Lab 3

Kevin Zhai

11/13/2016

1.

1a. The two explanatory variables I would remove from the model are INDUS and AGE. Based on p-values alone, it is clear that these are the only two variables that are insignificant in the full model. They have extremely high p-values, as opposed to the other predictors which have very low p-values.

```
reg.picked <- lm(MEDV~CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO + B + LSTAT,data=boston)
summary(reg.picked)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT, data = boston)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -15.5984  -2.7386  -0.5046   1.7273  26.2373
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  36.341145   5.067492   7.171 2.73e-12 ***
## CRIM         -0.108413   0.032779  -3.307 0.001010 **
## ZN            0.045845   0.013523   3.390 0.000754 ***
## CHAS          2.718716   0.854240   3.183 0.001551 **
## NOX         -17.376023   3.535243  -4.915 1.21e-06 ***
## RM            3.801579   0.406316   9.356  < 2e-16 ***
## DIS          -1.492711   0.185731  -8.037 6.84e-15 ***
## RAD           0.299608   0.063402   4.726 3.00e-06 ***
## TAX          -0.011778   0.003372  -3.493 0.000521 ***
## PTRATIO      -0.946525   0.129066  -7.334 9.24e-13 ***
## B             0.009291   0.002674   3.475 0.000557 ***
## LSTAT        -0.522553   0.047424 -11.019  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.736 on 494 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7348
## F-statistic: 128.2 on 11 and 494 DF,  p-value: < 2.2e-16
```

1b. The model improves slightly, with the adjusted r^2 value improving marginally. However, the F-statistic increases significantly, meaning the model has a better explanatory power.

```
anova(reg)
```

```
## Analysis of Variance Table
##
## Response: MEDV
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## CRIM        1  6440.8  6440.8 286.0300 < 2.2e-16 ***
## ZN          1  3554.3  3554.3 157.8452 < 2.2e-16 ***
```

```
## INDUS       1  2551.2  2551.2 113.2984 < 2.2e-16 ***
## CHAS        1  1529.8  1529.8  67.9393 1.543e-15 ***
## NOX         1    76.2    76.2   3.3861 0.0663505 .
## RM          1 10938.1 10938.1 485.7530 < 2.2e-16 ***
## AGE         1    90.3    90.3   4.0087 0.0458137 *
## DIS         1  1779.5  1779.5  79.0262 < 2.2e-16 ***
## RAD         1    34.1    34.1   1.5159 0.2188325
## TAX         1   329.6   329.6  14.6352 0.0001472 ***
## PTRATIO     1  1309.3  1309.3  58.1454 1.266e-13 ***
## B           1   593.3   593.3  26.3496 4.109e-07 ***
## LSTAT       1  2410.8  2410.8 107.0634 < 2.2e-16 ***
## Residuals 492 11078.8    22.5
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(reg.picked)
```

```
## Analysis of Variance Table
##
## Response: MEDV
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## CRIM        1  6440.8  6440.8 287.1259 < 2.2e-16 ***
## ZN          1  3554.3  3554.3 158.4500 < 2.2e-16 ***
## CHAS        1  1233.8  1233.8  55.0016 5.282e-13 ***
## NOX         1  1592.4  1592.4  70.9878 3.947e-16 ***
## RM          1 12091.0 12091.0 539.0070 < 2.2e-16 ***
## DIS         1  1122.0  1122.0  50.0186 5.234e-12 ***
## RAD         1    97.5    97.5   4.3478   0.03757 *
## TAX         1   669.3   669.3  29.8380 7.456e-08 ***
## PTRATIO     1  1519.7  1519.7  67.7494 1.666e-15 ***
## B           1   590.6   590.6  26.3273 4.149e-07 ***
## LSTAT       1  2723.5  2723.5 121.4111 < 2.2e-16 ***
## Residuals 494 11081.4    22.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
mse.reg <- 11078.8 / 492
mse.reg.picked <- 11081.4 / 494
mae.reg <- sum(abs(reg$residuals)) / 492
mae.reg.picked <- sum(abs(reg.picked$residuals)) / 494

print(mse.reg)
```

```
## [1] 22.51789
```

```r
print(mse.reg.picked)
```

```
## [1] 22.43198
```

```r
print(mae.reg)
```

```
## [1] 3.363936
```

```r
print(mae.reg.picked)
```

```
## [1] 3.351519
```

