Homework 1

Peter Williams

Problem #3

Give examples of hard-to-change factors. How do you reconcile the hard-to-change nature of the factor with the need for randomization?

Some examples of hard-to-change factors could include:

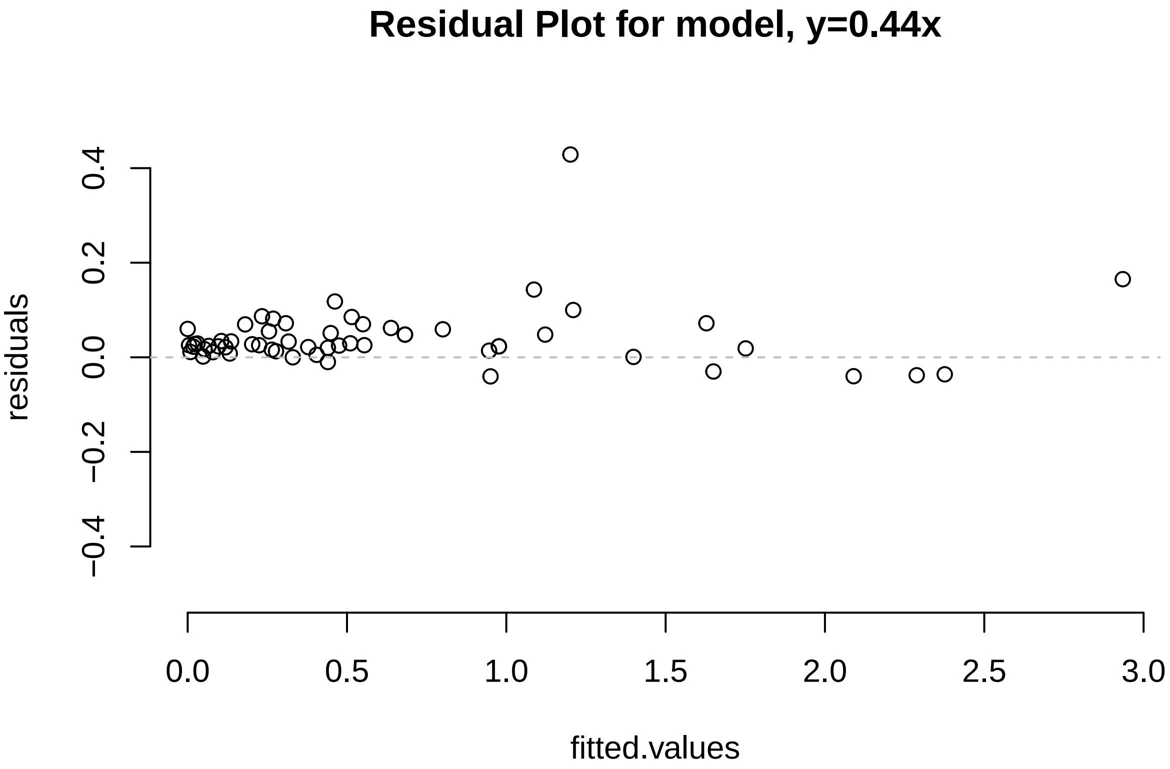
- The temperature in an oven in an experiment related to baking. If the oven takes time to adjust to a different temp, or an oven can only fit a limited amount of baked goods in it at a time, randomizing the temperature would take a lot of effort and time.
- The ordering of TV programs for a schedule at a TV network. Since each program on a TV schedule has separate licensing agreement and flight restrictions, re-ordering TV programs on a schedule requires significant effort.
- Setting up a complex assembly machine in a manufacturing plant. If a particular configuration on a piece of complex equipment takes significant time and effort to set-up, there is a trade-off between the benefits randomizing levels of the factor and effort, and time need to perform the experiment.
- The plot of land in an experiment to test seed and fertilizer types. Since fertilizer can only be applied to large areas it is hard to randomize the plots of land if there is only limited land, or there are only limited plots available.

