

Regression Analysis Tutorial

(회귀 분석 튜토리얼)

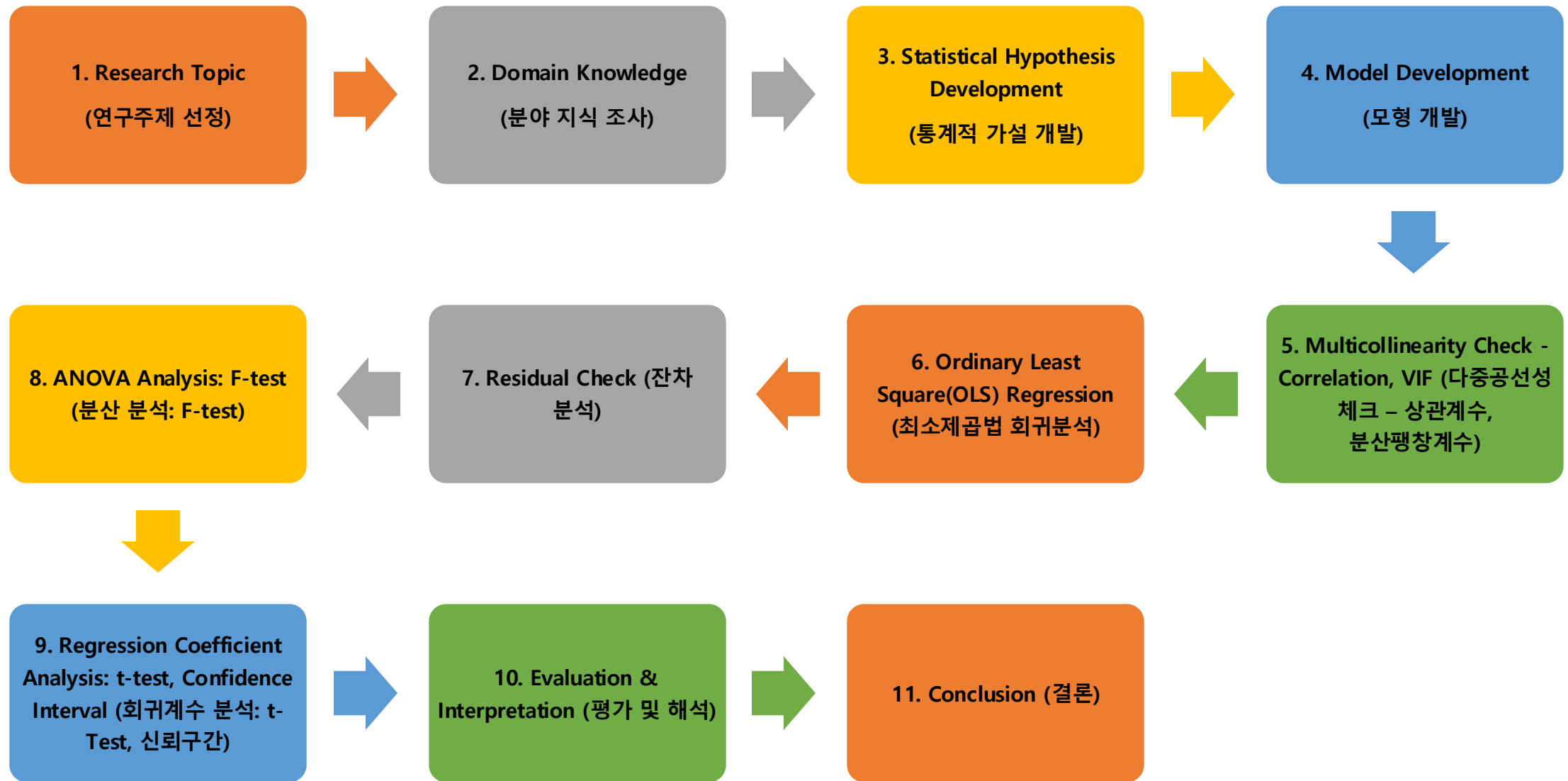
Eunseok Kim

Disclaimer

워크시트:
셀 주소:

- 이 자료는 학생들로 하여금 회귀 분석의 모든 과정을 따라해볼 수 있는 Tutorial 자료로 만들었습니다
- 함께 제공된 엑셀 스프레드 시트 "Regression Tutorial.xlsx"는 슬라이드에서 제공된 계산값과 분석 결과를 제공합니다. 엑셀 스프레드 시트 자료의 특정 부분을 참조해야 할 경우 이를 슬라이드 우측 상단에 표시해두었습니다
- 자료에 설명된 모든 과정을 직접 따라해보면서 스스로 엑셀을 활용하여 회귀분석을 수행할 수 있도록 연습하세요. 회귀분석은 혼자서도 잘할 수 있다는 자신감이 생기는 것이 목표입니다

Regression Analysis Workflow



1. Research Topic & 2. Domain Knowledge

□1. 연구주제 선정

- 연구자 본인이 규명하고자 하는 연구주제를 찾는 과정.
- 예) 자동차 연비에 영향을 주는 요인들을 분석하고 싶다. 어떤 요인들이 있을까?

□2. 분야 지식 조사

- 차량 연비에 영향을 주는 요인들에 대해서 알아보고, 이전에 관련 연구가 없었는지, 연구에 활용할 수 있는 데이터는 어떤 것이 있는지 조사
- 예) 자동차 연비에는 차량 무게, 엔진 출력 등 다양한 요소가 영향을 주는 것으로 알려져있다. 이러한 영향력을 통계적으로 검증하는 연구에 활용할 수 있는 데이터로는 MtCars라는 데이터가 있다.

3. Statistical Hypotheses Development

□3. 통계적 가설 개발

- 배경지식을 바탕으로 통계분석을 통해 입증하고자 하는 가설을 선정한다.
- 예) “차량의 연비는 엔진의 배기량에 반비례한다”는 배경지식을 바탕으로 다음의 가설을 선정
 - 귀무가설 (H_0): 1/배기량은 차량의 연비에 미치는 영향은 통계적으로 유의미하지 않다
 - 대안가설 (H_1): 1/배기량은 차량의 연비에 미치는 영향은 통계적으로 유의미하다

4. Model Development

□4. 모형개발

- ▣ 회귀분석을 통해 검증하고자 하는 회귀식을 정의

- 예) $(\text{연비}) = \beta_0 + \beta_1 \cdot \left(\frac{1}{\text{disp}}\right) + \epsilon$: 연비와 1/배기량 사이의 단순회귀 모형

- 종속변수는 연비, 독립변수는 (1/배기량)

- ▣ 그러나 연비에 영향을 미치는 다른 요인들이 있기 때문에, 보통의 경우 이 요소들의 영향력을 배제하고 순전히 배기량이 연비에 미치는 영향력을 분석하기 위해, 이들 요소들을 통제 변수(Control Variable)로 회귀모형에 추가해서 회귀분석을 진행

- 예) 차량의 무게도 연비에 반비례한 영향력을 준다고 알고 있을 경우: 1/무게를 통제 변수로 회귀 모형에 추가하여 1/무게가 연비에 미치는 영향을 배제한 상태에서, 1/배기량이 연비에 미치는 영향을 분석

- $(\text{연비}) = \beta_0 + \beta_1 \cdot \left(\frac{1}{\text{disp}}\right) + \beta_2 \cdot \left(\frac{1}{\text{wt}}\right) + \epsilon$: 다중 회귀 분석

- 종속변수는 연비, 독립변수는 (1/배기량), (1/무게)인 다중 회귀 모형

5. Multicollinearity Check

□5. 다중공선성 체크

- 다중공선성은 두 가지 방법으로 체크: 1) Correlation(상관관계) 검사, 2) VIF(분산 팽창 계수) 검사 – 두 검사 모두 통과해야함
- 1) 상관관계 검사: 엑셀 상관관계 분석 기능 활용 (부록 슬라이드 참조)
 - 회귀식에 포함된 독립변수들간의 강한 상관관계가 없어야 함. 일반적으로 상관관계값이 -0.7 보다 크고 0.7보다 작아야한다는 기준을 요구함
 - E.g., 앞서 예제 회귀식에서는 (1/배기량)이 독립변수로, (1/무게)가 통제변수로 포함되어있었음. 따라서 (1/배기량)과 (1/무게)간의 상관관계가 -0.7~0.7 이어야함

5. Multicollinearity Check

워크시트: Correlation Analysis
셀 주소: R6:AG21

□5. 다중공선성 체크

- 다중공선성은 두 가지 방법으로 체크: 1) Correlation(상관관계) 검사, 2) VIF(분산 팽창 계수) 검사 – 두 검사 모두 통과해야함
- 1) 상관관계 검사: 엑셀 상관관계 분석 기능 활용 (부록 슬라이드 참조)

	mpg	cyl	disp	1/disp
mpg	1			
cyl	-0.852162	1		
disp	-0.8475514	0.90203287	1	
1/disp	0.92719281	-0.8887598	-0.8897124	1
hp	-0.7761684	0.83244745	0.79094859	-0.760039
1/hp	0.85913212	-0.8506863	-0.7896366	0.86974675
drat	0.68117191	-0.6999381	-0.7102139	0.75460333
wt	-0.8676594	0.78249579	0.88797992	-0.8466647
1/wt	0.89161728	-0.7753757	-0.80166	0.89792694
qsec	0.41868403	-0.5912421	-0.4336979	0.42621937
vs	0.66403892	-0.8108118	-0.7104159	0.70012849
am	0.59983243	-0.522607	-0.591227	0.65990751
gear	0.48028476	-0.4926866	-0.5555692	0.49250578
carb	-0.5509251	0.52698829	0.39497686	-0.4667893
1/carb	0.59876953	-0.5933848	-0.4870424	0.56848903

(1/배기량)과 (1/무게) 간의 상관관계는 $0.89 > 0.7$
따라서 상관관계 검정 실패. 즉 (1/무게)는 통제변수로서
(1/배기량)와 함께 회귀식에서 활용이 불가능

5. Multicollinearity Check

워크시트: Correlation Analysis
셀 주소: R6:AG21

□5. 다중공선성 체크

- 현실적으로는 상관관계 분석을 먼저 실행하고 검사 기준을 통과하는 변수들을 우선적으로 추려서 이들을 통제변수로 포함시킴
 - (1/배기량)과 상관관계 r 이 $-0.7 < r < 0.7$ 인 변수들을 찾아보면 qsec(제로백 – 가속능력), am(자동변속기 장착여부), gear(변속기의 단수), carb(엔진 기화기 갯수), 1/carb (기화기 갯수의 역수) 임을 알 수 있음

	mpg	cyl	disp	1/dis	hp	1/hp	drat	wt	1/wt	qsec	vs	am	gear	carb	1/carb
mpg	1														
cyl	-0.852162	1													
disp	-0.8475514	0.90203287	1												
1/dis	0.92719281	-0.8887598	-0.8897124	1											
hp	-0.7761684	0.83244745	0.79094859	-0.760039	1										
1/hp	0.85913212	-0.8506863	-0.7896366	0.86974675	-0.8817825	1									
drat	0.68117191	-0.6999381	-0.7102139	0.75460333	-0.4487591	0.63040063	1								
wt	-0.8676594	0.78249579	0.88797992	-0.8466647	0.65874789	-0.7079556	-0.7124406	1							
1/wt	0.89161728	-0.7753757	-0.80166	0.89792694	-0.6331044	0.73002735	0.71131297	-0.9218133	1						
qsec	0.41868403	-0.5912421	-0.4336979	0.42621937	-0.7082234	0.6168629	0.09120476	-0.1747159	0.17022005	1					
vs	0.66403892	-0.8108118	-0.7104159	0.70012849	-0.7230967	0.73379084	0.44027846	-0.5549157	0.55013866	0.74453544	1				
am	0.59983243	-0.522607	-0.591227	0.65990751	-0.2432043	0.39948222	0.71271113	-0.6924953	0.70893377	-0.2298609	0.16834512	1			
gear	0.48028476	-0.4926866	-0.5555692	0.49250578	-0.1257043	0.2607923	0.69961013	-0.583287	0.55893221	-0.2126822	0.20602335	0.79405876	1		
carb	-0.5509251	0.52698829	0.39497686	-0.4667893	0.74981247	-0.6186142	-0.0907898	0.42760594	-0.4261184	-0.6562492	-0.5696071	0.05753435	0.27407284	1	
1/carb	0.59876953	-0.5933848	-0.4870424	0.56848903	-0.6524165	0.6259762	0.11738189	-0.5075022	0.4752289	0.61717132	0.66928749	0.13697107	-0.1143989	-0.838736	1

