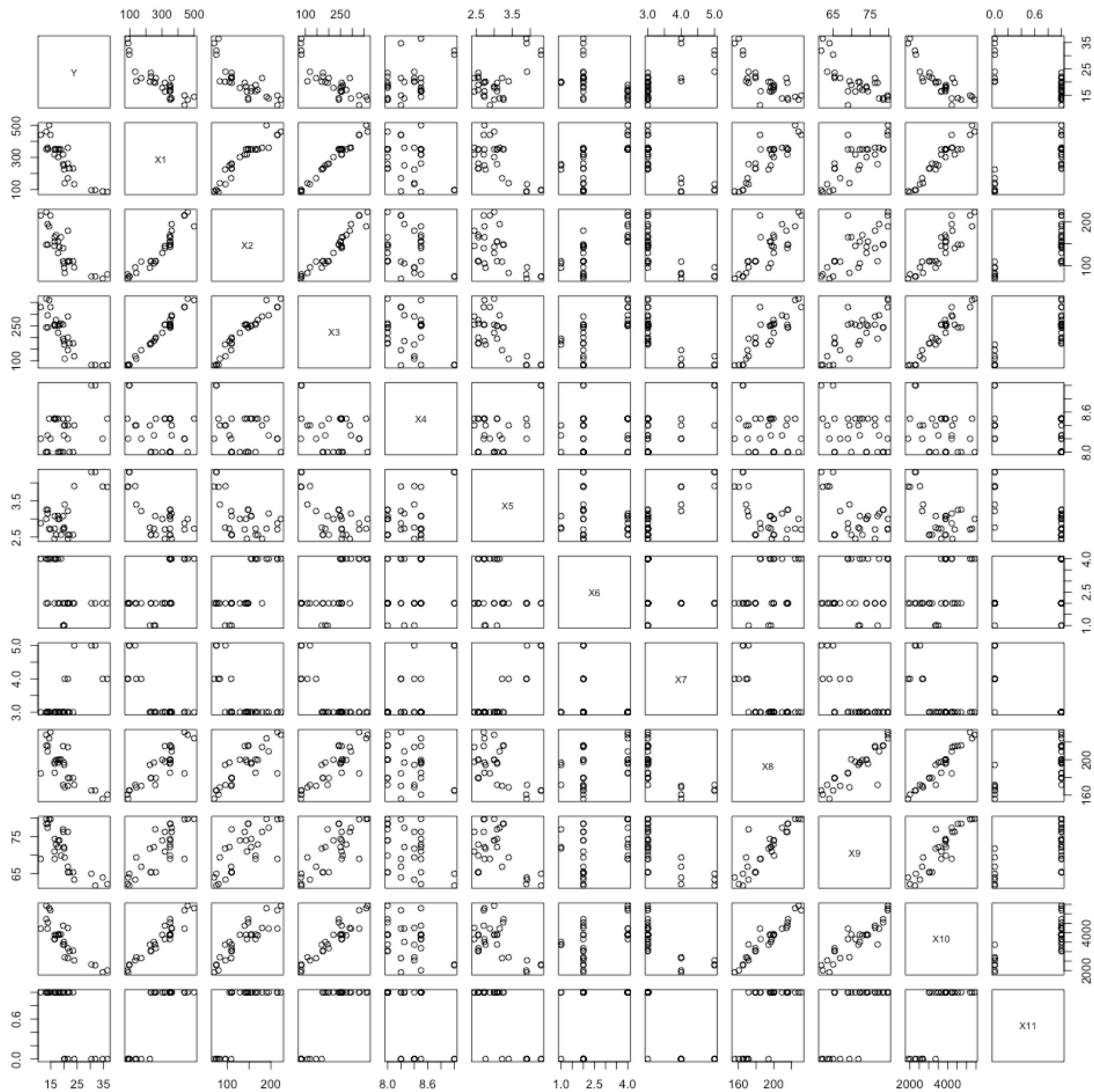
HW 5 | Shawn Xiang Li

Problem 1 (Textbook 9.3)
(a)

