Data Mining

HW#3

학번	182STG18
이름	이하경
제출일	2019.04.03

Description

Simple & Multiple Linear Regression

반응변수가 크기가 존재하는 양적변수의 형태일 때, 연속형 및 범주형 설명변수들이 반응변수와 선형 관계가 존재한다고 가정하고 간단하며 유용한 단순 또는 다중 선형회귀모형을 적합할 수 있다.

$$Simple\ LR:\ Y = \beta_0 + \beta_1 X_1 + \epsilon \rightarrow \hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1$$

$$Multiple\ LR:\ Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon \rightarrow \hat{y} = \widehat{\beta_0} + \widehat{\beta_1} x_1 + \cdots + \widehat{\beta_p} x_p$$

회귀계수의 추정을 위해서 RSS(Residual Sum of Squares), 즉 실제 Y 와 예측치의 차이의 제곱합을 최소화하는 추정치를 선택하는 Least Square Estimation 을 사용한다.

선형 회귀모형은 모형에 포함된 설명변수가 서로 연관되지 않고 독립적일 때 가장 이상적이며, 이 때 각 회귀계수는 다른 모든 설명변수가 동일하다고 가정할 때 X 가 1 단위 증가할 경우 Y 의 변화량이라고 해석할 수 있다. 그러나 많은 경우 설명변수들 사이의 상관관계가 존재하며 심한 경우 회귀계수의 분산이 커져 모형의 해석이 불안정해지므로, 가능한 모든 설명변수를 사용하는 것보다 서로 독립적이며 Y 에 대한 설명력이 큰 변수들로 적절히 모형을 구성하는 것이 좋다.

본 과제에서는 4 가지 예제를 통해 선형 회귀모형을 실제 데이터 셋 또는 랜덤하게 생성한 데이터 셋에 대해 적용해보고 다양한 plot 및 요약 통계량을 통해 모형의 결과를 해석한다. 또한 설명변수들 간의 상관관계, 교호작용 변수 포함 여부, 비선형 관계, 변수의 변환 등 선형 회귀모형에서 발생할 수 있는 다양한 Issue 에 대해 고려해본다.

Results

3.6 Lab: Linear Regression

Chapter 3 의 Lab 에서는 lm() 함수로 단순 및 다중 선형회귀모형을 적합하고 summary 및 diagnostic plot 등을 그려 모형의 결과를 해석하는 기초를 다졌다. linear model object 를 plot function 에 입력하는 것 외에 직접 몇 가지 함수를 사용해 잔차 대 예측치 산점도나 High Leverage Point 를 나타내는 plot 을 그려보고, 신뢰구간 및 예측구간을 계산하였다. 또한 lm 함수 안에서 변수의 변환을 하거나 교호작용 추가에 대해 직접 코드를 실행해 복습할 수 있었다. 다음은 잘 사용하지 않았거나 새롭게 알게된 유용한 함수들이다.

- **residuals**(model), **rstudent**(model), **hatvalues**(model): 잔차 및 표준화 잔차, hat matrix 를 계산해 모형 진단에 사용
- **vif**(model): Variance Inflation Factors 계산으로 변수 간 다중공선성 진단할 수 있음
- **poly**(x, n): x 의 1 차항부터 n 차항으로 이루어진 다항 행렬, lm() 함수 안에서 유용하게 사용
- **contrasts**(): 범주형 변수를 자동으로 0 과 1 로 이루어진 dummy 변수들로 자동 coding

Exercise 3.8

(a) Simple Linear Regression of mpg vs horsepower

Call: lm(formula = mpg ~ horsepower, data = Auto)

Residuals:

Min	1Q	Median	3Q	Max
-13.571	-3.259	-0.344	2.763	16.924

Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	39.936	0.717	55.66	<2e-16 ***
horsepower	-0.158	0.006	-24.49	<2e-16 ***

Residual standard error: 4.906 on 390 degrees of freedom

Multiple R-squared: 0.6059, Adjusted R-squared: 0.6049

F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16

i. 회귀계수 검정의 p-value가 0에 가까우며 horsepower와 mpg 사이의 상관관계가 존재한다.

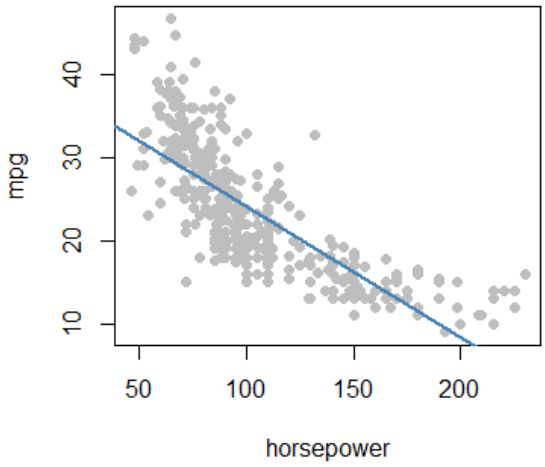
ii. 결정계수의 크기가 0.6059이며, 상관계수는 -0.7784으로 설명변수와 반응변수의 관계가 강하다.

iii. 회귀계수의 추정치 $\beta_1 = -0.158$으로 음수이므로 음의 관계가 있음을 알 수 있다.

