Semester Project
STAT 5532
Rongmin Xia

1. Basic analysis of data
    - Covariance
    - Correlation
    - Scatter plot
    - Histogram of each variable
    - Normality check, QQ-plot
2. Full model regression
    - Residual analysis
    - Equal variance assumption
    - Independent assumption check
3. Model Diagnostic
    - Multicollinearity
    - Principle-component regression analysis
4. Selecting a Regression Model
    - Forward
    - Stepwise
    - Variable selection(All possible Regressions)
5. Adequacy of the new models
    - Residual plot
    - Partial residual plots
    - Outlier detection, leverage and influence analysis
    - Lack of fit test
6. Inference of the parameters in the new model
7. Validation of new model
    - DUPLEX method to split dataset
8. Indicator variable analysis
9. CONCLUSION

1. Basic analysis of data

- Covariance

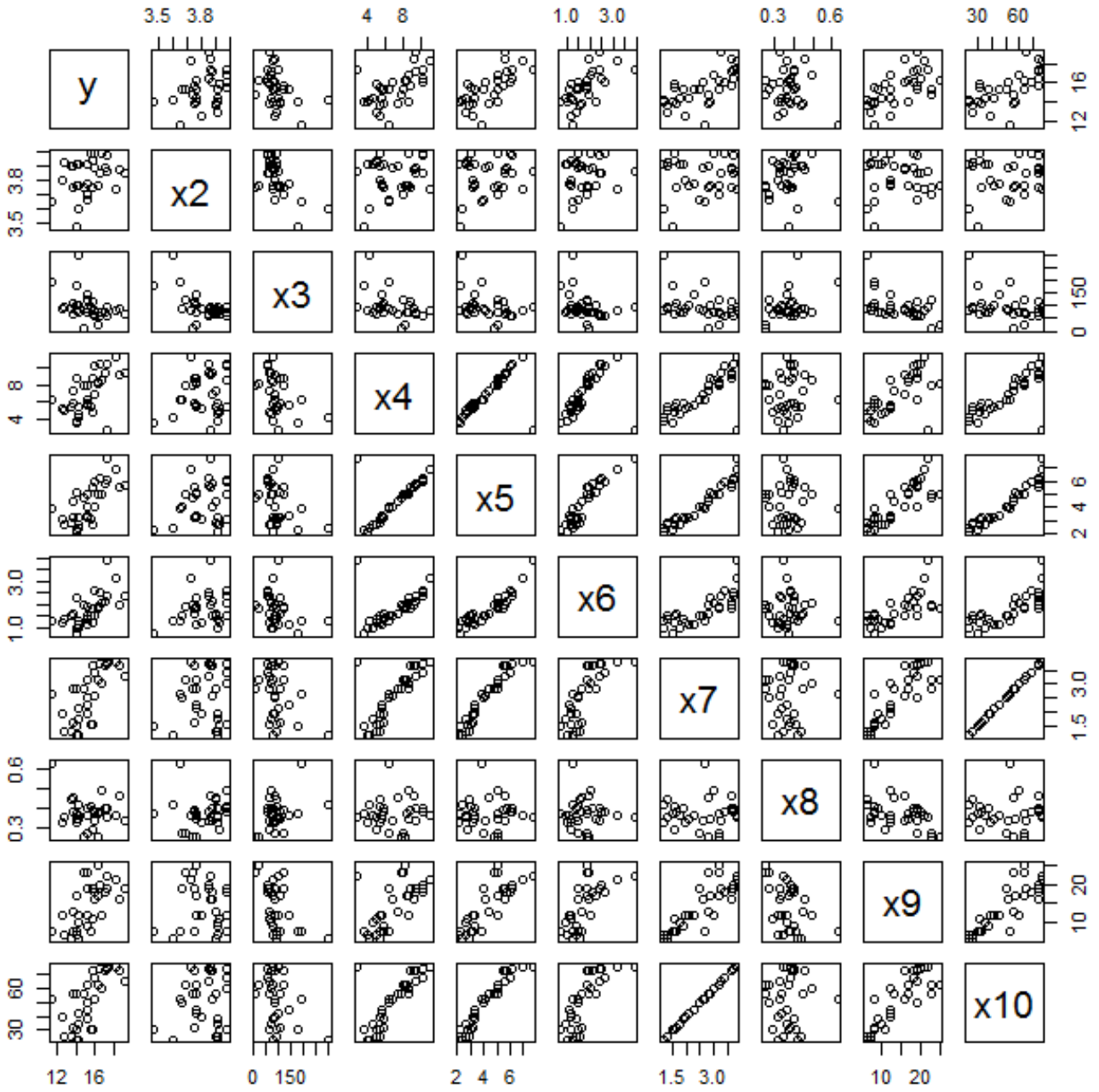
  y=quality; x1=varietal; x2=pH; x3=sulphates; x4=ClrDensity; x5=WineClr; x6=plyPigmentClr; x7=AnthColr; x8=Anthocyanin; x9=IonDegree; x10=IonAnthocyanin

	y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
y	3.14	-0.25	0.06	-34.71	2.26	1.85	0.77	1.08	-0.02	6.27	21.61
X1	-0.25	0.26	-0.01	3.22	0.10	-0.03	-0.06	0.03	0.00	0.42	0.60
X2	0.06	-0.01	0.01	-3.68	0.05	0.03	0.02	0.01	0.00	-0.03	0.19
X3	-34.71	3.22	-3.68	2719.92	-48.76	-28.61	-11.37	-17.24	1.59	-148.50	-344.80
X4	2.26	0.10	0.05	-48.76	4.95	2.34	0.79	1.56	0.01	8.20	31.13
X5	1.85	-0.03	0.03	-28.61	2.34	2.19	0.92	1.27	0.00	7.01	25.41
X6	0.77	-0.06	0.02	-11.37	0.79	0.92	0.45	0.47	0.00	2.66	9.36
X7	1.08	0.03	0.01	-17.24	1.56	1.27	0.47	0.80	0.00	4.36	16.05
X8	-0.02	0.00	0.00	1.59	0.01	0.00	0.00	0.00	0.01	-0.20	0.05
X9	6.27	0.42	-0.03	-148.50	8.20	7.01	2.66	4.36	-0.20	32.96	87.14
X10	21.61	0.60	0.19	-344.80	31.13	25.41	9.36	16.05	0.05	87.14	320.96

- Correlation

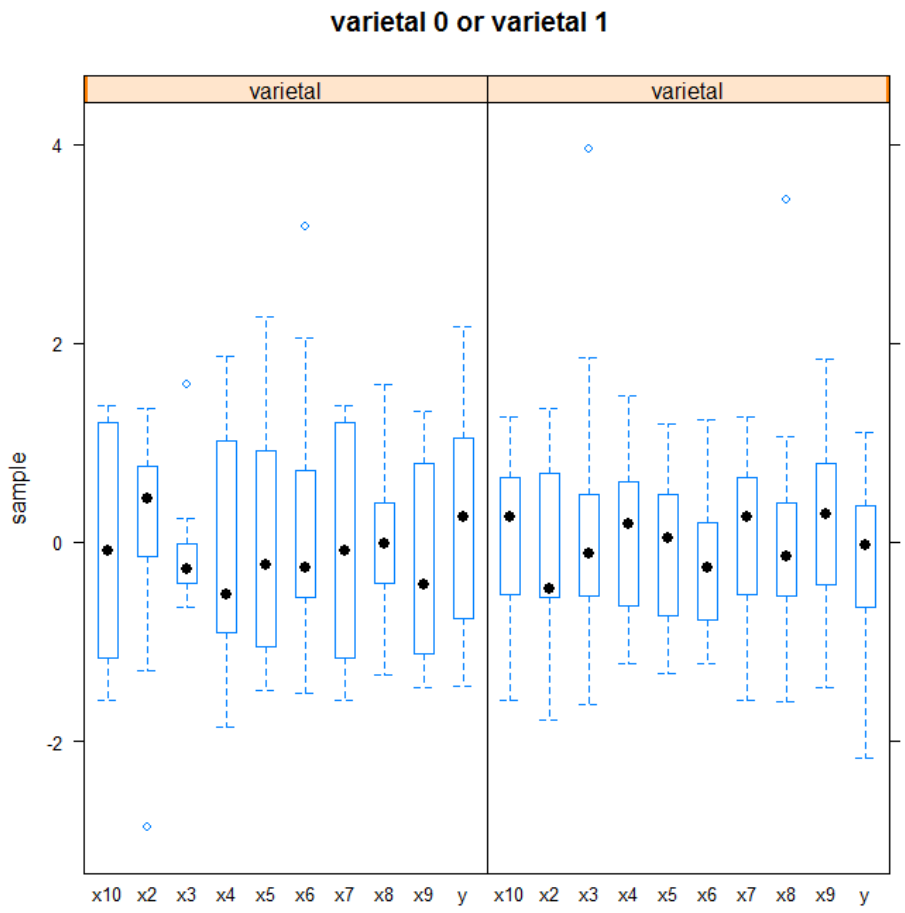
	y	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10
y	1	-0.282	0.277	-0.376	0.574	0.708	0.651	0.681	-0.168	0.617	0.681
X1	-0.282	1	-0.116	0.122	0.090	-0.042	-0.180	0.066	-0.026	0.145	0.066
X2	0.277	-0.116	1	-0.582	0.171	0.152	0.220	0.086	0.096	-0.049	0.086
X3	-0.376	0.122	-0.582	1	-0.420	-0.371	-0.325	-0.369	0.405	-0.496	-0.369
X4	0.574	0.090	0.171	-0.420	1	0.712	0.528	**0.781**	0.056	0.641	**0.781**
X5	0.708	-0.042	0.152	-0.371	0.712	1	**0.925**	**0.959**	0.003	0.826	**0.959**
X6	0.651	-0.180	0.220	-0.325	0.528	**0.925**	1	**0.780**	-0.043	0.691	0.780
X7	0.681	0.066	0.086	-0.369	0.781	0.959	0.780	1	0.037	**0.847**	**1.000**
X8	-0.168	-0.026	0.096	0.405	0.056	0.003	-0.043	0.037	1	-0.456	0.037
X9	0.617	0.145	-0.049	-0.496	0.641	0.826	0.691	0.847	-0.456	1	0.847
X10	0.681	0.066	0.086	-0.369	0.781	**0.959**	0.780	**1.000**	0.037	0.847	1

- Scatter plot

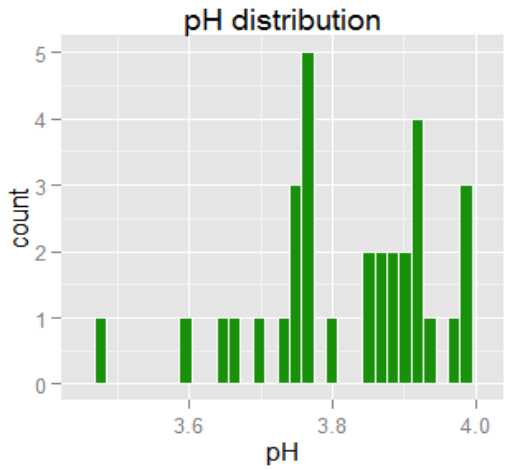


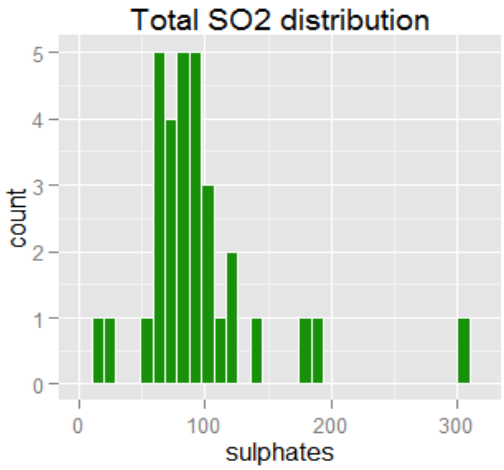
We can clearly see that x4, x5, x6, x8, x10 are highly correlated. Especially, the x7 with x10, and x4 with x5 are extremely correlated, which indicated that maybe x7with x10, and x4 with x5 are replaceable each other in the model.

