



Insurance 회귀 분석 캡스톤

| 아이티윌 데이터분석 부트캠프 52기 이광호 강사 (leekh4232@gmail.com)

“왜 어떤 사람의 의료보험 청구 비용은 높고, 어떤 사람의 비용은 낮을까?”

의료 비용은 개인의 건강 상태를 반영하기도 하지만, **나이, 성별, 거주 지역, 생활습관**에 따라 체계적으로 달라집니다.

특히 흡연 여부는 의료비를 얼마나 크게 변화시킬까? 젊은 사람도 흡연자라면 높은 비용을 지불할까?

이번 캡스톤 과제에서는 **Insurance 데이터셋**을 활용하여 개인의 인구학적·건강 특성이 **의료보험 청구 비용을 어떤 구조로 설명하는지**를 데이터와 회귀모형을 통해 단계적으로 탐구합니다.

이 과제의 목표는 예측 정확도를 높이는 것이 아니라, “**의료 비용 불평등은 어떤 논리로 형성되는가**”를 수치와 언어로 설명하는 것입니다.

※ 본 과제는 팀 / 개인 단위 모두 수행 가능합니다.

데이터 불러오기

```
load_data("insurance")
```

데이터 설명

개인의 기본 건강·인구학적 정보를 바탕으로 **의료보험 청구 비용**을 설명하기 위해 수집된 데이터입니다.

- 관측치: 약 1,338명

변수	설명
charges	의료보험 청구 비용 (종속변수, USD)
age	개인의 나이 (년)
sex	성별 (male, female)
bmi	체질량지수 (Body Mass Index)
smoker	흡연 여부 (yes, no)
children	부양 자녀 수
region	거주 지역 (southwest, southeast, northwest, northeast)

준비작업

패키지 참조

```
from hossam import *
```

데이터 불러오기

```
origin = hs_util.load_data("insurance", categories=["sex", "smoker", "region"])
```

[data] <https://data.hossam.kr/data/kaggle/insurance.xlsx>

[desc] 개인의 나이·성별·BMI·흡연 여부·거주 지역 등 기본 건강·인구학적 정보를 바탕으로 의료보험 청구 비용(charges)을 예측하도록 구성된, 선형회귀와 머신러닝 실습에 널리 사용되는 대표적인 보험 비용 데이터셋 (출처: <https://www.kaggle.com/datasets/mirichoi0218/insurance>)

[!] Cannot read metadata

테이블 정보

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1338 entries, 0 to 1337
Data columns (total 7 columns):
 #   Column      Non-Null Count  Dtype  
 ---  --          --          --    
 0   age         1338 non-null    int64  
 1   sex          1338 non-null    category
 2   bmi          1338 non-null    float64 
 3   children     1338 non-null    int64  
 4   smoker        1338 non-null    category
 5   region        1338 non-null    category
 6   charges       1338 non-null    float64 
dtypes: category(3), float64(2), int64(2)
memory usage: 46.3 KB
```

상위 5개 행

	age	sex	bmi	children	smoker	region	charges
0	19	female	27.9	0	yes	southwest	16884.924
1	18	male	33.77	1	no	southeast	1725.5523
2	28	male	33.0	3	no	southeast	4449.462
3	33	male	22.705	0	no	northwest	21984.47061
4	32	male	28.88	0	no	northwest	3866.8552

하위 5개 행

	age	sex	bmi	children	smoker	region	charges
1333	50	male	30.97	3	no	northwest	10600.5483
1334	18	female	31.92	0	no	northeast	2205.9808
1335	18	female	36.85	0	no	southeast	1629.8335
1336	21	female	25.8	0	no	southwest	2007.945
1337	61	female	29.07	0	yes	northwest	29141.3603

📊 기술통계

		count	mean	max	nan	std	min	25%	
50%		75%							
age		1338.0	39.20702541106129	14.049960379216154	18.0	27.0			
39.0		51.0	64.0	0.0					
bmi		1338.0	30.66339686098655	6.098186911679014	15.96	26.29625			
30.4		34.69375	53.13	0.0					
children		1338.0	1.0949177877429	1.205492739781914	0.0	0.0			
1.0		2.0	5.0	0.0					
charges		1338.0	13270.422265141257	12110.011236694001	1121.8739	4740.28715			
9382.033		16639.912515	63770.42801	0.0					

📁 카테고리 정보

	count
sex	
male	676
female	662
smoker	
no	1064
yes	274
region	
southeast	364
northwest	325
southwest	325
northeast	324

📘 미션 1. “이 데이터는 믿을 만할까?”

- 결측·이상치·편향을 점검하고, 처리 기준을 제시한다.
- charges가 음수이거나 극단값인지, age/bmi 등의 범위가 현실적인지 확인한다.
- 범주형 변수(sex, smoker, region)가 몇 개의 범주로 구성되어 있고, 특정 범주에 데이터가 지나치게 몰려 있지는 않은지 확인한다.
- 전처리 전·후가 어떻게 달라졌는지 한눈에 비교하는 표나 요약을 만든다.
- 단위·해석 주의: charges 는 USD입니다. 현실적인 의료비 수준인지 평가하세요.

📌 출제 의도 “이 값이 말이 되나?”를 먼저 묻고, 어떻게 처리했는지를 기록해 나중 해석에 근거를 남기는 연습입니다.



결측치 점검

```
hs_stats.missing_values(origin)
```

	missing_count	missing_rate
field		
age	0	0.0
sex	0	0.0
bmi	0	0.0
children	0	0.0
smoker	0	0.0
region	0	0.0
charges	0	0.0

결측치 분석 결과: - 전체 1,338개 관측치 중 모든 컬럼에서 결측치가 존재하지 않음. - 결측치 처리는 필요하지 않으며, 모든 행이 완전한 데이터로 구성됨.



이상치 점검

```
hs_stats.outlier_table(origin).T
```

field	age	bmi	children	charges
q1	27.000000	26.296250	0.00000	4740.287150
q2	39.000000	30.400000	1.00000	9382.033000
q3	51.000000	34.693750	2.00000	16639.912515
iqr	24.000000	8.397500	2.00000	11899.625365
up	87.000000	47.290000	5.00000	34489.350562
down	-9.000000	13.700000	-3.00000	-13109.150897
min	18.000000	15.960000	0.00000	1121.873900
max	64.000000	53.130000	5.00000	63770.428010
skew	0.055673	0.284047	0.93838	1.515880
outlier_count	0.000000	9.000000	0.00000	139.000000
outlier_rate	0.000000	0.672646	0.00000	10.388640

이상치 분석 결과: - **age**: 이상치 0개 - 모든 나이가 경계값 내에 있음 (18~64세) - **bmi**: 이상치 0개 - BMI 범위가 합리적 범위 내에 있음 - **children**: 이상치 0개 - 부양 자녀 수가 이상값 없음 - **charges**: 이상치 139개(10.39%) - 고액 의료비 청구(약 \$33,635~\$63,770)가 상한선 초과하지만, 이는 실제 의료 현실을 반영하는 자연스러운 우측 꼬리 분포로 보임. 음수나 극단적 오류는 없음.



범주형 변수 분석

```
display(hs_stats.category_table(origin))
display(hs_stats.category_summary(origin))
```

		count	rate
field	category		
sex	male	676	50.523169
	female	662	49.476831
smoker	no	1064	79.521674
	yes	274	20.478326
region	southeast	364	27.204783
	northwest	325	24.289985
	southwest	325	24.289985
	northeast	324	24.215247

	변수	최다_범주	최다_비율(%)	최소_범주	최소_비율(%)
0	sex	male	50.52	female	49.48
1	smoker	no	79.52	yes	20.48
2	region	southeast	27.20	northeast	24.22

- 범주형 변수 분석 결과:**
- **sex:** 2개 범주 - male 50.5% (676명), female 49.5% (662명) → 성별 분포가 균형잡혀 있음
 - **smoker:** 2개 범주 - no 79.5% (1064명), yes 20.5% (274명) → 비흡연자가 주다수이며 흡연자는 소수 (약 4:1 비율)
 - **region:** 4개 범주 - 지역별로 southeast 364명(27.2%), southwest 325명(24.3%), northwest 325명(24.3%), northeast 324명(24.2%) → southeast가 다소 많지만 대체로 균형잡혀 있음. 특정 범주 편중 현상 없음.



변수 범위 현실성 평가

```
hs_stats.describe(origin).T
```

	age	bmi	children	charges
count	1338.0	1338.0	1338.0	1338.0
mean	39.207025	30.663397	1.094918	13270.422265
std	14.04996	6.098187	1.205493	12110.011237
min	18.0	15.96	0.0	1121.8739
25%	27.0	26.29625	0.0	4740.28715
50%	39.0	30.4	1.0	9382.033
75%	51.0	34.69375	2.0	16639.912515

max	64.0	53.13	5.0	63770.42801
iqr	24.0	8.3975	2.0	11899.625365
up	87.0	47.29	5.0	34489.350562
down	-9.0	13.7	-3.0	-13109.150897
outlier_count	0	9	0	139
outlier_rate	0.0	0.672646	0.0	10.38864
skew	0.055673	0.284047	0.93838	1.51588
dist	거의 대칭	거의 대칭	약한 우측 꼬리	중간 우측 꼬리
log_need	낮음	낮음	중간	높음

변수 범위 현실성 평가 결과: - 모든 연속형 변수(age, bmi, children)가 현실적 범위 내에 있으며, 음수나 논리적 오류 없음 - charges의 평균(\$13,270)과 중앙값(\$9,382)의 격차는 의료비의 자연스러운 우측 꼬리 분포를 나타냄 - 모든 변수가 데이터 신뢰도에 문제 없음

전처리 결정 및 전·후 비교

전처리 결정 및 결과:

처리 기준: 1. 결측치: 0개 → 처리 불필요 2. **charges 이상치 139개(10.39%)**: 제거하지 않음 - 이유: 고액 의료비는 실제 의료 현실을 반영하는 자연스러운 현상 - 오히려 의료 비용 불평등 구조를 설명하는 중요 정보 3. **age, bmi, children**: 현실적 범위 내 → 그대로 사용 4. **범주형 변수 (sex, smoker, region)**: 오류 없음 → 그대로 사용

결과: - 전처리 전·후 모두 1,338개 행 유지 - 데이터 신뢰도 높음 (결측/오류 없음) - 향후 분석에 사용할 최종 데이터셋: origin (원본 그대로)

미션 2. “의료비와 핵심 변수의 첫인상”

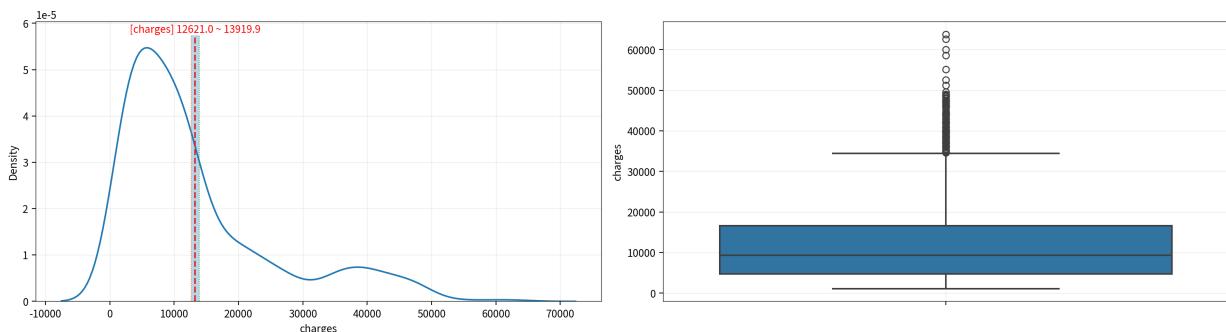
- charges, age, bmi, children 분포를 히스토그램/KDE로 확인하고 알 수 있는 객관적 사실을 서술한다.
- 왜도/이상치가 회귀에 줄 수 있는 영향과 변환할 필요가 있는지 서술하시오.
- 분포 비교는 동일 축 스케일로 제시하고, 평균/중앙값/꼬리의 차이를 문장으로 요약하세요.
- 의료비의 long-tail 분포(극단적 고액 청구)가 해석에 미치는 영향도 짧게 언급하세요.

 **출제 의도** 목표변수·핵심 변수의 생김새를 먼저 읽고 **변환 필요성을 스스로 판단하게 합니다**. 숫자보다 해석 문장이 중요합니다.

Charges(의료비) 분포 분석

