5.3 독립변수가 여럿인 모델

5.3.1 분석준비

```
# 수치 계산에 사용하는 라이브러리

import numpy as np
import pandas as pd
import scipy as sp
from scipy import stats

# 그래프를 그리기 위한 라이브러리
from matplotlib import pyplot as plt
import seaborn as sns
sns.set()

# 선형모델을 추정하는 라이브러리 (경고가 나올 수도)
import statsmodels. formula.api as smf
import statsmodels.api as sm
# 표시 자릿수 지정
%precision 3

#그래프를 주피터 노트북에 그리기 위한 설정
%matplotlib inline
```

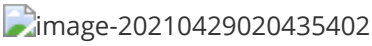
어느 가게의 가상의 매상 데이터를 읽어 들인다.

sales라는 변수에 저장

```
sales = pd.read_csv("5-3-1-lm-model.csv")
print(sales.head(3))
```

5.3.2 데이터로 그래프 그리기

```
sns.pairplot(data = sales, hue = "weather", palette = "gray")
```



왜 모양이 약간 다를까?

- 3행에 있는 그래프들이 Y축이 매상인 그래프이다.

그러나 이 그래프만 보면 판별하기 조금 힘들다. 이처럼 그래프만으로 바로 판단이 서지 않는 데이터들도 많이 존재한다.

- 한눈에 알 수 있는 것은 습도와 기온의 관계

기온이 높아지면 습도도 높아지는 관계를 예측할 수 있다.

5.3.3 나쁜 분석의 예 : 변수가 1개인 모델 만들기

- 여러 개의 독립변수가 필요한데, 1개만 사용한 모델을 작성
- 사용하는 소프트웨어 등의 제약으로 인해 단순회귀모델만 추정 가능

이러한 소프트웨어를 사용해서 억지로 분석을 진행하면 잘못된 고찰로 이끌리게 됨

독립변수에 가격만 사용하고 단순회귀모델 추정

가격의 계수가 양수가 됨

```
Im_dame = smf.ols("sales ~ price", sales).fit()
Im_dame.params
```

분산분석을 사용해서 검정하면 p값이 0.05보다 작게 나옴

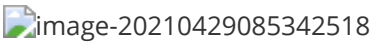
```
print(sm.stats.anova_lm(Im_dame, typ = 2))
```

위의 결과를 정리하면

- 가격은 매상에 대해 유의미한 영향이 있다.
- 가격이 오르면 매상도 증가한다

매상을 늘리고 싶다면 단순히 가격을 올리면 된다는 생각을 하게 됨

해당되는 회귀직선



- 우상향 직선이 그려져서 가격을 올리면 매상이 늘어날 것이라고 고찰하게 됨

5.3.4 독립변수 간의 관계 조사하기

위의 분석에서의 문제는 가게의 사정을 무시하고 **가격과 매장의 관계만 도출**했다는 것이다.

가게에서는 매일 가격을 어떻게 바꾸고 있을까?

```
print(sales.groupby("weather").mean())
```

출력 결과를 보면,

- 비오는 날에는 매상(sales)이 떨어졌다.
- 게다가 낮은 가격으로 억제되고 있다.

이는 비오는 날에는 매상이 떨어지기 때문에 그 대책으로 가격을 인하했다고 이해하면 된다.

그러나 가격을 인하한 날에도 매상이 떨어지는 것 처럼 보인다.

날씨가 같았을 때, 상품 가격이 매상에 미치는 영향에 대해 그래프로 살펴보자

맑은 날 쪽이 높은 매상을 보여주고 있다.

날씨로 보면 가격이 녹아지면 매상은 줄어든다는 것을 알 수 있다.

방금 전과 비교해서 완전히 반대의 결론이 되었다.

- 원래는 독립변수 간, 아무런 관련이 없는 데이터를 사용하는 것이 좋다.
- 하지만 분석을 간단하게 하기 위해 가게가 가격을 바꾼다고 가정하는 것은 현실적이지 않다.
- 이러한 데이터에 일반적인 검정을 실시하는 것은 위험하다.

예전에는 날씨 데이터를 먼저 분할하고 이에 대해서 회귀분석을 수행하는 방법을 사용한 적도 있었다.

- 맑은 날 끼리의 가격, 매상의 관계
- 비오는 날 끼리의 가격, 매상의 관계

그러나 이 방ㅂ버은 계수 검정을 할 때 2번의 검정을 반복하게 되어 **검정의 다중성 문제**가 발생한다.

각 요인의 영향을 올바르게 판단하기 위해서는 복수의 독립변수를 가지는 모델을 **한 번에 추정**해야한다.

5.3.5 복수의 독립변수를 가지는 모델

독립변수가 4개가 다 들어간 모델을 추정

```
Im_sales = smf.ols("sales ~ weather + humidity + temperature + price", data =
sales).fit()
Im_sales.params
```

가격 계수가 마이너스 - 가격이 오르면 매상이 떨어진다는 것을 알 수 있다.

5.3.6 나쁜 분석 예 : 일반적인 분산분석으로 검정하기

typ = 1파라미터를 넘기면 일반적인 분산분석 Type 1 ANOVA

```
print(sm.stats.anova_Im(Im_sales, typ =1).round(3))
```

이 검정 결과는 모든 독립변수가 유의미한 것 처럼 되어 있다.

