Linear Regression

두 변수 사이의 관계

- 대략적 파악 : 산점도(scatter plot)
- 상관분석(correlation analysis)
  - 두 변수 사이의 상관관계 분석
  - 확률변수 $X, Y \rightarrow \rho = \mathrm{Corr}(X, Y)$ - 직선적인 관련성 파악
- 회귀분석(regression analysis)
  - 두 변수 사이의 함수관계를 분석
  - x : 독립변수 또는 설명변수, Y : 종속변수 또는 반응변수

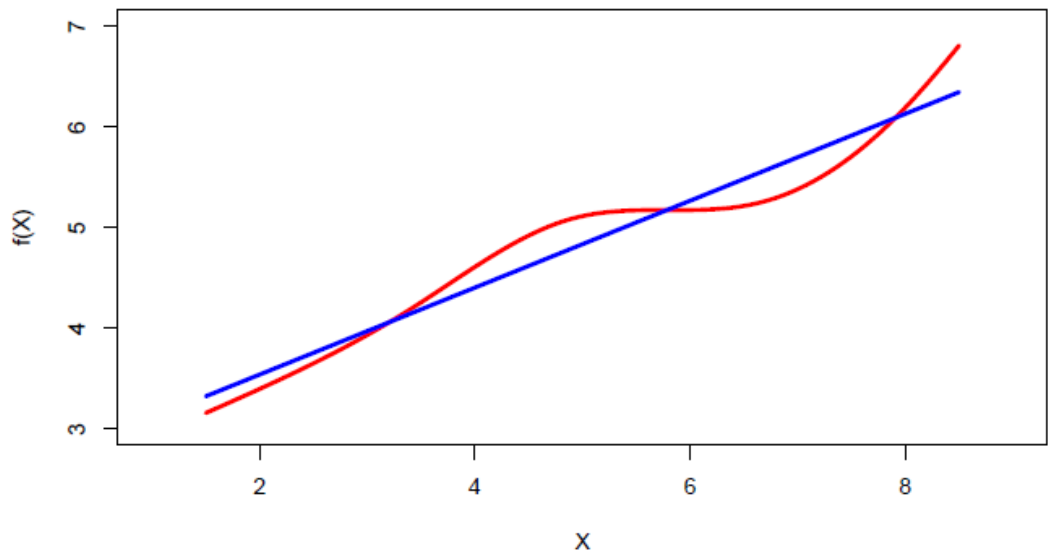
    $Y = f(x) + \epsilon$, ϵ : 오차항 $\rightarrow f(x)$?
  - 단순선형회귀분석 - 직선관계를 모형으로 분석

    $$(f(x) = a + bx)$$
  - 중회귀분석 - 두 개 이상의 설명변수 사용

    $$(f(x) = a + b_1 x_1 + \cdots + b_k x_k)$$

Logistic Regression

- 선형 회귀모형 : 대표적인 지도학습 (supervized learning)
- 가정 : Y와 $X_1, \ldots, X_p$ 사이에 선형관계가 있음
- 실제 회귀함수는 선형이 아니다



Example : Advertising data

	TV	radio	newspaper	sales
1	230.10	37.80	69.20	22.10
2	44.50	39.30	45.10	10.40
3	17.20	45.90	69.30	9.30
4	151.50	41.30	58.50	18.50
5	180.80	10.80	58.40	12.90
6	8.70	48.90	75.00	7.20
⋮	⋮	⋮	⋮	⋮

