KOREA 2020

DISCOVERY SUMMIT ONLINE

JMP를 이용한 제조/장치 산업의 Citizen Data Scientist 사례

Instructor : Kim Hanseong

(cvdkhs@wisemeca.com)

분석 방법론 개요

▣ 산업혁신 단계와 데이터 분석 방법의 변화

- 산업과 기술의 발전을 산업혁신 단계로 구분하고 현재를 산업혁신 4.0 단계로 구분하고 있음.
- 이러한 **산업혁신의 단계에 따라 지속적으로 Data 환경이 변화되고 이에 따른 Data의 분석 방법이 진화**되어 발전해 왔음.

구 분	핵심 내용	Data 환경	주요 방법론	Data 분석 원리
산업혁신 1.0	기계적 동력을 이용 (내연기관)	작동유무	Taylor 과학적 관리기법	Check List (합격/불합격 관리)
산업혁신 2.0	전기 동력을 이용 (컨베이어)	Only Y	SQC	Industry 1.0 + Visualization
산업혁신 3.0	IT 정보 기술을 활용	$Y=f(\text{Some Xs})$ (Small Data)	6Sigma	Industry 2.0 + Root Cause
산업혁신 4.0	ICT 기술의 융합	$Ys=f(\text{Many Xs})$ (Big Data)	Data Science	Industry 3.0 + Prediction

분석 방법론 개요

▣ Data Science의 용도에 의한 분류

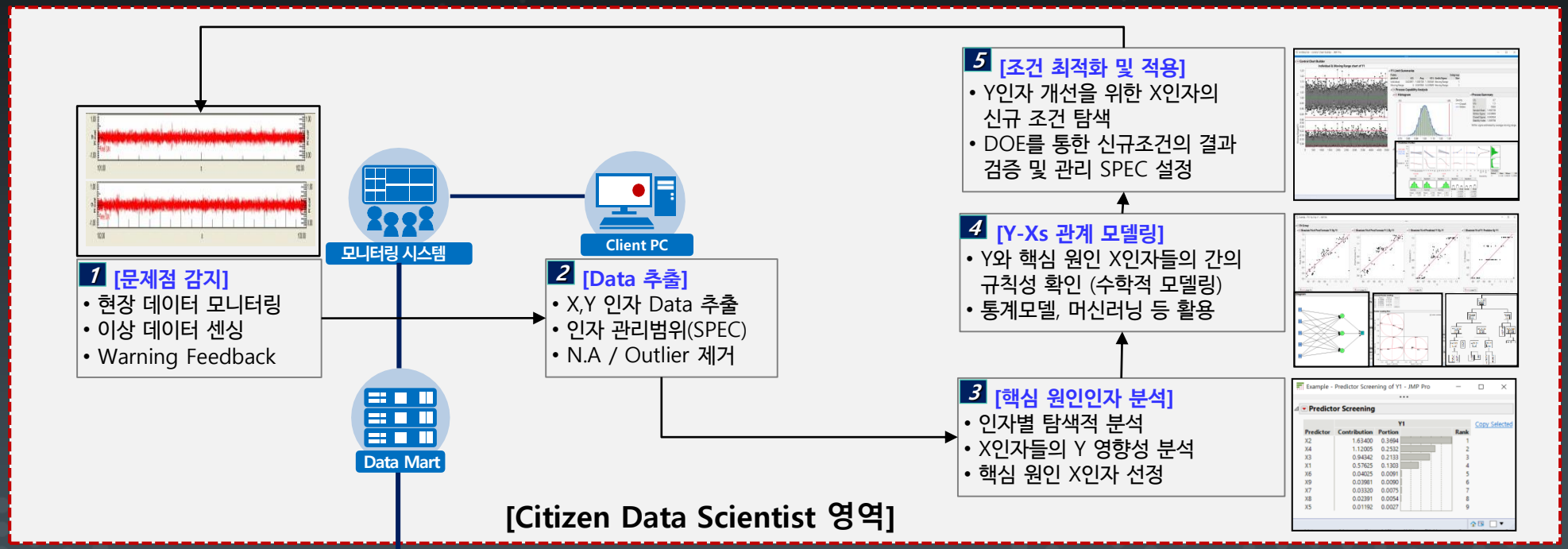
- Data Science는 다양한 유형의 Big Data를 활용하여 머신러닝/AI 개발등의 용도로 사용 되고 있음.
- 최근 **기업의 고질적인 현장 문제의 해결을 위하여 Data Science의 적용이 확산되고 있는 추세**임.
- Citizen Data Scientist는 미국의 Gartner에서 지칭한 해당 업종의 엔지니어에 의한 Data 분석 및 활용 전문가를 의미함.

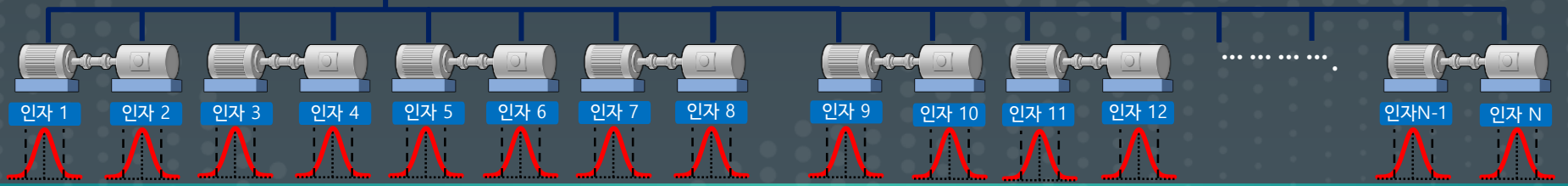
	Data Scientist	Citizen Data Scientist
추진범위 및 목적	• Big Data를 이용한 예측 기법을 활용, 머신 러닝 또는 딥 러닝 알고리즘의 개발 • 주요 의사 결정을 위해, Big Data를 이용한 각종 시각화 System 의 구축	• 기업 현장 문제를 해결하기 위해 Big Data를 이용 - 수율/불량의 원인 발굴 및 개선 - 설비 고장의 주요 원인 탐색 • 최적의 공정 조건 탐색 및 변수의 관리 SPEC 선정
추진인력	• Data Science Tool에 대해서 전문적으로 다룰 수 있는 인력 (Data Science 개발자) • 주로 Data Science를 전공하거나 상기 목적을 위해 고용된 인력을 의미	• Data Science를 교육받고 이를 업무에 적용하여 활용할 수 있는 인력 (Data Science 파워유저) • 주로 기존의 현장의 엔지니어 또는 관리자
주요기법	• 머신 러닝 또는 딥 러닝 등의 예측 기법 • 주로 Coding 언어 (Python, R 등)을 활용하여 분석	• 변수간 내재된 규칙(함수) 발견 또는 조건 최적화 기법 • 주로 Drag & Drop Tool (통계 S/W)을 활용하여 분석

분석 방법론 개요

▣ 기업 현장에서 문제 해결 패턴 (Citizen Data Scientist 방법)

- 최근 제조/장치 산업의 많은 기업은 현장 설비에 장착된 각종 Sensor를 통해서 Data를 집계하여 모니터링 하고 있습니다.
- 현장의 문제 발생시, 아래와 같은 형태로 발생되는 문제를 해결 및 개선하고 있음.



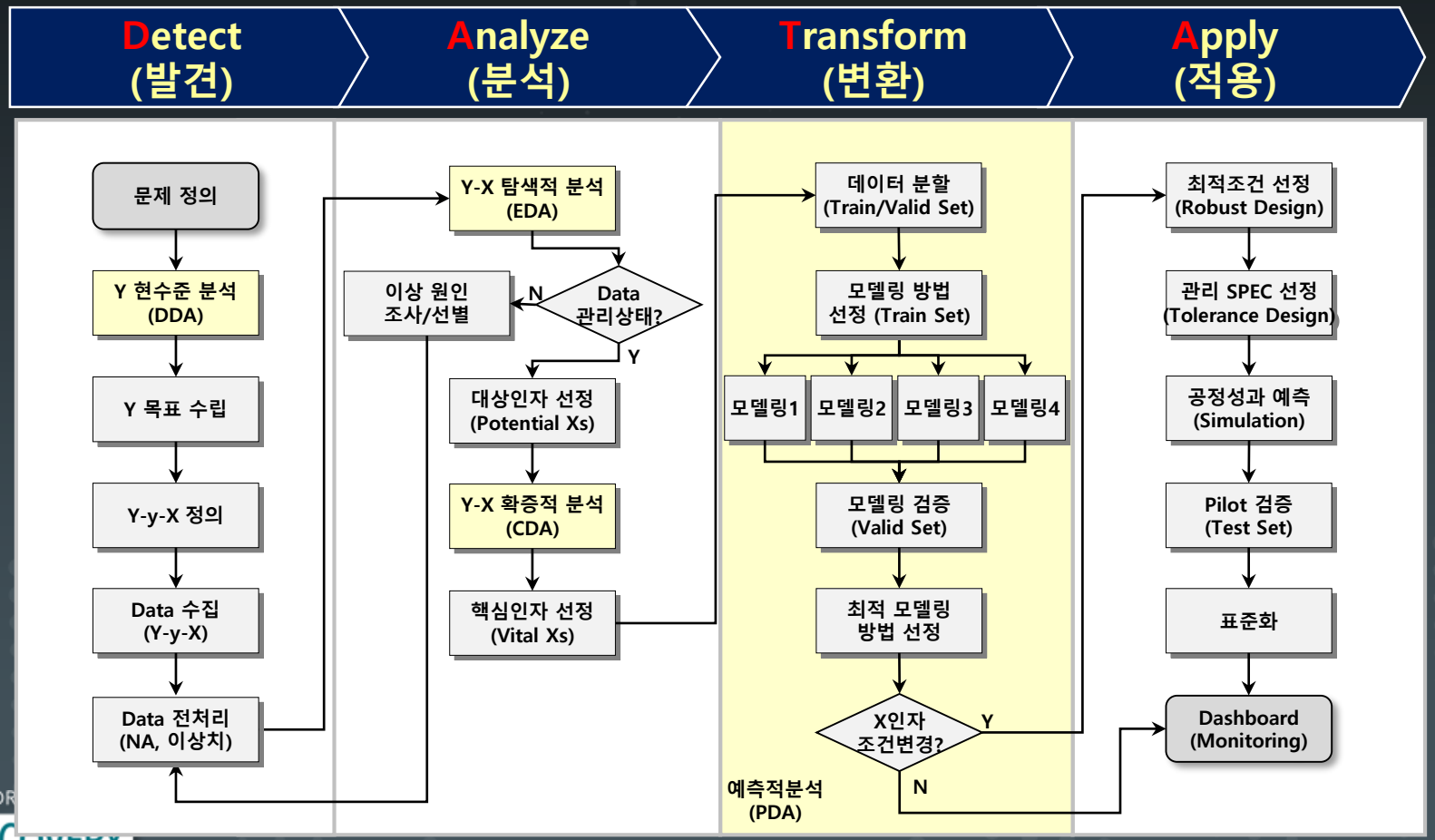


KOREA 2020 DISCOVERY SUMMIT ONLINE

분석 방법론 개요

▣ Citizen Data Scientist 문제해결 방법론 (DATA : Detect - Analyze - Transform - Apply)

- 따라서, Big Data 적용시 **기업의 현장 문제해결 방법을 표준화 하기 위하여 아래와 같은 문제해결 방법론을 개발 및 적용**하였음.
- 최초 S사 국내/베트남에 적용하여 수율개선 및 고장감소 등 적용 효과를 파악하고 여러 산업에 확대 적용 중.



구 분	세부 단계	세부 내용
발굴 (Detect)	Business case	• 현재 문제점 및 개선 기회에 대한 기술 • Y의 현재 수준 및 목표 기술
	Data Setup	• Y에 영향을 미치는 X변수들의 발굴 • Data수집 및 분석 전처리 (N.A 및 Outlier imputation)
분석 (Analyze)	Exploration	• 기초 통계량 및 Data의 분포 분석을 통한 현상 파악 • Graph분석을 통한 Y와 X인자들 간의 영향성 파악
	Screening	• 통계분석 (단변량/다변량)을 이용한 영향성 파악 • 핵심 X변수들의 선정 (Vital Few Xs)
변환 (Transform)	Modeling(Train)	• Data Set 구분 (Train/Valid Set) • Train Set을 통한 다양한 방법의 Modeling 분석
	Modeling(Valid)	• Valid Set을 통한 Modeling 분석 방법의 적합도 검증 • 현재 Process에 적합한 Best Modeling 선정
적용 (Apply)	Optimization	• Y의 최적화를 위한 X의 변수값 탐색 (Robust) • Y의 최적화를 위한 X의 관리범위 선정 (Tolerance)
	Feedback (Test)	• Modeling 및 X의 조건 현장 적용 결과 분석 • 표준화 및 Dashboard (시스템 연계)

Citizen Data Scientist
문제해결 방법론 DATA 사례

Instructor : Kim Hanseong

(cvdkhs@wisemeca.com)

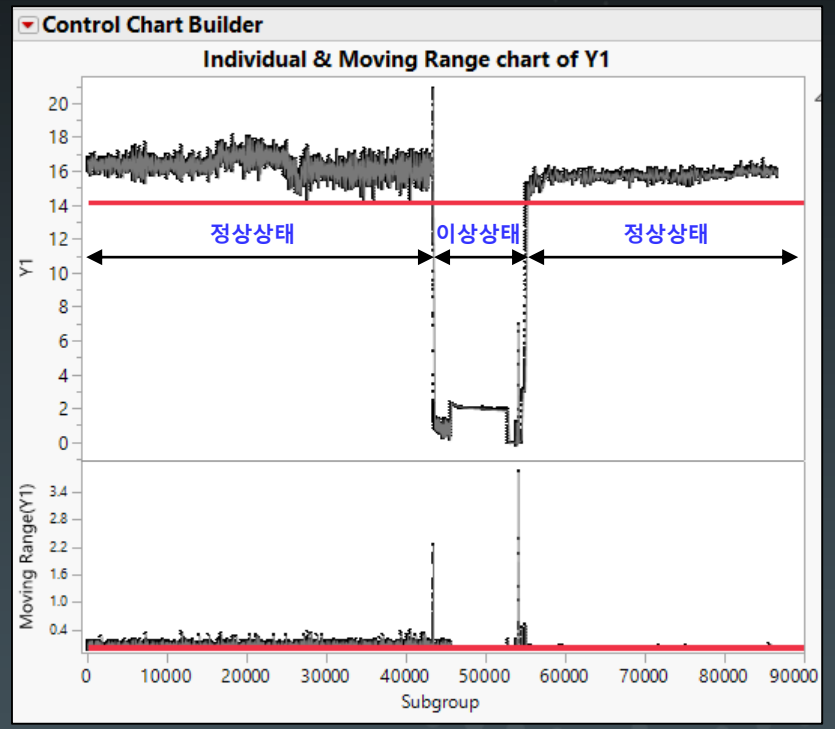
Detect Phase - ② Data Setup

▣ Data 항목 Setup

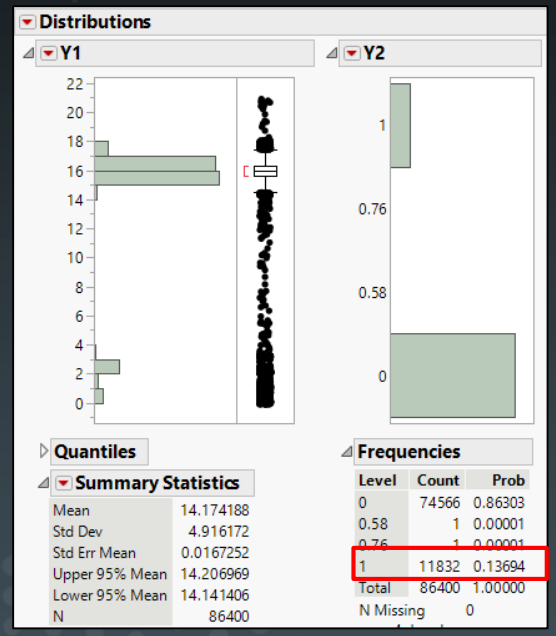
- Data에서 출력변수와 입력변수의 유형 및 Run Chart 확인을 통해 전체 Data에 대해서 파악함.
- 정상/이상 Data의 비율을 확인하고 만일 이상 Data가 충분하지 않을 경우 정상 Data의 Sampling 여부를 검토한다.

