# 차량 가격 예측에서의 다중 회귀 분석의 적용

# An Application of Multiple Regression Analysis in Predicting Car Prices

지도교수 김영호

2023 년 05 월

청주대학교 공과대학교

데이터사이언스학과

베네딕터스 에스라 헤르노오

# 목차

# 1. 배경

자동차 산업, 특히 자동차는 지난 수십 년 동안 기하급수적인 성장을 이루며 인간의 삶과 세계 경제의 일부가 되었습니다. 이는 사용 용도에 따라 다양한 모델과 차종이 생겨나며, 모든 속성과 가격이 다양해졌습니다. 이러한 다양하고 복잡한 요인으로 인해 자동차 가격을 예측하는 것은 딜러, 구매자 및 판매자들에게 우려가 될 수 있습니다.

자동차 가격을 정확하게 예측하는 작업은 자동차 가격에 영향을 주는 여러 요인으로 인해 도전적입니다. 이러한 요인에는 자동차의 모델, 엔진 종류, 엔진 사양 및 연비 등이 포함될 수 있습니다. 따라서 이러한 변수를 효과적으로 고려할 수 있는 모델을 구축하는 것은 시장 참여자들에게 가치 있는 통찰력과 도구를 제공할 수 있습니다.

통계적 방법, 특히 다중 회귀 분석은 이러한 도전에 대응하는 데 잠재력을 보여준 바 있습니다. 다중 회귀 분석은 두 개 이상의 독립 변수 값에 기초하여 종속 변수의 결과를 예측하는 통계 기법입니다. 이는 자동차 가격에 영향을 주는 여러 요인을 동시에 처리할 수 있어 이 문맥에서 강력한 도구가 될 수 있습니다.

데이터의 점점 더 많은 활용 가능성과 자동차 산업에서의 가격 예측의 중요성을 고려할 때, 다중 회귀 분석을 활용한 자동차 가격 예측에 초점을 맞춘 연구 프로젝트는 시기적절하고 관련성이 있습니다. 이 연구는 가격 산정에 더 과학적이고 데이터 기반의 기초를 제공하여 시장의 공정성과 투명성을 개선하고, 자동차 산업에서의 생산, 판매 및 마케팅과 관련된 전략적 결정에 영향을 줄 수 있습니다

## 2. 방법론

### 2.1. 데이터 수집

이 연구를 위한 데이터는 데이터 과학자와 머신 러닝 전문가들의 온라인 커뮤니티인 Kaggle 에서 수집되었습니다. Kaggle 은 다양한 도메인의 데이터셋을 제공하는 것으로 유명하며, 이 연구 프로젝트에 귀중한 자원으로 활용됩니다. 선택한 데이터셋은 다양한 자동차의 모델, 엔진 종류, 엔진 사양 및 연비와 관련된 다양한 특성을 포함하고 있습니다. 이 데이터셋은 다중 회귀 분석을 통한 자동차 가격 예측에 포괄적으로 적용할 수 있게 해줍니다.

### 2.2. 변수 설명

이 연구를 위한 데이터셋은 자동차 가격에 영향을 미칠 수 있는 다양한 매개 변수들을 포함하고 있습니다. 예측 변수로 사용될 변수들은 자동차 모델, 엔진 연료, 엔진 사양, 그리고 연비 등입니다. 이러한 변수들에 대한 좋은 이해는 효과적인 다중 회귀 모델 구축에 매우 중요합니다.

- 자동차 모델: 이 변수는 세단, 해치백, 왜건, 하드탑, 컨버터블과 같이 차량의 구체적인 바디 스타일을 나타냅니다. 각 바디 스타일은 다른 소비자 선호도를 반영하며, 시장 수요에 따라 차량 가격에 영향을 미칠 수 있습니다.
- 엔진 연료 종류: 이 변수는 차량의 엔진에 사용되는 연료 종류를 나타냅니다. 이 변수의 값으로는 가솔린과 디젤이 포함되며, 각각은 차량의 성능, 운영 비용 및 이에 따른 시장 가격에 영향을 줄 수 있습니다.
- 엔진 사양: 이 변수는 실린더 수, 엔진 배기량 및 마력과 같은 여러 엔진 관련 속성을 포함합니다. 이러한 사양은 차량의 성능을 결정하는 데 중요하며, 이는 차량의 시장 가치에 큰 영향을 미칠 수 있습니다.
- 연비: 이 변수는 연료 단위당 이동할 수 있는 거리를 나타냅니다. 연비가 높은 차량은 일반적으로 운영 비용이 낮아 선호되며, 이는 차량의 수요와 가격 형성에 매우 영향을 미치는 중요한 요소입니다.

