Math 445 HW 1

Erick Castillo

2/4/2021

A.

Below is the output for the linear model with medv as the response variable, and all other variables as predictors.

```
##
## Call:
## lm(formula = medv ~ ., data = BostonHousing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -15.595  -2.730  -0.518   1.777  26.199
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
## crim        -1.080e-01  3.286e-02  -3.287 0.001087 **
## zn           4.642e-02  1.373e-02   3.382 0.000778 ***
## indus        2.056e-02  6.150e-02   0.334 0.738288
## chas1        2.687e+00  8.616e-01   3.118 0.001925 **
## nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
## rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
## age          6.922e-04  1.321e-02   0.052 0.958229
## dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
## rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
## tax         -1.233e-02  3.760e-03  -3.280 0.001112 **
## ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
## b            9.312e-03  2.686e-03   3.467 0.000573 ***
## lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.745 on 492 degrees of freedom
## Multiple R-squared:  0.7406, Adjusted R-squared:  0.7338
## F-statistic: 108.1 on 13 and 492 DF,  p-value: < 2.2e-16
```

The coefficients for each of the respective predictors is given under the 'Estimate' column, and the corresponding p-values are under the 'Pr(>|t|)' column.

Observe that twelve predictors are significant at 1%. There are only two predictors that are not significant at all, these include the 'indus' and 'age' variables.

The model can be improved upon. There might be multicollinearity in the model present. Dropping the insignificant variables may improve the fit.

B.