Some reconciliation can be made to the hard-to-change nature of factors by systematically randomizing other factors of interest in an experiment. Ensuring that the form of randomization within the hard-to-change experimental factors is structured should help lead to an appropriate analysis.

Problem #9

(a) Plot the residuals y_i - $0.44x_i$ for the data, Do you observe any systematic pattern to question the validity of the formula $y = 0.44x$?

```r
rainfall <-
  read.table('http://www2.isye.gatech.edu/~jeffwu/wuhamadabook/data/rainfall.dat',
             header=T)
#residual plot
with(rainfall,{
  plot(0.44*x, y - 0.44*x, bty='n', main = 'Residual Plot for model, y=0.44x',
       xlab = 'fitted.values', ylab = 'residuals', ylim = c(-0.5,0.5));
  abline(h=0, lty=2, col='grey')
})
```

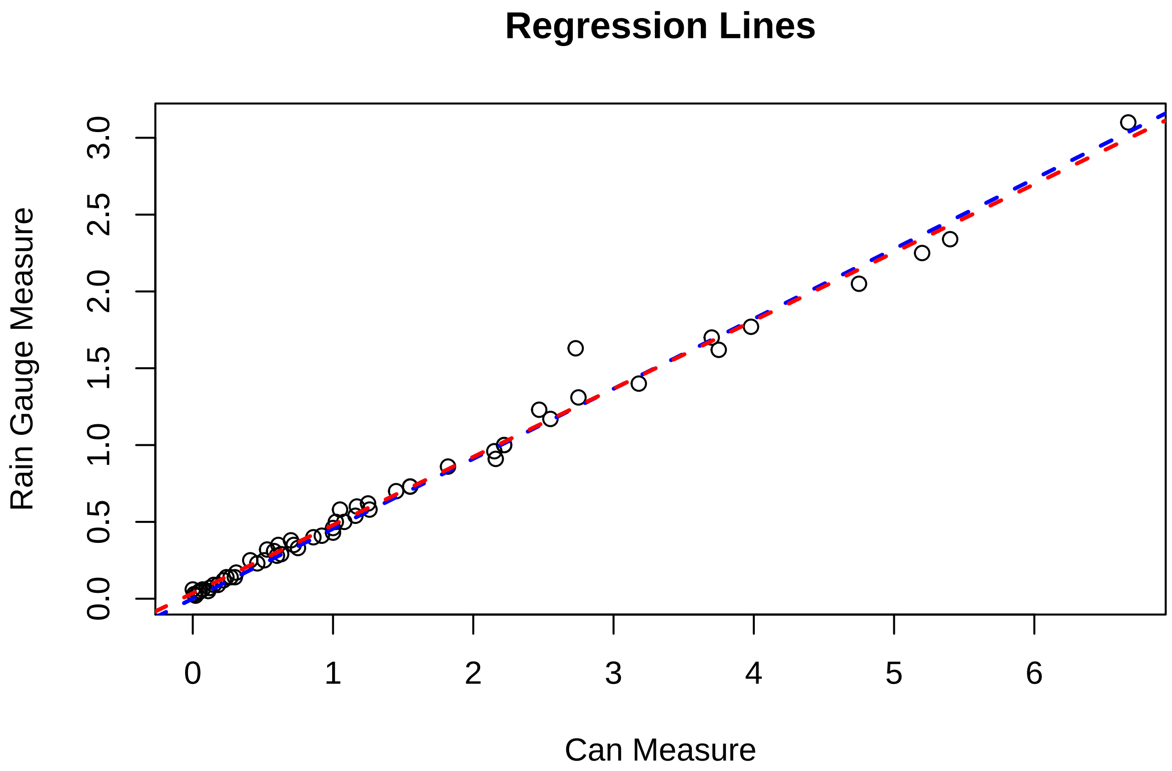


The model y = 0.44x consistently underestimates the amount of rain collected in the rain gauge, as the residual plot shows the majority of points falling above the zero line, especially at lower fitted.values.

