더미 회귀분석에서 참조변수에 대한 상세 설명

# 1. 이론적 배경

더미 변수란?

더미 변수는 범주형 데이터를 수치형 데이터로 변환하여 회귀분석에 사용할 수 있게 해주는 이진 변수입니다. 각 범주는 더미 변수로 변환되며, 이 변수는 해당 범주에 속할 때 1, 속하지 않을 때 0을 가집니다.

참조변수란?

참조변수(reference variable)는 더미 변수들 중에서 하나의 기준이 되는 범주를 말합니다. 회귀분석에서 모든 범주를 더미 변수로 변환하게 되면 다중공선성(multicollinearity) 문제가 발생할 수 있으므로, 하나의 범주를 기준으로 선택하고 나머지 범주만을 더미 변수로 포함합니다. 이 기준 범주가 참조변수가 됩니다.

# 2. 수식을 통한 설명

예시: 세 가지 범주를 가진 변수

색깔이라는 범주형 변수가 세 가지 범주를 가집니다: 빨강, 파랑, 녹색. 이 변수는 더미 변수로 변환됩니다. 이때 빨강을 참조변수로 선택하면 다음과 같은 수식으로 회귀분석이 설정됩니다.

1. 더미 변수 생성  
- D\_파랑 : 파랑이면 1, 그렇지 않으면 0.  
- D\_녹색 : 녹색이면 1, 그렇지 않으면 0.

2. 회귀식  
Y = β0 + β1 D\_파랑 + β2 D\_녹색 + ε

여기서:  
- β0: 빨강일 때의 종속변수 Y의 평균값.  
- β1: 파랑일 때 종속변수 Y의 평균값이 빨강과 얼마나 차이가 나는지.  
- β2: 녹색일 때 종속변수 Y의 평균값이 빨강과 얼마나 차이가 나는지.  
- ε: 오차(term).

3. 해석  
- β0: 빨강이 참조변수이므로, β0는 빨강일 때의 종속변수 Y의 평균을 의미합니다.  
- β1와 β2는 각각 파랑과 녹색이 참조변수인 빨강에 비해 종속변수 Y에 미치는 영향을 의미합니다.

# 3. 실제 사례

예시 데이터  
가정: 한 회사에서 마케팅 캠페인이 제품 판매에 미치는 영향을 분석하고자 합니다. 제품은 세 가지 색깔(빨강, 파랑, 녹색)로 출시되었으며, 주간 판매량이 종속변수 Y입니다.  
- 독립변수: 색깔 (범주형)  
- 종속변수: 주간 판매량 (연속형)

참조변수 선택: 빨강  
빨강을 참조변수로 설정하면, 다음과 같은 더미 변수가 생성됩니다:  
- D\_파랑: 파랑이면 1, 아니면 0.  
- D\_녹색: 녹색이면 1, 아니면 0.

회귀분석  
회귀식:  
판매량 = β0 + β1 D\_파랑 + β2 D\_녹색 + ε

결과 해석  
- β0: 빨강 제품의 평균 주간 판매량.  
- β1: 파랑 제품의 주간 판매량이 빨강에 비해 얼마나 더 높은지 (또는 낮은지).  
- β2: 녹색 제품의 주간 판매량이 빨강에 비해 얼마나 더 높은지 (또는 낮은지).

회귀 결과  
가정: 회귀 분석 결과가 다음과 같다고 하겠습니다.  
- β0 = 210  
- β1 = 30  
- β2 = -20

이 경우:  
- 빨강 제품의 평균 주간 판매량은 210개.  
- 파랑 제품은 빨강보다 평균적으로 30개 더 많이 팔림. 따라서 파랑의 예상 주간 판매량은 210 + 30 = 240개.  
- 녹색 제품은 빨강보다 평균적으로 20개 덜 팔림. 따라서 녹색의 예상 주간 판매량은 210 - 20 = 190개.

# 4. 참조변수 선택의 중요성

참조변수 선택은 해석의 기준을 결정합니다. 예를 들어, 파랑을 참조변수로 선택하면 회귀식과 해석이 다음과 같이 바뀝니다.  
1. 참조변수 변경 후 회귀식  
Y = β0 + β1 D\_빨강 + β2 D\_녹색 + ε

2. 결과 해석  
- β0: 파랑일 때 종속변수 Y의 평균값.  
- β1: 빨강일 때 종속변수 Y의 평균값이 파랑과 얼마나 차이가 나는지.  
- β2: 녹색일 때 종속변수 Y의 평균값이 파랑과 얼마나 차이가 나는지.