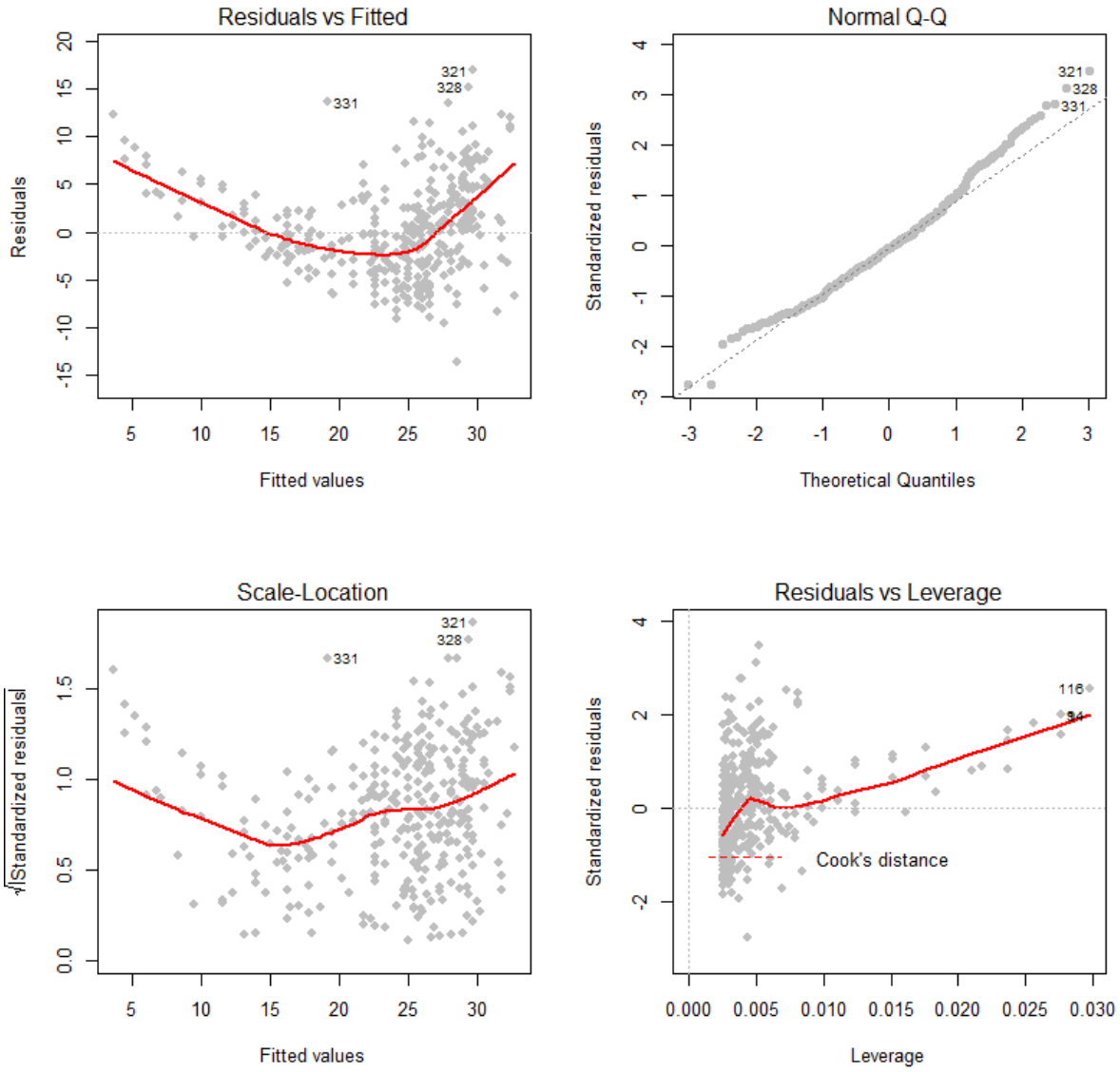
iv. horsepower가 98일 때 mpg의 예측 값과 95% CI, PI는 아래와 같다.

Interval	estimate	2.5%	97.5%
confidence	24.467	23.973	24.961
prediction	24.467	14.809	34.125

(b) Scatterplot & Regression Line



(c) Diagnostic Plots of the Least Squares Regression Fit



- 잔차(및 표준 잔차) 대 예측치 산점도에서 비선형 관계가 존재하므로 선형 모형 가정이 적절하지 않다.
- 116 번째 관측치와 34 번 관측치가 high leverage point 로 나타났다.

Exercise 3.9

(a) Scatterplot Matrix of all variables (except for 'name')

- 대부분의 설명변수가 반응변수 mpg 와 선형 또는 비선형 관계가 존재하는 것으로 보인다.
- 설명변수 중 cylinders 와 displacement, horsepower, weight 의 경우 특히 선형 상관관계가 존재하는 것으로 보인다. 다음의 (b)에서 해당 변수들 사이의 상관계수가 매우 크다. 이 설명변수들이 모형에 함께 포함될 경우 회귀계수 추정과 모형의 정확도에 악영향을 미칠 수 있다.

(b) Correlations between variables

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg	1.0000	-0.7776	-0.8051	-0.7784	-0.8322	0.4233	0.5805	0.5652
cylinders	-0.7776	1.0000	**0.9508**	**0.8430**	**0.8975**	-0.5047	-0.3456	-0.5689
displacement	-0.8051	**0.9508**	1.0000	**0.8973**	**0.9330**	-0.5438	-0.3699	-0.6145
horsepower	-0.7784	**0.8430**	**0.8973**	1.0000	**0.8645**	-0.6892	-0.4164	-0.4552
weight	-0.8322	**0.8975**	**0.9330**	**0.8645**	1.0000	-0.4168	-0.3091	-0.5850
acceleration	0.4233	-0.5047	-0.5438	-0.6892	-0.4168	1.0000	0.2903	0.2127
year	0.5805	-0.3456	-0.3699	-0.4164	-0.3091	0.2903	1.0000	0.1815
origin	0.5652	-0.5689	-0.6145	-0.4552	-0.5850	0.2127	0.1815	1.0000

(c) Multiple Linear Regression of mpg vs all predictors (-name)

Call: lm(formula = mpg ~ . - name, data = Auto)

Residuals:

Min	1Q	Median	3Q	Max
-9.590	-2.157	-0.117	1.869	13.060

Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	-17.2184	4.6443	-3.7074	0.0002 ***
cylinders	-0.4934	0.3233	-1.5261	0.1278
displacement	0.0199	0.0075	2.6474	0.0084 **
horsepower	-0.0170	0.0138	-1.2295	0.2196
weight	-0.0065	0.0007	-9.9288	< 2e-16 ***
acceleration	0.0806	0.0988	0.8152	0.4155
year	0.7508	0.0510	14.7288	< 2e-16 ***
origin	1.4261	0.2781	5.1275	4.67e-07 ***