5. Multicollinearity Check

□ 5. 다중공선성 체크

- 현실적으로는 상관관계 분석을 먼저 실행하고 검사 기준을 통과하는 변수들을 우선적으로 추려서 이들을 통제변수로 포함시킴
 - 예시로 우리는 이 중에 두 가지 am (자동변속기 장착여부), $gear$ (변속기의 단수)를 각각 통제변수로 회귀식에 추가해보기로 함. 그 결과 우리가 분석해야 하는 회귀식은 2개가 됨
 - $(연비) = \beta_0 + \beta_1 \cdot \left(\frac{1}{disp}\right) + \beta_2 \cdot (am) + \epsilon$
 - $(연비) = \beta_0 + \beta_1 \cdot \left(\frac{1}{disp}\right) + \beta_2 \cdot (gear) + \epsilon$
 - Note: 일반적으로 통제변수는 이전의 연구 사례나 관련 지식을 기반으로 여러 개를 포함시킴
 - 이 두 회귀식에 포함된 독립변수와 통제변수는 상관관계 분석은 통과했음 (애당초 통과하는 변수들을 모아서 회귀식을 구성했기 때문에). 이제 VIF를 검사할 차례

5. Multicollinearity Check

□ 5. 다중공선성 체크

- 다중공선성은 두 가지 방법으로 체크: 1) Correlation(상관관계) 검사, 2) VIF(분산 팽창 계수) 검사 – 두 검사 모두 통과해야함
- 2) VIF 검사: 엑셀의 회귀분석 기능 활용 필요 (부록 참조)
 - 회귀식에 포함된 독립 변수와 통제 변수간에 회귀분석을 수행하여 나온 unadjusted R squared value, R^2 를 활용하여 각각의 변수에 대해서 $VIF_{\text{변수}} = \frac{1}{1-R^2}$ 를 계산. 모든 VIF 값들이 5 이하로 나오면 통과
 - 예) $Y = \beta_0 + \beta_1 \cdot A + \beta_2 \cdot B + \beta_3 \cdot C + \epsilon$ 가 회귀식이라면 여기에 포함된 독립변수 & 통제변수는 A, B, C.
 - VIF_A 를 구하기 위해서는 $A = \beta_0 + \beta_1 \cdot B + \beta_2 \cdot C + \epsilon$ 로 회귀분석을 실행하고 이때의 R^2 을 결과창에서 확인 (이 값을 R_A^2 라고 하자). 그러면 $VIF_A = \frac{1}{1-R_A^2}$
 - 마찬가지로 VIF_B 를 구하기 위해서는 $B = \beta_0 + \beta_1 \cdot A + \beta_2 \cdot C + \epsilon$ 로 회귀분석을 실행하고 이때의 R^2 을 결과창에서 확인 (이 값을 R_B^2 라고 하자). 그러면 $VIF_B = \frac{1}{1-R_B^2}$
 - VIF_C 를 구하기 위해서는 $C = \beta_0 + \beta_1 \cdot A + \beta_2 \cdot B + \epsilon$ 로 회귀분석을 실행하고 이때의 R^2 을 결과창에서 확인 (이 값을 R_C^2 라고 하자). 그러면 $VIF_C = \frac{1}{1-R_C^2}$
 - VIF_A, VIF_B, VIF_C 모두 5 이하여야함

5. Multicollinearity Check

워크시트: VIF 1disp~Am, VIF
Am~1disp, VIF 1disp~Gear
셀 주소: B5

□5. 다중공선성 체크

▣ 2) VIF 검사: 엑셀의 회귀분석 기능 활용 필요 (부록 참조)

- 예) $(연비) = \beta_0 + \beta_1 \cdot \left(\frac{1}{disp}\right) + \beta_2 \cdot (am) + \epsilon$ 의 경우: $\left(\frac{1}{disp}\right) = \beta_0 + \beta_1 \cdot (am)$ 와 $(am) = \beta_0 + \beta_1 \cdot \left(\frac{1}{disp}\right)$ 의 R^2 를 계산해서 $VIF = \frac{1}{1-R^2}$ 를 계산해야 함

- 각각의 경우 $R_{disp}^2 = R_{am}^2 = 0.43547$ 이 나왔으며,
따라서 $VIF_{disp} = VIF_{am} = \frac{1}{1-0.4357} \cong 1.772$ 가 나옴

- 모든 VIF가 5 이하임으로 VIF 검사를 통과함

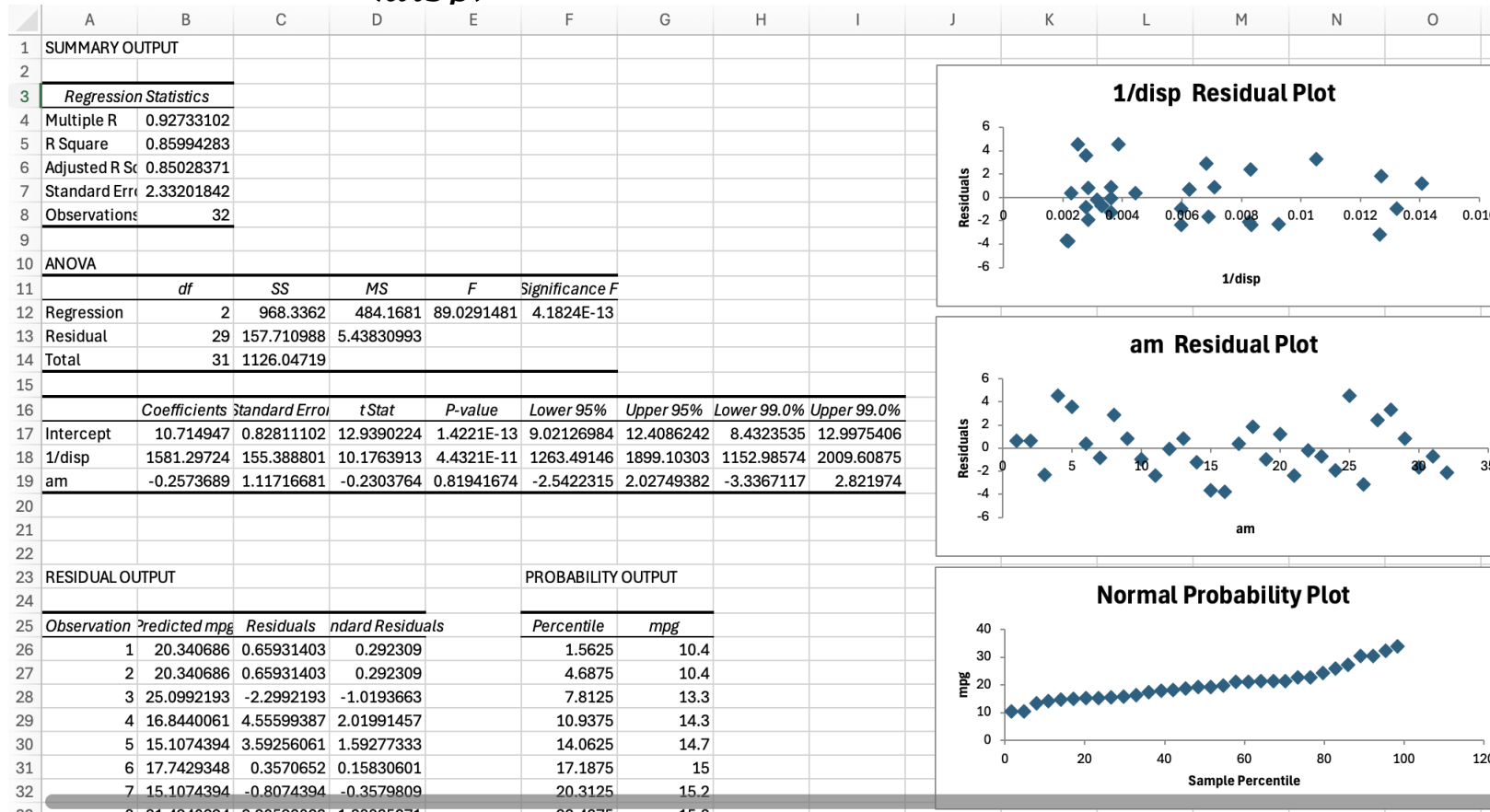
- c.f., $\left(\frac{1}{disp}\right)$ 와 $(gear)$ 의 경우 $R^2 = 0.2425$, $VIF \cong 1.32$
마찬가지로 모든 VIF가 5 이하이므로 VIF 검사를 통과

SUMMARY OUTPUT			SUMMARY OUTPUT		
Regression Statistics			Regression Statistics		
Multiple R	0.65990751		Multiple R	0.65990751	
R Square	0.43547793		R Square	0.43547793	
Adjusted R Square	0.41666052		Adjusted R Square	0.41666052	
Standard Error	0.00274001		Standard Error	0.38111261	
Observations	32		Observations	32	
ANOVA			ANOVA		
	df	SS		df	SS
Regression	1	0.00017374	Regression	1	3.36134525
Residual	30	0.00022523	Residual	30	4.35740475
Total	31	0.00039897	Total	31	7.71875
Coefficients			Coefficients		
Intercept	0.00406781	0.0006286	Intercept	-0.1440383	0.13275532
am	0.0047444	0.00098623	1/disp	91.7877584	19.0801387

6. Ordinary Least Square(OLS) Regression

□ 6. 최소제곱법 회귀 분석 수행: 엑셀의 회귀분석 기능 활용 필요 (부록 참조)

▫ 예) (연비) = $\beta_0 + \beta_1 \cdot \left(\frac{1}{disp}\right) + \beta_2 \cdot (am) + \epsilon$ 의 경우:



6. Ordinary Least Square(OLS) Regression

□6. 최소제곱법 회귀 분석 수행: 엑셀의 회귀분석 기능 활용 필요 (부록 참조)

- 한편 $(\text{연비}) = \beta_0 + \beta_1 \cdot \left(\frac{1}{\text{disp}}\right) + \beta_2 \cdot (\text{gear}) + \epsilon$ 의 경우 특이한 점은 *gear* 가 3,4,5 중의 하나의 값을 갖는 변수라는 점
 - 변속기의 단수인 Gear가 3,4,5로 늘어남에 따라 수치적으로 동일한 비율로 연비에 영향을 줄 것이라 판단한다면, Gear가 ratio scale를 따르는 quantitative variable로 간주하고 회귀분석을 진행
 - 이 경우 앞서 am을 통제변수로 했을 때와 동일한 방법으로 회귀분석 진행
 - 변속기의 단수인 Gear가 3,4,5는 각각 독자적인 영향력을 갖는 그룹으로 판단한다면, Gear가 Nominal scale을 따르는 categorical variable로 간주하고 회귀분석을 진행
 - 이 경우 Gear 를 0, 1의 값만 갖는 binary variable로 표현 방법(=encoding)을 바꿔줘야 함

