R을 이용한 통계 기초와 데이터 분석

남현진

한성대학교

2020

회귀분석

Definition

회귀분석이란 종속변수와 독립변수들간의 관련성을 설명할 수 있는 수학적 모델이다.

Important theorem

- 종속변수: 독립 변수에 의해 영향을 받는 변수
- 독립변수: 종속 변수에 영향을 주는 변수

Examples

- 종속변수: 아들의 키
- 독립변수: 아들의 몸무게, 나이, 아버지의 키, 아버지의 몸무게

회귀분석

Important theorem

회귀모형은 다음과 같은 측면에서 사용된다.

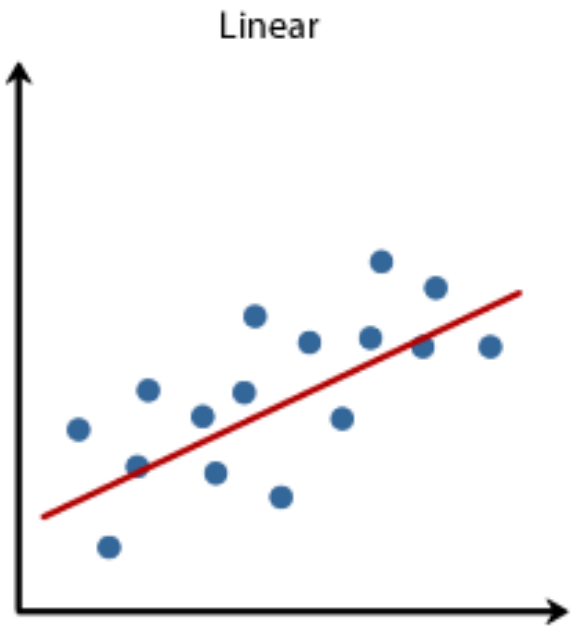
- 관측된 두 변수의 값을 이용하여 둘 간의 관계성을 확인한다.
- 확인된 관계성을 이용하여 독립변수를 가지고 종속변수 값을 예측한다.

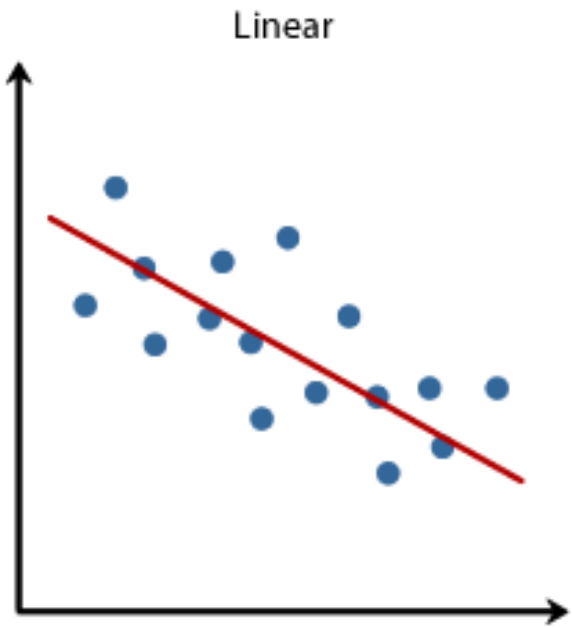
Examples

어떤 회사가 여러가지 매체에 광고를 보냈다고 한다. 이 때 매채별 광고 지출과 총 판매량간의 관계를 알아보고자 회귀분석을 사용할 수 있다. 회귀 분석을 사용하면 다음과 같은 것들을 확인할 수 있다.

- 광고 지출이 판매량과의 관계성
- 어떠한 매체(소셜미디어, 지하철, TV)가 가장 효과적이었는지. 각 매체에 대한 가중치
- 미래에 광고를 다시 보낸다고 할 때, 각 매체별 광고 지출에 따른 판매량의 예측치

선형 회귀





선형 회귀

Definition

선형 화귀는 i번째 관측값을 뜻하는 변수들이 $(X_{i1}, X_{i2}, X_{i3}, ..., X_{ip}, Y_i)$ 형태로 주어졌을 때 종속변수 Y_i의 p개의 독립변수 $X_{i1}, X_{i2}, X_{i3}, ..., X_{ip}$를 다음과 같은 선형 식으로 표현한다.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \epsilon_i$$

Important theorem

- $\beta_0, \beta_1, \beta_2, ..., \beta_p$: 회귀 모델의 계수
- ϵ_1: 오차(error)

선형 회귀

Important theorem

선형 회귀 모형에는 4가지 기본 가정이 있다.