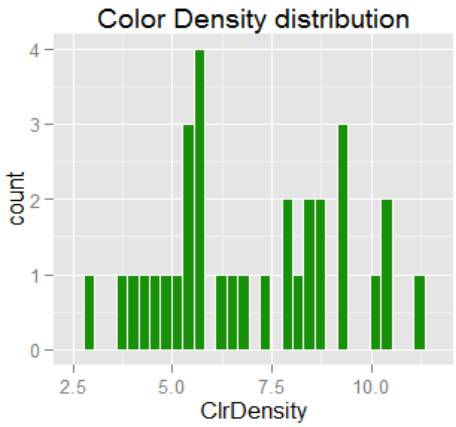
- Standardized box plot

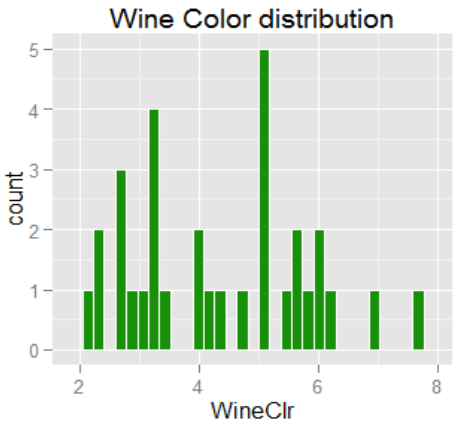


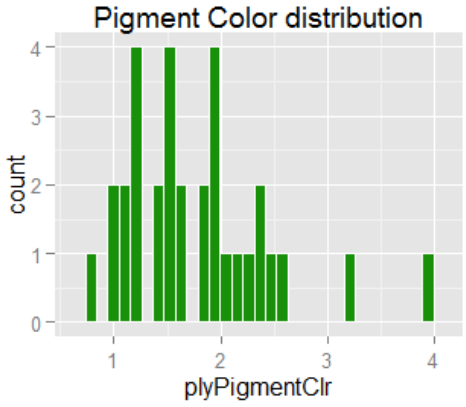
- Histogram of each variable

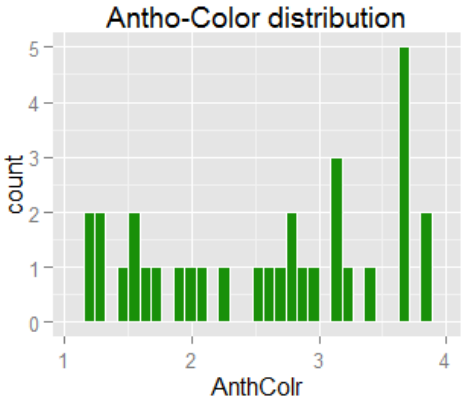


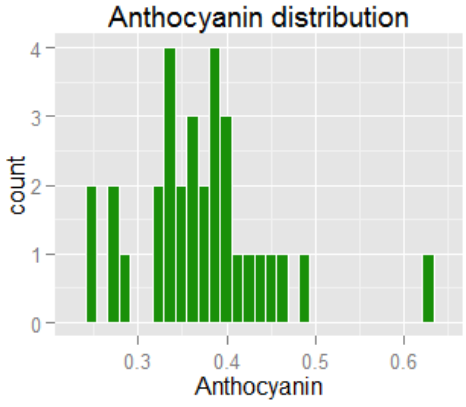


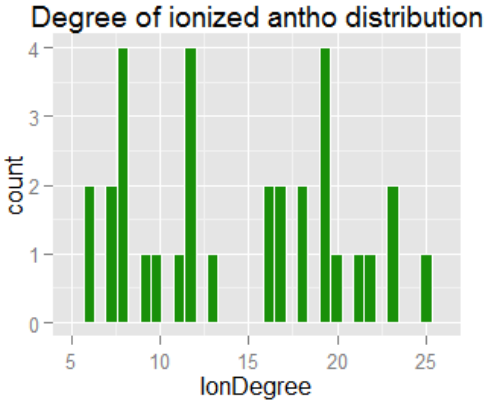

Color Density distribution
count
ClrDensity

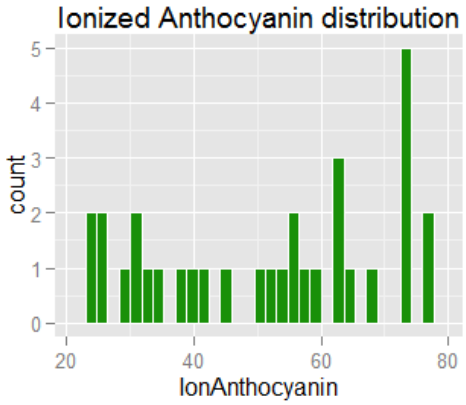

Wine Color distribution
count
WineClr

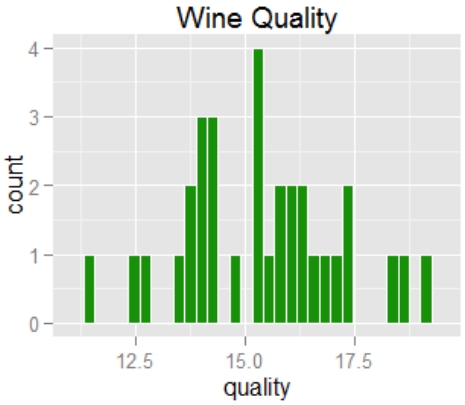

Pigment Color distribution
count
plyPigmentClr


Antho-Color distribution
count
AnthColr

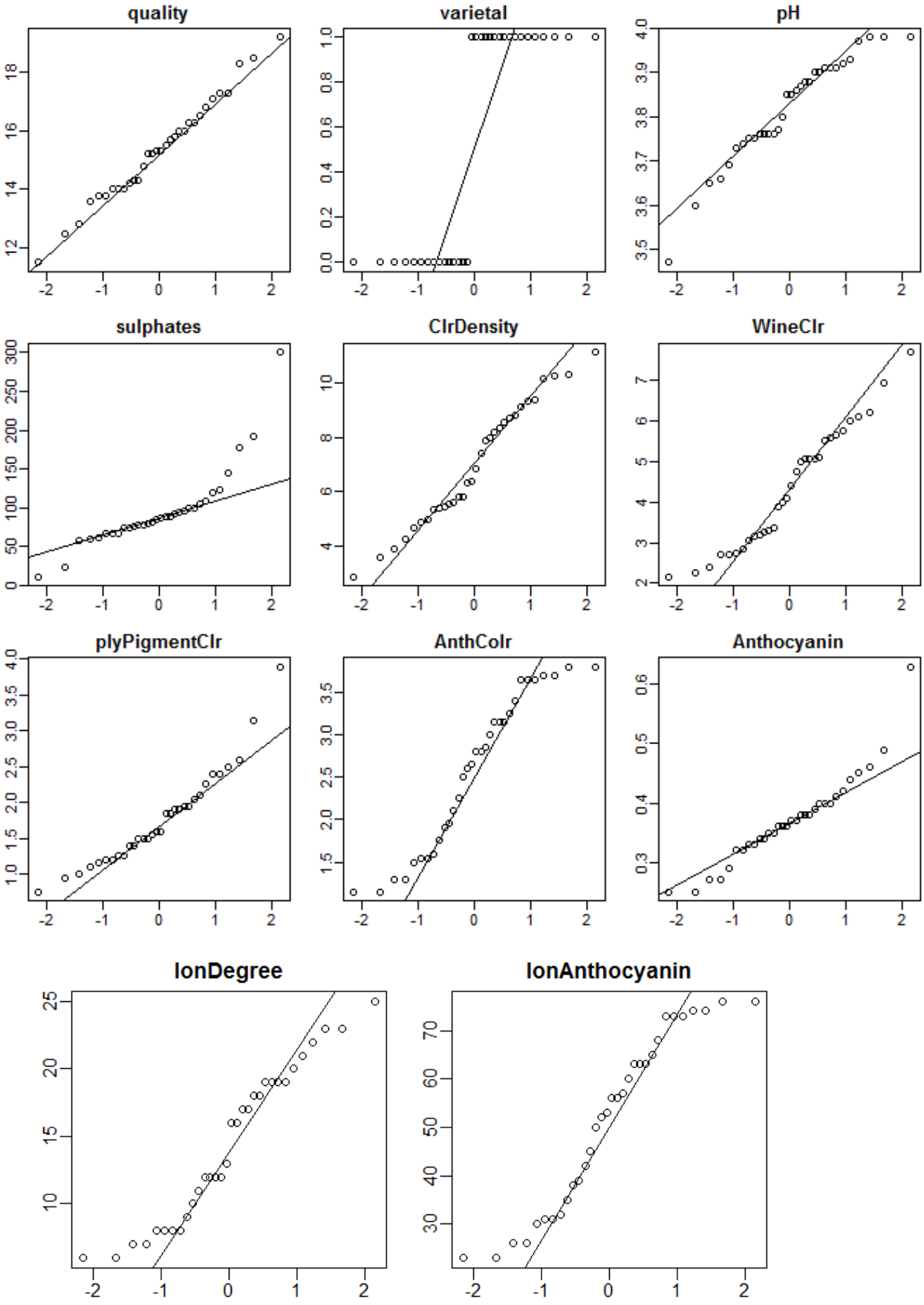

Anthocyanin distribution
count
Anthocyanin


Degree of ionized antho distribution
count
IonDegree


Ionized Anthocyanin distribution
count
IonAnthocyanin


Wine Quality
count
quality

- QQ plot


quality
varietal
pH
sulphates
ClrDensity
WineClr
plyPigmentClr
AnthColr
Anthocyanin
IonDegree
IonAnthocyanin

According to the QQ- plot and histogram, only the two variables (wine quality and color density) satisfy the normal distribution. Since the regression only require the y follow normal, we can think this dataset can used for regression analysis.

2. Full model regression

Modeling is to find the relationship between x1-x10 with y. One dummy variable only affect the intercept of model, it cannot affect the number of variables selected into model. Therefore, when we do the model select or variable selection, we treat the x1 as numeric variables to process whole dataset, instead of half-and -half splitting dataset. Once we find the best variables in the model, then we conduct the analysis of indicator variable, if there still have that dummy variables. Before we do this project, we test our hypothesis that the treating the dummy variable as numeric variable does not affect the model selection.

In software R, the function factor() is to define the indicator variable. We use the **factor(varietal)** in the full model analysis. In the output of the full model regression analysis, we find **factor(varietal) 1,** but we cannot find **factor(varietal) 0** , because the varietal has two level 0 and 1. Actually one dummy variable only affect the intercept, the **factor(varietal) 0** become a part of Intercept. Therefore, the dummy variable does not affect the selection of others variables.

This is the result of R for full model regression analysis:

```
Call:
lm(formula = quality ~ factor(varietal) + pH + sulphates + ClrDensity +
    WineClr + plyPigmentClr + AnthColr + Anthocyanin + IonDegree +
    IonAnthocyanin, data = wine)

Residuals:
     Min       1Q   Median       3Q      Max
-2.0242 -0.6722  0.1939  0.6135  1.8333

Coefficients: (2 not defined because of singularities)
                    Estimate Std. Error t value Pr(>|t|)
(Intercept)      -12.933532  13.038623  -0.992   0.3315
factor(varietal) 1  -1.359320   0.498872  -2.725   0.0121 *
pH                   6.988262   3.001919   2.328   0.0291 *
sulphates            0.016012   0.007709   2.077   0.0492 *
ClrDensity           0.150580   0.150681   0.999   0.3280
WineClr              1.472760   1.128563   1.305   0.2048
plyPigmentClr       -1.762277   1.309300  -1.346   0.1914
AnthColr                   NA         NA      NA       NA
Anthocyanin        -10.168000   7.358882  -1.382   0.1803
IonDegree            0.017610   0.206809   0.085   0.9329
IonAnthocyanin             NA         NA      NA       NA
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 1.092 on 23 degrees of freedom
```

Multiple R-squared: 0.718, Adjusted R-squared: 0.6198

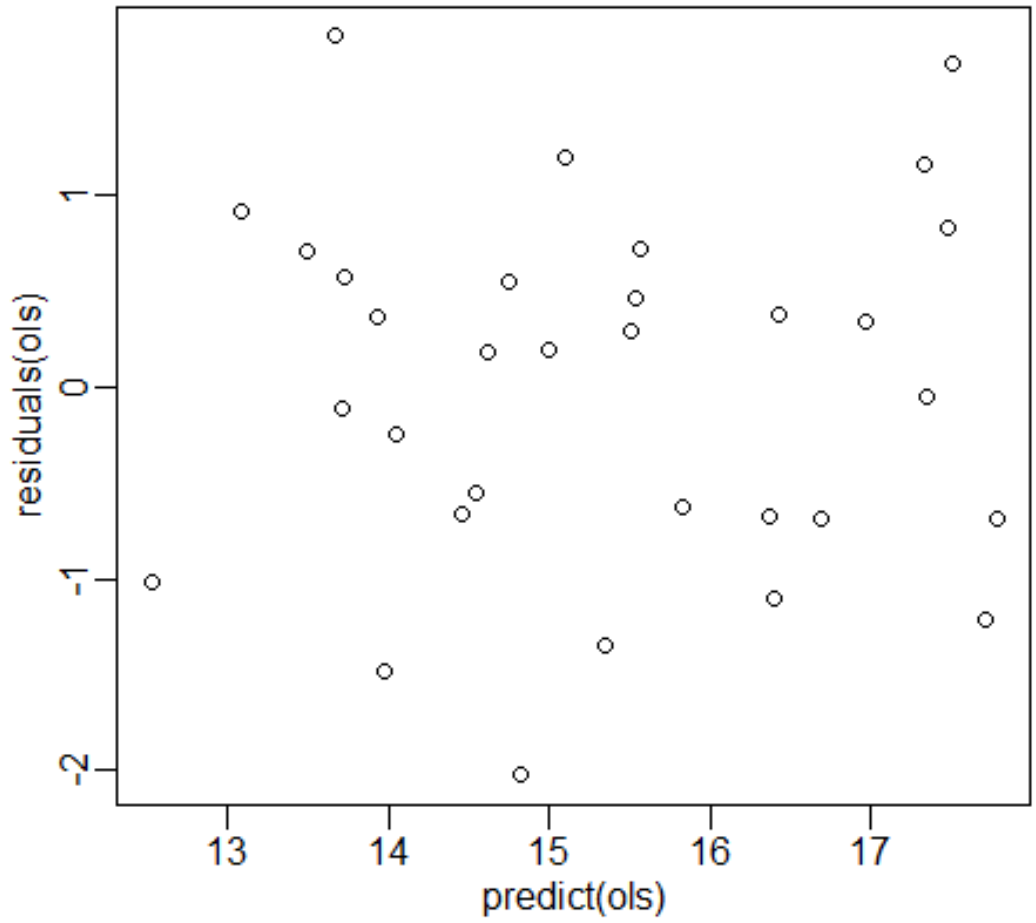
F-statistic: 7.318 on 8 and 23 DF, p-value: 7.922e-05