6. OLS Regression

워크시트: Selected Dataset
셀 주소: G1:L33

□6. 최소제곱법 회귀 분석 수행: 엑셀의 회귀분석 기능 활용 필요 (부록 참조)

- Categorical variable의 encoding을 binary variable로 바꾸는 방법:
 - 가능한 값들을 나열하고, 그 중 기준점이 되는 값을 지정: 예) Gear는 3,4,5 중 하나의 값을 가짐. 이중 Gear=3 인 경우를 기준으로 삼음
 - 기준 이외의 값에 해당하는 경우를 각각의 변수로 만들고, 해당 Gear 값을 만족할 때 1, 그렇지 않을 때 0으로 설정: 예) 두 개의 binary variable인 Gear=4, Gear=5을 만들고, Gear가 3일 때는 Gear=4, Gear=5 두 변수 모두 0, Gear가 4일 때는 Gear=4 는 1, Gear=5 는 0, Gear가 5일 때는 Gear=4 는 0, Gear=5 는 1로 처리.
 - 회귀분석은 이렇게 encoding을 바꾼 binary variable들을 독립변수로 포함시킨 후 실행

Gear		Gear=4	Gear=5
3		0	0
4	->	1	0
5		0	1

7. Residual Check

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

- 1) 선형성: 잔차의 평균이 0이고 특정 패턴이 존재하지 않을 것
- 2) 등분산성: 각각의 독립변수의 모든 값에 대하여 잔차의 분산이 동일해야함
- 3) 정규성: 잔차의 분포는 정규분포를 따라야한다
- 4) 각각의 잔차는 서로 독립이어야함

▣ 1)~3)의 항목을 위한 엄밀한 통계적 검정 방법이 존재하나 (선형성: Ramsey's RESET, 등분산성: Breusch-Pagan test, White test, 정규분포: Shapiro-Wilk test) 학부 통계 과목 수준에서는 시각적 방법으로만 점검

7. Residual Check

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

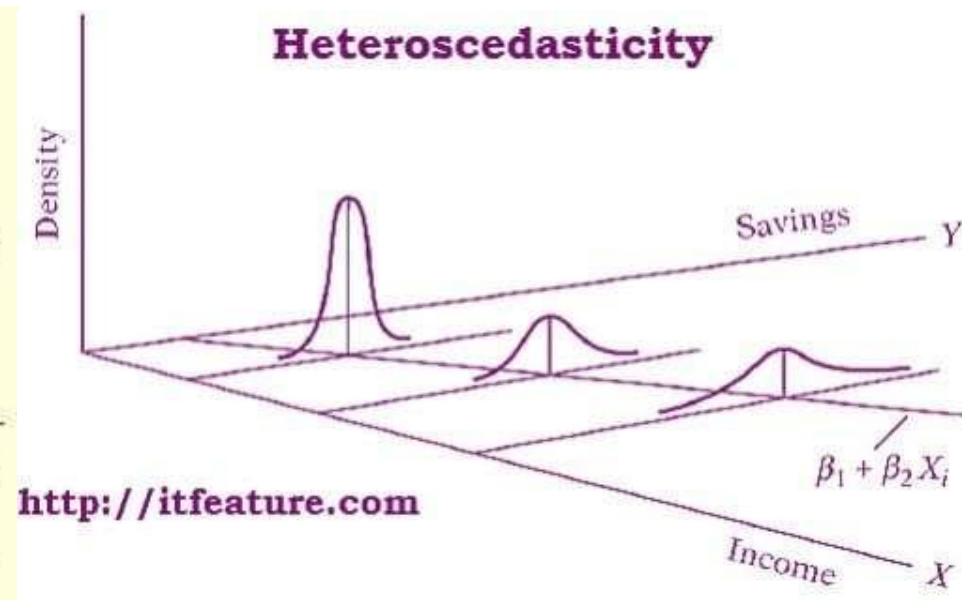
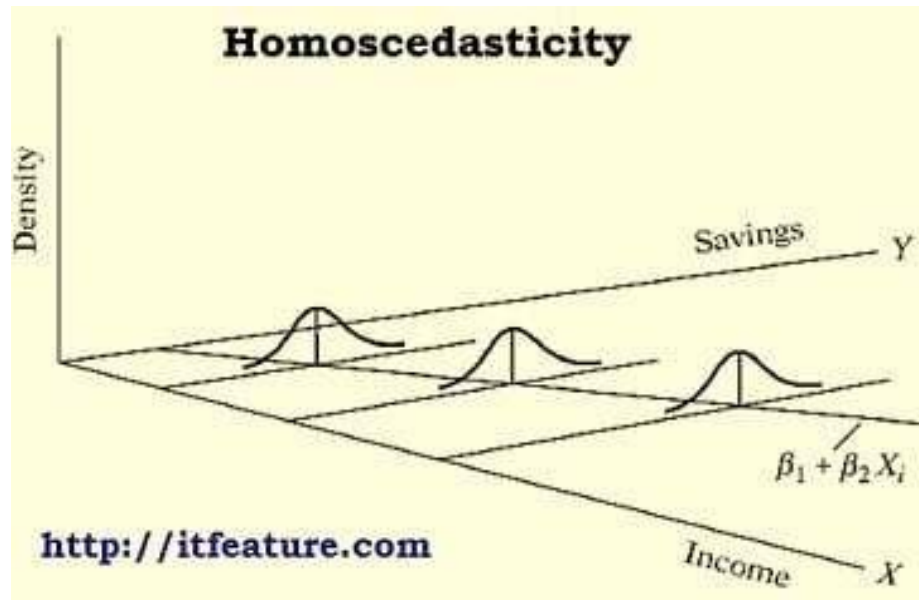
- 1) 선형성: 잔차의 평균이 0이고 특정 패턴이 존재하지 않을 것
- 2) 등분산성: 각각의 독립변수의 모든 값에 대하여 잔차의 분산이 동일해야함
- 3) 정규성: 잔차의 분포는 정규분포를 따라야한다
- 4) 각각의 잔차는 서로 독립이어야함

- ▣ 4) 잔차가 독립이어야한다는 전제 조건은 서로 다른 시점에 데이터가 수집되어 시간 흐름이나 순서에 따라 기록된 Time-series data나 Panel data 분석 시 인접한 시점의 데이터가 갖는 잔차가 서로 관련성이 있는 자기상관(autocorrelation) 현상으로 인해 깨지는 경우가 많음
 - 한 시점에 대한 데이터를 수집하는 cross-sectional data의 경우 데이터 내에서 클러스터링/그룹 구조가 존재하는 경우를 제외하면 대부분 이런 문제가 없어서 자기상관에 대한 검정이 필요하지 않음
 - 검정이 필요한 경우 데이터가 수집된 시간 순서대로 잔차 그래프를 그려서 잔차가 시간에 따라 영향을 받는지 확인하거나, Durbin-Watson test를 활용하여 검증. (Durbin-Watson test 같은 통계적 검정 방법은 본 과목의 범주를 넘어섬)
 - 현재 우리가 분석하는 Mtcars 데이터는 클러스터링 구조가 없는 cross-sectional data이므로 이러한 검증이 필요 없음

7. Residual Check

□ 7. 잔차 분석:

- 이상의 조건들을 전부 만족하지 않으면, 앞으로 진행할 F-test, T-test, Confidence Interval가 의미가 없어짐. 이들 과정은 전부 잔차가 평균이 0인 정규분포를 따른다는 가정하에 진행되는 것. 물론 이 정규분포들의 표준편차는 상황에 따라 다르지만, 독립변수의 값에 따라서 변하지 않고 일정하다는 것(등분산성 - Homoscedasticity)을 가정하고 있음. 즉, 등분산성이 성립되지 않는 경우(Heteroscedasticity), 표준편차에서 오차가 발생하며, 이 오차가 F statistics, T statistics, p-value, 그리고 신뢰구간 계산에 오차로 전이되므로 추후 검정 결과가 부정확해짐.

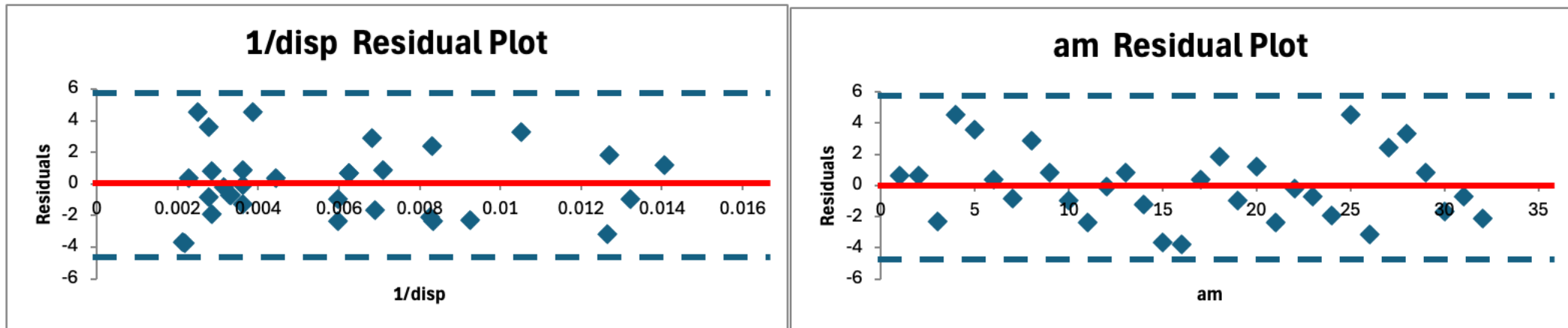


7. Residual Check

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

- 1) 선형성: 잔차의 평균이 0이고 특정 패턴이 존재하지 않을 것
- 2) 등분산성: 각각의 독립변수의 모든 값에 대하여 잔차의 분산이 동일해야함

▫ 예) $(\text{연비}) = \beta_0 + \beta_1 \cdot \left(\frac{1}{\text{disp}}\right) + \beta_2 \cdot (\text{am}) + \epsilon$ 의 경우:



7. Residual Check

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

3) 정규성: 잔차의 분포는 정규분포를 따라야한다

- 잔차 분포가 정규분포를 따르는지 시각적으로 확인하는 방법은 1차적으로는 잔차 정보를 바탕으로 히스토그램을 그리는 방법이 있다 (관련 엑셀 기능 존재). 그러나 히스토그램은 구간 설정에 따라 시각적 모양이 바뀌기 때문에 이를 정규분포와 비교하는데 어려움이 많음
- 따라서 우리는 잔차 분포가 정규분포를 따르는 지 시각적으로 확인하는 부분에 있어서 좀더 엄밀한 방법을 활용

7. Residual Check

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

3) 정규성: 잔차의 분포는 정규분포를 따라야한다

▣ 정규분포의 모양을 수치적으로 표현하는 방법 중 하나는 표준정규분포표를 거꾸로 적용하는 방법.

– 표준정규분포표에서는 z 값의 구간에 따라 확률을 알려주었음 ($0 \leq Z \leq 1.645 \rightarrow P[0 \leq Z \leq 1.645] = 45\%$).

– 반대로 확률을 가지고 z 값의 구간을 표현하는 방식을 생각해보자. Percentile의 경우 현재 데이터가 전체 데이터 중에서 어디에 위치하는지 알려준다. 예를 들어 특정 데이터가 1 percentile이라면 이 데이터는 전체 데이터를 크기 순으로 정렬했을 경우 하위 1%에 해당.