- 선형성: 예측하고자 하는 종속변수 y와 독립변수 x간에 선형성을 만족해야 한다.
- 독립성: 독립변수 X_i간의 상관관계가 없어야한다.
- 등분산성: 잔차의 분산은 모든 종속변수에 상관 없이 등분산이다.
- 정규성: 잔차가 정규분포를 만족해야 한다.

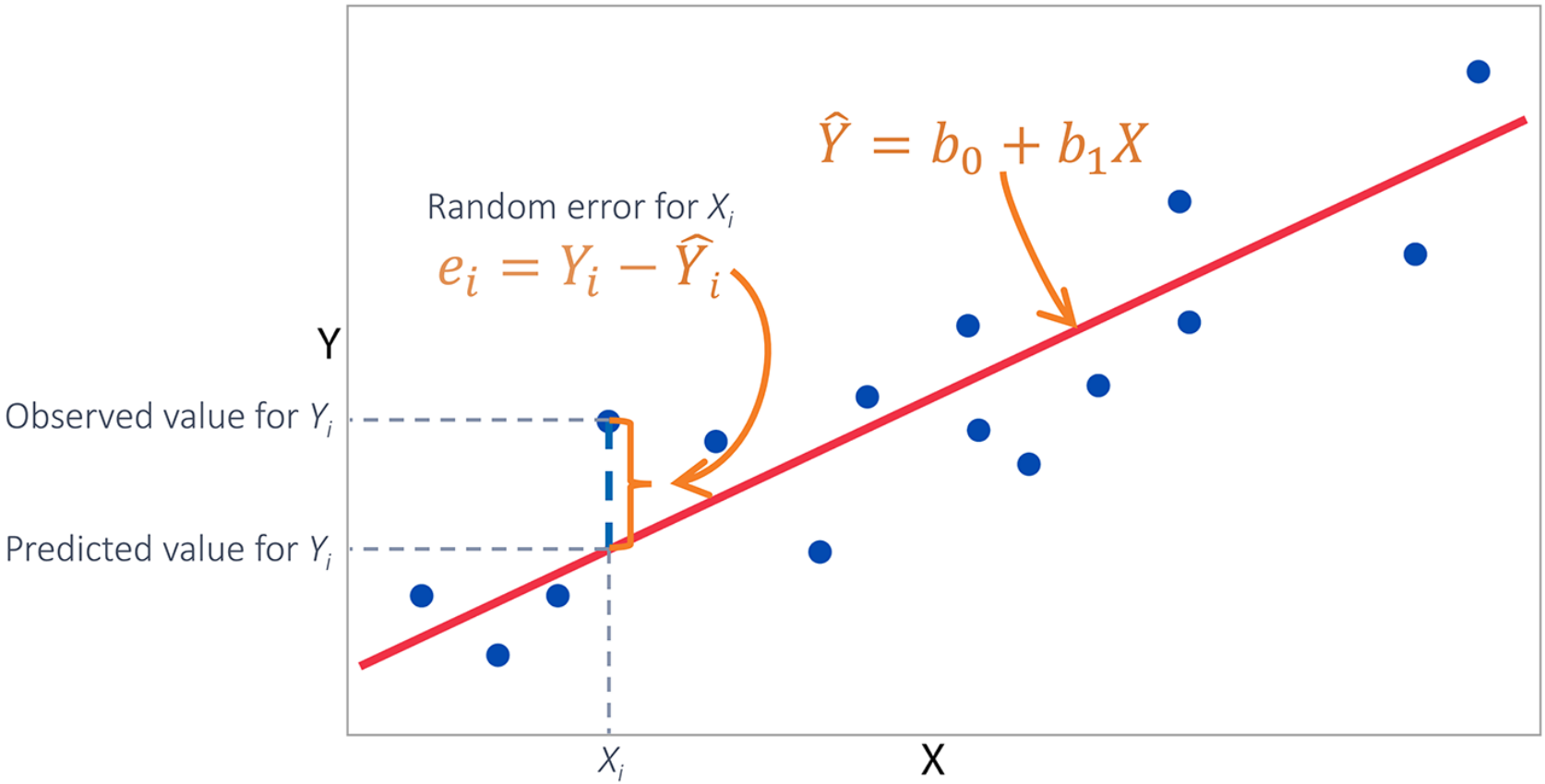
단순 선형 회귀

Definition

단순 선형 회귀는 종속변수를 하나의 독립변수와의 선형관계로 설명하는 회귀모형이다. 두 개 이상의 독립 변수로 설명하는 경우에는 다중회귀라도고 부른다. 단순 선형 회귀 모델은 다음과 같다.

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

단순 선형 회귀


$\hat{Y} = b_0 + b_1 X$
Random error for X_i
$e_i = Y_i - \hat{Y}_i$
Y
Observed value for Y_i
Predicted value for Y_i
X_i
X

최소 제곱법(Least Square Error)

Definition

선형 회귀의 회귀계수는 최소 제곱법(최소 자승법)으로 추정한다. 최소 제곱법이란 제곱의 합 $\Sigma\epsilon^2$ 이 최소가 되도록 값을 정하는 방법으로 선형회귀에서는 오차의 제곱 합이 최소가 되도록 회귀계수를 정한다. 예를 들어 단순 선형 회귀의 경우 다음을 최소로 만든다.

$$\Sigma(Y_i - (\hat{Y}_i))^2$$

Important theorem

최소 제곱법에 따르면 단순 선형 회귀의 계수 β_1, β_0 는 다음과 같이 정의할 수 있다.

- $\hat{\beta}_1 = \frac{\Sigma X_i(Y_i - \bar{Y})}{\Sigma x_i(X_i - \bar{X})}$
- $\hat{\beta}_0 = \bar{Y} - \beta_1 \bar{X}$

최소 제곱법(Least Square Error)

$$\hat{Y} = \beta_0 + \beta_1 X$$

$$\epsilon_i = Y_i - \hat{Y}_i$$

$$\begin{aligned} \Sigma\epsilon_i^2 &= \Sigma(Y_i - \hat{Y})^2 \\ &= \Sigma(Y_i - (\beta_0 + \beta_1 X))^2 \end{aligned}$$