인자 구분	인자 번호	Data 유형	관리 SPEC
출력변수	Y1	연속형	> 10.2
	Y2	이산형	0
입력변수	X1	연속형	
	X2	연속형	
	X3	연속형	
	X4	연속형	
	X5	연속형	
	X6	연속형	
	X7	연속형	
	X8	연속형	
	X9	연속형	
	X10	연속형	
	X11	연속형	
	X12	연속형	
	X13	연속형	
	X14	연속형	
	X15	연속형	

- 총 2개의 Y인자, 15개의 X인자로 구성됨.
- Data는 초당 1회씩 1일간 집계됨 (86,400 Row)



- Y1 Run Chart 확인시 설비의 Error에 의한 이상상태 발생후 Reset으로 정상으로 원위치 된 Data 임.



- 비정상 Data는 약 13.7% 수준으로 분석시 양호함.

※ 일반적인 Big Data분석시 이상 Data가 15~20% 수준 권장됨.

Detect Phase - ② Data Setup

▣ Data Status : Missing Value & Sample Size

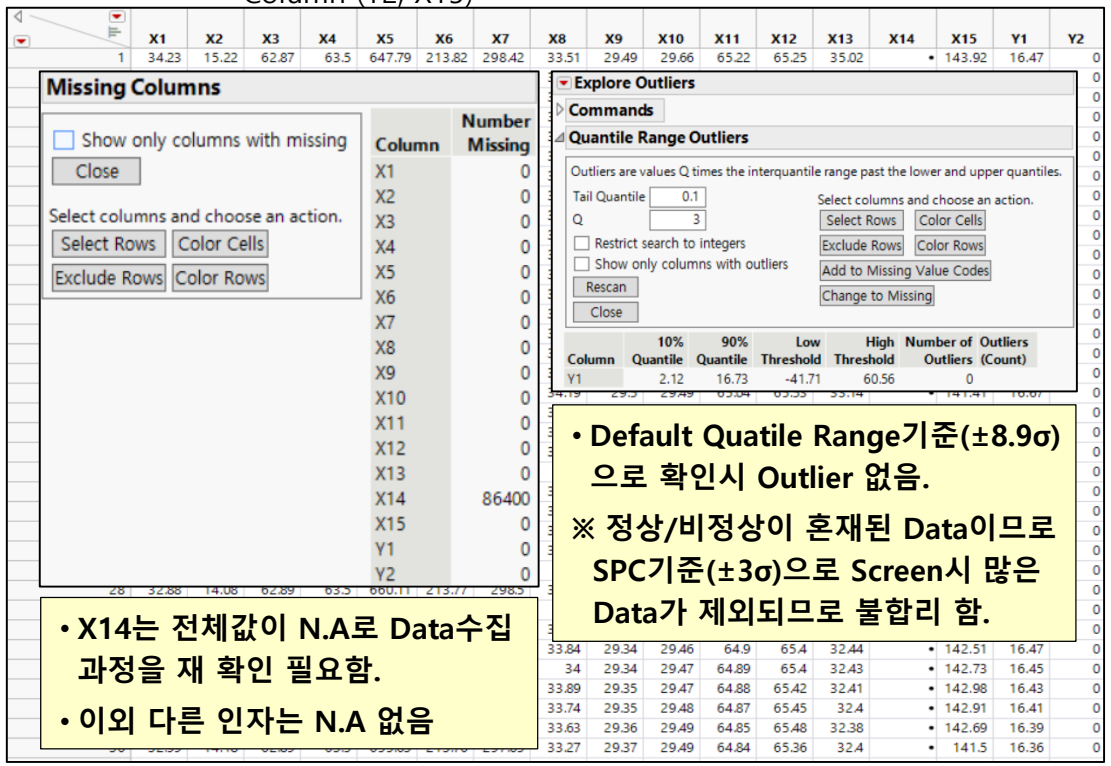
- Data 내의 Missing Value 및 Outlier를 확인하고, 현재 Sample Size가 분석에 충분한지를 확인함.
- 일반적 Missing Value 발생시 Imputation이 권장되며, Outlier는 제거를 검토한다.

Missing Value 및 Outlier

Initial	Col(17)×Row(86,400)	Final	Col(15)×Row(86,400)

Column (Y2, X15)

Row(86,400)

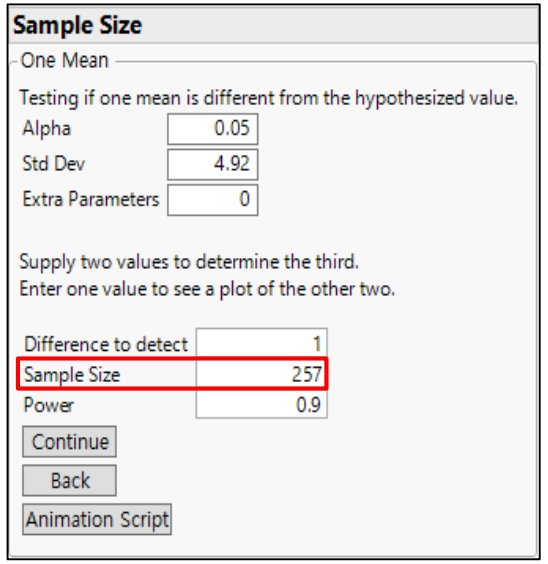


- X14는 전체값이 N.A로 Data수집 과정을 재 확인 필요함.
- 이외 다른 인자는 N.A 없음

- **Default Quatile Range기준(±8.9σ)으로 확인시 Outlier 없음.**
- **※ 정상/비정상이 혼재된 Data이므로 SPC기준(±3σ)으로 Screen시 많은 Data가 제외되므로 불합리 함.**

Sample Size

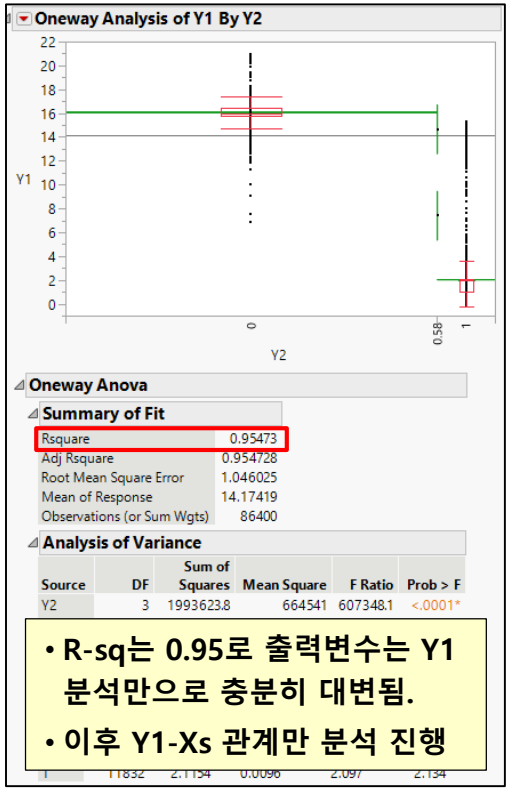
- 수집된 Data의 Power값을 바탕으로 결과를 충분히 신뢰할수 있을지 판단.



- **Result : sample size > 257로 현재 수집된 86,400개의 Data로 충분한 신뢰도 확보함.**

Y 변수간 관계

- 계량형 Y1과 계수형 Y2의 관계확인
  - ☞ Y1 > 10.2 경우 Y2 = 0 : 정상상태
  - Y1 < 10.2 경우 Y2 = 1 : 이상상태

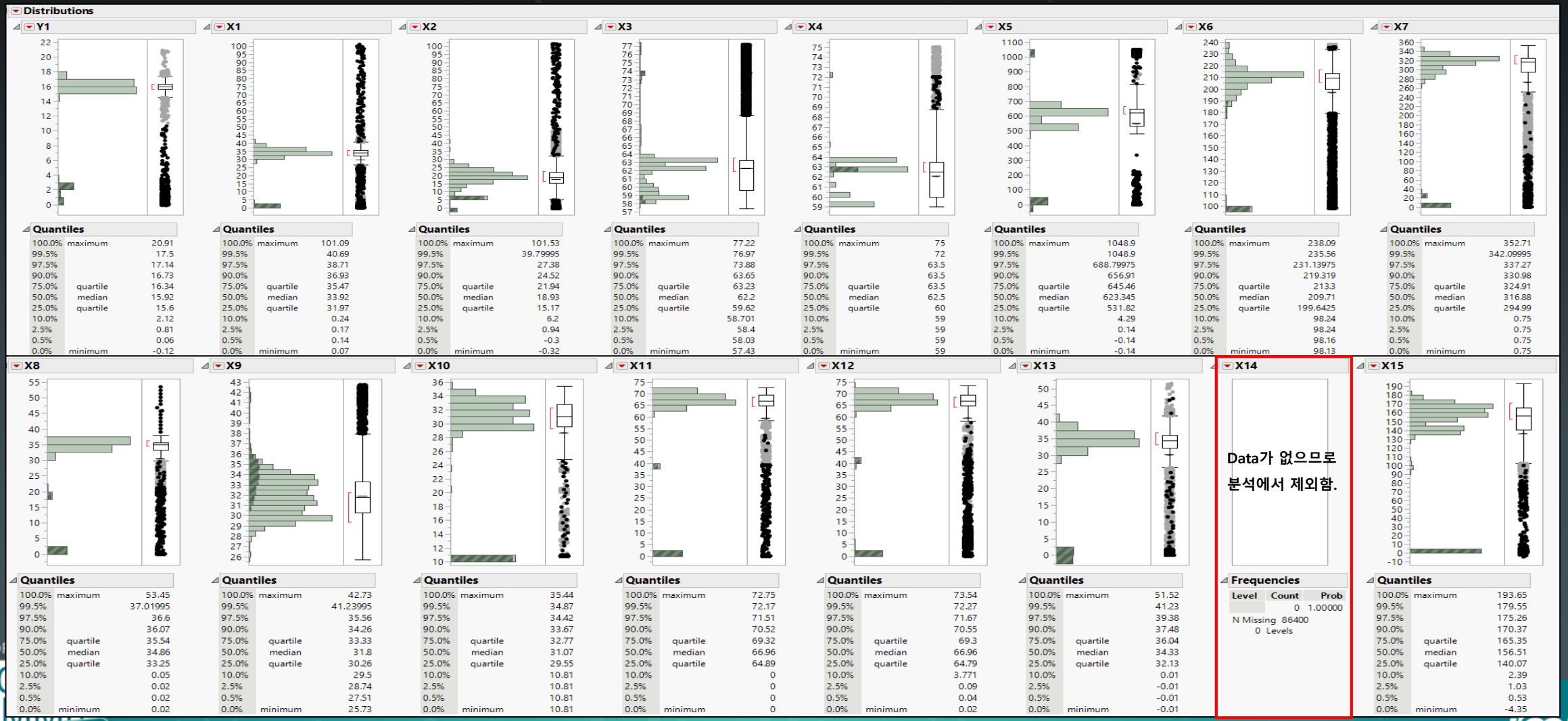


- **R-sq는 0.95로 출력변수는 Y1 분석만으로 충분히 대변됨.**
- **이후 Y1-Xs 관계만 분석 진행**

Analyze Phase - ③ Exploration

▣ Y-Xs Data 분포 Review (Histogram)

