Analytics for Competitive Advantage: Lab Exercise 3

Kristin Meier

November 15, 2016

Problem 1

Problem 1(a)

Build a 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.

```
housing <- read.table(paste(filepath,"ex3_bostonhousing.txt",sep=""),stringsAsFactors = F,header=T)

# MEDV is the response variable
reg <- lm(MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +
            DIS + RAD + TAX + PTRATIO + B + LSTAT, housing)
summary(reg)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + INDUS + CHAS + NOX + RM + AGE +
##     DIS + RAD + TAX + PTRATIO + B + LSTAT, data = housing)
##
## 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
```

```r
# Find which have the max 2 p-values
pvals.reg <- summary(reg)$coefficients[,4]
max2 <- sort(pvals.reg,decreasing=T)[1:2]
max.val <- rownames(summary(reg)$coefficients)[which(pvals.reg %in% max2)]
```

The two explanatory variables that are least likely to be in the best model are INDUS and AGE, based on the fact that the coefficient estimates for these predictors are not statistically significant and have the highest p-values of 0.958 and 0.738, respectively.

Problem 1(b)

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

```r
# exclude INDUS and AGE
reg.picked <- lm(MEDV ~ CRIM + ZN + CHAS + NOX + RM +
            DIS + RAD + TAX + PTRATIO + B + LSTAT, housing)
summary(reg.picked)
```

```
##
## Call:
## lm(formula = MEDV ~ CRIM + ZN + CHAS + NOX + RM + DIS + RAD +
##     TAX + PTRATIO + B + LSTAT, data = housing)
##
## 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
```

Problem 1(c)

For a regression model, the mean squared 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.

```
# MAE assigns equal weight to the data whereas MSE emphasizes the extremes.
# MAE gives equal weight to all errors, while RMSE gives extra weight to large errors.

# use predict to predict the values
actual <- housing$MEDV
pred.reg <- predict(reg, housing)
pred.reg.picked <- predict(reg.picked, housing)
error.reg <- actual - pred.reg
error.reg.picked <- actual - pred.reg.picked

# MSE
# mse.reg <- anova(reg)["Residuals","Mean Sq"]
# mse.reg.picked <- anova(reg.picked)["Residuals","Mean Sq"]
mse.reg <- mean(error.reg^2)
mse.reg.picked <- mean(error.reg.picked^2)

# MAE
mae.reg <- mean(abs(error.reg))
mae.reg.picked <- mean(abs(error.reg.picked))

round(data.frame(mse.reg,mse.reg.picked,mae.reg,mae.reg.picked),3)
```

```
##   mse.reg mse.reg.picked mae.reg mae.reg.picked
## 1  21.895           21.9   3.271          3.272
```

Based on MSE and MAE, I pick the model that minimizes these values, which is the reg model. The MSE and MAE are 21.895 and 3.271 (21.9 and 3.272 for the reg.picked model).

Problem 1(d)

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

```
# from lab
library(MASS)
```

```
##
## Attaching package: 'MASS'

## The following object is masked _by_ '.GlobalEnv':
##
##     housing
```

```r
reg = lm(MEDV~., data=housing)
reg.step = stepAIC(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
```

After running the step function to select a model, the result contains the same variables from part 1(b). (CRIM, ZN, B, CHAS, NOX, RM, DIS, RAD, TAX, PTRATIO, LSTAT)

Probelm 2

Problem 2(a)

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

```
lab <- read.table(paste(filepath,"ex3_labdata.txt",sep=""),stringsAsFactors = F,header=T)

# regression
reg <- lm(y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, lab)
summary(reg)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x3 + x4 + x5 + x6 + x7 + x8, 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
```

