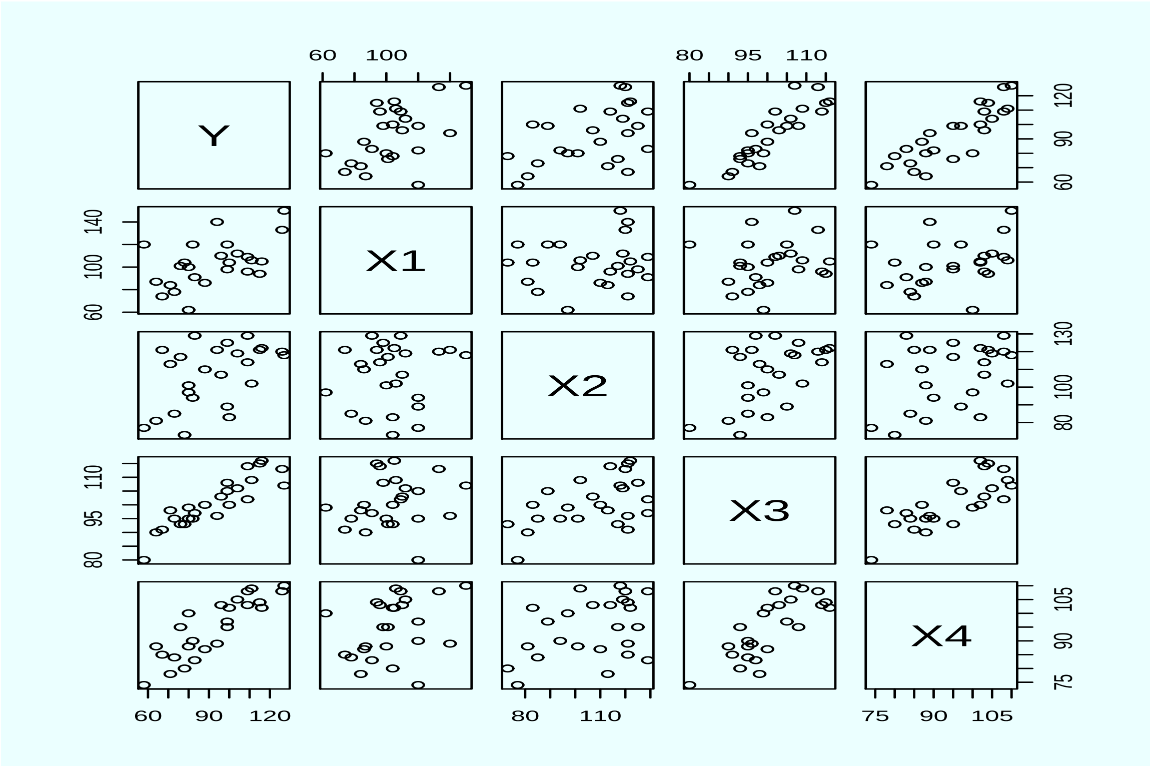
**MA 542 REGRESSION ANALYSIS**
**SPRING 2018**
HW - 9 - Solution Key

1. (Chapter 9 question 10)
b)




```
> cor(X)
          X1        X2        X3        X4
X1 1.0000000 0.1022689 0.1807692 0.3266632
X2 0.1022689 1.0000000 0.5190448 0.3967101
X3 0.1807692 0.5190448 1.0000000 0.7820385
X4 0.3266632 0.3967101 0.7820385 1.0000000
```

From the scatter plot matrix, it can be seen that the third and the Fourth test scores have strong positive linear relationships with the job proficiency score. The first and the second test scores have week relationships with the job proficiency score. From the correlation matrix it can be seen that the correlation between the third and the fourth test scores is high and that may be a serious multicolinearity problem.

c)

The fitted regression model is:

$$\hat{Y} = -124.38182 + 0.29573X_1 + 0.04829X_2 + 1.30601X_3 + 0.51982X_4$$

From the following summary output, we can see that the p-value for the t-test of the third test score is higher and so the third test score can be dropped from the model.