Table: Advertising data, $n = 200$

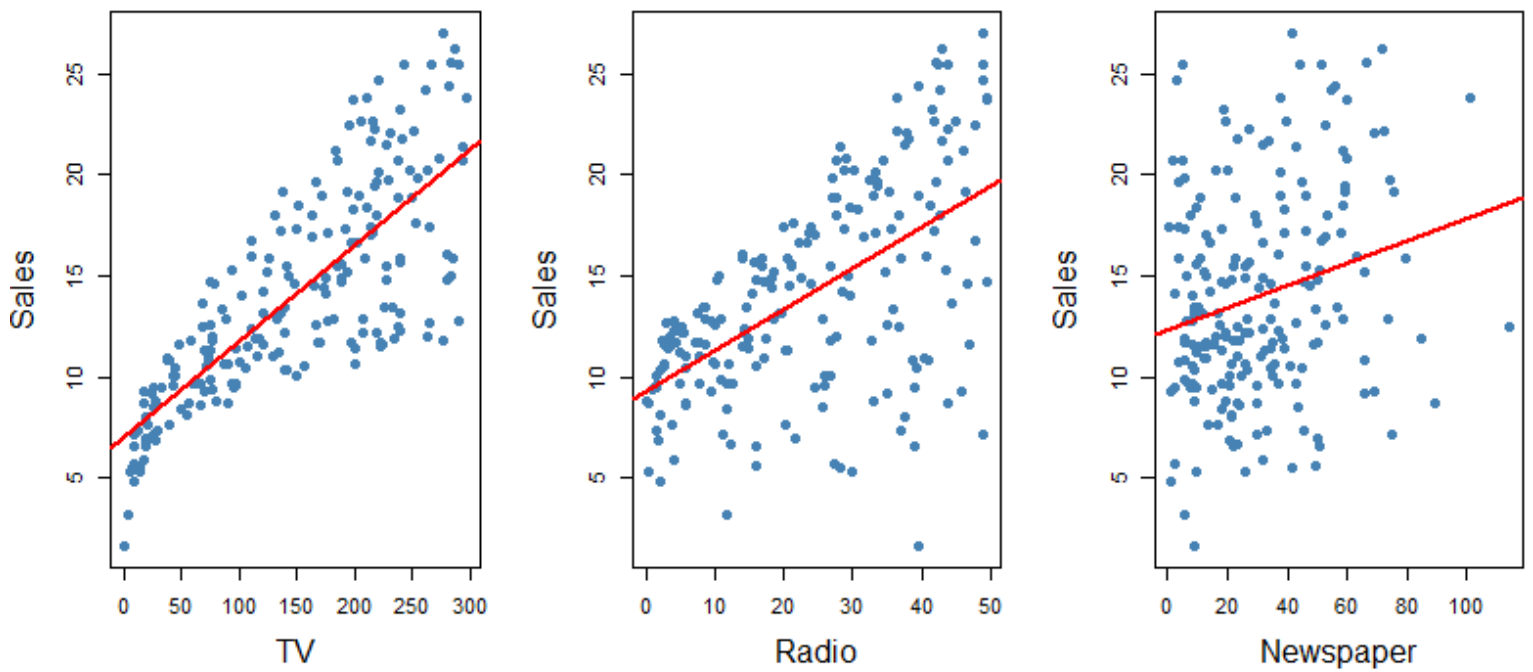
Linear Regression for Advertising data

- Questions :

  - 광고 예산과 판매량과의 관계가 있는가?
  - 광고 예산과 판매량과의 관계가 얼마나 강한가?
  - 어떤 미디어가 판매량에 기여하는가?
  - 미래 판매량을 얼마나 정확하게 예측할 수 있는가?
  - 선형 관계가 있는가?
  - 광고 매체간에 시너지 효과가 있는가?

Advertising data

Figure: Scatter plot



Simple Linear Regression Model

- Model

$$Y = \beta_0 + \beta_1 X + \epsilon$$

- ϵ : 오차항(random error)

  서로 독립이면서 평균이 0, 분산이 σ^2 인 확률 변수

- 회귀계수(regression coefficient) (or 모수, parameter)
  - β_0 : 상수항 또는 절편 (constant coefficient or intercept)
  - β_1 : 기울기 (slope)
- 예측 (Prediction) : $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$

Least Square Estimation (LSE)

- 최소제곱법(method of least squares)에 의한 추정

  - $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$: $X = x_i$ 일 때 Y의 예측값
  - 잔차(residual) : $e_i = y_i - \hat{y}_i$
  - 잔차 제곱합 (residual (or error) sum of squares : SSE)

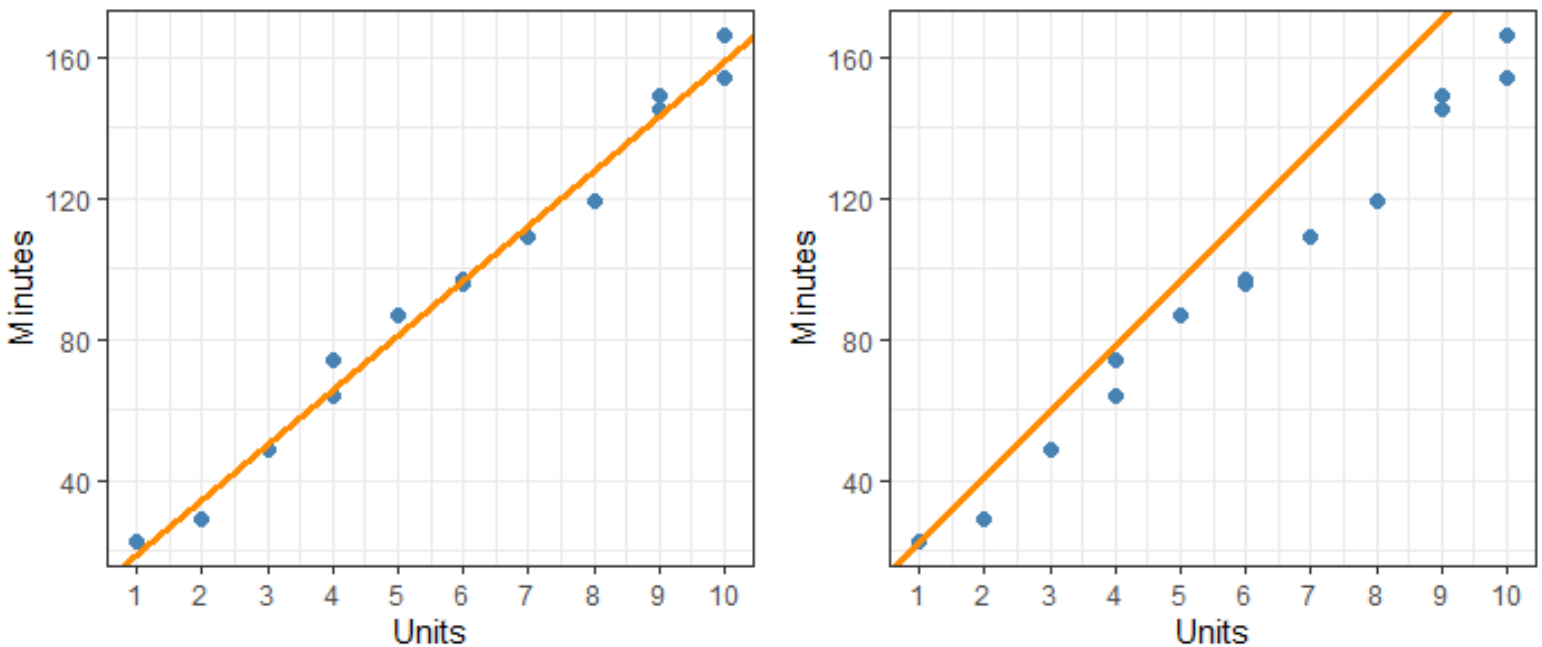
$$SSE = e_1^2 + \cdots + e_n^2 = \sum_{i=1}^{n} \left\{ y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_i) \right\}^2$$

