Homework 2

Asif Hasan - 301376671

2023-10-31

1. Problem Set 7, Applications

```
library(MASS) # for ridge
```

```
## Warning: package 'MASS' was built under R version 4.1.2
```

```
library(glmnet) # for LASSO
```

```
## Warning: package 'glmnet' was built under R version 4.1.2
```

```
## Loading required package: Matrix
```

```
## Warning: package 'Matrix' was built under R version 4.1.2
```

```
## Loaded glmnet 4.1-7
```

```
# get the data
air.data <- airquality
air.data2 <- na.omit(air.data[, 1:4])
air.data2$TWcp = air.data2$Temp*air.data2$Wind
air.data2$TWrat = air.data2$Temp/air.data2$Wind
```

1(a).

```
# fit a ridge regression
ridge <- lm.ridge(Ozone ~ ., lambda = seq(0, 100, .05), data = air.data2)
which.min(ridge$GCV)
```

```
##   0.20
##      5
```

```
(coef.ri.best <- coef(ridge)[which.min(ridge$GCV), ])
```

```
##                    Solar.R          Wind          Temp          TWcp
## -161.63123617    0.06284069    6.82529896    2.45651021   -0.10883212
##         TWrat
##    1.64685249
```

1(b).

```
# fit a least squares regression
ls <- lm(Ozone ~ ., data = air.data2)
ls
```

```
##
## Call:
## lm(formula = Ozone ~ ., data = air.data2)
##
## Coefficients:
## (Intercept)      Solar.R         Wind         Temp         TWcp        TWrat
##  -191.19856      0.06384      9.56187      2.89466     -0.14751      1.36619
```

The magnitude of coefficient estimates of all parameters (including the intercept) except 'TWrat' is smaller in ridge regression compared to least squares regression.

2(a).

```
# fit lasso
lasso <- cv.glmnet(y = as.matrix(air.data2[, 1]), x = as.matrix(air.data2[, c(2:6)]), famil
y = "gaussian")
# lambda min
lasso$lambda.min
```

```
## [1] 0.006698302
```

```
# lambda 1se
lasso$lambda.1se
```

```
## [1] 9.49467
```

2(b).

```
# get the coefficient estimates
coef(lasso, s = lasso$lambda.min) # lambda min
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                        s1
## (Intercept) -183.90117408
## Solar.R        0.06354556
## Wind           8.88369468
## Temp           2.78797591
## TWcp          -0.13800612
## TWrat          1.43082832
```

```
coef(lasso, s = lasso$lambda.1se) # lambda 1se
```

```
## 6 x 1 sparse Matrix of class "dgCMatrix"
##                       s1
## (Intercept) -37.1881577
## Solar.R       .
## Wind          .
## Temp          0.8031335
## TWcp          .
## TWrat         1.7845895
```

For the Lambda-min value, all parameters have non-zero coefficient estimates. Conversely, when using the Lambda-1SE value, three parameters have coefficient estimates equal to zero. Additionally, for Lambda-1SE, the intercept and 'Temp' have smaller magnitudes, while 'TWrat' has a larger magnitude compared to Lambda-min.

2(c).

```
# fit step-wise regression
initial <- lm(
  data = air.data2,
  formula = Ozone ~ 1
)
final <- lm(
  data = air.data2,
  formula = Ozone ~ .
)

step <- step(
  object = initial, scope = list(upper = final),
  k = log(nrow(air.data2))
)
```

```
## Start:  AIC=781.78
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + TWrat    1     64323  57479 703.13
## + Temp     1     59434  62367 712.19
## + Wind     1     45694  76108 734.29
## + TWcp     1     24804  96998 761.21
## + Solar.R  1     14780 107022 772.13
## <none>                 121802 781.78
##
## Step:  AIC=703.13
## Ozone ~ TWrat
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1     12916  44563 679.59
## + Solar.R  1      6542  50938 694.43
## <none>                  57479 703.13
## + TWcp     1      1256  56223 705.39
## + Wind     1       332  57147 707.20
## - TWrat    1     64323 121802 781.78
##
## Step:  AIC=679.59
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2964.5 41599 676.66
## <none>                 44563 679.59
## + TWcp     1     434.8 44128 683.21
## + Wind     1     222.1 44341 683.74
## - Temp     1   12916.3 57479 703.13
## - TWrat    1   17804.4 62367 712.19
##
## Step:  AIC=676.66
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                 41599 676.66
## - Solar.R  1    2964.5 44563 679.59
## + TWcp     1     508.1 41090 680.00
## + Wind     1     248.0 41351 680.70
## - Temp     1    9339.1 50938 694.43
## - TWrat    1   18045.8 59644 711.94
```

```
summary(step)
```

```
##
## Call:
## lm(formula = Ozone ~ TWrat + Temp + Solar.R, data = air.data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -56.168 -12.102  -4.424  11.403  77.471
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -93.30421   17.28283  -5.399 4.08e-07 ***
## TWrat         2.86326    0.42026   6.813 5.82e-10 ***
## Temp          1.25231    0.25551   4.901 3.41e-06 ***
## Solar.R       0.05960    0.02158   2.761  0.00678 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 19.72 on 107 degrees of freedom
## Multiple R-squared:  0.6585, Adjusted R-squared:  0.6489
## F-statistic: 68.77 on 3 and 107 DF,  p-value: < 2.2e-16
```

For Lambda-min value using LASSO, all five explanatory variables ('Solar.R', 'Wind', 'Temp', 'TWcp', and 'TWrat') are selected. For Lambda-1SE value using LASSO, only 'Temp' and 'TWrat' are selected. Using hybrid stepwise regression, 'Solar.R', 'Temp', and 'TWrat' are selected.