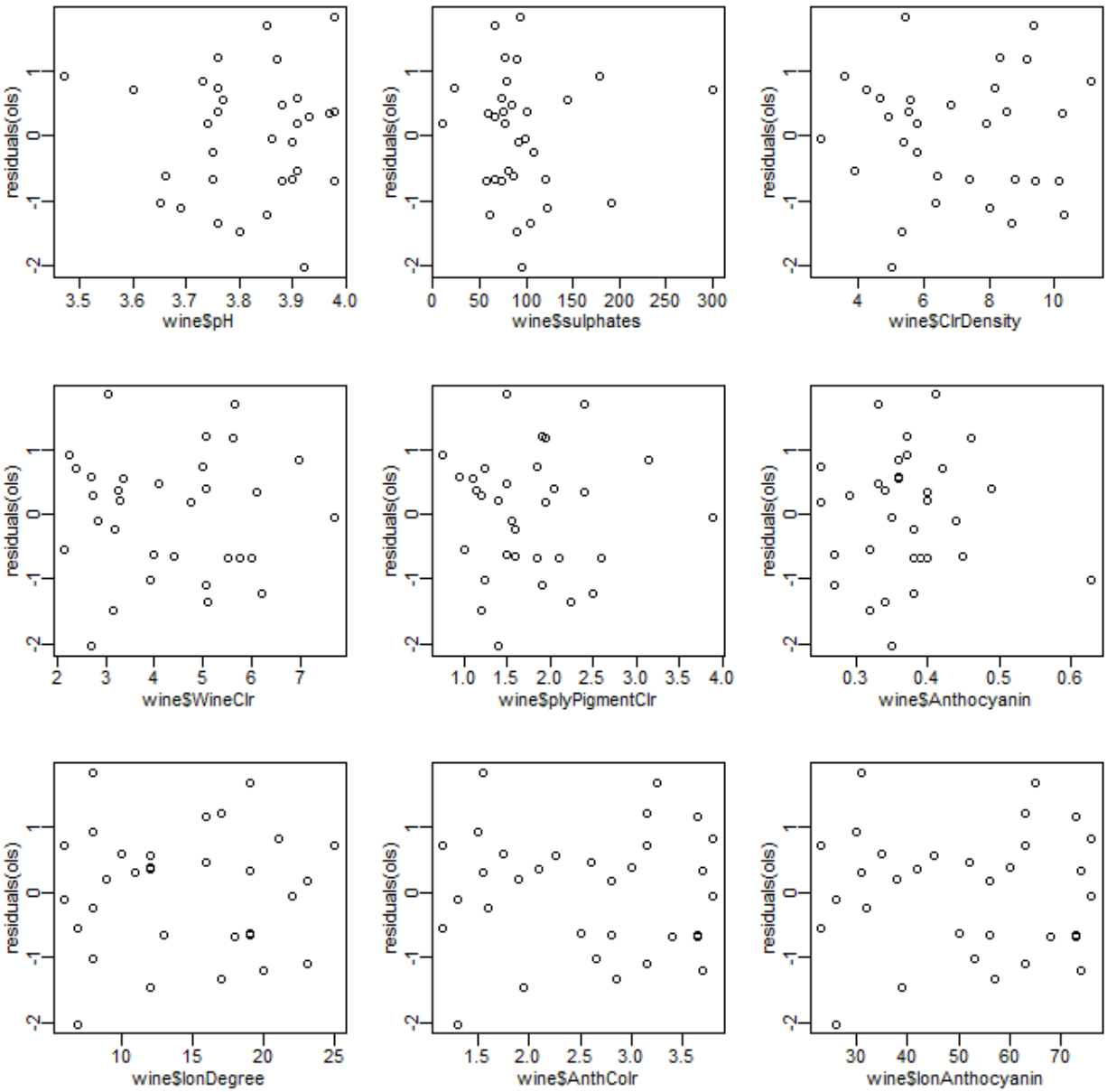
Only the varietal, pH and sulphates are significant, so it seems that this model can be simplified.

A two-level categorical variable (like gender) in SAS or R becomes a simple 0-1 recode and then treated as continuous. A three-level categorical variable has to become two variables. Therefore, we can treat one two level dummy variable as numeric variable in SAS or R.

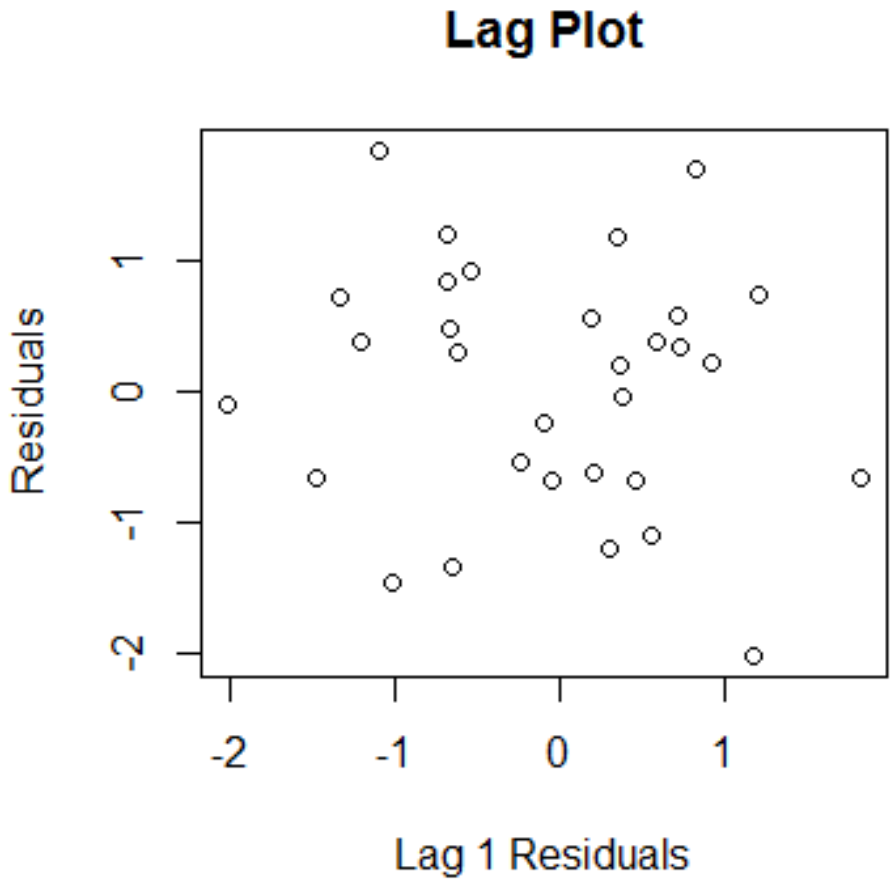
- Residual Analysis

We plot the residuals for each variables and predicted value, there are no obvious pattern. It is satisfy with the constant variance assumption, when we conduct the statistical regression analysis. Therefore, this data does not need to do transformation. Also the residual values randomly scatter in the Residual verse Predicted plot, therefore, we can believe that the experiment is well designed and the data follow the independence assumption.





- **Independent assumption check.**



Here the lagged residual plot is used for independent check of dataset. Any obvious pattern will indicate the dependence of the dataset. Checking this graph, we do not find any pattern, the residual scatter randomly, so this wine quality dataset satisfy the independent assumption.

3. Model Diagnostic

- Multicollinearity
  Here we conduct two methods, one is Eigensystem Analysis, and the other is the Variance Decomposition.
  - Eigenvector

x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
0.005	0.090	-0.231	0.350	0.421	0.369	0.418	-0.052	0.388	0.418
0.297	-0.571	0.614	0.066	0.102	0.004	0.166	0.373	0.014	0.166
-0.360	0.419	0.086	0.082	0.095	0.148	0.047	0.716	-0.364	0.047
0.801	0.383	-0.171	0.233	-0.107	-0.264	0.020	0.208	-0.059	0.020
0.319	0.306	0.240	-0.651	0.207	0.517	-0.044	-0.090	0.022	-0.044
0.018	0.211	0.466	0.612	-0.005	0.323	-0.250	-0.341	-0.132	-0.250
-0.180	0.434	0.497	-0.116	-0.045	-0.503	0.302	-0.245	0.147	0.302
-0.083	0.128	0.104	0.023	-0.164	0.000	-0.271	0.343	0.821	-0.271
0.000	0.000	0.000	0.000	-0.849	0.384	0.257	0.000	0.000	0.257
0.000	0.000	0.000	0.000	0.000	0.000	0.707	0.000	0.000	-0.707

Eigenvalue	5.308	1.527	1.352	0.995	0.477	0.208	0.114	0.019	0.000	0.000

λ9, λ10 equal to 0 indicate the

x10*(-0.707)=0

x5*(-0.849)+x6*(0.384)+x7*(0.257)+x10(0.257)=0

thus, x10=0; x5=x6*(0.452)+x7*(0.303)

λ8 also is small, thus

x7*(-0.251)+ x8*(0.343) + x9*(0.821)=0

Therefore, we can conclude that x5-x10 are involved in the multicollinear relationship.

- Variance decomposition

No.	**Eigen value**	**Condition Index**	**x1**	**x2**	**x3**	**x4**	**x5**	**x6**	**x7**	**x8**	**x9**	**x10**
1	5.308	1.000	0.000	0.000	0.002	0.008	0.000	0.000	0.000	0.000	0.001	0.000
2	1.527	1.865	0.035	0.062	0.059	0.001	0.000	0.000	0.000	0.011	0.000	0.000
3	1.352	1.981	0.058	0.038	0.001	0.002	0.000	0.000	0.000	0.048	0.003	0.000
4	0.995	2.310	0.387	0.043	0.007	0.019	0.000	0.000	0.000	0.005	0.000	0.000
5	0.477	3.334	0.128	0.057	0.029	0.303	0.000	0.000	0.000	0.002	0.000	0.000
6	0.208	5.057	0.001	0.063	0.249	0.617	0.000	0.000	0.000	0.070	0.002	0.000
7	0.114	6.815	0.170	0.479	0.514	0.041	0.000	0.000	0.000	0.066	0.005	0.000
8	0.019	16.901	0.221	0.258	0.139	0.010	0.000	0.000	0.000	0.797	0.989	0.000
9	0.000	2303977	0.000	0.000	0.000	0.000	1.000	1.000	0.117	0.000	0.000	0.117
10	0.000	2303977	0.000	0.000	0.000	0.000	0.000	0.000	0.883	0.000	0.000	0.883

η9, η10 are larger than 30, so there are 2 dependence in the columns of wine quality

Furthermore, π(10,7), π(10,10) are all exceed than 0.5, indicating x7 and x10 are involved in a multicollinear relationship. π(9,5) and π(9,6) are all exceed than 0.5, indicating intercept, x5 and x6 are involved in a multicollinear relationship. Also we notice that the π(8,8) and π(8,9) are also very large. Therefore, the x5-x7, and x8-x10 are involved in the multicollinear relationship. The result is same as the Eigensystem Analysis.