$$\frac{\partial \Sigma \epsilon^2}{\partial \beta_0} = 2N\beta_0 + 2\beta_1 \Sigma X_i - 2\Sigma Y_i$$

$$\begin{aligned} \beta_0 &= \frac{\Sigma Y_i - \beta_1 \Sigma X_i}{N} \\ &= \bar{Y} - \beta_1 \bar{X} \end{aligned}$$

$$\begin{aligned} \frac{\partial \Sigma e^2}{\partial \beta_1} &= 2\beta_1 \Sigma X_i^2 + 2\beta_0 \Sigma X_i - 2\Sigma Y_i X_i \\ &= 2\beta_1 \Sigma X_i^2 + 2(\bar{Y} - \beta_1 \bar{X})\Sigma X_i - 2\Sigma Y_i X_i \end{aligned}$$

$$\beta_1 = \frac{\Sigma X_i (Y_i - \bar{Y})}{\Sigma X_i (X_i - \bar{X})}$$

다중 선형 회귀

Definition

다중 회귀는 하나 이상의 독립 변수가 사용된 선형 회귀이다. 즉 종속변수가 p개의 독립 변수로 설명되는 다중 선형 회귀 모델은 다음과 같다.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + ... + \beta_p X_{pi} + \epsilon_i$$

$$Y = X\beta + \epsilon$$

$$\mathrm{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} \mathrm{x}_1^T \\ \mathrm{x}_2^T \\ \vdots \\ \mathrm{x}_n^T \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \boldsymbol{\epsilon} = \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{pmatrix}$$

Important theorem

최소 제곱법에 따르면 β 는 다음과 같이 정의할 수 있다.