Residual standard error: 3.328 on 384 degrees of freedom

Multiple R-squared: 0.8215, Adjusted R-squared: 0.8182

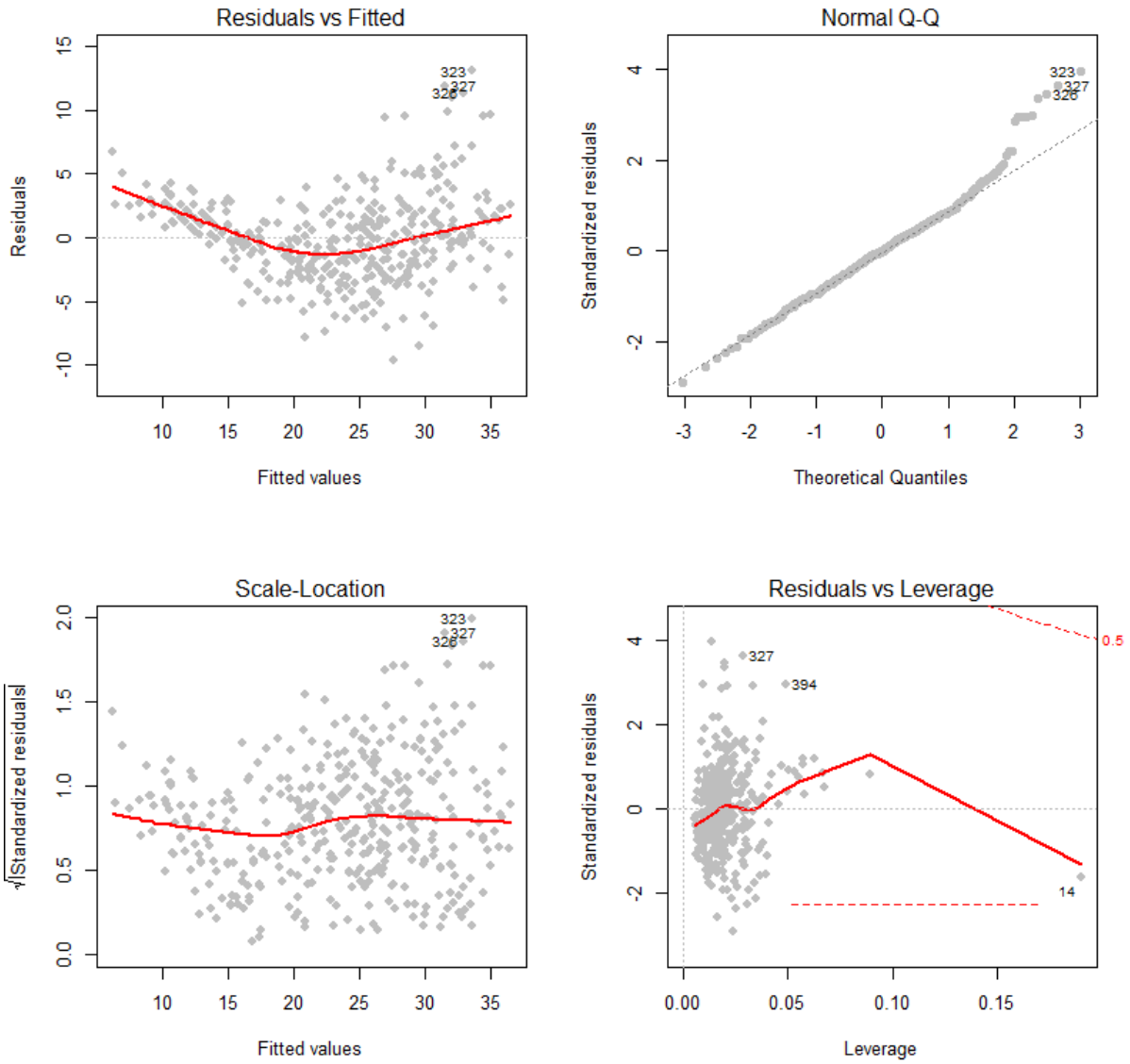
F-statistic: 252.4 on 7 and 384 DF, p-value: < 2.2e-16

i. 설명변수들과 반응변수 mpg 사이의 설명관계가 존재한다.

ii. 개별 회귀계수 검정의 p-value 를 확인해볼 때 displacement, weight, year, origin 이 유의하다.

iii. year 의 회귀계수는 0.7508 으로 다른 설명변수가 모두 동일할 때 model year(출시년도)가 1 년 증가함에 따라 mpg 가 약 0.75 증가한다고 할 수 있다. 따라서 자동차가 최신일수록 갤런당 마일수가 증가한다.

(d) Diagnostic Plots of the Linear Regression Fit



- 잔차 대 예측치 산점도에서 비선형 관계가 보이지만 표준화 잔차와의 산점도에서는 완화되었다. 그러나 예측치가 커질수록 잔차의 분산이 커지는 형태이므로 잔차의 등분산성 가정을 만족한다고 할 수 없다.
- 323, 326, 327 번째 관측치의 잔차가 눈에 띄게 커 outlier 로 발견되었다.
- 14 번째 관측치가 High leverage point 로 나타났다.

▷ Variance Inflation Factor

- 아래의 VIF 에서 cylinders, weight, acceleration 의 VIF 가 10 이상이며 displacement 의 VIF 는 20 이상으로 다중공선성 문제가 존재하는 것으로 판단된다.

cylinders	displacement	horsepower	weight	acceleration	year	origin
10.7375	21.8368	9.9437	10.8313	2.6258	1.2450	1.7724

(e) Regression Fit Including Interaction Effects

Call: lm(formula = mpg ~ . * ., data = Auto[, -9])

Residuals:

Min	1Q	Median	3Q	Max
-7.630	-1.448	0.060	1.274	11.139

Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	35.479	53.136	0.668	0.505
cylinders	6.989	8.248	0.847	0.397
displacement	-0.479	0.189	-2.527	0.012 *
horsepower	0.503	0.347	1.451	0.148
weight	0.004	0.018	0.235	0.814
acceleration	-5.859	2.174	-2.696	0.007 **
year	0.697	0.610	1.144	0.253
origin	-20.896	7.097	-2.944	0.003 **
cylinders:displacement	-0.003	0.007	-0.524	0.601
cylinders:horsepower	0.012	0.024	0.480	0.632
cylinders:weight	0.000	0.001	0.399	0.690
cylinders:acceleration	0.278	0.166	1.670	0.096
cylinders:year	-0.174	0.097	-1.793	0.074
cylinders:origin	0.402	0.493	0.816	0.415
displacement:horsepower	0.000	0.000	-0.294	0.769
displacement:weight	0.000	0.000	1.682	0.093
displacement:acceleration	-0.004	0.003	-1.041	0.299
displacement:year	0.006	0.002	2.482	0.014 *
displacement:origin	0.024	0.020	1.232	0.219
horsepower:weight	0.000	0.000	-0.673	0.501
horsepower:acceleration	-0.007	0.004	-1.939	0.053
horsepower:year	-0.006	0.004	-1.482	0.139
horsepower:origin	0.002	0.029	0.076	0.939
weight:acceleration	0.000	0.000	1.025	0.306
weight:year	0.000	0.000	-1.056	0.292
weight:origin	-0.001	0.002	-0.364	0.716
acceleration:year	0.056	0.026	2.174	0.030 *
acceleration:origin	0.458	0.157	2.926	0.004 **
year:origin	0.139	0.074	1.882	0.061

Residual standard error: 2.695 on 363 degrees of freedom

Multiple R-squared: 0.8893, Adjusted R-squared: 0.8182808

F-statistic: 104.2 on 7 and 363 DF, p-value: < 2.2e-16

- 설명변수들 간의 교호작용을 모두 포함한 Full 모형에서 개별 회귀계수 검정의 p-value 를 확인해볼 때 유의수준 0.05 에서 의미 있는 교호작용 효과는 displacement 와 year, acceleration 과 year, acceleration 과 origin 이다. 세 교호작용 효과의 계수는 모두 양수로 변수들 사이의 시너지 효과가 mpg 에 대해 긍정적인 영향을 미친다고 할 수 있다.

