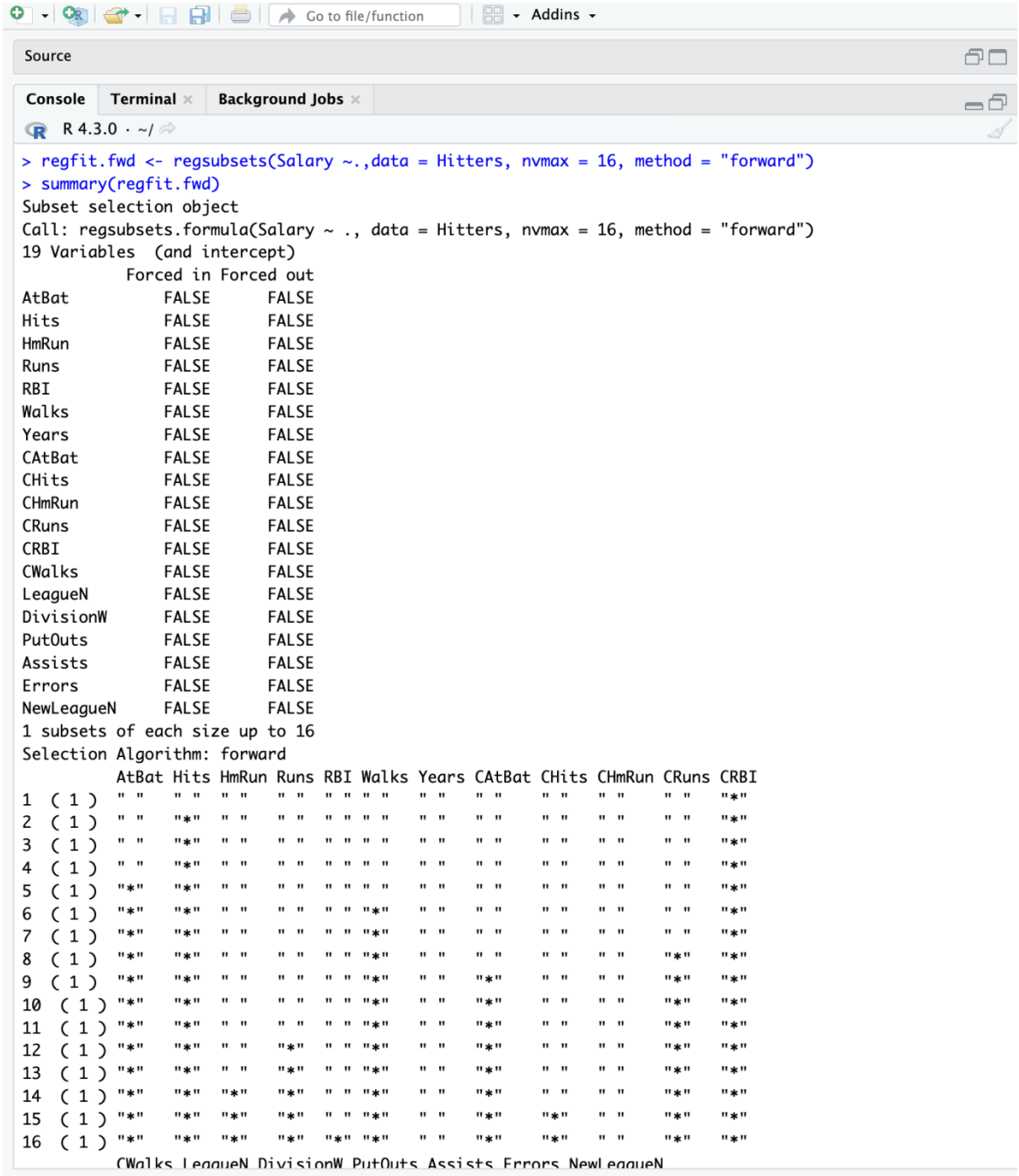
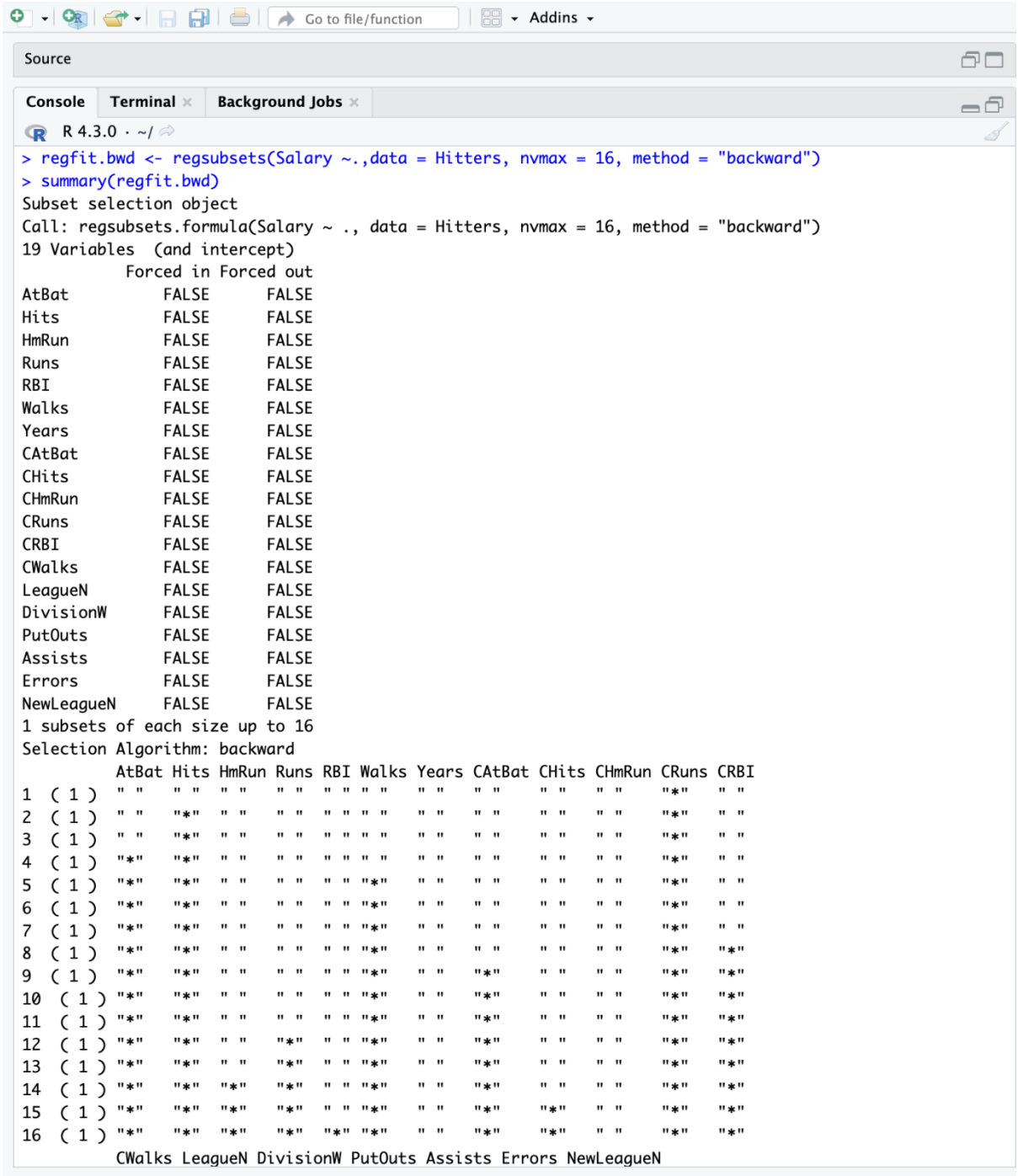



```
>
> library(ISLR2)

Attaching package: ‘ISLR2’

The following object is masked _by_ ‘.GlobalEnv’:

    Hitters

> library(leaps)
> library(dplyr)

Attaching package: ‘dplyr’

The following objects are masked from ‘package:stats’:

    filter, lag

The following objects are masked from ‘package:base’:

    intersect, setdiff, setequal, union

> library(ridge)
> library(glmnet)
Loaded glmnet 4.1-7
> library(tidyr)

Attaching package: ‘tidyr’

The following objects are masked from ‘package:Matrix’:

    expand, pack, unpack

>
```

PART – A

- When nvmax is set to 16, only 16 variables in each subset are shown in the summary() reports generated by the forward and backward methods, respectively.