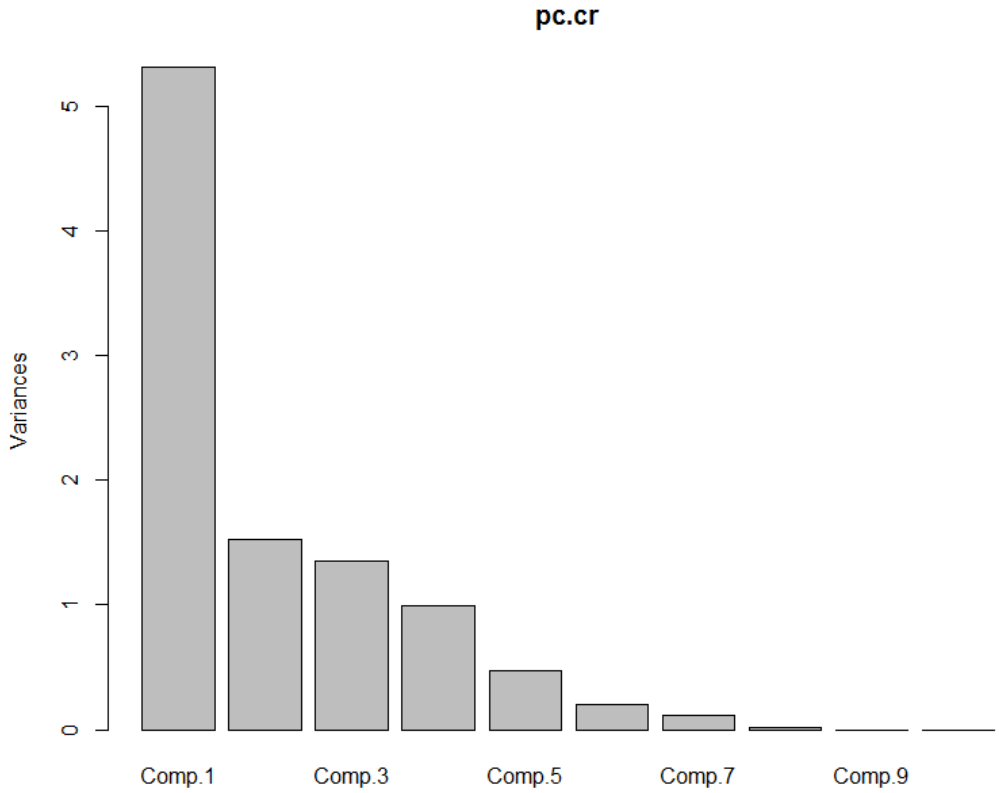
- Component analysis

```
proc princomp data=A out=pc_A std;
var x1 x2 x3 x4 x5 x6 x7 x8 x9 x10;
ods select eigenvectors eigenvalues;
ods trace on;
ods show;
run;
```

Based the above SAS code, we get the 10 components, and use them as new variables to conduct the regression analysis. From the tables and the Scree plot, the first four components contribute 92% to the model.

SAS output of principle component regression

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Variance Inflation
Intercept	**1**	15.35000	0.19300	79.53	<.0001	0
Prin1	**1**	1.28134	0.19609	6.53	<.0001	1.00000
Prin2	**1**	-0.27032	0.19609	-1.38	0.1813	1.00000
Prin3	**1**	0.17541	0.19609	0.89	0.3803	1.00000
Prin4	**1**	-0.37837	0.19609	-1.93	0.0661	1.00000
Prin5	**1**	-0.04549	0.19609	-0.23	0.8186	1.00000
Prin6	**1**	0.28060	0.19609	1.43	0.1659	1.00000
Prin7	**1**	0.52921	0.19609	2.70	0.0128	1.00000
Prin8	**1**	-0.03765	0.19609	-0.19	0.8494	1.00000
Prin9	**0**	0	.	.	.	.
Prin10	**0**	0	.	.	.	.



Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	5.30831121	3.78176029	0.5308	0.5308
2	1.52655092	0.17445055	0.1527	0.6835
3	1.35210037	0.35698435	0.1352	0.8187
4	0.99511602	0.51766939	0.0995	0.9182
5	0.47744663	0.26985335	0.0477	0.9660
6	0.20759329	0.09329515	0.0208	0.9867
7	0.11429814	0.09571471	0.0114	0.9981
8	0.01858343	0.01858343	0.0019	1.0000
9	0.00000000	0.00000000	0.0000	1.0000
10	0.00000000		0.0000	1.0000

The new model can be written as

Y= 15.35 + 1.28 Prin1-0.27Prin2 +0.175x3-0.378 Prin4
All the principle component can be computed from the eigenvectors.
Prin1=0.005x1+0.090x2-0.231x3+0.350x4+0.421x5+0.369x6+0.418x7-0.052x8+0.388x9+0.418x10. The Prin2, Prin3 and Prin4 also can obtained from the eigenvector in same way.

- ## Selecting a Regression Model
  - ### Forward Selection

SAS default has the 0.5 for significance level for entry into the model.

The table below shows which variable entry the model in every forward step, and computed the corresponding F value and p value.

	Step 1		**Step 2**		**Step 3**		**Step 4**		**Step 5**		**Step 6**		**Step 7**	
	F Value	Pr	F Value	Pr	F Value	Pr	F Value	Pr	F Value	Pr	F Value	Pr	F Value	Pr
Intercept	272.23	.0001	288.3	.0001	117.7	.0001	0.5	0.485	0.21	0.650	1.62	0.215	1.67	0.209
x1			4.26	0.0481	4.59	0.041	4.05	0.054	6.1	0.020	9.63	0.005	9.88	0.004
x2							1.87	0.183	2.81	0.106	7.69	0.010	7.58	0.011
x3											4.51	0.044	5.17	0.032
x4													1.04	0.319

x5	30.09	.0001	32.32	.000	33.66	.000	31.61	.0001	0	0.984	0.47	0.499	0.24	0.625
x7									2.28	0.143	5.71	0.025	3.18	0.087
x8					2.17	0.152	2.62	0.117	3.52	0.072	8.46	0.008	9.06	0.006

After 7 steps, we obtained the model with all possible variables at the level of 0.5 significant. The results shows that x5 entered at the first one, and x1, x8,x2, x7, x3, x4 are followed.

Summary of Forward Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	x5	1	0.5008	0.5008	12.7099	30.09	<.0001
2	x1	2	0.0639	0.5647	9.4971	4.26	0.0481
3	x8	3	0.0313	0.5960	8.9437	2.17	0.1519
4	x2	4	0.0262	0.6222	8.8070	1.87	0.1825
5	x7	5	0.0304	0.6526	8.3258	2.28	0.1433
6	x3	6	0.0530	0.7057	6.0004	4.51	0.0439
7	x4	7	0.0122	0.7179	7.0073	1.04	0.3189

The significant level 0.5 seems too high, and this model is too conservative. There are seven variable, plus intercept, this is almost same as the full model. If we change the entry level from 0.5 to 0.15, then only x5 and x1 entry the model. The SAS output is shown as below

Summary of Forward Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	x5	1	0.5008	0.5008	12.7099	30.09	<.0001
2	x1	2	0.0639	0.5647	9.4971	4.26	0.0481

The corresponding the regression result is

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	12.19512	0.71824	420.62297	288.29	<.0001
x1	-0.88379	0.42826	6.21348	4.26	0.0481
x5	0.83469	0.14682	47.15581	32.32	<.0001

This model is much simple, however, the Cp value is much larger than Number Variables in

model, which indicates that maybe too simple. In order to further exploit the possible reduced model, we conducted the stepwise method.

- Stepwise

Entry level is set as 0.5, and stay level is set as 0.25;

Summary of Stepwise Selection								
Step	**Variable Entered**	**Variable Removed**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	x5		1	0.5008	0.5008	12.7099	30.09	<.0001
2	x1		2	0.0639	0.5647	9.4971	4.26	0.0481
3	x8		3	0.0313	0.5960	8.9437	2.17	0.1519
4	x2		4	0.0262	0.6222	8.8070	1.87	0.1825
5	x7		5	0.0304	0.6526	8.3258	2.28	0.1433
6		x5	4	0.0000	0.6526	6.3262	0.00	0.9839
7	x3		5	0.0475	0.7001	4.4519	4.12	0.0527
8	x4		6	0.0148	0.7150	5.2421	1.30	0.2648
9		x4	5	0.0148	0.7001	4.4519	1.30	0.2648

After nine steps, SAS output a reasonable model at the level of 0.25, because only p-value<0.25 can stay inside the model.

Variable	**Parameter Estimate**	**Standard Error**	**Type II SS**	**F Value**	**Pr > F**
Intercept	-9.68494	8.72303	1.38185	1.23	0.2770
x1	-1.19946	0.38319	10.98373	9.80	0.0043
x2	6.22916	2.26463	8.48134	7.57	0.0107
x3	0.01291	0.00636	4.61802	4.12	0.0527
x7	1.62569	0.24847	47.98842	42.81	<.0001
x8	-9.47303	3.29706	9.25391	8.26	0.0080

Comparing with forward model, the model x1x2x3x7x8 selected by stepwise seems a little better. Since there are only 10 variables, the all-possible regression is possible to analysis. Here, we

will conduct the all – possible model, and select 2~3 best models to compare with models from stepwise and forward method.

- Variable selection(All possible Regressions)

11 variables has too many combinations, fortunately, SAS can compute all the possible models and corresponding AIC, BIC, R square, adjusted R square, Cp and MSres.

```
proc reg data=A outest=est0;
model y = x1-x10 / selection=adjrsq sse aic bic adjrsq cp rmse;    run;
```

here we only listed best 8 models for each criterion

1. Best model sorted by Cp, in these tables the character 'I' means the intercept.

No in Model	p	SSE	RSQ	ADJRSQ	Cp	AIC	BIC	Regressors
5	6	29.1457	0.70015	0.64248	4.4519	9.0103	14.524	x1x2x3x7x8
5	6	29.1457	0.70015	0.64248	4.4519	9.0103	14.524	x1x2x3x8x10
6	7	27.7036	0.71498	0.64658	5.24205	9.3865	16.3779	x1x2x3x4x8x10
6	7	27.7036	0.71498	0.64658	5.24205	9.3865	16.3779	x1x2x3x4x7x8
4	5	32.6022	0.66459	0.6149	5.35175	10.5966	14.2383	x1x2x3x9
5	6	30.2692	0.68859	0.6287	5.39442	10.2206	15.2067	x1x2x3x4x9x10
6	7	28.6075	0.70568	0.63505	6.00035	10.4138	16.858	x1x2x3x5x8x10
6	7	28.6075	0.70568	0.63505	6.00035	10.4138	16.858	x1x2x3x5x7x8

2. Best model sorted by adjusted AIC

No in Model	p	SSE	RSQ	ADJRSQ	Cp	AIC	BIC	Regressors
5	6	29.1457	0.70015	0.64248	4.4519	9.0103	14.524	12378
5	6	29.1457	0.70015	0.64248	4.4519	9.0103	14.524	123810
6	7	27.7036	0.71498	0.64658	5.24205	9.3865	16.3779	1234810
6	7	27.7036	0.71498	0.64658	5.24205	9.3865	16.3779	123478
5	6	30.2692	0.68859	0.6287	5.39442	10.2206	15.2067	12349
6	7	28.6075	0.70568	0.63505	6.00035	10.4138	16.858	1235810
6	7	28.6075	0.70568	0.63505	6.00035	10.4138	16.858	123578
6	7	28.6075	0.70568	0.63505	6.00035	10.4138	16.858	1236810

3. Best model sorted by BIC'

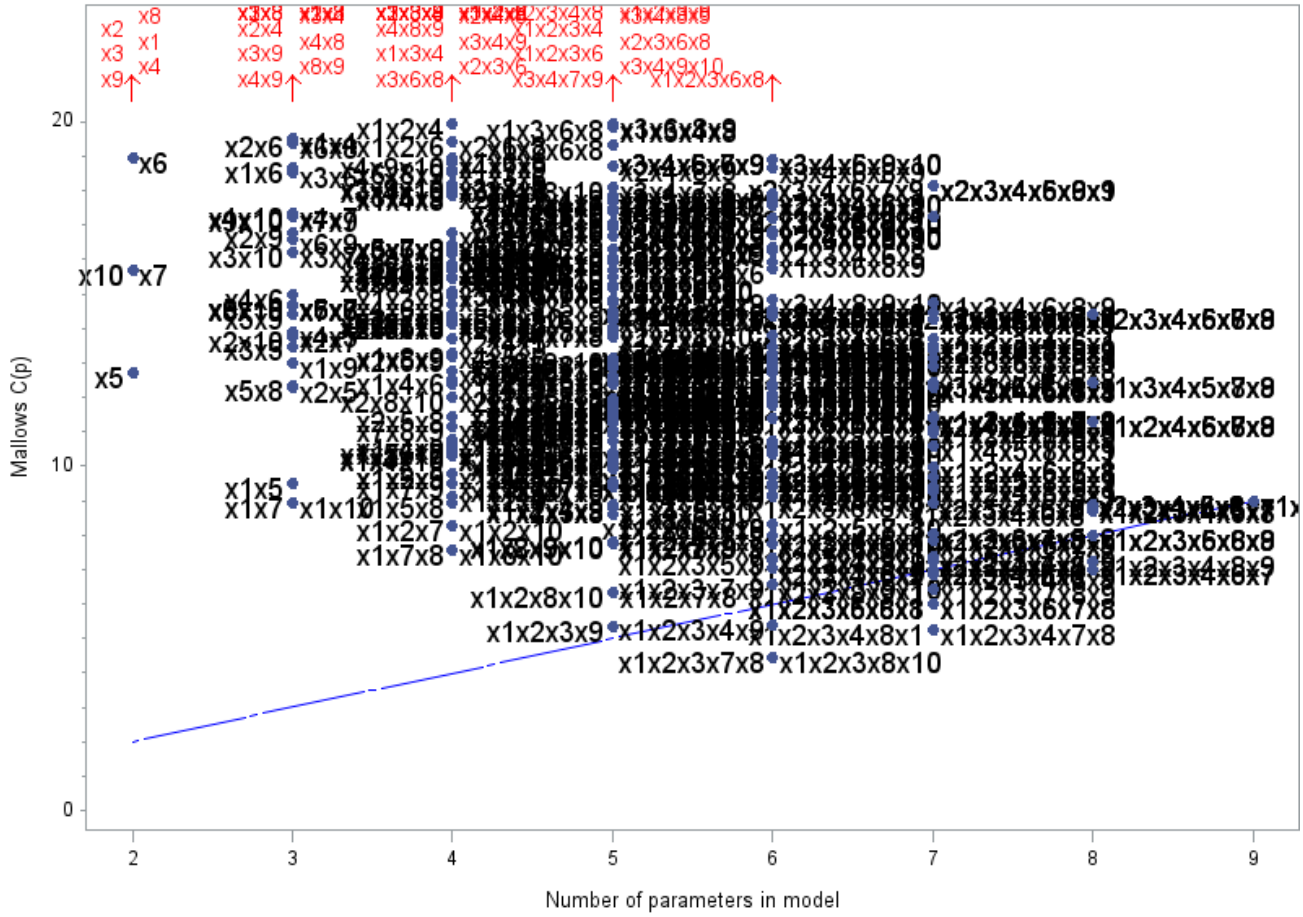
No in Model	p	SSE	RSQ	ADJRSQ	Cp	AIC	BIC	Regressors
4	5	32.6022	0.66459	0.6149	5.35175	10.5966	14.2383	1239
5	6	29.1457	0.70015	0.64248	4.4519	9.0103	14.524	12378
5	6	29.1457	0.70015	0.64248	4.4519	9.0103	14.524	123810