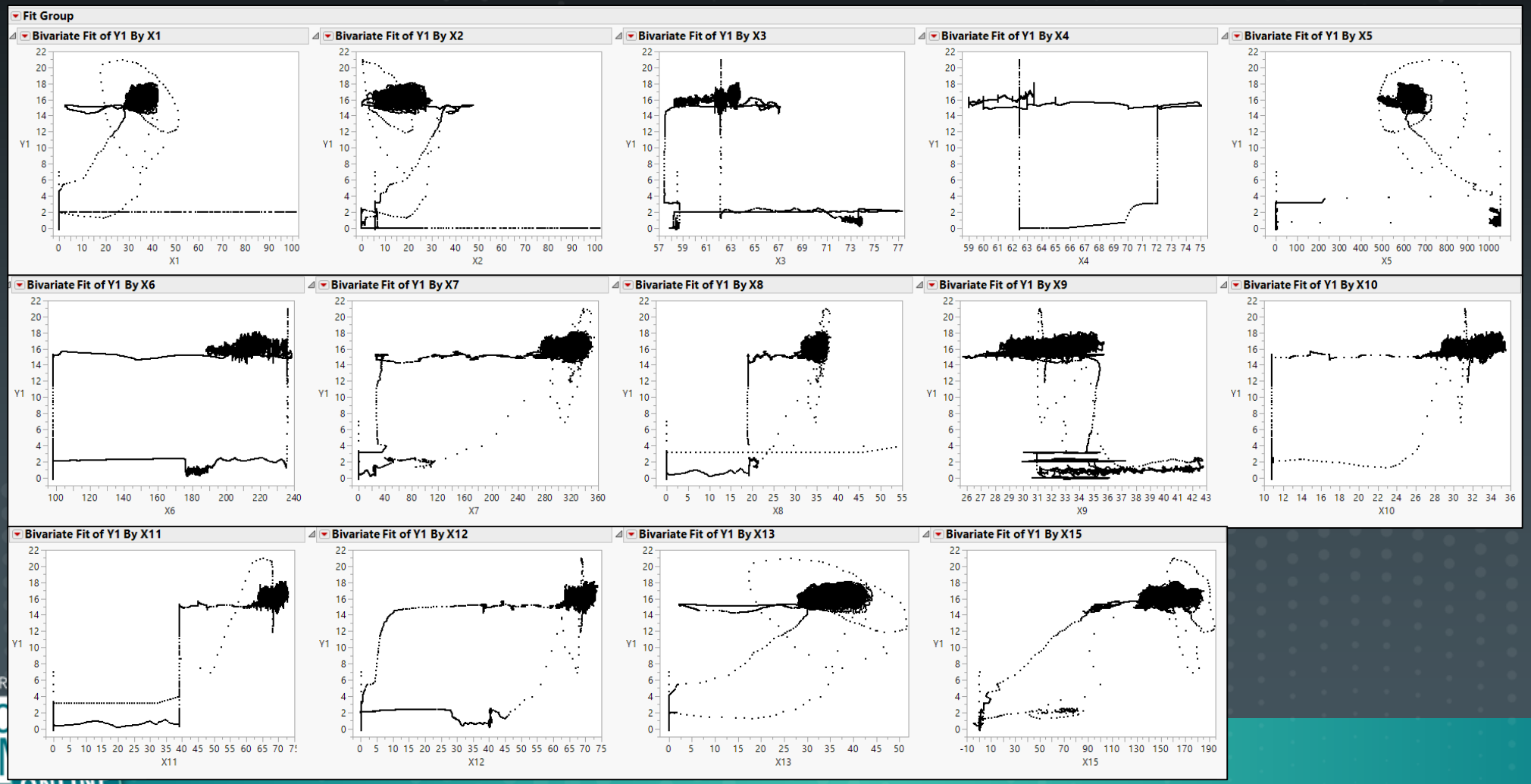
- Data의 분포에 대한 Histogram 확인을 통해 전체적인 분포에 대해 Study하고, Data가 없거나 하나의 값으로만 되어있는 Data를 제거한다.



Analyze Phase - ③ Exploration

▣ Y-Xs Data Review (산점도)

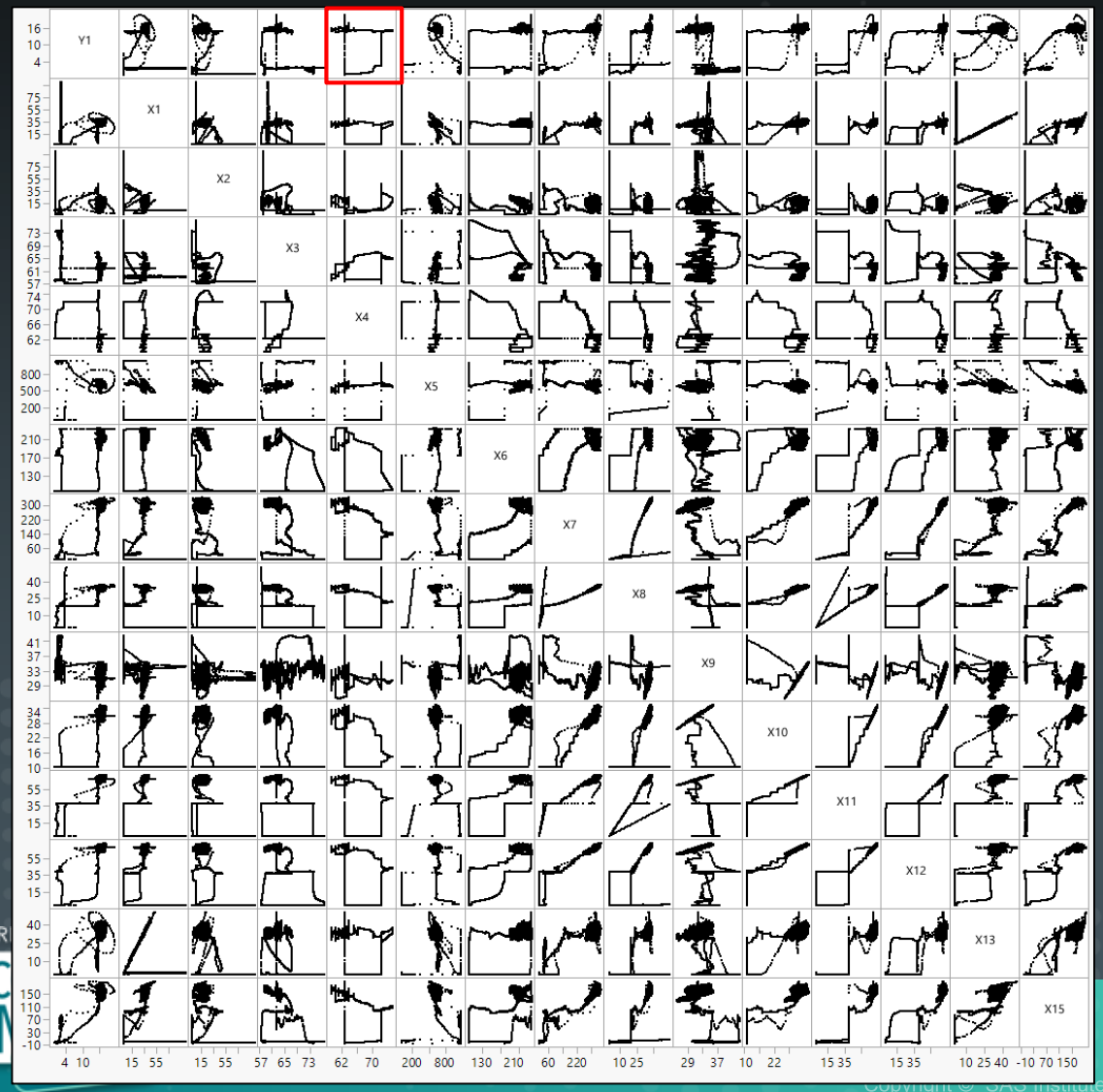
- Scatterplot으로 확인 시, 가운데 Data 집중된 형태가 발생하며, 이는 정상 운전상태로 확인됨.
- 초당 1회씩 Data가 집계되어 조건이 변화되는 선형의 패턴이 확인됨.



Analyze Phase - ④ Screening

▣ Multivariate Chart

- Y-Xs에 대하여 **Multivariate Chart 및 상관분석을 통해 통계적으로 의미가 없는 인자를 Screening**함.



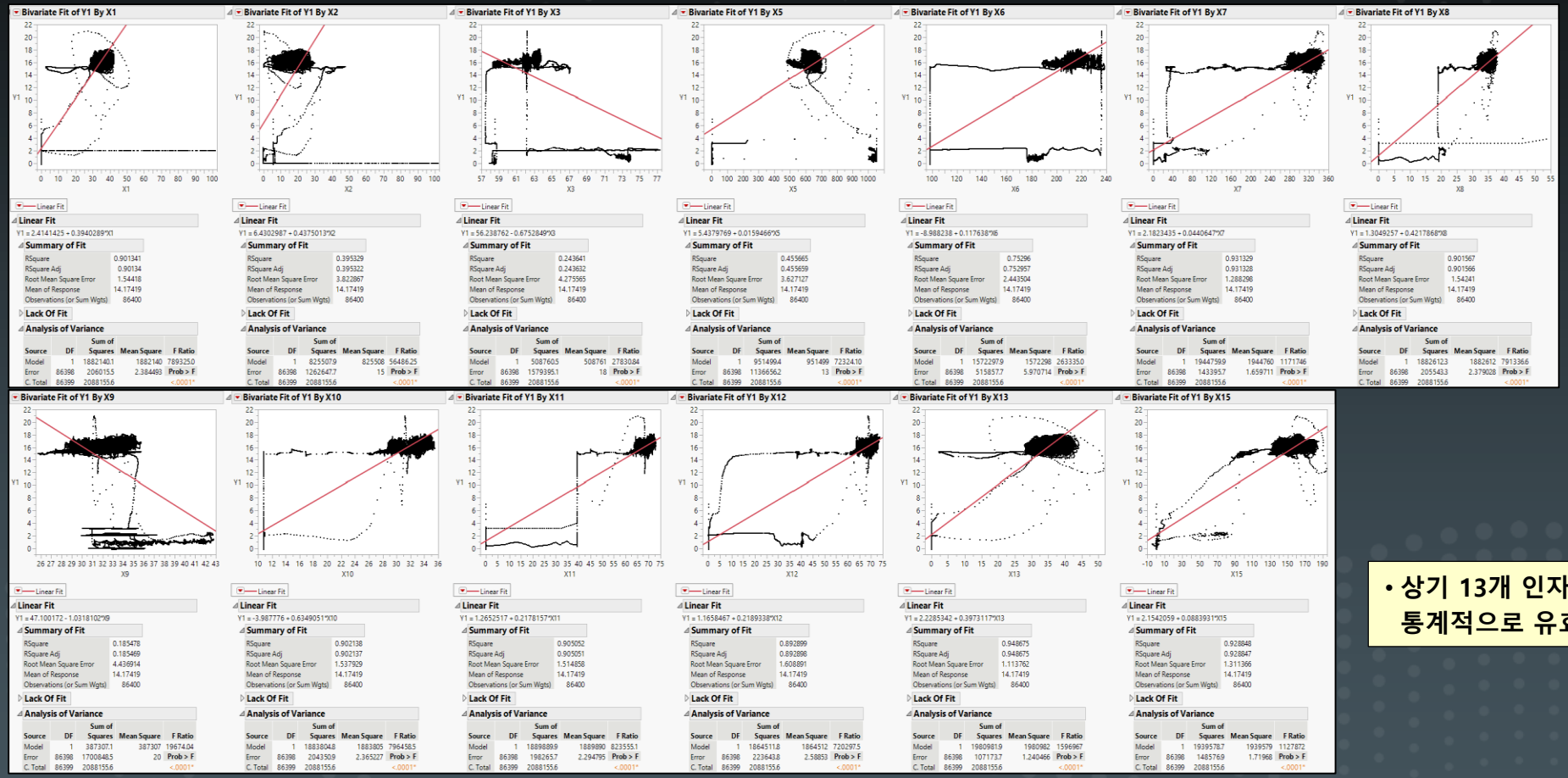
Corr	Y1	X1	X2	X3	X4	X5	X6	X7	X8	X9	X10	X11	X12	X13	X15
Y1	**1**	**0.949**	**0.629**	**-0.494**	**-0.122**	**0.675**	**0.868**	**0.965**	**0.950**	**-0.431**	**0.950**	**0.951**	**0.945**	**0.974**	**0.964**
X1	0.949	1	0.567	-0.505	-0.155	0.646	0.865	0.944	0.919	-0.376	0.941	0.923	0.917	0.971	0.931
X2	0.629	0.567	1	-0.506	-0.208	0.284	0.360	0.590	0.570	-0.323	0.565	0.562	0.551	0.593	0.624
X3	-0.494	-0.505	-0.506	1	0.424	-0.088	-0.321	-0.545	-0.475	0.356	-0.514	-0.452	-0.432	-0.501	-0.591
X4	-0.122	-0.155	-0.208	0.424	1	0.046	-0.204	-0.292	-0.236	0.150	-0.253	-0.208	-0.219	-0.162	-0.320
X5	0.675	0.646	0.284	-0.088	0.046	1	0.838	0.673	0.809	-0.109	0.648	0.823	0.824	0.671	0.638
X6	0.868	0.865	0.360	-0.321	-0.204	0.838	1	0.907	0.938	-0.289	0.897	0.944	0.954	0.891	0.872
X7	0.965	0.944	0.590	-0.545	-0.292	0.673	0.907	1	0.973	-0.363	0.988	0.969	0.967	0.970	0.984
X8	0.950	0.919	0.570	-0.475	-0.236	0.809	0.938	0.973	1	-0.314	0.952	0.997	0.992	0.948	0.960
X9	-0.431	-0.376	-0.323	0.356	0.150	-0.109	-0.289	-0.363	-0.314	1	-0.275	-0.296	-0.295	-0.387	-0.363
X10	0.950	0.941	0.565	-0.514	-0.253	0.648	0.897	0.988	0.952	-0.275	1	0.954	0.952	0.965	0.965
X11	0.951	0.923	0.562	-0.452	-0.208	0.823	0.944	0.969	0.997	-0.296	0.954	1	0.996	0.952	0.954
X12	0.945	0.917	0.551	-0.432	-0.219	0.824	0.954	0.967	0.992	-0.295	0.952	0.996	1	0.946	0.950
X13	0.974	0.971	0.593	-0.501	-0.162	0.671	0.891	0.970	0.948	-0.387	0.965	0.952	0.946	1	0.958
X15	0.964	0.931	0.624	-0.591	-0.320	0.638	0.872	0.984	0.960	-0.363	0.965	0.954	0.950	0.958	1

- 상관계수 (Correlation Coefficient)값을 기준으로 ±0.3 이하를 먼저 제외함.
  - 상관계수 기준 : ±0.3이하는 무의미, ±0.3~±0.7 유의미, ±0.7이상은 강한 관계 임
- X4는 상관계수가 -0.122 수준으로 Y와 상관성이 적음에 따라 인자를 Screening함

Analyze Phase - ④ Screening

▣ Regression 회귀분석

- 선정된 13개 인자에 대해서 회귀분석을 통해 P-Value 및 선형관계를 확인함.