(d) Transformations of the variables

Call: lm(formula = mpg ~ cylinders + poly(displacement, 2) + log(horsepower) + log(weight) + poly(acceration, 2) + year + origin)

Residuals:

Min	1Q	Median	3Q	Max
-9.860	-1.582	0.005	1.602	12.194

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	90.8024	13.4916	6.7303	6.22e-11 ***
cylinders	0.3265	0.3097	1.0544	0.2924
poly(displacement, 2)1	-13.0248	15.4139	-0.8450	0.3986
poly(displacement, 2)2	19.3083	4.3148	4.4749	1.01e-05 ***
log(horsepower)	-7.4481	1.4916	-4.9935	9.03e-07 ***
log(weight)	-11.8211	2.0699	-5.7110	2.26e-08 ***
poly(acceleration, 2)1	-11.7272	5.4386	-2.1563	0.0317 *
poly(acceleration, 2)2	10.0217	3.4398	2.9135	0.0038 **
year	0.7670	0.0447	17.1644	<2e-16 ***
origin	0.5298	0.2645	2.0030	0.0459 *

Residual standard error: 2.89 on 382 degrees of freedom

Multiple R-squared: 0.8661, Adjusted R-squared: 0.8629

F-statistic: 274.5 on 9 and 382 DF, p-value: < 2.2e-16

일부 변수에 2 차항 또는 로그 변환 등을 진행한 결과 적절하게 나타났다. displacement 와 acceleration 의 경우 1 차항보다 2 차항의 회귀계수 검정의 p-value 가 더 낮았으며 horsepower 와 weight 변수는 로그 변환을 하는 것이 더 mpg 를 잘 설명한다고 할 수 있다. 모형에서 결정계수와 조정 결정계수가 변환을 하지 않은 (c)의 다중선형회귀모형과 비교하였을 때 향상되었고 잔차의 표준오차 또한 줄어들었다.

Exercise 3.13

(a)-(c) Generate x, eps and y

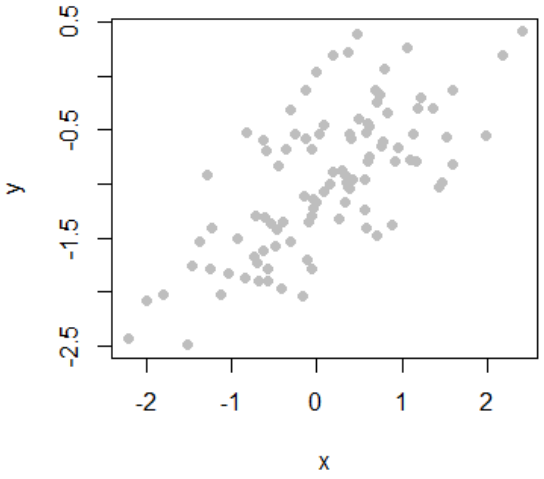
$$Y = -1 + 0.5X + \epsilon, \qquad \epsilon \sim iid\ N(0, 0.5^2)$$

$\mathrm{N}(0,1)$으로부터 100 개의 x(seed 로 고정)와 $\mathrm{N}(0,0.5^2)$로부터 100 개의 eps 를 랜덤으로 생성하여 x 와 y 로 이루어진 데이터를 만들었다. y vector 의 길이는 x 와 eps 과 동일하게 100 개의 점을 포함하고 있다.

위의 식에서 $\boldsymbol{\beta_0} = -\mathbf{1},\ \boldsymbol{\beta_1} = \mathbf{0.5}$이다.

(d) Scatterplot: (y, x)

오른쪽의 산점도에서 x 와 y 가 양의 선형관계로 이루어져 있음을 확인할 수 있다.



(e) Least Square Linear Regression Model

Call: lm(formula = y ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-0.93842	-0.30688	-0.06975	0.26970	1.17309

Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	**-1.0189**	0.04849	-21.010	<2e-16 ***
x	**0.4995**	0.05386	9.273	4.58e-15 ***

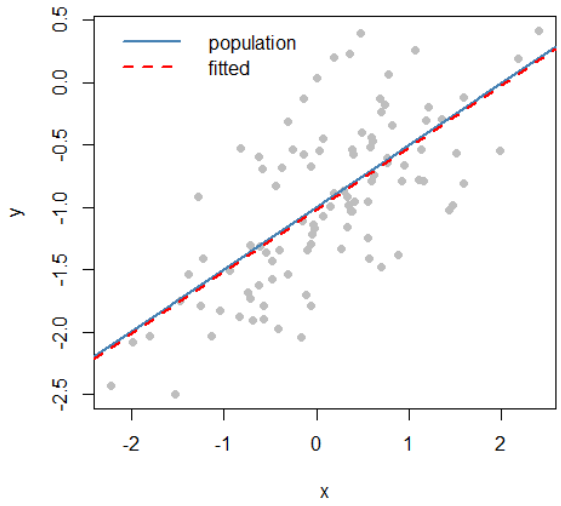
Residual standard error: 0.4814 on 98 degrees of freedom

Multiple R-squared: 0.4674, Adjusted R-squared: 0.4619

F-statistic: 85.99 on 1 and 98 DF, p-value: 4.58e-15

x 와 y 의 단순선형회귀모형에서 추정된 회귀계수는 $\widehat{\boldsymbol{\beta_0}} = -\mathbf{1.0189}$, $\widehat{\boldsymbol{\beta_1}} = \mathbf{0.4995}$으로 실제 값 -1 와 0.5 에 거의 가깝다. 회귀계수 검정의 p-value 역시 매우 작다.

(f) Scatterplot: Least Square Line & Population Regression Line



β_0, β_1의 실제 값과 추정치를 이용해 실제 회귀선과 추정된 회귀선을 산점도에 나타냈을 때 두 직선이 거의 동일하게 그려지는 것을 확인할 수 있다.

(g) Polynomial Regression Model with using x and x^2

▷ Summary of Fitted Model

Call: lm(formula = y ~ poly(x, 2))

Residuals:

Min	1Q	Median	3Q	Max
-0.98252	-0.31270	-0.06441	0.29014	1.13500

Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	**-0.9645**	0.0479	-20.134	<2e-16 ***
poly(x, 2)1	**4.4638**	0.4790	9.319	3.97e-15 ***
poly(x, 2)2	**-0.6720**	0.4790	-1.403	0.164

Residual standard error: 0.479 on 97 degrees of freedom

Multiple R-squared: 0.4779, Adjusted R-squared: 0.4672

F-statistic: 44.4 on 2 and 97 DF, p-value: 2.038e-14

▷Aalysis of Variance Table

Model 1: y ~ x

Model 2: y ~ poly(x, 2)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	98	22.709				
2	97	22.257	1	0.4516	1.9682	0.1638

x의 2차항을 추가하여 모형을 적합하였을 때 2차항의 추가 설명력이 존재하지 않으며 적합도가 향상되었다고 할 수 없다.