– Percentile은 하위 몇 퍼센트인지를 표현하기 때문에 수학적으로 표현하자면:

$$-\infty \leq Z \leq z_x \rightarrow P[-\infty \leq Z \leq z_x] = x/100 \text{ percentile}$$

• 포인트: 확률분포에서 주어진 Percentile 값에 해당하는 z 값이 존재

7. Residual Check

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

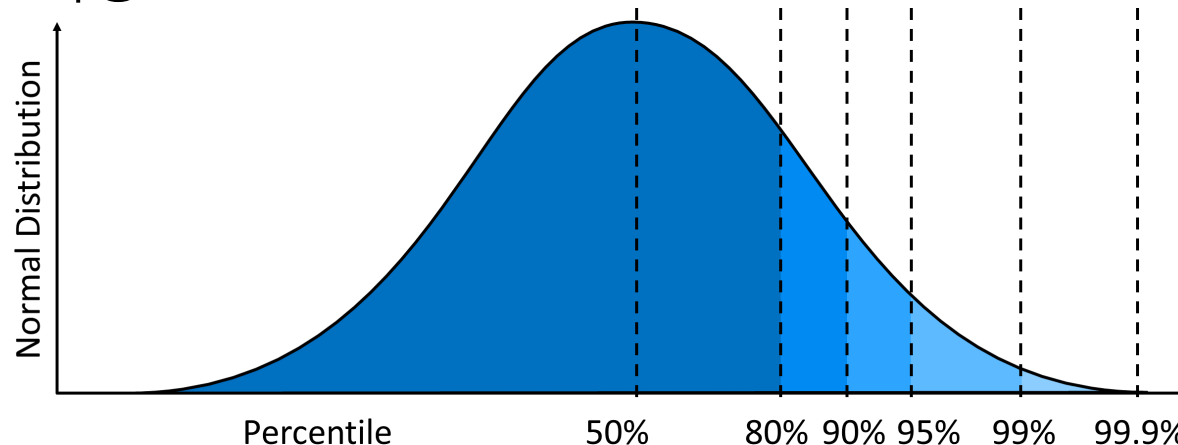
3) 정규성: 잔차의 분포는 정규분포를 따라야한다

▣ 정규분포의 모양을 수치적으로 표현하는 방법 중 하나는 표준정규분포표를 거꾸로 적용하는 방법.

– Percentile을 표준정규분포표에서 면적으로 생각하면 다음과 같다.

• 50 Percentile $\rightarrow -\infty \leq Z \leq 0$. 일반화하면 x Percentile $\rightarrow -\infty \leq Z \leq z_x$

– 이런 관계를 줄여서 표준정규분포는 x Percentile $\rightarrow z_x$ 관계식으로 표현 가능하며, 잔차의 분포 또한 이런 관계식으로 표현가능. 이 둘의 관계식을 비교하면 잔차의 분포가 정규분포에 얼마나 가까운지 비교가능.



x_1 Percentile $\rightarrow z_{x_1}$
 x_2 Percentile $\rightarrow z_{x_2}$
 x_3 Percentile $\rightarrow z_{x_3}$
...
 x_n Percentile $\rightarrow z_{x_n}$

7. Residual Check

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

3) 정규성: 잔차의 분포는 정규분포를 따라야한다

▣ 정규분포의 모양을 수치적으로 표현하는 방법 중 하나는 표준정규분포표를 거꾸로 적용하는 방법

– 데이터를 잔차 크기에 따라 정렬해서 각 데이터별 Percentile을 구하고, 해당 Percentile에 해당하는 표준정규분포표상 z 값과, 표준화 된 잔차 (= (잔차-잔차의 평균)/(잔차의 표준편차), 엑셀 기능으로 제공) 와 비교

잔차 크기로 오름차 순 정렬 데이터 번호	Percentile	Z (표준정규분포)	표준화 된 잔차
1	1.5625	-2.1538747	-1.663705
2	4.6875	-1.6759397	-1.6249574
3	7.8125	-1.4177971	-1.4072027
4	10.9375	-1.2298588	-1.0558244
5	14.0625	-1.0775156	-1.0418334
...



7. Residual Check

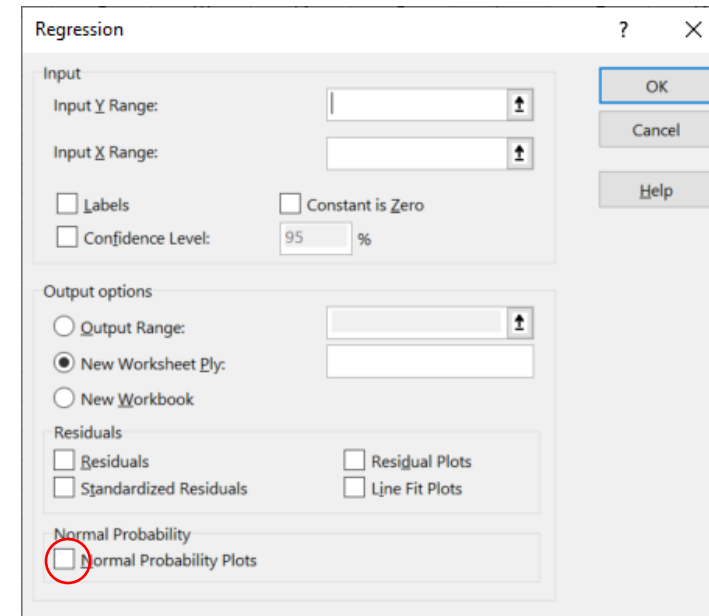
워크시트: Regression - 1dist,am Pt1 Resid
셀 주소: F23:G57

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

3) 정규성: 잔차의 분포는 정규분포를 따라야한다

□ 표준정규분포와 잔차를 x Percentile $\rightarrow z_x$ 관계식으로 표현하는 법:

– Percentile: 엑셀 회귀 분석 기능에서 Normal Probability Plot 출력을 선택했을 경우 잔차의 Percentile이 회귀분석 결과에 나옴 (우측 그림)



The image shows the 'Regression' dialog box in Excel. The 'Input' section has 'Input Y Range' and 'Input X Range' fields. The 'Output options' section has 'Output Range', 'New Worksheet Ply', and 'New Workbook' options. The 'Residuals' section has 'Residuals', 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots' options. The 'Normal Probability' section has a checkbox for 'Normal Probability Plots' which is checked and circled in red. The 'OK', 'Cancel', and 'Help' buttons are on the right.



PROBABILITY OUTPUT	
Percentile	mpg
1.5625	10.4
4.6875	10.4
7.8125	13.3
10.9375	14.3
14.0625	14.7
17.1875	15
20.3125	15.2
23.4375	15.2
26.5625	15.5
29.6875	15.8
32.8125	16.4
35.9375	17.3
39.0625	17.8
42.1875	18.1
45.3125	18.7
48.4375	19.2
51.5625	19.2
54.6875	19.7
57.8125	21
60.9375	21
64.0625	21.4
67.1875	21.4
70.3125	21.5
73.4375	22.8
76.5625	22.8
79.6875	24.4
82.8125	26
85.9375	27.3
89.0625	30.4
92.1875	30.4
95.3125	32.4
98.4375	33.9

7. Residual Check

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

3) 정규성: 잔차의 분포는 정규분포를 따라야한다

▣ 표준정규분포와 잔차를 x Percentile $\rightarrow z_x$ 관계식으로 표현하는 법:

- z_x : 정규분포에서 x Percentile에 따른 z 값. 엑셀함수 “=Norm.S.INV(Percentile)” 활용. 단, 엑셀은 Percentile을 0~1 사이값만 허용하기 때문에 우측에 나온 Percentile 값을 100으로 나눠주고, 이 값을 Norm.S.INV 함수에 적용

PROBABILITY OUTPUT		Apply "=Norm.S.INV(Percentile/100)"	
Divide by 100			
Percentile	mpg	Percentile/100	Z statistics
1.5625	10.4	0.015625	-2.153875
4.6875	10.4	0.046875	-1.67594
7.8125	13.3	0.078125	-1.417797
10.9375	14.3	0.109375	-1.229859
14.0625	14.7	0.140625	-1.077516
17.1875	15	0.171875	-0.946782
20.3125	15.2	0.203125	-0.830511
23.4375	15.2	0.234375	-0.724514
26.5625	15.5	0.265625	-0.626099
29.6875	15.8	0.296875	-0.53341
32.8125	16.4	0.328125	-0.445097
35.9375	17.3	0.359375	-0.36013
39.0625	17.8	0.390625	-0.27769
42.1875	18.1	0.421875	-0.197099
45.3125	18.7	0.453125	-0.11777
48.4375	19.2	0.484375	-0.039176
51.5625	19.2	0.515625	0.039176
54.6875	19.7	0.546875	0.11777
57.8125	21	0.578125	0.197099
60.9375	21	0.609375	0.27769
64.0625	21.4	0.640625	0.36013
67.1875	21.4	0.671875	0.445097
70.3125	21.5	0.703125	0.53341
73.4375	22.8	0.734375	0.626099
76.5625	22.8	0.765625	0.724514
79.6875	24.4	0.796875	0.830511
82.8125	26	0.828125	0.946782
85.9375	27.3	0.859375	1.077516
89.0625	30.4	0.890625	1.229859
92.1875	30.4	0.921875	1.417797
95.3125	32.4	0.953125	1.67594
98.4375	33.9	0.984375	2.153875

7. Residual Check

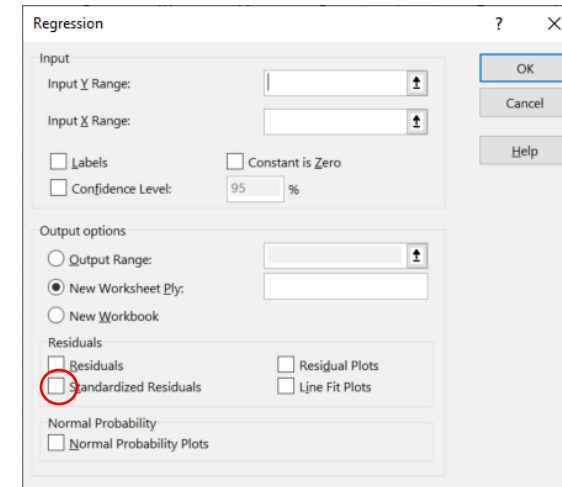
워크시트: Regression - 1dist,am Pt1 Resid
셀 주소: A23:E57

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

3) 정규성: 잔차의 분포는 정규분포를 따라야한다

▫ 표준정규분포와 잔차를 x Percentile $\rightarrow z_x$ 관계식으로 표현하는 법:

- 표준 잔차: 엑셀 회귀 분석 기능에서 Standardized Residual 출력을 선택했을 경우 표준화된 잔차가 회귀분석 결과에 나옴 (우측 그림)
- 단, 이 표준잔차는 잔차의 크기에 따라 정렬되지 않았으므로, 오름차순(ascending order)로 정렬이 필요



RESIDUAL OUTPUT			
Observation	Predicted mpg	Residuals	Standard Residuals
1	20.340686	0.65931403	0.292309
2	20.340686	0.65931403	0.292309
3	25.0992193	-2.2992193	-1.0193663
4	16.8440061	4.55599387	2.01991457
5	15.1074394	3.59256061	1.59277333
6	17.7429348	0.3570652	0.15830601
7	15.1074394	-0.8074394	-0.3579809
8	21.4940694	2.90593063	1.28835371
9	21.9457513	0.85424866	0.3787339
10	20.1498948	-0.9498948	-0.4211389
11	20.1498948	-2.3498948	-1.0418334
12	16.4484396	-0.0484396	-0.0214758
13	16.4484396	0.85156041	0.37754205
14	16.4484396	-1.2484396	-0.5534997
15	14.0651531	-3.6651531	-1.6249574
16	14.1525498	-3.7525498	-1.663705
17	14.3088044	0.39119558	0.17343782
18	30.5503005	1.84969951	0.82007024
19	31.3465775	-0.9465775	-0.4196682
20	32.6980458	1.20195423	0.53289028
21	23.881452	-2.381452	-1.0558244
22	15.6875799	-0.1875799	-0.0831641
23	15.9165827	-0.7165827	-0.3176993
24	15.2329392	-1.9329392	-0.8569748
25	14.6681902	4.53180984	2.0091925
26	30.473999	-3.173999	-1.4072027
27	23.6021937	2.39780631	1.06307516
28	27.0853095	3.31469053	1.46957873
29	14.9626985	0.83730146	0.3712203
30	21.3630764	-1.6630764	-0.7373303
31	15.7110574	-0.7110574	-0.3152496
32	23.5261505	-2.1261505	-0.9426357