	Y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11
Y	1.0000000	-0.8718188	-0.7965605	-0.8493416	0.42241460	0.6352323	-0.47192100	0.7078714	-0.7523967	-0.7624550	-0.8525706	-0.7216882
X1	-0.8718188	1.0000000	0.9406456	0.9895851	-0.34958682	-0.6714311	0.63996417	-0.7717815	0.8649023	0.8001582	0.9531271	0.8241409
X2	-0.7965605	0.9406456	1.0000000	0.9643592	-0.28989951	-0.5509642	0.76141897	-0.6259445	0.8027387	0.7105117	0.8878810	0.7086735
X3	-0.8493416	0.9895851	0.9643592	1.0000000	-0.32599915	-0.6728661	0.65312630	-0.7461800	0.8641224	0.7881284	0.9434871	0.8012765
X4	0.4224146	-0.3495868	-0.2898995	-0.3259992	1.00000000	0.4137808	0.03748643	0.5582357	-0.3041503	-0.3781736	-0.3584588	-0.4405457
X5	0.6352323	-0.6714311	-0.5509642	-0.6728661	0.41378081	1.0000000	-0.21952829	0.8717662	-0.5613315	-0.4534470	-0.5798617	-0.7546650
X6	-0.4719210	0.6399642	0.7614190	0.6531263	0.03748643	-0.2195283	1.00000000	-0.2756386	0.4220680	0.3003862	0.5203669	0.3954893
X7	0.7078714	-0.7717815	-0.6259445	-0.7461800	0.55823570	0.8717662	-0.27563863	1.0000000	-0.6552065	-0.6551300	-0.7058126	-0.8506963
X8	-0.7523967	0.8649023	0.8027387	0.8641224	-0.30415026	-0.5613315	0.42206800	-0.6552065	1.0000000	0.8831512	0.9554541	0.6824919
X9	-0.7624550	0.8001582	0.7105117	0.7881284	-0.37817358	-0.4534470	0.30038618	-0.6551300	0.8831512	1.0000000	0.8994711	0.6326677
X10	-0.8525706	0.9531271	0.8878810	0.9434871	-0.35845879	-0.5798617	0.52036693	-0.7058126	0.9554541	0.8994711	1.0000000	0.7530353
X11	-0.7216882	0.8241409	0.7086735	0.8012765	-0.44054570	-0.7546650	0.39548928	-0.8506963	0.6824919	0.6326677	0.7530353	1.0000000



HW 5 | Shawn Xiang Li

When two predictive variables have a high correlation and/or there is a linear trend in their scatter plot, these two variables have collinearity. We can identify some collinearites based evidence from correlation matrix and pairwise scatter plot. They are:
X1&x2; X1&x3, X1&x5 X1&x6 X1&x7 X1&x8 X1&x9 X1&x10 X1&x11 X2&X3 X2&X5 X2&X6 X2&X7 X2&X8 X2&X9 X2&X10 X2&X11 X3&X5 X3&X6 X3&X7 X3&X8 X3&X9 X3&X10 X7&X8 X7&X9 X7&X10 X7&X11 X8&X9 X8&X10 X8&X11 X10&X11

(d) If VIF(Xi) > 10, we consider Xi are affected by the presence of collinearity. Based on table below, we can see X1, X2, X3, X7, X8 and X10 are the cases.

```
> vif(fitted)
        X1         X2         X3         X4         X5         X6         X7         X8         X9
128.834832  43.921063 160.436093   2.057834   7.780750   5.326714  11.735038  20.585810   9.419449
       X10        X11
 85.675755   5.142547
```

Problem 2 (Textbook 9.4)

(a) Degree of freedom of error is n-(p+1); n is sample size and p is number of predictive variables in the model. So, in this question, the sample size is 21. Since df of error >= 0; p <= 20. Therefore, the maximum number of terms in a linear regression model that I can fit to these data.