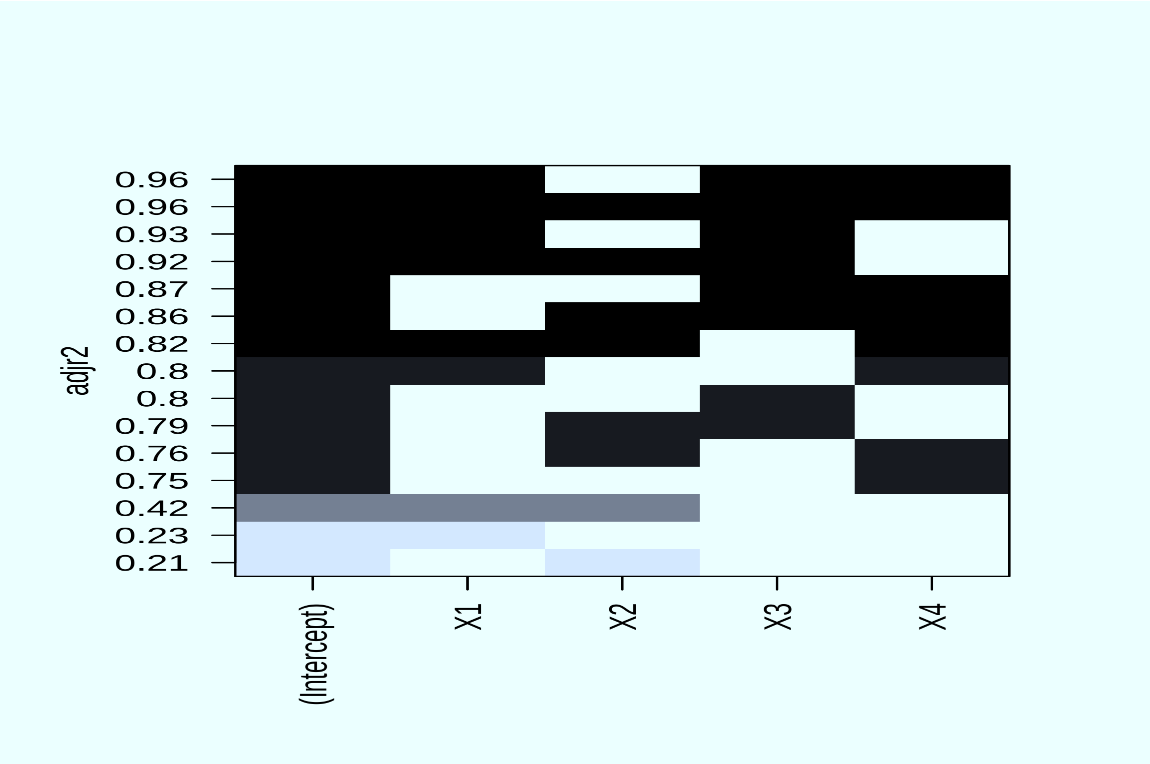
```
Call:
lm(formula = Y ~ X1 + X2 + X3 + X4, data = data)

Residuals:
    Min      1Q  Median      3Q     Max
-5.9779 -3.4506  0.0941  2.4749  5.9959

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -124.38182    9.94106 -12.512 6.48e-11 ***
X1             0.29573    0.04397   6.725 1.52e-06 ***
X2             0.04829    0.05662   0.853  0.40383
X3             1.30601    0.16409   7.959 1.26e-07 ***
X4             0.51982    0.13194   3.940  0.00081 ***
```