Below is the output from running an AIC model selection with a forward step.

```
## Start:  AIC=2246.51
## medv ~ 1
##
##           Df Sum of Sq   RSS    AIC
## + lstat    1   23243.9 19472 1851.0
## + rm       1   20654.4 22062 1914.2
## + ptratio  1   11014.3 31702 2097.6
## + indus    1    9995.2 32721 2113.6
## + tax      1    9377.3 33339 2123.1
## + nox      1    7800.1 34916 2146.5
## + crim     1    6440.8 36276 2165.8
## + rad      1    6221.1 36495 2168.9
## + age      1    6069.8 36647 2171.0
## + zn       1    5549.7 37167 2178.1
## + b        1    4749.9 37966 2188.9
## + dis      1    2668.2 40048 2215.9
## + chas     1    1312.1 41404 2232.7
## <none>                 42716 2246.5
##
## Step:  AIC=1851.01
## medv ~ lstat
##
##           Df Sum of Sq   RSS    AIC
## + rm       1    4033.1 15439 1735.6
## + ptratio  1    2670.1 16802 1778.4
## + chas     1     786.3 18686 1832.2
## + dis      1     772.4 18700 1832.5
## + age      1     304.3 19168 1845.0
## + tax      1     274.4 19198 1845.8
## + b        1     198.3 19274 1847.8
## + zn       1     160.3 19312 1848.8
## + crim     1     146.9 19325 1849.2
## + indus    1      98.7 19374 1850.4
## <none>                 19472 1851.0
## + rad      1      25.1 19447 1852.4
## + nox      1       4.8 19468 1852.9
##
## Step:  AIC=1735.58
## medv ~ lstat + rm
##
##           Df Sum of Sq   RSS    AIC
## + ptratio  1   1711.32 13728 1678.1
## + chas     1    548.53 14891 1719.3
## + b        1    512.31 14927 1720.5
## + tax      1    425.16 15014 1723.5
## + dis      1    351.15 15088 1725.9
```

```
## + crim     1    311.42 15128 1727.3
## + rad      1    180.45 15259 1731.6
## + indus    1     61.09 15378 1735.6
## <none>                 15439 1735.6
## + zn       1     56.56 15383 1735.7
## + age      1     20.18 15419 1736.9
## + nox      1     14.90 15424 1737.1
##
## Step:  AIC=1678.13
## medv ~ lstat + rm + ptratio
##
##         Df Sum of Sq   RSS    AIC
## + dis    1    499.08 13229 1661.4
## + b      1    389.68 13338 1665.6
## + chas   1    377.96 13350 1666.0
## + crim   1    122.52 13606 1675.6
## + age    1     66.24 13662 1677.7
## <none>               13728 1678.1
## + tax    1     44.36 13684 1678.5
## + nox    1     24.81 13703 1679.2
## + zn     1     14.96 13713 1679.6
## + rad    1      6.07 13722 1679.9
## + indus  1      0.83 13727 1680.1
##
## Step:  AIC=1661.39
## medv ~ lstat + rm + ptratio + dis
##
##         Df Sum of Sq   RSS    AIC
## + nox    1    759.56 12469 1633.5
## + b      1    502.64 12726 1643.8
## + chas   1    267.43 12962 1653.1
## + indus  1    242.65 12986 1654.0
## + tax    1    240.34 12989 1654.1
## + crim   1    233.54 12995 1654.4
## + zn     1    144.81 13084 1657.8
## + age    1     61.36 13168 1661.0
## <none>               13229 1661.4
## + rad    1     22.40 13206 1662.5
##
## Step:  AIC=1633.47
## medv ~ lstat + rm + ptratio + dis + nox
##
##         Df Sum of Sq   RSS    AIC
## + chas   1    328.27 12141 1622.0
## + b      1    311.83 12158 1622.7
## + zn     1    151.71 12318 1629.3
## + crim   1    141.43 12328 1629.7
## + rad    1     53.48 12416 1633.3
## <none>               12469 1633.5
## + indus  1     17.10 12452 1634.8
## + tax    1     10.50 12459 1635.0
## + age    1      0.25 12469 1635.5
##
## Step:  AIC=1621.97
```

```
## medv ~ lstat + rm + ptratio + dis + nox + chas
##
##         Df Sum of Sq   RSS    AIC
## + b      1   272.837 11868 1612.5
## + zn     1   164.406 11977 1617.1
## + crim   1   116.330 12025 1619.1
## + rad    1    58.556 12082 1621.5
## <none>               12141 1622.0
## + indus  1    26.274 12115 1622.9
## + tax    1     4.187 12137 1623.8
## + age    1     2.331 12139 1623.9
##
## Step:  AIC=1612.47
## medv ~ lstat + rm + ptratio + dis + nox + chas + b
##
##         Df Sum of Sq   RSS    AIC
## + zn     1   189.936 11678 1606.3
## + rad    1   144.320 11724 1608.3
## + crim   1    55.633 11813 1612.1
## <none>               11868 1612.5
## + indus  1    15.584 11853 1613.8
## + age    1     9.446 11859 1614.1
## + tax    1     2.703 11866 1614.4
##
## Step:  AIC=1606.31
## medv ~ lstat + rm + ptratio + dis + nox + chas + b + zn
##
##         Df Sum of Sq   RSS    AIC
## + crim   1    94.712 11584 1604.2
## + rad    1    93.614 11585 1604.2
## <none>               11678 1606.3
## + indus  1    16.048 11662 1607.6
## + tax    1     3.952 11674 1608.1
## + age    1     1.491 11677 1608.2
##
## Step:  AIC=1604.19
## medv ~ lstat + rm + ptratio + dis + nox + chas + b + zn + crim
##
##         Df Sum of Sq   RSS    AIC
## + rad    1   228.604 11355 1596.1
## <none>               11584 1604.2
## + indus  1    15.773 11568 1605.5
## + age    1     2.470 11581 1606.1
## + tax    1     1.305 11582 1606.1
##
## Step:  AIC=1596.1
## medv ~ lstat + rm + ptratio + dis + nox + chas + b + zn + crim +
##     rad
##
##         Df Sum of Sq   RSS    AIC
## + tax    1   273.619 11081 1585.8
## <none>               11355 1596.1
## + indus  1    33.894 11321 1596.6
## + age    1     0.096 11355 1598.1
```

```
##
## Step:  AIC=1585.76
## medv ~ lstat + rm + ptratio + dis + nox + chas + b + zn + crim +
##     rad + tax
##
##         Df Sum of Sq   RSS    AIC
## <none>               11081 1585.8
## + indus  1   2.51754 11079 1587.7
## + age    1   0.06271 11081 1587.8
```

The best model from the above output is the model with the lowest AIC, which in this case has predictors lstat, rm, ptratio, dis, nox, chas, b, zn, crim, rad, and tax.

C.

