Guided Question Set 9 Solutions

1)

```
Data<-read.table("nfl.txt", header=TRUE)
allreg <- regsubsets(y ~., data=Data, nbest=2)
summary(allreg)
```

```
## Subset selection object
## Call: regsubsets.formula(y ~ ., data = Data, nbest = 2)
## 9 Variables  (and intercept)
##    Forced in Forced out
## x1     FALSE      FALSE
## x2     FALSE      FALSE
## x3     FALSE      FALSE
## x4     FALSE      FALSE
## x5     FALSE      FALSE
## x6     FALSE      FALSE
## x7     FALSE      FALSE
## x8     FALSE      FALSE
## x9     FALSE      FALSE
## 2 subsets of each size up to 8
## Selection Algorithm: exhaustive
##          x1  x2  x3  x4  x5  x6  x7  x8  x9
## 1  ( 1 ) " " " " " " " " " " " " " " "*" " "
## 1  ( 2 ) "*" " " " " " " " " " " " " " " " "
## 2  ( 1 ) " " "*" " " " " " " " " " " "*" " "
## 2  ( 2 ) " " "*" " " " " " " " " "*" " " " "
## 3  ( 1 ) " " "*" " " " " " " " " "*" "*" " "
## 3  ( 2 ) "*" "*" " " " " " " " " " " "*" " "
## 4  ( 1 ) " " "*" " " " " " " " " "*" "*" "*"
## 4  ( 2 ) "*" "*" " " " " " " " " " " "*" "*"
## 5  ( 1 ) "*" "*" " " " " " " " " "*" "*" "*"
## 5  ( 2 ) " " "*" " " "*" " " " " "*" "*" "*"
## 6  ( 1 ) " " "*" "*" "*" " " " " "*" "*" "*"
## 6  ( 2 ) "*" "*" " " "*" " " " " "*" "*" "*"
## 7  ( 1 ) " " "*" "*" "*" " " "*" "*" "*" "*"
## 7  ( 2 ) "*" "*" " " "*" " " "*" "*" "*" "*"
## 8  ( 1 ) "*" "*" "*" "*" " " "*" "*" "*" "*"
## 8  ( 2 ) " " "*" "*" "*" "*" "*" "*" "*" "*"
```

2a)

The regression equation with the highest adjusted R^2 is $\hat{y} = -1.8217 + 0.0038x_2 + 0.2169x_7 - 0.0040x_8 - 0.0016x_9$.

```
coef(allreg, which.max(summary(allreg)$adjr2))
```

```
##  (Intercept)           x2           x7           x8           x9
## -1.821703427  0.003818572  0.216894094 -0.004014887 -0.001634926
```

2b)

The regression equation with the lowest Mallow’s C_p is $\hat{y} = -1.8084 + 0.0036x_2 + 0.1940x_7 - 0.0048x_8$.

```
coef(allreg, which.min(summary(allreg)$cp))
```

```
##  (Intercept)           x2           x7           x8
## -1.808372059  0.003598070  0.193960210 -0.004815494
```

2c)

The regression equation with the lowest BIC is $\hat{y} = -1.8084 + 0.0036x_2 + 0.1940x_7 - 0.0048x_8$.

```
coef(allreg, which.min(summary(allreg)$bic))
```

```
##  (Intercept)           x2           x7           x8
## -1.808372059  0.003598070  0.193960210 -0.004815494
```

3)