(h) Repeat (a)-(f) using *less* noise

$$Y_{less} = -1 + 0.5X + \epsilon, \qquad \epsilon \sim N(0, 0.1^2)$$

▷ Summary of Fitted Model

Call: lm(formula = y_less ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-0.2914	-0.0482	-0.0045	0.0649	0.2642

Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	**-0.9973**	0.0105	-95.25	<2e-16 ***
x	**0.5021**	0.0116	43.17	<2e-16 ***

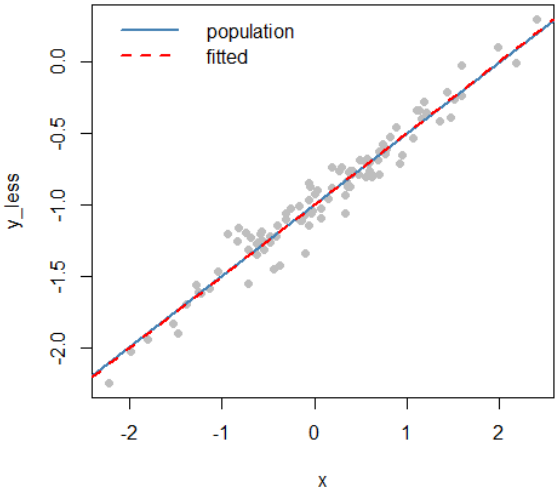
Residual standard error: 0.1039 on 98 degrees of freedom

Multiple R-squared: 0.9501, Adjusted R-squared: 0.9495

F-statistic: 1864 on 1 and 98 DF, p-value: < 2.2e-16

▷ **Scatterplot with Population & Fitted Regression Line**

Error 의 표준편차를 더 작게 **0.1** 로 가정하고 y 를 다시 생성한 후 단순선형회귀모형을 적합하였다. 추정된 회귀계수는 $\widehat{\beta_0} = -0.9973$, $\widehat{\beta_1} = 0.5021$이다. 산점도 그림 결과 점들이 더 좁은 폭으로 분포하며 실제 회귀선과 추정된 회귀선이 이전에 비해 더 겹쳐 보인다.



(i) Repeat (a)-(f) using *more* noise

$$Y_{more} = -1 + 0.5X + \epsilon, \qquad \epsilon \sim N(0, 1^2)$$

▷ **Summary of Fitted Model**

Call: lm(formula = y_more ~ x)

Residuals:

Min	1Q	Median	3Q	Max
-2.5163	-0.5453	-0.0378	0.6729	1.8789

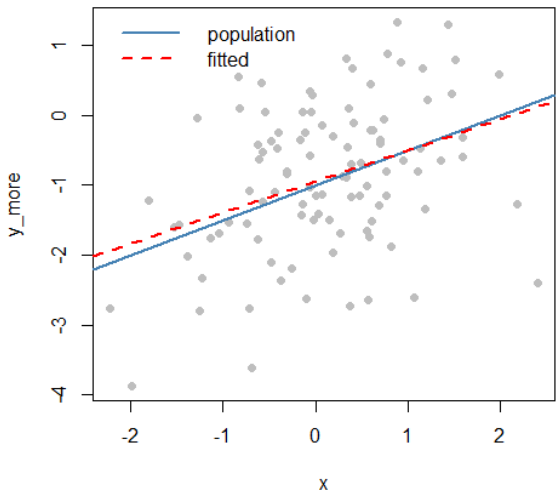
Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	**-0.9423**	0.1003	-9.397	2.47e-15 ***
x	**0.4443**	0.1114	3.989	0.0001 ***

Residual standard error: 0.9955 on 98 degrees of freedom

Multiple R-squared: 0.1397, Adjusted R-squared: 0.1309

F-statistic: 15.91 on 1 and 98 DF, p-value: 0.000128



▷ **Scatterplot with Population & Fitted Regression Line**

Error 의 분산을 **1** 으로 더 크게 가정하고 모형을 적합한 결과 유의수준 0.05 에서 x 가 y 와 선형관계가 존재한다고 할 수 있지만 추정의 정확도가 전에 비해 떨어지는 것을 알 수 있다. 추정 회귀계수는 $\widehat{\beta_0} = -0.9423$, $\widehat{\beta_1} = 0.4443$으로, 산점도에서 역시 점들이 매우 넓게 분포하여 양의 상관관계가 쉽게 보이지 않고 추정된 회귀선이 실제 선과 일치하지 않는 것이 보인다.

(j) Confidence Intervals for β_0, β_1 in 3 data sets

Variance of Noise	Coefficient	Estimate	2.5%	97.5%
0.25	$\widehat{\beta_0}$	-1.0188	-1.1151	-0.9226
	$\widehat{\beta_1}$	0.4995	0.3926	0.6064
0.01	$\widehat{\beta_0}$	-0.9973	-1.0180	-0.9765
	$\widehat{\beta_1}$	0.5021	0.4790	0.5252
1	$\widehat{\beta_0}$	-0.9423	-1.1413	-0.7433
	$\widehat{\beta_1}$	0.4443	0.2233	0.6654

- 분산이 작을수록 신뢰구간의 길이가 짧으며, 분산이 클수록 신뢰구간의 길이가 길다.
- 모든 경우 95% 신뢰구간이 실제 값인 -1 과 0.5 를 포함하고 있다.

Exercise 3.14

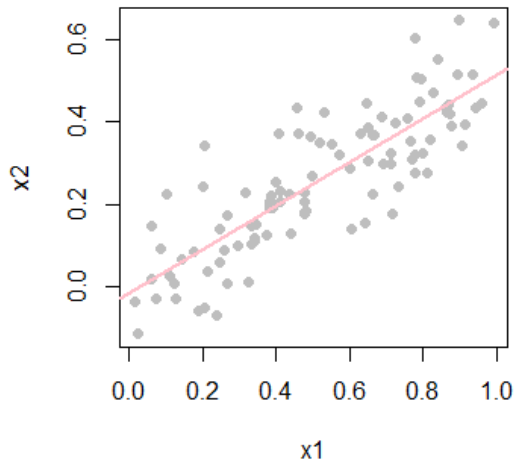
(a) Generate x_1, x_2 and y

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon, \quad \epsilon \sim N(0,1) \quad \rightarrow \quad Model: \quad E(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$$

회귀 계수의 True value 는 $\boldsymbol{\beta_0 = 2}$, $\boldsymbol{\beta_1 = 2}$, $\boldsymbol{\beta_2 = 0.3}$이다.

(b) Correlation between x_1 and x_2

$$\boldsymbol{Corr(x_1, x_2) = 0.8351}$$



(c) Least Squares Regression using x_1 and x_2

Call: lm(formula = y ~ x1 + x2)

Residuals:

Min	1Q	Median	3Q	Max
-2.8311	-0.7273	-0.0537	0.6338	2.3359

Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	**2.1305**	0.2319	9.188	7.61e-15 ***
x1	**1.4396**	0.7212	1.996	0.0487 *
x2	**1.0097**	1.1337	0.891	0.3754

Residual standard error: 1.056 on 97 degrees of freedom

Multiple R-squared: 0.2088, Adjusted R-squared: 0.1925

F-statistic: 12.8 on 2 and 97 DF, p-value: 1.164e-05

추정된 회귀계수는 $\widehat{\beta_0} = 2.1305$, $\widehat{\beta_1} = 1.4396$, $\widehat{\beta_2} = 1.0097$이며 실제 값 2, 2, 0.3 과 비교했을 때 β_0의 경우 비슷하지만 β_1, β_2와는 어느 정도 차이가 있다. 회귀계수 검정의 p-value 를 확인해볼 때 유의수준 0.05 에서 $H_0: \beta_1 = 0$을 기각할 수 있으나 $H_0: \beta_2 = 0$은 기각할 수 없다. 따라서 x_2는 설명력이 있다고 할 수 없다.