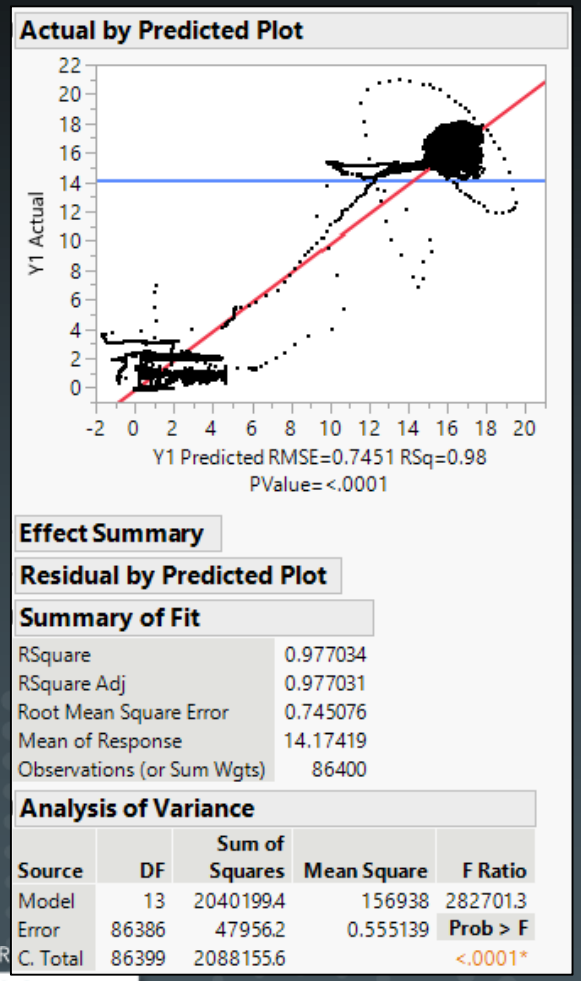
	R-sq	P-Value
X1	0.901	<0.0001
X2	0.395	<0.0001
X3	0.244	<0.0001
X5	0.456	<0.0001
X6	0.753	<0.0001
X7	0.931	<0.0001
X8	0.902	<0.0001
X9	0.185	<0.0001
X10	0.902	<0.0001
X11	0.905	<0.0001
X12	0.893	<0.0001
X13	0.949	<0.0001
X15	0.929	<0.0001

- 상기 13개 인자들 모두 R-sq>10%, P-Value<0.05로 통계적으로 유효함으로 인자 채택함.

Analyze Phase - ④ Screening

▣ 주인자 모델링 (SLS)

- 13개의 선정된 인자에 대해서 Y1에 대한 주인자 모델링을 통해 인자들의 Screening 여부를 확인함.



Term	Estimate	Std Error	t Ratio	Prob>|t|	Std Beta (%)		VIF
X15	0.049	0.00038	130.0	<.0001	0.537	18.2%	64.1
X11	0.103	0.00293	35.1	<.0001	0.449	15.2%	617.7
X10	0.227	0.00437	52.0	<.0001	0.340	11.5%	160.5
X13	0.128	0.00138	92.6	<.0001	0.313	10.6%	43.0
X6	-0.039	0.00034	-114.6	<.0001	-0.290	9.9%	24.2
X7	-0.013	0.00041	-31.4	<.0001	-0.282	9.6%	302.7
X8	-0.123	0.00427	-28.9	<.0001	-0.278	9.4%	347.0
X9	-0.345	0.00214	-161.1	<.0001	-0.144	4.9%	3.0
X5	0.003	0.00007	36.0	<.0001	0.111	3.8%	35.6
X3	0.128	0.00108	118.2	<.0001	0.093	3.2%	2.3
X1	0.035	0.00091	39.0	<.0001	0.085	2.9%	18.1
X12	0.005	0.00163	2.9	0.004	0.020	0.7%	186.1
X2	0.002	0.00062	3.2	0.0015	0.003	0.1%	2.9

- **주인자에 대한 SLS 모델링 확인시 모든 인자의 P-Value<0.05로 13개 인자를 채택함.**
- **Y1에 대한 주인자의 기여도는 Std Beta값의 %로 확인할수 있음.**
- **VIF값 확인시 다중공선성이 예상됨에 따라 이후 모델링시 차원의 축소 방법 사용 필요함.**

- 다중공선성 : X인자들 사이에 강한 관계가 존재하여 서로 영향을 주는 현상
- VIF (Variance Inflation Factor : 분산 팽창 계수)

$$VIF_i = \frac{1}{1 - R_i^2}$$

- 일반적으로 VIF>10 면 다중공선성이 존재함으로 해당 변수들의 차원을 축소시켜 모델링 필요
  → 주로 PCA, PLS, LASSO 방법이 사용됨

Analyze Phase - ④ Screening

▣ 각 X인자에 대한 Y의 기여도 확인

- 13개의 선정된 인자에 대해서 Y1에 대한 기여도에 대하여 Response Screening/Predict Screening을 통해 예측함.

Y	X	Count	PValue	LogWorth	FDR PValue	FDR LogWorth
Y1	X13	86400	0	55715.2	0	55714.1
Y1	X7	86400	0	50252.8	0	50252.0
Y1	X15	86400	0	49586.9	0	49586.2
Y1	X11	86400	0	44174.2	0	44173.7
Y1	X10	86400	0	43607.0	0	43606.6
Y1	X8	86400	0	43497.9	0	43497.5
Y1	X1	86400	0	43454.8	0	43454.6
Y1	X12	86400	0	41914.5	0	41914.3
Y1	X6	86400	0	26234.4	0	26234.2
Y1	X5	86400	0	11412.7	0	11412.6
Y1	X2	86400	0	9440.5	0	9440.4
Y1	X3	86400	0	5241.1	0	5241.1
Y1	X9	86400	0	3851.1	0	3851.1

- Response Screening 메뉴로 P-Value 및 FDR P-Value 확인시 모두 <0.05로 유효함
- Y에 대한 기여도는 FDR Logworth로 파악함.

Predictor Screening

Y1

Copy Selected

Predictor	Contribution	Portion	Rank
X15	320636	0.2441	1
X13	228575	0.1740	2
X7	222581	0.1694	3
X11	177444	0.1351	4
X1	106247	0.0809	5
X8	76578	0.0583	6
X10	74981	0.0571	7
X6	47764	0.0364	8
X12	39312	0.0299	9
X5	12851	0.0098	10
X3	3073	0.0023	11
X2	2512	0.0019	12
X9	1004	0.0008	13

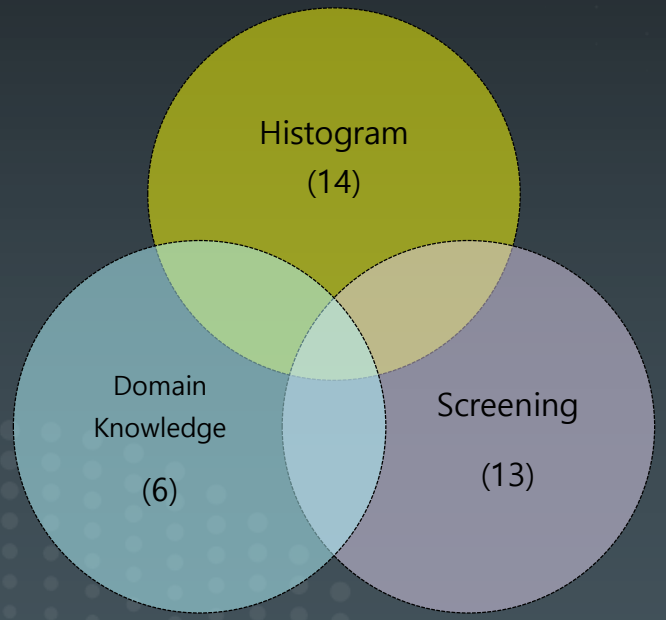
- Predict Screening 메뉴로 Y1에 대한 기여도를 확인함.
  - Bootstrap Forest에 의한 산정

Analyze Phase - ④ Screening

▣ X 인자에 대한 Screening 결과 - Vital X 선정

• Histogram, 상관/회귀분석, Prediction등의 통계 지표를 바탕으로 Vital X 인자를 선정 (총 13개 인자)

	Initial (Number)	Histogram	Correlation & Regression	Screening	Domain Knowledge	최종결정 (Final)
X인자	15	14	13	13	7	13



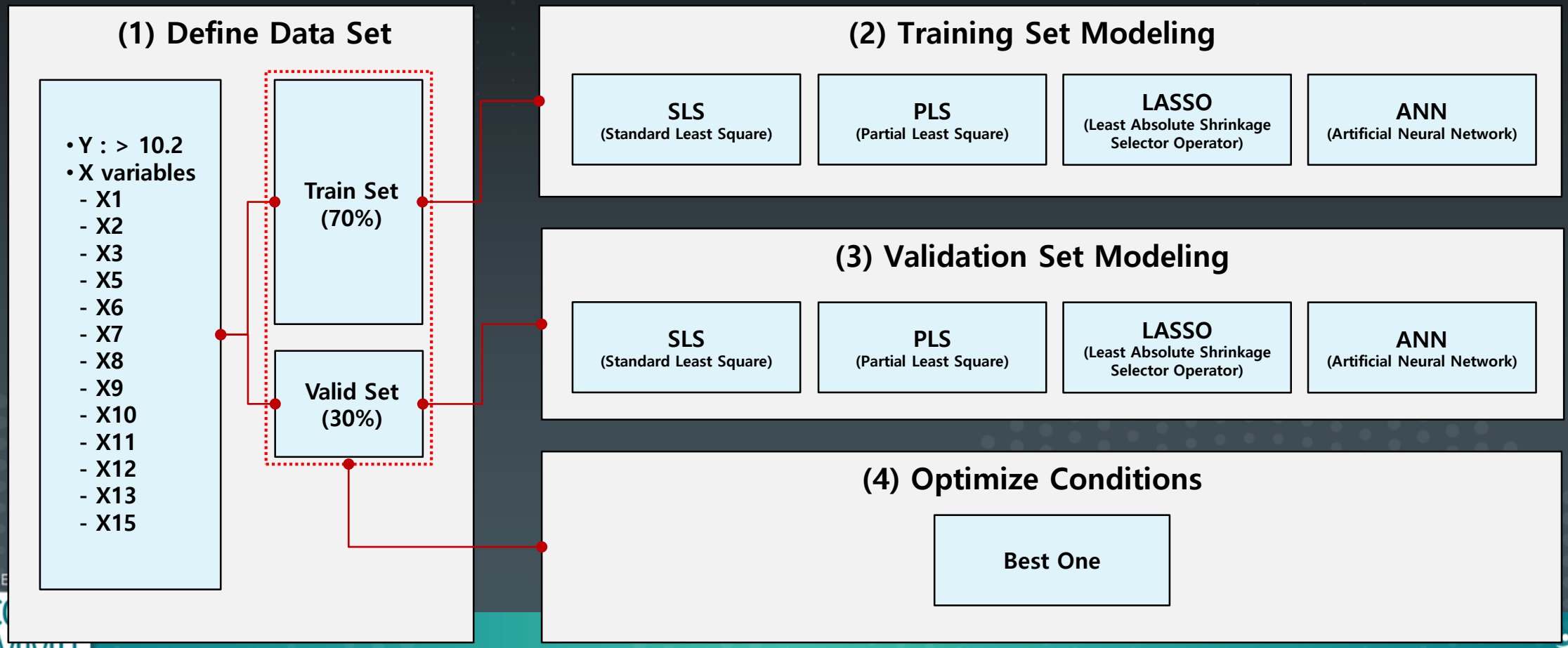
Variables	Histogram	Correlation Regression	Screening	Domain Knowledge	최종결정 (Final)
X1	●	●	●	●	●
X2	●	●	●	●	●
X3	●	●	●		●
X4	●				
X5	●	●	●		●
X6	●	●	●		●
X7	●	●	●	●	●
X8	●	●	●		●
X9	●	●	●		●
X10	●	●	●	●	●
X11	●	●	●	●	●
X12	●	●	●		●
X13	●	●	●	●	●
X14					
X15	●	●	●	●	●

KOREA 2020 DISCOVERY SUMMIT ONLINE

Transform Phase - ⑤ Modeling

▣ Modeling을 위한 Data Set 분할

- Data Science에서 모델링 분석시, 주어진 Data Set에서 지나치게 R-sq가 높게 분석되면 이후 Data Set이 바뀌게 되면 오히려 R-sq가 현저하게 떨어지는 현상이 나타나는 경향이 있음. (이를 과적합 (Overfitting) 이라고 함.)
- 따라서, **Data Set을 Train/Valid으로 분할하고, 각 Set을 여러가지 분석 Tool로 분석하여 가장 R-sq의 변화가 적은 모델을 최종 선정**함.



Transform Phase - ⑤ Modeling

▣ Modeling을 위한 Data Set 분할

- JMP에서는 Row State 라는 Data Set 분할 메뉴가 있어 이를 이용하여 Training Set, Validation Set으로 구분하여 분석이 용이하게 해줌.
- **일반적인 Train:Valid의 비율은 약 7:3 수준으로 분할하여 검증**

Make Validation Column

Stratified Validation Column

Randomly partitions the rows into training, validation and test set across levels of the stratification variable(s). Use this option when of a column's levels in each of the training, validation and test set

Stratification Columns: Y1

Specify rates or relative rates

		Adjusted Rates	Row Counts
Training Set	0.7	0.7	60480
Validation Set	0.3	0.3	25920
Test Set	0	0	0
Excluded Rows			0
Total Rows			86400