## 2.3. 모델 명세

이 연구에서는 주어진 예측 변수를 기반으로 자동차 가격을 예측하기 위해 다중 회귀 분석 모델을 사용합니다. 이 모델은 자동차 모델, 엔진 연료, 엔진 사양 및 연비와 같은 다중 독립 변수가 자동차 가격(반응 변수)에 영향을 미친다고 가정합니다.

이 사양을 통해 각 변수와 자동차 가격 간의 관계를 연구할 수 있습니다. 다중 회귀 분석을 적용함으로써 회귀 방정식에서 독립 변수의 계수를 추정할 수 있습니다. 이 계수는 가격의 평균 변화를 나타냅니다.

모델의 신뢰성과 타당성을 높이기 위해 데이터는 학습 세트와 테스트 세트로 분할되었습니다. 학습 세트는 모델을 구축하거나 '학습'하는 데 사용되고, 테스트 세트는 모델의 예측 성능을 평가하는 데 사용됩니다. 이 접근 방식은 모델의 평가가 훈련에 사용된 데이터가 아닌 새롭고 보지 못한 데이터를 예측하는 능력에 기반한다는 점에서 신뢰성이 향상됩니다..

모델 사양의 일반적인 형식은 다음과 같습니다:

$$\text{Car Price} = \beta 0 + \beta 1*(\text{Car Model}) + \beta 2*(\text{Type of Engine Fuel}) + \beta 3*(\text{Engine Specification}) + \beta 4*(\text{Fuel Economy}) + \varepsilon,$$

where:

- $\beta 0$ is the y-intercept (base price of cars when all variables are 0)
- $\beta_1$, $\beta_2$, $\beta_3$ 및 $\beta_4$ 는 각각의 독립 변수에 대한 계수로, 자동차 가격에 미치는 영향을 나타냅니다.
- $\varepsilon$ 는 오차 항으로, 모델에 포함되지 않은 자동차 가격에 영향을 주는 요소를 설명합니다.

# 3. 결과

## 3.2. 기술 통계

회귀 분석에 더 깊이 들어가기 전에, 각 변수의 기술적 통계량을 찾아 경향성, 분산 및 분포를 설명하는 주요 지표를 상세하게 제시하고자 합니다. 이는 모델을 구축하기 전에 중요한 초기 단계로, 복잡한 분석에 들어가기 전에 데이터의 일반적인 특성을 이해하는 데 도움이 됩니다. 자세한 설명 분석은 다음과 같습니다:

• Fuel type, Car body, Drive wheel: 이러한 변수는 문자 데이터를 나타냅니다. 데이터 세트에는 205 개의 관측치가 포함되어 있습니다. 주유방식은 휘발유차가 185 대, 경유차가 20 대가 대부분이다. 차체의 형태는 시장에서 유행하는 세단형 자동차가 대부분이다. 마지막으로 구동륜은 사륜구동, 실륜구동, 전륜구동 등 구동륜 시스템의 종류에 대한 정보를 제공한다. 이 가변 모드는 120 량의 전륜 구동 시스템입니다.

• **실린더 번호**: 이 변수는 엔진의 실린더 수를 제공합니다. 이 변수의 최소값은 2 이며 데이터 세트에서 실린더 수가 가장 적은 자동차에 실린더가 2 개 있음을 나타냅니다. 첫 번째 사분위수(데이터의 25%)에는 최소 4 개의 실린더가 있습니다. 중앙값(데이터의 50%)에는 실린더가 4 개 있고 평균은 약 4.38 이며, 이는 평균적으로 자동차에 실린더가 약 4-5 개 있음을 나타냅니다. 세 번째 사분위수(데이터의 75%)에는 4 개의 실린더가 있고 데이터 세트의 최대 실린더 수는 12 개입니다.