(d) Least Squares Regression using only x_1

Call: lm(formula = y ~ x1)

Residuals:

Min	1Q	Median	3Q	Max
-2.8950	-0.6687	-0. 0779	0.5922	2.4556

Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	**2.1124**	0.2307	9.155	8.27e-15 ***
x1	**1.9759**	0.3963	4.986	2.66e-06 ***

Residual standard error: 1.055 on 98 degrees of freedom

Multiple R-squared: 0.2024, Adjusted R-squared: 0.1942

F-statistic: 24.86 on 1 and 98 DF, p-value: 2.661e-06

x1 만을 이용한 단순선형회귀모형 적합 결과 귀무가설 $H_0: \beta_1 = 0$을 기각하며 x1 이 y 를 설명한다고 할 수 있다.

(e) Least Squares Regression using only x_2

Call: lm(formula = y ~ x2)

Residuals:

Min	1Q	Median	3Q	Max
-2.6269	-0.7516	-0. 0360	0.7238	2.4489

Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	**2.3899**	0.1949	12.26	< 2e-16 ***
x2	**2.8996**	0.6330	4.58	1.37e-05 ***

Residual standard error: 1.072 on 98 degrees of freedom

Multiple R-squared: 0.1763, Adjusted R-squared: 0.1679

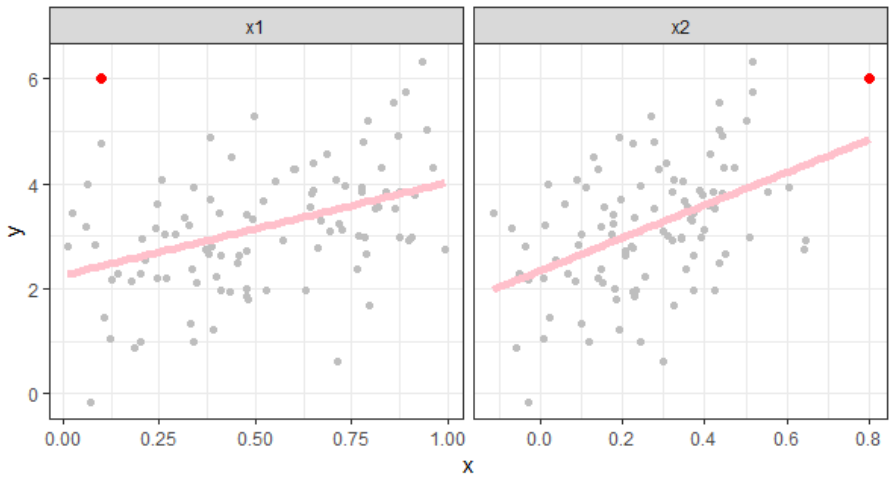
F-statistic: 20.98 on 1 and 98 DF, p-value: 1.366e-05

x2 만을 이용한 단순선형회귀모형 적합 결과 귀무가설 $H_0: \beta_2 = 0$을 기각하며, x2 역시 y 와 선형관계가 존재함을 확인할 수 있다.

(f) Results in (c)-(e)

(d)와 (e)에서 x1 과 x2 개별적으로 모형에 포함될 경우 모두 회귀계수 검정 p-value 가 0 에 가까워 각각 y 를 잘 설명하고 있다고 할 수 있지만 (c)에서 x1 과 x2 가 동시에 모형에 포함될 경우 x1 이 있는 모형에서 x2 는 더 이상 추가 설명력을 가지지 않으며, 추정된 회귀계수 또한 정확도가 떨어진다.

(g) Re-Fit the Models with Additional Observations



$(y, x_1, x_2) = (6, 0.1, 0.8)$을 추가 관측치로 포함하고 단순 및 다중 선형회귀모형을 적합하였다. (y, x_1), (y, x_2)의 개별 산점도에서 추가한 관측치가 다른 대부분의 점들과 다른 곳에 위치하여 Outlier 및 High Leverage Point 가 될 가능성이 보인다. 그러나 x2 의 경우 추정 회귀선과의 차이가 커보이지 않는다.

▷ **Model using x1 and x2**

Call: lm(formula = y ~ x1 + x2, data = data4g)

Residuals:

Min	1Q	Median	3Q	Max
-2.7335	-0.6932	-0.0526	0.6639	2.3062

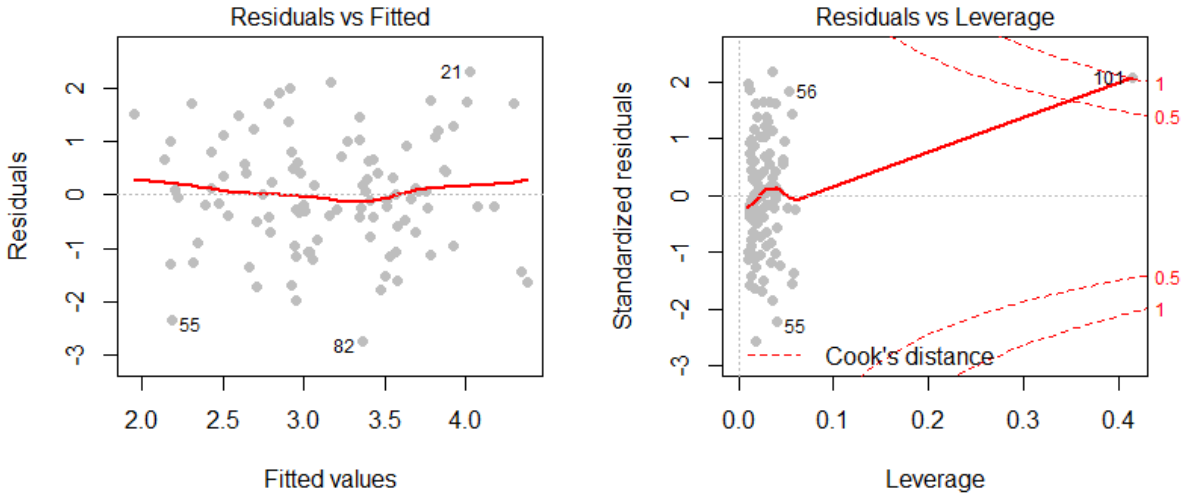
Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	**2.2267**	0.2314	9.624	< 7.91e-16 ***
x1	**0.5394**	0.5922	0.911	0.3646
x2	**2.5146**	0.8977x	2.801	0.0061 ***

Residual standard error: 1.075 on 98 degrees of freedom

Multiple R-squared: 0.2188, Adjusted R-squared: 0.2019

F-statistic: 13.72 on 2 and 98 DF, p-value: 5.564e-06



x1 과 x2 를 모두 포함한 모형에서 x2 는 설명력이 있으나 x1 는 x2 가 존재할 때 귀무가설 $H_0: \beta_1 = 0$을 기각할 수 없으며 y 에 대한 추가 설명력이 없다. 새로 추가한 관측치가 High Leverage Point 로 나타났다.