22/1 Cols

	X1	X2	X3	X5	X6	X7	X8	X9	X10	X11	X12	X13	X15	Y1	Validation	Train Set	Valid Set
1	•	15.22	62.87	647.79	213.82	298.42	33.51	29.49	29.66	65.22	65.25	35.02	143.92	16.47	Validation	•	•
2	34.19	15.2	62.87	647.54	213.82	300.09	33.69	29.48	29.65	65.21	65.36	34.91	143.71	16.49	Training	•	•
3	34.15	15.18	62.87	647.01	213.82	299.11	33.61	29.47	29.64	65.2	65.46	34.72	143.44	16.51	Training	•	•
4	34.11	15.16	62.87	646.6	213.82	297.54	33.46	29.46	29.63	65.19	65.44	34.49	143.2	16.53	Validation	•	•
5	34.07	15.14	62.87	646.24	213.81	299.6	33.53	29.46	29.62	65.18	65.36	34.41	144.26	16.54	Validation	•	•
6	34.03	15.12	62.87	646.92	213.81	300.36	33.57	29.47	29.61	65.17	65.29	34.43	144.33	16.56	Validation	•	•
7	33.99	15.1	62.88	648.53	213.81	299.3	33.54	29.48	29.6	65.16	65.22	34.45	143.43	16.58	Training	•	•
8	33.95	15.08	62.88	648.93	213.81	298.23	33.45	29.48	29.59	65.15	65.24	34.47	142.52	16.59	Training	•	•
9	33.91	15.07	62.88	647.95	213.81							4.49	142.39	16.61	Training	•	•
10	33.87	15.05	62.88	647.75	213.8							4.44	143.29	16.63	Training	•	•
11	33.83	15.03	62.88	647.74	213.8							4.23	144.19	16.64	Training	•	•
12	33.79	15.01	62.88	648.28	213.8							4.08	144.67	16.66	Training	•	•
13	33.75	14.99	62.88	648.87	213.8							3.94	143.97	16.68	Training	•	•
14	33.71	14.97	62.88	649.61	213.8							3.79	143.21	16.69	Training	•	•
15	33.65	14.95	62.88	651.6	213.8							3.61	142.48	16.71	Validation	•	•
16	33.58	14.93	62.88	653.87	213.79							3.42	141.77	16.7	Validation	•	•
17	33.52	14.91	62.88	656.05	213.79							3.23	141.12	16.69	Training	•	•
18	33.45	14.84	62.88	655.89	213.79							3.14	141.41	16.67	Validation	•	•
19	33.39	14.75	62.88	654.71	213.79							3.08	141.97	16.66	Training	•	•
20	33.32	14.66	62.88	653.53	213.79							33	142.11	16.64	Training	•	•
21	33.26	14.58	62.88	653.03	213.78							2.89	142.04	16.63	Validation	•	•
22	33.19	14.49	62.88	653.21	213.78							2.82	141.8	16.61	Validation	•	•
23	33.13	14.4	62.89	653.38	213.78							2.79	141.42	16.6	Training	•	•
24	33.06	14.32	62.89	653.56	213.78							2.77	141.04	16.58	Training	•	•
25	33	14.23	62.89	654.14	213.78							2.74	141.12	16.57	Training	•	•
26	32.93	14.14	62.89	656	213.77	303.84	33.51	29.39	29.43	64.95	65.53	32.7	141.91	16.56	Validation	•	•
27	32.88	14.07	62.89	658.45	213.77	301.08	33.7	29.37	29.43	64.94	65.58	32.59	142.27	16.54	Training	•	•
28	32.88	14.08	62.89	660.11	213.77	298.5	33.87	29.35	29.44	64.93	65.52	32.5	142.6	16.53	Training	•	•
29	32.9	14.09	62.89	657.21	213.77	296.74	34	29.35	29.45	64.92	65.47	32.47	142.8	16.51	Training	•	•
30	32.91	14.1	62.89	657.42	213.77	299.31	34.07	29.34	29.45	64.91	65.43	32.46	142.42	16.49	Training	•	•
31	32.92	14.11	62.89	655.7	213.76	302.34	33.84	29.34	29.46	64.9	65.4	32.44	142.51	16.47	Validation	•	•
32	32.93	14.13	62.89	657	213.76	303.64	34	29.34	29.47	64.89	65.4	32.43	142.73	16.45	Validation	•	•
33	32.95	14.14	62.89	657.98	213.76	304.08	33.89	29.35	29.47	64.88	65.42	32.41	142.98	16.43	Validation	•	•
34	32.96	14.15	62.89	658.1	213.76	295.74	33.74	29.35	29.48	64.87	65.45	32.4	142.91	16.41	Training	•	•

Data Filter f...

Data Filter

Clear | Favorites | Help

Select | Show | Include

60480 matching rows

Inverse

Validation (2)

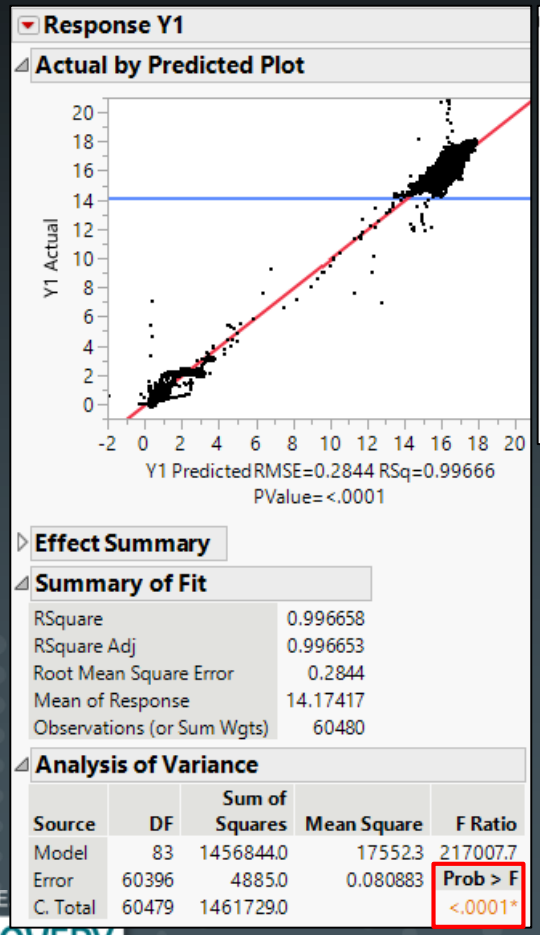
Training | Validation

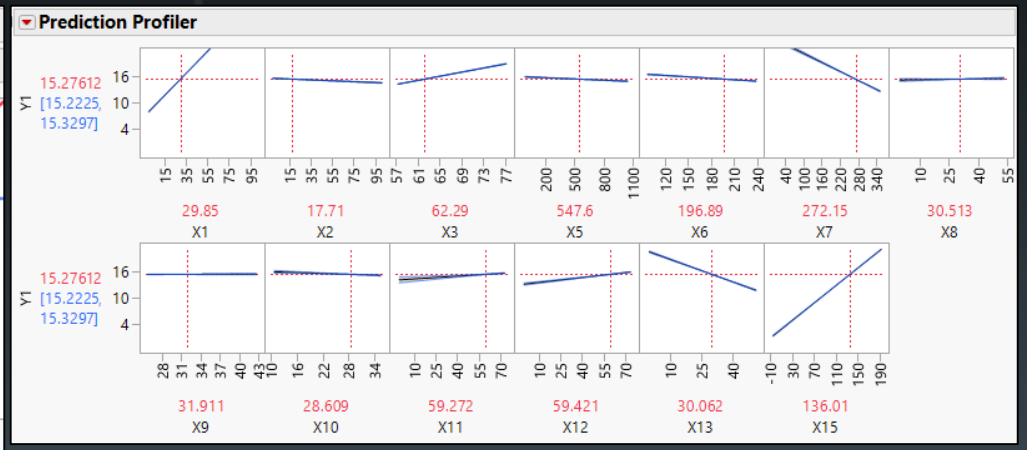
AND | OR

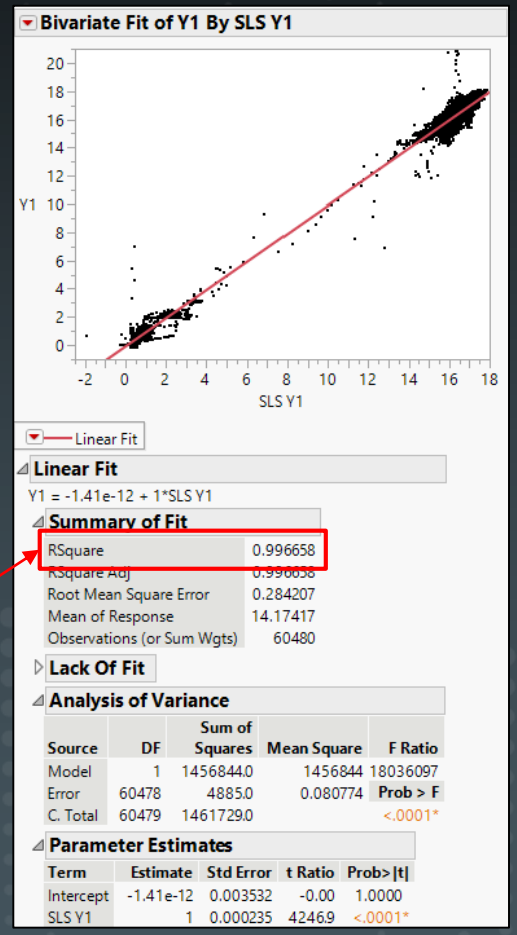
Transform Phase - ⑤ Modeling (Train Set)

▣ SLS (Standard Least Square, 표준최소자승법) 모델링

- 표준최소자승법으로 모델링 결과 X인자들 사이에 주효과 와 교호효과가 존재 확인됨
- 교효효과를 포함한 모델링의 결과, 실제값과 예측값의 일치도는 약 99.67% 수준임.







- **SLS 모델링으로 예측한 실제값과 예측값은 R-sq 99.67% 수준임.**

Transform Phase - ⑤ Modeling (Train Set)

▣ PLS (Partial Least Square, 부분최소자승법) 모델링

- X인자들이 많고, X인자들 사이에 상관성이 존재할때 (VIF값이 높을때) PCA(주효과분석)와 SLS를 동시에 수행하는 분석 모델임.
- PLS 분석 결과 실제값과 예측값의 일치도는 약 97.69% 임.

Partial Least Squares

Model Launch

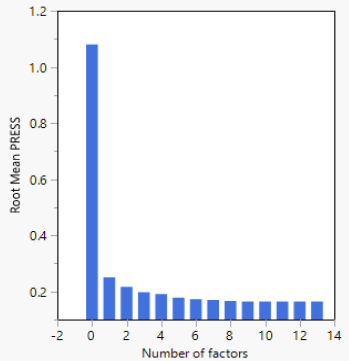
Model Comparison Summary

Method	Number of rows	Number of factors	Percent Variation Explained for Cumulative X	Percent Variation Explained for Cumulative Y	Number of VIP > 0.8
NIPALS	60480	13	100	97.685783	10

KFold Cross Validation with K= 7 and Method=NIPALS

Number of factors	Root Mean PRESS	van der Voet T^2	Prob > van der Voet T^2	Q^2	Cumulative Q^2	R^2X	Cumulative R^2X	R^2Y	Cumulative R^2Y
0	1.080147	10274.836	<.0001*	-0.000230	-0.000230	0.000000	0.000000	0.000000	0.000000
1	0.250362	2219.0472	<.0001*	0.946263	0.946263	0.760288	0.760288	0.946287	0.946287
2	0.217309	1070.5768	<.0001*	0.959516	0.997825	0.080136	0.840424	0.013268	0.959555
3	0.197986	504.89160	<.0001*	0.966395	0.999927	0.053910	0.894334	0.006890	0.966445
4	0.191509	350.72331	<.0001*	0.968558	0.999998	0.045596	0.939930	0.002157	0.968602
5	0.178487	103.64874	<.0001*	0.972688	1.000000	0.020241	0.960172	0.004139	0.972741
6	0.172938	172.81248	<.0001*	0.974360	1.000000	0.028131	0.988303	0.001684	0.974425
7	0.170421	171.24235	<.0001*	0.975101	1.000000	0.006195	0.994498	0.000720	0.975145
8	0.166859	57.538519	<.0001*	0.976131	1.000000	0.002036	0.996534	0.001022	0.976167
9	0.164874	8.133222	0.0040*	0.976696	1.000000	0.000845	0.997379	0.000570	0.976737
10	0.164653	4.598828	0.0240*	0.976758	1.000000	0.001126	0.998505	0.000072	0.976809
11	0.164494	0.205235	0.6690	0.976803	1.000000	0.001009	0.999514	0.000041	0.976850
12	0.164470	0.000199	0.9930	0.976810	1.000000	0.000206	0.999720	0.000008	0.976858
13	0.164469	0.000000	1.0000	0.976810	1.000000	0.000280	1.000000	0.000003	0.976861

Root Mean PRESS Plot

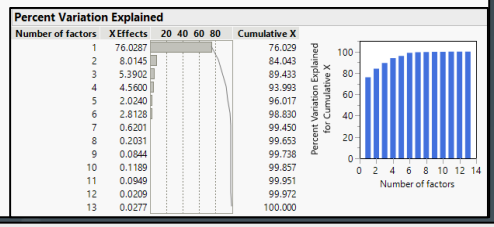


Note: The minimum root mean PRESS is 0.16447 and the minimizing number of factors is 13.

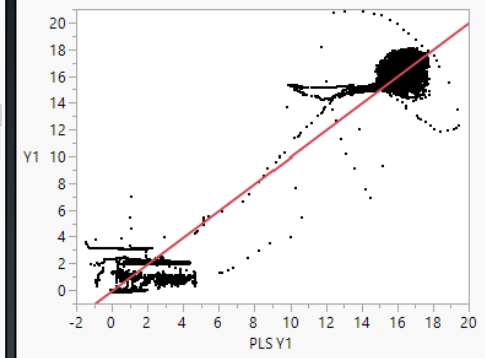
- 모델에 포함된 주성분은 13개임
  - Voet T^2의 P-value값이 가장 큼 (H0:이 모델이 최적값임)

Percent Variation Explained

Number of factors	X Effects	Cumulative X
1	76.0287	76.029
2	8.0145	84.043
3	5.3902	89.433
4	4.5600	93.993
5	2.0240	96.017
6	2.8128	98.830
7	0.6201	99.450
8	0.2031	99.653
9	0.0844	99.738
10	0.1189	99.857
11	0.0949	99.951
12	0.0209	99.972
13	0.0277	100.000



Bivariate Fit of Y1 By PLS Y1



Linear Fit

Linear Fit

Y1 = 5.4e-13 + 1*PLS Y1

Summary of Fit

RSquare	0.976858
RSquare Adj	0.976857
Root Mean Square Error	0.747888
Mean of Response	14.17417
Observations (or Sum Wgts)	60480

Lack Of Fit

Analysis of Variance

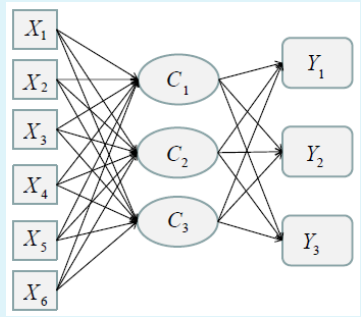
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1427901.5	1427901	2552846
Error	60478	33827.6	0.559337	Prob > F
C. Total	60479	1461729.0		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	5.4e-13	0.009378	0.00	1.0000
PLS Y1	1	0.000626	1597.8	<.0001*

Prediction Profiler

- PLS는 잠재변수의 도출과 Y와의 회귀분석을 동시에 실시하여 설명력 높은 주성분을 도출하는 방법
- PLS = PCA(주성분분석) + SLS (표준최소자승법)