```
> regfit.bwd <- regsubsets(Salary ~.,data = Hitters, nvmax = 16, method = "backward")
> summary(regfit.bwd)
Subset selection object
Call: regsubsets.formula(Salary ~ ., data = Hitters, nvmax = 16, method = "backward")
19 Variables  (and intercept)
           Forced in Forced out
AtBat          FALSE      FALSE
Hits           FALSE      FALSE
HmRun          FALSE      FALSE
Runs           FALSE      FALSE
RBI            FALSE      FALSE
Walks          FALSE      FALSE
Years          FALSE      FALSE
CAtBat         FALSE      FALSE
CHits          FALSE      FALSE
CHmRun         FALSE      FALSE
CRuns          FALSE      FALSE
CRBI           FALSE      FALSE
CWalks         FALSE      FALSE
LeagueN        FALSE      FALSE
DivisionW      FALSE      FALSE
PutOuts        FALSE      FALSE
Assists        FALSE      FALSE
Errors         FALSE      FALSE
NewLeagueN     FALSE      FALSE
1 subsets of each size up to 16
Selection Algorithm: backward
          AtBat Hits HmRun Runs RBI Walks Years CAtBat CHits CHmRun CRuns CRBI
1  ( 1 )  " "   " "  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "
2  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "
3  ( 1 )  " "   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "
4  ( 1 )  "*"   "*"  " "   " "  " " " "   " "   " "    " "   " "    "*"   " "
5  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " "
6  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " "
7  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   " "
8  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   " "    " "   " "    "*"   "*"
9  ( 1 )  "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*"
10  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*"
11  ( 1 ) "*"   "*"  " "   " "  " " "*"   " "   "*"    " "   " "    "*"   "*"
12  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"    " "   " "    "*"   "*"
13  ( 1 ) "*"   "*"  " "   "*"  " " "*"   " "   "*"    " "   " "    "*"   "*"
14  ( 1 ) "*"   "*"  "*"   "*"  " " "*"   " "   "*"    " "   " "    "*"   "*"
15  ( 1 ) "*"   "*"  "*"   "*"  " " "*"   " "   "*"    "*"   " "    "*"   "*"
16  ( 1 ) "*"   "*"  "*"   "*"  "*" "*"   " "   "*"    "*"   " "    "*"   "*"
          CWalks LeagueN DivisionW PutOuts Assists Errors NewLeagueN
```

Ridge Regression

- The last three digits of my student ID, 814, are substituted for the lambda value.