  - 최소제곱추정량 (LSE)

$$(\hat{\beta}_0, \hat{\beta}_1) = \underset{\beta_0, \beta_1 \in \mathbb{R}}{\operatorname{argmin}} \sum_{i=1}^{n} \{ y_i - (\beta_0 + \beta_1 x_i) \}^2$$

Least Square Estimation (LSE)

Figure: 회귀직선 비교



Least Square Estimation (LSE)

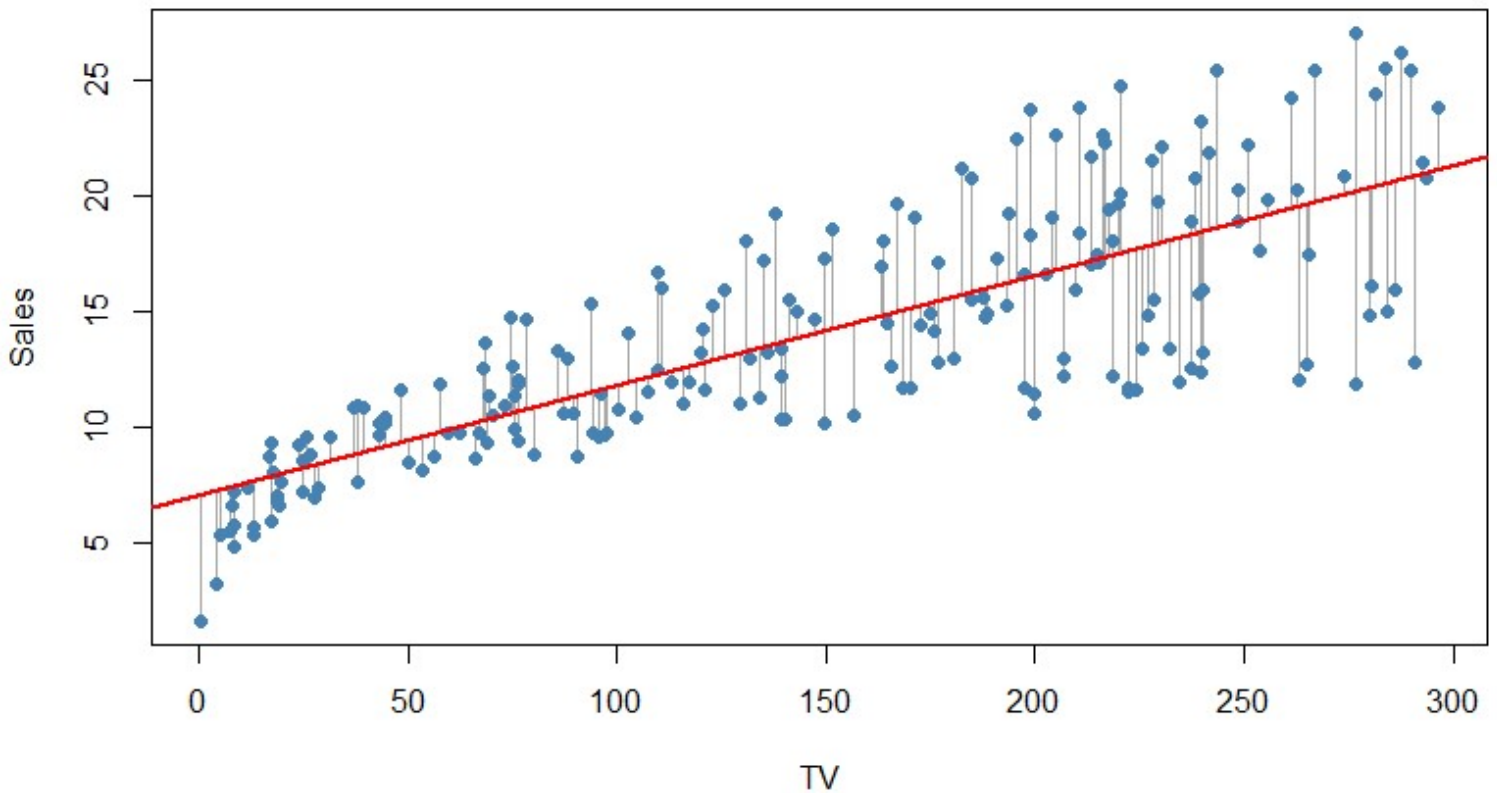
- 최소제곱추정량

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$\bar{x} \equiv \frac{1}{n}\sum_{i=1}^{n} x_i, \bar{y} \equiv \frac{1}{n}\sum_{i=1}^{n} y_i$: 표본평균 (sample mean)