3(a)

```
set.seed(2928893)
```

3(b)

```
# set number of folds
V <- 10
# sample the folds
n = nrow(air.data2)
folds <- floor((sample.int(n) - 1) * V / n) + 1
```

3(c)

```
## including ls and step for part e
# create matrix for MSPEs for 5 models
MSPEs.cv <- matrix(NA, nrow = V, ncol = 5)
colnames(MSPEs.cv) <- c("LS", "Step", "Ridge", "LASSO-min", "LASSO-1SE")
# run cross-validation in for-loop
for (v in 1:V) {

  # fit 5 models on fold == !v
  model.ls.cv <- lm(Ozone ~ ., data=air.data2[folds!=v, ])
  model.step.cv <- step <- step(
    object = lm(data=air.data2[folds!=v, ], formula = Ozone ~ 1),
    scope = list(upper = lm(data=air.data2[folds!=v, ], formula = Ozone ~ .)),
    k = log(nrow(air.data2[folds!=v, ]))
    )
  model.ridge.cv <- lm.ridge(Ozone ~ ., lambda=seq(0, 100, .05), air.data2[folds!=v, ])
  model.lasso.cv <- cv.glmnet(
    y = as.matrix(air.data2[folds!=v, 1]),
    x = as.matrix(air.data2[folds!=v, c(2:6)]),
    family = "gaussian"
    )

  # predict Ozone using the fitted models on fold == v
  pred.ls.cv <- predict(model.ls.cv, newdata=air.data2[folds==v, ])
  pred.step.cv <- predict(model.step.cv, newdata=air.data2[folds==v, ])
  pred.ridge.cv <- as.matrix(cbind(1, air.data2[folds==v, 2:6])) %*%
    coef(model.ridge.cv)[which.min(model.ridge.cv$GCV), ]
  pred.lasso.min.cv <- predict(model.lasso.cv, newx=as.matrix(air.data2[folds==v, c(2:6)]),
                               s=model.lasso.cv$lambda.min)
  pred.lasso.1se.cv <- predict(model.lasso.cv, newx=as.matrix(air.data2[folds==v, c(2:6)]),
                               s=model.lasso.cv$lambda.1se)

  # calculated MSPEs for 5 models for each v fold
  MSPEs.cv[v, 1] <- mean((air.data2[folds==v, "Ozone"] - pred.ls.cv)^2)
  MSPEs.cv[v, 2] <- mean((air.data2[folds==v, "Ozone"] - pred.step.cv)^2)
  MSPEs.cv[v, 3] <- mean((air.data2[folds==v, "Ozone"] - pred.ridge.cv)^2)
  MSPEs.cv[v, 4] <- mean((air.data2[folds==v, "Ozone"] - pred.lasso.min.cv)^2)
  MSPEs.cv[v, 5] <- mean((air.data2[folds==v, "Ozone"] - pred.lasso.1se.cv)^2)
}
```

```
## Start:  AIC=702.68
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + TWrat    1     60815  53471 632.08
## + Temp     1     56213  58073 640.25
## + Wind     1     43966  70320 659.19
## + TWcp     1     25493  88793 682.29
## + Solar.R  1     12398 101888 695.90
## <none>                 114286 702.68
##
## Step:  AIC=632.08
## Ozone ~ TWrat
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1     11732  41738 612.15
## + Solar.R  1      5995  47476 624.90
## <none>                  53471 632.08
## + TWcp     1       860  52610 635.06
## + Wind     1       430  53041 635.87
## - TWrat    1     60815 114286 702.68
##
## Step:  AIC=612.15
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2335.5 39403 611.04
## <none>                 41738 612.15
## + TWcp     1     329.7 41409 615.96
## + Wind     1     146.4 41592 616.39
## - Temp     1   11732.4 53471 632.08
## - TWrat    1   16334.3 58073 640.25
##
## Step:  AIC=611.04
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                 39403 611.04
## - Solar.R  1    2335.5 41738 612.15
## + TWcp     1     461.4 38942 614.47
## + Wind     1     199.8 39203 615.13
## - Temp     1    8073.2 47476 624.90
## - TWrat    1   17102.6 56506 642.14
## Start:  AIC=703.03
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + TWrat    1     59974  47975 626.54
## + Temp     1     51641  56308 642.55
## + Wind     1     38512  69437 663.51
## + TWcp     1     20036  87913 687.10
## + Solar.R  1     12164  95785 695.68
## <none>                 107949 703.03
```

```
##
## Step:  AIC=626.54
## Ozone ~ TWrat
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1      8751  39224 611.00
## + Solar.R  1      5464  42511 619.05
## + TWcp     1      3069  44906 624.53
## <none>                  47975 626.54
## + Wind     1       178  47797 630.77
## - TWrat    1     59974 107949 703.03
##
## Step:  AIC=611
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2844.4 36380 608.08
## <none>                  39224 611.00
## + Wind     1      40.7 39183 615.50
## + TWcp     1       4.1 39220 615.60
## - Temp     1    8750.6 47975 626.54
## - TWrat    1   17084.4 56308 642.55
##
## Step:  AIC=608.08
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                  36380 608.08
## - Solar.R  1    2844.4 39224 611.00
## + Wind     1      10.6 36369 612.66
## + TWcp     1       4.5 36375 612.67
## - Temp     1    6131.2 42511 619.05
## - TWrat    1   17501.0 53881 642.75
## Start:  AIC=697.92
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + TWrat    1     52346  50221 631.11
## + Temp     1     51128  51439 633.51
## + Wind     1     35862  66705 659.50
## + TWcp     1     18152  84415 683.04
## + Solar.R  1     13264  89303 688.67
## <none>                 102567 697.92
##
## Step:  AIC=631.11
## Ozone ~ TWrat
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1     12215  38006 607.85