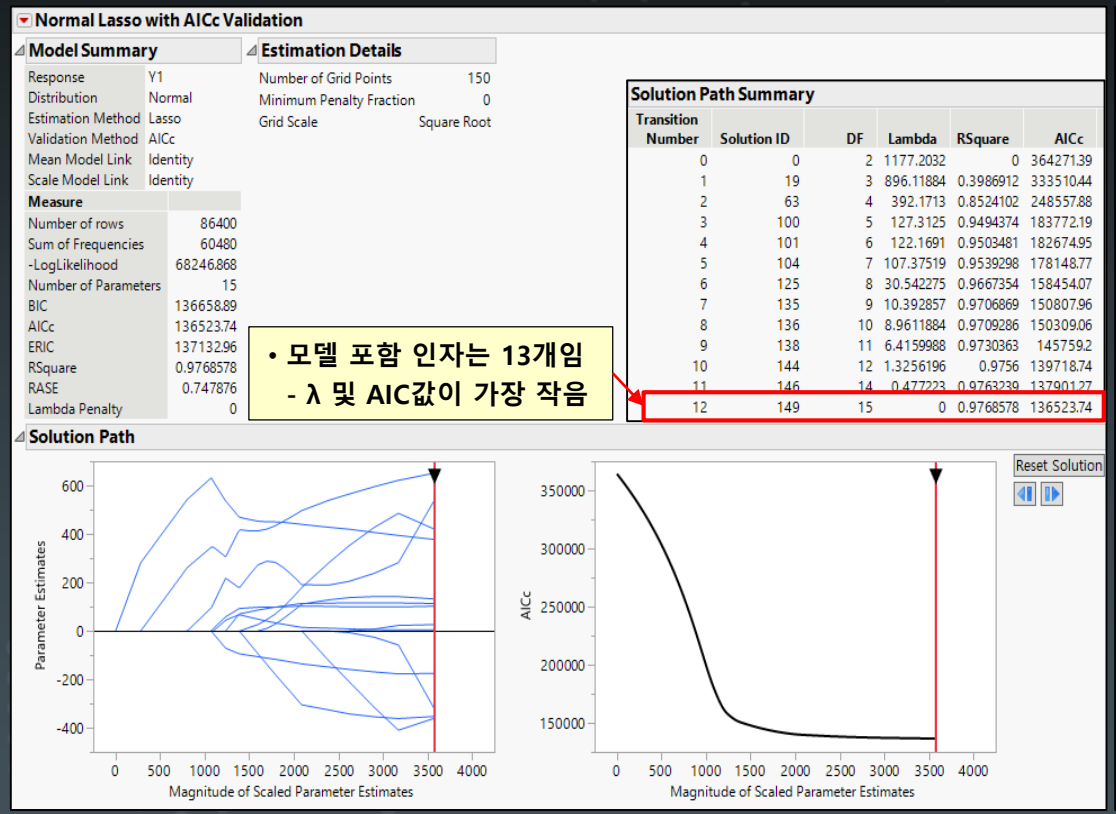


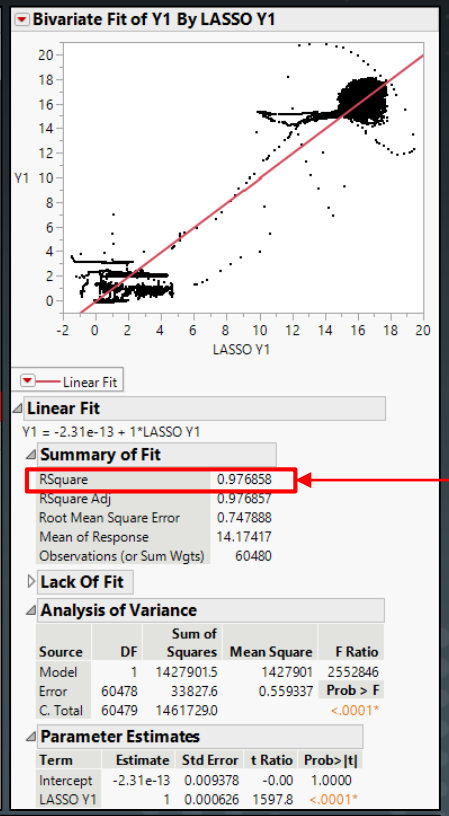
- **PLS 모델링으로 예측한 실제값과 예측값은 R-sq 97.69% 수준임.**

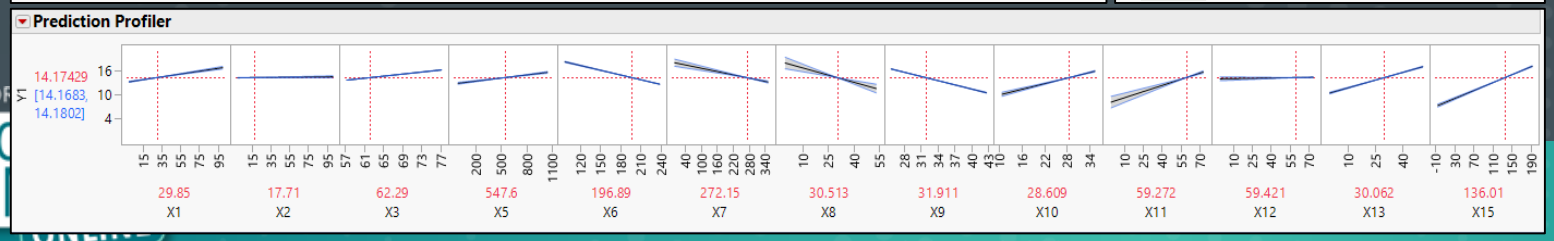
Transform Phase - ⑤ Modeling (Train Set)

▣ LASSO (Least Absolute Shrinkage Selector Operator) 회귀 모델링

- 많의 변수의 경우, 오차의 제곱 대신 절대값을 이용하여 불필요한 변수를 제거하는 방식의 회귀분석
- LASSO 분석 결과 실제값과 예측값의 일치도는 약 97.69% 임.







- 많은 변수에 대한 회귀분석시 오차의 제곱합이 과장되게 커지는 것을 억제하기 위하여 절대값을 이용한 패널티를 부여하는 방식

$$\min_{\beta}\left\{\sum_{i=1}^{I}\left(y_i-\sum_{j=1}^{J}\beta_j x_{i,j}\right)^2+\lambda\sum_{j=1}^{J}|\beta_j|\right\},$$

- 중요하지 않은 변수의 회귀 계수를 0으로 수축시킴으로써 차원을 축소 시킴.

- **LASSO 모델링으로 예측한 실제값과 예측값은 R-sq 97.69% 수준임.**

Transform Phase - ⑤ Modeling (Train Set)

▣ ANN (Artificial Neural Network, 인공신경망) 모델링

- 머신러닝 기법의 하나로 인체의 신경구조와 유사한 형태의 분석 알고리즘을 이용한 모델링 예측 기법임.
- ANN 분석 결과 실제값과 예측값의 일치도는 약 99.75% 임.

Neural

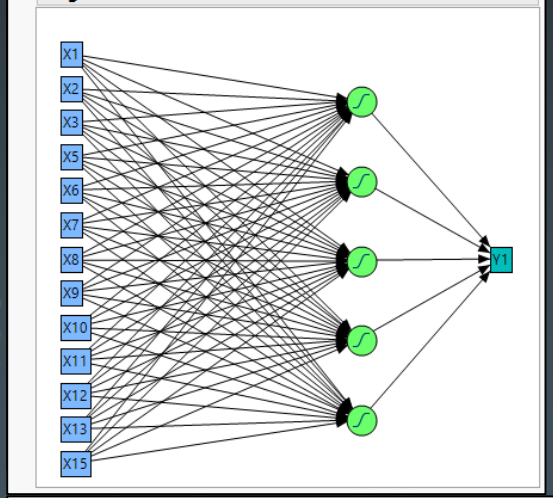
Validation: Random Holdback

Model Launch

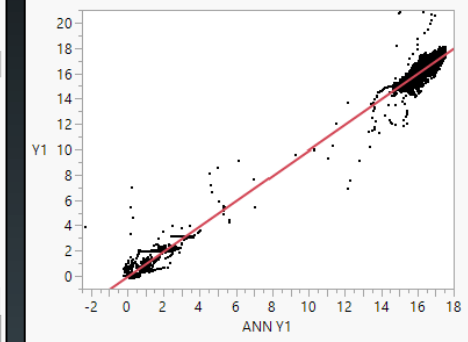
Model NTanH(5)

Training Y1		Validation Y1	
Measures	Value	Measures	Value
RSquare	0.9975948	RSquare	0.9973253
RMSE	0.2410569	RMSE	0.2543475
Mean Abs Dev	0.1631608	Mean Abs Dev	0.1654111
-LogLikelihood	-152.5682	-LogLikelihood	1005.5746
SSE	2343.048	SSE	1304.0743
Sum Freq	40322	Sum Freq	20158

Diagram



Bivariate Fit of Y1 By ANN Y1



Linear Fit

Linear Fit

Y1 = 0.0039905 + 0.999662*ANN Y1

Summary of Fit

RSquare	0.997505
RSquare Adj	0.997505
Root Mean Square Error	0.245564
Mean of Response	14.17417
Observations (or Sum Wgts)	60480

Lack Of Fit

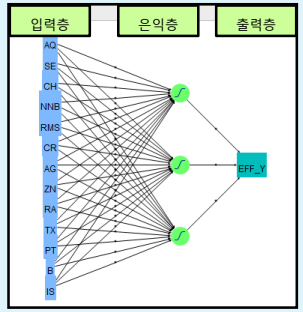
Analysis of Variance

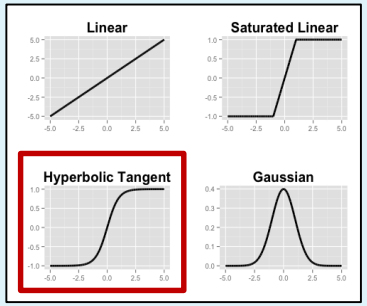
Source	DF	Sum of Squares	Mean Square	F Ratio
Model	1	1458082.1	1458082	24179847
Error	60478	3646.9	0.060302	Prob > F
C. Total	60479	1461729.0		<.0001*

Parameter Estimates

Term	Estimate	Std Error	t Ratio	Prob>|t|
Intercept	0.0039905	0.00305	1.31	0.1907
ANN Y1	0.999662	0.000203	4917.3	<.0001*

- 사람의 신경망인 Neuron의 Network 구조와 유사한 형태의 분석 방법

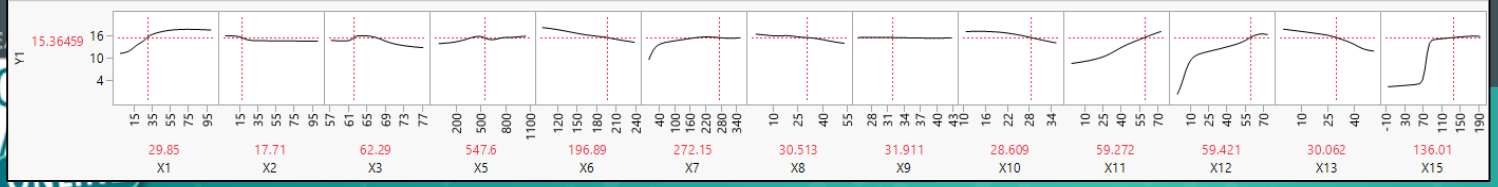




- 주로 은익층 Node의 수를 변화시키면서 모델의 정확성을 개선시키는 방향으로 분석함.

- **ANN 모델링으로 예측한 실제값과 예측값은 R-sq 99.75% 수준임.**

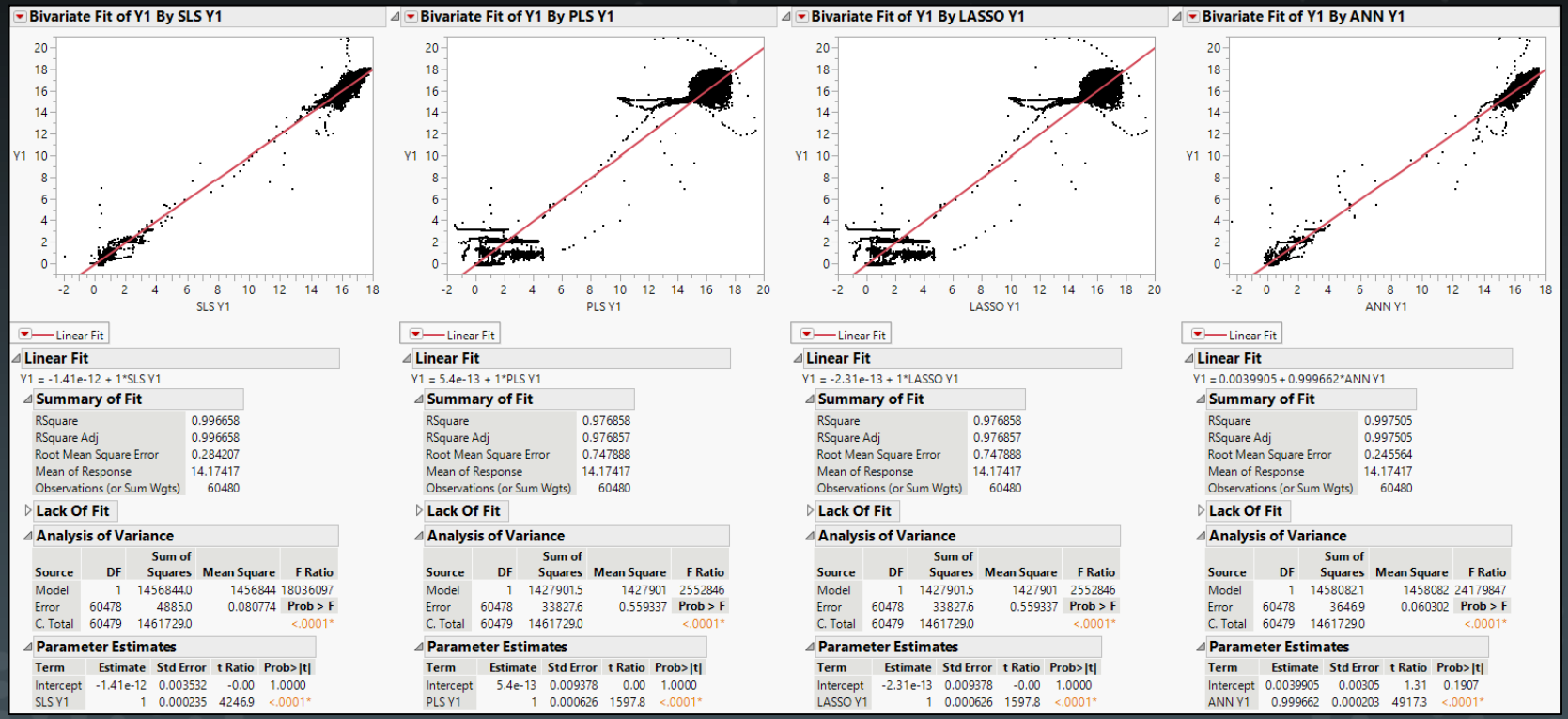
Prediction Profiler



Transform Phase - ⑤ Modeling (Train Set)

