\*\* 부분회귀계수 집합검정.

Beta1, Beta2

k개의 회귀계수에서 0인지 아닌지 가설검정에서

SSR(beta1|beta2) = SSR(beta1,beta2,…,betak| beta0) = SSR

추가 회귀 제곱합이 순수한 유의성 t검정.

(개별회귀계수검정)

SSR(b1,b2) – ssr(b1) = SSR(b2|b1) \*\*

3번의 특별한 경우

다중회귀분석 회귀유의성, 개별회귀계수, 부분집합검정에서

세번째의 특별한 경우는 1번경우와 같아짐(회귀유의성검정과 같아짐)

예제1.

다중선형회귀모형으로 적합시킬건데

SAS코드를 보면

SAS에는 2가지 스텝이 있는데

1. Data step
2. Proc step

SAS: 통계학적인 어플리케이션 위주였는데, 분야가 방대해져서 회사경영 데이터마이닝 등등 많은 것을 다룸.

해당하는 통계적 분석 개발 : Proc.

* 매뉴얼을 확인해서 다중회귀엔 어떻게 해야하는지 찾아봐야함. (보통 anova, glm 많이 씀)

데이터를 분석하기 위한 데이터를 만들어주는 스텝: data

* 읽어주는 것도 되지만, 프로그래밍 랭귀지, R처럼 할 수 있음.
* SAS 언어랭귀지.

Data 데이터스텝의 이름(형성된 데이터셋이 담길 이름으로 저장);

Input x1 x2 y;

(입력할 때 이름 줘야함,x1,x2,y)

Cards;

(밑에 데이터 카드가 들어간다. 위 형식으로 읽어들여라)

80 8 2256

.

.

.

Run; (끝에는 run 이라고 씀. proc시스템에선 안넣어도 돌아감. 일반적으론 넣게 되어있음.)

Proc gplot

(다중회귀모형적합시킬텐데, proc g : 그림을 그린다는 것. Proc plot은 해상도가 떨어짐. 훨씬 그림 이쁘게 나오는건 g.. 따로 사용료 받음)

만약 proc gplot ; 이렇게 닫으면 바로 앞에있는 것만 그림을 그리거든

그래서

proc gplot data=데이터이름;

이걸 써주는게 좋다.

Plot Y좌표\* 좌표 = 1;

(산점도를 그리는데 symbol 1번 형식으로 그리는 것. Star, square 이런식으로 캐릭터 모양을 지정할 수 있다.)

만약 두개 써주면 2개 그려줌.

Run;

Regression 하기 전에.

Proc corr data=이름;

Var x1 x2 y ;

상관분석을 해줌. 데이터의 변수를 지정해줘야 함

Var 뒤에 3 변수로 상관계수 행렬을 보여달라 이런 것임.

각 변수의 최대값 최소값 그런것도 알려주기도 함.

SAS에서 가장 간단한 시스템 중 하나.

Proc reg data = ;

(복잡한 프로시져. 옵션도 많음.)

다중회귀모형을 적합시킬려고 하는 것.

Model y=x1 x2;

Run;

(상수항 절편은 아무것도 안쓰면 절편이 있는거라고 생각하고 함)

절편없는경우에는 no constant라는 걸 적어줘야함

비교돌림은

Proc reg data=;

Model y=x1

Proc reg data=;

Model y=x2

맨 마지막엔 quit; 혹은 run;을 써야함

Quit;이 맨 마지막이니까 더 좋긴해.

창에 출력창, 로그창, 입력창, Output이 나오는 창이 잇음(viewer)

\*\*\* 분석결과보기

선형모형 적합시키려고 하는거잖아요

Y = x1 + x2 + e

1. Scatter plot

Y – x1선형관계가 강하긴 한데 산포가 점점 넓어지는 느낌이 드네

Y – x2 상대적으로 뚜렷한 관계가 아닌거 같고 곡선관계 같기도하고

1. 상관관계

상관계수들의 3x3 행렬이 나옴.

대각선 상엔 1이나오게끔 되어있고

서로다른 x1,x2 상관계구있고 밑에는 유의확률이 있음.

p-value 정의 정확히 익히기\*\*

유의수준은 사람마다 좀 다를수있잖아요. 그래서 오늘날은 p-value만 제시함.

