Project I : Wind Chill Analysis

2014122006 허승민

2019년 4월 18일

1 개요

주어진 데이터는 바람의 속도, 온도에 의해서 체온의 변화를 나타내는 데이터입니다. 데이터를 활용하여 체온과 바람의 속도, 온도와의 관계를 파악한다면 이와 관련된 재난에 빠진 인원을 구조하는데 도움을 주는 등의 효용이 예상됩니다.

2 데이터 진단

기본적인 변수 간의 관계를 파악하기 위해 산점도, 상관관계 파악을 해보았습니다.(Appendix 표1) 먼저, 설명변수인 바람의 속도와 온도는 경우의 수를 모두 따져 체온 변화를 파악하기 위한 목적이므로 균일하게 분포하도록 의도된 데이터로 보여집니다. 때문에 상관계수는 존재하지 않습니다. 종속변수와 설명변수의 관계를 확인해보면 체온 변화는 바람의 속도가 빨라지면 내려가는 관계이며 온도는 경우 온도의 증감 방향과 체온의 증감방향이 일치한다는 것을 확인할 수 있습니다. 종속변수와 설명변수의 상관관계가 존재하는 것으로 보여 간단한 선형회귀 모델로 설명해보겠습니다.

3 모델링

3.1 1차 모델링

두개의 설명변수만을 고려하여 기본적인 선형회귀 모형을 적용하면 통계적인 수치적 근거는 좋은 모델이라고 판단할 수 있습니다. 하지만 세 변수간의 관계와 모형을 그림으로 파악해보면 비슷하지 않은 것을 확인할 수 있습니다. 또한 예측한 결과의 오차가 일정하지 않고 변동이 큽니다.(Appendix 2) 따라서 간단한 선형관계는 좋은 모형이 아니라는 것을 확인할 수 있습니다. 조금 더 복잡한 모형으로 설명해야할 필요가 있어보입니다.

3.2 모델 수정

모델의 개선방향은 설명변수와 관련하여 두가지 방향으로 생각할 수 있습니다.(Appendix 3) 설명변수는 진단 단계에서 말씀드렸듯이 의도적으로 조정한 데이터로 보이기 때문에 두 변수간의 관련성이 없다고 말하기는 어렵습니다. 상식적으로 기온이 낮을 때 바람의 속도가 빠르면 더 춥게 느끼게 됩니다. 두 변수 간에 상호작용 관계가 있을 것으로 보이며 이를 반영합니다. 또한 바람의 속도와 체온변화의 관계를 보면 선형이라기 보다는 조금 아래로 굽어있는 형태라는 것을 볼 수 있습니다.

따라서 두 변수의 관계를 선형으로 만들도록 펴주는 변환을 고려해봐야합니다. 정리하면 다음과 같습니다.

- 설명변수간의 상호작용을 반영해준다.
- 설명변수가 더 쉽게 설명하도록 변수 모양을 바꿔준다.

3.3 2차 모델링

수정된 모형은 모양을 바꾼 변수를 추가하고 상호작용을 반영해주는 방향으로 구성되었습니다. 결과적으로, 상호작용을 반영한 변수는 매우 좋은 결과를 내었고, 모양을 바꾼 변수는 좋지 않은 영향을 미쳐 제거되었습니다. 따라서 기존 두 변수에 상호작용을 더한 결과가 가장 좋은 결과를 내었고 균일한 오차를 가진 예측을 하는 건강한 모형이라는 것을 확인하였습니다.(Appendix 4)

4 모형 분석

지금까지 만들어낸 모형의 오차는 대부분 화씨 10도를 넘어서는 경우가 많지 않습니다. 섭씨 기준 5도 정도 차이를 보이는 것으로 체온의 변동폭을 생각하면 크지 않은 정도입니다. 따라서 설명력이 충분하고, 통계적인 근거 또한 충분합니다. 데이터 진단을 통해 확인한 기본적인 상관관계는 같지만 상호작용을 추가하였기 때문에 각각의 변수의 영향력을 명확하게 설명하기는 어렵습니다. 조금 다른 해석도 가능해집니다. 기온이 어떤 값 이상(약 4도)이라면 바람의 속도가 증가하면 체온이 기존의 관계처럼 감소하는 것이 아니라 증가하는 경향이 있을 수 있습니다. 이러한 이유때문에 만들어진 그림의 모형을 나타내는 평면이 휘어있고 원래 데이터에 보다 더 잘 들어맞는 것을 볼 수 있습니다.(Appendix 5)