1c. In both cases, the MSE and the MAE are lower for the model reg.picked, so I would choose that model.

```r
library(MASS)
reg.step = step(object=reg, direction='both')
```

```
## Start:  AIC=1589.64
## MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## - AGE      1      0.06 11079 1587.7
## - INDUS    1      2.52 11081 1587.8
## <none>                 11079 1589.6
## - CHAS     1    218.97 11298 1597.5
## - TAX      1    242.26 11321 1598.6
## - CRIM     1    243.22 11322 1598.6
## - ZN       1    257.49 11336 1599.3
## - B        1    270.63 11349 1599.8
## - RAD      1    479.15 11558 1609.1
## - NOX      1    487.16 11566 1609.4
## - PTRATIO  1   1194.23 12273 1639.4
## - DIS      1   1232.41 12311 1641.0
## - RM       1   1871.32 12950 1666.6
## - LSTAT    1   2410.84 13490 1687.3
##
## Step:  AIC=1587.65
## MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + DIS + RAD + TAX +
##     PTRATIO + B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## - INDUS    1      2.52 11081 1585.8
## <none>                 11079 1587.7
## + AGE      1      0.06 11079 1589.6
## - CHAS     1    219.91 11299 1595.6
## - TAX      1    242.24 11321 1596.6
## - CRIM     1    243.20 11322 1596.6
## - ZN       1    260.32 11339 1597.4
## - B        1    272.26 11351 1597.9
## - RAD      1    481.09 11560 1607.2
## - NOX      1    520.87 11600 1608.9
## - PTRATIO  1   1200.23 12279 1637.7
## - DIS      1   1352.26 12431 1643.9
## - RM       1   1959.55 13038 1668.0
## - LSTAT    1   2718.88 13798 1696.7
##
## Step:  AIC=1585.76
## MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX + PTRATIO +
##     B + LSTAT
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1585.8
