MSiA 400 Lab Assignment 2

Oct 27, 2014

- Due: 11:59pm Nov 3, 2014
- This is an open book assignment.
- Please submit one report file that includes : short answer, related code and print for each problem if necessary.

Problem 1

Data set *bostonhousing.txt*, created byHarrison and Rubinfeld (1978), concerns housing values in suburbs of Boston. The attributes include

MEDV	Median value of owner-occupied homes in $1000's
CRIM	per capita crime rate by town
ZN	proportion of residential land zoned for lots over 25,000 sq.ft.
INDUS	proportion of non-retail business acres per town
CHAS	Charles River dummy variable (= 1 if tract bounds river; 0 otherwise)
NOX	nitric oxides concentration (parts per 10 million)
RM	average number of rooms per dwelling
AGE	proportion of owner-occupied units built prior to 1940
DIS	weighted distances to five Boston employment centres
RAD	index of accessibility to radial highways
TAX	full-value property-tax rate per $10,000
PTRATIO	pupil-teacher ratio by town
B	$1000(Bk - 0.63)^2$ where Bk is the proportion of blacks by town
LSTAT	% lower status of the population,

in which MEDV is the response variable. The summary of the data set is below.

Name of the data set	bostonhousing
Number of observations	506
Number of attributes	14 (1 response variable and 13 explanatory variables)

problem 1(a)

Build regression model reg and display summary() of the model. Pick two explanatory variables that are least likely to be in the best model, and support your suggestion in one sentence.

Solution
Based on p-values, INDUS and AGE are least likely to be in the best model.

```
> y = bostonhousing[,1];
> x = bostonhousing[,2:14];
> reg = lm(y~., data = x);
> summary(reg);

Call:
lm(formula = y ~ ., data = x)
```

```
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
CHAS         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
```

problem 1(b)

Build regression model reg.picked by excluding the two explanatory variables selected in problem 1(a). Display summary() of the model.

Solution

```
> reg.picked = lm(y~ x$CRIM + x$ZN + x$CHAS + x$NOX + x$RM + x$DIS + x$RAD + x$TAX
             + x$PTRATIO + x$B + x$LSTAT)
> summary(reg.picked)

Call:
lm(formula = y ~ x$CRIM + x$ZN + x$CHAS + x$NOX + x$RM + x$DIS +
    x$RAD + x$TAX + x$PTRATIO + x$B + x$LSTAT)

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 ***
x$CRIM       -0.108413   0.032779  -3.307 0.001010 **
x$ZN          0.045845   0.013523   3.390 0.000754 ***
x$CHAS        2.718716   0.854240   3.183 0.001551 **
x$NOX       -17.376023   3.535243  -4.915 1.21e-06 ***
x$RM          3.801579   0.406316   9.356  < 2e-16 ***
x$DIS        -1.492711   0.185731  -8.037 6.84e-15 ***
x$RAD         0.299608   0.063402   4.726 3.00e-06 ***
x$TAX        -0.011778   0.003372  -3.493 0.000521 ***
x$PTRATIO    -0.946525   0.129066  -7.334 9.24e-13 ***
x$B           0.009291   0.002674   3.475 0.000557 ***
```

```
x$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
```

problem 1(c)

For a regression model, the mean squared error (MSE) is defined as $\frac{SSE}{n-1-p}$, in which p is the number of explanatory variables used in the model. The mean absolute error (MAE) is similarly defined: $\frac{SAE}{n-1-p}$. Display MSE and MAE for regression models reg and reg.picked from the previous problems. Based on MSE and MAE, pick one model you prefer.

Solution
pick reg.picked since it has smaller MSE and MAE.

```
> e.reg = resid(reg);
> MSE.reg = sum(e.reg^2)/(506-1-13);
> MSE.reg
[1] 22.51785
> e.reg.picked = resid(reg.picked);
> MSE.reg.picked = sum(e.reg.picked^2)/(506-1-11);
> MSE.reg.picked
[1] 22.43191
> MAE.reg = sum(abs(e))/(506-1-13);
> MAE.reg
[1] 3.363936
> MAE.reg.picked = sum(abs(e.reg.picked))/(506-1-11);
> MAE.reg.picked
[1] 3.351519
```

problem 1(d)

Run step() using regression model reg in problem 1(a). Compare the model with reg.picked in problem 1(b).