(b) Use regression analysis to analyzed the data in Table 1.10 by assuming a general β_0 and $\beta_0 = 0$. How well do the two models fit the data? Is the intercept term significant?

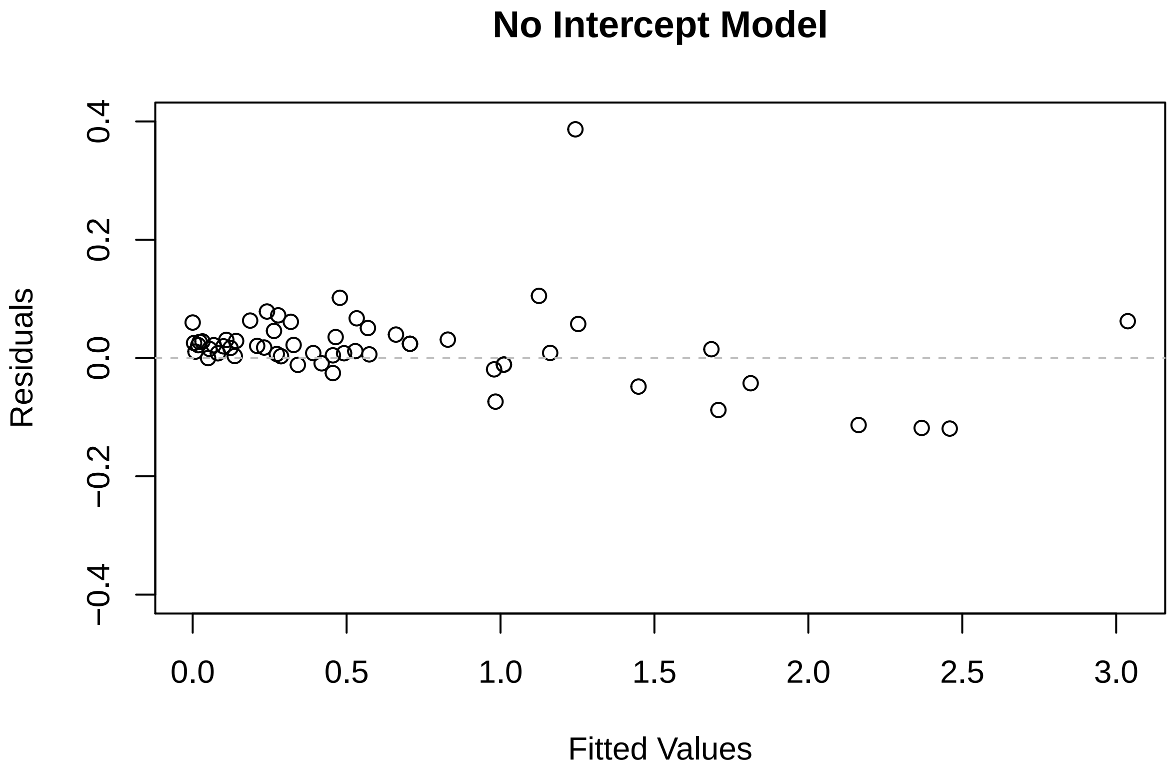
Model fit comparisons:

```r
noIntercept <- lm(y~x-1, data=rainfall)
withIntercept <- lm(y~x, data=rainfall)
with(rainfall,{
    plot(x,y, main='Regression Lines', ylab = 'Rain Gauge Measure', xlab = 'Can Measure');
    abline(noIntercept, col='blue', lty=2, lwd=2);
    abline(withIntercept, col='red', lty=2, lwd=2);
})
```