(내가 귀무가설을 지지않지 않을 그 최대허용치: 사람마다 다름. 내가 오류를 범할 가능성, 위험을 가지고 있는 것.)

실험한 사람 같은 경우 x는 설명변수로 하면 안됨.

X,y 둘다 확률변수로 볼 수 있는데. 확률변수 주어진 값에서의 확률변수 y에 대한 설명할 수 있는 관계모형을 따지는 거니까 x는 다중선형회귀모형에선 상수취급

다만 상관분석에선 xy 둘다 확률변수로 해야 맞다.

회귀모형을 적합시킬건데

아노바 분산분석표\*

GLM ANOVA … 보면 exam9\_2의 경우.

모형제곱합(여기선회귀제곱합)

에러제곱합

수정총제곱합

변동계수(평균분에 표준편차) ; coeffvar

Intercept에 대한가설검정

일반적으론 관심없다. 상수항을 넣는건 회귀직선을 적합시키기위해 넣는거임.

원래 회귀분석에서 계수 추정하면 절편에 대한 추정치는 나오게 되어있음.

절편이 있어야 하거든. 일반적으론.

그래서 b0에 대한 건. x데이터들이 y축에 가까이있을수록 beta0가 잘 추정이 되고

X데이터들이 y축에 멀면 b0의 추정이 어렵기도 하고 정확도가 떨어진다.

(y축에 데이터가 멀리 떨어질주록 신뢰구간의 폭이 넓어지니까. 등간격으로 신뢰구간의 폭이 형성되는게 아님. 회귀직선에 대한 신뢰구간이…. 절편이 멀리 떨어질수록 어렵지. 더 먼 미래를 알아맞추기 어려운거랑 같은거야)

만약, intercept에서 p-value를 보고 아 p-value가 크다고 해서 막 제거하면 안됨.

(유의확률로 따지기 보단, 절편값이 추정ㄱ해보니 0에가깝게 나왔어, 게다가 유의확률도 엄청 크게 나왔어. 그럼 beta0를 무시해도 괜찮을 수도 있지. 근데 이게 일반적인 것은 아니야)

상수항에 대한 유의성을 따지는 건 위의 케이스 외엔 안된다.

근본적으로 뺄 수 있는 건

절편이 없는게 그냥 이론상 맞다 라고 하면 그게 맞는거지

그런 경우 아니면 넣어야해. 유의하지않다고 무조건 버리면 큰일남.

(0에 가깝되, 유의확률도 엄청 크게 나올때만 빼는게 가능해)

실험계획에선 자유도가 분할되서 합치면 위에것과 맞아 떨어지는데

근데 회귀분석에서는 모델-> 개별회귀계수 합친거랑 달라.

왜냐면 beta0는 (intercept)는 귀무가설에선 포함 안되서 1개가 더 많이 나올 수도 있어.

(그래서 b0,b1,b2에 대한거랑 SSR은 다르다)

X1만 돌린건 유의햇고

X2만 돌린건 유의하지않았어

근데 0.3088이 상관계수확률에서 유의확률가 같음

상관계수의 유의확률이 같다.

X1, x2랑 두개를 다 넣는게 좋은거 같기는 한데

X2혼자있을 때 별볼일 없는게 같이 들어가니까 영향력이 있음.

Ex. 반응온도 혼자 설명하는거보다 촉매공급률이 들어가니까 선형관계가 더 잘 설명되더라. => 아마도 점도의 어떤 데이터의 변동성 중에서 반응온도에 의해 설명하고 남은부분은 촉매공급률에 의해 설명되는데 상당히 도움이 된다는 의미

* 하나의 설명변수에서 다른 반응변수를 regression에서 설명하고 남은 잔차를 갖다가 나머지 다른 설명변수의 관계에서 보면 다른 설명변수가 유의하면 회귀하고 남은 요 잔차를 잘 설명하기 때문(이론적으로 되어있음)
* X2는 y변동성이 커서 설명이 턱도없이 부족해.
* 찌꺼기를 설명하기엔 상당히 의미가 있음.
* 그래서 유의성이 올라갈 수 있지

추가회귀제곱합

X1가지고 SSR

X1,x2가지고 SSR 검정할때

* 추가회귀제곱합을 계산하게 되면 그 나오는 유의성. Beta2가 추가로 들어갔을때의 F검정의 유의성은 x2라는 회귀계수의 유의성하고 똑같이 나와야 함.