The regression equation from forward selection is $\hat{y} = -1.8217 + 0.0038x_2 + 0.2169x_7 - 0.0040x_8 - 0.0016x_9$.

```
##intercept only model
regnull <- lm(y~1, data=Data)
##model with all predictors
regfull <- lm(y~., data=Data)
step(regnull, scope=list(lower=regnull, upper=regfull), direction="forward")
```

```
## Start:  AIC=70.81
## y ~ 1
##
##        Df Sum of Sq    RSS    AIC
## + x8    1   178.092 148.87 50.785
## + x1    1   115.068 211.90 60.669
## + x7    1    97.238 229.73 62.931
## + x5    1    86.116 240.85 64.255
## + x2    1    76.193 250.77 65.385
## + x9    1    30.167 296.80 70.104
## <none>               326.96 70.814
## + x4    1    21.844 305.12 70.878
## + x6    1    16.411 310.55 71.372
## + x3    1     2.135 324.83 72.631
##
## Step:  AIC=50.78
## y ~ x8
##
##        Df Sum of Sq     RSS    AIC
## + x2    1    64.934  83.938 36.741
## + x5    1    11.607 137.265 50.512
## <none>               148.872 50.785
## + x1    1     6.636 142.236 51.508
## + x3    1     6.368 142.504 51.561
## + x4    1     6.345 142.527 51.565
## + x7    1     0.974 147.898 52.601
## + x6    1     0.487 148.385 52.693
## + x9    1     0.008 148.864 52.783
##
## Step:  AIC=36.74
## y ~ x8 + x2
##
##        Df Sum of Sq    RSS    AIC
## + x7    1   14.0682 69.870 33.604
## + x1    1   11.1905 72.748 34.734
## + x3    1    8.9010 75.037 35.602
## + x5    1    5.8147 78.124 36.730
## <none>               83.938 36.741
## + x9     1    2.0256 81.913 38.057
## + x6     1    1.3216 82.617 38.296
## + x4     1    0.0161 83.922 38.735
##
## Step:  AIC=33.6
## y ~ x8 + x2 + x7
##
##        Df Sum of Sq    RSS    AIC
## + x9    1    4.8657 65.004 33.583
## <none>              69.870 33.604
## + x3    1    1.3873 68.483 35.043
## + x4    1    0.9792 68.891 35.209
## + x1    1    0.9022 68.968 35.240
## + x6    1    0.4879 69.382 35.408
## + x5    1    0.2987 69.571 35.484
##
## Step:  AIC=33.58
## y ~ x8 + x2 + x7 + x9
##
##        Df Sum of Sq    RSS    AIC
## <none>              65.004 33.583
## + x1    1   1.86452 63.140 34.768
## + x4    1   1.74260 63.262 34.822
## + x3    1   0.70148 64.303 35.279
## + x6    1   0.45071 64.554 35.388
## + x5    1   0.32667 64.678 35.442
```

```
##
## Call:
## lm(formula = y ~ x8 + x2 + x7 + x9, data = Data)
##
## Coefficients:
## (Intercept)           x8           x2           x7           x9
##   -1.821703    -0.004015     0.003819     0.216894    -0.001635
```

4)

Backward elimination pick the same model as forward selection.

```
step(regfull, scope=list(lower=regnull, upper=regfull), direction="backward")
```

5)

Stepwise regression picks the same model as forward selection and backward elimination.

```
step(regnull, scope=list(lower=regnull, upper=regfull), direction="both")
```

6)

```
PRESS <- function(linear.model) {
  ## get the residuals from the linear.model.
  ## extract hat from lm.influence to obtain the leverages
  pr <- residuals(linear.model)/(1-lm.influence(linear.model)$hat)
  ## calculate the PRESS by squaring each term and adding them up
  PRESS <- sum(pr^2)

  return(PRESS)
}
```

7)

The PRESS statistic is 87.46. The $R^2_{prediction}$ is 0.7325. The R^2 is 0.7863.

The model might be able to explain 73.25% of the variability in the new observations. The $R^2$is 0.7863. Both values are fairly high and close to each other, so the model has good predictive ability.

```
result<-lm(y~x2+x7+x8, data=Data)
PRESS(result)
```

```
## [1] 87.46123
```

```
##Find SST
anova_result<-anova(result)
SST<-sum(anova_result$"Sum Sq")
##R2 pred
Rsq_pred<-1-PRESS(result)/SST
Rsq_pred
```

```
## [1] 0.7325052
```

```
summary(result)
```

```
##
## Call:
## lm(formula = y ~ x2 + x7 + x8, data = Data)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.0370 -0.7129 -0.2043  1.1101  3.7049
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.808372   7.900859  -0.229 0.820899
## x2           0.003598   0.000695   5.177 2.66e-05 ***
## x7           0.193960   0.088233   2.198 0.037815 *
## x8          -0.004816   0.001277  -3.771 0.000938 ***
## ---
## Signif. codes:  0 ’***’ 0.001 ’**’ 0.01 ’*’ 0.05 ’.’ 0.1 ’ ’ 1
##
## Residual standard error: 1.706 on 24 degrees of freedom
## Multiple R-squared:  0.7863, Adjusted R-squared:  0.7596
## F-statistic: 29.44 on 3 and 24 DF,  p-value: 3.273e-08
```