```
>
> x <- model.matrix(Salary ~.,Hitters)[,-1]
> y <- Hitters$Salary
> dim(Hitters)
[1] 263  20
> Hitters = na.omit(Hitters)
> with(Hitters,sum(is.na(Salary)))
[1] 0
> dim(Hitters)
[1] 263  20
> x <- model.matrix(Salary ~.-1, data = Hitters)
>
> y <- Hitters$Salary
> ridge.mod = glmnet(x,y,alpha = 0,)
> grid = 10^seq(10,-2,length = 350)
>
> ridge.mod = glmnet(x,y,alpha = 0,lambda = grid)
> dim(coef(ridge.mod))
[1]  21 350
> ridge.mod$lambda[814]
[1] NA
> coef(ridge.mod)[,50]
  (Intercept)         AtBat          Hits         HmRun          Runs
 5.359167e+02  2.634408e-06  9.556189e-06  3.850614e-05  1.616022e-05
          RBI         Walks         Years        CAtBat         CHits
 1.707026e-05  2.009067e-05  8.216187e-05  2.261892e-07  8.324411e-07
       CHmRun         CRuns          CRBI        CWalks       LeagueA
 6.277752e-06  1.670062e-06  1.723541e-06  1.823490e-06  2.806875e-05
      LeagueN     DivisionW       PutOuts       Assists        Errors
-2.806875e-05 -3.778420e-04  1.055173e-06  1.723477e-07 -8.036230e-07
   NewLeagueN
-5.574925e-06
>
>
>
```

When the 814 returns an error for the function coef() (index greater than maximum)

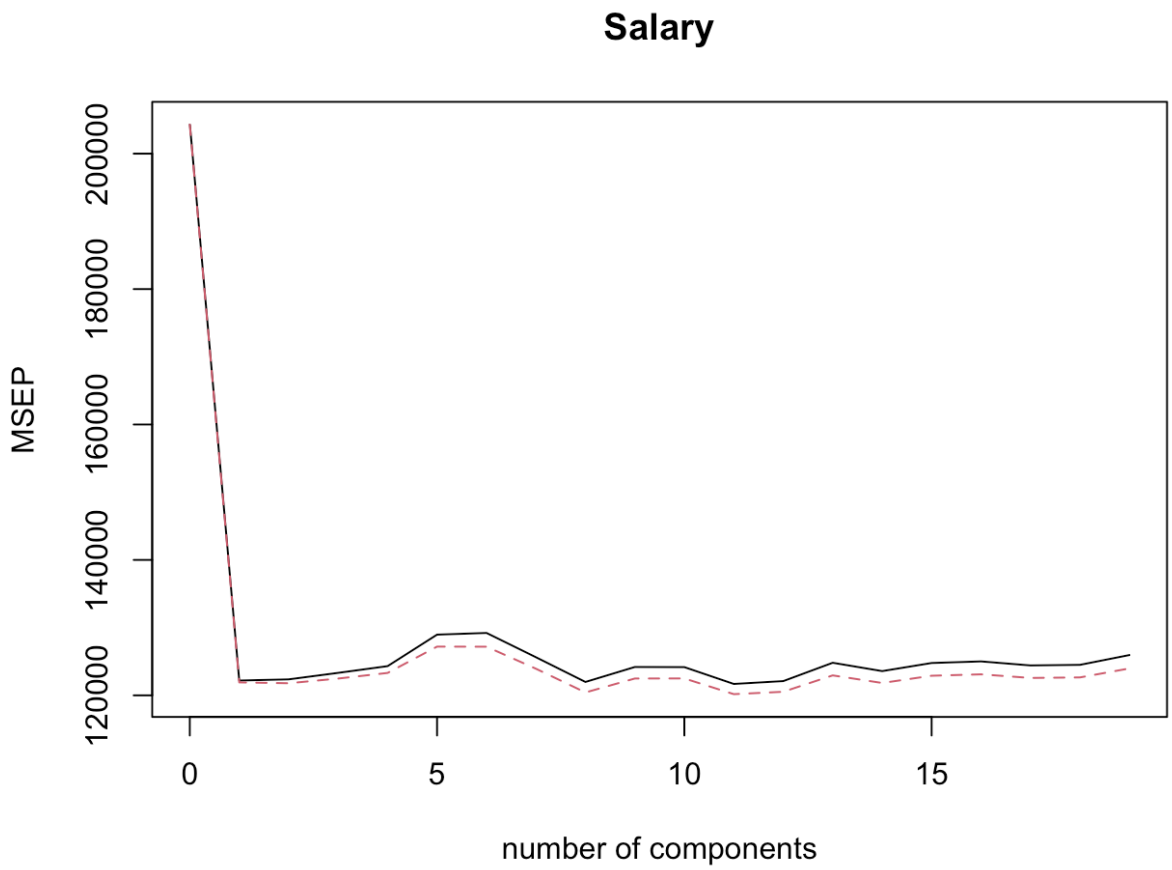
```
>
> coef(ridge.mod)[,814]
Error in intI(j, n = x@Dim[2], dn[[2]], give.dn = FALSE) :
  index larger than maximal 350
> coef(ridge.mod)[,50]