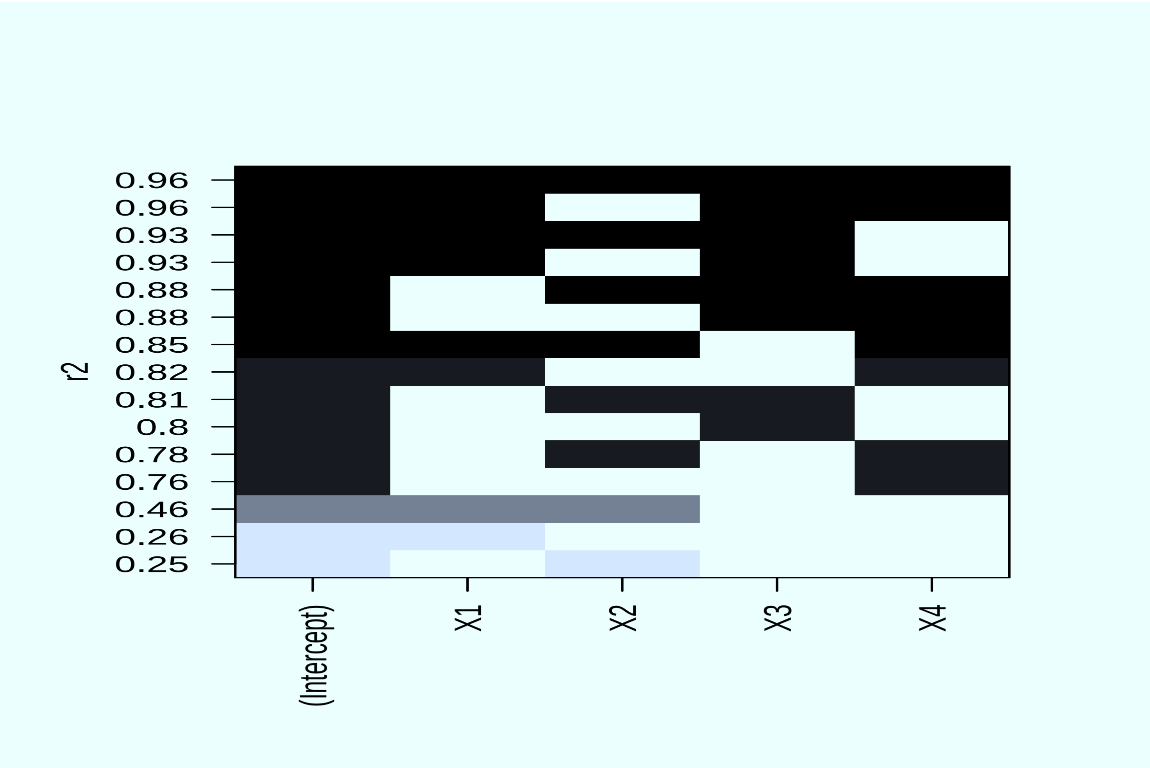
2. (Chapter 9 question 11)

a)




According to the plot, based on the $R^2_{a,p}$ criteria, the best four subsets are $(X_1,\ X_3,\ X_4)$, $(X_1,\ X_2,\ X_3,\ X_4)$, $(X_1,\ X_3,)$, $(X_1,\ X_2,\ X_3)$.

b) Since there is an evidence of multicollinearity between $X_3$ and $X_4$, I would use only one of them in the model to avoid the problem. So I would choose the model with $X_1$ and $X_3$ as the best model. We also can use $R^2_P$ as the selection criteria which gives the same decision ($R^2_p$ values for the four best models are close).



r2
0.96
0.96
0.93
0.93
0.88
0.88
0.85
0.82
0.81
0.8
0.78
0.76
0.46
0.26
0.25
(Intercept)
X1
X2
X3
X4

3. (Chapter 9 question 18)

```
> null=lm(Y~1, data=data)
>
> full=lm(Y~X1+X2+X3+X4,data=data)
>
> step(null, scope = list(upper=full), data=data, direction="both",levels='(0.05,0.01)')
Start:  AIC=149.3
Y ~ 1

       Df Sum of Sq    RSS    AIC
+ X3    1    7286.0 1768.0 110.47
+ X4    1    6843.3 2210.7 116.06
+ X1    1    2395.9 6658.1 143.62
+ X2    1    2236.5 6817.5 144.21
<none>              9054.0 149.30

Step:  AIC=110.47
Y ~ X3

       Df Sum of Sq    RSS     AIC
+ X1    1    1161.4  606.7  85.727
+ X4    1     656.7 1111.3 100.861
<none>              1768.0 110.469
+ X2    1      12.2 1755.8 112.295
- X3    1    7286.0 9054.0 149.302

Step:  AIC=85.73
Y ~ X3 + X1

       Df Sum of Sq    RSS     AIC
+ X4    1     258.5  348.2  73.847
<none>               606.7  85.727
+ X2    1       9.9  596.7  87.314
- X1    1    1161.4 1768.0 110.469
- X3    1    6051.5 6658.1 143.618

Step:  AIC=73.85
Y ~ X3 + X1 + X4

       Df Sum of Sq     RSS     AIC
<none>               348.20  73.847
+ X2    1     12.22  335.98  74.954
- X4    1    258.46  606.66  85.727
- X1    1    763.12 1111.31 100.861
- X3    1   1324.39 1672.59 111.081

Call:
lm(formula = Y ~ X3 + X1 + X4, data = data)

Coefficients:
(Intercept)           X3           X1           X4
  -124.2000       1.3570       0.2963       0.5174
```

So using forward stepwise regression, the best subset of predictors is $(X_1,\ X_3,\ X_4)$.

b) For this problem, both method chose the same set $(X_1,\ X_3,\ X_4)$ as the best subset.