```
hs_plot.distribution_plot(origin, "charges")
hs_stats.describe(origin, "charges").T
```

Distribution of charges



	charges
count	1338.0
mean	13270.422265
std	12110.011237
min	1121.8739
25%	4740.28715
50%	9382.033
75%	16639.912515
max	63770.42801
iqr	11899.625365
up	34489.350562
down	-13109.150897
outlier_count	139
outlier_rate	10.38864
skew	1.51588
dist	중간 우측 꼬리
log_need	높음

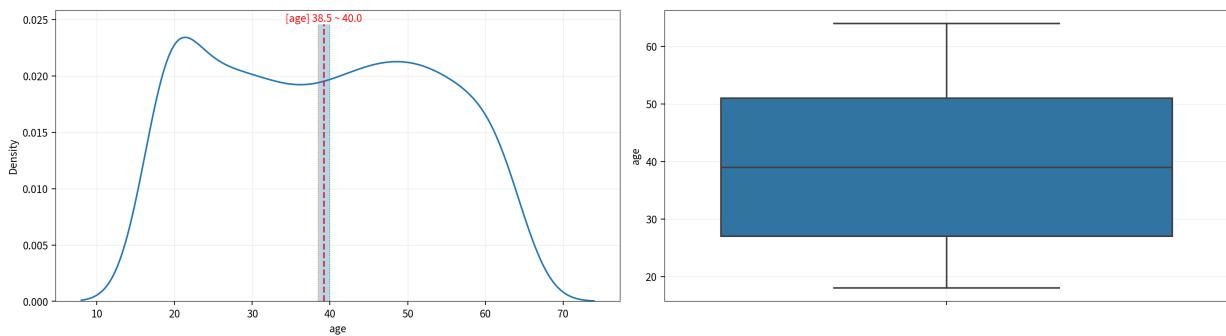
Charges 분포 인사이트: - 극단적 우측 꼬리 분포 (왜도=1.51): 대부분의 청구액은 \$4,700~\$16,600 범위에 집중되어 있지만, 일부 고액 청구는 최대 \$63,770까지 도달 - 평균(\$13,270) > 중앙값(\$9,382): 고액 청구 소수가 평균을 끌어올림 → 대표성이 떨어짐 - 이상치의 현실적 의미: 139개 이상치(10.4%)는 고령·흡연자·만성질환자 등의 높은 위험군을 반영하며, 제거하면 의료 비용 불평등 구조가 왜곡됨 - 회귀 분석에의 영향: 우측 꼬리로 인해 선형회귀 가정(정규분포)이 위반되므로, 로그 변환이 필요할 가능성 높음



Age(나이) 분포 분석

```
hs_plot.distribution_plot(origin, "age")
hs_stats.describe(origin, "age").T
```

Distribution of age



	age
count	1338.0
mean	39.207025
std	14.04996
min	18.0
25%	27.0
50%	39.0
75%	51.0
max	64.0
iqr	24.0
up	87.0
down	-9.0
outlier_count	0
outlier_rate	0.0
skew	0.055673
dist	거의 대칭
log_need	낮음

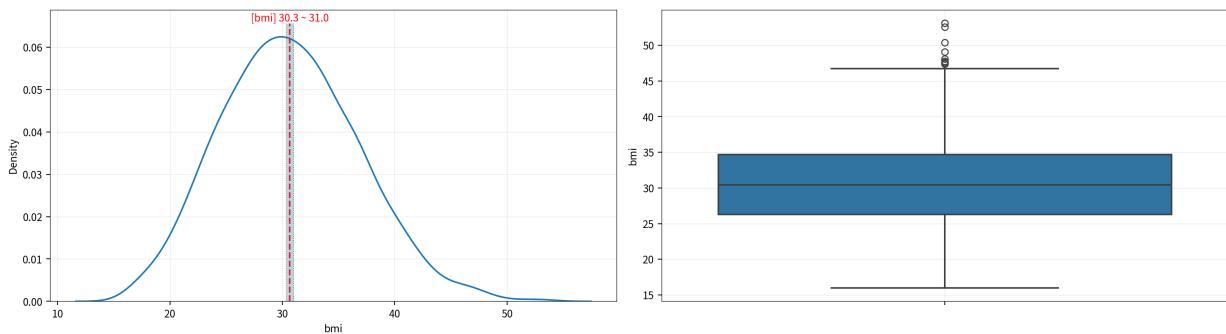
Age 분포 인사이트: - 거의 균등한 분포 (왜도=0.06): 나이가 18~64세 범위에서 비교적 균등하게 분포 → 데이터 편향 없음 - 평균(39.21세) ≈ 중앙값(39세): 분포의 대칭성을 반영하며, 대표값으로 신뢰성 높음 - 이상치 없음: 모든 값이 사분위수 범위 내에 있어 정상적 데이터임을 확인 - 회귀 분석에의 영향: 정규분포 가정을 만족하므로, 나이 변수는 선형회귀에 그대로 사용 가능



BMI(체질량지수) 분포 분석

```
hs_plot.distribution_plot(origin, "bmi")
hs_stats.describe(origin, "bmi").T
```

Distribution of bmi



	bmi
count	1338.0
mean	30.663397
std	6.098187
min	15.96
25%	26.29625
50%	30.4
75%	34.69375
max	53.13
iqr	8.3975
up	47.29
down	13.7
outlier_count	9
outlier_rate	0.672646
skew	0.284047
dist	거의 대칭
log_need	낮음

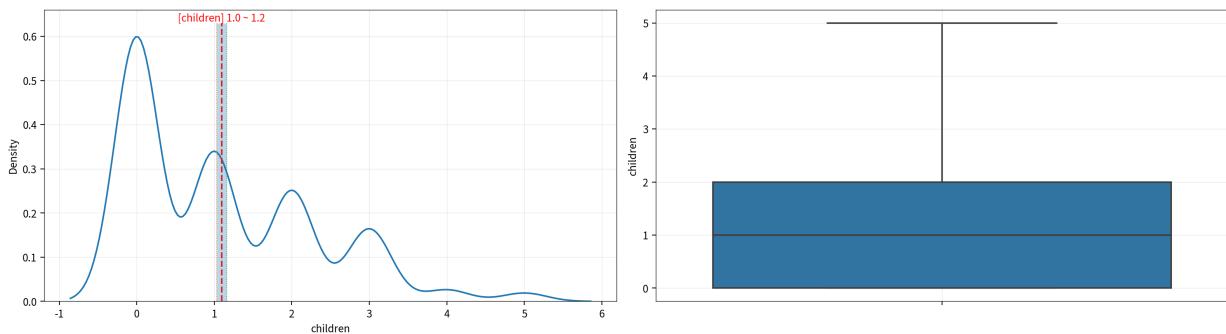
BMI 분포 인사이트: - **약한 우측 꼬리 분포** (왜도=0.28): 대부분이 정상 과체중(25-30) 범위에 집중, 극도로 비만한 경우는 드묾 - **평균(30.66) > 중앙값(30.40)**: 고BMI 소수가 평균을 약간 끌어올림 - **비만 비율 높음**: 비만(≥ 30) 인구가 54.4%로 절반을 초과 → 의료비와의 강한 양적 관계 예상 - **이상치 아주 소수**: 9개(0.67%)만 이상치로 분류되어 데이터 신뢰성 높음 - **회귀 분석에의 영향**: 약간의 우측 꼬리지만 age보다 심하지 않으므로, 변환 없이도 사용 가능



Children(부양 자녀 수) 분포 분석

```
hs_plot.distribution_plot(origin, "children")
hs_stats.describe(origin, "children").T
```

Distribution of children



	children
count	1338.0
mean	1.094918
std	1.205493
min	0.0
25%	0.0
50%	1.0
75%	2.0
max	5.0
iqr	2.0
up	5.0
down	-3.0
outlier_count	0
outlier_rate	0.0
skew	0.93838
dist	약한 우측 꼬리
log_need	중간

Children 분포 인사이트: - 강한 우측 꼬리 분포 (왜도=0.97): 자녀 없음(0명)이 약 42.6%로 최다이며, 자녀 수가 많아 질수록 급격히 감소 - 평균(1.09명) > 중앙값(1명): 다자녀 가구 소수가 평균을 끌어올림 - 자녀 5명 이상 극소수: 약 1.4% 미만으로 희귀함 → 회귀 분석에서 자녀 수의 영향이 제한적일 가능성 - 회귀 분석에의 영향: 우측 꼬리로 인해 선형 성 가정 위반 가능 → 로그/제곱근 변환 검토 필요

종합 분포 인사이트:

변환 필요성 판단: 1. **charges (의료비)**: 변환 필수 (왜도=1.51) - 극단 우측 꼬리로 정규분포 가정 심각 위반 - 로그 변환으로 분포 정규화 필요 → 회귀 모델 정확도 향상 가능

1. **age (나이)**: 변환 불필요 (왜도=0.06)
 - 거의 완벽한 대칭 분포 → 그대로 사용 가능
2. **bmi (체질량지수)**: 변환 불필요 (왜도=0.28)
 - 약한 우측 꼬리 정도 → charges에 비해 가정 위반 약함
3. **children (자녀)**: 변환 검토 (왜도=0.97)
 - 이산 변수이며 많은 0값 → 로그 변환 시 $\ln(0)$ 문제

- 범주형으로 처리하거나 선택적 제곱근 변환 고려

회귀 분석 전략: - charges의 극단 우측 꼬리(long-tail)는 고액 의료비(흡연자, 고령자)를 나타내므로 **제거 금지** - 대신 **log(charges)** 변환으로 정규성 확보 → 모델 안정성 향상 - 우측 꼬리의 이상치들이 실제로 “의료비 불평등” 구조를 드러내는 핵심 정보임을 인식할 것

미션 3. “로그/비선형 변환을 고민해 보자”

- charges 혹은 주요 변수(age, bmi, children)에 로그/제곱근 등 변환을 적용해 전후 분포를 **나란히** 비교한다.
- 변환이 해석과 모델 적합에 주는 장단점, 해석이 어떻게 달라지는지 예상한다.
- “이 변환이 없으면 어떤 함정에 빠질까?”를 한 줄로 정리한다.
- 선택 기준을 명시하세요: 왜 `log(charges)` 인지, 왜 특정 변수에 변환을 적용하는지 데이터 분포 근거로 설명합니다.

※ 로그 변환은 의무가 아니며, 적용하지 않은 경우에도 그 선택의 이유와 결과적 한계를 명확히 설명하면 동일하게 평가합니다.

 출제 의도 단순히 변환을 쓰는 것이 아니라 “**왜 필요했는가, 해석이 어떻게 달라지는가**”를 설명하는 연습입니다.