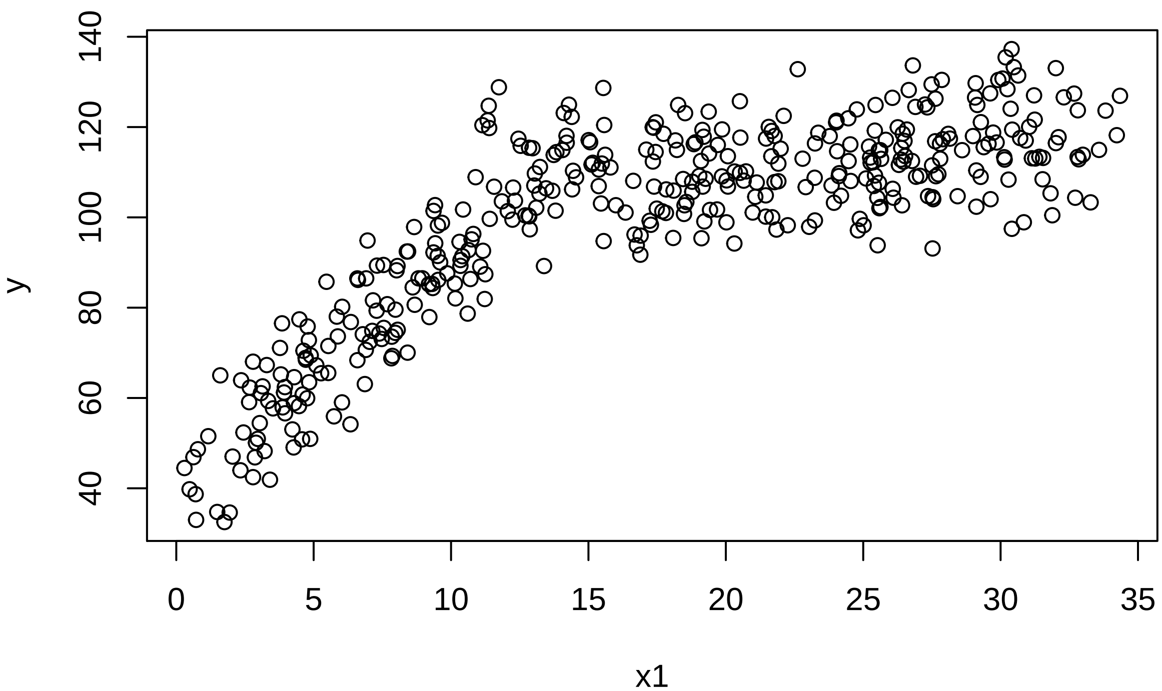
Problem 2(b)

For each explanatory variable, plot it against the response variable. Based on the scatter 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.

```
#par(omi = c(.75,.75,.75,.75), mar = c(5,4,2,2))
#layout(matrix(c(1,2),ncol=1, byrow = TRUE))
'
for(i in 2:ncol(lab)){
  plot(x=lab[,i],
       y=lab[,1],
       xlab = colnames(lab)[i],
       ylab = colnames(lab)[1])
}
'
```

```
## [1] "\nfor(i in 2:ncol(lab)){\n  plot(x=lab[,i],\n       y=lab[,1],\n       xlab = colnames(lab)[i],\
```

```
# just plot x1
  plot(x=lab[,2],
       y=lab[,1],
       xlab = colnames(lab)[2],
       ylab = colnames(lab)[1])
```



The explanatory variable most likely to be used in a piecewise regression model is x1. From the scatter plot it is clear that the data display different patterns before and after a cricial point (around x1=15). The other variables do not have a clear break in their relationship with y.

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.

```
var.picked <- "x1"
var.mean <- mean(lab[,var.picked])

#install.packages("segmented")
library(segmented)

reg.piece = segmented(reg, seg.Z = ~x1, psi=var.mean)
summary(reg.piece)
```

```
##
##  ***Regression Model with Segmented Relationship(s)***
##
## Call:
## segmented.lm(obj = reg, seg.Z = ~x1, psi = var.mean)
##
## Estimated Break-Point(s):
##    Est. St.Err
## 12.585  0.097
##
## Meaningful coefficients of the linear terms:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.896778   1.181852  -1.605    0.109
## x1           5.421324   0.057178  94.814   <2e-16 ***
## x2           1.009138   0.018602  54.248   <2e-16 ***
## x3           0.978814   0.018574  52.699   <2e-16 ***
## x4           0.011381   0.019577   0.581    0.561
## x5           0.004714   0.008323   0.566    0.571
## x6          -0.017313   0.033733  -0.513    0.608
## x7          -0.018195   0.035011  -0.520    0.604
## x8           0.007425   0.035295   0.210    0.833
## U1.x1       -4.907611   0.062084 -79.048       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.215 on 389 degrees of freedom
## Multiple R-Squared: 0.9908,  Adjusted R-squared: 0.9906
##
## Convergence attained in 5 iterations with relative change -5.955294e-16
```

```
# interpret as coef.x1*x1 + coef.U1.x1*1_x1<mean

# SSE
sse.reg <- round(anova(reg)["Residuals","Sum Sq"],3)
sse.piece <- round(anova(reg.piece)["Residuals","Sum Sq"],3)
# R2
r2.reg <- round(summary(reg)$r.squared,3)
```

```r
r2.piece <- round(summary(reg.piece)$r.squared,3)
# F value
f.reg <- round(summary(reg)$fstatistic[1],3)
f.piece <- round(summary(reg.piece)$fstatistic[1],3)
# num predictors significant

sse.piece < sse.reg
```

```
## [1] TRUE
```

```r
r2.piece > r2.reg
```

```
## [1] TRUE
```

```r
f.piece > f.reg
```

```
## value
##  TRUE
```

The regpiece model is better than the reg model because although less variables are significant, it outperforms the reg model by all other measures, namely SSE, R^2, and F-value.