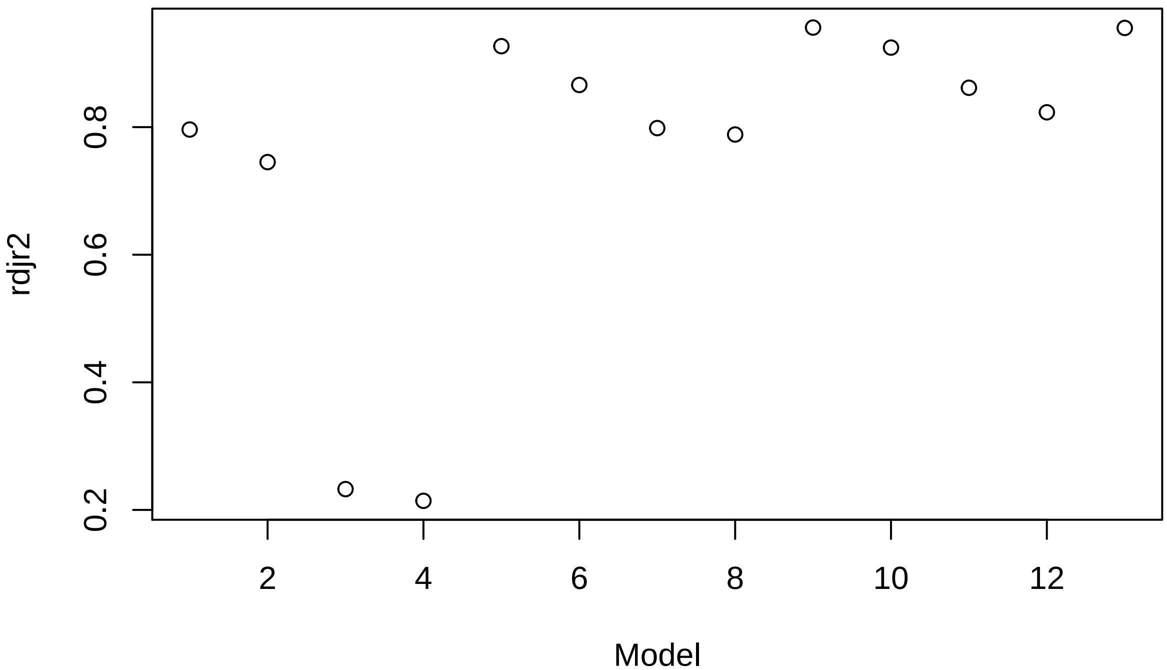
Homework Chapter 9

Jinhong Du 15338039

9.11. Refer to Job proficiency Problem 9.10.

a. Using only first-order terms for the predictor variables in the pool of potential X variables, find the four best subset regression models according to the $R^2_{a,p}$ criterion.

```
data1 <- read.table("CH09PR10.txt",head=FALSE,col.names = c('Y',
'X1','X2','X3','X4'))
library(leaps)
regfit.full = regsubsets(Y~.,data=data1,nbest = 4)
reg.summary=summary(regfit.full)
plot(reg.summary$adjr2,xlab="Model",ylab="rdjr2")
```




```
coef(regfit.full,order(reg.summary$adjr2,decreasing=TRUE)[1:4])
```

```
## [[1]]
##  (Intercept)          X1          X3          X4
## -124.2000166   0.2963260   1.3569675   0.5174211
##
## [[2]]
##   (Intercept)            X1            X2            X3            X4
## -124.38182058    0.29572537    0.04828772    1.30601100    0.51981909
##
## [[3]]
##  (Intercept)          X1          X3
## -127.5956876   0.3484575   1.8232055
##
## [[4]]
##   (Intercept)            X1            X2            X3
```

```
## -127.77378375    0.34813384    0.04353454    1.77921293
```

```r
reg.summary$adjr2[order(reg.summary$adjr2,decreasing=TRUE)[1:4]]
```