4	5	33.7637	0.65264	0.60118	6.3262	11.7168	14.9802	1278
4	5	33.7637	0.65264	0.60118	6.3262	11.7168	14.9802	12810
5	6	30.2692	0.68859	0.6287	5.39442	10.2206	15.2067	12349
3	4	37.64	0.61276	0.57127	7.57818	13.1946	15.3011	178
3	4	37.64	0.61276	0.57127	7.57818	13.1946	15.3011	1810

4. Best model sorted by RMSE

No in Model	p	SSE	RSQ	ADJRSQ	Cp	AIC	BIC	Regressors
6	7	27.7036	0.71498	0.64658	5.24205	9.3865	16.3779	1234810
6	7	27.7036	0.71498	0.64658	5.24205	9.3865	16.3779	123478
5	6	29.1457	0.70015	0.64248	4.4519	9.0103	14.524	12378
5	6	29.1457	0.70015	0.64248	4.4519	9.0103	14.524	123810
7	8	27.4237	0.71786	0.63557	7.00725	11.0616	19.0099	12345810
7	8	27.4237	0.71786	0.63557	7.00725	11.0616	19.0099	1234578
7	8	27.4237	0.71786	0.63557	7.00725	11.0616	19.0099	12346810
7	8	27.4237	0.71786	0.63557	7.00725	11.0616	19.0099	1234678

Mallows's Cp plot



The blue dots represent models, and black texts indicate the model components. There are too many

models, so the name of model are overlapped each other. But we still can find the best Cp model, that is x1x2x3x9, x1x2x3x4x9, x1x2x3x7x8.

In the table of AIC and BIC, model x1x2x3x7x8 and model x1x2x3x9 are among the best model, and Cp prefers x1x2x3x9 and x1x2x3x4x9. Model x1x2x3x7x8 also appeared in the table of RMSE. Therefore, the best model we choose is x1x2x3x7x8. For comparison, we also selected model x1x2x3x9.
In the processing of model selection of stepwise and forward, we got two models **x1x2x3x7x8 and x1x5.** Therefore, we have total 3 models for comparison.

- Model Comparison of **x1x2x3x7x8, x1x2x3x9 and x1x5**

I. Model x1x2x3x9

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Variance Inflation
Intercept	**1**	**-18.97399**	**9.48401**	**-2.00**	**0.0556**	**7539.92000**	**4.83300**	**0**
x1	**1**	**-1.42225**	**0.40477**	**-3.51**	**0.0016**	**7.73302**	**14.90792**	**1.08124**
x2	**1**	**7.74975**	**2.28593**	**3.39**	**0.0022**	**5.89785**	**13.87819**	**1.96957**
x3	**1**	**0.01489**	**0.00623**	**2.39**	**0.0242**	**5.94709**	**6.89027**	**2.71212**
x9	**1**	**0.28363**	**0.04645**	**6.11**	**<.0001**	**45.01982**	**45.01982**	**1.82566**

Rsq=0.6646, and AdjRsq=0.6149, RMSE=1.2075 PRESS= 46.96838

II. Model x1x2x3x7x8

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Variance Inflation
Intercept	**1**	-9.68494	8.72303	-1.11	0.2770	7539.92000	1.38185	0
x1	**1**	-1.19946	0.38319	-3.13	0.0043	7.73302	10.98373	1.04378
x2	**1**	6.22916	2.26463	2.75	0.0107	5.89785	8.48134	2.08220
x3	**1**	0.01291	0.00636	2.03	0.0527	5.94709	4.61802	3.04295
x7	**1**	1.62569	0.24847	6.54	<.0001	39.22242	47.98842	1.36992
x8	**1**	-9.47303	3.29706	-2.87	0.0080	9.25391	9.25391	1.69885

Rsq=0.7001, and AdjRsq=0.6425, RMSE=1.0588 PRESS=51.25296

III. Model selected by Forward: x1x5

Parameter Estimates

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	Type I SS	Type II SS	Variance Inflation
Intercept	**1**	12.19512	0.71824	16.98	<.0001	7539.92000	420.62297	0
x1	**1**	-0.88379	0.42826	-2.06	0.0481	7.73302	6.21348	1.00173
x5	**1**	0.83469	0.14682	5.69	<.0001	47.15581	47.15581	1.00173

Rsq=0.5647, and AdjRsq=0.5347, RMSE=1.208 **PRESS**=50.95

Comparing the above three models, the models have similar VIF values, PRESS, and p-Value. By comparing R square, RMSE, first two models are similar, and better than Model X1X5. Between the first two models, the model x1x2x3x9 has less parameter. Therefore, the model x1x2x3x9 maybe the best model we can chose. Before drawing a conclusion, we also conduct anova to compare these models.

Model	RSS	DF	Sum of Sq	F	Pr(>F)
Full	27.415	23			
x1x2x3x7x8	29.146	26	-1.731	0.484	0.697
x1x2x3x9	32.602	27	-5.187	1.088	0.386
x1x5	42.311	29	-14.896	2.083	0.095

From this table, at the 20% level, x1x2x3x9 works same as the full model. This one also can verified in Deviance table. Also we notice that AIC and BIC in full model are in the worst case, maybe due to collinearnarity.

Model	AIC	BIC	Deviance	Difference in Deviance compared to Full model	P-Value
Full	105.864	120.521	157.344		
x1x2x3x7x8	101.822	112.083	158.927	1.583	0.663
x1x2x3x9	103.409	112.203	163.033	5.689	0.224
x1x5	107.750	113.613	170.362	7.328	0.026

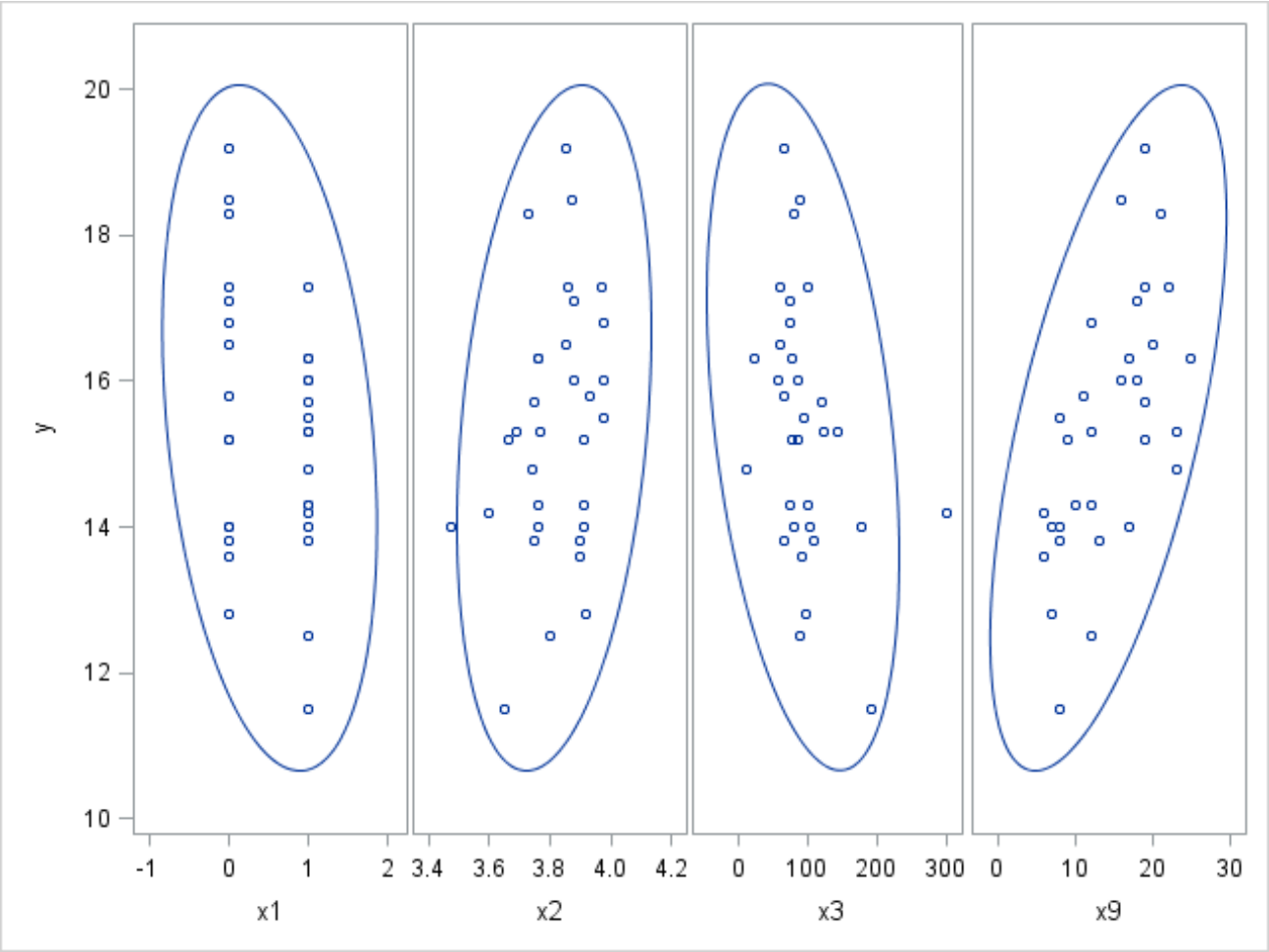
Now we can propose a new reduced model:

$$\underline{y = -18.97399 - 1.42225\,x1 + 7.74975\,x2 + 0.01489\,x3 + 0.28363\,x9}$$