(b) The model is: fitted2<-lm(V~Year+I+D1+D2+W+G:I+P+N+P:N+W:D). Correlation matric, VIFs and Conditional Number are showed below.

```
> cor(a)
           Year          I           D          W          G           P          N
Year  1.0000000 -0.2046969 -0.11637147 -0.3146266  0.3085929 -0.18482213 -0.3161760
I    -0.2046969  1.0000000  0.81744307  0.3892495  0.1369564  0.11921266  0.2650860
D    -0.1163715  0.8174431  1.00000000  0.2876780  0.3230490 -0.07290826  0.2835083
W    -0.3146266  0.3892495  0.28767798  1.0000000 -0.2168366  0.64831150  0.2718636
G     0.3085929  0.1369564  0.32304903 -0.2168366  1.0000000 -0.58368979  0.2617113
P    -0.1848221  0.1192127 -0.07290826  0.6483115 -0.5836898  1.00000000 -0.1670507
N    -0.3161760  0.2650860  0.28350827  0.2718636  0.2617113 -0.16705075  1.0000000
```

```
> sqrt(kappa(cor(a),exact=TRUE))
[1] 4.05107
```

the conditional number is 4.05107.

```
> vif(fitted1)
    Year         I         D         W         P         N       I:G       I:D       P:N
1.682418  4.386550  4.297923 13.255400 24.563454  9.501884  2.099456  1.368038 27.249283
     D:W
8.600832
```

Problem 3 (Textbook 9.5)

(a) The model is: fitted2<-lm(V~Year+I+D1+D2+W+G:I+P+N). Correlation matric, VIFs and Conditional Number are showed below.

```
> cor(b)
             I          W           P          N         D1          D2        G...I
I    1.0000000  0.3892495  0.11921266  0.2650860  0.7479576 -0.66332496  0.20560659
W    0.3892495  1.0000000  0.64831150  0.2718636  0.2401922 -0.25819889 -0.18685769
P    0.1192127  0.6483115  1.00000000 -0.1670507 -0.1036814  0.01942015 -0.38056872
N    0.2650860  0.2718636 -0.16705075  1.0000000  0.2824205 -0.20532004  0.29265096
D1   0.7479576  0.2401922 -0.10368142  0.2824205  1.0000000 -0.49613894  0.35454668
D2  -0.6633250 -0.2581989  0.01942015 -0.2053200 -0.4961389  1.00000000  0.02216247
G...I 0.2056066 -0.1868577 -0.38056872  0.2926510  0.3545467  0.02216247  1.00000000
```

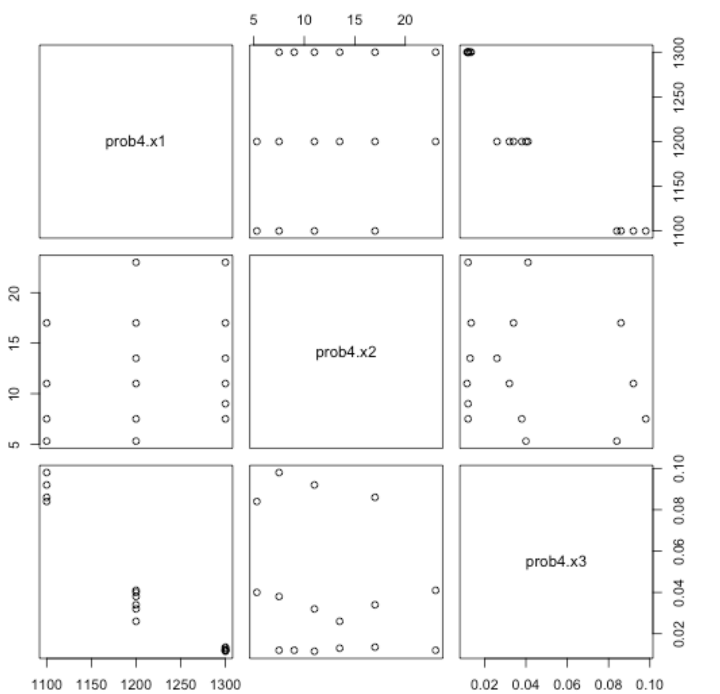
> sqrt(kappa(cor(b), exact = TRUE))
[1] 4.030008