- $\hat{\beta} = (X^t X)^{-1} X^t Y$

선형 회귀의 모형 검정

- 회귀 계수의 T 검정
- 회귀 모델의 F 검정
- 결정계수 R^2

회귀 계수의 T 검정

Definition

T 검정의 경우 각 독립변수가 개별적으로 얼마나 유의한지를 판단하는 것이다. 이 때 사용되는 귀무가설은 '계수가 0 이다'이고 대립가설은 '계수가 0이 아니다'이다. 만약 해당 회귀계수의 값이 유의하지 않는다고 나온다면 그 회귀계수는 사실상 0으로 간주해도 된다.

Assumptions

- 귀무가설 H0: $\beta_i = 0$
- 대립가설 H1: $\beta_i \neq 0$

회귀 모델의 F 검정

Definition

F 검정은 T 검정과 다항 회귀식 전체에 대한 유의성을 검정한다. F 통계량은 MSR/MSE의 비율로 모델이 통계적으로 얼마나 의미가 있는지를 설명한다. 즉 모든 회귀계수가 0 이라는 귀무가설의 기각 여부를 검정하는 것인데 귀무가설이 기각되지 않고 채택된다면 해당 회귀 모델은 의미가 없게 된다.

Assumptions

- 귀무가설 H0: $\beta_0 = \beta_1 = ... = \beta_p = 0$
- 대립가설 H1: $\beta_0, \beta_1, ..., \beta_p$ 중 적어도 하나는 0이 아니다

결정계수 R^2

Definition

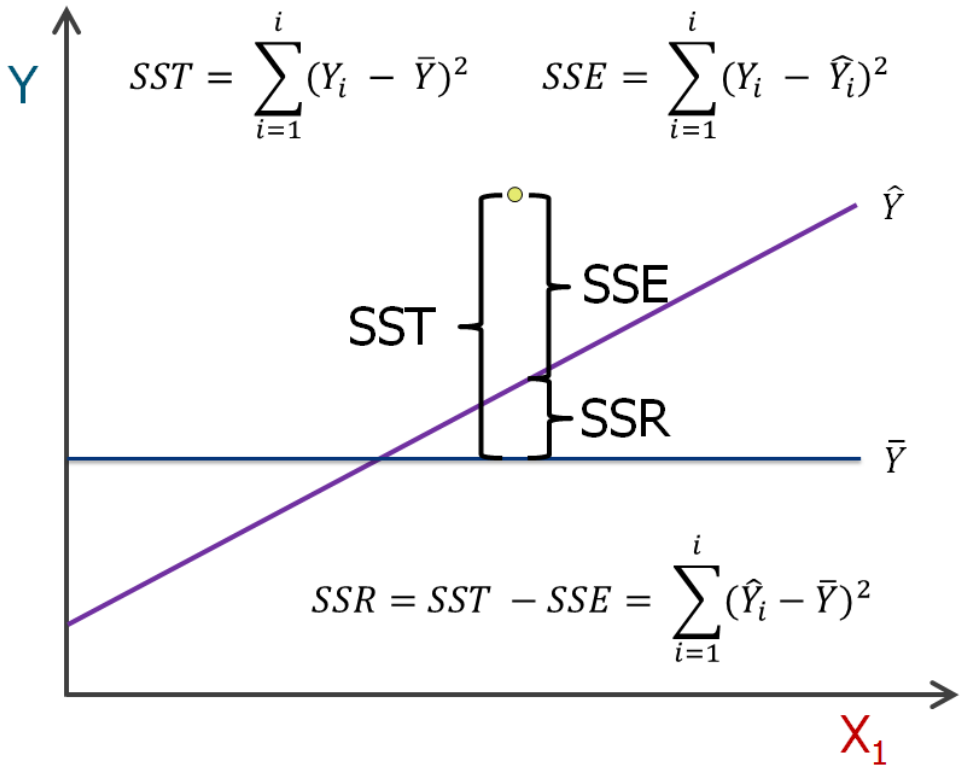
회귀모델의 검증력은 R^2로 볼 수 있다. R^2는 전체 변동 중 설명된 변동의 비율로 범위는 $0 \leq R^2 \leq 1$을 만족하며, 1에 가까울수록 회귀 모델이 데이터를 더 잘 설명한다고 말한다.