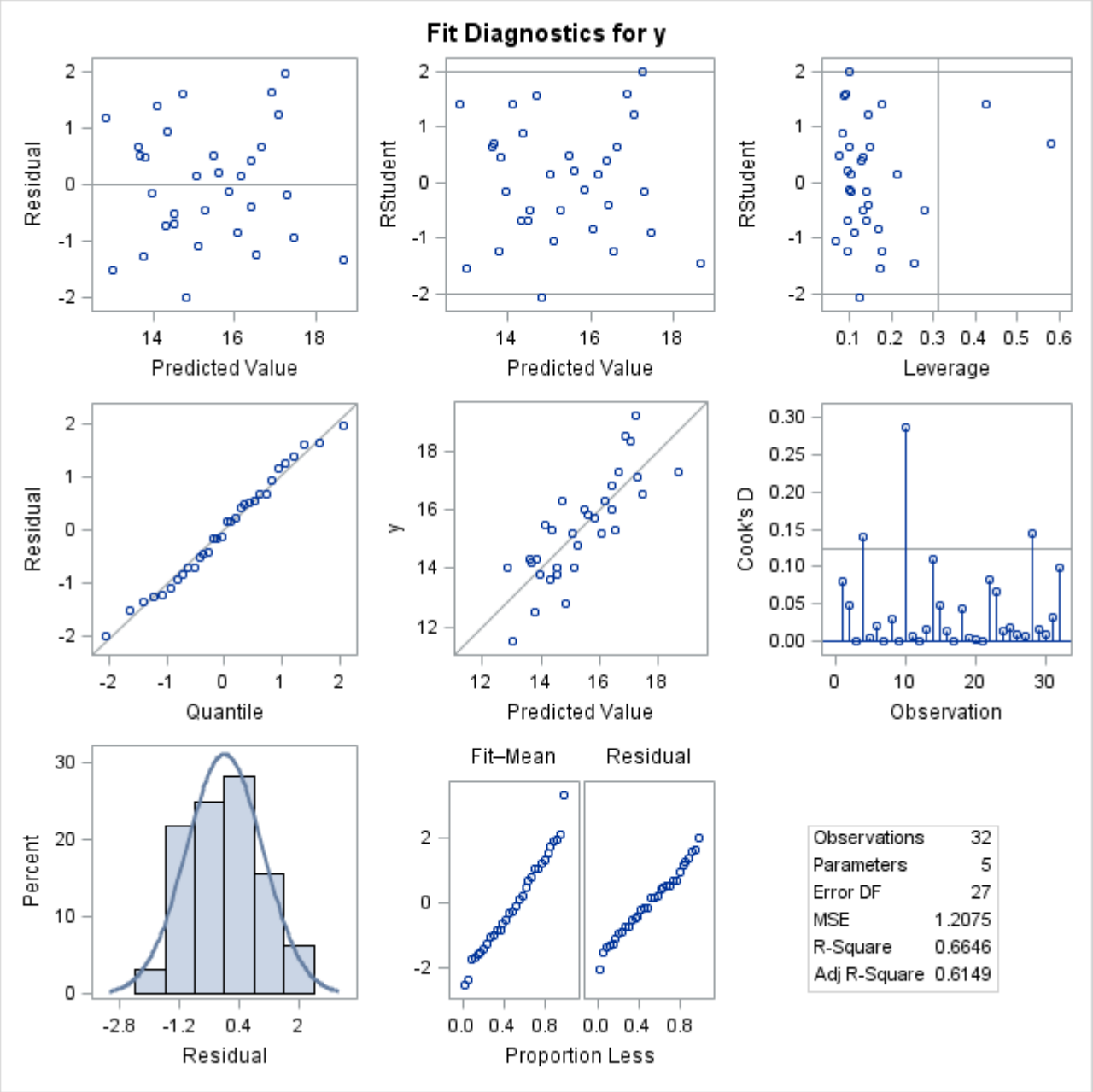
5. Adequacy of the new models

1. Scatter plot with 95% confidential interval

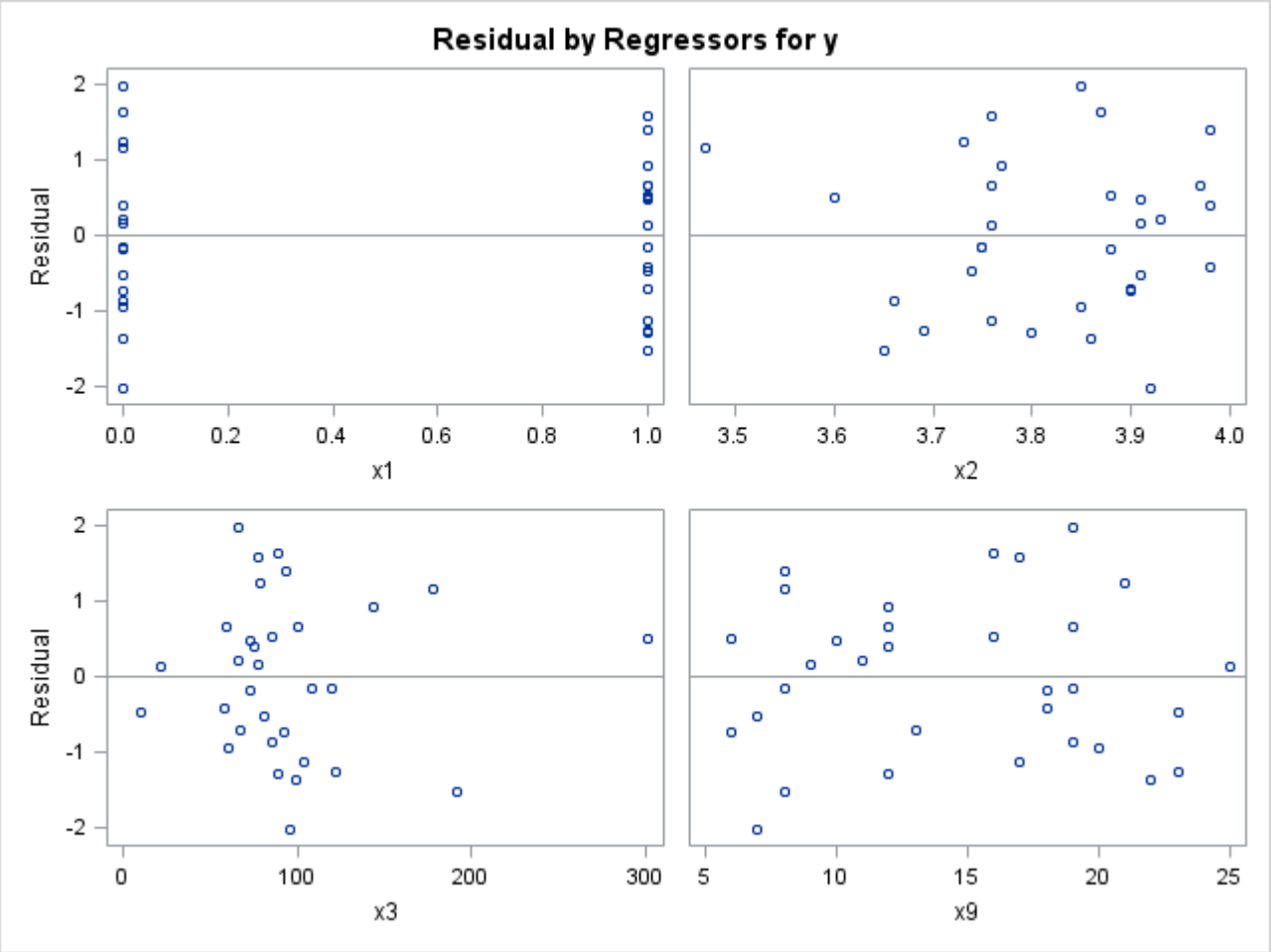
   Scatter plot can be easy to identify the relationship between dependant and independent variables, 95% confidential interval can reveal the possible outliers. From the scatter plot, it can tell us that x1, x2, x3 and x9 have certain relationships, and one or two observations are suspicious.



2. Residual plot

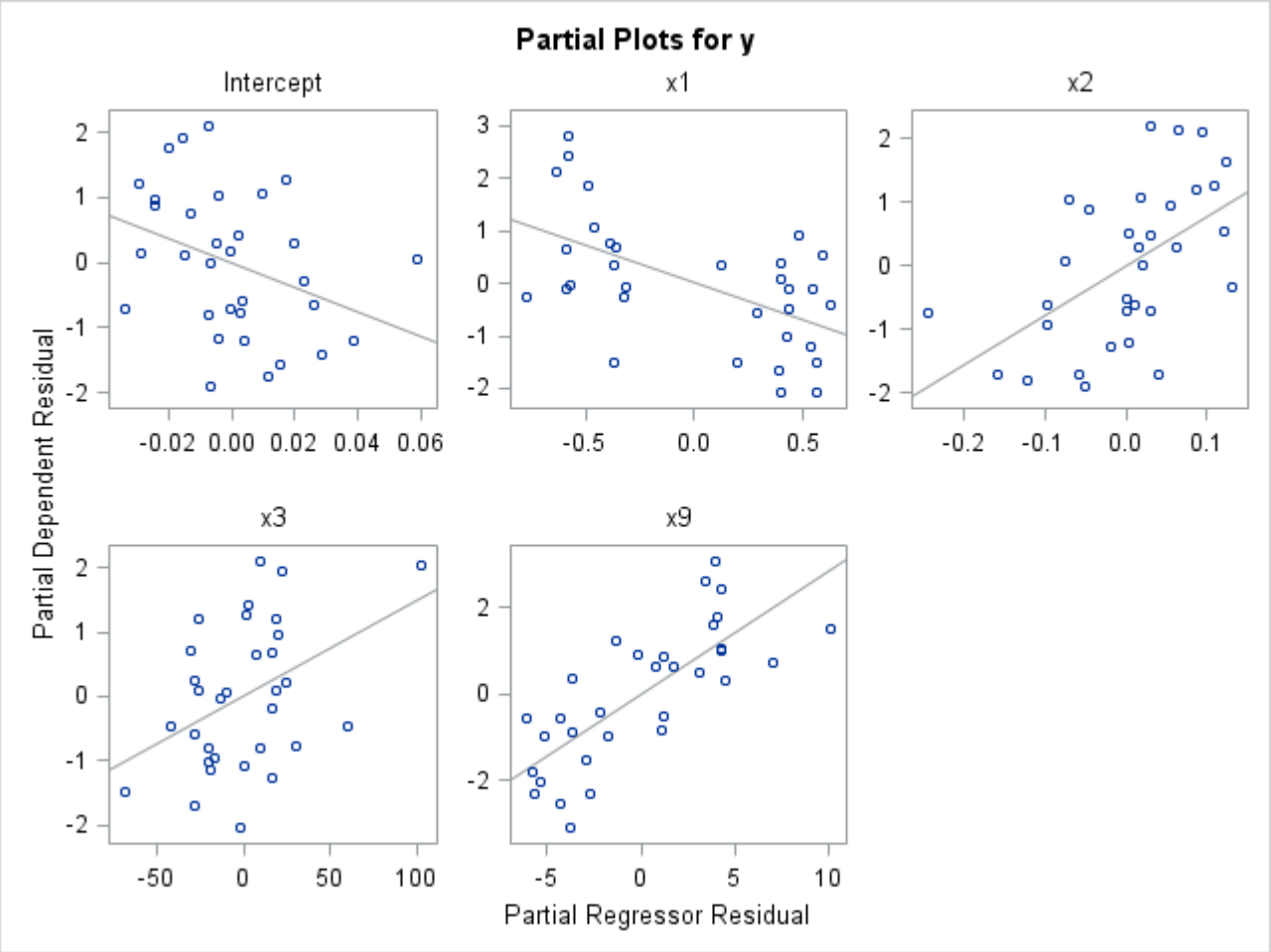


From the plot of predicted ~ residuals, there are not obvious pattern, so we can think the variance is constant. The QQplot and histogram of residuals show that it satisfied with normality assumption. The R-Student plot shows the same as residual plot, but it shows the 3 possible outliers, which also be shown in Cook'D and leverage plot. We will conduct outliers analysis. In order to check residuals, we plot another four residual plots, x1~residuals, x2~reisduals, x3~reisduals, and x9~reisduals. There are no scatter patterns in all the plots, and the results further verified our new model.



3. Partial residual plots

Plotting residuals versus a repressor is not always the most effective way to reveal whether a curvature effect is required for that variable in the model. The below figure presents the partial regression plots for wine quality,. The linear relationship between y and (x1,x2 x3, x9) is clearly evident in these plots.



The linear relationship between the four variables (x1 x2 x3 and x7) with y is clearly evident in all of those plots.

4. Outlier or influential analysis

Here we computed Residual, Studentized Residual, R-student residual, PRESS, Cook's Distance, Hat matrix, Covriance Ratio, DFFITS, DFBETAS.

Obs	Residual	R Student	Studentized R	Hat Diag H	Cook D	Cov Ratio	DFFITS	PRESS
1	1.966	1.886	1.987	0.100	0.079	0.662	0.664	8.705
2	1.235	1.214	1.226	0.143	0.049	1.064	0.501	5.470
3	-0.187	-0.180	-0.177	0.104	0.001	1.340	-0.060	0.466
4	-1.354	-1.428	-1.457	0.256	0.140	1.096	-0.854	0.098
5	0.410	0.399	0.393	0.127	0.005	1.343	0.150	1.400
6	-0.943	-0.911	-0.908	0.112	0.021	1.164	-0.323	0.013
7	0.215	0.206	0.202	0.097	0.001	1.327	0.066	0.687

8	-0.859	-0.859	-0.855	0.171	0.030	1.268	-0.388	0.003
9	0.159	0.153	0.150	0.105	0.001	1.344	0.051	0.545
10	1.163	1.395	1.421	0.424	0.286	1.442	1.218	9.312
11	-0.519	-0.507	-0.500	0.134	0.008	1.329	-0.197	0.006
12	-0.164	-0.161	-0.158	0.140	0.001	1.397	-0.064	0.169
13	-0.721	-0.708	-0.702	0.141	0.016	1.280	-0.284	0.093
14	-2.020	-1.965	-2.083	0.125	0.111	0.638	-0.788	2.803
15	1.619	1.547	1.591	0.093	0.049	0.837	0.509	6.560
16	0.662	0.654	0.647	0.149	0.015	1.311	0.271	0.012
17	0.139	0.143	0.140	0.214	0.001	1.530	0.073	0.343
18	1.589	1.514	1.554	0.088	0.044	0.850	0.483	0.934
19	-0.417	-0.410	-0.404	0.146	0.006	1.371	-0.167	1.354
20	0.524	0.495	0.488	0.075	0.004	1.247	0.139	0.009
21	-0.141	-0.135	-0.132	0.098	0.000	1.335	-0.044	0.362
22	1.384	1.387	1.413	0.176	0.082	1.013	0.653	0.397
23	-1.240	-1.243	-1.257	0.176	0.066	1.092	-0.582	3.339
24	0.932	0.886	0.882	0.083	0.014	1.136	0.266	0.155
25	-0.460	-0.493	-0.486	0.279	0.019	1.600	-0.302	2.573
26	0.665	0.637	0.630	0.098	0.009	1.242	0.208	0.012
27	0.472	0.461	0.454	0.132	0.006	1.337	0.177	0.207
28	0.514	0.722	0.716	0.580	0.144	2.610	0.842	0.649
29	-1.113	-1.048	-1.050	0.066	0.016	1.051	-0.279	3.052
30	-0.712	-0.682	-0.675	0.097	0.010	1.226	-0.221	2.569
31	-1.281	-1.226	-1.238	0.095	0.032	1.003	-0.402	4.818
32	-1.518	-1.521	-1.560	0.175	0.098	0.936	-0.718	6.062