Advertising data



$$\hat{\beta}_0 = 7.0326, \ \hat{\beta}_1 = 0.0475$$

Estimation of error variance

- 잔차 (residual) : $e_i = y_i - \hat{y}_i,\ (\sum_{i=1}^n e_i = 0, \sum_{i=1}^n x_i e_i = 0)$

- 오차분산 (σ^2)의 추정:

  - 잔차(오차) 제곱합 :

$$SSE = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \sum_{i=1}^{n} e_i^2$$

  - 평균제곱오차 (mean squared error) : $MSE = \dfrac{SSE}{n-2}$
  - 오차분산의 추정값 : $\hat{\sigma}^2 = MSE$

Decomposition of deviations

- 총편차의 분해

  - $y_i - \bar{y} = (y_i - \hat{y}_i) + (\hat{y}_i - \bar{y}), \ \ \forall i$
  - 총편차(total deviation) $= y_i - \bar{y}$
  - 추측값의 편차 $= (\hat{y}_i - \bar{\hat{y}}) = (\hat{y}_i - \bar{y})$

    $\Rightarrow$ 총편차 $=$ 잔차 $+$ 추측값의 편차

Decomposition of sum of squares

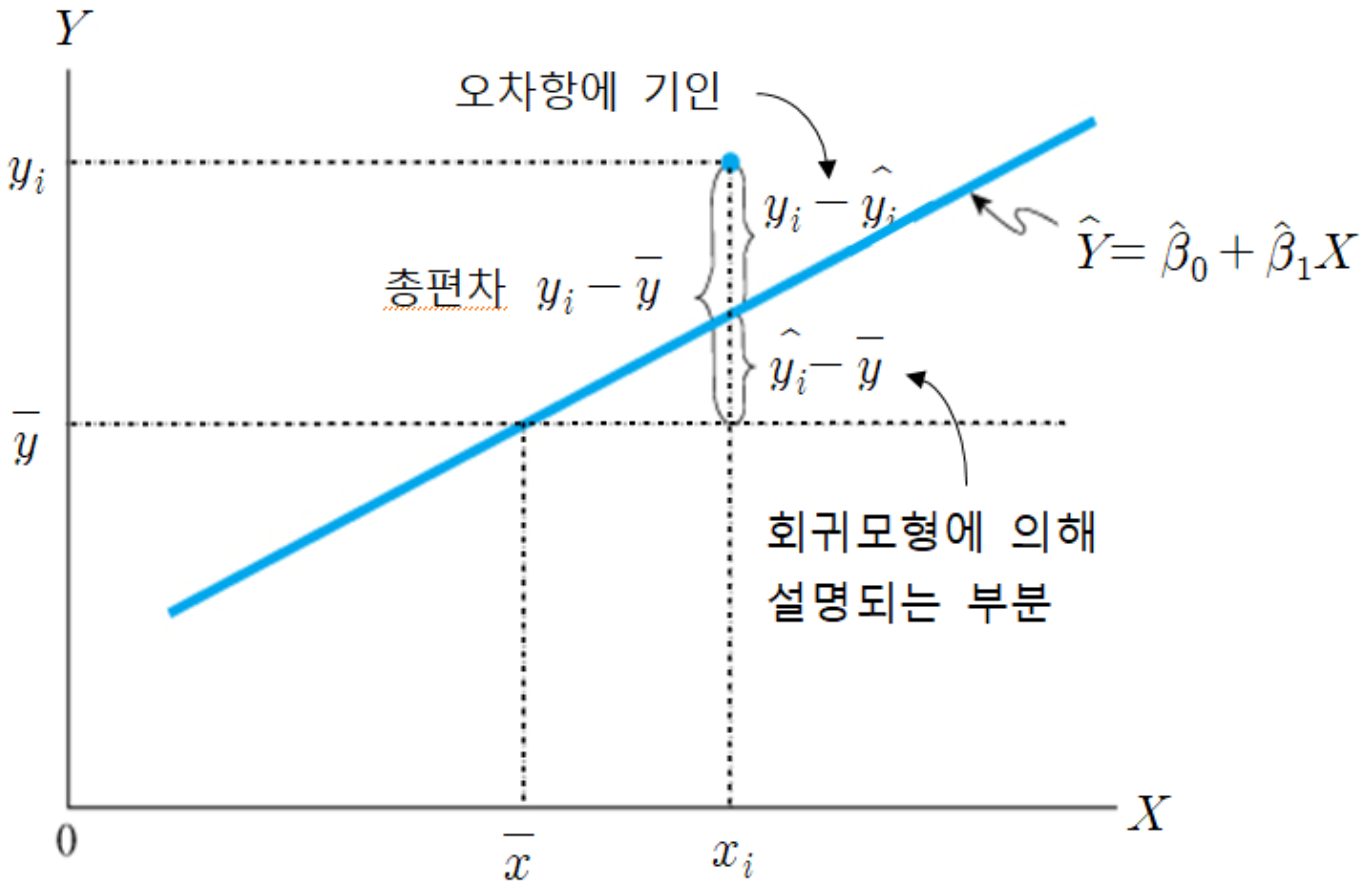
- 제곱합의 분해 : SST = SSE + SSR

$$\sum_{i=1}^{n}(y_i - \bar{y})^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