## + Solar.R  1      5921  44300 623.17
## <none>                  50221 631.11
## + TWcp     1      1695  48526 632.28
## + Wind     1       155  50066 635.41
## - TWrat    1     52346 102567 697.92
```

```
##
## Step:  AIC=607.85
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2256.7 35749 606.33
## <none>                 38006 607.85
## + TWcp     1     111.4 37894 612.16
## + Wind     1      28.3 37978 612.38
## - Temp     1   12215.3 50221 631.11
## - TWrat    1   13432.6 51439 633.51
##
## Step:  AIC=606.33
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                 35749 606.33
## - Solar.R  1    2256.7 38006 607.85
## + TWcp     1     178.1 35571 610.44
## + Wind     1      50.5 35699 610.80
## - Temp     1    8550.6 44300 623.17
## - TWrat    1   13786.9 49536 634.34
## Start:  AIC=700.65
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + TWrat    1     54886  50521 631.71
## + Temp     1     47981  57426 644.52
## + Wind     1     37187  68220 661.74
## + TWcp     1     19016  86390 685.36
## + Solar.R  1     14196  91210 690.79
## <none>                 105407 700.65
##
## Step:  AIC=631.71
## Ozone ~ TWrat
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1     10331  40190 613.44
## + Solar.R  1      6711  43810 622.06
## <none>                  50521 631.71
## + TWcp     1      1602  48919 633.09
## + Wind     1        81  50440 636.15
## - TWrat    1     54886 105407 700.65
##
## Step:  AIC=613.44
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    3642.8 36548 608.54
## <none>                 40190 613.44
## + TWcp     1     180.4 40010 617.59
## + Wind     1      85.2 40105 617.83
## - Temp     1   10330.5 50521 631.71
## - TWrat    1   17235.4 57426 644.52
```

```
##
## Step:  AIC=608.54
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                 36548 608.54
## + TWcp     1     242.7 36305 612.48
## + Wind     1     123.9 36424 612.81
## - Solar.R  1    3642.8 40190 613.44
## - Temp     1    7262.2 43810 622.06
## - TWrat    1   17199.5 53747 642.50
## Start:  AIC=690.95
## Ozone ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + Temp     1     51807 43862 617.57
## + TWrat    1     48459 47210 624.93
## + Wind     1     35136 60533 649.79
## + TWcp     1     17692 77977 675.11
## + Solar.R  1     12912 82757 681.06
## <none>                 95669 690.95
##
## Step:  AIC=617.57
## Ozone ~ Temp
##
##           Df Sum of Sq   RSS    AIC
## + TWrat    1     10632 33230 594.42
## + TWcp     1      9593 34269 597.50
## + Wind     1      7685 36176 602.91
## <none>                 43862 617.57
## + Solar.R  1      1638 42224 618.37
## - Temp     1     51807 95669 690.95
##
## Step:  AIC=594.42
## Ozone ~ Temp + TWrat
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2041.6 31188 592.68
## <none>                 33230 594.42
## + TWcp     1     787.7 32442 596.63
## + Wind     1     400.5 32829 597.81
## - TWrat    1   10631.5 43862 617.57
## - Temp     1   13980.0 47210 624.93
##
## Step:  AIC=592.68
## Ozone ~ Temp + TWrat + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                 31188 592.68
## - Solar.R  1    2041.6 33230 594.42
## + TWcp     1     828.4 30360 594.60
## + Wind     1     402.1 30786 595.99
## - Temp     1    9954.5 41143 615.78
## - TWrat    1   11035.4 42224 618.37
```

```
## Start:  AIC=710.17
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + TWrat    1     60836  55107 640.40
## + Temp     1     57604  58339 646.10
## + Wind     1     43129  72814 668.26
## + TWcp     1     23462  92481 692.17
## + Solar.R  1     13584 102359 702.32
## <none>                 115943 710.17
##
## Step:  AIC=640.4
## Ozone ~ TWrat
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1     13828  41279 616.11
## + Solar.R  1      6051  49056 633.37
## <none>                  55107 640.40
## + TWcp     1      1218  53889 642.77
## + Wind     1       378  54729 644.31
## - TWrat    1     60836 115943 710.17
##
## Step:  AIC=616.11
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2736.2 38543 613.86
## <none>                 41279 616.11
## + TWcp     1     519.9 40759 619.45
## + Wind     1     296.5 40983 619.99
## - Temp     1   13827.9 55107 640.40
## - TWrat    1   17059.7 58339 646.10
##
## Step:  AIC=613.86
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                 38543 613.86
## - Solar.R  1    2736.2 41279 616.11
## + TWcp     1     502.6 38041 617.15
## + Wind     1     257.0 38286 617.79
## - Temp     1   10513.2 49056 633.37
## - TWrat    1   17085.3 55628 645.94
## Start:  AIC=710.48
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + TWrat    1     67132  49163 628.98