  (Intercept)         AtBat          Hits         HmRun          Runs
 5.359167e+02  2.634408e-06  9.556189e-06  3.850614e-05  1.616022e-05
          RBI         Walks         Years        CAtBat         CHits
 1.707026e-05  2.009067e-05  8.216187e-05  2.261892e-07  8.324411e-07
       CHmRun         CRuns          CRBI        CWalks       LeagueA
 6.277752e-06  1.670062e-06  1.723541e-06  1.823490e-06  2.806875e-05
      LeagueN     DivisionW       PutOuts       Assists        Errors
-2.806875e-05 -3.778420e-04  1.055173e-06  1.723477e-07 -8.036230e-07
   NewLeagueN
-5.574925e-06
> sqrt(sum(coef(ridge.mod)[-1,814]^2))
Error in intI(j, n = d[2L], dn[[2L]], give.dn = FALSE) :
  index larger than maximal 350
> predict(ridge.mod, s = 8147, type = "coefficients")[1:20, ]
  (Intercept)         AtBat          Hits         HmRun          Runs
369.401629026   0.047799448   0.180934447   0.671081031   0.300564597
          RBI         Walks         Years        CAtBat         CHits
  0.310595953   0.378810202   1.412705388   0.004037015   0.015089235
       CHmRun         CRuns          CRBI        CWalks       LeagueA
  0.113184766   0.030272457   0.031260197   0.032118997  -0.327451698
      LeagueN     DivisionW       PutOuts       Assists        Errors
  0.327453963  -8.509127495   0.022181541   0.003462085  -0.032222072
>
```

Partial Least Squares

- Change the seed to 814 after installing the "pls" package.