• **엔진 크기**: 가장 작은 엔진 크기는 61 개이며 첫 번째 사분위수는 엔진 크기가 97 개입니다. 평균 엔진 크기는 120 이고 평균은 약 126.9 단위입니다. 세 번째 사분위수는 141 개이고 데이터 세트의 최대 엔진 크기는 326 개입니다.

• **마력**: 마력의 범위는 최소 48 에서 최대 288 이며 중앙값은 95 이고 평균은 104.1 입니다. 1 사분위수와 3 사분위수는 각각 70 과 116 입니다.

• **시내 MPG**: 시내 주행 시 갤런당 마일(MPG)의 범위는 13~49 이며 MPG 중앙값은 24 이고 평균은 약 25.22 입니다. 1 사분위수와 3 사분위수는 각각 19 와 30 입니다.

• **고속도로 MPG**: 고속도로 주행의 경우 MPG 범위는 16 에서 54 까지이며 MPG 중앙값은 30 이고 평균은 30.75 입니다. 1 사분위수와 3 사분위수는 각각 25 와 34 입니다.

• **가격**: 데이터 세트의 자동차 가격 범위는 최소 $5,118 에서 최대 $45,400 입니다. 첫 번째 사분위수(데이터의 25%) 비용은 $7,788 이하이고 중간 가격은 $10,295 이며 세 번째 사분위수(데이터의 75%) 비용은 $16,503 이하입니다. 데이터 세트에 있는 자동차의 평균 가격은 약 $13,277 입니다.

### 3.3. 모델 추정

기술 통계를 통해 데이터에 대한 명확한 이해를 얻은 후, 다음 단계는 다중 회귀 모델을 추정하는 것입니다. 이 과정은 R 프로그래밍 언어를 사용하는 통계 소프트웨어인 R Studio 를 사용하여 독립 변수의 계수를 계산하는 것을 포함합니다. 이 계수는 종속 변수인 자동차 가격에 대한 각 독립 변수의 영향력을 정량화합니다.

모델 추정은 이전에 설정한 방정식을 따릅니다:

$$\text{Car Price} = \beta 0 + \beta 1*(\text{Car Model}) + \beta 2*(\text{Type of Engine Fuel}) + \beta 3*(\text{Engine Specification}) + \beta 4*(\text{Fuel Economy}) + \varepsilon$$

이 모델의 출력은 추정된 계수, R-제곱 값, t-값 및 p-값을 제공합니다. 이러한 지표들은 각 변수의 중요성과 자동차 가격에 미치는 영향을 이해하는 데 도움이 됩니다.

모델 추정 결과는 자동차 가격에 대한 예측 모델뿐만 아니라 다양한 자동차 특성과 가격 간의 관계를 더 깊게 이해하는 데 도움을 줍니다. 이 정보를 통해 자동차 구매자는 보다 명확한 결정을 내릴 수 있습니다.

### 3.4. 모델 평가

#### 3.4.1. 모델 요약

다중 회귀 모델을 추정한 후, 모델의 성능과 신뢰성을 평가하는 것이 다음 단계입니다. 아래는 출력의 기술적 해석입니다:

- **계수(Coefficients)**: 이는 각 변수에 대한 추정된 $\beta$ 값입니다. 자동차 가격에 강력한 영향을 미치는 변수는 엔진 크기로, 매우 유의미한

결과($p < 0.001$)를 나타냅니다. 가스 연료 차량, 해치백 모델, 그리고 마력도 자동차 가격에 영향을 미치는 요인이지만, 엔진 크기만큼 유의미하지는 않습니다. 이들 변수는 각각 $p < 0.01$ 의 유의수준을 가지고 있습니다.