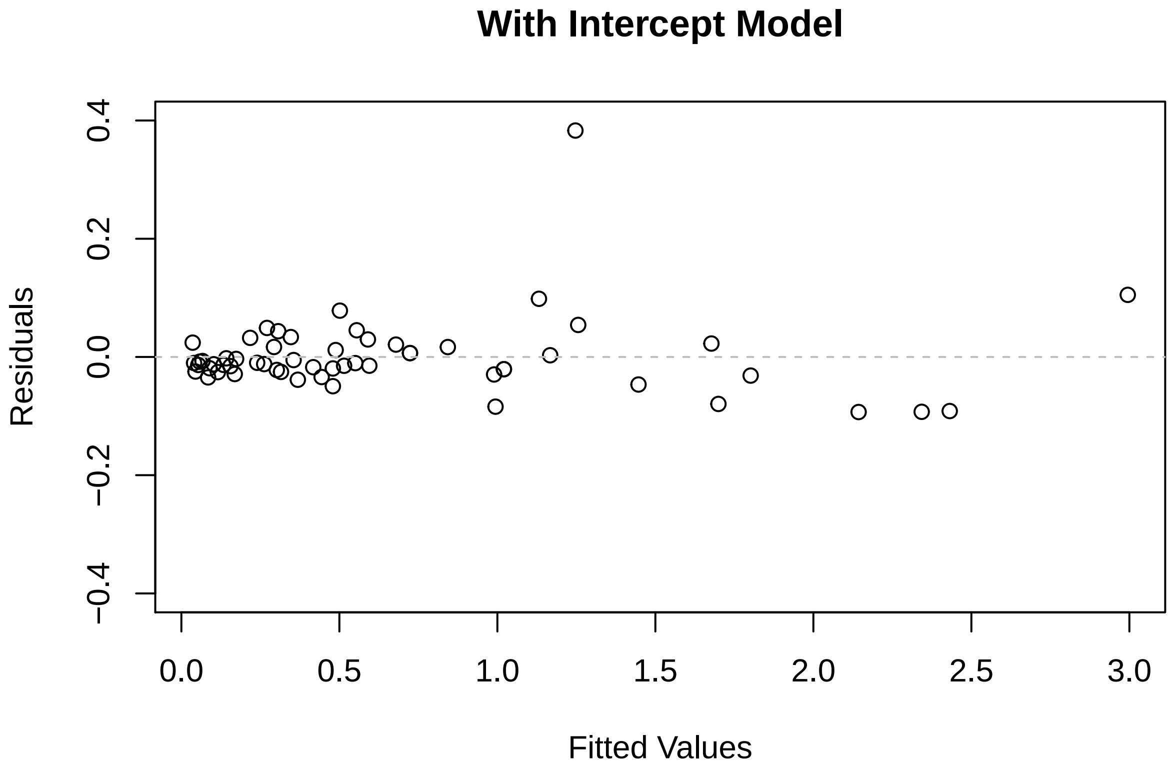


Residual plots:

```
plot(as.numeric(noIntercept$fitted.values), as.numeric(resid(noIntercept)), ylim =c(-0.4,0.4),
     main='No Intercept Model', ylab='Residuals', xlab='Fitted Values')
abline(h=0, lty=2, col='grey')
```



```
plot(as.numeric(withIntercept$fitted.values), as.numeric(resid(withIntercept)), ylim =c(-0.4,0.4),
     main='With Intercept Model', ylab='Residuals', xlab='Fitted Values')
abline(h=0, lty=2, col='grey')
```



```
summary(withIntercept)
```

```
## 
## Call:
## lm(formula = y ~ x, data = rainfall)
## 
## Residuals:
##      Min       1Q   Median       3Q      Max 
## -0.09314 -0.02529 -0.01205  0.01689  0.38304 
## 
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)    
## (Intercept) 0.035787   0.012210   2.931  0.00491 ** 
## x           0.443652   0.005817  76.264  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## 
## Residual standard error: 0.06668 on 55 degrees of freedom
## Multiple R-squared:  0.9906, Adjusted R-squared:  0.9905 
## F-statistic:  5816 on 1 and 55 DF,  p-value: < 2.2e-16
```

The scatter plot with lines show the fit of the no intercept model in blue, and the with intercept model in red shows the fact that the 'no intercept' model goes right through the origin (0,0). As a result of this, the residual plot for the no intercept model shows a systematic underestimation of the model at lower fitted values. The model with the intercept does not have this bias as seen in its residual plot. The summary of the model with an intercept shows a parameter estimate with a significant estimated effect.