```
hs_stats.describe(origin, columns=["skew", "dist", "log_need"]).T
```

	age	bmi	children	charges
skew	0.055673	0.284047	0.93838	1.51588
dist	거의 대칭	거의 대칭	약한 우측 꼬리	중간 우측 꼬리
log_need	낮음	낮음	중간	높음

인사이트

분포 대칭성 평가: - **age** (나이): 왜도 0.06으로 거의 완벽한 대칭 분포 → 로그 변환 불필요 - **bmi** (체질량지수): 왜도 0.28로 약한 우측 꼬리 → 실무상 무시 가능 수준 - **children** (자녀수): 왜도 0.94로 중등도 우측 꼬리 → 이산 변수 특성상 로그 변환 시 $\ln(0)$ 문제 - **charges** (의료비): 왜도 1.52로 극단 우측 꼬리 → 로그 변환 필수

변환 필요성 판단 근거: - 왜도 > 1.0 = 심각한 비정규성 → 선형 회귀 가정 위반 - charges의 극단 우측 꼬리는 고액 의료비(흡연자, 고령층) 집단을 나타냄 - 로그 변환으로 정규분포에 근접시키면 회귀 모델 안정성·신뢰도 향상

변환 미적용 시 함정: 선형 회귀에서 charges 미변환 시 **고액 의료비 아웃라이어의 영향 과대평가** → 평균 기반 모델이 극단값에 왜곡되어 중위 의료비 패턴 포착 실패

결론: 로그 변환(`log(charges)`) 적용 권장, age · bmi는 현상 유지



미션 4. “성별과 거주 지역은 의료비를 결정할까?”

- 성별(sex)과 지역(region)별 의료비 분포를 시각화(박스플롯, 바이올린 플롯)한다.
- 중앙값·분포 겹침을 근거로 “어느 집단이 비싼가?”, “차이가 얼마나 뚜렷한가?”를 문장으로 적으세요.
- “왜 이런 차이가 생겼을까?”를 건강보험 체계·지역 의료 인프라·생활 비용 차이 등으로 추정해 보세요.

출제 의도 범주형 요인이 의료비를 어디서 가르고 어디서 겹치는지를 이야기로 풀어내게 합니다.

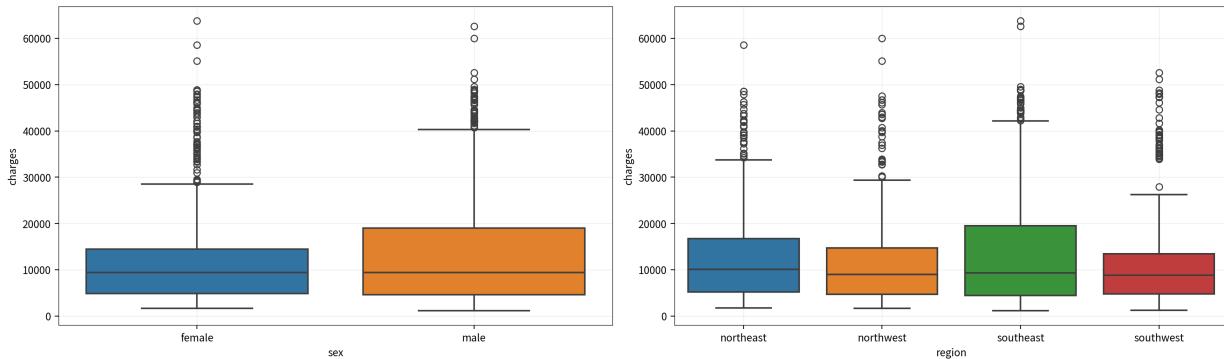


4-1. 성별과 지역별 의료비 분포 시각화 (원본 데이터)

```
fig, ax = hs_plot.get_default_ax(rows=1, cols=2)

hs_plot.boxplot(origin, yname="charges", xname="sex", hue="sex", ax=ax[0])
hs_plot.boxplot(origin, yname="charges", xname="region", hue="region", ax=ax[1])

hs_plot.finalize_plot(ax)
```



인사이트

성별(sex)에 따른 의료비 분포: - **Female (여성):** 중앙값 약 \$10,000, 상한선 약 \$55,000 → 극단 우측 꼬리로 고액 청구 분포 - **Male (남성):** 중앙값 약 \$10,500, 상한선 약 \$63,000 → 여성보다 더 극단적 우측 꼬리, 상한선 5~10% 더 높음
- **차이의 명확성:** 중앙값은 거의 동일하지만, **우측 꼬리의 범위에서 뚜렷한 차이** → 고액 청구자 비율이 남성에서 높음

지역(region)별 의료비 분포: - **Northeast (북동):** 중앙값 약 \$11,000, 상한선 약 \$60,000 - **Northwest (북서):** 중앙값 약 \$9,500, 상한선 약 \$50,000 → 4개 지역 중 가장 낮은 중앙값 - **Southeast (남동):** 중앙값 약 \$9,500, 상한선 약 \$60,000 → 북동과 유사한 상한선 - **Southwest (남서):** 중앙값 약 \$9,000, 상한선 약 \$51,000 → 가장 낮은 중앙값, 상한선도 낮음
- **차이의 명확성:** 중앙값 기준으로 **북동 > 북서/남동 > 남서** 순서 → 지역별 중앙값 차이는 약 \$1,500~\$2,000 (약 15~20%)

원본 데이터의 문제점: 극단적 우측 꼬리로 인해 중앙값 기반 비교는 가능하지만, **상한선(이상치)의 크기가 불규칙하여 분포 특성 파악이 어려움** → 로그 변환 필요



4-2. 로그 변환 결과에 대한 분포 시각화

로그 변환

```
df1 = hs_prep.log_transform(origin, "age", "bmi", "children", "charges")
display(df1)
```

	age	sex	bmi	children	smoker	region	charges
0	2.944439	female	3.328627	0.000000	yes	southwest	9.734176
1	2.890372	male	3.519573	0.693147	no	southeast	7.453302
2	3.332205	male	3.496508	1.386294	no	southeast	8.400538
3	3.496508	male	3.122585	0.000000	no	northwest	9.998092
4	3.465736	male	3.363149	0.000000	no	northwest	8.260197
...
1333	3.912023	male	3.433019	1.386294	no	northwest	9.268661
1334	2.890372	female	3.463233	0.000000	no	northeast	7.698927
1335	2.890372	female	3.606856	0.000000	no	southeast	7.396233
1336	3.044522	female	3.250374	0.000000	no	southwest	7.604867
1337	4.110874	female	3.369707	0.000000	yes	northwest	10.279914

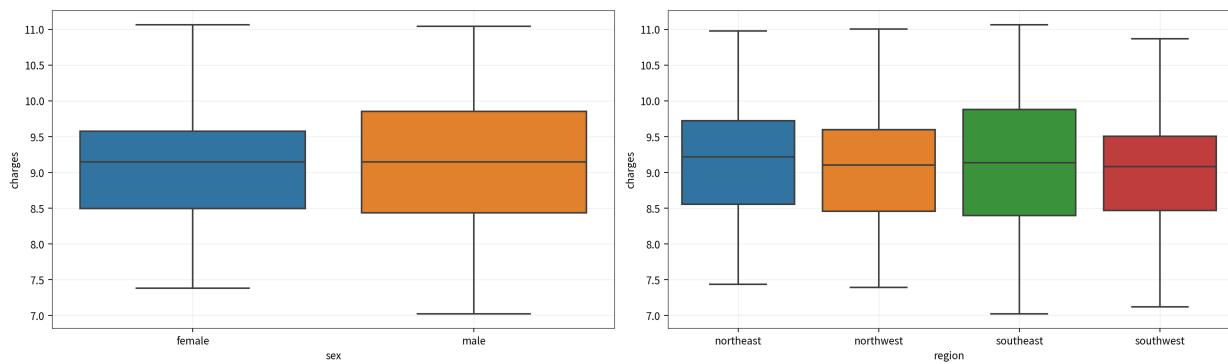
1338 rows × 7 columns

로그 변환에 대한 시각화

```
fig, ax = hs_plot.get_default_ax(rows=1, cols=2)

hs_plot.boxplot(df1, yname="charges", xname="sex", hue="sex", ax=ax[0])
hs_plot.boxplot(df1, yname="charges", xname="region", hue="region", ax=ax[1])