7. Residual Check

워크시트: Regression - 1dist,am Pt1 Resid
셀 주소: A23:E57

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

3) 정규성: 잔차의 분포는 정규분포를 따라야한다

▫ 표준정규분포와 잔차를 x Percentile $\rightarrow z_x$ 관계식으로 표현하는 법:

– 정렬 방법:

- 1) 표준화된 잔차 컬럼을 Copy & Paste로 복제하고, 이렇게 복제한 데이터를 선택
- 2) 선택한 데이터에 엑셀에서 필터 적용. (필터 기능은 선택된 데이터들을 오름차순/내림차순으로 정렬 하는 기능을 제공)
- 3) 필터 기능을 통해 표준화 된 잔차를 오름차순(Ascending)으로 정렬

The image shows an Excel spreadsheet with a table titled 'RESIDUAL OUTPUT'. The table has five columns: 'Observation', 'Predicted mpg', 'Residuals', 'Standard Residuals', and 'Standard Residuals'. The data is sorted by 'Standard Residuals' in ascending order. A red box highlights the 'Filter' button in the 'Sort & Filter' task pane. Below the task pane, a 'Sorted Standard Residuals' dialog box is open, showing 'Ascending' selected under the 'Sort' section. The 'Filter' section is also visible, with 'By color' set to 'None'.

Observation	Predicted mpg	Residuals	Standard Residuals	Standard Residuals
340686	0.65931403	0.292309	-1.663705	
340686	0.65931403	0.292309	-1.6249574	
992193	-2.2992193	-1.0193663	-1.4072027	
440061	4.55599387	2.01991457	-1.0558244	
074394	3.59256061	1.59277333	-1.0418334	
429348	0.3570652	0.15830601	-1.0193663	
074394	-0.8074394	-0.3579809	-0.9426357	
940694	2.90593063	1.28835371	-0.8569748	
457513	0.85424866	0.3787339	-0.7373303	
498948	-0.9498948	-0.4211389	-0.5534997	
498948	-2.3498948	-1.0418334	-0.4211389	
484396	-0.0484396	-0.0214758	-0.4196682	
484396	0.85156041	0.37754205	-0.3579809	
484396	-1.2484396	-0.5534997	-0.3176993	
651531	-3.6651531	-1.6249574	-0.3152496	
525498	-3.7525498	-1.663705	-0.0831641	
17	14.3088044	0.39119558	0.17343782	-0.0214758
18	30.5503005	1.84969951	0.82007024	0.15830601
75	-0.9465775	-0.4196682	0.17343782	
58	1.20195423	0.53289028	0.292309	
52	-2.381452	-1.0558244	0.292309	
99	-0.1875799	-0.0831641	0.3712203	
27	-0.7165827	-0.3176993	0.37754205	
92	-1.9329392	-0.8569748	0.3787339	
02	4.53180984	2.0091925	0.53289028	
99	-3.173999	-1.4072027	0.82007024	
37	2.39780631	1.06307516	1.06307516	
95	3.31469053	1.46957873	1.28835371	
85	0.83730146	0.3712203	1.46957873	
64	-1.6630764	-0.7373303	1.59277333	
74	-0.7110574	-0.3152496	2.0091925	
32	23.5261505	-2.1261505	-0.9426357	2.01991457

7. Residual Check

워크시트: Regression - 1dist,am Pt1 Resid
셀 주소: L25:N57, P36:V53

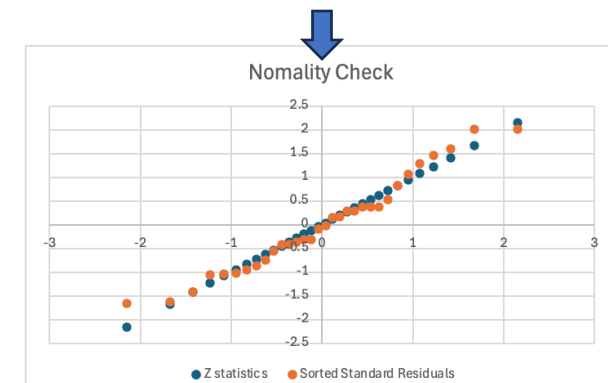
□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

3) 정규성: 잔차의 분포는 정규분포를 따라야한다

▫ 정규분포와 잔차 분포의 시각적 비교를 위한 Scatter Plot을 만드는 법

- Z statistics는 복사해서 두 개의 열에 같은 값을 나란히 붙여넣기 (이때 선택하여 붙여넣기 기능을 통해 값만 붙여넣도록 함)
- 그리고 그 옆에 앞서 정렬한 표준 잔차를 복사해서 붙여넣기
- 이렇게 3개의 열로 이뤄진 데이터를 전체 선택하고, 삽입 (Insert) - 분산도(Scatter Plot)을 선택하면 시각적 비교를 위한 Scatter Plot이 만들어짐

Z statistics	Z statistics	Standard Res
-2.1538747	-2.1538747	-1.663705
-1.6759397	-1.6759397	-1.6249574
-1.4177971	-1.4177971	-1.4072027
-1.2298588	-1.2298588	-1.0558244
-1.0775156	-1.0775156	-1.0418334
-0.9467818	-0.9467818	-1.0193663
-0.8305109	-0.8305109	-0.9426357
-0.7245144	-0.7245144	-0.8569748
-0.626099	-0.626099	-0.7373303
-0.5334097	-0.5334097	-0.5534997
-0.4450965	-0.4450965	-0.4211389
-0.3601299	-0.3601299	-0.4196682
-0.2776904	-0.2776904	-0.3579809
-0.1970991	-0.1970991	-0.3176993
-0.1177699	-0.1177699	-0.3152496
-0.0391761	-0.0391761	-0.0831641
0.03917609	0.03917609	-0.0214758
0.11776987	0.11776987	0.15830601
0.19709908	0.19709908	0.17343782
0.27769044	0.27769044	0.292309



7. Residual Check

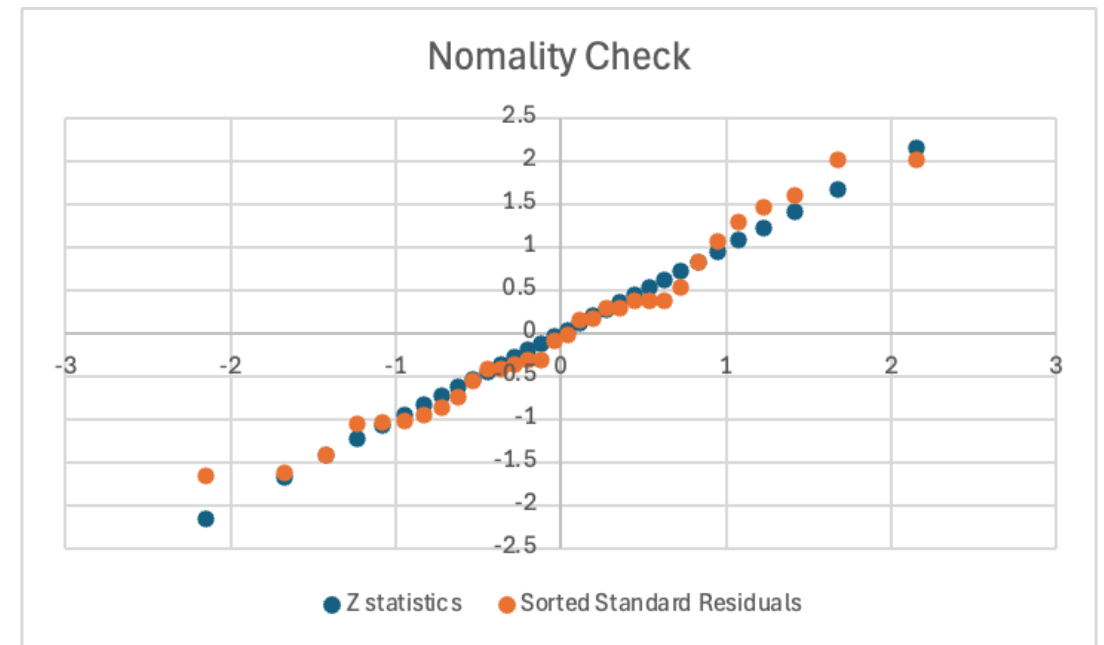
워크시트: Regression - 1dist,am Pt1 Resid
셀 주소: P36:V53

□ 7. 잔차 분석: 잔차와 관련해서 검증이 요구되는 전제 조건은 다음과 같다:

3) 정규성: 잔차의 분포는 정규분포를 따라야한다

▫ 이렇게 그린 Scatter Plot을 통해 같은 Percentile 에서 정규 분포의 Z statistics와 표준화된 잔차를 시각적으로 비교가능. 이 두 분포가 많이 떨어져 있으면 잔차의 분포가 정규 분포를 따르지 않는 것

Z (표준정규분포)	표준화 된 잔차
-2.1538747	-1.663705
-1.6759397	-1.6249574
-1.4177971	-1.4072027
-1.2298588	-1.0558244
-1.0775156	-1.0418334
...	...



8. ANOVA Analysis: F-test

□8. 분산 분석 (Analysis of Variance – ANOVA)

- 원래 분산분석은 집단 간 평균 차이가 통계적으로 유의미한가를 검정하는 과정. 하지만 우리는 이 기법을 우리가 제안한 회귀 모형과 아무런 의미가 없는 모형과 다른지 검정하는데 활용:

$$\text{모형1: } y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n$$

$$\text{모형2: } y = 0 + 0 \cdot x_1 + \cdots + 0 \cdot x_n$$

- 만일 검정결과 모형 1과 모형 2가 통계적으로 유의미하게 다르지 않다면 모형1의 β_0, \dots, β_n 를 전부 0으로 간주해야하며, 회귀분석의 결과로 나온 β_0, \dots, β_n 값들을 신뢰할 수 없게 된다

8. ANOVA Analysis: F-test

□8. 분산 분석 (Analysis of Variance – ANOVA)

$$\text{모형1: } y = \beta_0 + \beta_1 \cdot x_1 + \cdots + \beta_n \cdot x_n$$

$$\text{모형2: } y = 0 + 0 \cdot x_1 + \cdots + 0 \cdot x_n$$

– How?