그러나 이것은 틀렸다. 이것은 **독립변수를 넣는 순서를 바꾸면 검정 결과가 바뀐다.**

이를 확인하기 위해 **완전히 똑같은 4개의 독립변수**를 가지는 일반화 설명 모델을 하나 추정해보자

변수의 순서를 바꾸어서

```
Im_sales2 = smf.ols("sales ~ weather + temperature + humidity+ price", data =
sales).fit()
Im_sales.params

print(sm.stats.anova_Im(Im_sales2, typ =1).round(3))
```

독립변수의 순서를 바꾸었을 뿐이다.

그래서 추정된 계수의 값은 일치하지만 **검정 결과는 일치하지 않는다.**

습도의 p값이 0.6정도로 매상에 유의미한 영향을 끼치지 않는다는 결과가 나왔다.

5.3.7 회귀계수의 t검정

```
Im_sales.summary().table[1]

Im_sales2.summary().tables[1]
```

회귀계수의 t검정에서는 독립변수의 순서가 초래하는 문제가 발생하지 않는다.

그러나 카테고리형 변수가 맑음, 비 **두 가지** 였기 때문에 가능한 것이다. **(다중성의 문제)**

5.3.8 Type 2 ANOVA

독립변수를 넣는 순서를 바꾸어도 검정 결과가 변하지 않는 분산분석

5.3.9 모델 선택과 분산분석

아래의 순서로 변수를 포함시키는 모델에 대해 생각해보자

(처음 나오는 1은 절편임)

- 변수를 1개씩 늘려나가 보겠습니다.

1. 맨 처음에는 독립변수가 없는 **Null 모델**의 잔차제곱합의 크기를 구한다.
2. 독립변수에 날씨만 넣은 모델의 잔차제곱합을 구한다
3. 잔차제곱합의 차를 구한다.

날씨의 변화에 따른 **군간 편차제곱합**은

모델에 날씨라는 독립변수를 추가하는 것으로 인해 **감소하는 잔차제곱합**과 일치한다고 할 수 있다.

- 독립변수 humidity를 추가

1. 잔차 제곱합을 구한다.
2. (날씨만 있는 모델의 잔차제곱합) - (날씨 + 습도가 들어간 모델의 잔차제곱합)

....

많은 독립변수를 가지는 경우,

분산분석은 독립변수를 1개씩 늘려나가면서 독립변수가 늘어남

감소한 잔차제곱합의 크기에 기반하여 독립변수가 가지는 효과의 크기를 계산

그래서 독립변수 추가 순서에 따라 sum_sq값이 크게 바뀌게 된다.

그래서 독립변수가 여러게 있을 경우 이것을 사용하면 잘못된 결과를 얻을 가능성이 있습니다.

5.3.10 Type 2 ANOVA와 수정제곱합

Type 1 ANOVA은 아래와 같이 잔차제곱합을 비교합니다.

각각 하나 씩 (0이랑 1, 1이랑 2 ...)

Type 2 ANOVA은 아래와 같이 잔차제곱합을 비교합니다.

이것은 0이랑 1, 0이랑 2...로 모두 모델 0의 잔차제곱합과 비교한다.

- 독립변수가 줄어들면서 증가하는 잔차제곱합의 크기에 기반해서 독립변수가 갖는 효과의 크기를 정량화하고 있다고..... 이런 방식이라면 변수 추가 순서를 바꾸어도 검정 결과는 변하지 않는다.
- 이 방법으로 계산된 군간 편차제곱합 : **수정제곱합**

5.3.11 Type 2 ANOVA (실습)

수정제곱합 계산

5.3.12 Type 2 ANOVA 해석

습도의 영향을 검정한 결과 는 *다른 독립변수가 있어도 습도가 매상에 영향을 끼치고 있는지 알 수 있는가?* 이다.

습도는 기온과 강한 상관관계가 있었다.

그렇다면 기온이라는 독립변수가 모델에 포함되어 있으면 습도는 매상에 영향을 끼친다고 볼 수 없게 되는 상황이 있을 수도 있다.

5.3.13 변수 선택과 모델 해석

그래서 다시 계속해서 변수 선택을 진행

- 습도는 모델에 필요 없다는 것을 알 수 있었으므로 이것을 빼고 다시 Type 2 ANOVA 수행

5.3.14 AIC를 이용한 변수 선택

- 이것은 분산분석처럼 계산 방법을 바꿀 필요가 없음.
- 모델을 만들고 AIC를 비교하면 됨.

*모든 변수를 포함한 모델*과 *습도를 제외한 모델*의 AIC를 비교해보자

```
print("모든 변수를 포함한 모델 :", mod_full.aic.round(3))
print("습도를 제외한 모델 :", mod_non_humi.aic.round(3))
```

습도를 제외한 모델의 AIC가 더 작아졌다.

- 이 때문에 습도는 매상 예측 모델에서 제외해야한다는 결론

AIC는

- **검정의 다중성**을 걱정할 필요가 없다.
- **검정의 비대칭성**도 마찬가지
- 기계적으로 수행 가능 - 그래서 **편리**하다
- 그러나 역시 너무 믿으면 안됨

5.3.15 다중공선성

다중공선성 - 독립변수 간에 강한 상관관계가 있을 때 나타나는 문제

이에 대한 대처로서 가장 간단한 것 - **상관관계가 강한 변수 중, 어느 한쪽 모델을 제거하는 것**

다중공선성이 있으면 추정된 계수의 해석이 어려워진다.

그래서 변수 선택 후, 그 결과를 이용해서 해석 하는 것이 중요하다.

강한 상관관계가 있으면 p값 해석도 어려워진다.

그래서 사전에 어느 한쪽의 변수를 제외하고 모델링 하거나 7장에서 배우는 리지 회귀 등으로 해당 문제를 완화하려 한다.