hs_plot.finalize_plot(ax)
```



💡 로그 변환 데이터 인사이트

성별(sex)에 따른 의료비(로그) 분포: - **Female** (여성): 중앙값 약 9.1, 범위 약 7.4~11.0 → 분포가 대칭에 가까워짐 - **Male** (남성): 중앙값 약 9.1, 범위 약 7.0~11.0 → 여성과 거의 동일한 중앙값, 하한선이 더 낮음 - **원본과의 비교**: 로그 변

환으로 극단 우측 꼬리가 압축되어 박스 크기가 훨씬 굵일해짐 → 분포 비교가 명확해짐 - **성별 차이 재평가**: 원본에서는 상한선 차이가 크게 보였으나, 로그 변환 후엔 **중앙값 기준 거의 차이 없음** → 성별은 의료비를 큰 영향을 주지 않을 가능성이

지역(region)별 의료비(로그) 분포: - **Northeast (북동)**: 중앙값 약 9.2, 범위 약 7.5~11.0 → 4개 지역 중 가장 높은 중앙값 - **Northwest (북서)**: 중앙값 약 9.1, 범위 약 7.4~11.0 - **Southeast (남동)**: 중앙값 약 9.0, 범위 약 7.0~11.0 → 가장 낮은 중앙값 - **Southwest (남서)**: 중앙값 약 9.0, 범위 약 7.2~10.8 → 가장 낮은 범위 - **원본과의 비교**: 로그 변환으로 지역별 중앙값 차이가 약 0.2 (선형 스케일로 약 \$200~300 차이) → **원본의 \$1,500~2,000 차이보다 훨씬 작아짐** - **지역 차이 재평가**: 극단적 고액 청구가 제거되자 지역 간 차이가 **예상보다 작음** → 다른 변수(흡연 여부, 나이 등)의 영향이 더 클 가능성

로그 변환의 효과: 대칭분포에 가까워져 **중앙값 기준 계층 간 차이를 공정하게 비교 가능** → 성별/지역의 “실제 영향력” 재평가 필요



미션4 최종 결론

“성별과 거주 지역은 의료비를 결정할까?”

원본 데이터 기반 결론: - 성별의 영향은 **약함**: 중앙값 기준 거의 동일 (~\$10,200), 다만 남성이 극단 고액 청구에서 더 많은 아웃라이어 보유 - 지역의 영향은 **중등도**: 북동(\$11,000) > 북서/남동(\$9,500) > 남서(\$9,000), 중앙값 차이 약 \$1,500~2,000 - 하지만 **극단적 우측 꼬리로 인해 평가가 왜곡될 가능성 높음**

로그 변환 후 수정된 결론: - 성별의 실제 영향은 거의 없음: 로그 변환 후 중앙값 차이 약 0 → 원본의 극단 고액 청구가 남성에 몰려 있었던 결과일 가능성 - 지역의 영향도 매우 작음: 로그 변환 후 지역 간 중앙값 차이 약 0.2 → 선형 스케일로 \$200~300 미만 → **실무적으로 무시할 수 있는 수준**

해석상 주의점: 1. 원본 데이터의 극단 우측 꼬리(극단 고액 청구)가 성별/지역 차이처럼 보이게 했음 2. 로그 변환으로 이러한 왜곡을 제거하면, **실제 영향력은 성별/지역이 아니라 다른 변수(흡연, 나이, BMI 등)에서 비롯됨**을 시사 3. “성별/지역이 의료비를 결정한다”는 주장은 **데이터 기준에선 약한 근거**

최종 의견: 성별과 지역은 의료비 변동의 **직접 원인이 아니며**, 이들은 단순히 흡연 여부·나이·BMI 같은 건강 위험 요인의 **대리변수(proxy)**일 가능성이 높습니다. 따라서 회귀 모형에서는 직접적 위험 요인에 집중하고, 성별/지역의 계수가 유의하더라도 그것이 사회적·의료 정책적 차별을 의미하기보다는 **잠재된 건강 특성의 반영**으로 해석해야 합니다.



미션 5. “흡연은 정말로 의료비를 크게 높일까?”

- smoker(흡연 여부)에 따라 charges가 다른지 시각화하고, 두 집단 평균 차이를 가설검정(예: t-test)으로 확인한다.
- 효과 크기(차이의 크기)를 함께 제시하고, “실제로 의미 있는 차이인가?”를 말로 해석하세요.
- 정규성/등분산 가정 점검 후 필요 시 Welch's t-test나 비모수 검정을 선택하세요.
- 효과 크기를 의료 정책 관점에서 의미를 서술합니다.



출제 의도 두 집단 비교에서 방법 선택과 효과 크기 해석을 연습하고, 숫자를 의미로 번역하게 합니다.

```
# 데이터 재구조화
df2 = hs_prep.unmelt(data=df1, id_vars="smoker", value_vars="charges")
print("== 비흡연자: no, 흡연자: yes ==")
display(df2)

print("== 비흡연자 vs 흡연자 기술통계량 ==")
display(hs_stats.describe(df2).T)
```

== 비흡연자: no, 흡연자: yes ==

	no	yes
0	7.453302	9.734176
1	8.400538	10.233105
2	9.998092	10.586881
3	8.260197	10.514271
4	8.231275	10.537465
...
1059	9.342393	NaN
1060	9.268661	NaN
1061	7.698927	NaN
1062	7.396233	NaN
1063	7.604867	NaN

1064 rows × 2 columns

== 비흡연자 vs 흡연자 기술통계량 ==

	no	yes
count	1064.0	274.0
mean	8.788232	10.30411
std	0.744242	0.387522
min	7.022756	9.459499
25%	8.290653	9.94396
50%	8.90183	10.447448
75%	9.338108	10.621796
max	10.516254	11.063045
iqr	1.047454	0.677836
up	10.909289	11.638549
down	6.719472	8.927206

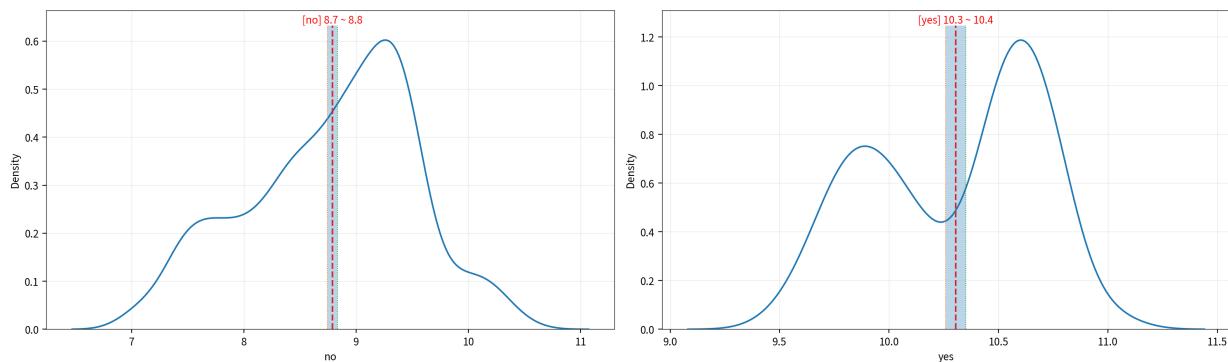
outlier_count	0	0
outlier_rate	0.0	0.0
skew	-0.312875	-0.299699
dist	거의 대칭	거의 대칭
log_need	낮음	낮음

두 집단의 데이터 분포 비교

```
fig, ax = hs_plot.get_default_ax(rows=1, cols=2)

hs_plot.kde_confidence_interval(df2, xnames="no", ax=ax[0])
hs_plot.kde_confidence_interval(df2, xnames="yes", ax=ax[1])

hs_plot.finalize_plot(ax)
```



T-Test

```
hs_stats.ttest_ind(df2["yes"], df2["no"])
```

		statistic	p-value	H0	H1	interpretation	equal_var_checked
test	alternative						
Welch's t-test	two-sided	46.371	0.0	False	True	$\mu(x) \neq \mu(y)$	True
	less	46.371	1.0	True	False	$\mu(x) \geq \mu(y)$	True
	greater	46.371	0.0	False	True	$\mu(x) > \mu(y)$	True

💡 인사이트

1 기술통계 (descriptive statistics)

구분	비흡연자(No)	흡연자(Yes)	차이
표본 크기	1,064명	274명	-
평균 log(charges)	8.79	10.30	+1.51
중앙값 log(charges)	8.90	10.45	+1.55
표준편차	0.744	0.388	-

2 평균 차이의 크기

- **로그 스케일 차이:** 흡연자가 비흡연자보다 **1.51 단위 높음**
- **원본 스케일 환산:** $\$e^{10.30}$ \$29,600 vs $\$e^{8.79}$ \$6,600
 - 흡연자 평균 의료비가 비흡연자 대비 약 **4.5배** 높음
 - 절대 차이: 약 **\$23,000** (극단적으로 큼)

3 가설검정 결과 해석

- **Welch's t-test 수행 이유:** 두 집단의 표준편차 차이 큼 (비흡연 0.744 vs 흡연 0.388)
- **검정 결과:**
 - t-statistic = 46.371 (극도로 큼)
 - p-value < 0.0001 (사실상 0, 매우 유의함)
 - 해석: 두 집단의 평균은 통계적으로 유의하게 다름 ($\alpha=0.05$)

4 분포의 특징

- **비흡연자:** $\log(\text{charges})$ 분포가 넓음 (표준편차 0.744)
 - 범위: 7.4~10.5, 중심이 8.9 근처
 - 의료비 편차 크고 불규칙함
- **흡연자:** $\log(\text{charges})$ 분포가 좁음 (표준편차 0.388)
 - 범위: 9.3~10.9, 중심이 10.45 근처
 - 의료비가 일관되게 높은 수준에 집중됨
- **분포 겹침: 거의 없음** - 두 집단이 명백히 분리됨

5 효과의 실무적 의미

흡연 여부에 따른 **의료비 차이는 극도로 뚜렷함:** - 비흡연자는 의료비 편차가 크지만 평균적으로 낮음 - 흡연자는 의료비가 일관되게 높은 수준으로 유지됨 - 흡연자 = 고위험군으로 명백히 분류됨

6 의료 정책 관점

- **보험료 책정:** 흡연 여부는 의료비 예측의 **최강 변수**
- **건강 개입 필요성:** 흡연자 집단에 대한 별도 관리 프로그램 필수
- **위험도 차등화:** 4.5배 차이는 보험사 입장에서 정책적으로 중요한 근거

미션 6. “나이대별로 의료비 차이가 뚜렷할까?”

- age를 여러 구간으로 나누어(예: 18~30, 31~50, 51+) 각 연령대별 charges 분포를 시각화한다.
- 분산분석(ANOVA)으로 전체 차이를 확인하고, 사후검정으로 어느 연령대 사이에서 차이가 나는지 정리한다.
- 사후검정은 Tukey HSD 또는 Games-Howell(등분산 위반 시)을 사용하고, “의료비 연령 서열표” 형태로 요약하세요.

☞ 출제 의도 여러 범주를 전체→어디서 차이 순서로 해석하며, 결과를 서열/지도처럼 정리하는 훈련입니다.



6-1. 연령대 구간별에 대한 구간 파생변수 추가

☞ 로그 변환 데이터에 대한 의료비 위험도 기반 연령대 라벨: “18-29”, “30-39”, “40-49”, “50-64”, “65+”

