PSTAT 126 Project 6

Melanie Flandes

3/15/2018

```
library(faraway)
data(state)
state.x77 = as.data.frame(state.x77)
colnames(state.x77)[4] = "Life.Exp"
```

Problem 1a | Fitting Linear Models

```
mod0 = lm(Life.Exp ~ 1, data = state.x77)
mod.all = lm(Life.Exp ~ ., data = state.x77 )
step(mod0, scope = list(lower = mod0, upper = mod.all))
```

```
## Start:  AIC=30.44
## Life.Exp ~ 1
##
##              Df Sum of Sq    RSS     AIC
## + Murder      1    53.838 34.461 -14.609
## + Illiteracy  1    30.578 57.721  11.179
## + `HS Grad`   1    29.931 58.368  11.737
## + Income      1    10.223 78.076  26.283
## + Frost       1     6.064 82.235  28.878
## <none>                    88.299  30.435
## + Area        1     1.017 87.282  31.856
## + Population  1     0.409 87.890  32.203
##
## Step:  AIC=-14.61
## Life.Exp ~ Murder
##
##              Df Sum of Sq    RSS     AIC
## + `HS Grad`   1     4.691 29.770 -19.925
## + Population  1     4.016 30.445 -18.805
## + Frost       1     3.135 31.327 -17.378
## + Income      1     2.405 32.057 -16.226
## <none>                    34.461 -14.609
## + Area        1     0.470 33.992 -13.295
## + Illiteracy  1     0.273 34.188 -13.007
## - Murder      1    53.838 88.299  30.435
##
## Step:  AIC=-19.93
## Life.Exp ~ Murder + `HS Grad`
```

```
##
##               Df Sum of Sq    RSS     AIC
## + Frost        1    4.3987 25.372 -25.920
## + Population   1    3.3405 26.430 -23.877
## <none>                     29.770 -19.925
## + Illiteracy   1    0.4419 29.328 -18.673
## + Area         1    0.2775 29.493 -18.394
## + Income       1    0.1022 29.668 -18.097
## - `HS Grad`    1    4.6910 34.461 -14.609
## - Murder       1   28.5974 58.368  11.737
##
## Step:  AIC=-25.92
## Life.Exp ~ Murder + `HS Grad` + Frost
##
##               Df Sum of Sq    RSS     AIC
## + Population   1     2.064 23.308 -28.161
## <none>                     25.372 -25.920
## + Income       1     0.182 25.189 -24.280
## + Illiteracy   1     0.172 25.200 -24.259
## + Area         1     0.026 25.346 -23.970
## - Frost        1     4.399 29.770 -19.925
## - `HS Grad`    1     5.955 31.327 -17.378
## - Murder       1    32.756 58.128  13.531
##
## Step:  AIC=-28.16
## Life.Exp ~ Murder + `HS Grad` + Frost + Population
##
##               Df Sum of Sq    RSS     AIC
## <none>                     23.308 -28.161
## + Income       1     0.006 23.302 -26.174
## + Illiteracy   1     0.004 23.304 -26.170
## + Area         1     0.001 23.307 -26.163
## - Population   1     2.064 25.372 -25.920
## - Frost        1     3.122 26.430 -23.877
## - `HS Grad`    1     5.112 28.420 -20.246
## - Murder       1    34.816 58.124  15.528

##
## Call:
## lm(formula = Life.Exp ~ Murder + `HS Grad` + Frost + Population,
##     data = state.x77)
##
## Coefficients:
## (Intercept)       Murder    `HS Grad`        Frost   Population
##   7.103e+01   -3.001e-01    4.658e-02   -5.943e-03    5.014e-05
```

```
mod.AIC = lm(Life.Exp ~ Murder + `HS Grad` + Frost, data = state.x77)
summary(mod.AIC)
```

```
##
## Call:
## lm(formula = Life.Exp ~ Murder + `HS Grad` + Frost, data = state.x77)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.5015 -0.5391  0.1014  0.5921  1.2268
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 71.036379   0.983262  72.246  < 2e-16 ***
## Murder      -0.283065   0.036731  -7.706 8.04e-10 ***
## `HS Grad`    0.049949   0.015201   3.286  0.00195 **
## Frost       -0.006912   0.002447  -2.824  0.00699 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7427 on 46 degrees of freedom
## Multiple R-squared:  0.7127, Adjusted R-squared:  0.6939
## F-statistic: 38.03 on 3 and 46 DF,  p-value: 1.634e-12
```