Important theorem

- $SST = \Sigma(Y_i - \bar{Y})^2$
- $SSR = \Sigma(\hat{Y}_i - \bar{Y})^2$
- $R^2 = \frac{SST}{SSR}$

결정계수 R^2

SST는 관측된 Y_i 값들이 평균 $\hat{Y}$로부터 얼마나 떨어져 있는지를 뜻하며, SSR은 추정치 $\hat{Y}_i$가 평균 $\bar{Y}$로부터 얼마나 떨여져 있는지를 뜻한다. 따라서 이 둘의 비율인 R^2은 Y_i의 총 변동에 대비해 회귀 모델이 얼마나 그 변동을 설명하는지를 알려준다.



변수 선택

Definition

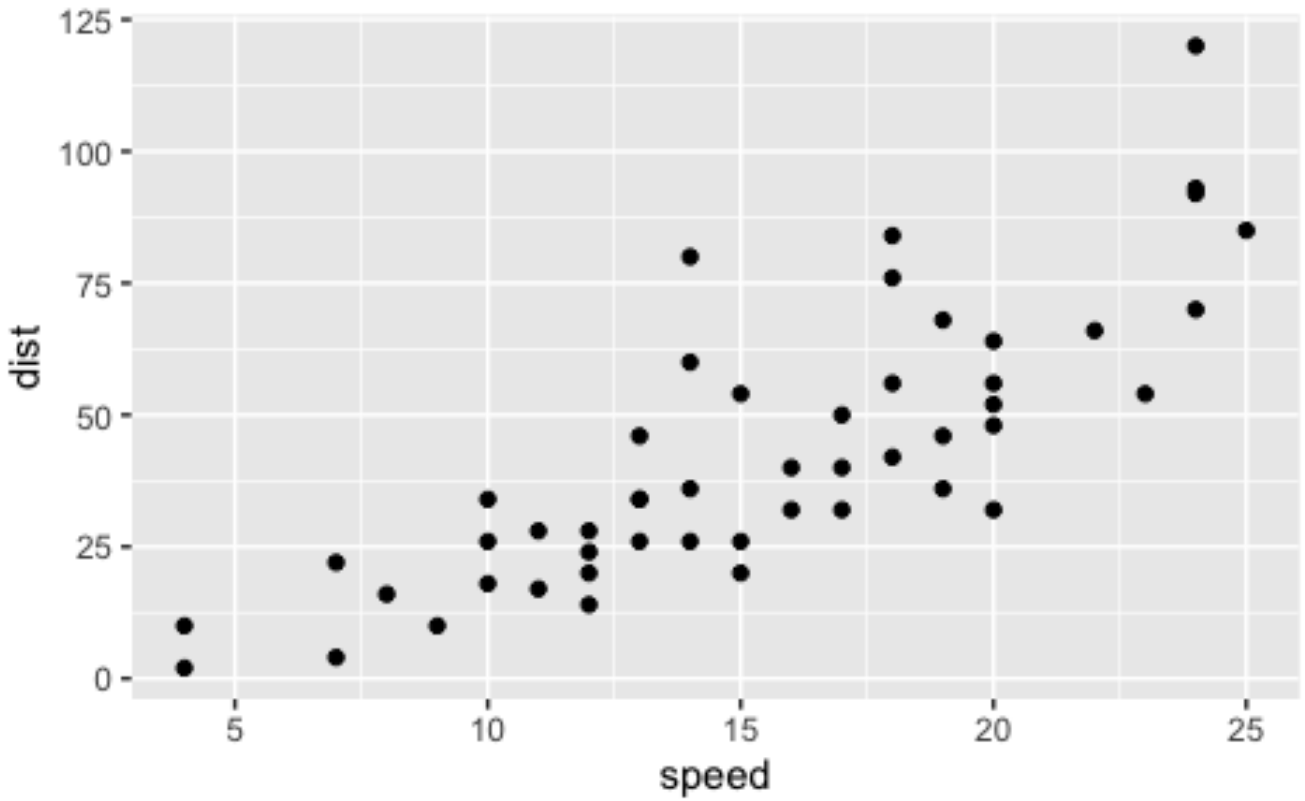
다중 회귀 모형에서 설명 변수를 선택하는 방법은 F 통계량이나 AIC을 사용하여 통계량을 높일 수 있는 변수를 하나씩 택하거나 제거하는 방식이다.