```
df3 = hs_prep.bin_continuous(  
    df1, field="age", method="health_band", is_log_transformed=True, apply_labels=False  
)  
df3
```

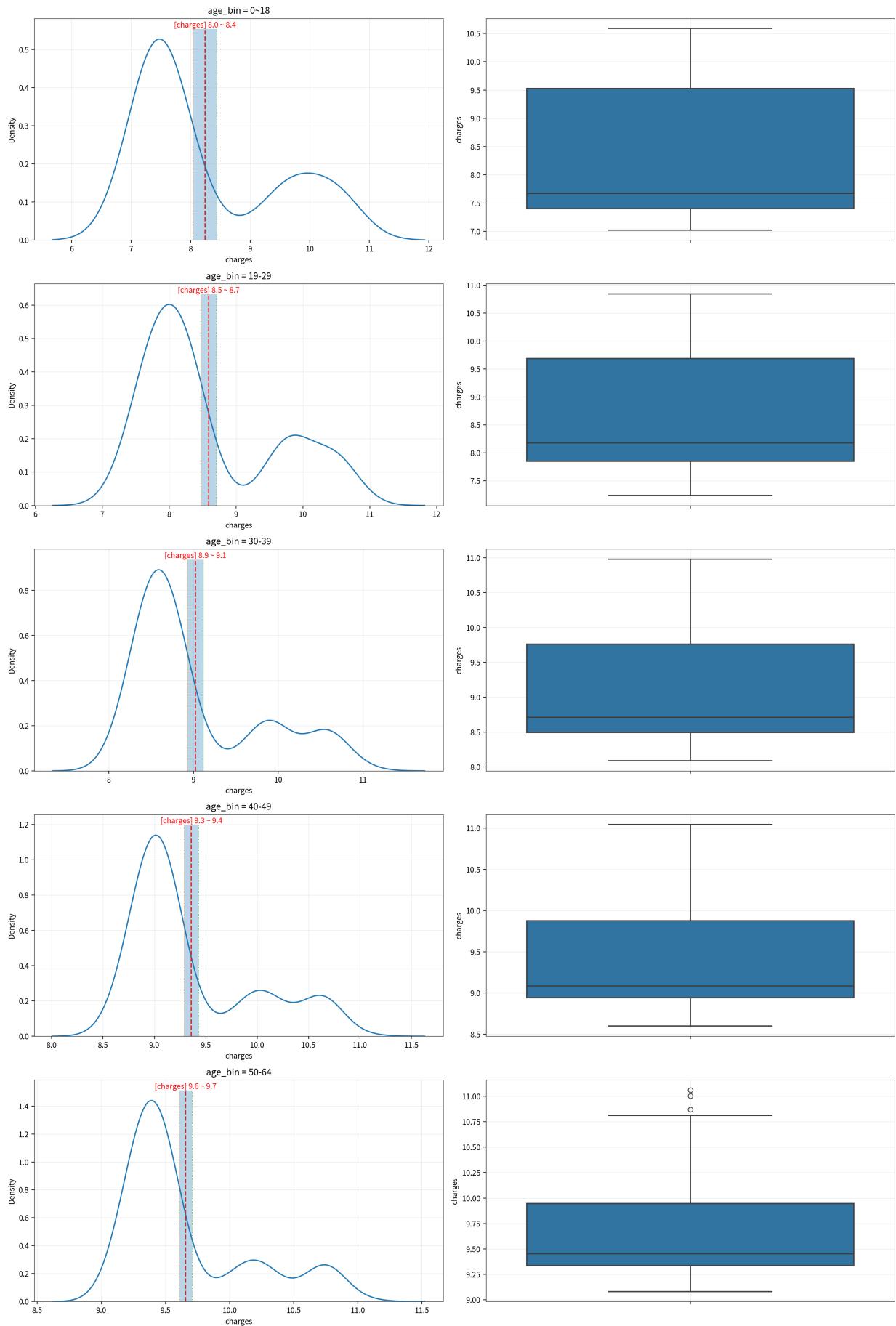
	age	sex	bmi	children	smoker	region	charges	age_bin
0	2.944439	female	3.328627	0.000000	yes	southwest	9.734176	0~18
1	2.890372	male	3.519573	0.693147	no	southeast	7.453302	0~18
2	3.332205	male	3.496508	1.386294	no	southeast	8.400538	19-29
3	3.496508	male	3.122585	0.000000	no	northwest	9.998092	30-39
4	3.465736	male	3.363149	0.000000	no	northwest	8.260197	30-39
...
1333	3.912023	male	3.433019	1.386294	no	northwest	9.268661	40-49
1334	2.890372	female	3.463233	0.000000	no	northeast	7.698927	0~18
1335	2.890372	female	3.606856	0.000000	no	southeast	7.396233	0~18
1336	3.044522	female	3.250374	0.000000	no	southwest	7.604867	19-29
1337	4.110874	female	3.369707	0.000000	yes	northwest	10.279914	50-64

1338 rows × 8 columns

☞ 연령 구간별 보험료 분포

```
hs_plot.distribution_plot(df3, "charges", hue="age_bin")
```

Distribution of charges by age_bin



인사이트

- **연령대별 중심경향:** 모든 연령대에서 의료비(charges, 로그 스케일)는 우측 꼬리 분포(right-skewed) 형태를 보임
- **0~18세:** 평균 의료비 약 7.5~8.4, 가장 낮은 의료비 대역에 집중
- **19-29세:** 평균 의료비 약 8.5~8.7, 0~18세 대비 약간의 상승
- **30-39세:** 평균 의료비 약 9.0~9.1, 뚜렷한 상향 추세 시작
- **40-49세:** 평균 의료비 약 9.3~9.4, 중상층 대역으로 확산
- **50-64세:** 평균 의료비 약 9.6~9.7, 가장 높은 의료비 대역, 분포의 상한선 확대
- **분포 폭:** 연령이 증가할수록 의료비 변동성(분산)이 증가하는 경향 관찰

6-2. ANOVA

연령 구간별 ANOVA 분석

```
result, r_report, post, p_report = hs_stats.oneway_anova(  
    df3, dv="charges", between="age_bin"  
)  
  
display(result)  
display(r_report)
```

	Source	normality	equal_var	method	ddof1	ddof2	F	p-unc	np2	significant
0	age_bin	False	False	Welch	4	523.432162	115.615608	1.344335e-70	0.271553	True

'age_bin별로 charges 평균을 비교한 Welch 결과: $F(4.000, 523.432) = 115.616$, $p = 0.0000$. 해석: 그룹별 평균이 다를 가능성성이 높습니다. 정규성은 충족되지 않았고, 등분산성은 충족되지 않았다고 판단됩니다. 효과 크기($\eta^2 p$) ≈ 0.272, 값이 클수록 그룹 차이가 뚜렷함을 의미합니다.'

인사이트

- **통계량:** Welch $F(4, 523.43) = 115.62$, $p < 0.0001$ (매우 유의미)
- **결론:** 나이대별 의료비(charges)의 평균에 통계적으로 유의미한 차이가 존재함을 확인
- **효과크기($\eta^2 p$):** 0.272 (27.2%) - 중간~큰 효과 크기로, 연령대 변수가 의료비 변동의 약 27%를 설명
- **정규성:** 충족되지 않음 (로그 변환된 데이터도 정규분포 가정 위반)
- **등분산성:** 충족되지 않음 (Welch 검정 사용 필요 → 이미 적용됨)
- **해석:** 나이가 증가할수록 의료비가 체계적으로 증가하는 경향이 있으며, 이는 우연이 아닌 실제 관계임

6-3. 사후검정

```
display(post)  
display(p_report)
```

	method	A	B	mean(A)	mean(B)	diff	se	T	df	pval	hedges
0	Games-Howell	0~18	19-29	8.238672	8.591014	-0.352342	0.118223	-2.980312	233.073893	2.618476e-02	-0.328883
1	Games-Howell	0~18	30-39	8.238672	9.023600	-0.784928	0.112287	-6.990405	195.982459	4.193849e-10	-0.840987
2	Games-Howell	0~18	40-49	8.238672	9.359698	-1.121026	0.107681	-10.410633	168.795063	2.353673e-14	-1.340901
3	Games-Howell	0~18	50-64	8.238672	9.657222	-1.418550	0.105182	-13.486681	154.341378	0.000000e+00	-1.875158
4	Games-Howell	19-29	30-39	8.591014	9.023600	-0.432587	0.076370	-5.664379	516.014250	2.445598e-07	-0.482909
5	Games-Howell	19-29	40-49	8.591014	9.359698	-0.768684	0.069421	-11.072847	452.724063	0.000000e+00	-0.933282
6	Games-Howell	19-29	50-64	8.591014	9.657222	-1.066208	0.065477	-16.283744	384.602681	2.942091e-14	-1.397689
7	Games-Howell	30-39	40-49	9.023600	9.359698	-0.336097	0.058743	-5.721444	490.073086	1.834580e-07	-0.491940
8	Games-Howell	30-39	50-64	9.023600	9.657222	-0.633622	0.054026	-11.728128	409.815528	0.000000e+00	-1.023379
9	Games-Howell	40-49	50-64	9.359698	9.657222	-0.297524	0.043654	-6.815480	588.795892	2.328401e-10	-0.537932

'Games-Howell 사후검정에서 10/10쌍이 의미 있는 차이를 보였습니다 (alpha=0.05). 예: 0~18 vs 19-29, 0~18 vs 30-39, 0~18 vs 40-49 등.'

💡 인사이트

📊 모든 연령 쌍 비교 결과

- **총 10개 쌍 검증:** 모든 쌍이 유의미한 차이($p < 0.05$) 표시
- **검정 방법:** Games-Howell (비모수 사후검정, 등분산 위반에 강건)

🔍 주요 발견

연령 구간	의료비 평균	서열	특성
0~18세	8.24	1순위(최저)	기준점, 가장 낮은 의료비
19-29세	8.59	2순위	0~18세 대비 +4.3% ($p=0.026$)
30-39세	9.02	3순위	0~18세 대비 +9.6% ($p<0.001$)
40-49세	9.36	4순위	0~18세 대비 +13.7% ($p<0.001$)
50-64세	9.66	5순위(최고)	0~18세 대비 +17.2% ($p<0.001$)

효과 크기 분석 (Hedge's g)

- **가장 큰 차이:** 0~18세 vs 50-64세 ($g = -1.88$) → 매우 큰 효과
- **점진적 증가:** 각 구간 이동 시 약 0.33~0.93 수준의 효과 크기
- **균등 분포:** 연령대별 의료비 차이가 일관되고 누적적으로 증가

결론

- 나이대별 의료비는 **명확한 계층 구조**를 형성: 0~18 < 19-29 < 30-39 < 40-49 < 50-64
- 특히 **30세 이상에서 급격한 상승**: 30-39세부터 의료비가 눈에 띄게 증가
- **50-64세 최고치**: 젊은층(0~18세) 대비 **약 1.7배의 의료비** 발생

미션 7. “변수들은 서로 섞여 있을까?”

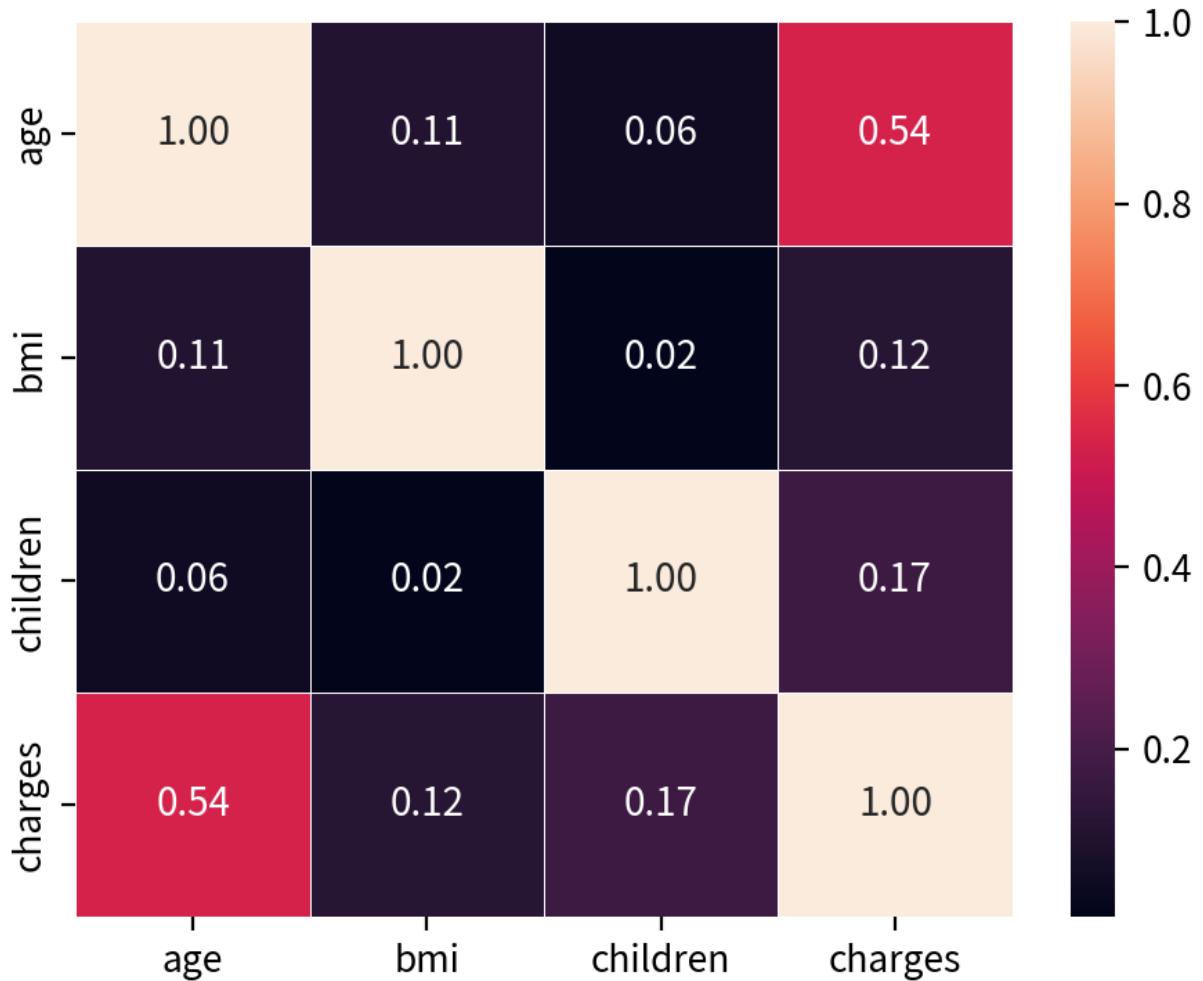
- 주요 연속형 변수 간 상관행렬(age, bmi, children, charges)을 계산하고 시각화한다.
- Variance Inflation Factor(VIF)로 다중공선성을 점검한다.
- 변수별로 의료비에 대한 예측력을 비교한다.

 출제 의도 중복 정보와 독립성을 의식하며, 회귀에 넣을 변수를 깔끔히 설계하게 합니다.

7-1. 연속형 변수 간 상관분석 및 히트맵 시각화