Problem 1b | Plotting using Model Selection

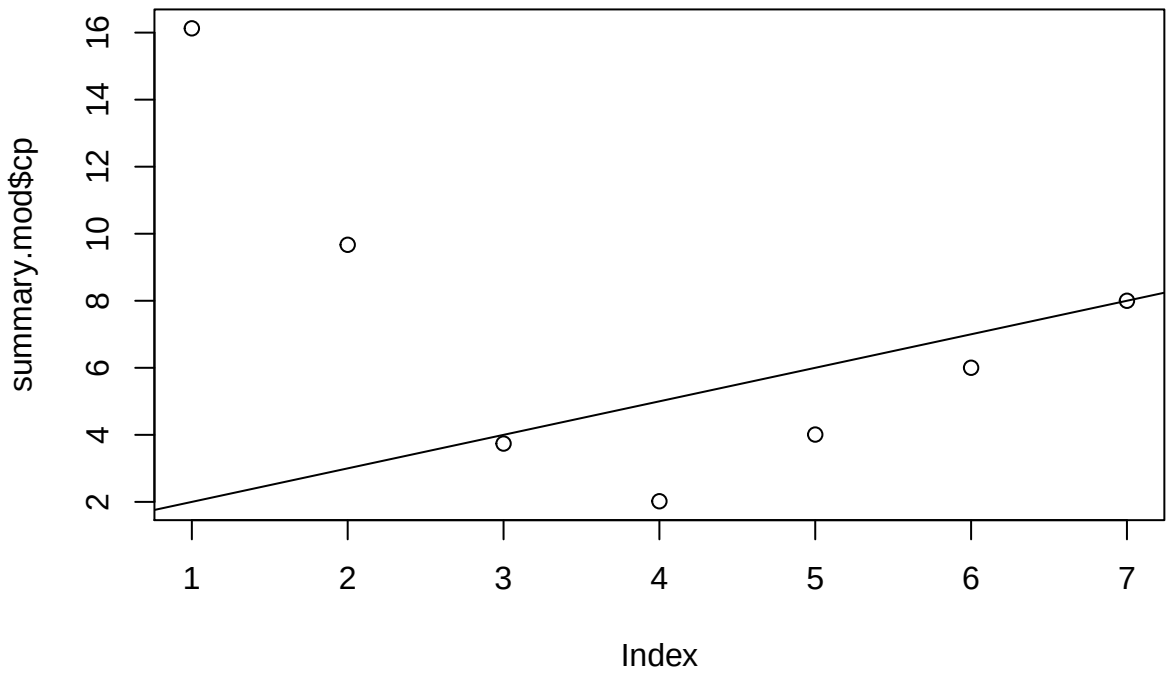
```
library(leaps)
Population = state.x77$Population
Income = state.x77$Income
Illiteracy = state.x77$Illiteracy
Murder = state.x77$Murder
`HS Grad` = state.x77$`HS Grad`
Frost = state.x77$Frost
Area = state.x77$Area
Life.Exp = state.x77$Life.Exp

mod = regsubsets(cbind(Population,Income, Illiteracy, Murder, `HS Grad`, Frost, Area), Life.Exp)
summary.mod = summary(mod)
names(summary(mod))
```

```
## [1] "which"  "rsq"    "rss"    "adjr2"  "cp"     "bic"    "outmat" "obj"
```

```
summary.mod$which
```

```
##   (Intercept) Population Income Illiteracy Murder HS Grad Frost  Area
## 1        TRUE      FALSE  FALSE      FALSE   TRUE   FALSE FALSE FALSE
## 2        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE FALSE FALSE
## 3        TRUE      FALSE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
## 4        TRUE       TRUE  FALSE      FALSE   TRUE    TRUE  TRUE FALSE
## 5        TRUE       TRUE   TRUE      FALSE   TRUE    TRUE  TRUE FALSE
## 6        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE FALSE
## 7        TRUE       TRUE   TRUE       TRUE   TRUE    TRUE  TRUE  TRUE
```

```
plot(summary.mod$cp)
abline(1,1)
```




```
mod.cp = lm(Life.Exp ~ Murder + `HS Grad` + Frost, data = state.x77)
```