- 두 모형으로 부터 계산되는 y 값 평균의 차이는 전체 데이터의 변동성을 회귀 모형이 설명하는 분산 (집단 1 - SSR)과 그렇지 못한 잔차의 분산(집단 2- SSE)의 차이로 연결됨
- 따라서 모형의 분산과 잔차의 분산의 비율을 F-test를 통해 비교함으로써 우리가 제안한 회귀모형이 의미 없는 모형과 다른지 검정하는 방법

H_0 (귀무가설): 모형1과 모형2는 같다 $\rightarrow \beta_0 = \beta_1 = \cdots = \beta_n = 0$

H_0 (대안가설): 모형1과 모형2는 다르다 \rightarrow 최소한 한개 이상의 회귀계수 $\beta \neq 0$

8. ANOVA Analysis: F-test

워크시트: Regression - 1dist,am Pt2 Ftest
셀 주소: A10:F14

□ 8. 분산 분석 (Analysis of Variance – ANOVA)

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	968.3362	484.1681	89.0291481	4.1824E-13
Residual	29	157.710988	5.43830993		
Total	31	1126.04719			

□ 자유도:

- Degree of Freedom for Regression: $k = \# \text{ of Independent Variable} = 2$
- Degree of Freedom for Residual: $df = n - k - 1 = 29$
- Degree of Freedom for Total: $n - 1 = 31$

8. ANOVA Analysis: F-test

워크시트: Regression - 1dist,am Pt2 Ftest
셀 주소: A10:F14

□8. 분산 분석 (Analysis of Variance – ANOVA)

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	968.3362	484.1681	89.0291481	4.1824E-13
Residual	29	157.710988	5.43830993		
Total	31	1126.04719			

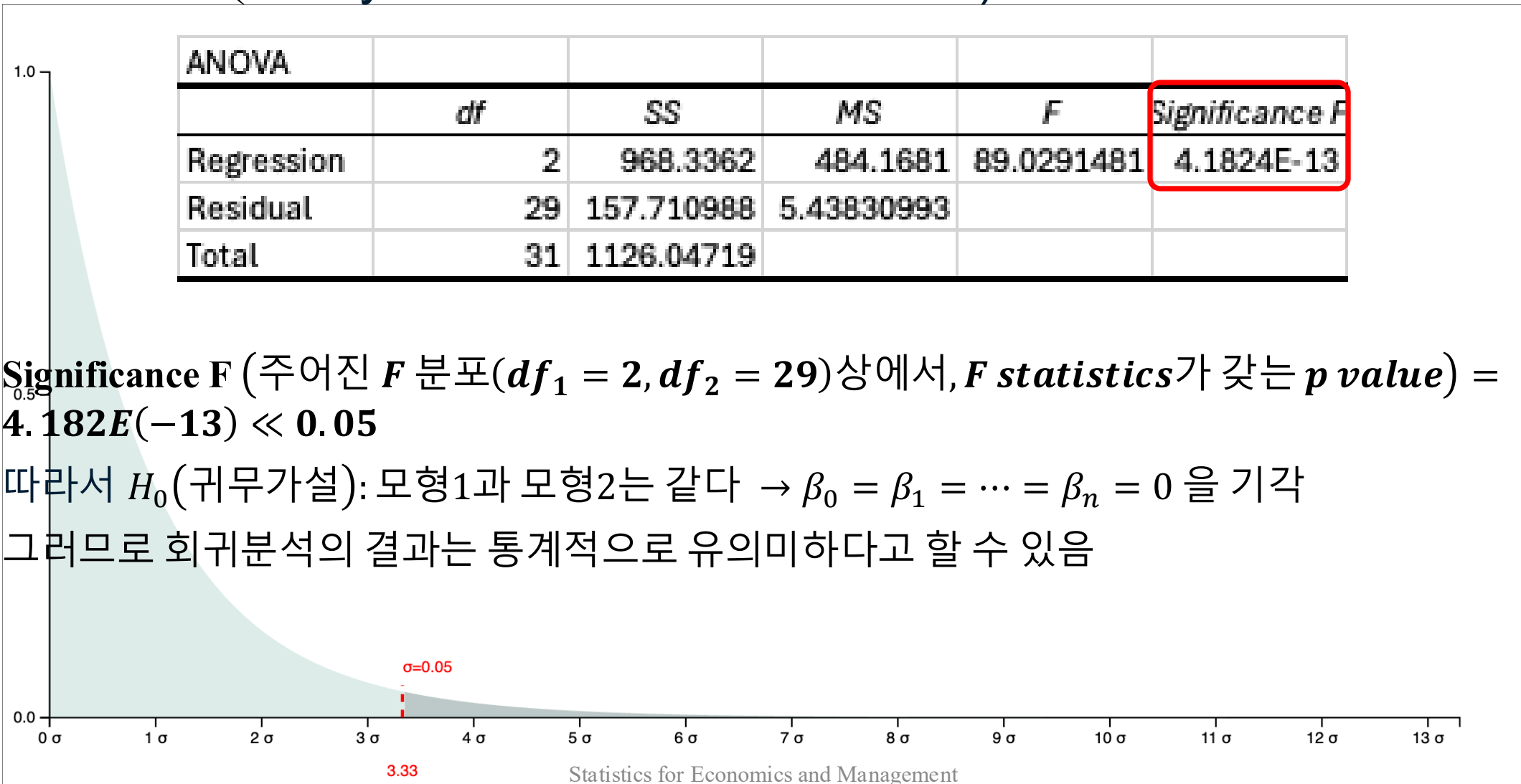
- MSR (회귀분석이 설명한 변동성 분산의 평균) = $\frac{SSR}{k} = \frac{968.3361}{2} = 484.1681$
- MSE (회귀분석이 설명하지 못한 잔차 분산의 평균) = $\frac{SSE}{n-k-1} = \frac{157.7109}{29} = 5.4383$
- 제안한 회귀분석 모형이 설명한 변동성 분산 대비 오차 분산 비율:
(클수록 회귀분석이 설명한 분산의 비율이 커지며, 모형이 유의미할 확률이 커짐)

$$F \text{ statistics} = \frac{MSR}{MSE} = \frac{484.1681}{5.4383} = 89.029148$$

8. ANOVA Analysis: F-test

워크시트: Regression - 1dist,am Pt2 Ftest
셀 주소: A10:F14

□8. 분산 분석 (Analysis of Variance – ANOVA)



9. Regression Coefficient Analysis

□9. 회귀계수 분석: t-test

- 모형 전체에 대한 검정을 통과했을 경우, 다음 단계로는 각각의 회귀 계수가 통계적으로 유의미한지 검정할 차례

- 회귀 계수에 대한 가설검정은 다음과 같이 진행:

$$H_0(\text{귀무가설}): \text{회귀 계수 } \beta_i = 0$$

$$H_1(\text{대안가설}): \text{회귀 계수 } \beta_i \neq 0$$

- 검정방법은 $t \text{ statistics} = \left(\frac{\text{Coefficients} - 0}{\text{standard error}} \right)$ 를 활용한 t-test를 진행 : 해당 t statistics의 p-value가 유의수준(0.05)를 넘기는지 점검

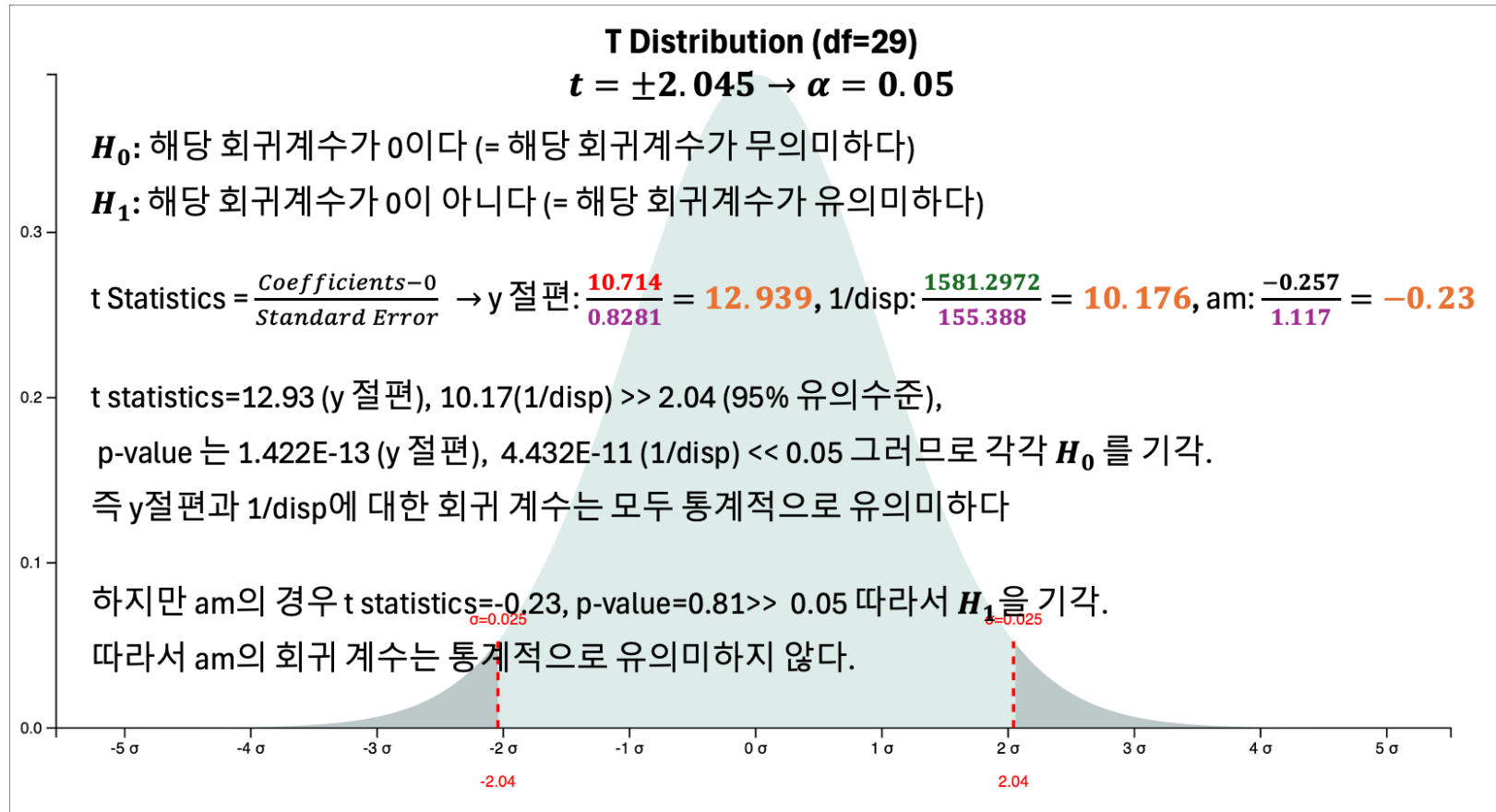
- t statistics는 표준화($= \frac{\text{값} - \text{평균}}{\text{표준편차}}$)된 통계량으로, 분모의 0은 가설에서 제시한 회귀계수 값인 0, $\text{standard error} = \beta_i$ 의 표준편차

9. Regression Coefficient Analysis

워크시트: Regression - 1dist,am Pt3 Ttest
셀 주소: A16:I19

□9. 회귀계수 분석: t-test

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	10.714947	0.82811102	12.9390224	1.4221E-13	9.021269841	12.4086242	8.4323535	12.99754059
1/dis	1581.29724	155.388801	10.1763913	4.4321E-11	1263.491463	1899.10303	1152.98574	2009.608746
am	-0.2573689	1.11716681	-0.2303764	0.81941674	-2.542231536	2.02749382	-3.3367117	2.821973997



9. Regression Coefficient Analysis

워크시트: Regression - 1dist,am Pt3 Ttest
셀 주소: A16:I19

□9. 회귀계수 분석: t-test

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	10.714947	0.82811102	12.9390224	1.4221E-13	9.021269841	12.4086242	8.4323535	12.99754059
1/disp	1581.29724	155.388801	10.1763913	4.4321E-11	1263.491463	1899.10303	1152.98574	2009.608746
am	-0.2573689	1.11716681	-0.2303764	0.81941674	-2.542231536	2.02749382	-3.3367117	2.821973997

□ 결론

Estimated Regression Equation: $\text{mpg} = 10.714 + 1581.297 \cdot \left(\frac{1}{\text{disp}}\right) - 0.257 \cdot (\text{am})$.