## + INDUS    1      2.52 11079 1587.7
## + AGE      1      0.06 11081 1587.8
## - CHAS     1    227.21 11309 1594.0
## - CRIM     1    245.37 11327 1594.8
## - ZN       1    257.82 11339 1595.4
```

```
## - B        1    270.82 11352 1596.0
## - TAX      1    273.62 11355 1596.1
## - RAD      1    500.92 11582 1606.1
## - NOX      1    541.91 11623 1607.9
## - PTRATIO  1   1206.45 12288 1636.0
## - DIS      1   1448.94 12530 1645.9
## - RM       1   1963.66 13045 1666.3
## - LSTAT    1   2723.48 13805 1695.0
```

```r
anova(reg.step)
```

```
## Analysis of Variance Table
##
## Response: MEDV
##            Df  Sum Sq Mean Sq  F value    Pr(>F)
## CRIM        1  6440.8  6440.8 287.1259 < 2.2e-16 ***
## ZN          1  3554.3  3554.3 158.4500 < 2.2e-16 ***
## CHAS        1  1233.8  1233.8  55.0016 5.282e-13 ***
## NOX         1  1592.4  1592.4  70.9878 3.947e-16 ***
## RM          1 12091.0 12091.0 539.0070 < 2.2e-16 ***
## DIS         1  1122.0  1122.0  50.0186 5.234e-12 ***
## RAD         1    97.5    97.5   4.3478   0.03757 *
## TAX         1   669.3   669.3  29.8380 7.456e-08 ***
## PTRATIO     1  1519.7  1519.7  67.7494 1.666e-15 ***
## B           1   590.6   590.6  26.3273 4.149e-07 ***
## LSTAT       1  2723.5  2723.5 121.4111 < 2.2e-16 ***
## Residuals 494 11081.4    22.4
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

1d. The model that the stepwise regression picks is the exact same as the model reg.picked, where AGE and INDUS are taken out. Therefore, the SSE is the exact same for both models.

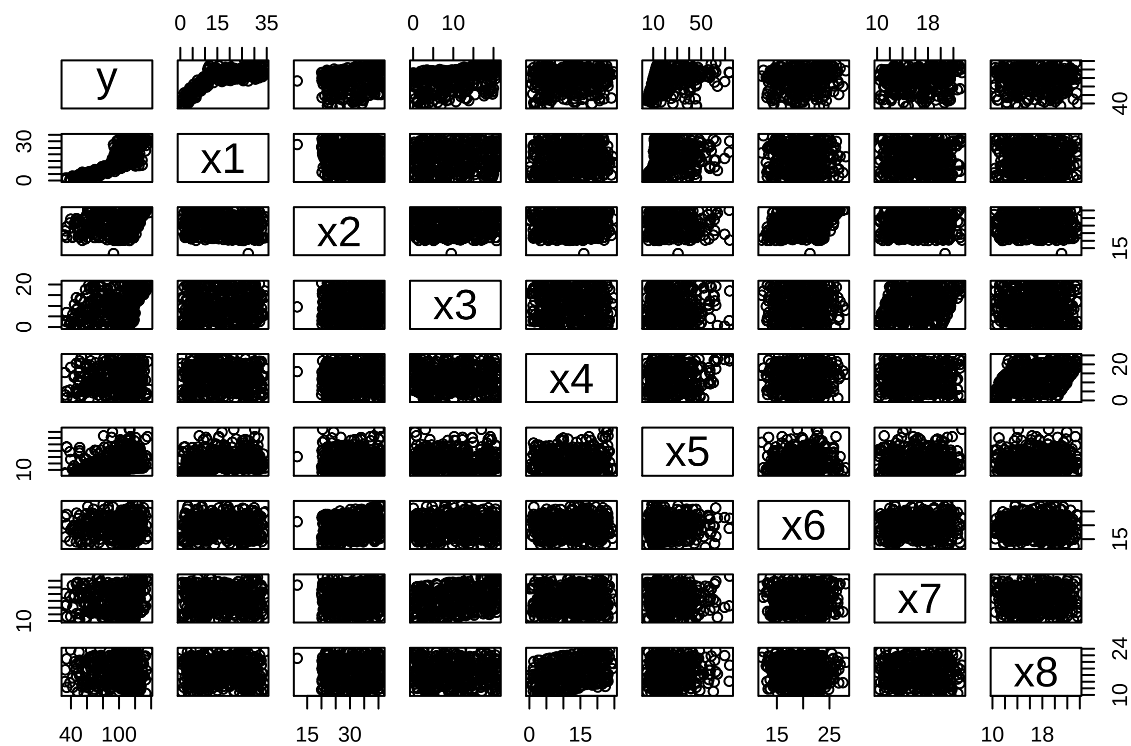
2.