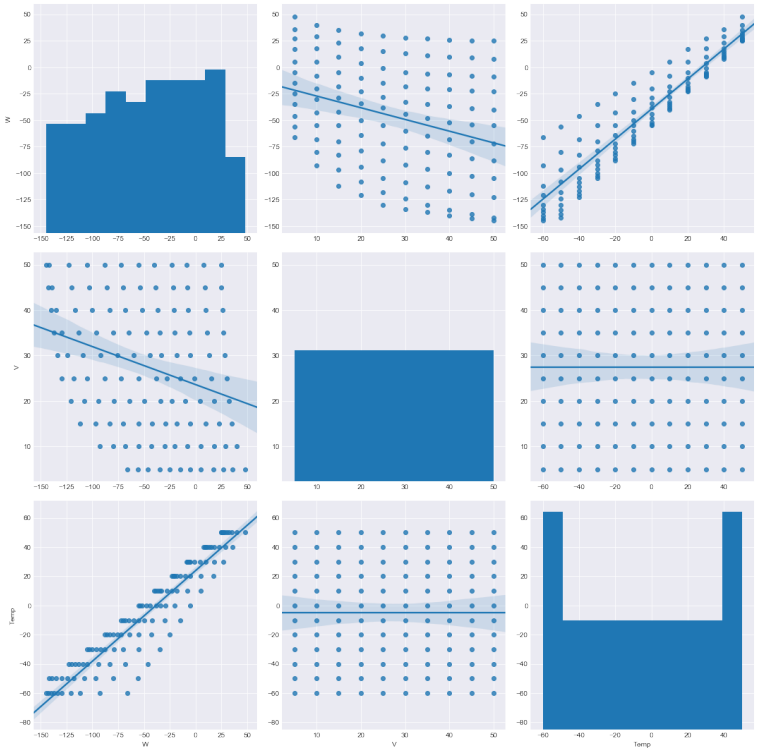
5 결론

지금까지 바람의 속도, 기온과 체온의 변화의 관계를 데이터를 통해 확인해보았습니다. 최종적으로 획득한 식은 다음과 같습니다.

$$W = -10.6967 - 1.0458V + 0.6111Temp + 0.2588logV * Temp$$

결론적으로 바람의 속도와 기온은 체온의 변화와 각각 반비례, 정비례의 관계를 갖는다고 생각할 수 있지만 기온이 일정 온도(화씨 4도) 이상으로 올라가게 되면 바람의 속도와 체온의 변화가 정비례의 관계로 바뀔 수 있다는 사실을 확인했습니다. 앞서 말했듯, 이러한 관계를 밝힘으로서 재난상황에 빠진 사람의 저체온 가능성, 날씨 조건에 따른 맞춤 옷 추천 등 활용할 가치가 높은 모형이라고 생각할 수 있습니다.

Appendix

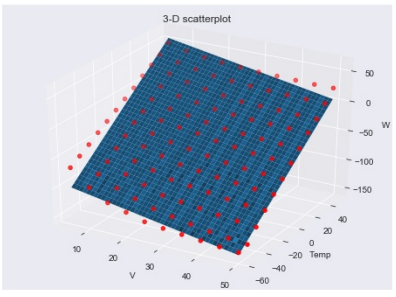


	W	V	Temp
W	1.00000	-0.30417	0.93829
V	-0.30417	1.00000	0.00000
Temp	0.93829	0.00000	1.00000

표 1: Scatter plot Matrix 와 상관계수. Temp와 V 변수는 상관관계가 없는 변수이며, Temp 변수는 종속변수와 선형적인 관계가 있는 것으로 보인다. 하지만 V 변수의 경우는 약간의 곡선형태가 보이므로 추가적인 변수변환이 필요할 것으로 생각된다.