제곱합의 종류	정의 및 기호	자유도
총제곱합 (total sum of squares)	$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2$	$n-1$
잔차제곱합 (residual sum of squares)	$SSE = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$n-2$
회귀제곱합 (regression sum of squares)	$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$	1

Decomposition of sum of squares

Figure: 편차의 분해



Coefficient of determination

- 결정계수 (Coefficient of determination)
  - 정의 : $R^2 = \frac{SSR}{SST} = 1 - \frac{SSE}{SST}$
  - 의미 : 회귀직선의 기여율

    (총변동 가운데 회귀직선으로 설명되는 변동의 비율)
  - 성질
    - $0 \leq R^2 \leq 1$
    - R^2 값이 1에 가까울수록 회귀에 의한 설명이 잘 됨을 뜻함
    - $R^2 = r^2$ (r : sample correlation)

      (단순선형회귀모형에서만 성립)

회귀직선의 유의성 검정

- Model : $Y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \;\; i = 1, 2, \ldots, n, \;\; \epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

- 회귀직선의 유의성 검정 (F-test)

  - 가설 : $H_0 : \beta_1 = 0 \; vs. \; H_1 : \beta_1 \neq 0$
  - 검정통계량 : $F = \frac{MSR}{MSE} = \frac{SSR/1}{SSE/(n-2)} \sim_{H_0} F(1, n-2)$
  - 검정통계량의 관측값 : f
  - 유의수준 α에서의 기각역 : $f \geq F_\alpha(1, n-2)$
  - 유의확률 $= P(F \geq f)$

회귀직선의 유의성 검정

- 회귀직선의 유의성 검정을 위한 분산분석표

요인	제곱합(SS)	자유도(df)	평균제곱(MS)	f	유의확률
회귀	SSR	1	$MSR = \frac{SSR}{1}$	$f = \frac{MSR}{MSE}$	$P(F \geq f)$
잔차	SSE	$n-2$	$MSE = \frac{SSE}{n-2}$		
계	SST	$n-1$			

Advertising data

- $R^2 = \dfrac{SSR}{SST} = 0.6099$
- 회귀직선의 유의성 검정 : $H_0 : \beta_1 = 0 \; vs. \; H_1 : \beta_1 \neq 0$

요인	제곱합	자유도	평균제곱	f	유의확률
회귀	3314.6	1	3314.6	312.14	< 0.0001
잔차	2102.5	198	$10.6 (= \hat{\sigma}^2)$		
계	5417.1	199			

회귀계수에 대한 추론

모회귀계수(기울기) β_1에 대한 추론

- β_1의 최소제곱추정량 : $\hat{\beta}_1 = \frac{S_{xY}}{S_{xx}}$
- 추정값 : $\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}$
- 추정량 $\hat{\beta}_1$ 의 분포 : $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{S_{xx}}\right)$
- studentized $\hat{\beta}_1$의 분포 : $\frac{\hat{\beta}_1 - \beta_1}{\hat{\sigma}/\sqrt{S_{xx}}} \sim t(n-2)$, $\hat{\sigma} = \sqrt{MSE}$
- $\hat{\beta}_1$의 $100(1-\alpha)\%$ 신뢰구간 : $\hat{\beta}_1 \pm t_{\alpha/2}(n-2)\hat{\sigma}/\sqrt{S_{xx}}$

회귀계수에 대한 추론

모회귀계수(기울기) β_1에 대한 추론

- 가설검정 : $H_0 : \beta_1 = \beta_1^0$
- 검정통계량 : $T = \dfrac{\hat{\beta}_1 - \beta_1^0}{\hat{\sigma}/\sqrt{S_{xx}}} \sim_{H_0} t(n-2)$, 관측값 : t

대립가설	유의확률	유의수준 α 기각역
$H_1 : \beta_1 > \beta_1^0$	$P(T \geq t)$	$t \geq t_\alpha(n-2)$
$H_1 : \beta_1 < \beta_1^0$	$P(T \leq t)$	$t \geq t_\alpha(n-2)$
$H_1 : \beta_1 \neq \beta_1^0$	$P(\|T\| \geq \|t\|)$	$\|t\| \geq t_{\alpha/2}(n-2)$

회귀계수에 대한 추론

모회귀계수(절편) β_0에 대한 추론

- β_0의 최소제곱추정량 : $\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{x}$
- 추정값 : $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- 추정량 $\hat{\beta}_0$ 의 분포 : $\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2\left(\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}\right)\right)$

$$\frac{\hat{\beta}_0 - \beta_0}{s.e.(\hat{\beta}_0)} \sim t(n-2), \quad s.e.(\hat{\beta}_0) = \hat{\sigma}\sqrt{\frac{1}{n} + \frac{\bar{x}^2}{S_{xx}}}$$

- $\hat{\beta}_0$의 $100(1-\alpha)\%$ 신뢰구간 : $\hat{\beta}_0 \pm t_{\alpha/2}(n-2)s.e.(\hat{\beta}_0)$