```r
lab <- read.csv(file="labdata.txt",head=TRUE,sep="\t")
labreg <- lm(y~. ,data=lab)
summary(labreg)
```

```
##
## Call:
## lm(formula = y ~ ., data = lab)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.7138  -7.3129  -0.1718   7.4281  23.8909
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 17.58565    5.10223   3.447 0.000629 ***
## x1           1.91936    0.05492  34.951  < 2e-16 ***
## x2           0.89747    0.08389  10.699  < 2e-16 ***
## x3           1.07895    0.08370  12.890  < 2e-16 ***
## x4           0.23834    0.08759   2.721 0.006798 **
## x5           0.10141    0.03725   2.723 0.006766 **
## x6           0.29608    0.15153   1.954 0.051421 .
```

```
## x7          -0.06268    0.15824  -0.396 0.692262
## x8          -0.01515    0.15846  -0.096 0.923860
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 10.01 on 391 degrees of freedom
## Multiple R-squared:  0.8113, Adjusted R-squared:  0.8074
## F-statistic: 210.1 on 8 and 391 DF,  p-value: < 2.2e-16
```

```r
plot(lab)
```

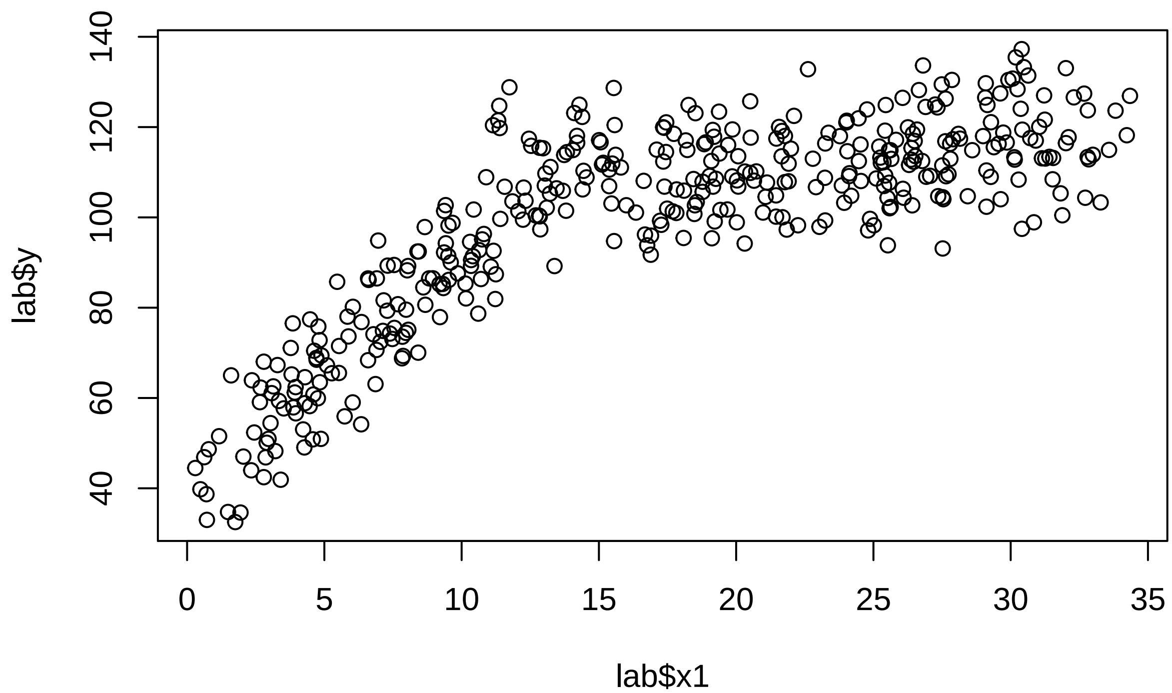



```r
cor(lab)
```

```
##              y          x1          x2           x3           x4
## y  1.000000000  0.80533240  0.19670658  0.357489045  0.102390521
## x1 0.805332395  1.00000000 -0.08474350  0.078455900  0.044997692
## x2 0.196706581 -0.08474350  1.00000000  0.032190768  0.010550315
## x3 0.357489045  0.07845590  0.03219077  1.000000000 -0.023429594
## x4 0.102390521  0.04499769  0.01055031 -0.023429594  1.000000000
## x5 0.215721371  0.19510364 -0.01520499 -0.018719329  0.110846689
## x6 0.083910285 -0.02965772  0.24679088 -0.001009249 -0.001815613
## x7 0.096340682  0.03255074  0.03925291  0.233860531  0.011634563
## x8 0.004553459 -0.01039468 -0.03037249 -0.012508475  0.392708258
##              x5           x6          x7           x8
## y   0.215721371  0.083910285 0.096340682  0.004553459
## x1  0.195103635 -0.029657719 0.032550741 -0.010394675
## x2 -0.015204993  0.246790884 0.039252914 -0.030372494
```

```
## x3 -0.018719329 -0.001009249 0.233860531 -0.012508475
## x4  0.110846689 -0.001815613 0.011634563  0.392708258
## x5  1.000000000  0.051007118 0.004495165  0.052027308
## x6  0.051007118  1.000000000 0.014613630 -0.057075892
## x7  0.004495165  0.014613630 1.000000000  0.025789695
## x8  0.052027308 -0.057075892 0.025789695  1.000000000
```

```r
plot(lab$x1,lab$y)
```



2b. Based on the matrix scatter plot and the pairwise correlations, the relationship between y and x1 seems to be the strongest. I would use x1 as the best predictor of y.