## + Temp     1     55714  60581 649.87
## + Wind     1     46340  69955 664.25
## + TWcp     1     23729  92565 692.26
## + Solar.R  1     15806 100488 700.47
## <none>                 116294 710.48
##
```

```
## Step:  AIC=628.98
## Ozone ~ TWrat
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1     10889  38273 608.55
## + Solar.R  1      5659  43504 621.36
## <none>                   49163 628.98
## + TWcp     1      1535  47628 630.42
## + Wind     1       164  48999 633.25
## - TWrat    1     67132 116294 710.48
##
## Step:  AIC=608.55
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2463.7 35810 606.50
## <none>                  38273 608.55
## + TWcp     1     354.0 37919 612.23
## + Wind     1     183.2 38090 612.67
## - Temp     1   10889.4 49163 628.98
## - TWrat    1   22307.2 60581 649.87
##
## Step:  AIC=606.5
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                  35810 606.50
## - Solar.R  1    2463.7 38273 608.55
## + TWcp     1     414.9 35395 609.94
## + Wind     1     203.8 35606 610.54
## - Temp     1    7694.2 43504 621.36
## - TWrat    1   22096.6 57906 649.96
## Start:  AIC=710.68
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + TWrat    1     63520  53013 636.52
## + Temp     1     56525  60007 648.92
## + Wind     1     48353  68180 661.68
## + TWcp     1     29046  87487 686.62
## + Solar.R  1     13418 103115 703.05
## <none>                  116533 710.68
##
## Step:  AIC=636.52
## Ozone ~ TWrat
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1     10768  42245 618.42
## + Solar.R  1      4493  48519 632.27
## <none>                   53013 636.52
## + Wind     1       936  52076 639.35
## + TWcp     1       348  52665 640.47
## - TWrat    1     63520 116533 710.68
##
```

```
## Step:  AIC=618.42
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2356.9 39888 617.29
## <none>                 42245 618.42
## + TWcp     1     837.1 41407 621.03
## + Wind     1     508.4 41736 621.82
## - Temp     1   10768.2 53013 636.52
## - TWrat    1   17762.9 60007 648.92
##
## Step:  AIC=617.29
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                 39888 617.29
## - Solar.R  1    2356.9 42245 618.42
## + TWcp     1     730.7 39157 620.04
## + Wind     1     405.2 39483 620.87
## - Temp     1    8631.7 48519 632.27
## - TWrat    1   17159.3 57047 648.46
## Start:  AIC=711.46
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + TWrat    1     61399  56049 642.09
## + Temp     1     57702  59745 648.48
## + Wind     1     45089  72359 667.63
## + TWcp     1     25377  92070 691.72
## + Solar.R  1     14633 102814 702.76
## <none>                 117448 711.46
##
## Step:  AIC=642.09
## Ozone ~ TWrat
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1     12664  43385 621.08
## + Solar.R  1      7059  48989 633.23
## <none>                  56049 642.09
## + TWcp     1       914  55135 645.05
## + Wind     1       523  55526 645.76
## - TWrat    1     61399 117448 711.46
##
## Step:  AIC=621.08
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    3464.7 39920 617.37
## <none>                 43385 621.08
## + TWcp     1     517.6 42867 624.49
## + Wind     1     293.0 43092 625.01
## - Temp     1   12664.1 56049 642.09
## - TWrat    1   16360.6 59745 648.48
##
```

```
## Step:  AIC=617.37
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                 39920 617.37
## + TWcp     1     635.7 39284 620.37
## - Solar.R  1    3464.7 43385 621.08
## + Wind     1     347.6 39572 621.10
## - Temp     1    9069.4 48989 633.23
## - TWrat    1   16703.5 56624 647.72
## Start:  AIC=698.44
## Ozone ~ 1
##
##           Df Sum of Sq    RSS    AIC
## + TWrat    1     52732  50375 631.42
## + Temp     1     48457  54650 639.56
## + Wind     1     38890  64217 655.70
## + TWcp     1     22477  80630 678.46
## + Solar.R  1     10598  92509 692.20
## <none>                 103107 698.44
##
## Step:  AIC=631.42
## Ozone ~ TWrat
##
##           Df Sum of Sq    RSS    AIC
## + Temp     1     10214  40161 613.36
## + Solar.R  1      5153  45222 625.23
## <none>                  50375 631.42
## + Wind     1       556  49818 634.91
## + TWcp     1       500  49875 635.03
## - TWrat    1     52732 103107 698.44
##
## Step:  AIC=613.36
## Ozone ~ TWrat + Temp
##
##           Df Sum of Sq   RSS    AIC
## + Solar.R  1    2506.3 37655 611.53
## <none>                 40161 613.36
## + TWcp     1     692.5 39469 616.23
## + Wind     1     461.0 39700 616.81
## - Temp     1   10213.6 50375 631.42
## - TWrat    1   14489.1 54650 639.56
##
## Step:  AIC=611.53
## Ozone ~ TWrat + Temp + Solar.R
##
##           Df Sum of Sq   RSS    AIC
## <none>                 37655 611.53
## - Solar.R  1    2506.3 40161 613.36
## + TWcp     1     753.1 36902 614.11
## + Wind     1     487.5 37167 614.83
## - Temp     1    7567.3 45222 625.23
## - TWrat    1   14825.1 52480 640.12
```