회귀계수에 대한 추론

모회귀계수(기울기) β_0에 대한 추론

- 가설검정 : $H_0 : \beta_0 = \beta_0^0$
- 검정통계량 : $T = \dfrac{\hat{\beta}_0 - \beta_0^0}{s.e.(\hat{\beta}_0)} \sim_{H_0} t(n-2)$, 관측값 : t

대립가설	유의확률	유의수준 α 기각역
$H_1 : \beta_0 > \beta_0^0$	$P(T \geq t)$	$t \geq t_\alpha(n-2)$
$H_1 : \beta_0 < \beta_0^0$	$P(T \leq t)$	$t \geq t_\alpha(n-2)$
$H_1 : \beta_0 \neq \beta_0^0$	$P(\lvert T\rvert \geq \lvert t\rvert)$	$\lvert t\rvert \geq t_{\alpha/2}(n-2)$

Advertising data

- $\hat{\beta}_0$와 $\hat{\beta}_1$의 95% 신뢰구간 ($t_{0.05/2}(198) \approx 2$)

$$\begin{aligned}
\hat{\beta}_0 \pm t_{\alpha/2} s.e.(\hat{\beta}_0) &= 7.0326 \pm 2 \times 0.4578 \\
&= 7.0326 \pm 0.9156 = (6.117, 7.9482) \\
\hat{\beta}_1 \pm t_{\alpha/2} s.e.(\hat{\beta}_1) &= 0.0475 \pm 2 \times 0.0027 \\
&= 0.0475 \pm 0.0054 = (0.0421, 0.0529)
\end{aligned}$$

Advertising data

- $H_0 : \beta_0 = 0 \; vs. \; H_1 : \beta_0 \neq 0$ 에 대한 가설검정 $(\alpha = 0.05)$
  - 검정통계량 관측값 : $t = \dfrac{\hat{\beta}_0 - 0}{s.e.(\hat{\beta}_0)} = \dfrac{7.0326}{0.4578} = 15.3617$
  - 기각역 : $|t| \geq t_{0.05/2}(198) \approx 2$
  - 결과 : 기각!, 유의확률 $= < 0.001$
- $H_1 : \beta_1 = 0 \; vs. \; H_1 : \beta_1 \neq 0$ 에 대한 가설검정 $(\alpha = 0.05)$
  - 검정통계량 관측값 : $t = \dfrac{\hat{\beta}_1 - 0}{s.e.(\hat{\beta}_1)} = \dfrac{0.0475}{0.0027} = 17.5926$
  - 기각역 : $|t| \geq t_{0.05/2}(198) \approx 2$
  - 결과 : 기각!, 유의확률 $= < 0.001$

Advertising data

- 회귀계수의 유의성 검정

	Estimate	Std. Error	t	p-value
(Intercept)	7.032594	0.457843	15.36	< 0.0001
TV	0.047537	0.002691	17.67	< 0.0001

중회귀모형(Multiple Linear Regression)

설명변수가 p개인 다중(선형)회귀모형

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots + \beta_p x_{ip} + \epsilon_i \quad i = 1, 2, \ldots, n$$

- 회귀모수 : $\beta_0, \beta_1, \ldots, \beta_p$
- 설명변수(독립변수) :

  $X_1 = (x_{11}, \ldots, x_{n1})^T, \ldots, X_p = (x_{1p}, \ldots, x_{np})^T$
- 반응변수(종속변수) : $Y = (y_1, \ldots, y_n)^T$
- 오차항 : $\epsilon_1, \ldots, \epsilon_n$, $(\sim_{i.i.d} N(0, \sigma^2))$

중회귀모형(Multiple Linear Regression)

- Model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_p X_p + \epsilon$$

- β_j : the average effect on Y of a one unit increase in X_j, holding all other predictors fixed
- Advertising data

$$\text{Sales} = \beta_0 + \beta_1 \times \text{TV} + \beta_2 \times \text{radio} + \beta_3 \times \text{newspaper} + \epsilon$$

중회귀모형(Multiple Linear Regression)

- 기본 가정 : 설명변수들 사이에 상관관계가 없다.

  (uncorrelated)
- 설명변수들 사이에 상관관계가 있는 경우 : 다중공선성
  - 회귀계수의 분산이 증가
  - X_j가 변할 때, 다른 변수들도 변함

    $\Rightarrow$ 회귀계수의 설명이 어려워짐

중회귀모형(Multiple Linear Regression)

- George Box

  "Essentially, all models are wrong, but some are useful"

- Fred Mosteller and John Tukey

  "The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively"

Least Square Estimation

- 최소제곱추정량 :