Problem 1c | Summary

```
summary.mod$adjr2
```

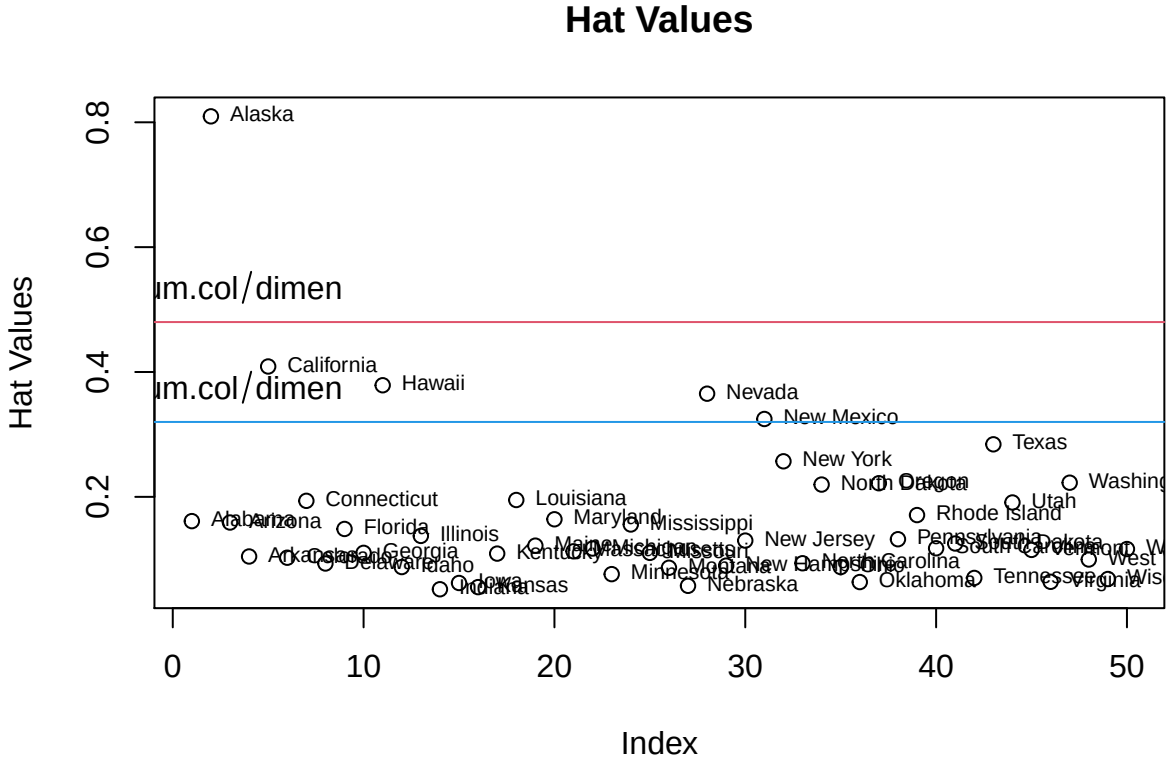
```
## [1] 0.6015893 0.6484991 0.6939230 0.7125690 0.7061129 0.6993268 0.6921823
```