Solution
They are same.

```
> step(reg);
Start:  AIC=1589.64
y ~ 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
y ~ 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
y ~ 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

Call:
lm(formula = y ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD + TAX +
    PTRATIO + B + LSTAT, data = x)

Coefficients:
(Intercept)         CRIM           ZN         CHAS          NOX           RM
  36.341145    -0.108413     0.045845     2.718716   -17.376023     3.801579
        DIS          RAD          TAX      PTRATIO            B        LSTAT
  -1.492711     0.299608    -0.011778    -0.946525     0.009291    -0.522553
```

Problem 2

Import *labdata.txt*. The summary of the data set is below.

Name of the data set	labdata
Number of observations	400
Number of attributes	9 (1 response variable and 8 explanatory variables)

Column y is the response variable and remaining attributes x1,x2,... are the explanatory variables.

problem 2(a)

Build regression model reg and display summary() of the model

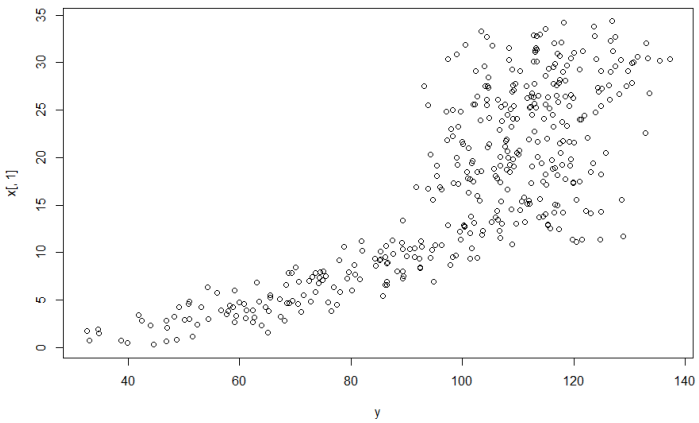
Solution

```
> y = labdata[,1]
> x = labdata[,2:9]
> reg = lm(y~., data=x)
> summary(reg)

Call:
lm(formula = y ~ ., data = x)

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
```

problem 2(b)

For each explanatory variable, plot it against the response variable. Based on the scartter plots, pick one variable that is most likely to be used in a piecewise regression model. Attach one plot associated with the variable you pick.

Solution
Pick x1



problem 2(c)

Calculate the mean of the variable you pick in problem 2(b) and build piecewise regression model reg.piece using the mean. Is model reg.piece better than model reg in problem 2(a)? Support your argument in one sentence.

SOLUTION
reg.piece is better, it has higher r^2 (and adjusted r^2).

```
> mean(x[,1])
[1] 17.19417
> reg.piece = lm(y~(x[,1]<17.19417)*x[,1]+x[,2]+x[,3]+x[,4]+x[,5]+x[,6]+x[,7]+x[,8])
> summary(reg.piece)

Call:
lm(formula = y ~ (x[, 1] < 17.19417) * x[, 1] + x[, 2] + x[,
    3] + x[, 4] + x[, 5] + x[, 6] + x[, 7] + x[, 8])

Residuals:
     Min       1Q   Median       3Q      Max
-14.3914  -1.3793  -0.1569   1.3062  14.5014

Coefficients:
                              Estimate Std. Error t value Pr(>|t|)
(Intercept)                  61.403705   2.254025  27.242   <2e-16 ***
x[, 1] < 17.19417TRUE       -57.094813   1.444014 -39.539   <2e-16 ***
x[, 1]                        0.517863   0.051891   9.980   <2e-16 ***
x[, 2]                        0.989106   0.029237  33.831   <2e-16 ***
x[, 3]                        1.032202   0.029060  35.520   <2e-16 ***
x[, 4]                        0.018861   0.030815   0.612    0.541
x[, 5]                       -0.017325   0.013135  -1.319    0.188
x[, 6]                       -0.006076   0.052914  -0.115    0.909
x[, 7]                       -0.053892   0.054904  -0.982    0.327
x[, 8]                       -0.038638   0.055426  -0.697    0.486
x[, 1] < 17.19417TRUE:x[, 1]  4.097539   0.078416  52.254   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 3.475 on 389 degrees of freedom
Multiple R-squared:  0.9774,Adjusted R-squared:  0.9768
F-statistic:  1682 on 10 and 389 DF,  p-value: < 2.2e-16
```

References

Harrison, D. and Rubinfeld, D.L. (1978) Hedonic prices and the demand for clean air. *Economics & Management*, **5**, 81–102.