(c) Because of evaporation during the summer and the can being made of metal, the formula y = 0.44x may not fit the rainfall data collected in the summer. An argument can be made that supports the model with an intercept. Is this supported by your analyses in (a) and (b)?

Yes, there is statistical evidence that the parameter for the intercept is worth including. Visualization of the data, and the model residuals also highlights a better less biased description of the data from a model with an intercept. The added information about the can and evaporation in the summer provides further, domain-specific knowledge to support the model specification.

Problem #13

Data:

1. Minority: minority percentage
2. Crime: rate of serious crimes per 100 population
3. Poverty: percentage poor
4. Language: percentage having difficultty speaking or wriing English
5. Highschool: percentage age 25 or older who had not finished high school
6. Housing: percentage of housing in small multi-unit buildings
7. City a factor with two levels: "city" (major city), "state" (state remainder).
8. Conventional: percentage of households counted by conventional personal enumeration

The response is the undercount (in terms of percentage). Use regression to investigate the relationship between undercount and the eight predictors.

(a) Perform regression analysis using all the predictors except city. Show the regression residuals. Which predictors seems to be important? Draw the residual plot against the fitted value. What can you conclude from this plot

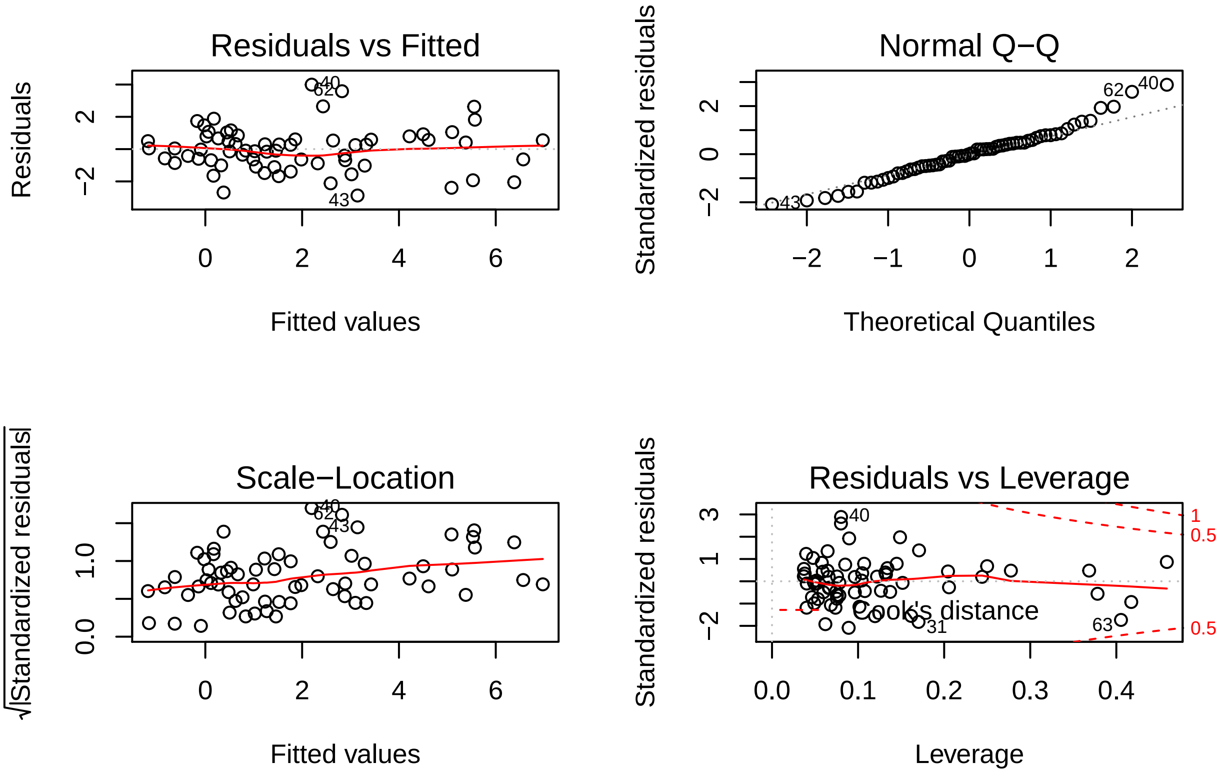
```
allModel <-
  lm(Undercount~Minority+Crime+Poverty+Language+Highschool+Housing+Conventional,
     data=ericksenData)
summary(allModel)
```

```
##
## Call:
## lm(formula = Undercount ~ Minority + Crime + Poverty + Language +
##     Highschool + Housing + Conventional, data = ericksenData)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8742 -0.8182  0.0064  0.6476  3.9922
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -2.217678   1.364658  -1.625 0.109568
## Minority      0.093366   0.020970   4.452 3.92e-05 ***
## Crime         0.034687   0.012775   2.715 0.008712 **
## Poverty      -0.173520   0.085775  -2.023 0.047696 *
## Language      0.230780   0.092615   2.492 0.015589 *
## Highschool    0.058204   0.045213   1.287 0.203100
## Housing      -0.022518   0.023454  -0.960 0.340995
## Conventional  0.036120   0.009335   3.869 0.000279 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.442 on 58 degrees of freedom
## Multiple R-squared:  0.6961, Adjusted R-squared:  0.6594
## F-statistic: 18.98 on 7 and 58 DF,  p-value: 6.62e-13
```