- The graph is largely flat, with a high value of 100,000 after changing the seed, according to the data.

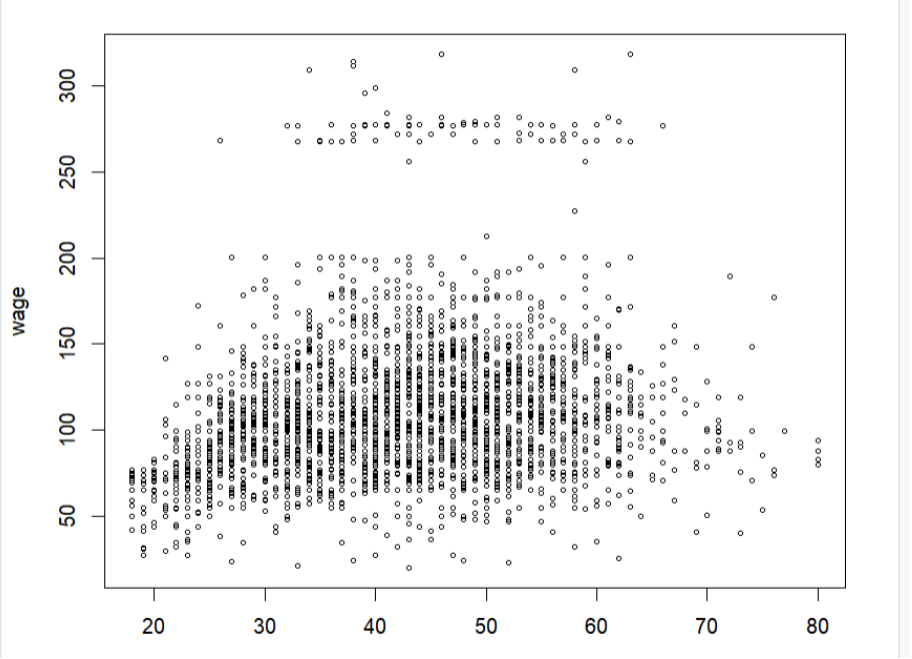
```
> install.packages("pls")
trying URL 'https://cran.rstudio.com/bin/macosx/big-sur-arm64/contrib/4.3/pls_2.8-1.tgz'
Content type 'application/x-gzip' length 1179346 bytes (1.1 MB)
==================================================
downloaded 1.1 MB


The downloaded binary packages are in
        /var/folders/0_/_b9nlh615b70tbmgn8l92k680000gn/T//RtmpHMoC9g/downloaded_packages
> library(pls)

Attaching package: ‘pls’

The following object is masked from ‘package:stats’:

    loadings

> set.seed(814)
> pls.fit <- plsr(Salary ~ .,data = Hitters, scale = TRUE , validation = "CV")
> summary(pls.fit)
Data:   X dimension: 263 19
        Y dimension: 263 1
Fit method: kernelpls
Number of components considered: 19

VALIDATION: RMSEP
Cross-validated using 10 random segments.
       (Intercept)  1 comps  2 comps  3 comps  4 comps  5 comps  6 comps
CV             452    349.6    349.8    351.2    352.6    359.1    359.5
adjCV          452    349.2    349.0    350.0    351.2    356.6    356.6
       7 comps  8 comps  9 comps  10 comps  11 comps  12 comps  13 comps
CV       354.4    349.3    352.4     352.4     348.8     349.4     353.3
adjCV    352.0    347.0    350.0     350.0     346.7     347.2     350.6