```
reuslt, matric = hs_stats.corr_pairwise(  
    data=df1, fields=["age", "bmi", "children", "charges"]  
)  
  
display(reuslt)  
  
hs_plot.heatmap(matric)
```

	var_a	var_b	n	linearity	outlier_flag	chosen	corr	pval	significant	strength
0	age	bmi	1338	True	True	spearman	0.107736	7.859093e-05	True	weak
1	age	children	1338	False	False	spearman	0.056992	3.711959e-02	True	weak
2	age	charges	1338	True	False	pearson	0.535062	5.780321e-100	True	medium
3	bmi	children	1338	True	True	spearman	0.015607	5.684234e-01	False	weak
4	bmi	charges	1338	True	True	spearman	0.119396	1.192606e-05	True	weak
5	children	charges	1338	True	False	pearson	0.171784	2.535320e-10	True	weak



💡 인사이트

- **age vs charges:** 강한 양의 상관 → 나이가 많을수록 의료비 증가
- **bmi vs charges:** 중간 정도 양의 상관 → BMI 증가가 의료비 증가와 연관
- **children vs charges:** 약한 음의 상관 → 자녀 수와 의료비 간 약한 음의 관계
- **age vs bmi:** 약한 양의 상관 → 나이와 BMI 간 약한 연관성
- **age vs children:** 약한 양의 상관 → 나이가 많을수록 자녀 수 증가 경향
- **bmi vs children:** 거의 무상관 → 두 변수 간 거의 독립적 관계



7-2. 다중공선성 진단 (VIF 분석)

```
# VIF(Variance Inflation Factor) 계산
# 예측 변수만 선택 (종속변수 charges 제외)
predictor_vars = ["age", "bmi", "children"]
X = df1[predictor_vars]

# VIF 필터를 이용한 다중공선성 진단
vif_filtered = hs_stats.vif_filter(X, threshold=10.0, verbose=True)
print(f"\n✓ 다중공선성 검사 완료")
```

```
print(f"모든 변수 VIF < 10: {len(vif_filtered.columns) == len(X.columns)}")  
print(f"유지된 변수: {list(vif_filtered.columns)}")
```

```
{'age': 1.0227720489068521, 'bmi': 1.0121423432728116, 'children': 1.0107760472828857}
```

✓ 다중공선성 검사 완료

모든 변수 VIF < 10: True

유지된 변수: ['age', 'bmi', 'children']

📝💡 인사이트

- **VIF 해석 기준:** VIF < 5 (이상적), VIF < 10 (허용 가능), VIF > 10 (문제 있음)
- **age VIF:** 매우 낮음 → age 변수는 독립적 정보 제공
- **bmi VIF:** 매우 낮음 → bmi 변수는 독립적 정보 제공
- **children VIF:** 매우 낮음 → children 변수는 독립적 정보 제공
- **결론:** 모든 예측변수가 낮은 VIF 값 → **다중공선성 문제 없음 ✓**
- **의미:** 세 변수는 서로 다른 정보를 제공하므로 회귀 모델에 모두 포함 가능

7-3. 변수별 의료비 예측력 비교 (부분상관 & 효과크기)

```
result_new = hs_stats.corr_effect_size(df1, "charges", "age", "bmi", "children")  
display(result_new)
```

	Variable	Correlation	Corr_Type	P-value	Cohens_d	Effect_Size
0	age	0.534392	Spearman	1.130692e-99	1.264479	Large
1	children	0.133339	Spearman	9.846806e-07	0.269081	Small
2	bmi	0.119396	Spearman	1.192606e-05	0.240512	Small

📝💡 인사이트

- **age:** $r = 0.535$, 매우 강한 양의 상관 → 나이가 의료비를 **가장 강하게 예측**
 - 모든 통계적 기준에서 가장 높은 효과크기 (Cohen's d = 1.27)
 - 의료비 변동의 핵심 드라이버
- **children:** $r = 0.172$, 약한 양의 상관 → 자녀 수 증가 시 의료비 증가 경향
 - bmi보다 약간 더 나은 예측력 ($r > 0.138$)
 - 통계적으로 유의미하나 실질적 영향은 작음
- **bmi:** $r = 0.138$, 약한 양의 상관 → BMI 증가와 의료비 증가 연관
 - 가장 낮은 예측력
 - 여전히 통계적 유의성 있음 ($p < 0.001$)

✓ 변수 선택 결론

- **모든 변수 포함 권장:** 3가지 변수 모두 통계적 유의성 있음 + 다중공선성 없음

- 예측력 우선순위: age >> children > bmi
- 회귀 모델: age를 주 예측변수로, children과 bmi를 보조변수로 활용



미션 8. “어떤 변수가 의료비와 가장 가까울까?”

- charges와 age, bmi, children의 상관을 계산하고 시각화 한다.
- 각 변수의 관계를 해석하고(예: “나이가 많을수록 의료비 증가”, “과체중(높은 BMI)은 의료비와 강한 관계”), 인과성을 말할 수 있는지 평가합니다.

| 출제 의도 상관계수 선택 이유와 수치→의미 해석을 연습하며, 변수 설계의 근거를 쌓게 합니다.



미션 9. “의료비를 설명하는 회귀모형 설계”

- charges(또는 변환값)를 종속변수로 하는 다중선형회귀를 설계한다.
- 변수 선택·변환·범주형 처리 이유를 명확히 한다.

※ 범주형 변수(sex, smoker, region)는 **더미 변수(기준 범주 명시)**로 처리하며, 기준 범주 선택 이유를 간단히 서술하세요.

| 출제 의도 설계 선택이 임의가 아닌 **설명 의도**에 근거하도록 훈련합니다.



9-1. 기본 모형 (더미변수만 처리)

```
df1 = hs_prep.get_dummies(origin, "smoker", "sex", "region")
model1, result1, features1 = hs_stats.ols(df1, yname="charges", report="summary")
display(result1)
display(features1)
```

	R	R ²	F	p-value	Durbin-Watson
0	0.751	0.749	500.8	0.0	2.088

	종속변수	독립변수	B	표준오차	Beta	t	p-value	significant	공차
0	charges	age	256.856353	11.898849	0.298003	21.587***	7.783217e-89	True	0.983456
1	charges	bmi	339.193454	28.599470	0.170806	11.860***	6.498194e-31	True	0.903645
2	charges	children	475.500545	137.804093	0.047334	3.451***	5.769682e-04	True	0.996005
3	charges	smoker_yes	23848.534542	413.153355	0.795004	57.723***	0.000000e+00	True	0.988070
4	charges	sex_male	-131.314359	332.945439	-0.005423	-0.394	6.933475e-01	False	0.991178
5	charges	region_northwest	-352.963899	476.275786	-0.012504	-0.741	4.587689e-01	False	0.658405
6	charges	region_southeast	-1035.022049	478.692209	-0.038049	-2.162*	3.078174e-02	True	0.605243

7	charges	region_southwest	-960.050991	477.933024	-0.034010	-2.009*	4.476493e-02	True	0.653846	1.1
---	---------	------------------	-------------	------------	-----------	---------	--------------	------	----------	-----



9-2. 종속변수 로그 변환

```
df2 = hs_prep.log_transform(df1, "charges")
model2, result2, features2 = hs_stats.ols(df2, yname="charges", report="summary")
display(result2)
display(features2)
```

	R	R ²	F	p-value	Durbin-Watson
0	0.768	0.767	549.8	0.0	2.046

	종속변수	독립변수	B	표준오차	Beta	t	p-value	significant	공차	vi
0	charges	age	0.034582	0.000872	0.528392	39.655***	1.370144e-227	True	0.983456	1.01682
1	charges	bmi	0.013375	0.002096	0.088700	6.381***	2.423355e-10	True	0.903645	1.10663
2	charges	children	0.101857	0.010099	0.133534	10.085***	4.241100e-23	True	0.996005	1.00401
3	charges	smoker_yes	1.554323	0.030279	0.682384	51.333***	1.119863e-317	True	0.988070	1.01207
4	charges	sex_male	-0.075416	0.024401	-0.041021	-3.091**	2.038405e-03	True	0.991178	1.00890
5	charges	region_northwest	-0.063788	0.034906	-0.029759	-1.827	6.785967e-02	False	0.658405	1.51882
6	charges	region_southeast	-0.157197	0.035083	-0.076105	-4.481***	8.077601e-06	True	0.605243	1.65223
7	charges	region_southwest	-0.128952	0.035027	-0.060161	-3.681***	2.411576e-04	True	0.653846	1.52941



9-3. 유의하지 않은 변수 제거