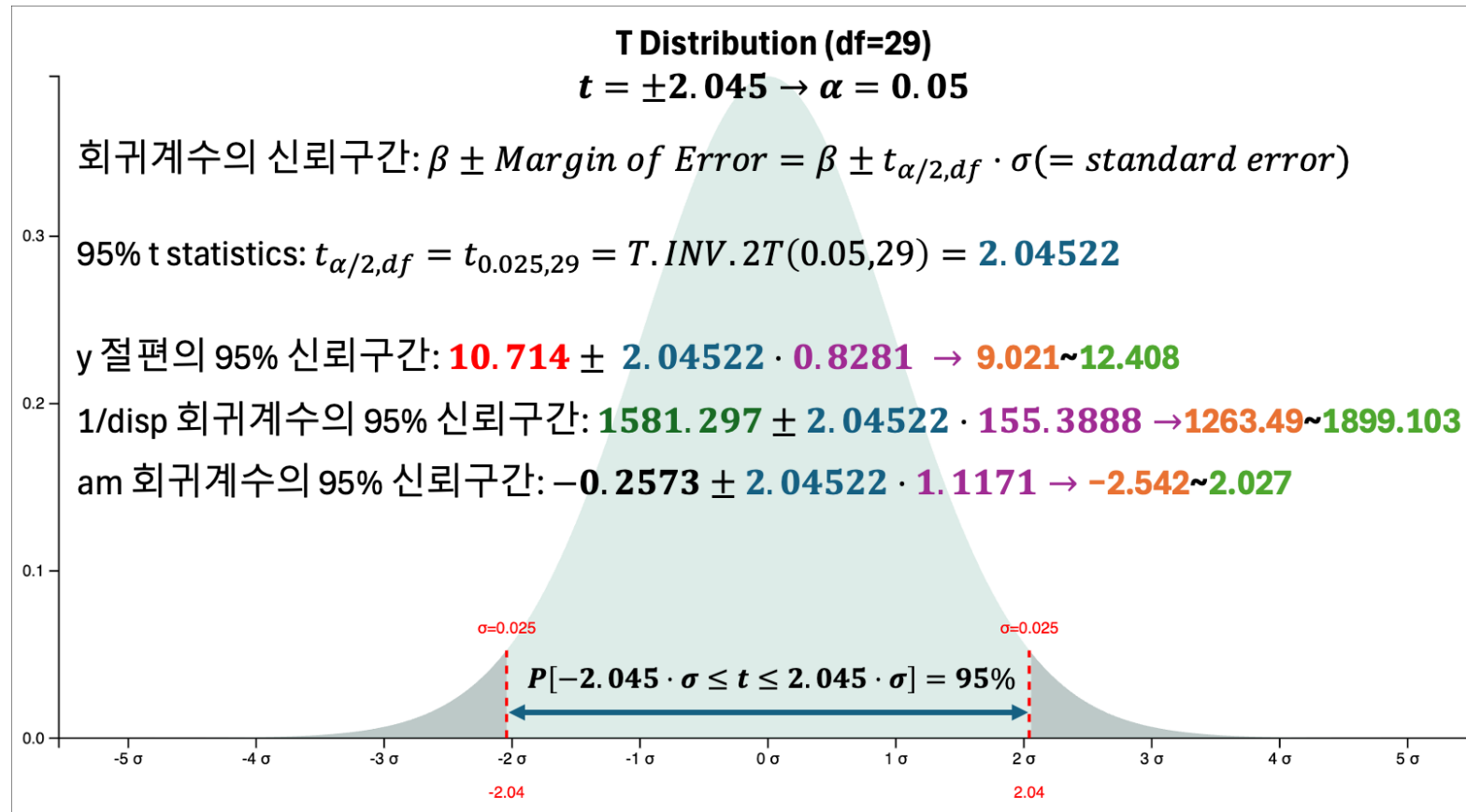
하지만 t-test 결과 am의 회귀계수가 유효하지 않았음

9. Regression Coefficient Analysis

워크시트: Regression - 1dist,am Pt4 C.I.
셀 주소: A16:I19

□9. 회귀계수 분석: Confidence Interval (신뢰구간)

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	10.714947	0.82811102	12.9390224	1.4221E-13	9.021269841	12.4086242	8.4323535	12.99754059
1/disp	1581.29724	155.388801	10.1763913	4.4321E-11	1263.491463	1899.10303	1152.98574	2009.608746
am	-0.2573689	1.11716681	-0.2303764	0.81941674	-2.542231536	2.02749382	-3.3367117	2.821973997



10. Evaluation & Interpretation

워크시트: Regression - 1dist,am Pt5 Eval
셀 주소: A3:B8

□10. 평가 및 해석

$$R^2 = 1 - \frac{SSE}{SST} = \frac{SSR}{SST} = \frac{968.047}{1126.047} = 0.8599$$

$$Adj R^2 = 1 - \frac{SSE/n - k - 1}{SST/(n - 1)} = 1 - \frac{157.710/29}{1126.047/31} = 0.8502$$

Standard Deviation of Residuals: $s = \sqrt{MSE} = \sqrt{\frac{SSE}{df}} = \sqrt{5.438} = 2.322$

Regression Statistics					
Multiple R	0.92733102				
R Square	0.85994283	R^2			
Adjusted R Square	0.85028371	$Adj R^2$			
Standard Error	2.33201842	s			
Observations	32				
ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	2	968.3362	484.1681	89.0291481	4.18242E-13
Residual	29	157.710988	5.43830993		
Total	31	1126.04719			

다중회귀모형에서는 Adjusted R squared를 평가 기준으로 활용한다.

평가: 이 회귀 모형은 전체 데이터가 가진 변동성의 85.02%를 설명한다.

10. Evaluation & Interpretation

워크시트: Regression - 1dist,am Pt5 Eval
셀 주소: A16:I19

□10. 평가 및 해석

	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 99.0%	Upper 99.0%
Intercept	10.714947	0.82811102	12.9390224	1.4221E-13	9.02126984	12.4086242	8.4323535	12.99754059
1/disp	1581.29724	155.388801	10.1763913	4.4321E-11	1263.49146	1899.10303	1152.98574	2009.608746
am	-0.2573689	1.11716681	-0.2303764	0.81941674	-2.5422315	2.02749382	-3.3367117	2.821973997

Estimated Regression Equation: $\text{mpg} = 10.714 + 1581.297 \cdot \left(\frac{1}{\text{disp}}\right) - 0.257 \cdot (\text{am})$

하지만 t-test 결과 am의 회귀계수가 유효하지 않았음

해석: 자동차의 연비에 있어서 (1/배기량)은 유의미한 영향을 미치는 것이 확인됨.

기본 연비 10.714 mile per gallon을 기준으로 (1/배기량)이 1씩 커질때마다

연비는 1581.297 mile per gallon 증가함. 단 (1/배기량) 증가할때 연비가 증가하는 것이므로, 실제 배기량은 작아져야 연비가 증가하는 반비례 관계를 확인할 수 있음.

Note: am의 회귀계수는 통계적으로 유효하지 않기 때문에 이와 같은 종속변수에 미치는 영향력의 방향성과 크기에 대한 통계적 해석은 수행할 수 없음. 만약 am의 회귀계수가 유효했다면, am=1, 즉 자동변속기가 적용되었을 경우 연비가 0.257 mile per gallon 감소하는 것으로 해석.

10. Evaluation & Interpretation

워크시트: Regression - 1dist,gear
셀 주소:

□10. 평가 및 해석

Regression Statistics																	
Multiple R	0.8505	통제변수로 Gear를 Ratio scale로 간주하여 회귀분석을 시도한 모형의 경우:															
R Square	0.7233	잔차 분석 결과 잔차에 요구되는 모든 전제를 만족시키는 것으로 판명															
Adjusted R Square	0.7023	F-test 결과 p-value는 3.97E-13으로 전체 모형은 유의미한 것으로 판명															
Standard Error	2.32800801																
Observations	32	회귀분석 결과는 다음과 같음:															
Estimated Regression Equation: $\text{mpg} = 9.966 + 1531.85 \cdot \left(\frac{1}{\text{disp}}\right) + 0.254 \cdot (\text{gear})$.																	
ANOVA	하지만 t-test 결과 gear의 회귀계수는 통계적으로 유효하지 않았음																
	df	SS	MS	F	Significance F												
Regression	2	157.118817	78.559408	15.474175	3.97E-13	해석: 자동차의 연비에 있어서 (1/배기량)은 유의미한 영향을 미치는 것이 확인됨.											
Residual	29	137.118817	4.728235			기본 연비 9.966 mile per gallon을 기준으로 (1/배기량)이 1씩 커질때마다											
Total	31	294.237634				연비는 1531.85 mile per gallon 증가함. 단 (1/배기량) 증가할 때 연비가 증가하는 것이므로, 실제 배기량은 작아져야 연비가 증가하는 반비례 관계를 확인할 수 있음.											
평가: 이 회귀 모형은 전체 데이터가 가진 변동성의 85.07%를 설명한다.																	
Intercept	9.96680669	2.1634713	4.60685876	7.55483E-05	5.54201107	14.3916023	4.00344491	15.9301685									
gear	0.25491799	0.65116196	0.3914817	0.69830304	-1.0768577	1.58669372	-1.5399356	2.04977162									
1/disp	1531.8536	133.918049	11.4387329	15.474175	1257.86043	1805.74676	1162.72378	1900.98342									

Statistics for Economics and Management

10. Evaluation & Interpretation

워크시트: Regression - 1dist,gear(ctgr)
셀 주소:

□10. 평가 및 해석

통제변수로 Gear를 Nominal scale로 간주하여 Categorical variable로 간주하여 회귀분석을 시도한 모형의 경우:
잔차 분석 결과 잔차에 요구되는 모든 전제를 만족시키는 것으로 판명
F-test 결과 p-value는 3.21E-13으로 전체 모형은 유의미한 것으로 판명

회귀분석 결과는 다음과 같음:

Estimated Regression Equation: $\text{mpg} = 10.576 + 1612.58 \cdot \left(\frac{1}{\text{disp}}\right) - 0.64 \cdot (\text{Gear} = 4) + 0.51 \cdot (\text{Gear} = 5)$

하지만 t-test 결과 Gear=4, Gear=5의 회귀계수는 통계적으로 유효하지 않았음

해석: 자동차의 연비에 있어서 (1/배기량)은 유의미한 영향을 미치는 것이 확인됨.
기본 연비 10.576 mile per gallon을 기준으로 (1/배기량)이 1씩 커질때마다
연비는 1612.58 mile per gallon 증가함. 단 (1/배기량) 증가할때 연비가 증가하는 것이므로,
실제 배기량은 작아져야 연비가 증가하는 반비례 관계를 확인할 수 있음.

평가: 이 회귀 모형은 전체 데이터가 가진 변동성의 84.86%를 설명한다.