3(d)

```
# get the MSPEs for each 10 folds
MSPEs.cv[, 3:5]
```

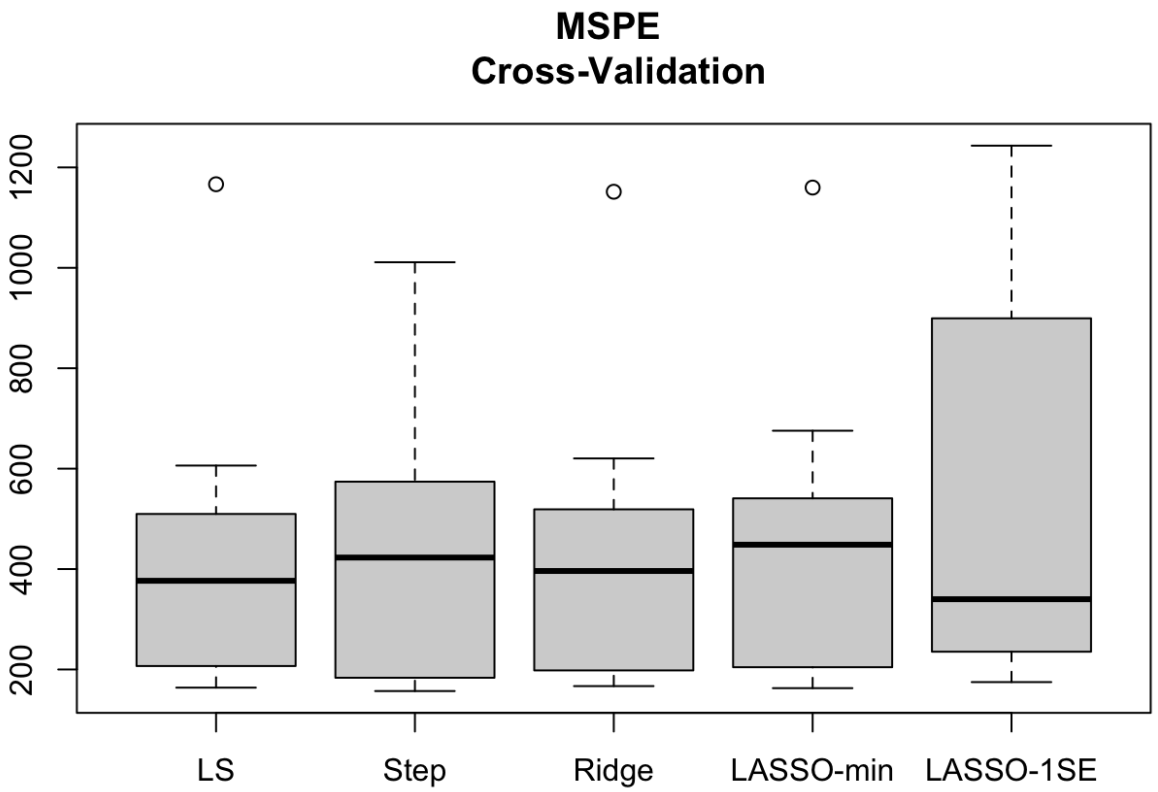
```
##           Ridge LASSO-min LASSO-1SE
##  [1,]  198.1305  204.4936  277.7900
##  [2,]  518.9598  499.7676  402.1023
##  [3,]  516.3355  675.5505 1160.4274
##  [4,]  427.7802  517.4057  899.4312
##  [5,] 1151.5464 1159.8831 1243.3303
##  [6,]  278.8979  285.3114  233.5950
##  [7,]  620.3692  540.8619  235.3842
##  [8,]  183.9873  187.7806  250.7959
##  [9,]  166.8732  162.7871  174.8679
## [10,]  364.5256  397.2974  749.3817
```

```
# get the mean MSPEs
MSPEcv <- apply(X = MSPEs.cv, MARGIN = 2, FUN = mean)
MSPEcv[3:5]
```

```
##     Ridge LASSO-min LASSO-1SE
##  442.7406  463.1139  562.7106
```

3(e)

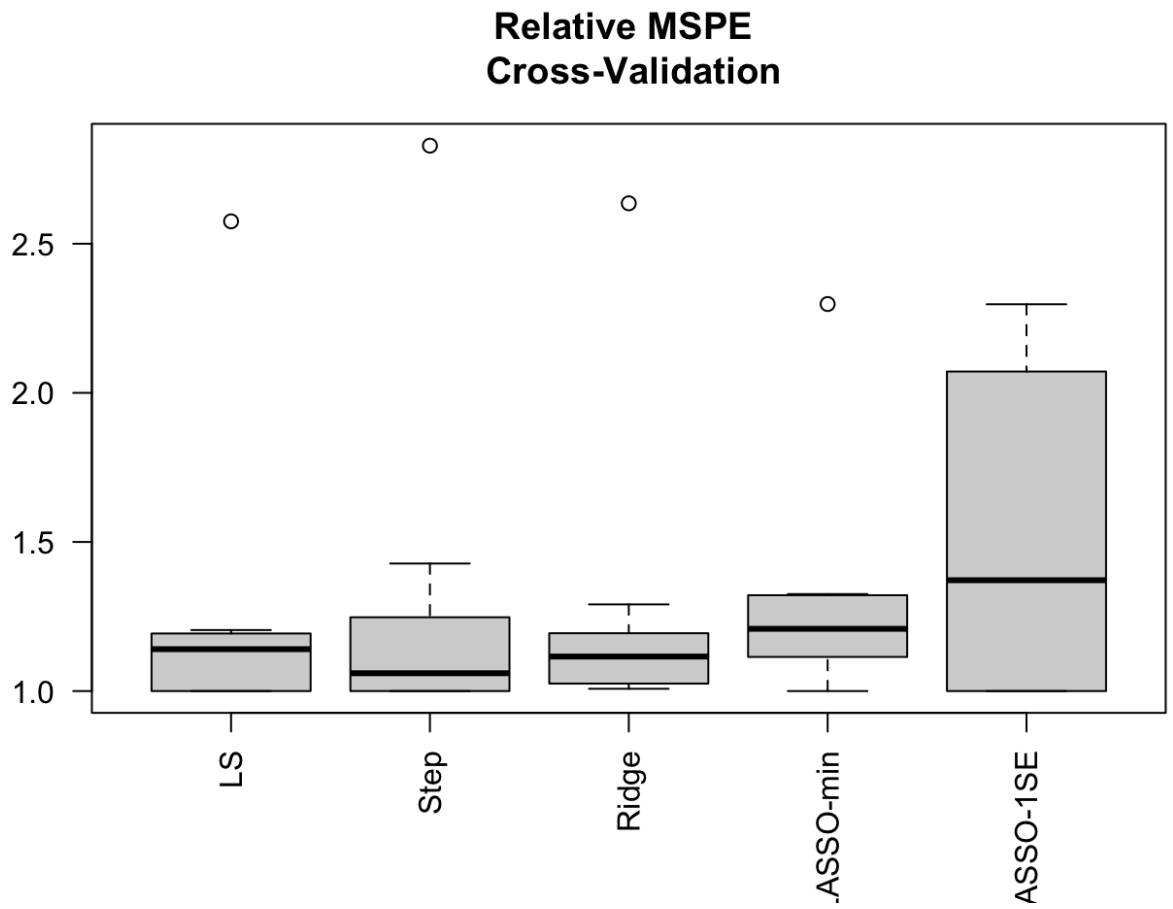
```
# create boxplots for MSPEs
boxplot(MSPEs.cv, main = "MSPE \n Cross-Validation")
```