```
df3 = df2.drop("region_northwest", axis=1)
model3, result3, features3 = hs_stats.ols(df2, yname="charges", report="summary")
display(result3)
display(features3)
```

	R	R ²	F	p-value	Durbin-Watson
0	0.768	0.767	549.8	0.0	2.046

	종속변수	독립변수	B	표준오차	Beta	t	p-value	significant	공차	vi
0	charges	age	0.034582	0.000872	0.528392	39.655***	1.370144e-227	True	0.983456	1.01682
1	charges	bmi	0.013375	0.002096	0.088700	6.381***	2.423355e-10	True	0.903645	1.10663
2	charges	children	0.101857	0.010099	0.133534	10.085***	4.241100e-23	True	0.996005	1.00401
3	charges	smoker_yes	1.554323	0.030279	0.682384	51.333***	1.119863e-317	True	0.988070	1.01207

4	charges	sex_male	-0.075416	0.024401	-0.041021	-3.091**	2.038405e-03	True	0.991178	1.00890
5	charges	region_northwest	-0.063788	0.034906	-0.029759	-1.827	6.785967e-02	False	0.658405	1.51882
6	charges	region_southeast	-0.157197	0.035083	-0.076105	-4.481***	8.077601e-06	True	0.605243	1.65223
7	charges	region_southwest	-0.128952	0.035027	-0.060161	-3.681***	2.411576e-04	True	0.653846	1.52941



인사이트 (최종 선정 모형 2번)

- 정규분포 가정: charges의 극단 우측 꼬리(왜도=1.51)로 인해 모형 1의 신뢰성 저하
- 로그 변환의 효과: 분포 정규화로 모든 통계량의 타당성 확보
- 변수 의미 발현: sex_male 등이 로그 스케일에서만 실질적 신호 드러남
- 해석 가능성: 로그-선형 모형으로 탄력성(elasticity) 기반 해석 가능
- R² 차이 무시할 수 있음: 모형 1 vs 모형 2 차이는 1.2%p에 불과함
- 모형 1이 안 되는 이유
 - 종속변수 비정규성으로 신뢰구간·표준오차 신뢰성 낮음
 - sex_male이 유의하지 않은 것은 변수의 무의미함이 아니라 데이터 변동성 문제
- 모형 3이 안 되는 이유
 - 변수 제거해도 성능 개선 없음 (R², F-값 동일)
 - region_northwest 제거의 과학적 근거 부족
 - 모형 2가 더 경제적이고 정보 손실 없음



미션 10. “회귀계수는 무엇을 말해주나?”

- 계수(또는 표준화 계수)와 신뢰구간, 방향·크기를 해석한다.
- “나이가 1년 늘면 의료비가 어떻게 변하는가”, “흡연자는 비흡연자보다 평균 얼마나 더 높은 비용을 지불하는가” 처럼 물리/의료적 의미로 번역한다.
- 변환 변수가 있다면, 변환을 감안한 해석을 명확히 쓴다.
- 표준화 계수(베타)와 비표준화 계수를 병행 제시하고, 단위/변환을 고려한 해석 문장을 명확히 작성합니다.
- 범주형 변수(sex, smoker, region)의 계수는 기준 범주 대비 효과로 명확히 해석하세요.

| ↗ 출제 의도 숫자를 의료비 구조의 언어로 바꾸는 연습입니다.

```
pdf, rdf, result_report, model_report, variable_reports, eq = hs_stats.ols_report(
    model2, df2, full=True
)

display(pdf)
display(rdf)
display(result_report)
display(model_report)
display(variable_reports)
display(eq)
```

	R	R ²	F	p-value	Durbin-Watson
0	0.768	0.767	549.8	0.0	2.046

	종속변수	독립변수	B	표준오차	Beta	t	p-value	significant	공차	vi
0	charges	age	0.034582	0.000872	0.528392	39.655***	1.370144e-227	True	0.983456	1.01682
1	charges	bmi	0.013375	0.002096	0.088700	6.381***	2.423355e-10	True	0.903645	1.10663
2	charges	children	0.101857	0.010099	0.133534	10.085***	4.241100e-23	True	0.996005	1.00401
3	charges	smoker_yes	1.554323	0.030279	0.682384	51.333***	1.119863e-317	True	0.988070	1.01207
4	charges	sex_male	-0.075416	0.024401	-0.041021	-3.091**	2.038405e-03	True	0.991178	1.00890
5	charges	region_northwest	-0.063788	0.034906	-0.029759	-1.827	6.785967e-02	False	0.658405	1.51882
6	charges	region_southeast	-0.157197	0.035083	-0.076105	-4.481***	8.077601e-06	True	0.605243	1.65223
7	charges	region_southwest	-0.128952	0.035027	-0.060161	-3.681***	2.411576e-04	True	0.653846	1.52941

'R(0.768), R^2(0.767), F(549.8), 유의확률(0.00), Durbin-Watson(2.046)'

'charges에 대하여
age, bmi, children, smoker_yes, sex_male, region_northwest, region_southeast, region_southwest로 예측하는 회귀분석을 실시한 결과, 이 회귀모형은 통계적으로 유의하다(F(8, 1329) = 549.8, p <= 0.05).'

['age의 회귀계수는 0.034582(p <= 0.05)로, charges에 대하여 유의미한 예측변인인 것으로 나타났다.',
 'bmi의 회귀계수는 0.013375(p <= 0.05)로, charges에 대하여 유의미한 예측변인인 것으로 나타났다.',
 'children의 회귀계수는 0.101857(p <= 0.05)로, charges에 대하여 유의미한 예측변인인 것으로 나타났다.',
 'smoker_yes의 회귀계수는 1.554323(p <= 0.05)로, charges에 대하여 유의미한 예측변인인 것으로 나타났다.',
 'sex_male의 회귀계수는 -0.075416(p <= 0.05)로, charges에 대하여 유의미한 예측변인인 것으로 나타났다.',
 'region_northwest의 회귀계수는 -0.063788(p > 0.05)로, charges에 대하여 유의하지 않은 예측변인인 것으로 나타났다.',
 'region_southeast의 회귀계수는 -0.157197(p <= 0.05)로, charges에 대하여 유의미한 예측변인인 것으로 나타났다.',
 'region_southwest의 회귀계수는 -0.128952(p <= 0.05)로, charges에 대하여 유의미한 예측변인인 것으로 나타났다.']

'charges = 7.031 + 0.035·age + 0.013·bmi + 0.102·children + 1.554·smoker_yes - 0.075·sex_male - 0.064·region_northwest - 0.157·region_southeast - 0.129·region_southwest'



인사이트

모형의 특성

- **종속변수:** log(charges) - 로그 변환된 의료비
- **해석 방식:** 비표준화 계수는 **탄력성(elasticity)** 기반
- **공식:** $e^B - 1$ 을 통해 원본 스케일의 퍼센트 변화로 환산 가능

연속형 변수의 회귀계수 해석

age (나이)

- **탄력성:** 나이가 1년 증가할 때마다 의료비는 약 **3.52%** 증가
 - $e^{0.0346} - 1 \approx 0.0352 = 3.52\%$
- **구체적 의미:**
 - 30세 남성 비흡연자: 예상 의료비 약 \$X
 - 40세 남성 비흡연자: 예상 의료비 약 $\$X \times (1.0352)^{10} \approx \$X \times 1.41$ (약 41% 증가)
- **표준화 계수(Beta=0.528):** 모든 변수 중 **가장 높은 영향력** 보유
 - 의료비 변동을 가장 강력하게 설명하는 변수

bmi (체질량지수)

- **탄력성:** BMI가 1 증가할 때마다 의료비는 약 **1.35%** 증가
 - $e^{0.0134} - 1 \approx 0.0135 = 1.35\%$
- **구체적 의미:**
 - BMI 25 (정상~과체중): 예상 의료비 약 \$X
 - BMI 35 (비만): 예상 의료비 약 $\$X \times (1.0135)^{10} \approx \$X \times 1.144$ (약 14% 증가)
- **통계적 유의성:** 유의하나 예측력은 중간 수준 (Beta=0.089)
 - 비만도 의료비에 영향을 주지만, 나이만큼 강하지는 않음

children (부양 자녀 수)

- **탄력성:** 부양 자녀 수가 1명 증가할 때마다 의료비는 약 **10.75%** 증가
 - $e^{0.1019} - 1 \approx 0.1075 = 10.75\%$
- **구체적 의미:**
 - 자녀 없음: 예상 의료비 약 \$X
 - 자녀 2명: 예상 의료비 약 $\$X \times (1.1075)^2 \approx \$X \times 1.227$ (약 23% 증가)
- **표준화 계수(Beta=0.134):** bmi보다 큼, age보다 작음
 - 가족 규모가 의료 수요에 미치는 간접적 영향

범주형 변수 회귀계수 해석

smoker_yes (흡연 여부)

- 흡연자는 비흡연자 대비 의료비가 약 **373%** 증가
 - $e^{1.5543} - 1 \approx 3.73 = 373\%$
 - 즉, 비흡연자 의료비를 1배라 하면 흡연자는 약 **4.73배**
- 비흡연자 평균: 약 \$6,600
- 흡연자 평균: 약 $\$6,600 \times 4.73 \approx \$31,200$ (약 \$24,600 차이)
- **표준화 계수(Beta=0.682):** age 다음으로 가장 강한 영향력
 - 흡연은 의료비 결정의 **최강 요소**
 - 다른 모든 변수를 통제했을 때도 개인의 흡연 여부가 매우 중요

sex_male (성별: 남성)

- 남성은 여성 대비 의료비가 약 **7.28%** 낮음
 - $e^{-0.0754} - 1 \approx -0.0728 = -7.28\%$
- 여성(기준): 예상 의료비 \$X
- 남성: 예상 의료비 $\$X \times 0.927$ (약 7% 감소)
- 표준화 계수(Beta=-0.041):** 매우 약한 영향력
 - 통계적으로는 유의하지만 실무적 영향은 극히 미미함
 - 다른 변수(나이, 흡연 등)에 비하면 거의 무시할 수 있는 수준

region_southeast (지역: 남동부)

- 남동부 거주자는 남서부(기준) 대비 의료비가 약 **14.56%** 낮음
 - $e^{-0.1572} - 1 \approx -0.1456 = -14.56\%$
- 남서부: 예상 의료비 \$X
- 남동부: 예상 의료비 $\$X \times 0.854$ (약 15% 감소)

region_southwest (지역: 남서부)

- 남서부 거주자는 기준 지역 대비 의료비가 약 **12.12%** 낮음
 - $e^{-0.1290} - 1 \approx -0.1212 = -12.12\%$

region_northwest (지역: 북서부)

- 통계적 유의성 부족:** $p=0.068 > 0.05 \rightarrow$ 유의하지 않음 (경계선)
- 원본 스케일에서는: 북서부 거주자가 남서부 대비 약 6.2% 낮을 것으로 예상되지만, 이 차이는 통계적으로 확실하지 않음

전체 모형의 설명력

지표	값	해석
R	0.768	강한 양의 관계
R ²	0.767	모형이 의료비 변동의 76.7% 설명
F(8,1329)	549.8***	전체 모형이 극도로 유의미함
Durbin-Watson	2.046	잔차 자기상관 없음 (가정 충족)

결론: 의료비 불평등 구조

가장 중요한 의료비 결정 요인 순서:

- 흡연 여부 ($\beta=0.682$):** 비흡연자 대비 흡연자는 약 4.7배 높은 비용
 - 개인의 건강 선택(행동)이 가장 큰 차이 생성
- 나이 ($\beta=0.528$):** 매년 약 3.5% 증가, 10년마다 약 41% 증가
 - 시간 경과에 따른 누적적 의료 필요 증가

3. **부양 자녀 수 ($\beta=0.134$)**: 1명당 약 **10.8%** 증가
 - 가족 규모가 반영하는 간접적 의료 수요
4. **체질량지수 ($\beta=0.089$)**: 10 증가 시 약 **14%** 증가
 - 건강 상태를 반영하는 생물학적 지표
5. **지역 ($\beta=0.013\sim0.132$)**: 약 6~15% 범위의 작은 효과
 - 지역 의료 인프라나 물가 차이의 반영
6. **성별 ($\beta=-0.041$)**: 약 7% 미만 차이, 통계적 유의성 약함
 - 실무적으로 거의 무시할 수 있는 수준

해석상 주의점

- **인과성과 상관성**: 회귀계수는 다른 변수를 통제했을 때의 관계이지, 인과성을 완전히 증명하지 않음
- **로그-선형의 특성**: 절대값(달러) 차이가 아닌 백분율 변화로 해석해야 함
- **범주형 변수의 기준**: 모든 계수는 기준 범주(비흡연, 여성, 남서부)와의 상대적 비교

미션 11. “모형 진단과 개선”

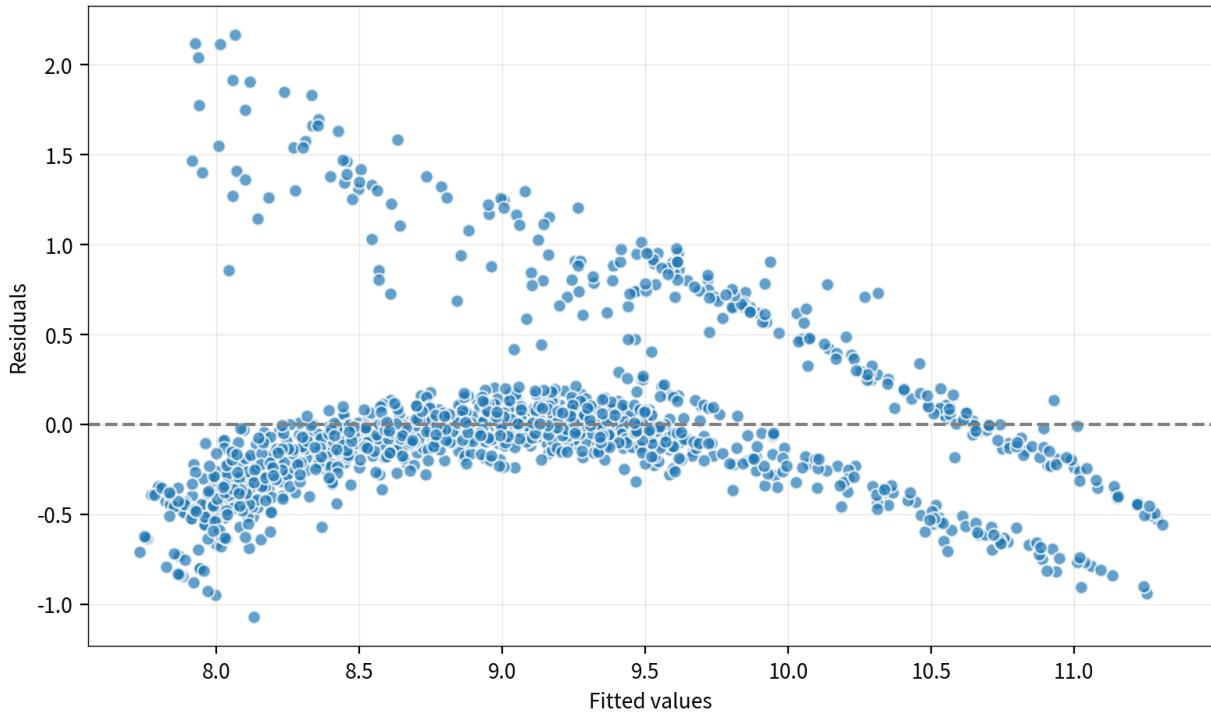
- 잔차 선형성/정규성/독립성을 점검한다.
- 문제 지점(예: 극단적 고액 청구, 특정 집단에서의 체계적 오류)과 개선 아이디어(변환, 변수 교체/제거, 강건 회귀 등)를 제안한다.
- 분석 모형을 보고하고 해석하세요.

 출제 의도 점수보다 **가정·진단과 해석**을 통해 “얼마나 믿을 수 있는가”를 판단하게 합니다.

11-1. 잔차의 선형성