Dep. Variable:	W	**R-squared:**	0.973
Model:	OLS	**Adj. R-squared:**	0.972
Method:	Least Squares	**F-statistic:**	2101.
Date:	Mon, 15 Apr 2019	**Prob (F-statistic):**	2.10e-92
Time:	18:20:10	**Log-Likelihood:**	-428.36
No. Observations:	120	**AIC:**	862.7
Df Residuals:	117	**BIC:**	871.1
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-9.0566	1.720	-5.267	0.000	-12.462	-5.651
V	-1.1055	0.055	-19.989	0.000	-1.215	-0.996
Temp	1.4187	0.023	61.661	0.000	1.373	1.464



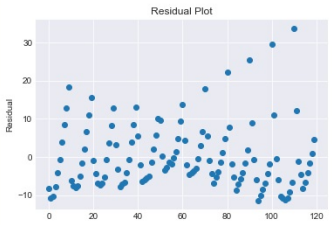


표 2: 모형 1의 결과. 통계적으로 R^2와 p-value 등이 좋은 모델이라는 근거를 주지만 그림으로는 그렇지 않다. 모형과 실제 산점도가 차이를 보이고 분산이 나비넥타이 모양의 이분산 형태를 보이므로 좋은 모형이 아니다.

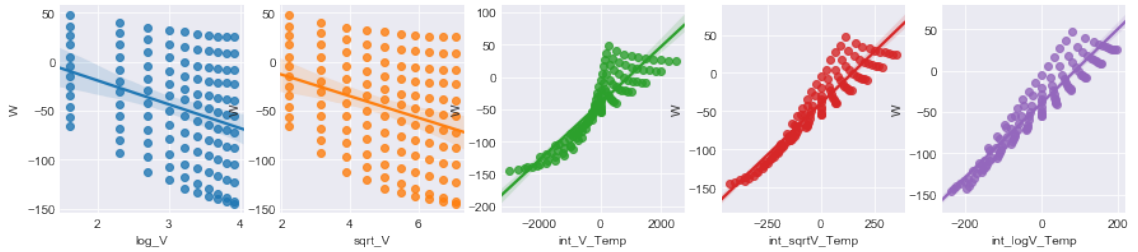
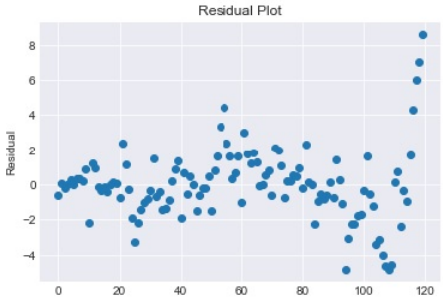


표 3: 새로운 변수를 찾기 위한 산점도. V 변수의 경우 선형성을 더 강화하기 위해 변수 변환 시 log 변환이 좀더 선형성에 가까운 것으로 보인다. 상호작용을 위해 V 변수의 변환들과 Temp 변수 사이의 관계를 관찰. log 변환한 V 변수와 Temp의 상호작용 변수는 선형성이 짙은 것으로 보인다. 결과적으로, logV 와 Temp의 상호작용, logV 변수 두가지를 추가하여 새로운 모형을 만든다.

Dep. Variable:	W	**R-squared:**	0.998
Model:	OLS	**Adj. R-squared:**	0.998
Method:	Least Squares	**F-statistic:**	1.878e+04
Date:	Tue, 16 Apr 2019	**Prob (F-statistic):**	7.36e-161
Time:	00:25:19	**Log-Likelihood:**	-255.87
No. Observations:	120	**AIC:**	521.7
Df Residuals:	115	**BIC:**	535.7
Df Model:	4		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	37.6741	1.706	22.086	0.000	34.295	41.053
V	0.1533	0.043	3.554	0.001	0.068	0.239
Temp	0.6433	0.025	25.386	0.000	0.593	0.693
log_V	-26.0739	0.892	-29.236	0.000	-27.840	-24.307
int_logV_Temp	0.2485	0.008	31.352	0.000	0.233	0.264



