Stat 6021: Homework Set 7 Solutions

1. (a) The following variables are categorical:

   - low (binary)
   - race
   - smoke (binary)
   - ht (binary)
   - ui (binary)

   Note: some write that the variables ptl and ftv are also categorical. They are discrete, but still quantitative, since arithmetic operations can be performed on these variables. We cannot perform arithmetic operations on variables such as race and smoking status, so they are categorical.

   (b) I agree with my classmate since *low* is a binary version of response variable, so it is directly based on the response variable.

   (c) Change race to categorial variable.

```
race<-factor(race)
levels(race)<-c("W", "B", "O")

##perform all possible regressions (1st order)
allreg <- regsubsets(bwt ~age+lwt+race+smoke+ptl+ht+ui+ftv,
                     data=data, nbest=9)
```

   Since race has 3 classes, we expect 2 coefficients associated with the race variable.

   i. The best model according to adjusted R^2 has the predictors *lwt*, *race* (both indicator variables), *smoke*, *ht*, and *ui*.

```
(Intercept)         lwt       raceB       raceO       smoke          ht 
 2837.26392     4.24155  -475.05760  -348.15038  -356.32095  -585.19312 
         ui 
 -525.52390 
```

   ii. The best model according to C_p has the predictors *lwt*, *race* (both indicator variables), *smoke*, *ht*, and *ui*.

   iii. The best model according to BIC has the predictors *lwt*, *race* (both indicator variables), *smoke*, *ht*, and *ui*.
   Note that in this question, all criteria lead to the same model.

(d)

```
> regnull <- lm(bwt~1, data=data)
> regfull <- lm(bwt~age+lwt+race+smoke+ptl+ht+ui+ftv, data=data)
>
> step(regfull, scope=list(lower=regnull, upper=regfull), direction="backward")
Start:  AIC=2458.21
bwt ~ age + lwt + race + smoke + ptl + ht + ui + ftv

        Df Sum of Sq      RSS    AIC
- ftv    1     38708 75741025 2456.3
- age    1     58238 75760555 2456.3
- ptl    1     95285 75797602 2456.4
<none>               75702317 2458.2
- lwt    1   2661604 78363921 2462.7
- ht     1   3631032 79333349 2465.1
- smoke  1   4623219 80325536 2467.4
- race   2   6578597 82280914 2470.0
- ui     1   5839544 81541861 2470.2

Step:  AIC=2456.3
bwt ~ age + lwt + race + smoke + ptl + ht + ui

        Df Sum of Sq      RSS    AIC
- age    1     79115 75820139 2454.5
- ptl    1     91560 75832585 2454.5
<none>               75741025 2456.3
- lwt    1   2623988 78365013 2460.7
- ht     1   3592430 79333455 2463.1
- smoke  1   4606425 80347449 2465.5
- race   2   6552496 82293521 2468.0
- ui     1   5817995 81559020 2468.3

Step:  AIC=2454.5
bwt ~ lwt + race + smoke + ptl + ht + ui

        Df Sum of Sq      RSS    AIC
- ptl    1    117366 75937505 2452.8
<none>               75820139 2454.5
- lwt    1   2545892 78366031 2458.7
- ht     1   3546591 79366731 2461.1
- smoke  1   4530009 80350149 2463.5
- race   2   6571668 82391807 2466.2
- ui     1   5751122 81571261 2466.3

Step:  AIC=2452.79
```

```
bwt ~ lwt + race + smoke + ht + ui

        Df Sum of Sq      RSS    AIC
<none>                75937505 2452.8
- lwt    1   2674229 78611734 2457.3
- ht     1   3584838 79522343 2459.5
- smoke  1   4950633 80888138 2462.7
- race   2   6630123 82567628 2464.6
- ui     1   6353218 82290723 2466.0

Call:
lm(formula = bwt ~ lwt + race + smoke + ht + ui, data = data)

Coefficients:
(Intercept)          lwt        raceB        raceO        smoke           ht
   2837.264        4.242     -475.058     -348.150     -356.321     -585.193
         ui
   -525.524
```

So the regression equation is

$$y = 2837.26 + 4.24lwt - 475.06I_1 - 348.15I_2 - 356.32smoke - 585.19ht - 525.52ui$$

where $I_1 = 1$ for black mothers, $I_2 = 1$ for mothers of other races. White mothers are the reference class.

2. (a) Based on forward selection, the model selected had price, discount, and promo as the predictors.

   (b) i. The algorithm starts out with a model with none of the five predictors in the model.

      ii. The algorithm then considers adding one of the predictors into the model. For each predictor added, the algorithm calculates the AIC, which measures the model in terms of fit and complexity. A small value for the AIC is desirable.

      iii. The algorithm selects the predictor, which, after adding to the model, results in the following two situations: (i) results in the smallest AIC and (ii) reduces the AIC when compared to the previous model, which in this case is the model with none of the predictors. In this dataset, adding the predictor discount accomplishes both of these criteria. Once a predictor is added to the model, it is never removed in forward selection.

      iv. The algorithm repeats steps 2(b)ii and 2(b)iii, by considering adding a predictor to the model with discount in it.

      v. The algorithm stops when criteria (ii) in step 2(b)iii cannot be fulfilled by adding any of the remaining predictors.

(c) Some pieces of advice:

i. All possible regression only consider first order models. Need consultation with subject matter expert if models of higher order should be considered.

ii. Need to check with original purpose of study. Is there a specific predictor that has to be evaluated. Will this model answer the questions of interest?

iii. Automated search procedures should not be viewed as the end of the model building process. In fact, using other search procedures or with different starting points may yield in a different model.

iv. We need to verify the regression assumptions are met by examining residual plots, ACF plots of the residuals, and QQ plots of the residuals.

3. Advantage of R^2: has a nice geometrical interpretation. R^2 measures the proportion of the variance in the response variable that can be explained by our model. The adjusted R^2 does not have this interpretation

   Advantage of adjusted R^2: the addition of predictors that do not help further explain the response variable will lead to a decrease in adjusted R^2. On the other hand, R^2 always increases, regardless of whether the predictor helps further explain the variance in the response variable or not. So R^2 will always pick a larger model, while adjusted R^2 balances between model fit and simplicity.

4. An example code is below and I suspect most code will be similar.

```
PRESS <- function(linear.model) {
  ## calculate the predictive residuals
  pr <- residuals(linear.model)/(1-lm.influence(linear.model)$hat)
  ## calculate the PRESS
  PRESS <- sum(pr^2)

  return(PRESS)
}
```