LS (Least Squares) and Ridge perform the best, LASSO-1SE performs the worst, and Stepwise Regression (Step) and LASSO-min are worse than LS and Ridge.

3(f)

```
# create boxplots for relative MSPEs
low.cv <- apply(MSPEs.cv, 1, min)
boxplot(MSPEs.cv / low.cv,
        las = 2,
        main = "Relative MSPE \n Cross-Validation"
)
```



2. Problem Set 8, Applications (OZONE DATA)

```
library(pls)
```

```
##
## Attaching package: 'pls'
```

```
## The following object is masked from 'package:stats':
##
##     loadings
```

```
# create matrix for number of optimal PCs
opt.pc.cv <- matrix(NA, nrow = V, ncol = 1)
colnames(opt.pc.cv) <- c("optimal number of PCs")
# create matrix for MSPEs for pls
MSPEs.pls.cv <- matrix(NA, nrow = V, ncol = 1)
colnames(MSPEs.pls.cv) <- c("PLS")
# run cross-validation in for-loop
for(v in 1:V) {
  # fit pls
  model.pls <- plsr(Ozone ~ ., data=air.data2[folds!=v, ], ncomp=5, validation="CV")
  CVpls <- model.pls$validation
  pls.comps <- CVpls$PRESS

  # get the lowest RMSEP
  opt.comps <- which.min(pls.comps)
  opt.pc.cv[v] <- opt.comps

  ## 1(c)
  # predict Ozone using the fitted models and number of components on fold == v
  pred.pls <- predict(model.pls, n_comp=opt.comps,newdata=air.data2[folds==v, ])

  # calculated MSPEs for each v fold
  MSPEs.pls.cv[v] <- mean((air.data2[folds==v, "Ozone"] - pred.pls)^2)
}
```

1(a)

```
# get the number of optimal PCs in each folds
opt.pc.cv
```

```
##       optimal number of PCs
##  [1,]                     5
##  [2,]                     3
##  [3,]                     5
##  [4,]                     3
##  [5,]                     5
##  [6,]                     3
##  [7,]                     3
##  [8,]                     5
##  [9,]                     4
## [10,]                     3
```

1(c)

```
# get the MSPEs for each 10 folds
MSPEs.pls.cv
```

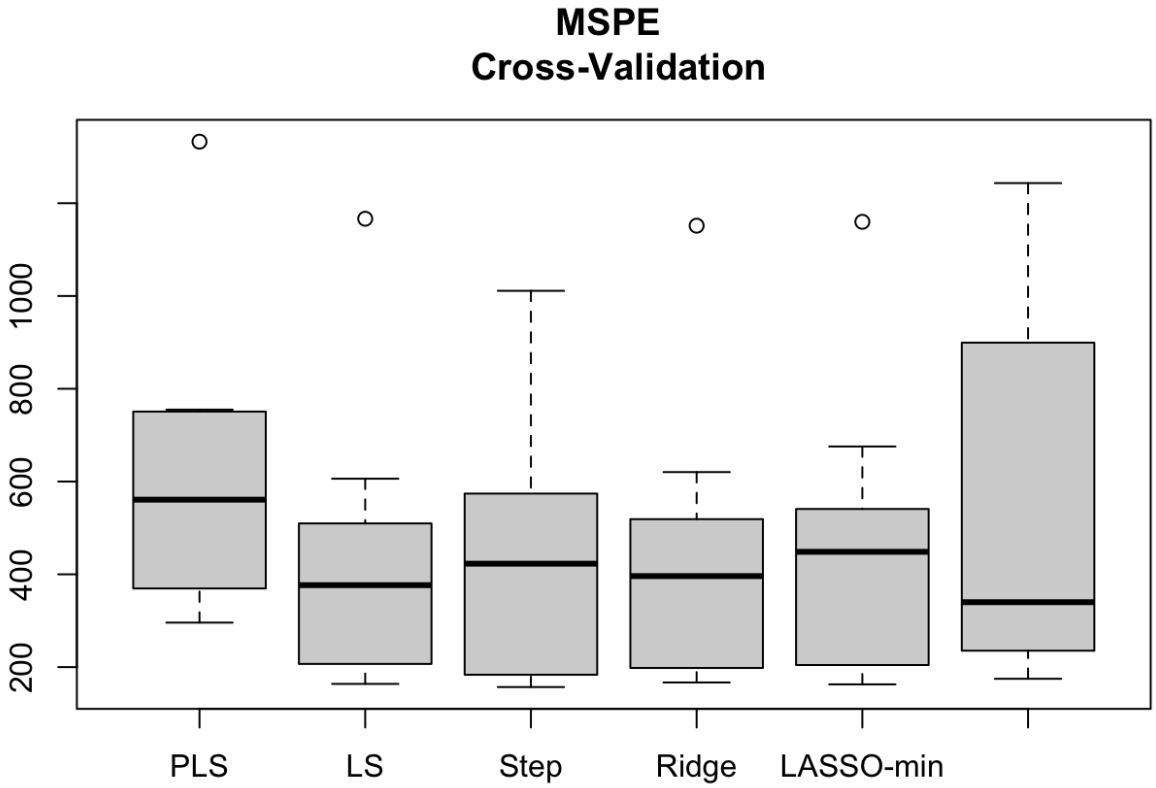
```
##             PLS
##  [1,]  369.5422
##  [2,]  572.0789
##  [3,]  755.0097
##  [4,]  628.8825
##  [5,] 1332.6429
##  [6,]  295.9189
##  [7,]  549.8411
##  [8,]  381.0518
##  [9,]  300.2433
## [10,]  750.8445
```

```
# get the mean MSPEs
(MSPEpls <- apply(X = MSPEs.pls.cv, MARGIN = 2, FUN = mean))
```

```
##      PLS
## 593.6056
```