Problem 1d | Plot

```
dimen = dim(state.x77)[1]
num.col = 8
hat_val = hatvalues(mod.all)
which(hat_val == max(hat_val))
```

```
## Alaska
##      2
```

```
states = rownames(state.x77)
plot(hat_val, ylab = 'Hat Values', main = 'Hat Values')
text(1:dimen, hat_val, labels = states, cex = .7, pos = 4)
avg.hat = num.col/dimen

abline(h=2*avg.hat, col = 4)

abline(h=3*avg.hat, col = 2)
text(2, y=2*avg.hat, expression(2 %*% num.col/dimen), pos=3)
text(2, y=3*avg.hat, expression(3 %*% num.col/dimen), pos=3)
```




```
rstand = rstandard(mod.all)
which(rstand == max(rstand))
```

```
## Hawaii
##     11
```

```
c = 2*sqrt((num.col+1)/(dimen-num.col-1))
c
```

```
## [1] 0.9370426
```

```r
which(dffits(mod.all) > c)
```

```
## Hawaii
##     11
```

Alaska is the state with the highest leverage, Hawaii is the state with the largest externally studentized leverage, The degrees of freedom for the fit is Hawaii.

Problem 1e | Plotting Fitted Model

```r
fit2 = lm(Life.Exp ~ ., data = state.x77[-11,])
summary(mod.all)
```

```
##
## Call:
## lm(formula = Life.Exp ~ ., data = state.x77)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.48895 -0.51232 -0.02747  0.57002  1.49447
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.094e+01  1.748e+00  40.586  < 2e-16 ***
## Population   5.180e-05  2.919e-05   1.775   0.0832 .
## Income      -2.180e-05  2.444e-04  -0.089   0.9293
## Illiteracy   3.382e-02  3.663e-01   0.092   0.9269
## Murder      -3.011e-01  4.662e-02  -6.459 8.68e-08 ***
## `HS Grad`    4.893e-02  2.332e-02   2.098   0.0420 *
## Frost       -5.735e-03  3.143e-03  -1.825   0.0752 .
## Area        -7.383e-08  1.668e-06  -0.044   0.9649
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7448 on 42 degrees of freedom
## Multiple R-squared:  0.7362, Adjusted R-squared:  0.6922
## F-statistic: 16.74 on 7 and 42 DF,  p-value: 2.534e-10
```

```r
summary(fit2)
```