       14 comps  15 comps  16 comps  17 comps  18 comps  19 comps
CV        351.5     353.2     353.6     352.7     352.9     354.9
adjCV     349.0     350.6     350.9     350.1     350.2     352.1

TRAINING: % variance explained
        1 comps  2 comps  3 comps  4 comps  5 comps  6 comps  7 comps  8 comps
X         38.08    51.03    65.98    73.93    78.63    84.26    88.17    90.12
Salary    43.05    46.40    47.72    48.71    50.53    51.66    52.34    53.26
        9 comps  10 comps  11 comps  12 comps  13 comps  14 comps  15 comps
X         92.92     95.00     96.68     97.68     98.22     98.55     98.98
Salary    53.52     53.77     54.04     54.20     54.32     54.47     54.54
        16 comps  17 comps  18 comps  19 comps
```

```
> validationplot(pls.fit, val.type = "MSEP")
> pls.fit <- plsr(Salary ~ .,data = Hitters, scale = TRUE , ncomp = 1)
> summary(pls.fit)
Data:   X dimension: 263 19
        Y dimension: 263 1
Fit method: kernelpls
Number of components considered: 1
TRAINING: % variance explained
        1 comps
X         38.08
Salary    43.05
>
```

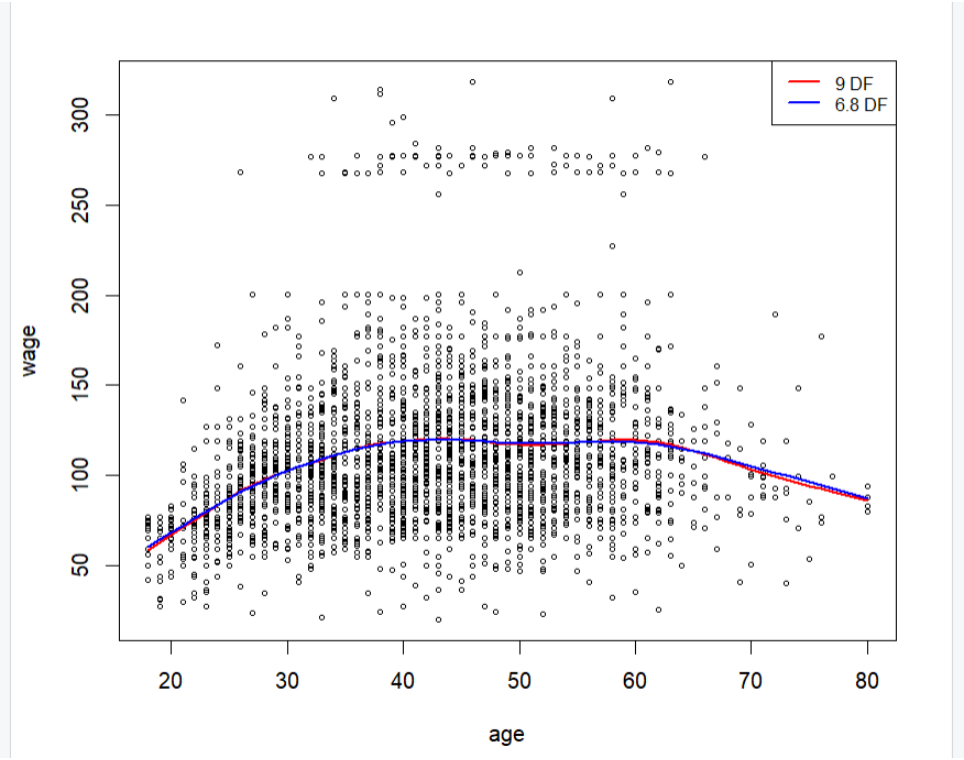


PART -B

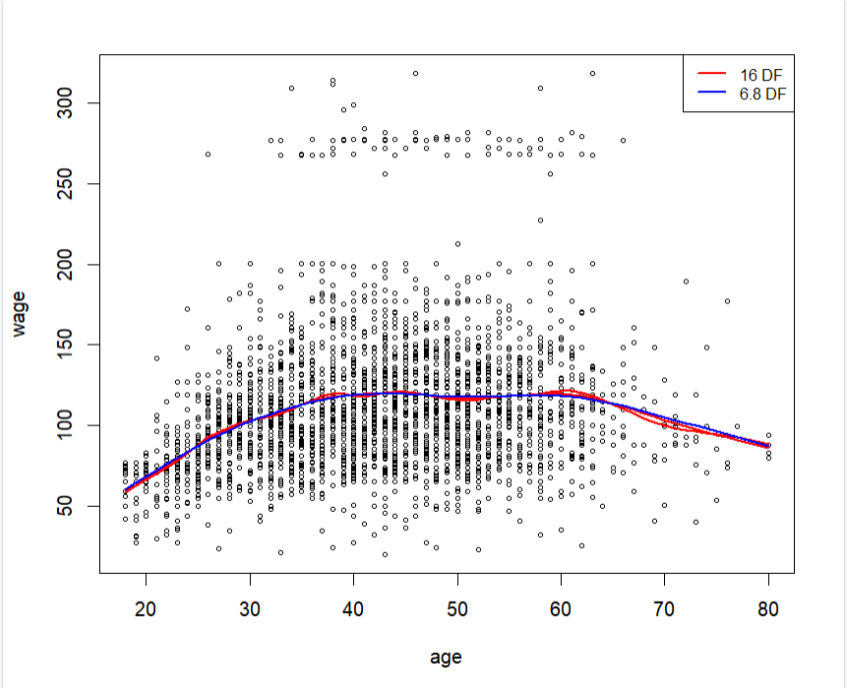
```
> library(ISLR2)
> attach(Wage)
> agelims <- range(age)
>
> plot(age, wage , xlim = agelims , cex = .5, col = "black")
>
```



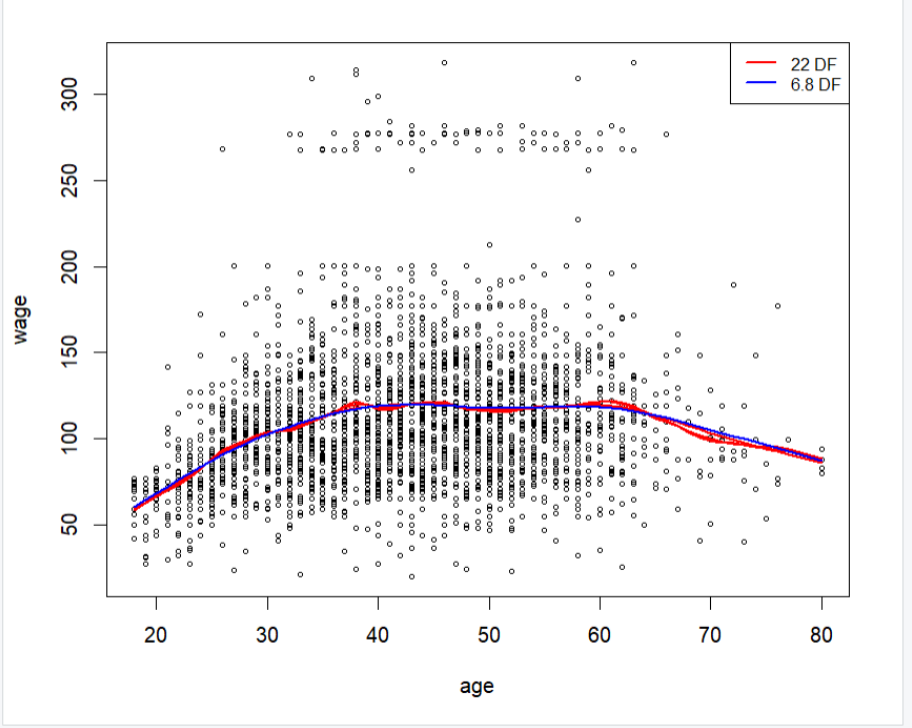

```
>
> fit <- smooth.spline(age, wage, df = 9)