1(d)

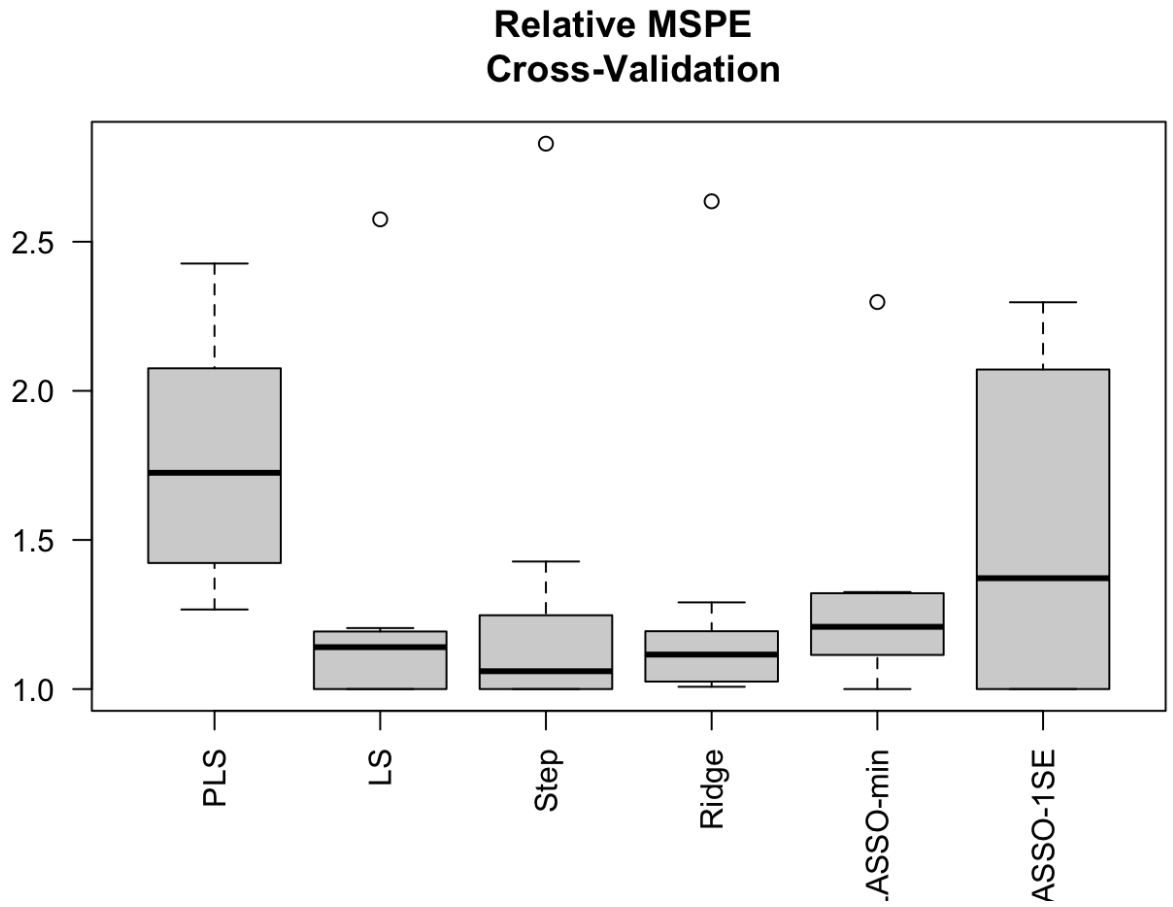
```
# create boxplots for MSPEs
boxplot(cbind(MSPEs.pls.cv, MSPEs.cv), main = "MSPE \n Cross-Validation")
```



PLS has the highest MSPE among all the models; however, it exhibits lower variability than LASSO-1SE.

1(e)

```
# create boxplots for relative MSPEs
low.cv <- apply(cbind(MSPEs.pls.cv, MSPEs.cv), 1, min)
boxplot(cbind(MSPEs.pls.cv, MSPEs.cv) / low.cv,
        las = 2,
        main = "Relative MSPE \n Cross-Validation"
)
```



3. Problem Set 9, Concepts

1(a)

β_0 represents the “baseline” or the first region/interval of X, and it measures the mean value of Y in the first region/interval of X.