▷ Model using only x1

Call: lm(formula = y ~ x1, data = data4g)

Residuals:

Min	1Q	Median	3Q	Max
-2.8897	-0.6556	-0.0909	0.5682	325665

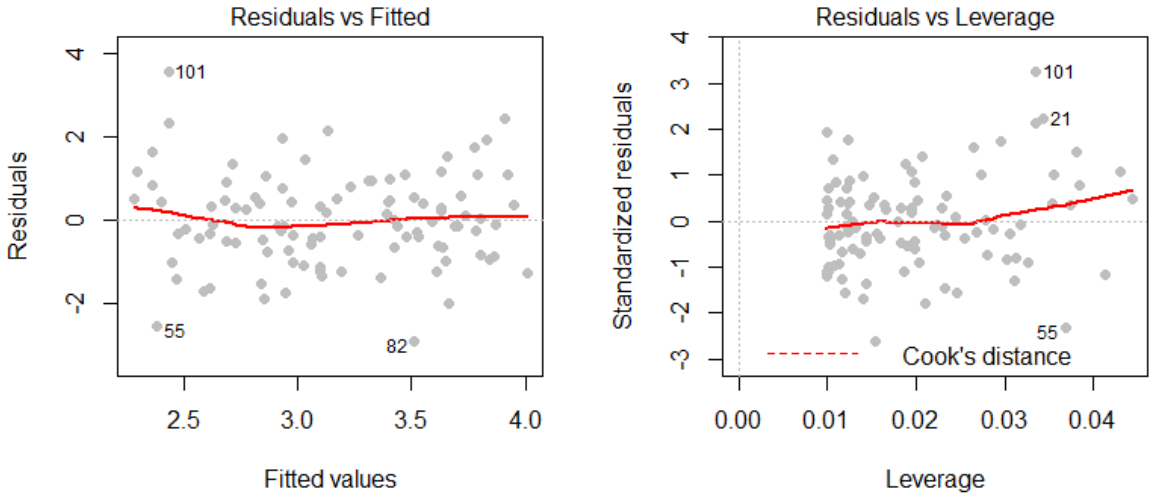
Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	**2.2569**	0.2390	9.445	1.78e-15 ***
x1	**1.7657**	0.4124	4.282	4.29e-05 ***

Residual standard error: 1.111 on 99 degrees of freedom

Multiple R-squared: 0.1562, Adjusted R-squared: 0.1477

F-statistic: 18.33 on 1 and 99 DF, p-value: 4.295e-05



x1 단순선형회귀모형에서 새로 추가된 관측치는 high leverage point 이면서 outlier 로 판정되었다. 관측치 추가 이전의 x1 모형보다 결정계수가 감소하였고 p-value 와 RMSE 는 증가하여 설명력이 감소하였으나 여전히 유의한 결과이다.

▷ Model using only x2

Call: lm(formula = y ~ x2, data = data4g)

Residuals:

Min	1Q	Median	3Q	Max
-2.6473	-0.7102	-0.0690	0.7270	2.3807

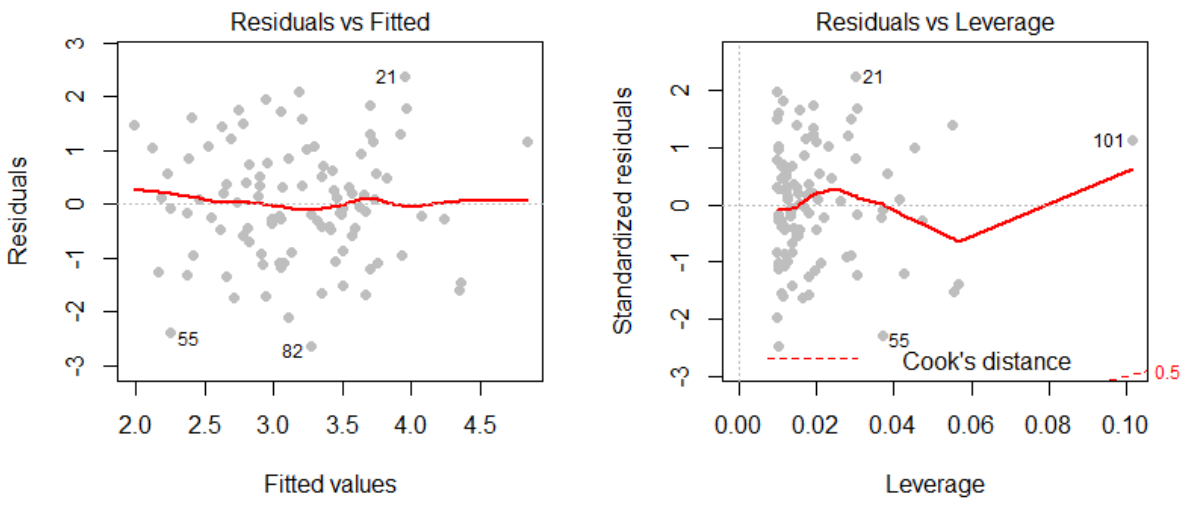
Coefficients:

	Estimate	Std. Error	t-value	P(>|t|) (p-value)
(Intercept)	**2.3451**	0.1912	12.264	< 2e-16 ***
x2	**3.1190**	0.6040	5.164	1.25e-06 ***

Residual standard error: 1.074 on 99 degrees of freedom

Multiple R-squared: 0.2122, Adjusted R-squared: 0.2042

F-statistic: 26.66 on 1 and 99 DF, p-value: 1.253e-06



x2 의 단순선형회귀모형에서 역시 p-value 가 작아 H_0: $\beta_2 = 0$을 기각한다. 추가한 관측치는 High Leverage Point 로 나타났으나, 모형의 설명력에는 안좋은 영향을 미치지 않았다.

Discussion

과제의 예제를 통해 선형 회귀모형을 실제 또는 랜덤 생성한 데이터에 적합하여 결과를 해석하였다. 실제 Auto 데이터 셋을 이용한 모형에서는 모형 진단에서 비선형성이 발견되었고 등분산성의 가정 역시 만족하지 않는 것을 알 수 있었다. 또한 설명변수들 중 일부의 상관관계가 강하여 모든 변수를 포함한 모형의 경우 일부 변수들의 회귀계수 검정 결과 추가 설명력이 없는 것으로 나타났다. 또한 교호작용 추가와 설명변수 변환의 효과를 확인해보았다.

랜덤하게 생성한 Data Point들에 선형모형을 적용한 결과를 해석하며 Noise의 분산이 작을수록 신뢰구간의 크기가 줄어들고 추정의 정확도가 향상되는 것을 확인하였다. 또다른 랜덤 생성 데이터에서는 서로 강한 상관관계가 있는 두 설명변수를 모두 포함한 모형과 개별 단일변수만을 이용한 모형의 결과를 비교하여 다중공선성의 문제를 확인하였다.