순수하게 늘어나는 것(맨 앞의 식을 참고하기)

B2=0이냐 아니냐 검정하는거랑 같은거니까

B2의 t검정이나 f 검정이나 같아지는거지.

이 F검정과 t검정이 유의확률이 같아야하고 t^2 = F여야 함.

편회귀계수랑 같은 것임.

t검정이란 것은 편이란 개념이 들어가 있는 것임.

나만의 순수한 기여도.

추가회귀계산해보면 딱 맞음.

::예제::2

실험데이터

8가지 조합에서 실험을했는데 반복을 안함. 3원배치법

이걸 가지고 하나하나를 각 요인에 두가지 수준중에 하나를 해석해서

3원배치법으로 보면, 온도, 압력,농도 => 과연 생산량 수율에 영향을 미칠까?

온도압력, 압력농도, 온도농도 상호작용이있는가 정도는 가능.

반복이없으니까 삼요인상호작용은 오차항으로 간주함

120,160,140은 그 수준자체에 의미를 둔거임.

계량형 변수로 본거임

* 그래서 양적인 변수로 적합시키면 이제부터 얘내들끼리 y생산량에 영향을 미치느냐에다가 추가로 얼마나 영향을 미치는가, 생산량이 증가하냐 감소하냐 그게 더 추가가 된거지.
* 질적차이에다가….. 양적차이까지 추가로 더 봄

SAS에는 있는그대로 일단 먼저 쳐놓고

손으로분석하는 경우에는 코딩을 해야해

원래 120, 160 중간의 값인 140추가

* 오차항의 분산을 추정하기위해 집어넣은 것임.
* 손으로 계산하는게 시간이 많이 안걸리는데
* 데이터로 일단 코딩을 함.
* (중간값을빼주고, 120-160의 값 차이에 2로 나눠)
* 그러면 -1,1,0으로 코딩이 됨.(부호변수) 그러면 손으로 계산하기 편리함.(9.3예제)
* solve(X’X)에 뭐 어쩌구저쩌구 계산하기 어려운데 이걸 사용하면 편하게 계산할 수 있음.
* (노트참고하기) : 매우 간단해짐.
* 이와 같이 9.4로 봤을땐 두 수준으로 실험했는데 모든 가능한 조합에서 실험하잖아요? 그럼(서로다른 변수에대한 코딩한 값들이 내적이 0, 직교한 열로 구성된 그런 실험에서만 대각행렬이 나타나기 때문에 쉽게 계산할 수 있음)
* 다음에할것이2수준, 3수준 실험.
* 2수준의 실험도 다 이런 것임(손으로 계산할 수 있음)

9.3 correlation

Y = x1 x2 x3;

상관계수가 0이 나오면 세개의 요인들이 직교하기 때문에 상관계수행렬에서 상관계수가 0이 나옴.

상관계수에서 분자가 공분산이고 각각 두개의 변수의 표준편차의 곱인데, 공분산이 0이라서 0인거거든요?

행렬의 열들이 직교하면 공분산이 0이되는데 이건 회귀분석의 경우이고

위의 예의 경우에

* Cov(b) = sigma^2(x’x)-1
* (X’x)-1 : 대각행렬, 그래서 서로다른 회귀계수의 공분산은 0이다
* : Cov( bi, bj)=0. 그 회귀계수의 공분산이 0이면 변수간의 상관계수도 똑같이 0이됨

온도, 압력, 농도중에서 압력은 수율에 압도적으로 영향을 미치는데,(유의성도)

X1, x3는 수율에 그렇게 깊은 영향을 주는 것 같진 않음

그럼에도 분석간 결과를 보면 x1,x2는 유의. 그래서 x3는 안유의하니까 빼고 돌려보자.

Y = x1 x2 ;

모델이 뭐 어느게 더 나을까 보기위핸 Adj R^2 를 비교해봄.

RMSE가 작을수록 R^2가 커짐\*

왜냐면 수정결정계수는 1-SSE/(n-k-1) / SST/(n-1) = 1 – MSE / SST/(n-1)

SST/(n-1)은 고정, MSE랑 adj R^2는 반대, 하지만 비교방법은 서로가 같은거임

둘중에 하나만 쓰면 됨.