```r
mean(lab$x1)
```

```
## [1] 17.19417
```

```r
library(segmented)
labreg.x1 = lm(y ~ x1, data=lab)
labreg.piece = segmented(labreg.x1, seg.Z = ~x1, psi=17.19)
anova(labreg)
```

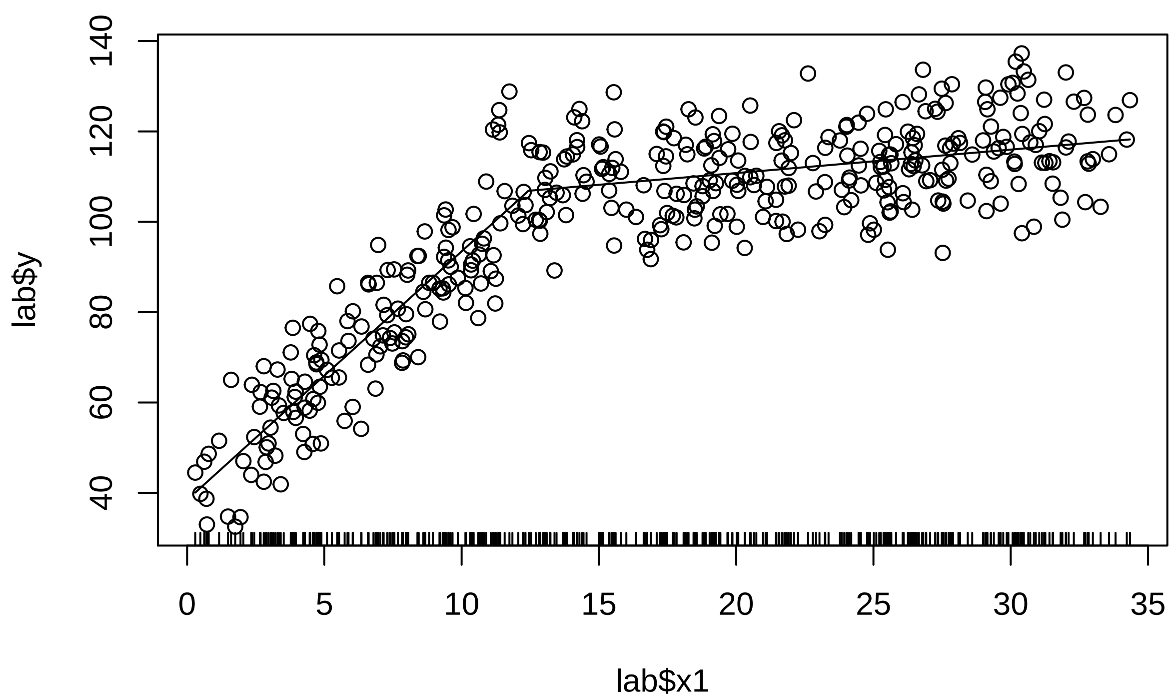
```
## Analysis of Variance Table
##
## Response: y
##             Df Sum Sq Mean Sq   F value    Pr(>F)
## x1           1 134777  134777 1343.7867 < 2.2e-16 ***
## x2           1  14694   14694  146.5037 < 2.2e-16 ***
## x3           1  16884   16884  168.3379 < 2.2e-16 ***
## x4           1   1027    1027   10.2397  0.001487 **
```

```
## x5           1    810     810    8.0783  0.004715 **
## x6           1    385     385    3.8390  0.050784 .
## x7           1     16      16    0.1590  0.690267
## x8           1      1       1    0.0091  0.923860
## Residuals 391  39216     100
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
anova(labreg.piece)
```

```
## Analysis of Variance Table
##
## Response: y
##            Df Sum Sq Mean Sq F value Pr(>F)
## x1          1 134777  134777  1620.8 <2e-16 ***
## U1.x1       1  40104   40104   482.3 <2e-16 ***
## psi1.x1     1      0       0     0.0      1
## Residuals 396  32928      83
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
plot(lab$x1, lab$y)
plot(labreg.piece, add=T)
```



At first glance at the SSE, the SSE of the reg.piece model is much lower. It also has a higher DF, meaning the MSE will be much lower for the piecewise model, so I would choose that one.