The conditional number is 4.030008.

```
> vif(fitted2)
       I       D1       D2        W        P        N      I:G
3.555492 2.678628 2.026857 2.724643 2.612056 1.476388 1.535663
```

Problem 4

(a)




```
> cor(c)
           prob4.x1   prob4.x2   prob4.x3
prob4.x1  1.0000000  0.2236278 -0.9582041
prob4.x2  0.2236278  1.0000000 -0.2402310
prob4.x3 -0.9582041 -0.2402310  1.0000000
```

The absolute(correlation) between x1 and x3 is almost 1, which shows they are highly correlated. It indicated they have a multicollinearity. I can conclude the same finding from their scatter plot.

HW 5 | Shawn Xiang Li

(b) The model is fitted3<- lm(y~x1+x2+x3+x1*x2+x1*x3+x2*x3+I(x1^2)+I(x2^2)+I(x3^2)).

If VIF(Xi) > 10, we consider Xi are affected by the presence of collinearity. The VIFs of all variables are highly larger than 10, which shows highly multicollinearity exist in the model.

```
> vif(fitted3)
          x1           x2           x3      I(x1^2)      I(x2^2)      I(x3^2)        x1:x2        x1:x3        x2:x3
2.856749e+06 1.095614e+04 2.017163e+06 2.501945e+06 6.573359e+01 1.266710e+04 9.802903e+03 1.428092e+06 2.403594e+02
```

(c) After centred x1, x2 and x3, I run the regression again and get VIFs of the new model:

```
> vif(fitted4)
         x1center          x2center          x3center     I(x1center^2)     I(x2center^2)
       375.247759          1.740631        680.280039       1762.575365          3.164318
    I(x3center^2) x1center:x2center x1center:x3center x2center:x3center
      1156.766284         31.037059       6563.345193         35.611286
```

Comparing to the VIFs table from (b), we can see VIFs of all variables drop down. And VIF of x2 and x2^2 are less than 10,which shows there is no multicollinearity of them. Therefore, centering made the multicollinearity problem less severe.

(d) I don't think at least squares fit of the above model will give reliable results since some these variable has multicollinearity, we need to exclude all multicollinearity first.

Problem 5

(a)

Variables in Model	SSE_p	p	Error d.f.	MSE_p	Adj. r^2_p	C_p
None	950	0	19	50	0	20
x1	720	1	18	40	0.2	12.8
x2	630	1	18	35	0.3	9.2
x3	540	1	18	30	0.4	5.6
x1, x2	595	2	17	35	0.3	9.8
x1, x3	425	2	17	25	0.5	3
x2, x3	510	2	17	30	0.4	6.4
x1, x2, x3	400	3	16	25	0.5	4

(b)

Choosing the best model is based on maximizing adjusted R^2_p and minimizing C_p. So model with variables x1 and x3 is the best one.

(c)

Variables in Model	Fp(add 1st)
None	
x1	5.75
x2	9.142857143
x3	13.66666667

Since their Fp are all larger than 4; we choose the variable with the largest Fp, which is x3. Therefore, we add x3 into the model first. **F-to-enter is 13.667.**

(d) After X3 added in to the model, we are considering add second variable into the model.

Variables in Model	Fp(add 2nd)
x1,X3	4.6
x2,X3	1

After add X1, Fp is larger than 4. So we add X1 into the model.
F-to-enter is **4.6**
Ry,x1|x3 = 0.4614791

(e) Since after first variable X3 added into the model, we add second variable X1 into the model and we get the **partial F-test is 4.6**, which is larger than 4. So these two variable all significant in the model. So, we should not remove x3.

(f) Right now, we have x3 and x1 in the model, we wan to check if we should add x2 in the model also and get the full model. The **partial F statistics is 1**, which is less than 4. So X2 is not significant in the model. So we should add it into the model as 3rd variable. We should not choose the full model.