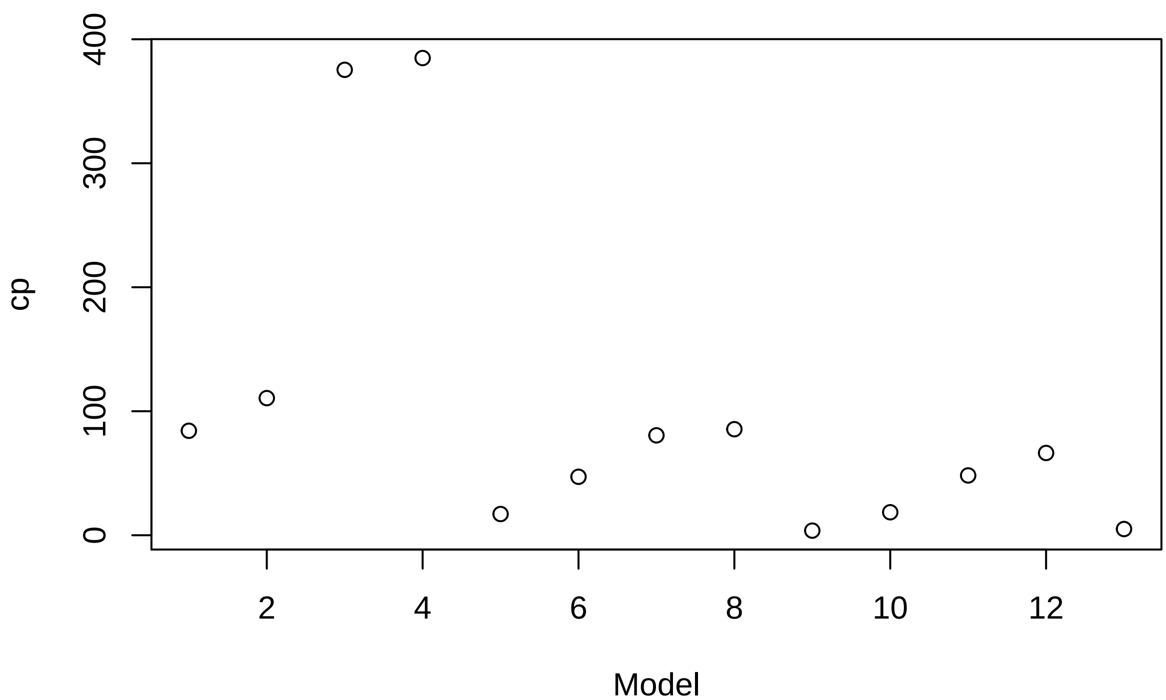
```
## [1] 0.9560482 0.9554702 0.9269043 0.9246779
```

$$R^2_{a,p} = 1 - \frac{MSE_p}{\frac{SSTO}{n-1}}$$

Top Model	$R^2_{a,p}$
X_1, X_3, X_4	0.9560482
X_1, X_2, X_3, X_4	0.9554702
X_1, X_3	0.9269043
X_1, X_2, X_3	0.9246779

b. Since there is relatively little difference in $R^2_{a,p}$ for the four best subset models, what other criteria would you use to help in the selection of the best model? Discuss.

```r
plot(reg.summary$cp,xlab="Model",ylab="cp")
```




```r
coef(regfit.full,order(reg.summary$cp,decreasing=FALSE)[1:4])
```

```
## [[1]]
##  (Intercept)           X1           X3           X4
## -124.2000166    0.2963260    1.3569675    0.5174211
##
## [[2]]
##   (Intercept)            X1            X2            X3            X4
## -124.38182058    0.29572537    0.04828772    1.30601100    0.51981909
##
## [[3]]
```

```
##  (Intercept)          X1          X3
## -127.5956876   0.3484575   1.8232055
##
## [[4]]
##   (Intercept)            X1            X2            X3
## -127.77378375   0.34813384   0.04353454   1.77921293
```

```
reg.summary$cp[order(reg.summary$cp,decreasing=FALSE)[1:4]]
```

```
## [1]  3.727399  5.000000 17.112978 18.521465
```

Top Model	C_p
X_1, X_3, X_4	3.727399
X_1, X_2, X_3, X_4	5.000000
X_1, X_3	17.112978
X_1, X_2, X_3	18.521465

C_p criteria considers both bias and variance, and estimate Γ_p. $\mathbb{E}C_p \approx p$ indicates a good model. Therefore, we should choose model X_1, X_2, X_3, X_4.