Dep. Variable:	W	**R-squared:**	0.987
Model:	OLS	**Adj. R-squared:**	0.987
Method:	Least Squares	**F-statistic:**	2961.
Date:	Mon, 15 Apr 2019	**Prob (F-statistic):**	2.15e-109
Time:	19:50:30	**Log-Likelihood:**	-383.80
No. Observations:	120	**AIC:**	775.6
Df Residuals:	116	**BIC:**	786.7
Df Model:	3		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	-10.6967	1.200	-8.913	0.000	-13.074	-8.320
V	-1.0458	0.039	-27.041	0.000	-1.122	-0.969
Temp	0.6111	0.073	8.349	0.000	0.466	0.756
int_logV_Temp	0.2588	0.023	11.304	0.000	0.213	0.304

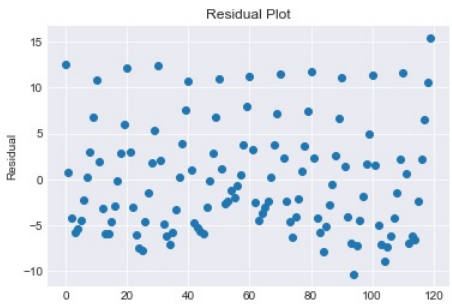


표 4: 수정한 모형들의 결과. 앞서 산점도를 통해 추가하기로 결정한 두 변수를 추가해본 결과가 좌측이다. 좌측의 경우 통계적인 분석결과는 매우 좋으나 잔차도를 그려보면 잔차가 일정하지 않으므로 가정에 위배된다. logV 변수의 변화가 매우 크기때문이라고 판단하여 변수를 제거하였고 우측의 결과를 얻었다. 통계적인 결과도 매우 우수하며 분산도 고른 형태를 보인다. 따라서 우측모형을 채택한다.

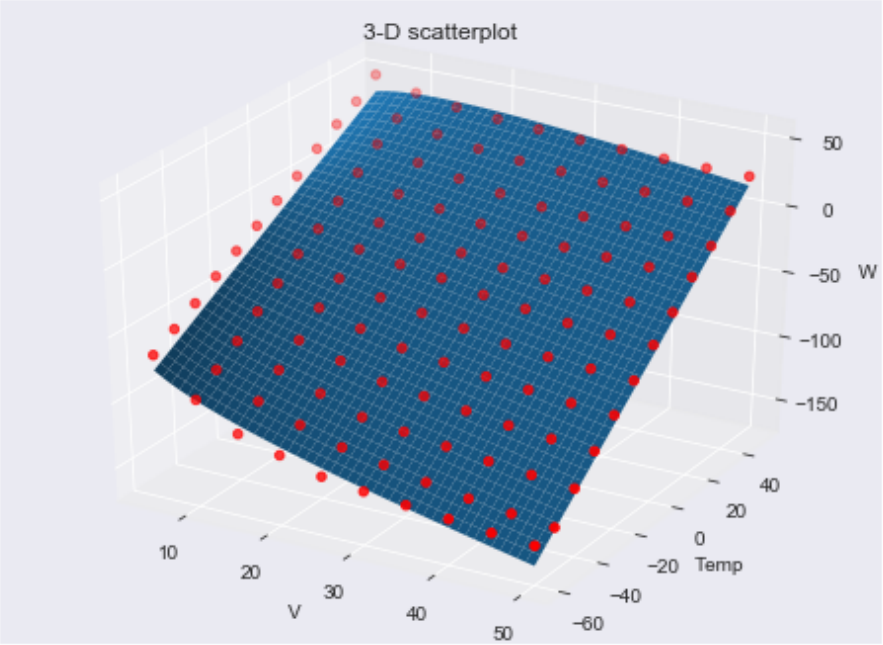


표 5: 최종적으로 얻은 모형과 주어진 변수의 산점도. 앞서 관찰했을 때 잘 맞지 않았던 모습과는 달리 평면이 휘어지면서 산점도와 비슷해 진 것을 확인할 수 있다. 상호작용과 변수변환을 통하여 만들어낸 모형의 모습이다.