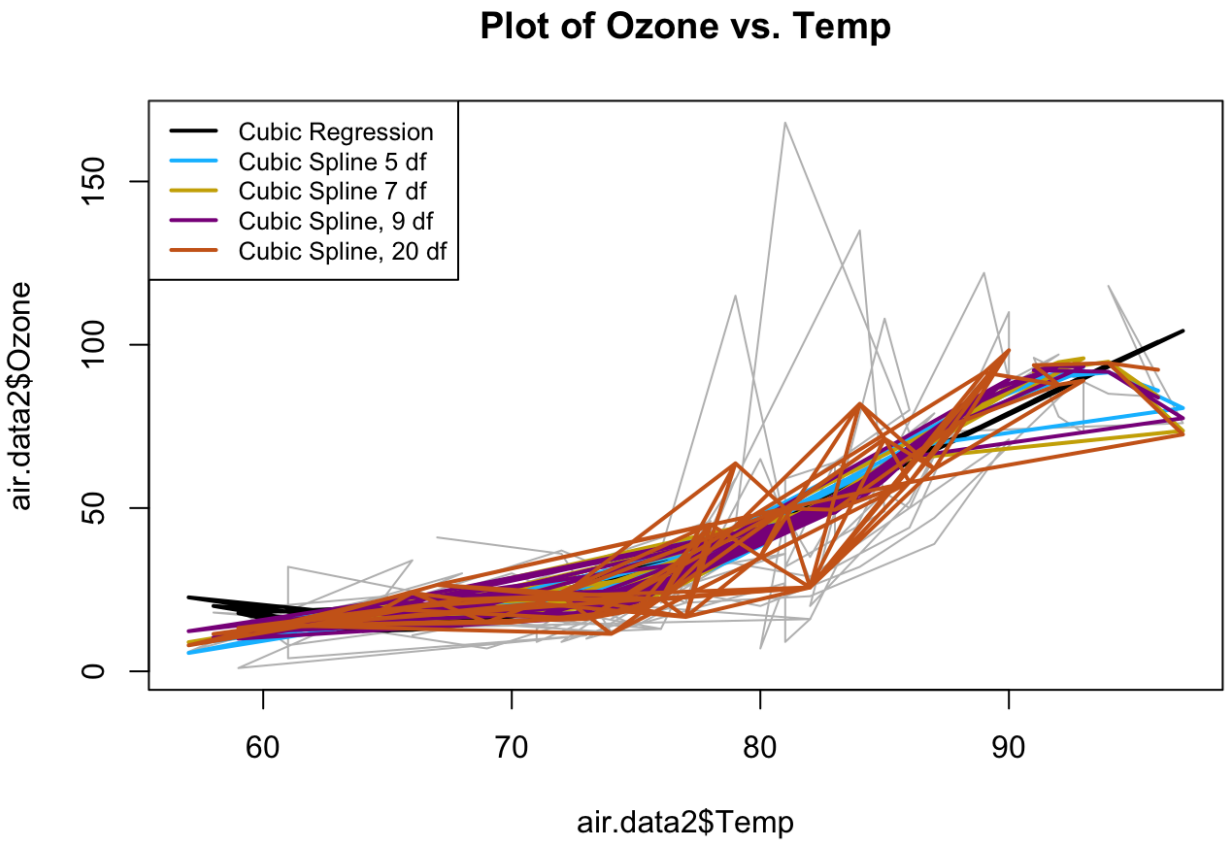
1(b)

β_K represents the last region/interval of X, and it measures the mean value of Y in the last region/interval of X relative to the baseline B_0. β_K quantifies the difference in the mean value of Y when the last region represented by C_k(X) is considered, compared to the first region, which serves as the baseline (where all other indicator variables are zero).

4. Problem Set 10, Applications

1(a)

```
library(splines)
plot(
  x = air.data2$Temp, y = air.data2$Ozone, type = "l", col = "gray",
  main = "Plot of Ozone vs. Temp"
)
legend(
    "topleft", legend = c(
    "Cubic Regression", "Cubic Spline 5 df",
    "Cubic Spline 7 df", "Cubic Spline, 9 df",
    "Cubic Spline, 20 df"),
  lty = "solid", col = colors()[c(24, 121, 145, 84, 55)],
  lwd = 2, cex=0.8
)
# add cubic polynomial to plot (3 df model)
poly3 <- lm(data = air.data2, Ozone ~ poly(x=Temp, degree = 3))
lines(x = air.data2$Temp, y = predict(poly3, newdata = air.data2), col = colors()[24], lwd
= 2)
# 5 DF spline
cub.spl.5 <- lm(data = air.data2, Ozone ~ bs(Temp, df = 5))
lines(x = air.data2$Temp, y = predict(cub.spl.5, newdata = air.data2),
      col = colors()[121], lwd = 2)
# 7 DF spline
cub.spl.7 <- lm(data = air.data2, Ozone ~ bs(Temp, df = 7))
lines(x = air.data2$Temp, y = predict(cub.spl.7, newdata = air.data2),
      col = colors()[145], lwd = 2)
# 9 DF spline
cub.spl.9 <- lm(data = air.data2, Ozone ~ bs(Temp, df = 9))
lines(x = air.data2$Temp, y = predict(cub.spl.9, newdata = air.data2),
      col = colors()[84], lwd = 2)
# 20 DF spline
cub.spl.20 <- lm(data = air.data2, Ozone ~ bs(Temp, df = 20))
lines(x = air.data2$Temp, y = predict(cub.spl.20, newdata = air.data2),
      col = colors()[55], lwd = 2)
```



1(b)

Cubic Regression

1(c)

Functions with higher degrees of freedom, e.g., cubic splines with 20 DF, have a tendency to overfit. Overfitting can be observed when the curve fits the data points very closely, showing a lot of variability.

1(d)

Cubic splines with 7 DF because it strikes a balance between capturing the underlying trend in the data while avoiding excessive overfitting.