- 전진 소거법(Forward selection): 절편만 있는 모델에서 통계치를 가장 많이 개션시키는 변수를 차례대로 추가하는 방법이다.
- 변수 소거법(Backward elimination): 모든 변수가 포함된 모형에서 통계치에 가장 도움이 안되는 변수를 하나씩 제거하는 방법이다.
- 단계적 방법(Stepwise selection): 변수의 추가와 삭제를 반복하며 통계치가 최대가 되는 지점을 찾는다.

Important theorem

AIC 란 회귀 모델을 평가하는 척도 중 하나이다. AIC는 절대적인 모델의 성능에 대해서는 알려주지 못하지만 여러가지 모형이 있었을 때 어떤 모형이 더 성능이 좋은지 상대적인 비교에서는 이용할 수 있다. 따라서 전반적인 모형을 선택한 다음에 그 모형을 발전시키기 위해 변수를 선택하는 과정에서 사용하기에 적합하다.

모형 해석



모형 해석

자동차 주행 속도와 제동 거리에 대한 선형 회귀 모델을 보고 이를 해석해 보자.

모형 해석

	Dependent variable:
	dist (Std.Error)
Constant	-17.579^{**} (6.758)
speed	3.932^{***} (0.416)
Observations	50
R^2	0.651
Adjusted R^2	0.644
Residual Std. Error	15.380 (df = 48)
F Statistic	89.567^{***} (df = 1; 48)
Note:	*p<0.1; **p<0.05; ***p<0.01