4. (Chapter 9 question 21)

```
> PRESS=0
> for (i in 1 : dim(data)[1]){
+   newdata=data[-i,] #data frame without the ith row.
+   newmodel=lm(Y~X1+X3+X4,data=newdata)
+   newx=data.frame(X1=X1[i],X3=X3[i],X4=X4[i])
+   pred=predict.lm(newmodel,newx)[1]
+   PRESS=PRESS+(pred-data$Y[i])^2
+ }
> PRESS
      1
471.452
```

```
> fit2=lm(Y~X1+X3+X4,data=data)
> anova(fit2)
Analysis of Variance Table

Response: Y
          Df Sum Sq Mean Sq F value    Pr(>F)
X1         1 2395.9  2395.9 144.496 7.054e-11 ***
X3         1 6051.5  6051.5 364.969 9.359e-15 ***
X4         1  258.5   258.5  15.588 0.0007354 ***
Residuals 21  348.2    16.6
```

So $PRESS = 471.452$ and $SSE = 348.2$. Since $PRESS$ is comparatively larger than $SSE$, the validity of $MSE$ as an indicator of the predictive ability of the fitted model is low.

5. (Chapter 9 question 22)

a)

```
> cor(Valid_X)
           TX1        TX2       TX3       TX4
TX1 1.00000000 0.01057088 0.1772891 0.3196395
TX2 0.01057088 1.00000000 0.3437441 0.2207638
TX3 0.17728907 0.34374413 1.0000000 0.8714466
TX4 0.31963945 0.22076377 0.8714466 1.0000000
```

Yes two correlation matrices are reasonably similar.

b) The following is the summary output for the selected model for the validation data set.

```
> summary (VFIT)

Call:
lm(formula = TY ~ TX1 + TX3 + TX4, data = Vdata)

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 ***
TX1            0.31238    0.04729   6.605 1.54e-06 ***
TX3            1.40676    0.23262   6.048 5.31e-06 ***
TX4            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
```

The following table shows the comparison of coefficients for two models.

| | Model Building data set | Validation data set |
|---|---|---|
| $b_0$ | -124.20002 | -122.76705 |
| $S\{b_0\}$ | 9.87406 | 11.84783 |
| $b_1$ | 0.29633 | 0.31238 |
| $S\{b_1\}$ | 0.04368 | 0.04729 |
| $b_3$ | 1.35697 | 1.40676 |
| $S\{b_3\}$ | 0.15183 | 0.23262 |
| $b_4$ | 0.51742 | 0.42838 |
| $S\{b_4\}$ | 0.13105 | 0.19749 |
| MSE | 16.58081 | 18.35493 |
| $R^2$ | 0.9615 | 0.9489 |

As we can see from the table, all the coefficients and the standard errors are approximately same for both models. $MSE$ and the $R^2$ values are also similar.

c)

The mean squared prediction error is 15.71 and that is very close to the MSE of the model for the model building data set. This is not same as the conclusion we made before for the previous question.

d)

```
> summary(Cmodel)

Call:
lm(formula = Y ~ X1 + X3 + X4, data = total)

Residuals:
    Min      1Q  Median      3Q     Max
-9.7192 -2.7369  0.1278  2.0971  7.0657

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -123.44104    7.16508 -17.228  < 2e-16 ***
X1             0.30364    0.03072   9.886 5.86e-13 ***
X3             1.36906    0.12280  11.148 1.15e-14 ***
X4             0.48735    0.10475   4.652 2.79e-05 ***
---
Signif. codes:  0   ***   0.001   **   0.01   *   0.05   .   0.1       1

Residual standard error: 4.006 on 46 degrees of freedom
Multiple R-squared:  0.9567,    Adjusted R-squared:  0.9539
F-statistic: 338.9 on 3 and 46 DF,  p-value: < 2.2e-16
```

So the fitted model for the combined data set is :

$$\hat{Y} = -123.44104 + 0.30364X_1 + 1.36906X_3 + 0.48735X_4.$$

Yes the estimated standard deviations of estimated coefficients are reduced. This is because the more data are considered now.