> fit2 <- smooth.spline(age, wage, cv = TRUE)
Warning message:
In smooth.spline(age, wage, cv = TRUE) :
  cross-validation with non-unique 'x' values seems doubtful
> fit2$df
[1] 6.794596
> lines(fit, col = "red", lwd = 2)
> lines(fit2, col = "blue", lwd = 2)
> legend("topright", legend = c("9 DF", "6.8 DF"), col = c("red", "blue"), lty = 1, lwd = 2, cex =
.8)
>
```




```
>
> fit <- smooth.spline(age, wage, df = 16)
> fit2 <- smooth.spline(age, wage, cv = TRUE)
Warning message:
In smooth.spline(age, wage, cv = TRUE) :
  cross-validation with non-unique 'x' values seems doubtful
> fit2$df
[1] 6.794596
> lines(fit, col = "red", lwd = 2)
> lines(fit2, col = "blue", lwd = 2)
> legend("topright", legend = c("16 DF", "6.8 DF"), col = c("red", "blue"), lty = 1, lwd = 2, cex =
.8)
>
```



```
> fit <- smooth.spline(age, wage, df = 22)
> fit2 <- smooth.spline(age, wage, cv = TRUE)
Warning message:
In smooth.spline(age, wage, cv = TRUE) :
  cross-validation with non-unique 'x' values seems doubtful
> fit2$df
[1] 6.794596
> lines(fit, col = "red", lwd = 2)
> lines(fit2, col = "blue", lwd = 2)
> legend("topright", legend = c("22 DF", "6.8 DF"), col = c("red", "blue"), lty = 1, lwd = 2, cex =
.8)
```



Based on the plots at DF = 9, 16, and 22. After age 65, it is seen that as the DF increases, there is a growing gap between the df line (RED) and the true line (BLUE).