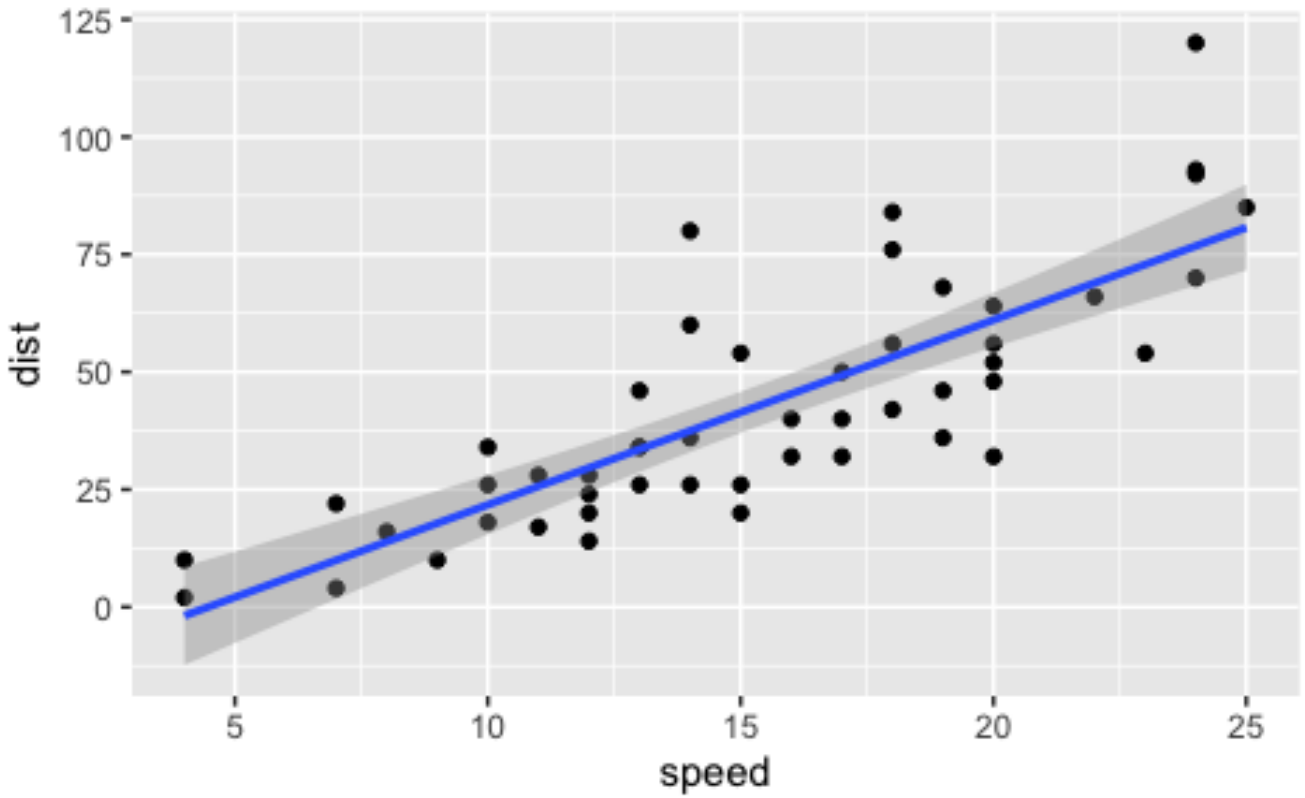
모형 해석

회귀식은 다음과 같이 정의할 수 있다.

$$Dist = -17.579 + 3.932 Speed$$

- *, **, ***로 표시된 문자열은 p-value의 범위를 뜻한다. 만약 유의한 계수가 있다면 그 정도에 따라 별로 표시되며 아무런 표시가 없거나 점이면 통계적으로 유의하지 않다는 의미이다.
- 상수항은 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 상수항의 계수가 0이 아니라고 말할 수 있다.
- Speed는 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 speed의 계수가 0이 아니라고 말할 수 있다.
- F 통계량은 89.567 로 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 β_0, β_1 중 적어도 하나는 0 이 아니라고 결론을 내릴 수 있다.
- R^2 는 0.651이다. 즉 모델은 자료의 65 퍼센트의 변동성을 설명할 수 있다고 말할 수 있다.

모형 해석



모형 해석

모형 해석

회귀식은 다음과 같이 정의할 수 있다.

$$Sales = 8.12 + 0.05 Speed$$

- 상수항은 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 상수항의 계수가 0이 아니라고 말할 수 있다.
- TV는 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 TV의 계수가 0이 아니라고 말할 수 있다.
- F 통계량은 210.8로 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 β_0, β_1 중 적어도 하나는 0 이 아니라고 결론을 내릴 수 있다.
- R^2 는 0.6373이다. 즉 모델은 자료의 63 퍼센트의 변동성을 설명할 수 있다고 말할 수 있다.

모형 해석

회귀식은 다음과 같이 정의할 수 있다.

$$Sales = 3.39 + 0.05\,TV + 0.19\,Radio - 0.01\,Newspaper$$

- 상수항은 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 상수항의 계수가 0이 아니라고 말할 수 있다.
- TV, Radio는 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 TV의 계수가 0이 아니라고 말할 수 있다.
- Newspaper는 p-value가 0.05보다 크다. 따라서 귀무가설을 기각하지 못하고 Newspaper의 계수가 0이라고 말할 수 있다.
- F 통계량은 445.9로 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 $\beta_0, \beta_1, \beta_2, \beta_3$ 중 적어도 하나는 0 이 아니라고 결론을 내릴 수 있다.
- R^2 는 0.9189이다. 즉 모델은 자료의 91.89 퍼센트의 변동성을 설명할 수 있다고 말할 수 있다.

모형 해석

회귀식은 다음과 같이 정의할 수 있다.

$$Sales = 3.18 + 0.05\,TV + 0.19\,Radio$$

- 상수항은 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 상수항의 계수가 0이 아니라고 말할 수 있다.
- TV, Radio는 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 TV의 계수가 0이 아니라고 말할 수 있다.
- F 통계량은 658.2로 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 $\beta_0, \beta_1, \beta_2$ 중 적어도 하나는 0 이 아니라고 결론을 내릴 수 있다.
- R^2 는 0.9179이다. 즉 모델은 자료의 91.79 퍼센트의 변동성을 설명할 수 있다고 말할 수 있다.