Note: 만약 Gear=4, Gear=5의 회귀계수는 통계적으로 유효했다면, 자동차의 연비는 Gear = 3 일 때를 기준으로
Gear = 4 면 연비가 0.64 감소, Gear = 5 면 0.51 증가한다고 해석할 수 있음

11. Conclusion

□ 11. 결론

지금까지 우리는 배기량이 연비에 반비례하는 영향을 미치는지 판단하기 위해 3가지 모델을 회귀분석에 적용했음.
다음의 3가지 모델 모두 회귀분석에서 통계적으로 요구되는 모든 조건들을 만족시킴:

1) Estimated Regression Equation: $\text{mpg} = 10.714 + 1581.297 \cdot \left(\frac{1}{\text{disp}}\right) - 0.257 \cdot (\text{am})$

하지만 t-test 결과 am의 회귀계수가 통계적으로 유효하지 않았음
이 회귀 모형은 전체 데이터가 가진 변동성의 85.02%를 설명

2) Estimated Regression Equation: $\text{mpg} = 9.966 + 1531.85 \cdot \left(\frac{1}{\text{disp}}\right) + 0.254 \cdot (\text{gear})$

하지만 t-test 결과 gear의 회귀계수는 통계적으로 유효하지 않았음
이 회귀 모형은 전체 데이터가 가진 변동성의 85.07%를 설명

3) Estimated Regression Equation: $\text{mpg} = 10.576 + 1612.58 \cdot \left(\frac{1}{\text{disp}}\right) - 0.64 \cdot (\text{Gear} = 4) + 0.51 \cdot (\text{Gear} = 5)$

하지만 t-test 결과 Gear=4, Gear=5의 회귀계수는 통계적으로 유효하지 않았음
이 회귀 모형은 전체 데이터가 가진 변동성의 84.86%를 설명

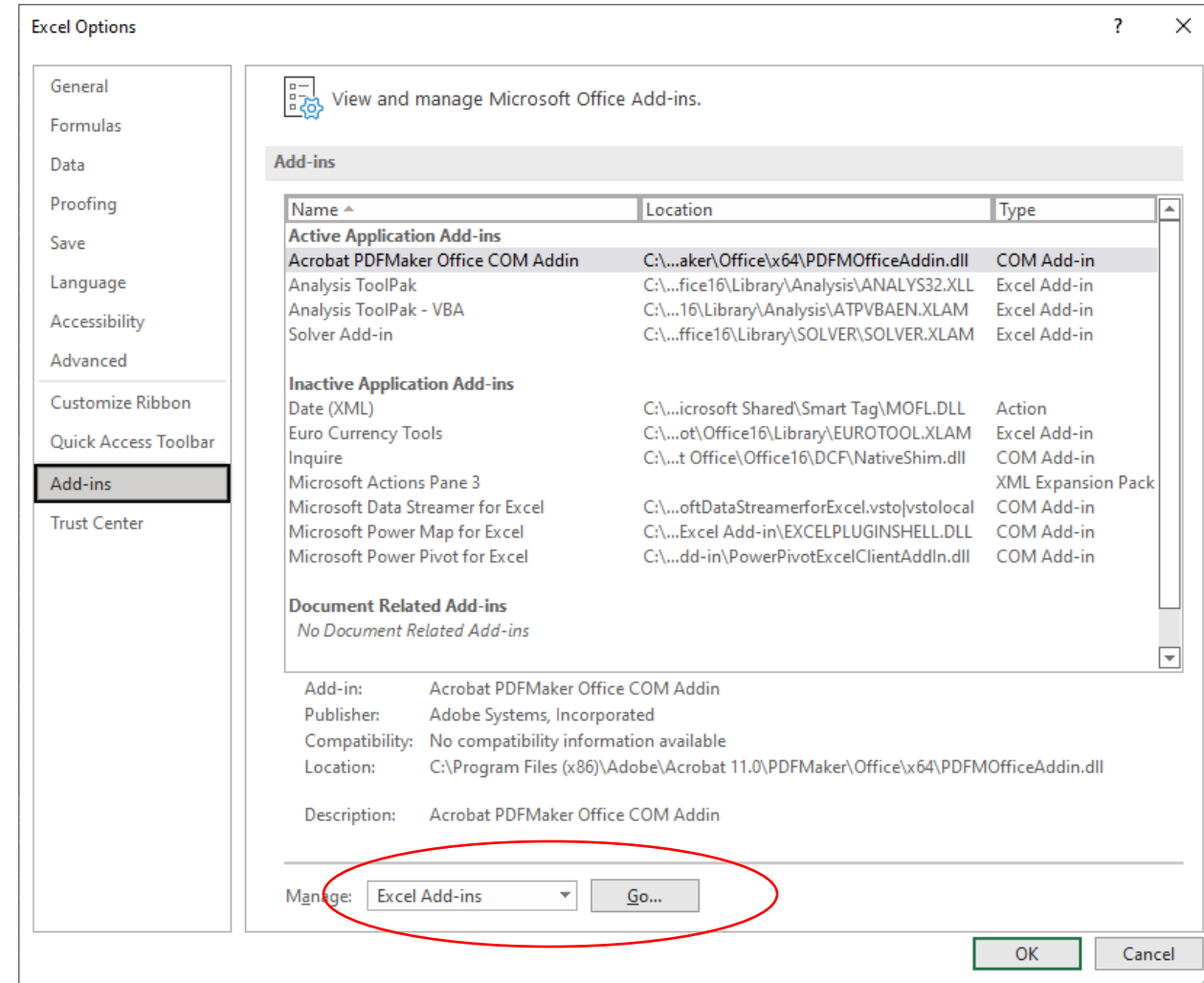
세 모형 모두 배기량이 연비에 반비례하는 영향이 통계적으로 유의미함을 입증함

Adjusted R squared 값에 기반하여 세 모형의 성능을 평가했을 때, 2)번 모형이 가장 높은 설명력을 지닌 것으로 평가됨

부록

How to Run Regression Analysis with Excel

- Step 0) Activate Data Analysis Tools:
 - Click File menu of Excel
 - Go to Options
 - Go to “Add-ins” Tab
 - Manage: “Excel Add-ins”, click “Go...”
 - Activate “Analysis ToolPak”, “Analysis ToolPak - VBA”, “Solver Add-in (Will be used after 1st exam)”, click OK



How to Run Regression Analysis with Excel

□ Step 2) In the regression window:

▫ Input

- Input Y range: Dependent Variable Column
- Input X range: All independent variables columns
- Labels if input range contains variable names
- Confidence level and set 95%

Check Label if your input Y/X range contains variable name (i.e., B1, C1, D1)

	A	B	C	D
1	Obs	y	x1	x2
2	1	9.95	2	50
3	2	24.45	8	110
4	3	31.75	11	120
5	4	35	10	550
6	5	25.02	8	295
7	6	16.86	4	200
8	7	14.38	2	375
9	8	9.6	2	52
10	9	24.35	9	100
11	10	27.5	8	300

How to Run Regression Analysis with Excel

□ Step 2) In the regression window:

▫ Output Options - For residuals, checks:

- Residuals
- Standardized Residuals
- Residual Plots
- Line Fit Plots

□ Note: You can try any output option you want. Try each option and see the result

The screenshot shows the 'Regression' dialog box in Excel. The 'Input' section has 'Input Y Range' and 'Input X Range' text boxes. Below these are checkboxes for 'Labels' and 'Confidence Level' (set to 95%), and a 'Constant is Zero' checkbox. The 'Output options' section has radio buttons for 'Output Range', 'New Worksheet Ply.' (selected), and 'New Workbook'. The 'Residuals' section has checkboxes for 'Residuals', 'Standardized Residuals', 'Residual Plots', and 'Line Fit Plots'. The 'Normal Probability' section has a checkbox for 'Normal Probability Plots'. Several checkboxes are circled in red: 'Labels', 'Confidence Level', 'Residuals', 'Standardized Residuals', 'Residual Plots', 'Line Fit Plots', and 'Normal Probability Plots'. The 'OK' button is highlighted with a blue border.

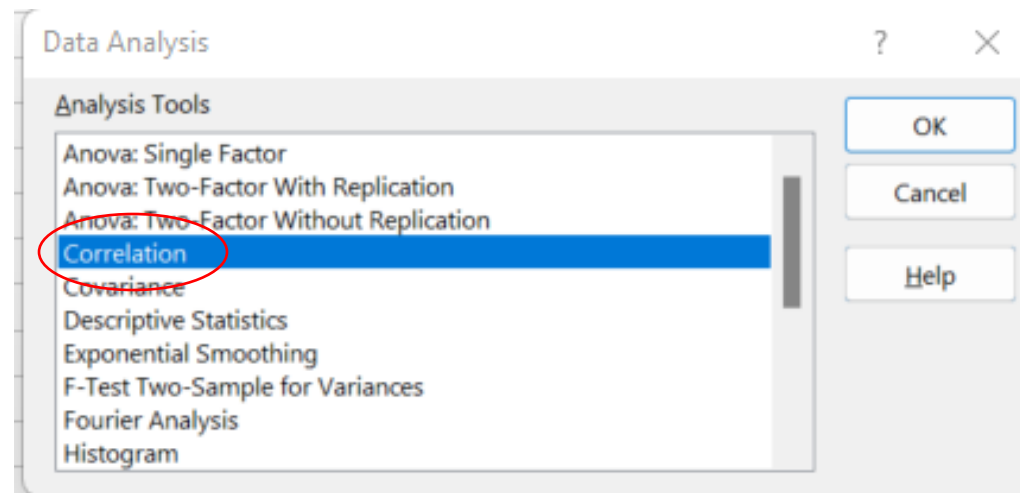
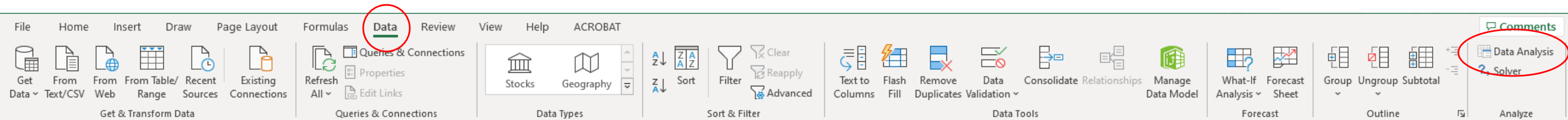
Regression – Excel Output

□ For sample regression equation, $y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \varepsilon$

SUMMARY OUTPUT									
Regression Statistics									
Multiple R	0.990523843								
R Square	0.981137483								
Adjusted R Square	0.979422709								
Standard Error	2.288046833								
Observations	25								
ANOVA									
	df	SS	MS	F	Significance F				
Regression	2	5990.771221	2995.385611	572.1671503	1.07546E-19				
Residual	22	115.1734828	5.235158308						
Total	24	6105.944704							
	Coefficients	Standard Error	t Stat	P-value	Lower 95%	Upper 95%	Lower 95.0%	Upper 95.0%	
Intercept	2.263791434	1.060066238	2.135518851	0.044099451	0.065348613	4.462234256	0.065348613	4.462234256	β_0
x1	2.744269643	0.093523844	29.34299438	3.90691E-19	2.550313061	2.938226226	2.550313061	2.938226226	β_1
x2	0.012527811	0.002798419	4.476746229	0.000188266	0.006724246	0.018331377	0.006724246	0.018331377	β_2

How to Run Correlation Analysis with Excel

□ Step 1) Click “Data” tab – Click “Data Analysis” – Choose “Correlation”



How to Run Correlation Analysis with Excel

□ Step 2) In the correlation window:

▫ Input:

- Input Range: Data for correlation analysis
- Grouped By: Check “Columns” if the records of each variable are contained in columns
- Check “Labels in First Row” if Input/Bin ranges contain variable names

The screenshot shows the 'Correlation' dialog box in Excel. The 'Input Range' field is empty. The 'Grouped By' section has 'Columns' selected with a radio button. The 'Labels in First Row' checkbox is checked. The 'Output options' section has 'New Worksheet Ply:' selected. A red arrow points from the 'Labels in First Row' checkbox to a data table below the dialog box. The table has columns A, B, and C, and rows 1 through 10. Row 1 contains labels 'y', 'x1', and 'x2'. Rows 2 through 10 contain numerical data.

Check "Labels in First Row" if your Input/Bin ranges contain variable names (i.e., A1:C1)

	A	B	C
1	y	x1	x2
2	9.95	2	50
3	24.45	8	110
4	31.75	11	120
5	35	10	550
6	25.02	8	295
7	16.86	4	200
8	14.38	2	375
9	9.6	2	52
10	24.35	9	100