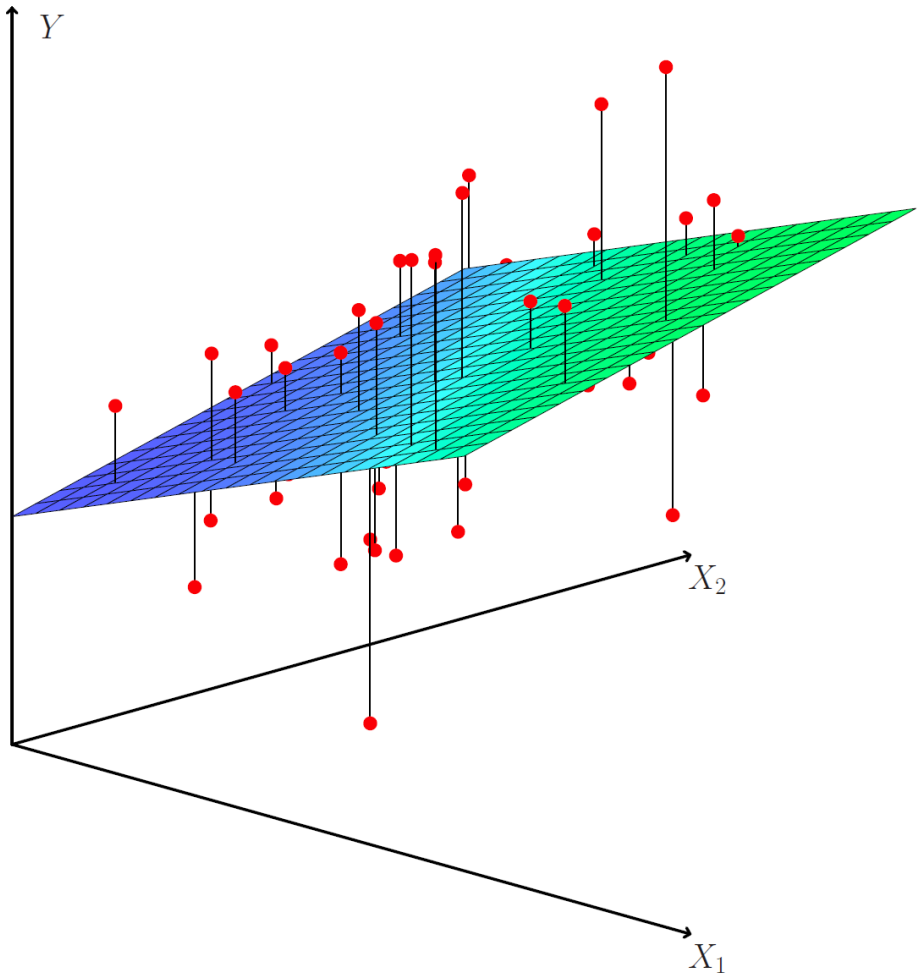
$$\left(\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right) =$$

$$\underset{(\beta_0,\ldots,\beta_p)\in\mathbb{R}^{p+1}}{\operatorname{argmin}} \sum_{i=1}^{n} \{y_i - (\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})\}^2$$

- 예측값 :

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_p x_p$$

Multiple Linear Regression



Advertising data

- 회귀계수의 유의성 검정

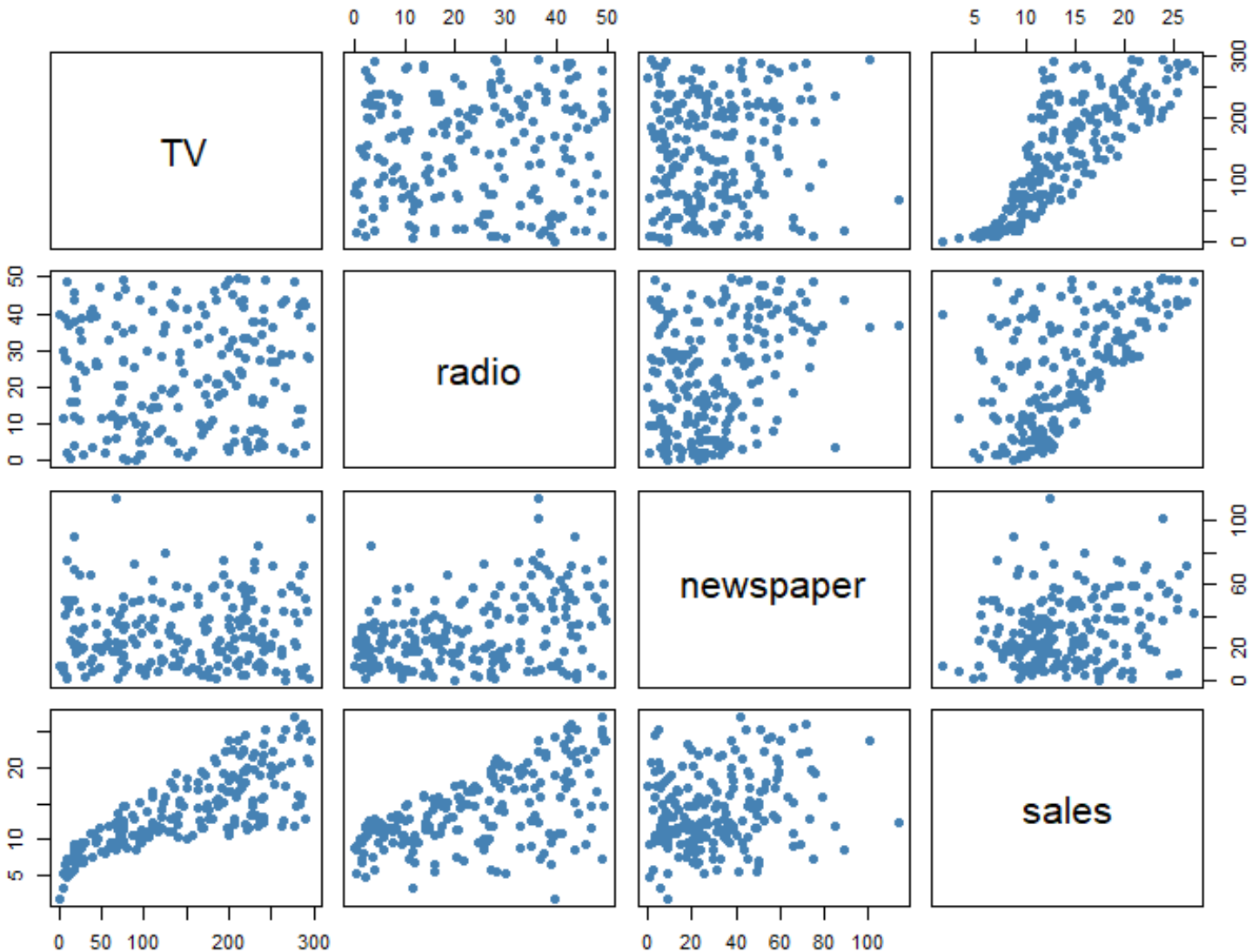
	Estimate	Std. Error	t	p-value
(Intercept)	2.94	0.31	9.42	< 0.0001
TV	0.05	0.00	32.81	< 0.0001
radio	0.19	0.01	21.89	< 0.0001
newspaper	-0.00	0.01	-0.18	0.8599