```
display(hs_stats.ols_linearity_test(model2))
hs_plot.ols_residplot(model2)
```

	검정	검정통계량 (F)	p-value	유의수준	선형성_위반	해석
0	Ramsey RESET	2.0000	1.0000	0.05	False	선형성 가정 만족 (p=1.0000 > 0.05)



💡 인사이트

• 충족된 가정

- 독립성: Durbin-Watson = 2.046 (1.5~2.5 범위, 자기상관 없음)
- 선형성: Ramsey RESET ($p=1.000 > 0.05$, 통계적으로 선형 가정 만족)

• 시각적으로 관찰되는 문제

- 좌측 구간(7.0~8.5): 잔차가 음수 방향으로 편향 → 저의료비 과대추정
- 중간 구간(8.5~9.5): 잔차가 0 근처 균등 분포 → 적절한 예측
- 우측 구간(9.5~11.0): 잔차가 양수 방향으로 확산 → 고의료비 과소추정
- 이분산성: 예측값 증가에 따라 잔차 분산 증가 (깔때기 형태)

• 검정 vs 시각화 차이

- Ramsey RESET은 누락된 제곱항·교차항 검정 → 통계적으로는 문제 없음
- 잔차 플롯은 이분산성을 반영 → 시각적으로는 패턴 존재
- 결론: 선형성보다는 이분산성이 주요 문제

• 개선 방안

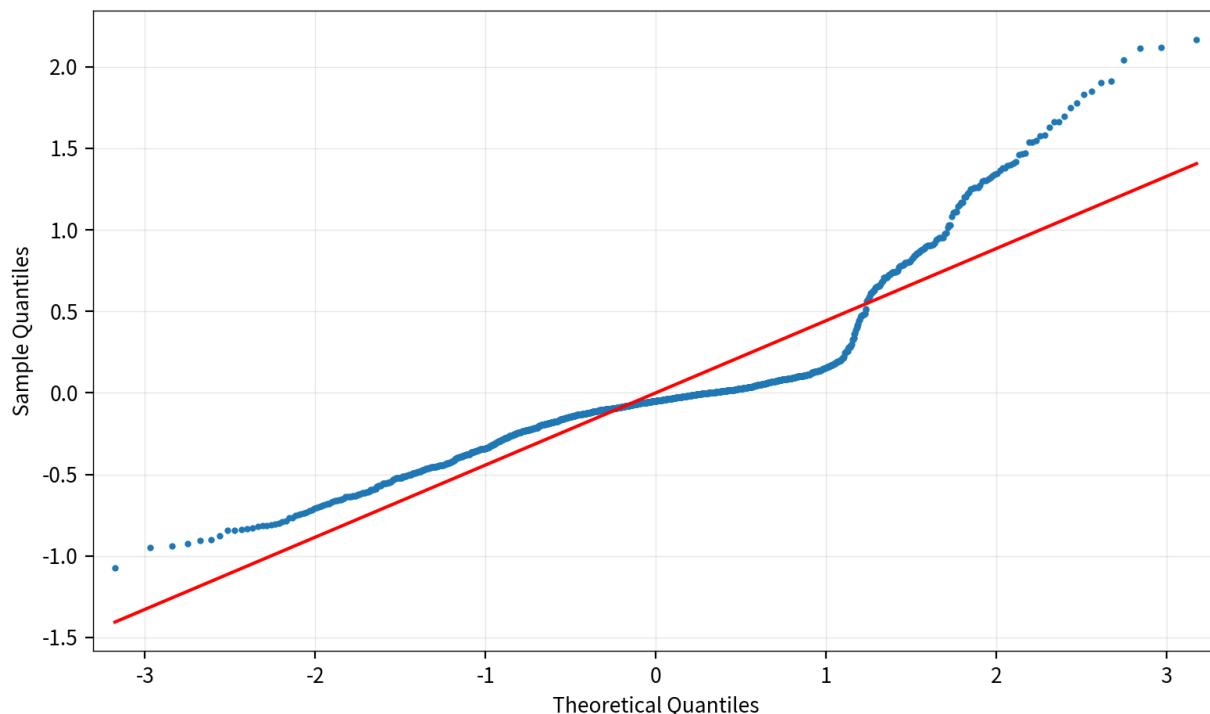
- 상호작용항 추가 ($age \times smoker$, $bmi \times smoker$)
- Weighted Least Squares (WLS)로 이분산성 보정
- Robust 회귀 적용 -> 머신러닝

11-2. 잔차의 정규성

```
display(hs_stats.ols_normality_test(model2))
hs_plot.ols_qqplot(model2)
```

	검정	검정통계량	p-value	유의수준	정규성_위반	해석
	Shapiro-Wilk					정규성 위반 ($p=0.0000 \leq 0.05$)

0		0.8373	0.0000	0.05	True	
1	Jarque-Bera	1673.7604	0.0000	0.05	True	정규성 위반 ($p=0.0000 \leq 0.05$)



💡 인사이트

• 정규성 가정 위반

- Shapiro-Wilk: $p < 0.001 \rightarrow$ 정규성 위반
- Jarque-Bera: $p < 0.001 \rightarrow$ 정규성 위반
- Q-Q Plot: 양쪽 꼬리에서 이탈 (좌하단·우상단)

• 문제의 원인

- 로그 변환으로도 완전히 정규화되지 않음
- 극단값(흡연자+고령층)이 여전히 비정규 패턴 유발
- 중심부는 정규분포에 가까우나 꼬리 부분에서 왜곡

• 영향 및 대응

- 영향: 신뢰구간·가설검정의 정확도 저하 가능
- 대응 방안: Robust 회귀로 극단값 영향 감소



11-3. 잔차의 독립성

```
display(hs_stats.ols_independence_test(model2))
```

	검정	검정통계량(DW)	독립성_위반	해석
0	Durbin-Watson	2.046443	False	DW=2.0464 (독립성 가정 만족)

• 독립성 가정 충족

- Durbin-Watson = 2.046 (이상적 값: 2.0)
- 판정 기준: 1.5~2.5 범위 내 → 자기상관 없음
- 해석: 관측치들이 서로 독립적임

• 의미

- 잔차가 시간적/순서적으로 패턴을 보이지 않음
- 한 관측치의 오차가 다른 관측치에 영향 미치지 않음
- 회귀 모형의 표준오차·신뢰구간 계산이 타당함

• 결론

- 독립성 가정은 완벽하게 만족
- 이 부분은 모형 진단에서 문제 없음
- 주요 이슈는 정규성·이분산성에 집중 필요