- **수정된 R-제곱(Adjusted R-squared)**: 이 측정은 모델 이 예측 변수의 개수에 대해 R-제곱을 보정합니다. 자유도를 고려하여 일반적으로 모델의 효과를 더 정확하게 나타냅니다. 수정된 R-제곱은 0.8351 로, 약 83.51%를 나타내며, 예측 변수의 개수를 고려할 때에도 모델이 상당히 효과적임을 시사합니다.
- **F-통계량과 그 p-값**: F-통계량은 모든 회귀 계수가 영인, 즉 예측 변수가 통계적으로 유의하지 않다는 귀무 가설을 검정하는 데 사용됩니다. 높은 F-통계량(60.94)과 매우 작은 p-값(< 2.2e-16)은 이 귀무 가설을 기각함을 시사합니다. 이는 적어도 하나의 예측 변수가 영과 유의하게 다르며, 자동차 가격 예측에 기여한다는 것을 나타냅니다.

### 3.4.2. 모델 검증

모델의 예측 성능을 평가하기 위해 모델 추정 중에 사용되지 않은 테스트 데이터 세트가 적용되었습니다. 이 과정은 모델이 새로운, 이전에 보지 못한 데이터에 대해 얼마나 일반화할 수 있는지를 평가하는 것으로, 예측 모델의 중요한 측면입니다.

이를 위해 R 에서 predict() 함수를 사용하고, 모델과 테스트 데이터 세트를 인자로 전달합니다. 이 함수는 모델을 기반으로 테스트 데이터의 자동차 가격에 대한 예측을 생성합니다. 모델의 성능을 시각적으로 평가하는 유용한 방법은 이러한 예측값을 실제 가격과 비교하는 것입니다. 예측값과 관측값을 산점도로 그려 비교할 수 있습니다.