Based on the model summary, the predictors of Minority, Crime, Poverty, Language, and Conventional all are estimated to have significant effects based on the t-statistics associated with the estimated betas. Highschool and Housing are estimated to have insignificant effects.

Minority, Crime, and Language, and Conventional all have a (+) estimated effect on undercount, and these variables are associated with higher expected undercount. An increase in Poverty is estimated to have a (-) effect on Undercount, when considering all the other variables in the model.

```
par(mfrow=c(2,2))
plot(allModel)
```



There aren't any strong patterns present in the residual plots, although the standardized residuals may slightly increase for higher fitted values. South Carolina and Philadelphia both stand out as places with higher than expected Undercount given the model. Texas had much lower than expected undercount given the model. More research into those deviations would be worthwhile if conducting further analysis.

(b) Explain how the variable “City” differs from the others?

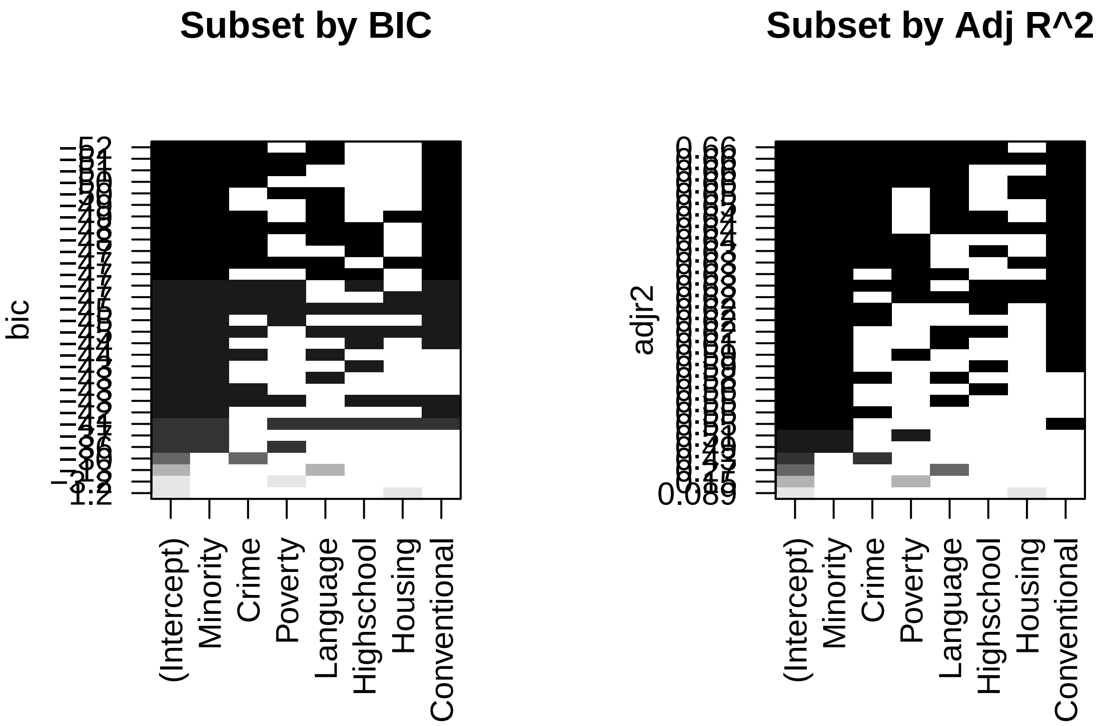
“City” is qualitative factor with two unique levels, ‘city’ or ‘state’. The other variables in the model take values over a continuous range, albeit on different scales.