Advertising data

- 분산분석표 (ANOVA table)

	Df	Sum Sq	Mean Sq	F value	p-value
TV	3314.62	1	3314.62	1166.73	0.0000
radio	1545.62	1	1545.62	544.05	0.0000
newspaper	0.09	1	0.09	0.03	0.8599
Residuals	556.83	196	2.84	$(= \hat{\sigma}^2)$	
Total	5417.1	199			

Advertising data



Advertising data

- Correlation

	TV	radio	newspaper	sales
TV	1.0000	0.0548	0.0566	0.7822
radio		1.0000	0.3541	0.5762
newspaper			1.0000	0.2283
sales				1.0000

Some important question

- Is at least one of the predictors $X_1, X_2, \ldots, X_p$ useful in predicting the response?
- Do all the predictors help to explain Y, or is only a subset of the predictors useful?
- How well does the model fit the data?
- Given a set of predictor values, what response value should we predict, and how accurate is our prediction?

Coefficient of determination

- 결정계수 (Coefficient of determination)

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

- 수정된 결정계수 (Adjusted multiple correlation coefficient)

$$R^2_{adj} = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

Advertising data

- Sales $= \beta_0 + \beta_1 \times$ TV $+ \beta_2 \times$ radio $+ \beta_3 \times$ newspaper $+ \epsilon$
  - 결정계수 $R^2 = 0.8972106$
  - 수정된 결정계수 $R^2_{adj} = 0.8956$
- Sales $= \beta_0 + \beta_1 \times$ TV $+ \beta_2 \times$ radio $+ \epsilon$
  - 결정계수 $R^2 = 0.8971943$
  - 수정된 결정계수 $R^2_{adj} = 0.8962$

Model Selection

- Statistics used in model selection

  - Residual mean squares error (MSE) : $MSE = \dfrac{SSE_p}{(n-p-1)}$
  - coefficient of determination : $R^2 = \dfrac{SSR}{SST} = 1 - \dfrac{SSE_p}{SST}$
  - Adjusted R^2 : $R^2_{adj} = 1 - \dfrac{SSE_p/(n-p-1)}{SST/(n-1)}$
  - Partial F-test statistics

Variable Selection

- Variable selection
  - All possible regression : 모든 가능한 회귀
  - Backward Elimination : 후진 제거법
  - Forward Selection : 전진 선택법
  - Stepwise regression : 단계별 회귀

Variable Selection

- All possible regression

  - 모든 가능한 변수들의 조합(2^p)을 회귀분석하여 결과 비교
  - 시간이 오래 걸림
  - R^2또는 MSE 사용

Variable Selection

- Backward Elimination

  (step 0) 모든 변수를 포함한 회귀모형 적합 (Full Model)

  (step 1) p-value가 가장 큰 변수 제거

  (step 2) 제거된 변수를 제외한 나머지 변수를 이용하여 회귀모형 적합

  (step 3) 적당한 정지 규칙 (stopping rule)을 만족시킬 때까지 step1,2 반복
  (예 : 모든 회귀계수가 유의한 경우 멈춤, 최종모형으로 선택)

Variable Selection

- Forward Selection

  (step 0) 변수 하나하나씩에 대한 회귀모형 적합 후 R^2 를 가장 크게 하는 설명변수 선택

  (step 1) 변수를 하나하나씩 추가하여 R^2를 가장 크게 하는 변수 선택하여 모형에 포함

  (step 2) 적당한 정지 규칙 (stopping rule)을 만족시킬 때까지 step1 반복 (예 : 모형에 포함된 변수에 대한 p-value가 어느 수준 이상 (ex. 0.1) 이 되면 변수를 추가하지 않고 멈춤. 최종모형으로 선택)

etc.

- 회귀모형에서 더 생각해보아야 하는 것
  - 이상점 (Outlier), 영향점 (Leverage)
  - 설명변수가 범주형 자료인 경우의 분석법
  - 다중공선성(Collinearity)
  - 잔차분석
  - 교호작용
  - etc.