Below is the output from running the stepAIC function in the backwards direction.

```
## 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
## - 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
```

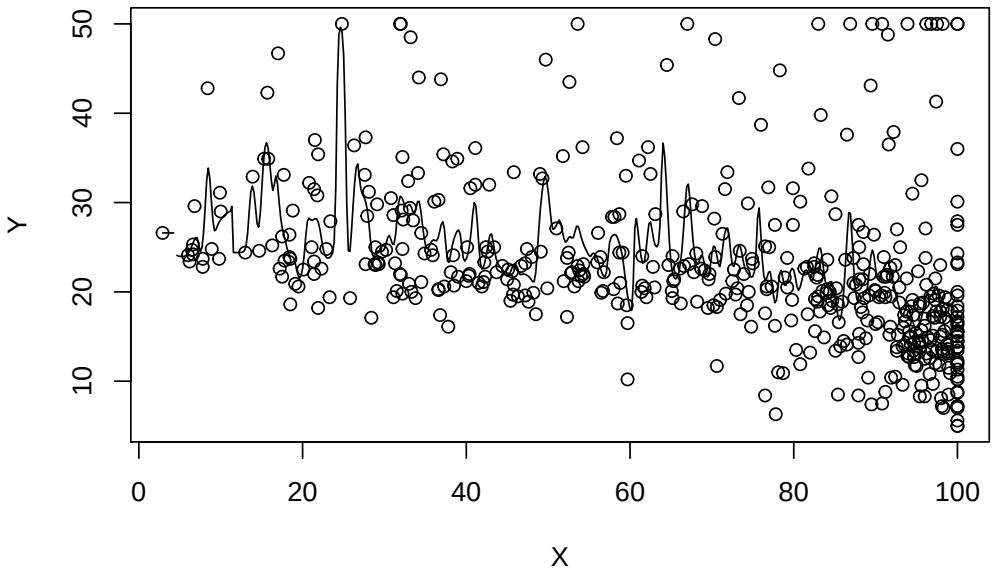
```
## - dis      1   1352.26 12431 1643.9
## - rm       1   1959.55 13038 1668.0
## - lstat    1   2718.88 13798 1696.7
##
## Step:  AIC=1585.76
## medv ~ crim + zn + chas + nox + rm + dis + rad + tax + ptratio +
##     b + lstat
##
##           Df Sum of Sq   RSS    AIC
## <none>                 11081 1585.8
## - 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
```

Once again, the model with the lowest AIC is the best model, which in this case has predictors crim, zn, chas, nox, rm, dis, rad, tax, ptratio, b, and lstat.

Both the forward and backward selection approaches ended up at the same model.

D.

Below is the plot of medv~age in a nonparametric regression. The bandwidth appears to be sufficient in predicting the response variables.



E.

Below is the code to create the dummy column:

```
BostonHousing$medv.dum = ifelse(BostonHousing$medv > 30, 1, 0)
```

Below is the output for the logistic regression

```
##
## Call:
## glm(formula = medv.dum ~ . - medv, family = binomial(), data = BostonHousing)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -3.4016  -0.0854  -0.0155  -0.0005   2.5627
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -2.460276   7.699396  -0.320  0.74932
## crim         0.047581   0.044580   1.067  0.28582
## zn           0.023898   0.012448   1.920  0.05488 .
## indus       -0.175563   0.096314  -1.823  0.06833 .
## chas1        0.182414   0.832590   0.219  0.82658
## nox          1.445091   6.118455   0.236  0.81329
## rm           2.354947   0.580370   4.058 4.96e-05 ***
## age          0.027123   0.016072   1.688  0.09149 .
## dis         -0.305451   0.195522  -1.562  0.11823
```

```
## rad          0.372120   0.116117   3.205  0.00135 **
## tax         -0.010361   0.005117  -2.025  0.04289 *
## ptratio     -0.607343   0.191184  -3.177  0.00149 **
## b            0.004366   0.013629   0.320  0.74871
## lstat       -0.683904   0.138736  -4.930 8.24e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 454.89  on 505  degrees of freedom
## Residual deviance: 117.64  on 492  degrees of freedom
## AIC: 145.64
##
## Number of Fisher Scoring iterations: 10
```

Once again, the 'Estimate' column displays the estimated coefficients for each of the predictors. This time around, eight of the twelve predictors are significant at the 10% level.

As expected, the more significant the variable, the more impact it has in predicting the log odds of a house being above 30 medv. This model can be improved by dropping non-significant predictors to preserve parsimony. Once the non-significant predictors are dropped, a both direction AIC would refine the model.