(c) Use both best subset regression and stepwise regression to select variables from all the predictors (excluding the variable “City”). Compare you final models obtained by the two methods.

```
library(leaps)
subModels <-
  regsubsets(
    Undercount~Minority+Crime+Poverty+Language+Highschool+Housing+Conventional,
                        data=ericksenData,nbest=5)
#Mallows CP Summary
cpSummary <- data.frame(cbind(summary(subModels)$which, summary(subModels)$cp),
                        row.names=NULL);
colnames(cpSummary)[ncol(cpSummary)] <- 'MallowsCP'
#part of results
tail(cpSummary[order(cpSummary$MallowsCP,decreasing=T),2:ncol(cpSummary)],n=5)
```

```
##    Minority Crime Poverty Language Highschool Housing Conventional
## 31        1     1       1        1          1       1            1
## 27        1     1       1        1          0       1            1
## 16        1     1       0        1          0       0            1
## 26        1     1       1        1          1       0            1
## 21        1     1       1        1          0       0            1
##    MallowsCP
## 31  8.000000
## 27  7.657168
## 16  7.137160
## 26  6.921773
## 21  6.010091
```

```
#Graphs for other criteria
par(mfrow=c(1,2))
plot(subModels,scale='bic', main='Subset by BIC')
plot(subModels,scale='adjr2', main='Subset by Adj R^2')
```