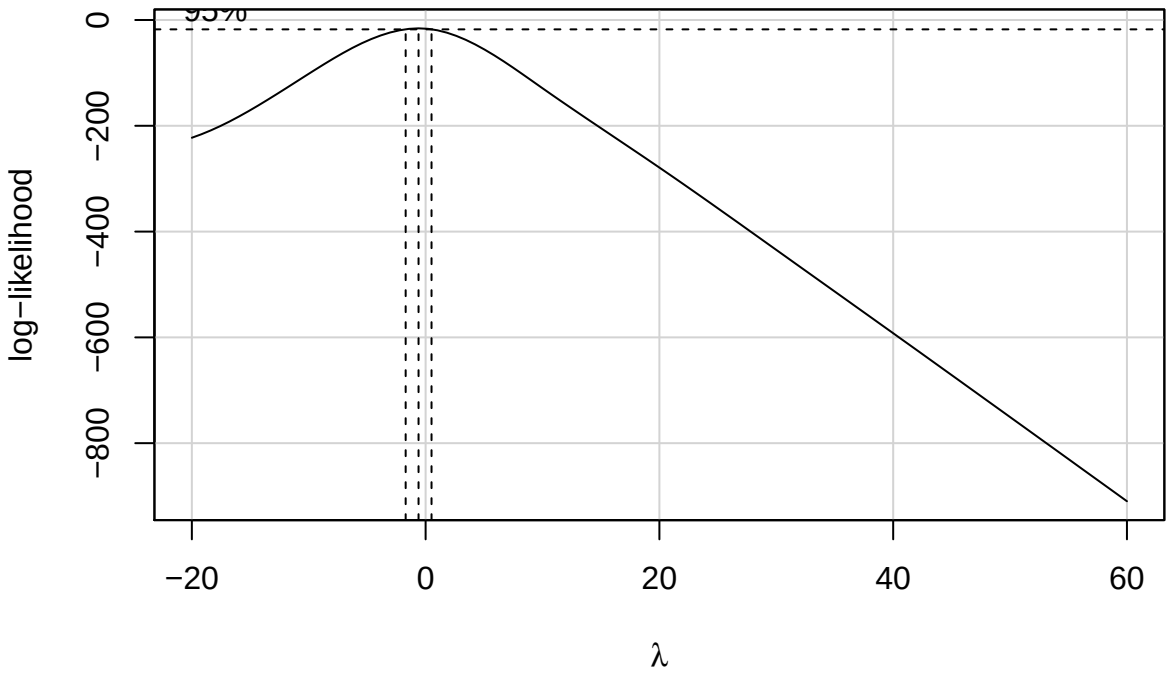
```
##
## Call:
## lm(formula = Life.Exp ~ ., data = state.x77[-11, ])
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1.60342 -0.46441 -0.05849  0.49517  1.02433
##
## Coefficients:
```

```
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)  7.270e+01  1.750e+00  41.551  < 2e-16 ***
## Population   6.773e-05  2.778e-05   2.438   0.0192 *
## Income      -1.749e-04  2.343e-04  -0.746   0.4597
## Illiteracy  -3.107e-01  3.634e-01  -0.855   0.3976
## Murder      -2.884e-01  4.364e-02  -6.608  5.9e-08 ***
## `HS Grad`    2.694e-02  2.315e-02   1.164   0.2512
## Frost       -4.095e-03  2.986e-03  -1.371   0.1778
## Area         1.441e-06  1.648e-06   0.874   0.3871
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6932 on 41 degrees of freedom
## Multiple R-squared:  0.756,  Adjusted R-squared:  0.7143
## F-statistic: 18.15 on 7 and 41 DF,  p-value: 1.002e-10
```

The new model has a stronger relationship because the r squared is increased.

Problem 2a | Log Likelihood for BoxCox

```
library(alr4)
data(lathe1)
attach(lathe1)
life = Life[-c(7,8,10)]
life1 = log(life)
speed1 = Speed[-c(7,8,10)]

boxCox(life1 ~ speed1, data = lathe1, seq(-20,60,10))
```



Problem 2b | Removing Influential Points

```
L = lathe1[-c(3,4),]
mod.all1 = lm(Life ~ Speed + Feed, data = lathe1)
fit2 = lm(Life ~ Speed + Feed, data = L)
summary(mod.all1)
```

```
##
## Call:
## lm(formula = Life ~ Speed + Feed, data = lathe1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.873 -11.420  -9.990   9.941  41.412
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   14.620      3.645   4.011 0.000906 ***
## Speed        -21.548      4.706  -4.579 0.000267 ***
## Feed         -10.494      4.706  -2.230 0.039521 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

```
## Residual standard error: 16.3 on 17 degrees of freedom
## Multiple R-squared:  0.6041, Adjusted R-squared:  0.5575
## F-statistic: 12.97 on 2 and 17 DF,  p-value: 0.00038
```

```
summary(fit2)
```

```
##
## Call:
## lm(formula = Life ~ Speed + Feed, data = L)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -14.674 -13.299  -5.374  13.474  33.844
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   16.874      3.930   4.293 0.000641 ***
## Speed        -25.305      5.322  -4.755 0.000256 ***
## Feed          -6.737      5.322  -1.266 0.224884
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 16.21 on 15 degrees of freedom
## Multiple R-squared:  0.6545, Adjusted R-squared:  0.6084
## F-statistic: 14.21 on 2 and 15 DF,  p-value: 0.0003456
```

Once the influential points are removed it made the relationship stronger.