▣ Train Set의 일치도 결과 종합

- Train Set으로 분석한 4종류의 모델링 분석 결과에 대한 일치도 종합 결과

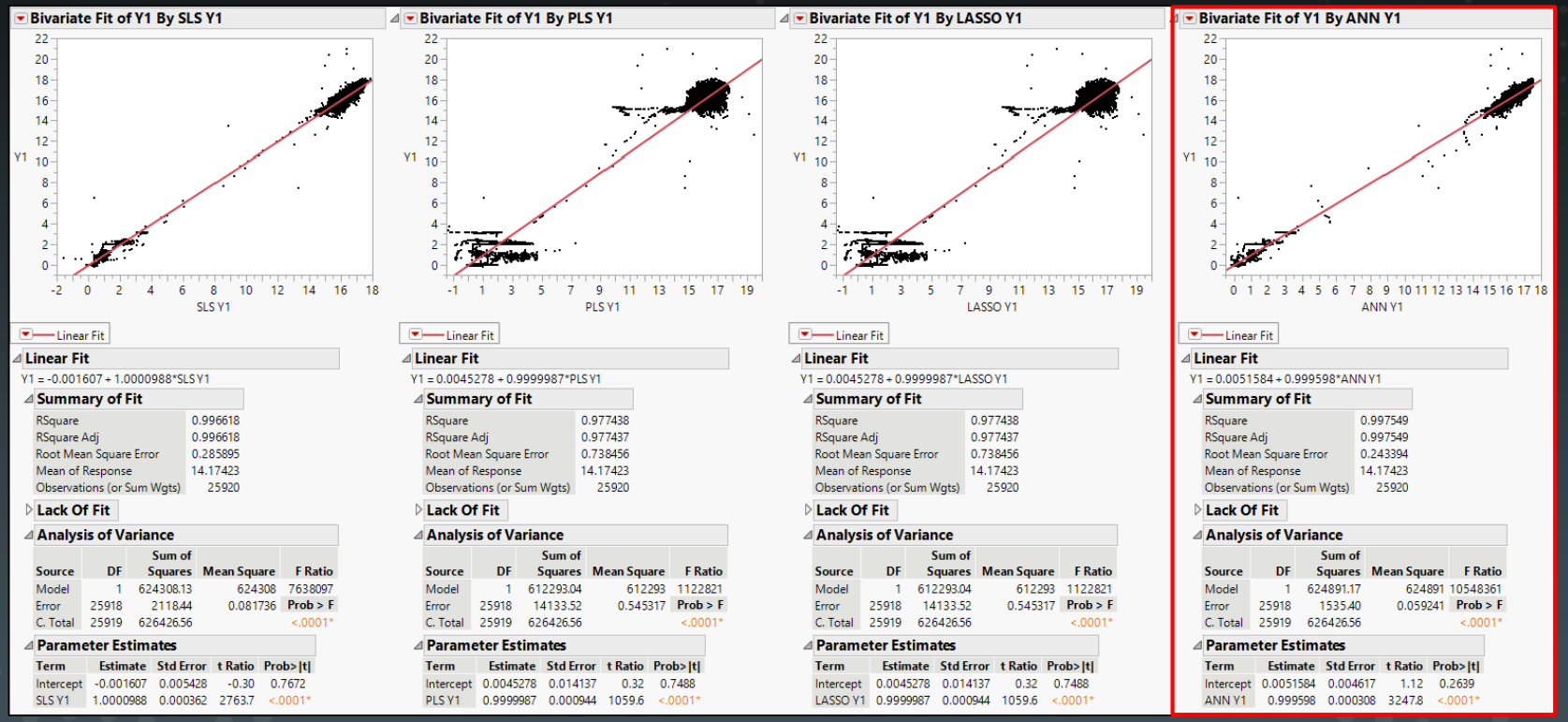


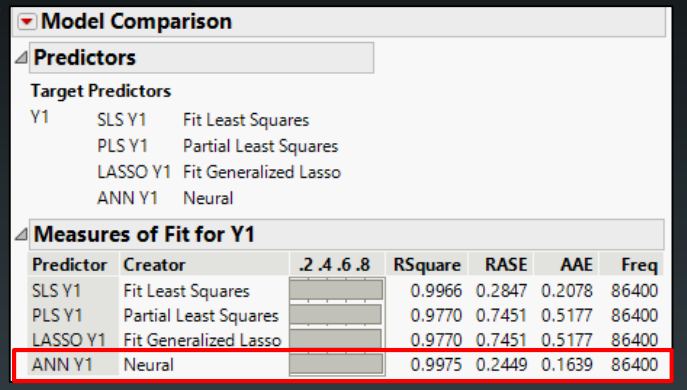
Modeling	Train Set		Valid Set		Conclusion
	P-Value	R Square	P-Value	R-Square	
SLS	< 0.0001	0.9967			
PLS	< 0.0001	0.9769			
LASSO	< 0.0001	0.9769			
ANN	< 0.0001	0.9975			

Transform Phase - ⑥ Modeling (Valid Set)

▣ Valid Set의 일치도 분석 결과 종합

- Valid Set으로 분석한 4종류의 모델링 분석 결과에 대한 일치도 종합 결과





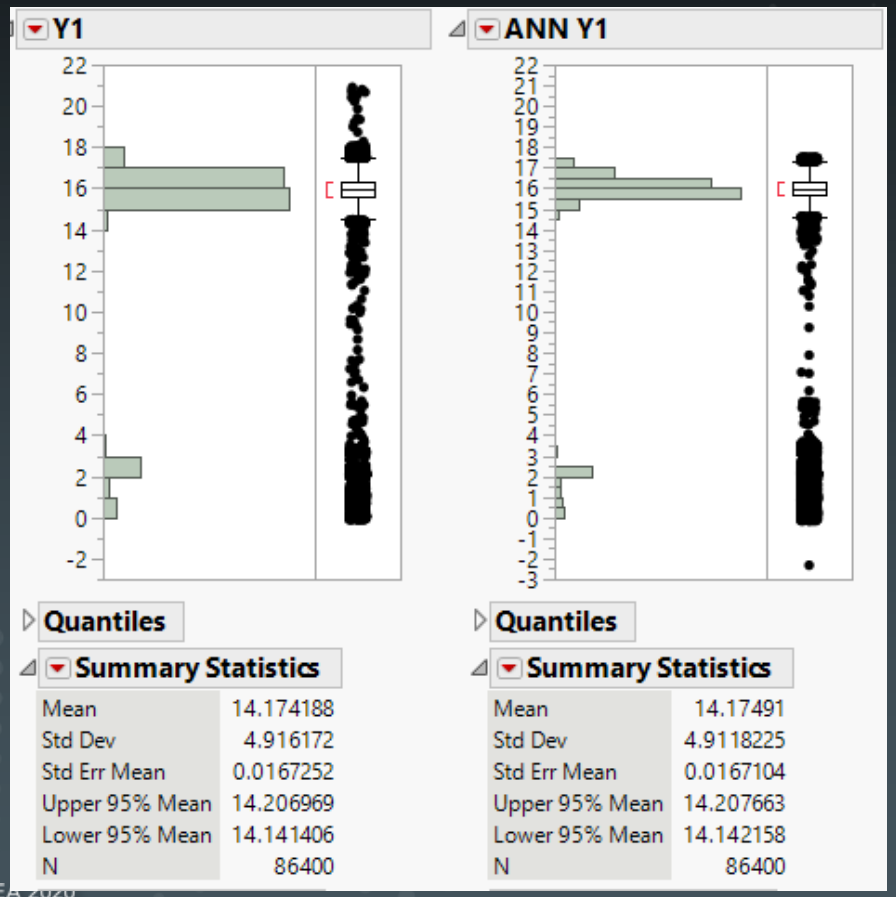
- 결론적으로 머신러닝 모델인 ANN이 현재 Data 에서 가장 예측력이 높은 것으로 분석됨.
- 차선으로 SLS 모델도 매우 양호한 결과를 보임

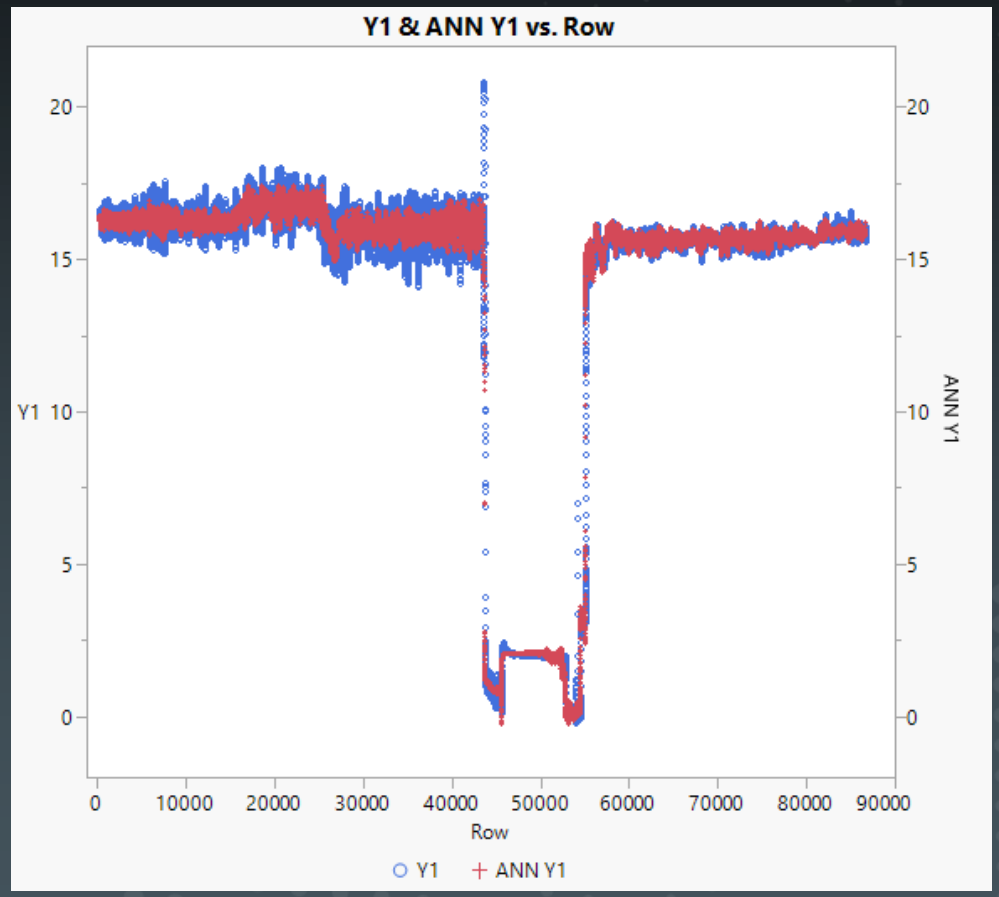
Modeling	Train Set		Valid Set		Conclusion
	P-Value	R Square	P-Value	R-Square	
SLS	< 0.0001	0.9967	< 0.0001	0.9967	
PLS	< 0.0001	0.9769	< 0.0001	0.9774	
LASSO	< 0.0001	0.9769	< 0.0001	0.9774	
ANN	< 0.0001	0.9975	< 0.0001	0.9975	적합모델 선정

Transform Phase - ⑥ Modeling (Valid Set)

▣ Y1과 ANN Y1의 그래프상의 일치도 확인

- 실제 Y1과 ANN에 의한 예측된 Y1의 그래프 및 기술통계량 일치도 비교
- 비교 결과 그래프 및 분포와 기술 통계량에서 거의 유사한 값을 보임

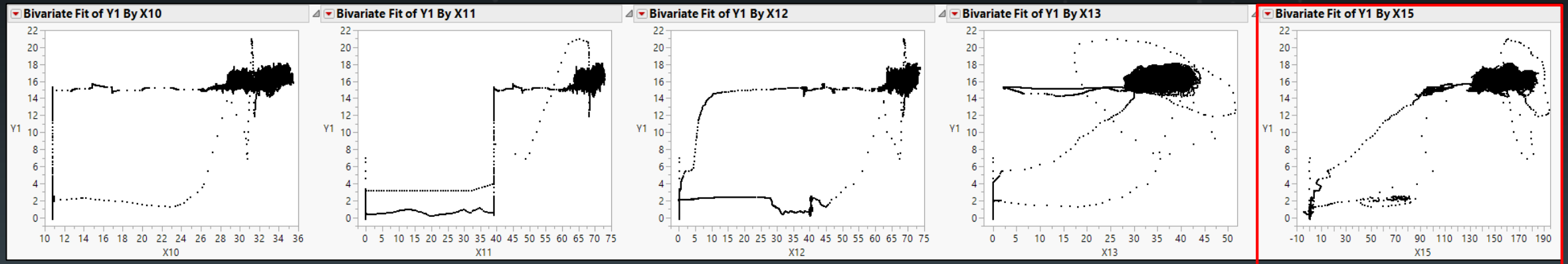




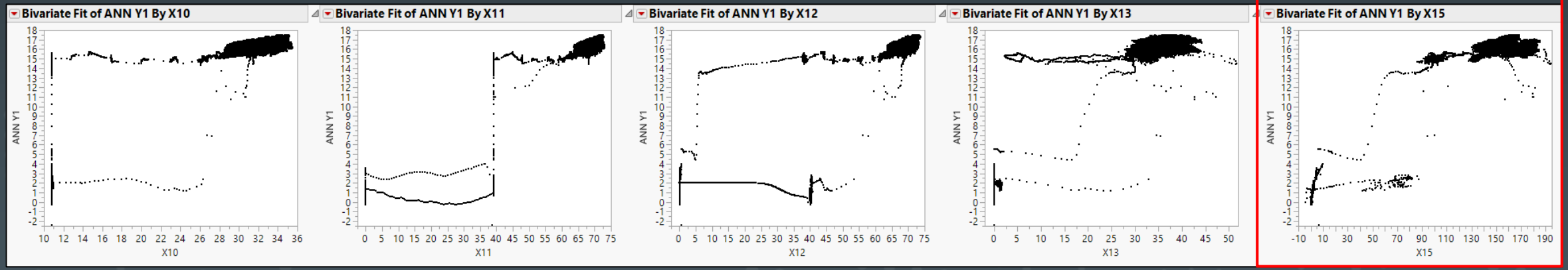
Transform Phase - ⑥ Modeling (Valid Set)