Using the subsets approach, the BIC criterion yields a model excluding Poverty, Highschool, and Housing. The minimum Mallows CP criterion value yields an option excluding Highschool and Housing (as seen above), the and Adjusted R^2 criterion yields a model excluding Housing.

```
#stepwise from null model
nullModel <- lm(Undercount~1, data=ericksenData)
step(nullModel, scope = list(lower=nullModel, upper=allModel), direction='both')
```

```
## Start:  AIC=120.39
## Undercount ~ 1
##
##                  Df Sum of Sq    RSS     AIC
## + Minority        1   195.816 200.96  77.489
## + Crime           1   175.454 221.33  83.859
## + Language        1   111.017 285.76 100.724
## + Poverty         1    63.743 333.04 110.828
## + Housing         1    40.754 356.03 115.233
## <none>                        396.78 120.386
## + Highschool      1     8.113 388.67 121.022
## + Conventional    1     0.041 396.74 122.379
##
## Step:  AIC=77.49
## Undercount ~ Minority
##
##                  Df Sum of Sq    RSS     AIC
## + Highschool      1    30.175 170.79  68.751
## + Language        1    29.667 171.30  68.947
## + Crime           1    29.171 171.79  69.138
## + Conventional    1    26.851 174.11  70.023
## + Poverty         1    12.136 188.83  75.378
## <none>                        200.96  77.489
## + Housing       1     2.218 198.75  78.757
## - Minority      1   195.816 396.78 120.386
##
## Step:  AIC=68.75
## Undercount ~ Minority + Highschool
##
##                Df Sum of Sq    RSS     AIC
## + Conventional  1    13.876 156.91  65.159
## + Language      1    11.886 158.90  65.990
## + Crime         1    10.907 159.88  66.396
## <none>                      170.79  68.751
## + Housing       1     3.231 167.56  69.491
## + Poverty       1     0.005 170.78  70.749
## - Highschool    1    30.175 200.96  77.489
## - Minority      1   217.877 388.67 121.022
##
## Step:  AIC=65.16
## Undercount ~ Minority + Highschool + Conventional
##
##                Df Sum of Sq    RSS     AIC
## + Crime         1    16.193 140.72  59.970
## + Language      1    15.133 141.78  60.466
## <none>                      156.91  65.159
## + Poverty       1     3.560 153.35  65.644
## + Housing       1     2.526 154.39  66.088
## - Conventional  1    13.876 170.79  68.751
## - Highschool    1    17.199 174.11  70.023
## - Minority      1   229.443 386.36 122.629
##
## Step:  AIC=59.97
## Undercount ~ Minority + Highschool + Conventional + Crime
##
##                Df Sum of Sq    RSS    AIC
## + Language      1     9.885 130.84 57.163
## + Poverty       1     6.862 133.86 58.671
## - Highschool    1     3.093 143.81 59.405
## <none>                      140.72 59.970
## + Housing       1     0.339 140.38 61.811
## - Crime         1    16.193 156.91 65.159
## - Conventional  1    19.161 159.88 66.396
## - Minority      1    55.555 196.28 79.931
##
## Step:  AIC=57.16
## Undercount ~ Minority + Highschool + Conventional + Crime + Language
##
##                Df Sum of Sq    RSS    AIC
## + Poverty       1     8.325 122.51 54.824
## - Highschool    1     0.440 131.28 55.385
## <none>                      130.84 57.163
## + Housing       1     1.733 129.10 58.284
## - Language      1     9.885 140.72 59.970
## - Crime         1    10.945 141.78 60.466
## - Conventional  1    21.216 152.05 65.082
## - Minority      1    39.344 170.18 72.515
```

```
##
## Step:  AIC=54.82
## Undercount ~ Minority + Highschool + Conventional + Crime + Language +
##     Poverty
##
##                Df Sum of Sq    RSS    AIC
## - Highschool    1     2.263 124.77 54.032
## <none>                      122.51 54.824
## + Housing       1     1.917 120.59 55.784
## - Poverty       1     8.325 130.84 57.163
## - Language      1    11.347 133.86 58.671
## - Crime         1    13.782 136.29 59.860
## - Conventional  1    29.360 151.87 67.003
## - Minority      1    47.669 170.18 74.515
##
## Step:  AIC=54.03
## Undercount ~ Minority + Conventional + Crime + Language + Poverty
##
##                Df Sum of Sq    RSS    AIC
## <none>                      124.77 54.032
## + Highschool    1     2.263 122.51 54.824
## - Poverty       1     6.502 131.28 55.385
## + Housing       1     0.734 124.04 55.643
## - Language      1     9.400 134.17 56.826
## - Crime         1    11.535 136.31 57.868
## - Conventional  1    29.971 154.75 66.240
## - Minority      1    51.832 176.61 74.962
```

```
##
## Call:
## lm(formula = Undercount ~ Minority + Conventional + Crime + Language +
##     Poverty, data = ericksenData)
##
## Coefficients:
##  (Intercept)      Minority  Conventional         Crime      Language
##     -0.79269       0.10097       0.02933       0.02435       0.18385
##      Poverty
##     -0.11003
```

Both a forward and backwards stepwise fitting routine selection procedures are run. The backward elimination procedure yields the lowest model AIC (54.03), modeling the response (Undercount) by Minority, Crime, Poverty, Language, and Conventional. Similar to the results of the best subset procedure when employing BIC as the criteria for variable selection.