변이계수, 변동계수\*\*

{Y의 평균/sqrt(MSE)} \* 100

MSE가지고만 하면 측정값에 따라(데이터의 측정단위에 따라 상대적으로 더 커질 수 있으니까. 그래서 그 단위를 좀 정량화 시킬려고 변이계수, 변동계수를 씀. Coef Var.)

* Coef Var = 6.56
* 평균에서 6.56정도

시그마 제곱이 평균에 대해 6.56배정도 되는구나 이정도를 알 수 있음.

9.3c, 9.3b 숙제1,2번.

9.3a (노트와 책에도 없는 예임)

진공포장할 때,

X1 = sealing temp(동봉할 때 온도)

X2 = cooling bar temp(차가운봉에 돌리는 거)

X3 = polyethylene (봉지에 폴리에틀렌 몇퍼센트로?

Y = seal strength 봉함힘

선형모형을 적합시켜보자

예제9.3과 다르게 직교한 형식으로 접근할 수는 없음

세개 변수에 대해 1차,2차항, 상호작용 효과까지 다 넣어봄(3 변수에대해)

* 그래서 9개 항까지 들어갈 수 잇는 모형을 만듬.
* (1차모형이 잘 맞으면 필요가없고, 회귀모형 부분집합 그거 해볼려고 하는 짓임)

데이터 읽어들였는데, 그것만 가지고 분석하는게 아니라 제곱합,

실험설계모형에서 제곱항 넣고싶음 x1\*x1 이러면 됨.(glm에선, ANOVA는 아님. 대신 데이터를 양적변수로도 할거니까 $ sign넣으면 망함. 숫자로 읽겠다면 x1$ 이거 큰일남.=>에러남. X1\*x2 항 모양을 만들어야함 그래서 data quadratic ; 이란 표현이 있음. )

Set 원래데이터셋;

원래 데이터셋을 이용해서 quadratic을 새로 만들겠다.

X1(square) = x1\*x1;

X2(squre) = x2\*x2;

.

.

.

* 귀찮지만 6개의 항을 미리 만들어 냄.
* 그래서 데이터스텝에서 읽어 들인 다음에 또 다른 데이터 스텝에서 제곱항하고 곱하기항을 추가로 만들어 냄.
* quadratic데이터셋을 이용해서 regression을 돌림
* 9개 항을 넣음(y)

ANOVA의 경우에는

Model y = x1}x2}x3 @2; 원래는 알아서 x1,x2,x3 제곱 알아서 다 분석을 함.

근데 회귀분석은 불가능함.

회귀계수 부분집합 검정을 할 때, 순수 2차항 3개, 서로다른 교호작용3개 해서 6개의 항이 다 필요없다, 필요하다 따져볼 수 있잖아

1차항만 갖고하냐 다른게 필요하냐? 이런거

Y = b0 + b1x1 + b2x2+ b3x3 + b11bx1^2 + b22x2 ^2… 노트참고

교호작용term (x1x2=0, x2x3=0…. ) 증명하기.

Test문장이 없을 때 해 볼 방법.

3개의 1차항 ->적합시킨 모형

상호작용 빼버린 모형이 낫지 않냐? 하는 6개 모형을 적합시켜서 9개 다 들은 모형이랑 비교를 해봄.

두개의 차이를 보고, 교호작용도 나눠서 F검정 해보면 cross-term이 들어가야하냐 안들어가야하냐를 체크해볼 수 있음.

* 교호작용은 다 0이 되어버렸고, 1차항만 필요하단건 아님. 3개의 순수한이차항과 3개의 일차항 총 6개의 항이 필요함

결과:: SAS System. Model

추가회귀제곱합 / 자유도 = 6.555 (numerator)

Denominator 1.18678 = MSE (완전모형의 MSE임)

교호작용항이 필요하냐 안하냐를 검정한 것

Test 2 results for dependent variable

* 유의확률 보니까 교호작용항은 별로 설명에 도움이 안되네, 빼도 괜찮겠네?
* 1차항만 갖고는 parameter estimates에서 x2는 정말 설명력이 떨어지네!

완전모형의 -노트참고 \*\* 앞의 설명이 이해될 것.

(결론)

그래서 크로스 텀을 다 빼버렸더니 회귀계수들이 다 유의하게 나옴.