9.18. Refer to Job proficiency Problems 9.10 and 9.11.

a. Using forward stepwise regression, find the best subset of predictor variables to predict job proficiency. Use α limits of .05 and .10 for adding or deleting a variable, respectively.

```
lm.full <- lm(Y~.,data = data1)
lm.null <- lm(Y ~ 1, data = data1)
add1(lm.null, ~X1+X2+X3+X4,test='F')
```

```
## Single term additions
##
## Model:
## Y ~ 1
##        Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>              9054.0 149.30
## X1      1    2395.9 6658.1 143.62  8.2763  0.008517 **
## X2      1    2236.5 6817.5 144.21  7.5451  0.011487 *
## X3      1    7286.0 1768.0 110.47 94.7824 1.264e-09 ***
## X4      1    6843.3 2210.7 116.06 71.1978 1.699e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
## Now X3 is the least significant and p(X3)<0.05
drop1(update(lm.null, ~ . +X3), test = "F")
```

```
## Single term deletions
##
## Model:
## Y ~ X3
##        Df Sum of Sq  RSS    AIC F value    Pr(>F)
## <none>              1768 110.47
## X3      1      7286 9054 149.30  94.782 1.264e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Now p(X3)<0.10, no need to drop them.
add1(update(lm.null, ~ . +X3), scope = ~ X1+X2+X3+X4, test = "F")
```

```
## Single term additions
##
## Model:
## Y ~ X3
##        Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>              1768.02 110.469
## X1      1   1161.37  606.66  85.727  42.116 1.578e-06 ***
## X2      1     12.21 1755.81 112.295   0.153   0.69946
## X4      1    656.71 1111.31 100.861  13.001   0.00157 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Now X1 is the least significant and p(X1)<0.05
drop1(update(lm.null, ~ . +X3+X1), test = "F")
```

```
## Single term deletions
##
## Model:
## Y ~ X3 + X1
##        Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>               606.7  85.727
## X3      1    6051.5 6658.1 143.618 219.453 6.313e-13 ***
## X1      1    1161.4 1768.0 110.469  42.116 1.578e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Now p(X1)<0.10 and p(X3)<0.10, no need to drop them.
add1(update(lm.null, ~ . +X3+X1), scope = ~ X1+X2+X3+X4, test = "F")
```

```
## Single term additions
##
## Model:
## Y ~ X3 + X1
##        Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>              606.66 85.727
## X2      1     9.937 596.72 87.314  0.3497 0.5605965
## X4      1   258.460 348.20 73.847 15.5879 0.0007354 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Now X4 is the least significant and p(X4)<0.05
drop1(update(lm.null, ~ . +X3+X1+X4), test = "F")
```

```
## Single term deletions
##
## Model:
## Y ~ X3 + X1 + X4
##        Df Sum of Sq     RSS     AIC F value    Pr(>F)
## <none>               348.20  73.847
## X3      1   1324.39 1672.59 111.081  79.875 1.334e-08 ***
## X1      1    763.12 1111.31 100.861  46.024 1.040e-06 ***
## X4      1    258.46  606.66  85.727  15.588 0.0007354 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
## Now p(X1)<0.10, p(X3)<0.10 and p(X4)<0.10, no need to drop them.
add1(update(lm.null, ~ . +X3+X1+X4), scope = ~ X1+X2+X3+X4, test = "F")
```

```
## Single term additions
##
## Model:
## Y ~ X3 + X1 + X4
##        Df Sum of Sq    RSS    AIC F value Pr(>F)
## <none>               348.20 73.847
## X2      1     12.22 335.98 74.954  0.7274 0.4038
```

```r
## Now X2 is the least significant and p(X2)>0.05, no need to add it.
## Therefore, no new variables can be entered and no old variables need to be removed.
## The regression process stop.
```

The best model given by forward stepwise regression is $Y \sim \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4$

b. How does the best subset according to forward stepwise regression compare with the best subset according to the $R^2_{a,p}$ criterion obtained in Problem 9.11a?