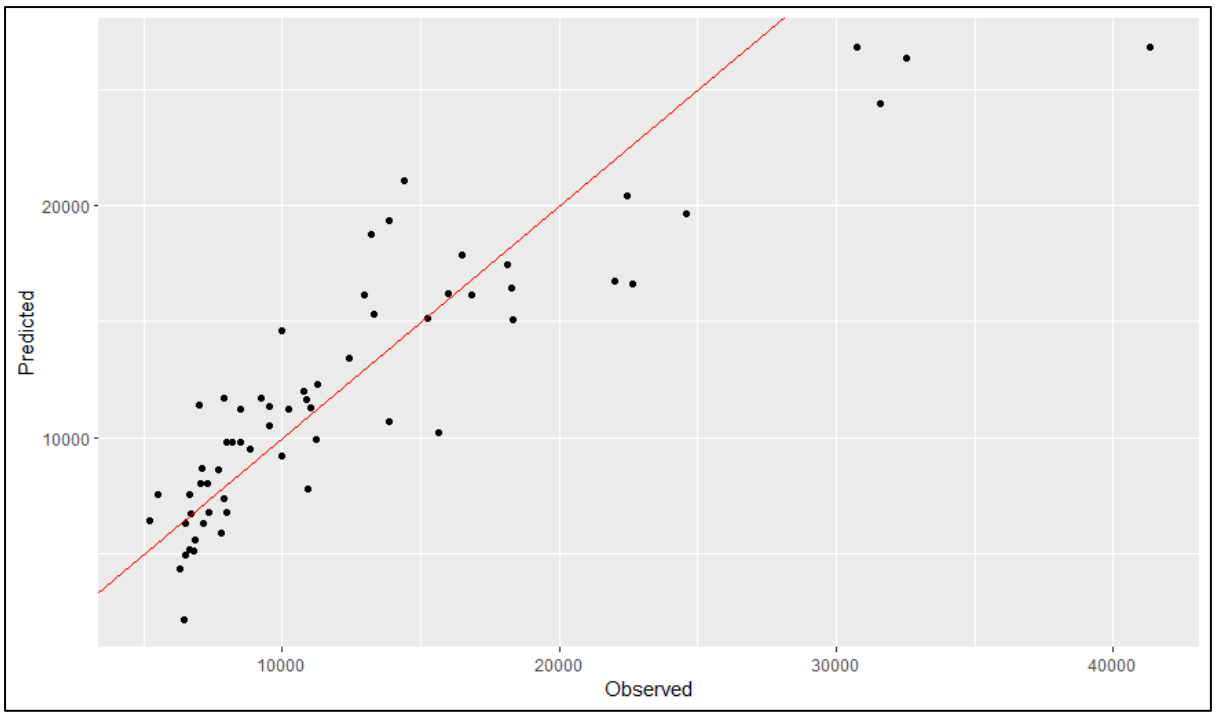


*Figure 1: Scatterplot of the model*

Figure 1 에 나타난 것처럼, 이 그래프의 각 점은 테스트 데이터 세트의 자동차를 나타내며, 한 축에는 관측된 가격이, 다른 축에는 (모델에서 예측한) 예측 가격이 표시됩니다. 모델이 완벽하게 정확하다면, 모든 점은 관측값과 예측값이 동일한 y = x 인 빨간색 기준선 위에 위치할 것입니다. 점들이 이 선에 가까울수록 모델의 예측이 더 좋습니다.

### 3.4.3. 성능 지표

모델의 정확도는 Root Mean Square Error (RMSE)와 같은 통계적 지표를 사용하여도 평가되었습니다. RMSE 는 예측값이 평균적으로 실제 가격에서 어느 정도 벗어났는지를 정량화합니다. RMSE 가 작을수록 모델의 정확도가 높습니다.

이 경우, RMSE 는 대략 $3433.87 입니다. 테스트 데이터의 중앙값인 $9984.5 를 고려하면, 모델의 예측은 평균적으로 약 $3433.87 정도 벗어난다는 것을 나타냅니다.

이 수치적 측정 값은 이전의 시각적 및 통계적 분석과 함께 모델의 성능을 종합적으로 평가하는 데 도움을 줍니다. 이를 통해 모델이 어디에서 뛰어나며 어디에서 개선이 필요한지를 파악할 수 있습니다.

## 4. 결론

이 연구는 자동차 모델, 엔진 연료, 엔진 사양 및 연비와 같은 다양한 요소를 기반으로 자동차 가격을 예측하기 위해 다중 회귀 분석을 적용하는 데 집중되었습니다. Kaggle 의 데이터를 활용하여 모델을 개발하고 통계적 및 그래픽 평가를 통해 효과를 평가했습니다.

모델은 R-제곱 값, RMSE 및 관측값과 예측값의 산점도를 통해 예측 정확도가 어느 정도 허용 가능한 수준임을 나타냈습니다. 이러한 결과는 자동차 가격에 영향을 미치는 요소들의 복잡한 상호작용을 강조하며, 자동차 산업에서 자동차 가격을 이해하고 예측하는 도구로서 모델의 가치를 강조합니다.

이 연구는 자동차 가격에 영향을 미치는 여러 요소들을 강조하며, 이를 이해하는 것이 자동차 시장에서 구매자와 판매자 모두에게 도움을 줄 수 있다고 주장합니다. 이 연구는 예측 도구를 제공하는데 그치지 않고, 자동차 가격에 영향을 미치는 요소들에 대한 통찰력을 제공하여 구매자와 판매자가 정보를 활용한 결정을 내릴 수 있도록 도움을 줍니다.

이 연구 결과를 바탕으로, 향후 연구에서는 보다 복잡한 모델에 더 깊이 집중하거나 추가적인 변수를 통합하여 예측 정확도를 더욱 향상시킬 수 있습니다. 저는 계속해서 이해를 깊게 하고 모델을 개선시키는 과정을 거치며, 더 많은 모델을 예측하기 위한 점점 정밀한 도구를 제공하는 것을 목표로 하고 있습니다.

# 참고문헌

"How to interpret RMSE (simply explained)". (2022 년 Augustus 월 24 일). Stephen Allwright: https://stephenallwright.com/interpret-rmse/에서 검색됨

"How to validate a predictive model?" (2019 년 January 월 3 일). aspexit: https://www.aspexit.com/how-to-validate-a-predictive-model/에서 검색됨