Obs	DFBETAS				
	00	x1	x2	x3	x9
1	-0.131	-0.446	0.128	0.111	0.346
2	0.112	-0.307	-0.128	0.012	0.237
3	0.025	0.040	-0.024	-0.020	-0.033
4	0.493	0.479	-0.461	-0.572	-0.720
5	-0.087	-0.072	0.095	0.040	0.014
6	0.062	0.209	-0.059	-0.051	-0.182
7	-0.009	-0.030	0.014	-0.012	-0.020
8	-0.231	0.197	0.240	0.102	-0.048
9	-0.001	-0.021	0.005	-0.012	-0.024
10	0.947	-0.336	-0.955	-0.320	-0.482
11	-0.015	0.063	-0.001	0.060	0.121
12	-0.038	0.023	0.035	0.028	0.041
13	-0.019	0.089	-0.003	0.069	0.179
14	0.126	0.302	-0.182	0.027	0.355
15	-0.220	-0.354	0.221	0.214	0.242

16	-0.179	0.103	0.181	0.077	0.119
17	0.031	0.025	-0.032	-0.037	0.012
18	0.243	0.286	-0.241	-0.241	-0.092
19	0.109	-0.069	-0.111	-0.040	-0.058
20	-0.056	0.081	0.057	0.022	0.026
21	0.008	-0.015	-0.006	-0.019	-0.025
22	-0.264	0.337	0.302	0.023	-0.239
23	0.048	-0.098	-0.007	-0.241	-0.411
24	-0.036	0.135	0.033	0.098	-0.006
25	-0.192	-0.113	0.191	0.223	0.043
26	0.111	0.132	-0.105	-0.105	-0.120
27	0.009	0.112	0.003	-0.073	-0.106
28	-0.233	0.050	0.199	0.643	0.178
29	-0.038	-0.156	0.045	-0.006	-0.049
30	0.004	-0.148	-0.016	0.081	0.088
31	-0.171	-0.269	0.155	0.211	0.238
32	-0.171	-0.252	0.181	-0.159	0.196

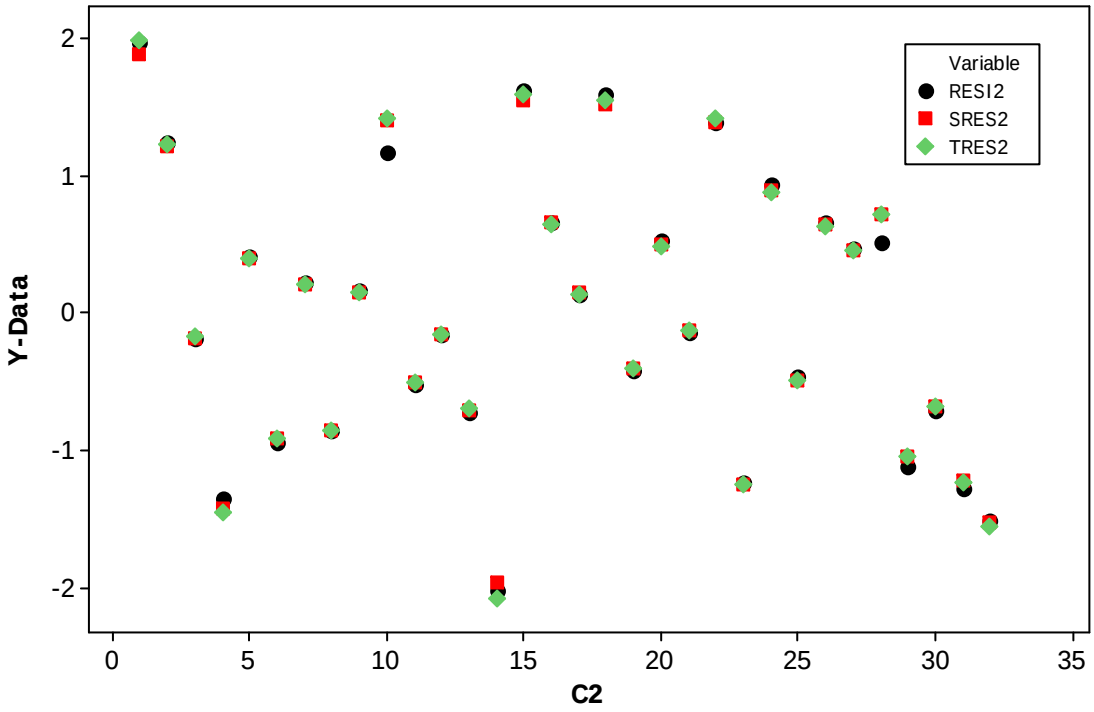
The studentized residuals are a first means for identifying outliers. Attention should be paid to studentized residuals that exceed +2 or -2 and get even more concerned about residuals that exceed 2 and even yet more concerned about residuals that exceed 3. In this case, point 1 and 14 may be the outliers.

The cutoff of leverage point is Hat greater than 2p/n = 10/32=0.3125, since Points 10 and 28 is much larger than 0.3125, they are the influential observation. At the same time, we find that PRESS of point 10 is 9.312, which contributes 15% of the total PRESS (63.18). The formal cutoff value for DIFFIT is the $2\sqrt{p/n}$ $=2\sqrt{5/32}=0.79$, both point 4, 33, 28 and 10 have the values of DIFFTS that exceed this value, and additionally point 14, 32 is very close to the cutoff.

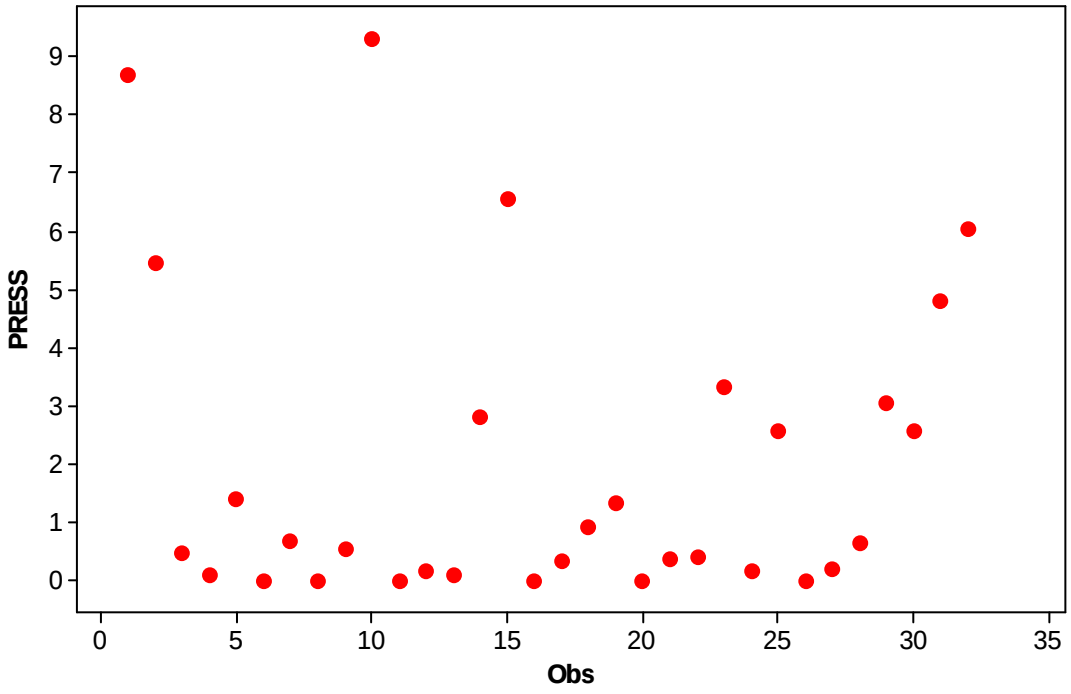
The formal cutoff value for DFBETAS is the $2/\sqrt{n}$ $=2/\sqrt{32}=0.353$, we immediately notice that point 10 in x2 has very large effect, and point 28 in x3 also large effects.

The observation 10 has the largest cook's Distance 0.286, which indicates that deletion of observation 10 would move the least square estimate to approximately the boundary of a 12% confidence region around β-hat. The second largest value is D28=0.144, and deletion of point 28 will move the estimate of β to approximately the edge of a 4% confidence region. Ideally we would like each estimate to stay within the boundary of a 10% confidence region. Therefore, we would conclude that **observation 10** is an influential using the cutoff of unity, and observation 28 is not influential.

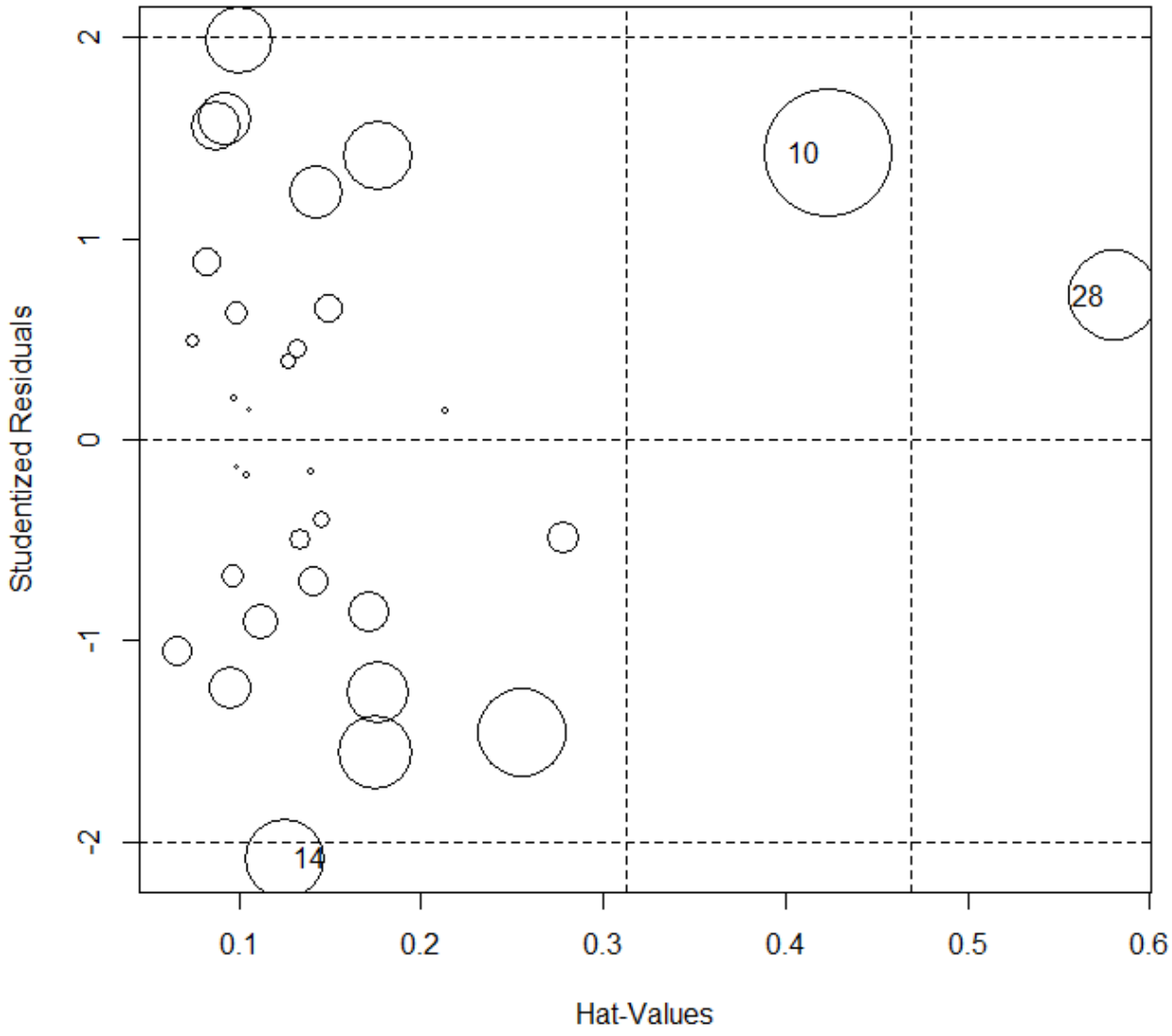
For better observation of the above table, we plot several residual parameters.