▣ Y1과 예측된 Y1(ANN)에 의한 실제값 Scatter Plot 비교

- Y1과 각 X인자들과의 산점도 비교
- 실제 Y1과 각 X인자들의 산점도



- ANN에 의해 예측된 Y1과 각 X인자들의 산점도



KOREA 2020

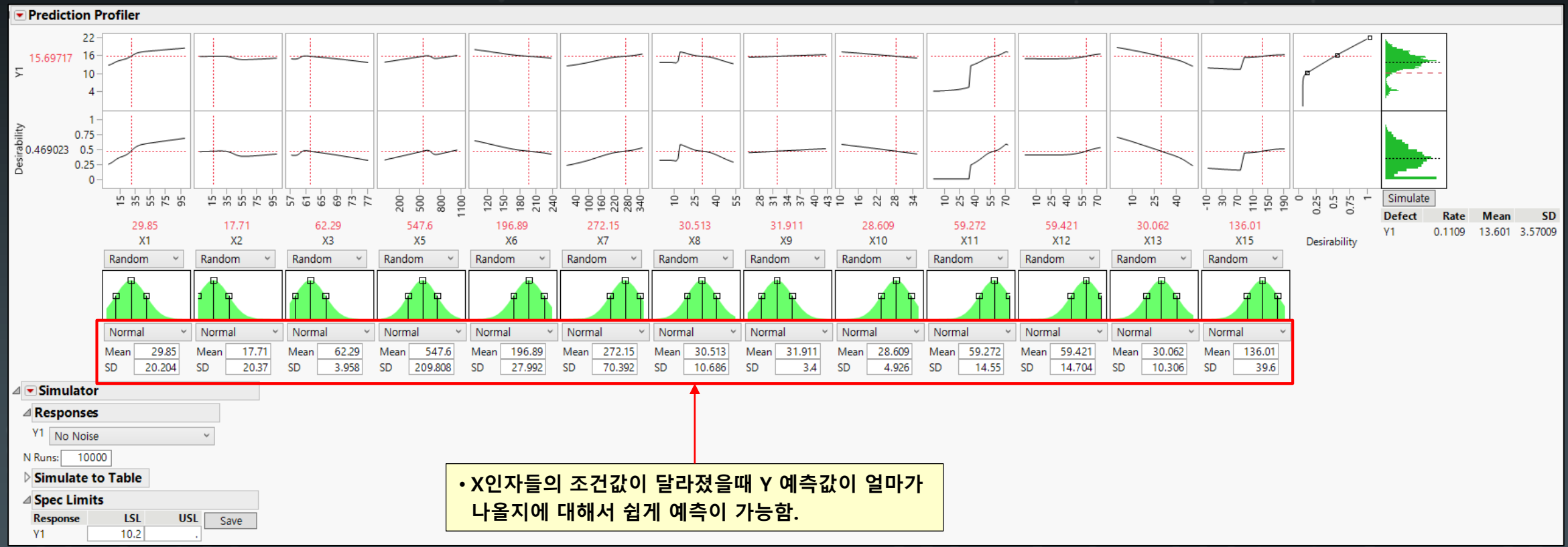
DISCOVERY SUMMIT ONLINE

- 실제 현장의 원인 조자 결과 X15에 의한 문제 발생으로 확인되어 예측한 인자들의 결과와 매칭됨.

Apply Phase - ⑦ Optimization

▣ ANN 결과에 대한 Profiler 예측 (Total Set)

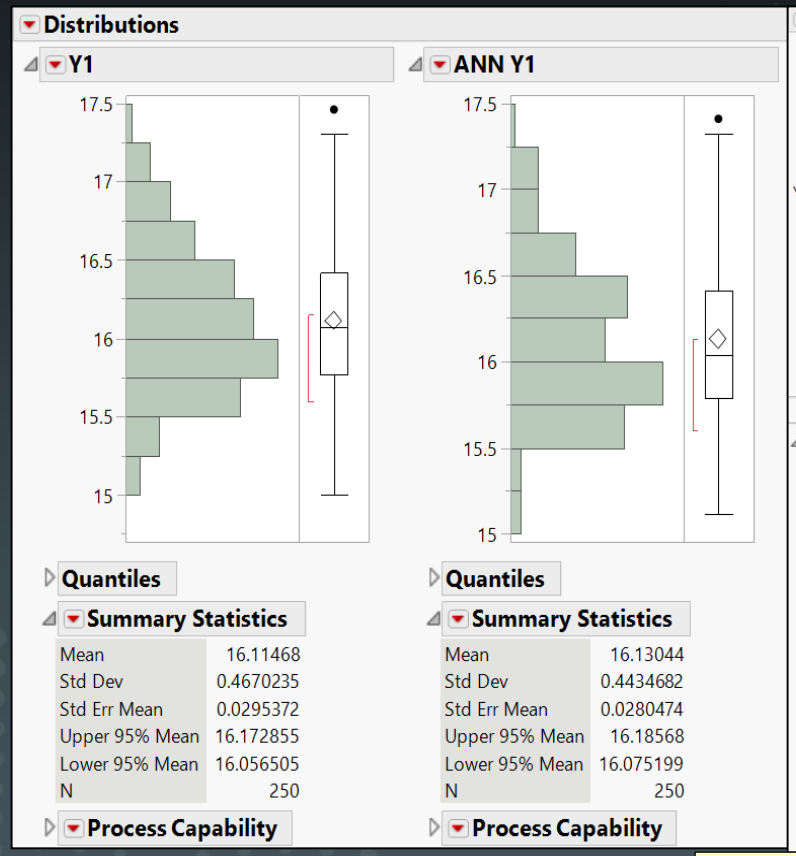
- **Profiler 예측 Tool을 이용하여 각 X인자들의 값의 변화에 따라 Y의 변화에 대해서 감지가 가능**하며,
- 만일 공정조건의 변경이 필요시 가장 안정된 조건에 대해서 산출 가능. (X인자들의 가용범위, 제약조건에 대한 정보 반영 필요)

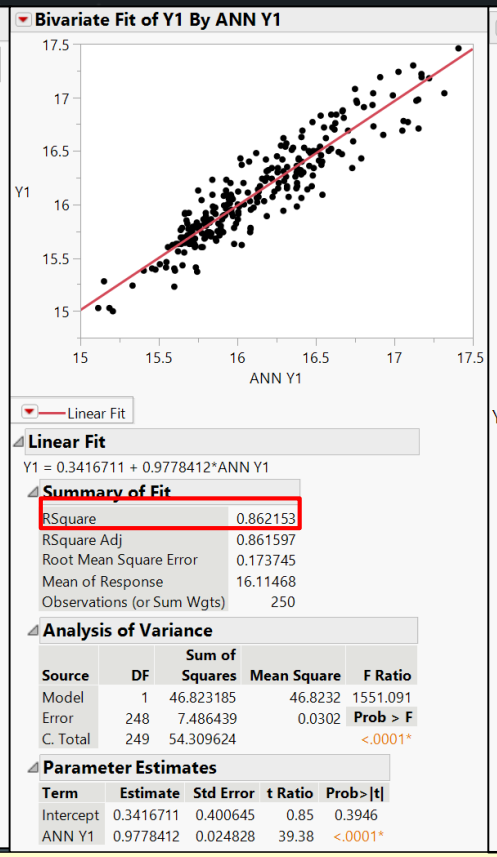


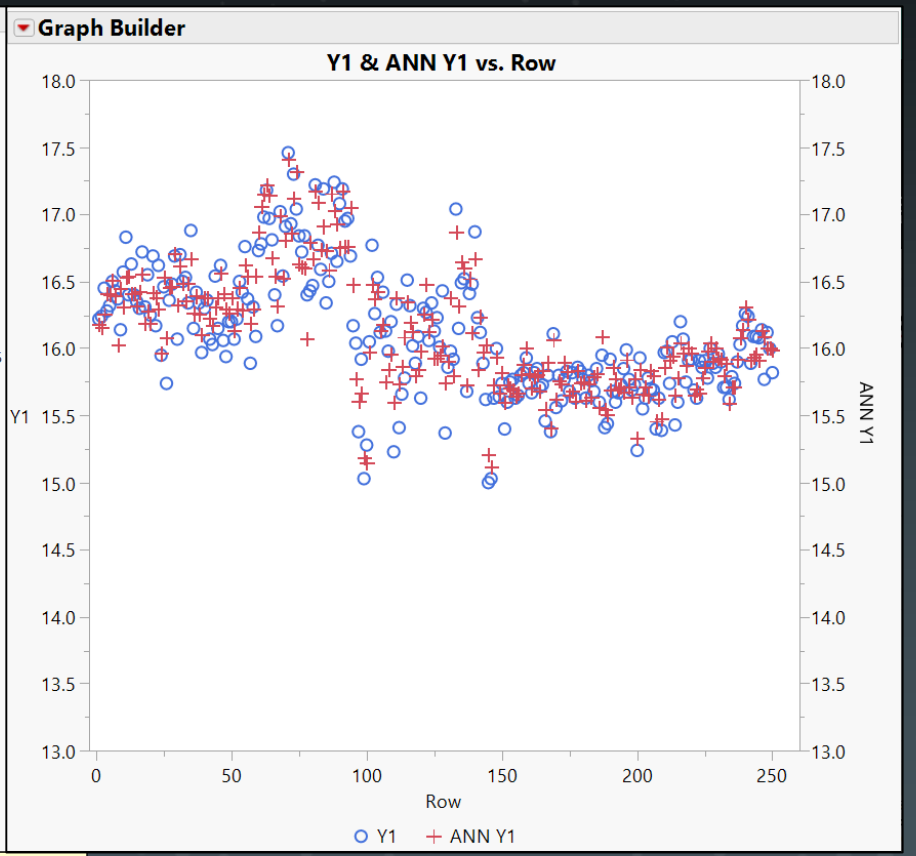
Apply Phase - ⑧ Feedback (Test Set)

▣ Test Set 을 통한 Final Confirm

- 최종 예측값의 확인을 위하여 이후 안정된 구간에 진행된 250개의 Test Set을 준비하여 확인함.
- 만일 공정 조건 변경이 발생했을 경우, 변경된 조건으로 Test Set을 준비한다.







- **250개의 Test Set으로 최종 예측 Test를 진행결과 양호한 결과를 얻음**

Apply Phase - ⑧ Feedback (Test Set)

▣ Vital X인자들을 이용한 예측 관리 시스템 구축 (Dashboard)

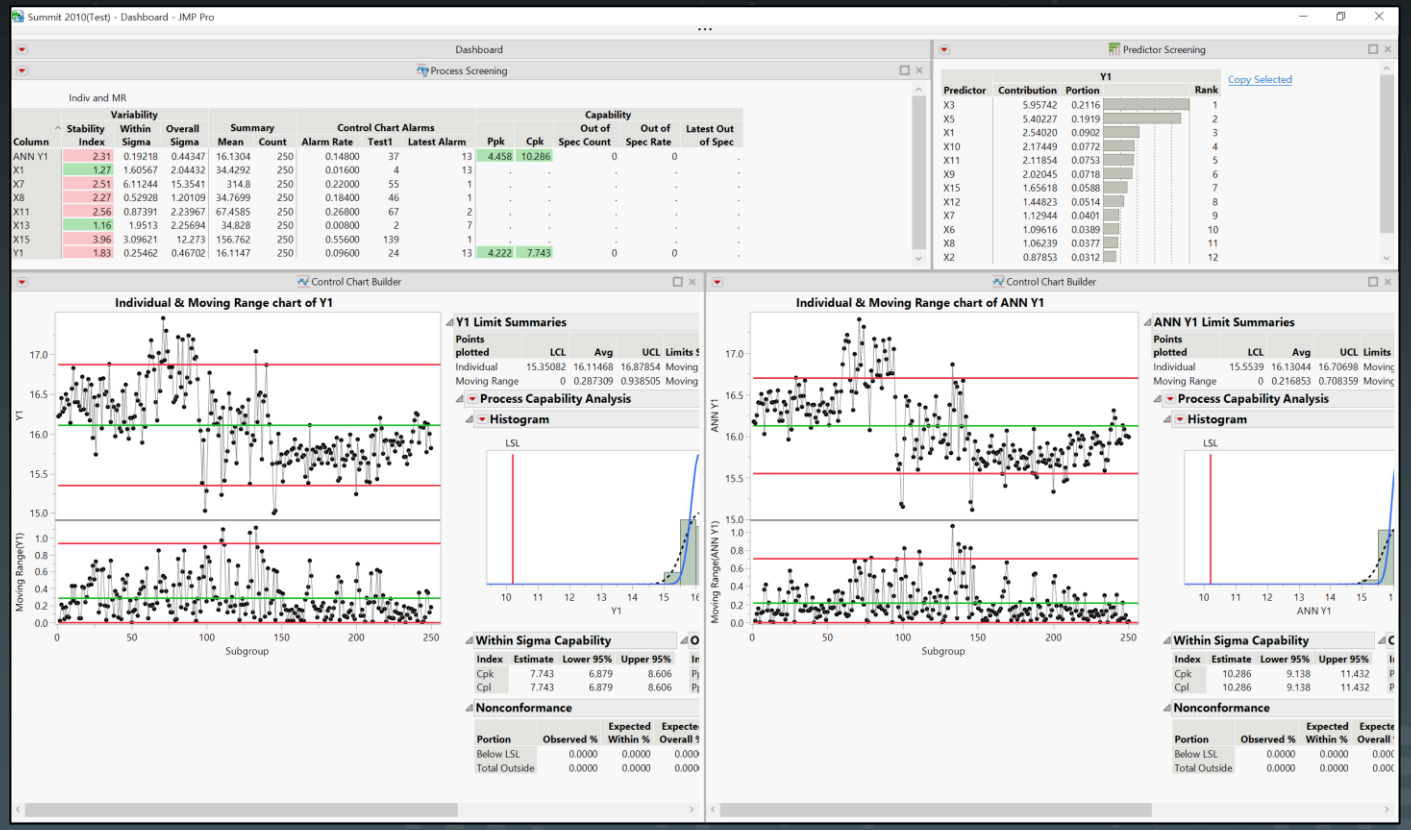
- 금번 분석에서 예측된 Y의 알고리즘을 바탕으로 Vital X 인자들에 대하여 SPC 관리를 통한 사전 예측이 가능함.
- **JMP의 Dashboard 기능을 통해 필요한 항목을 등록하여 관리 시스템과 연계** 시킴.

Predictor Screening

Copy Selected

Predictor	Y1 Contribution	Y1 Portion	Y1 Rank
X15	320636	0.2441	1
X13	228575	0.1740	2
X7	222581	0.1694	3
X11	177444	0.1351	4
X1	106247	0.0809	5
X8	76578	0.0583	6
X10	74981	0.0571	7
X6	47764	0.0364	8
X12	39312	0.0299	9
X5	12851	0.0098	10
X3	3073	0.0023	11
X2	2512	0.0019	12
X9	1004	0.0008	13

- 최초 누적 80% 수준의 Vital X인자들에 대하여 Dashboard 기능을 통한 SPC 관리



- Dashboard에 관리 필요한 항목을 등록하여 관리 함.
  - Process Screening, Predict Screening
  - Y1, ANN Y1, 중요 X인자의 SPC 등