R^2만 가지고 너무 비교하면 큰일나니까, 이렇게 상세하게 봐줘야함.

수정결정계수는 MSE 작은걸 사용한다는 당연한 얘기임. 그래야 설명력이 당연히 좋다는 의미임.

* 적합도는 변수의 수까지 다 보고 체크해야함.

9.3C

@@; 끝날때까지읽어라

대칭모양이면 이차항

비대칭이면 삼차항모형 회귀분석으로 해야할 수도 있다

1. Ax+b

잔차그림이 저런 모양이면 2차 회귀분석을 해야함.

이차항으로 회귀분석 하니까 R^2가 커짐

삼승항까지 넣었더니, d건 유의하지 않았다 -> 필요없는 항. ( 눈에띄게 R^2가 늘지않아, 그러면 수정결정계수는 줄어들지) => 빼야함.

**9.3b**

변수가 두개이상의 조합이면 골뱅이를 꼭 @@쳐야함.

* 요인, 변수가 2개 이상 들어갈 때 첫번째 인풋에대해선 @하나만, (숫자가 아니면 $ @; )
* 두번째 인풋문장쓸 때 @@ 두개 씀.

Exam5\_1

* 한줄에 여러 개 넣으면 @한 개 @@ 두개 이렇게 해도 됨.
* 한 문장에 데이터가 한 세트씩 들어갈 때, 골뱅이 안 써도 됨.

\*\* 소나무 데이터 9.3b

임업

X3 : 1에이커당 소나무의 그루 수.(지역에 따라 빽빽하게 심겨졋거나 널널하게 심겨졌거나) : 밀도

* 가장 좋은모형을 찾자.
* 다 비교해보자. 2^3-1 개. =>7개모형
* 설명변수가 수십개가 되는 경우에, 비교가 불가능. (차선책-부분집합을 비교해본다.) STEPWISE..
* 여기선 7가지 모형을 다 돌려보라는 얘기임.

Test x1=0; x1이 필요하냐 안 필요하냐

Test x2=0; x2가 필요하냐 안 필요하냐.

. . . .

* 할 필요가없음, 회귀계수 t검정이랑 같으니까/
* 추가회귀제곱합의 개념.
* 근데 SAS에서 parameter estimate part에서 교호작용 부분은 안나오거든
* 그래서 x1=0, x3=0; 이런건 돌려봐야함.
* 3개 다 넣고 돌린거 2개 다 넣고 돌린거 이런거 추가회귀제곱합 이런거!!!! 가 중요하거든
* 이거 SAS가 안알려주거든!!!!! 그래서 그게 중요한거임!!!!!!!!!!!!!!! 10장. 공분산분석에서 중요함.

RMSE, R^2로 비교해보장

다중회귀분석

X1만 적합시키면, 나이만가지고 밑둥(직경)을 설명하려면 상당히 설명력이 좋은 편이긴 함.

X2는 소나무의 높이를 가지고 하면 조금 더 설명력이 높음

X3 = 빽빽할수록, 성장하는데 방해를 받겠지? (기울기가 음의 방향 느낌)

이게 reg procedure model1

X2가지고만 설명하는덴 한계가있따. 그래서 x1도 같이 넣고 돌려보자

* X2는 유의한데 x1는 유의하지않아.
* Correlation이 높았음. => x2의 설명할수 있는 영향력이 x1에서 좀 많이 갖고있음 => 그래서 x1의 설명력이 별로 안좋아 => 그래서 별 볼일 없음. 확 떨어짐.(편상관계수)
* X3은 혼자서 설명력은 없어, 그렇긴 하지만??? X3이 있잖아!!! 그래서 x1을 그냥 무시해볼 수는 없어. 글서 한번 더 해보자 (높이가 설명하지 못하는 나머지 찌꺼기가 있잖아, 그 못하는 x2가지고 설명한 나머지 찌꺼기를 가지고 3번째 변수 x3를 한번 회귀를 시키면 그게 회귀를 시키면 x3가 되고, y,x1을 회귀시킨 찌꺼기를 해보면, x3가 유의해져버림.
* 이건 모순되는게 아니라 y가 x1가지고 상당부분을 설명하고 남은 부분을 보니 설명을 많이 하더라!
* X1,x3 설명이 괜찮고 x2,x3에선 설명이 잘 안되네!! (p-value check)