Plot of Standardized residual, studentized residual, and R-student residual



PRESS plot



Influence plot

- **Lack of fit test**

Calculation of Dii for the new model

obs	pred	resid	Delta0	Dii0	Delta1	Dii1	Delta2	Dii2	Delta3	Dii3	Delta4	Dii4
10	12.84	1.16	2.68	3.32	0.50	8.04	0.65	7.69	2.44	10.94	0.69	13.59
32	13.02	-1.52	2.18	3.22	2.03	4.70	0.24	5.46	1.99	6.23	1.35	3.73
26	13.64	0.66	0.15	11.09	1.95	2.50	0.19	1.52	0.83	1.96	0.72	2.41
28	13.69	0.51	1.80	12.64	0.04	14.59	0.68	9.90	0.87	15.05	1.24	14.44
31	13.78	-1.28	1.75	0.92	1.12	2.13	2.67	1.62	0.56	2.44	2.21	3.00
27	13.83	0.47	0.64	3.44	0.91	0.32	1.19	2.01	0.46	4.30	1.18	0.08

12	13.96	-0.16	1.55	4.34	0.56	1.43	1.10	4.33	0.55	3.17	0.35	3.81
22	14.12	1.38	2.11	2.26	0.45	5.05	2.10	0.52	1.90	4.35	0.21	9.12
13	14.32	-0.72	1.65	5.41	0.01	1.86	0.20	2.43	2.31	9.35	1.30	6.68
24	14.37	0.93	1.64	2.00	1.45	5.78	0.66	7.49	2.95	9.88	0.77	3.72
30	14.51	-0.71	0.19	4.11	2.30	7.66	1.31	8.51	0.87	1.97	0.40	5.49
11	14.52	-0.52	2.11	9.46	1.50	6.71	0.68	0.27	0.59	7.16	0.06	6.44
18	14.71	1.59	3.61	9.68	1.43	3.06	2.70	4.40	2.05	3.24	1.07	3.99
14	14.82	-2.02	2.18	0.33	0.91	7.22	1.56	7.04	2.54	5.04	2.23	1.84
9	15.04	0.16	1.27	7.18	0.62	6.36	0.37	4.99	0.06	1.71	0.30	7.54
29	15.11	-1.11	0.65	4.04	1.64	4.05	1.33	5.04	0.97	4.32	0.25	2.23
25	15.26	-0.46	0.98	5.27	0.68	5.71	0.32	6.49	0.40	3.05	0.60	2.44
20	15.48	0.52	0.31	3.53	0.66	5.33	1.38	4.08	0.38	3.84	0.11	13.45
7	15.59	0.22	0.36	8.09	1.07	3.70	0.08	5.87	0.19	11.40	0.63	4.21
21	15.84	-0.14	0.72	2.29	0.28	4.17	0.55	15.94	0.28	5.74	1.10	1.85
8	16.06	-0.86	1.00	5.32	1.27	16.37	0.44	9.31	0.38	3.62	1.52	7.65
17	16.16	0.14	0.27	15.86	0.56	5.04	1.38	3.74	0.52	3.51	1.48	3.70
5	16.39	0.41	0.83	4.13	1.65	7.93	0.25	2.79	1.21	1.24	0.83	4.78
19	16.42	-0.42	0.82	6.60	1.08	1.07	2.04	3.05	1.65	6.53	2.38	2.79
23	16.54	-1.24	1.90	5.69	2.86	4.09	2.48	3.76	3.21	3.79	1.05	3.98
16	16.64	0.66	0.96	2.94	0.57	6.28	1.30	2.67	0.85	2.18	1.61	2.66
15	16.88	1.62	0.38	2.66	0.35	0.38	1.81	0.12	2.56	0.43	2.97	0.29
2	17.06	1.24	0.73	1.01	1.42	1.19	2.18	1.04	2.59	1.18		
1	17.23	1.97	2.15	0.12	2.91	0.27	3.32	0.47				
3	17.29	-0.19	0.76	0.34	1.17	0.41						
6	17.44	-0.94	0.41	0.54								
4	18.65	-1.35										

Calculation of $\hat{\sigma}$ for the new model, here only list part of data.

Obs	n	Delta0	Dii0	n0	accu_delta
1	30	1.18406	0.0782	27	1.04908
2	3	1.80655	0.1186	15	1.32484
3	3	2.15308	0.1204	1	1.51910
4	9	0.67741	0.2681	11	1.28937
5	6	2.90919	0.2711	1	1.54701
6	4	2.97313	0.2898	15	1.72821
7	22	0.91214	0.3247	27	1.59677
8	9	2.17822	0.3309	14	1.63841

9	6	0.75611	0.3377	3	1.53080
10	1	0.34652	0.3835	15	1.40842
11	4	1.16658	0.4105	3	1.37435
12	6	2.56267	0.4303	15	1.44903
13	4	3.31965	0.4713	1	1.56381
14	30	2.09620	0.5188	22	1.58477
15	4	0.41046	0.5365	6	1.50336
16	27	1.75297	0.9153	31	1.50647
17	1	0.73082	1.0137	2	1.45595
18	6	2.17837	1.0422	2	1.48228
19	16	1.07898	1.0711	19	1.45458
20	4	2.58884	1.1805	2	1.49654
21	3	1.42226	1.1923	2	1.48528

The calculations yield $\hat{\sigma}$=1.51, from the SAS regression analysis, we find that $\sqrt{\mathrm{MSE}} = \sqrt{1.332} = 1.154$. Now if there is no appreciable lack of fit, we would expect to find that $\hat{\sigma} = \sqrt{\mathrm{MSE}}$. Currently, we do not consider the existence of lack of fit, since the $\hat{\sigma} > \sqrt{\mathrm{MSE}}$.

6. Inference of the parameters in the new model

Confidence interval for β of each variables:

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > |t|	95% Confidence Limits	
Intercept	1	-18.97399	9.48401	-2.00	0.0556	-38.43356	0.48559
x1	1	-1.42225	0.40477	-3.51	0.0016	-2.25277	-0.59173
x2	1	7.74975	2.28593	3.39	0.0022	3.05941	12.44010
x3	1	0.01489	0.00623	2.39	0.0242	0.00210	0.02767
x9	1	0.28363	0.04645	6.11	<.0001	0.18832	0.37893

Confidence interval for Mean and Variance of each variables

	Parameter	Estimate	95% Confidence Limits	
y	Mean	15.35000	14.71158	15.98842
	Std Deviation	1.77073	1.41960	2.35415
	Variance	3.13548	2.01526	5.54202
X2	Mean	3.81656	3.77288	3.86025
	Std Deviation	0.12117	0.09714	0.16109
	Variance	0.01468	0.00944	0.02595
X3	Mean	95.12500	76.32189	113.92811
	Std Deviation	52.15285	41.81111	69.33614
	Variance	2720	1748	4808
X9	Mean	14.40625	12.33641	16.47609
	Std Deviation	5.74096	4.60255	7.63249
	Variance	32.95867	21.18347	58.25497

10. Validation of new model

- DUPLEX method to split dataset

According to previous parts of this project, we know that best model is x1x2x7, we still use those three variables to conduct DUPLEX research. Here two equal sub dataset generated from original dataset has 16 observations each. One subset was treated as model dataset, that has observation number 28, 4, 25, 15, 14, 22, 26, 8, 29, 16, 7, 6, 12, 21, 30, 20. The others data was used as evaluation dataset.

Here is the R code to extract the DUPLEX dataset:

library(prospectr); dup = duplex(X=wine, k=16)

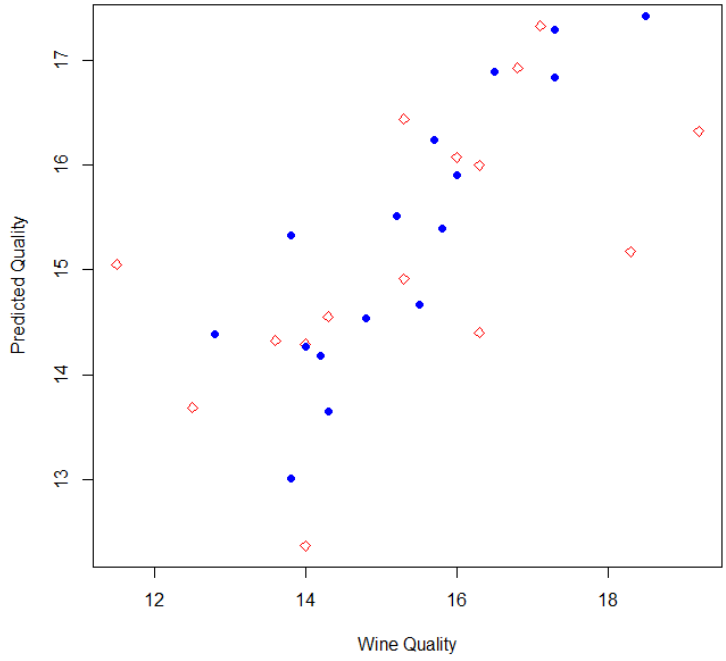
Now, we can redo the regression in R, based on the new dataset

Here is the result of comparison regression between estimation dataset and all dataset

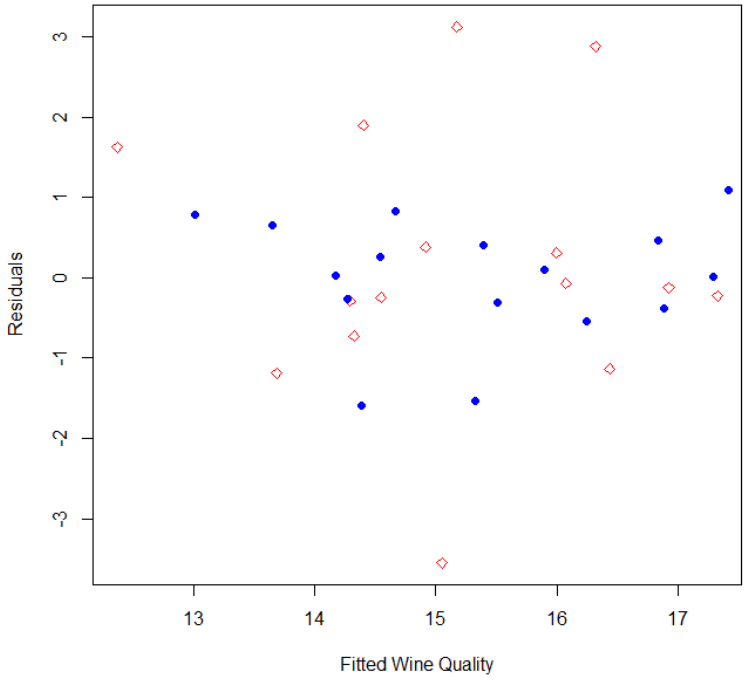
	Estimation data					ALL data				
	DF	Parameter Estimate	Standard Error	t	Pr	DF	Parameter Estimate	Standard Error	t	Pr
Intercept	1	61.27	31.39	1.95	0.08	1	-18.97	9.48	-2.00	0.06
x1	1	-1.73	0.64	-2.72	0.02	1	-1.42	0.40	-3.51	0.00

x2	**1**	12.89	1.82	7.09	.0001	**1**	7.75	2.29	3.39	0.00
x3	**1**	-9.47	8.78	-1.08	0.30	**1**	0.01	0.01	2.39	0.02
x9	**1**	32.98	15.74	2.10	0.06	**1**	0.28	0.05	6.11	.0001
MSres 12.0 Adj-Rsq 0.841						**MSres** 1.21 Adj-Rsq 0.615				

Predict value plot and residual plot may show how DUPLEX splits the original points into estimation and prediction data. The blue points are from predicted dataset, and the red points are from estimation dataset. The result shows that the prediction data set contains both interpolation and extrapolation points.



Linear regression fitted value verse original value of y based on estimation and prediction dataset.



Residuals plot of linear regression fit based on estimation and prediction dataset. In these plots, Dot is the model dataset, and Diamond is the prediction dataset. The residual plot seems normal, random distribution in the plot. Therefore, we can think this model works pretty well.

8. Indicator variable analysis

From the beginning, we already noticed that x1 is a indicator variable. Since it only has two level, 0 and 1, In analysis of SAS or R, these dummy variable can be treated as numerical variable. That is the reason, we did not do indicator variable analysis in the stage of model analysis.
Now, we consider x1x2x7 as our model, and indicator variable need to be considered, using SAS GLM package, it can be easily to do the analysis.

```
proc GLM data = A;
 class x1;
  model y= x1 x2 x3 x9 / solution;
run;
```

Parameter	Estimate		Standard Error	t Value	Pr > |t|
Intercept	-20.39623644	B	9.52078011	-2.14	0.0413
x1 0	1.42224769	B	0.40477014	3.51	0.0016
x1 1	0.00000000	B	.	.	.
x2	7.74975429		2.28593171	3.39	0.0022
x3	0.01488722		0.00623214	2.39	0.0242
x9	0.28362615		0.04645004	6.11	<.0001

Anova table:

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	64.59778120	16.14944530	13.37	<.0001
Error	27	32.60221880	1.20748959		
Corrected Total	31	97.20000000			

R-Square	Coeff Var	Root MSE
0.664586	7.158686	1.098858

9. Conclusion

Now we can propose a model for different wine location:
Wine in Cabernet:

Y0=-18.974+ 7.750 x2 + 0.015 x3+0.284x9

Wine in Shiraz:

Y1=-20.400 + + 7.750 x2 + 0.015 x3+0.284x9

Here, y is wine quality; x2 is the pH; x3=sulphates; x9=IonDegree

Or you can write as

y = -18.97399 - 1.42225 x1 + 7.74975 x2 + 0.01489 x3 + 0.28363 x9

here the x1 is treated as numerical number (0,1), 0 is the Cabernet, and 1 is the Shiraz.
Observation 10 is an influential, and it need to be checked.

.