It is the same.

9.22. Refer to Job proficiency Pmblems 9.10 and 9.18. To assess externally the validity of the regression model identified in Problem 9.18. 25 additional applicants for entry-level clerical positions in the agency were similarly tested and hired irrespective of their test scores.

b. Fit the regression model identified in Problem 9.18a to the validation data set. Compare the estimated regression coefficients and their estimated standard deviations to those obtained in Problem 9.18a. Also compare the error mean squares and coefficients of multiple determination. Do the estimates for the validation data set appear to be reasonably similar to those obtained for the model-building data set?

```r
data2 <- read.table("CH09PR22.txt",head=FALSE,col.names = c('Y',
'X1','X2','X3','X4'))
lm_val <- lm(Y~X1+X3+X4,data=data2)
lm_train <- lm(Y~X1+X3+X4,data=data1)
summary(lm_val)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3 + X4, data = data2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4619 -2.3836  0.6834  2.1123  7.2394
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -122.76705   11.84783 -10.362 1.04e-09 ***
## X1             0.31238    0.04729   6.605 1.54e-06 ***
## X3             1.40676    0.23262   6.048 5.31e-06 ***
## X4             0.42838    0.19749   2.169   0.0417 *
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.284 on 21 degrees of freedom
## Multiple R-squared:  0.9489, Adjusted R-squared:  0.9416
## F-statistic:   130 on 3 and 21 DF,  p-value: 1.017e-13
```

```
summary(lm_train)
```

```
##
## Call:
## lm(formula = Y ~ X1 + X3 + X4, data = data1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.4579 -3.1563 -0.2057  1.8070  6.6083
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -124.20002    9.87406 -12.578 3.04e-11 ***
## X1             0.29633    0.04368   6.784 1.04e-06 ***
## X3             1.35697    0.15183   8.937 1.33e-08 ***
## X4             0.51742    0.13105   3.948 0.000735 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.072 on 21 degrees of freedom
## Multiple R-squared:  0.9615, Adjusted R-squared:  0.956
## F-statistic:   175 on 3 and 21 DF,  p-value: 5.16e-15
```

```
MSE_train <- sum(lm_train$residuals^2)/lm_train$df.residual
MSE_val <- sum(lm_val$residuals^2)/lm_val$df.residual
MSE_train
```

```
## [1] 16.58081
```

```
MSE_val
```

```
## [1] 18.35493
```

```
lm_train.aov <- anova(lm_train)
lm_val.aov <- anova(lm_val)
1 - MSE_train * lm_train$df.residual / sum(lm_train.aov[, 2])
```

```
## [1] 0.9615422
```

```
1 - MSE_val*lm_val$df.residual / sum(lm_val.aov[, 2])
```

```
## [1] 0.948888
```

	Train	Val
b_0	-124.20002	-122.76705
b_1	0.29633	0.31238
b_3	1.35697	1.40676
b_4	0.51742	0.42838
$s\{b_0\}$	9.87406	11.84783
$s\{b_1\}$	0.04368	0.04729
$s\{b_3\}$	0.15183	0.23262
$s\{b_4\}$	0.13105	0.19749

	Train	Val
MSE	16.58081	18.35493
R^2	0.9615422	0.948888

c. Calculate the mean squared prediction error in (9.20) and compare it to MSE obtained form the model-building date set. Is there evidence of a substantial bias problem in MSE here?

```
lm.fit <- lm(Y~X1+X3+X4,data=data1)
lm.MSE <- sum(lm.fit$residuals^2)/lm.fit$df.residual
data4 <- data.frame(X1=data2$X1,X4=data2$X3,X5=data2$X4)
lm.predMSE <- sum((predict(lm.fit,data2)-data2$Y)^2)/length(data2$Y)
lm.predMSE
```

```
## [1] 15.70972
```

```
lm.MSE
```

```
## [1] 16.58081
```

The mean squared prediction error in validation set is close to MSE in training set. Therefore, there is no substantial bias problem in MSE here.