따라서 선형 회귀모형을 적용할 때 모형의 설명력 및 예측의 정확도를 높이기 위해서는 가능한 서로 독립적인 변수들로 모형을 구성하고 Y를 잘 설명하는 설명변수의 형태를 찾는 등 다양한 고려가 필요함을 알게 되고 정리할 수 있었다.

[Appendix] R code

3.6 Lab: Linear Regression (생략)

Exercise 3.8

```
# plot options
  plotlm <- function(model) {
    par(mfrow = c(2, 2))
    plot(model, pch = 19, col = 'gray', lwd = 2)
  }
  myplotlm <- function(model) {
    par(mfrow = c(1,2))
    plot(model, pch = 19, col = 'gray', lwd = 2, which = c(1,5))
  }

attach(Auto)

# (a)
  summary(lm1a <- lm(mpg ~ horsepower))
  cor(mpg, horsepower)
  coef(lm1a)[2]

  predict(lm1a, newdata = data.frame(horsepower = 98), interval = 'confidence')
  predict(lm1a, newdata = data.frame(horsepower = 98), interval = 'prediction')

# (b)
  plot(horsepower, mpg, pch = 19, col = 'gray')
  abline(lm1a, col = 'steelblue', lwd = 2)

# (c)
# plot all
  plotlm(lm1a)

# residuals vs fitted
  plot(predict(lm1a), residuals(lm1a), col = 'gray', lwd = 2)
  plot(predict(lm1a), rstudent(lm1a), col = 'gray', lwd = 2)
# leverage
  plot(hatvalues(lm1a), col = 'gray', lwd = 2)
  points(which.max(hatvalues(lm1a)), max(hatvalues(lm1a)), col = 'red', pch = 16)
```

Exercise 3.9

```
# (a)
  pairs(Auto[,-9], col = 'lightblue')
# (b)
  cor(Auto[,-9])
# (c)
  summary(lm2c <- lm(mpg ~ . -name, Auto))
# (d)
  plotlm(lm2c)
  plot(predict(lm2c), residuals(lm2c), col = 'gray', lwd = 2)
  plot(predict(lm2c), rstudent(lm2c), col = 'gray', lwd = 2)
  plot(hatvalues(lm2c), col = 'gray', lwd = 2)
  points(which.max(hatvalues(lm2c)), max(hatvalues(lm2c)), col = 'red', pch = 16)
# (e)
  summary(lm2e <- lm(mpg ~ .*., Auto[,-9]))
  coef2e <- as.data.frame(summary(lm2e)$coefficients) %>% round(4)
# (f)
  summary(lm2f <- lm(mpg ~ cylinders + poly(displacement, 2) +
                       log(horsepower) + log(weight) + poly(acceleration, 2) + year + origin))
```

Exercise 3.13

```
n <- 100
# (a)
  set.seed(1)
  x <- rnorm(n)
# (b)
  eps <- rnorm(n, 0, 0.5)
# (c)
  beta0 <- -1 ; beta1 <- 0.5
  y <- beta0 + beta1*x + eps
  length(y)

# (d)
  par(mfrow = c(1, 1))
  plot(x, y, col = 'gray', pch = 19, main = 'Scatterplot (y, x)')

# (e)
  summary(lm3e <- lm(y ~ x))
  plotlm(lm3e)

# (f)
  par(mfrow = c(1, 1))
  plot(x, y, col = 'gray', pch = 19)
  abline(-1, 0.5, col = 'steelblue', lwd = 2)
  abline(lm3e, col = 'red', lwd = 2, lty = 2)
  legend('topleft', legend = c('population', 'fitted'), col = c('steelblue', 'red'),
         lty = c(1, 2), lwd = 2, bty = 'n')

# (g)
  summary(lm3g <- lm(y ~ poly(x, 2)))
```

```
  plotlm(lm3g)

  anova(lm3e, lm3g, test = 'F')

  par(mfrow = c(1, 1))
  plot(x, y, col = 'gray', pch = 19)
  abline(-1, 0.5, col = 'steelblue', lwd = 2)
  abline(x = sort(x), y = sort(fitted(lm3g)), col = 'red', lwd = 2, lty = 2)
  legend('topleft', legend = c('population', 'fitted'), col = c('steelblue', 'red'),
         lty = c(1, 2), bty = 'n', lwd = 2)


# (h)
  eps_less <- rnorm(n, 0, 0.1)
  y_less <- beta0 + beta1*x + eps_less

  summary(lm3h <- lm(y_less ~ x))

  plotlm(lm3h)

  par(mfrow = c(1, 1))
  plot(x, y_less, col = 'gray', pch = 19)
  abline(-1, 0.5, col = 'steelblue', lwd = 2)
  abline(lm3h, col = 'red', lwd = 2, lty = 2)
  legend('topleft', legend = c('population', 'fitted'), col = c('steelblue', 'red'),
         lty = c(1, 2), lwd = 2, bty = 'n')

# (i)
  eps_more <- rnorm(n, 0, 1)
  y_more <- beta0 + beta1*x + eps_more

  summary(lm3i <- lm(y_more ~ x))

  plotlm(lm3i)

  par(mfrow = c(1, 1))
  plot(x, y_more, col = 'gray', pch = 19)
  abline(-1, 0.5, col = 'steelblue', lwd = 2)
  abline(lm3i, col = 'red', lwd = 2, lty = 2)
  legend('topleft', legend = c('population', 'fitted'), col = c('steelblue', 'red'),
         lty = c(1, 2), lwd = 2, bty = 'n')

# (j)
  rbind(
    cbind(coef(lm3e), confint(lm3e, interval = 'confidence')), # original
    cbind(coef(lm3h), confint(lm3h, interval = 'confidence')), # less
    cbind(coef(lm3i), confint(lm3i, interval = 'confidence'))  # more
  ) %>% round(4)
```

Exercise 3.14

```
# (a)
  set.seed(1)
  x1 <- runif(100)
  x2 <- 0.5*x1 + rnorm(100)/10

  beta <- c(2, 2, 0.3)
  y <- beta[1] + beta[2]*x1 + beta[3]*x2 + rnorm(100)

# (b)
  cor(x1, x2)
  plot(x1, x2, col = 'gray', pch = 19)
  abline(lm(x2 ~ x1), col = 'pink', lwd = 2)

# (c)
  summary(lm4c <- lm(y ~ x1 + x2))
  rbind(beta, coef(lm4c))
  confint(lm4c, interval = 'confidence') %>% cbind(estimate = coef(lm4c))

# (d)
  summary(lm4d <- lm(y ~ x1))

# (e)
  summary(lm4e <- lm(y ~ x2))

# (f) comment
  anova(lm4c)

# (g)
  data4g <- rbind(data.frame(y, x1, x2), c(6, 0.1, 0.8))

  summary(lm4g12 <- lm(y ~ x1 + x2, data4g)) ; myplotlm(lm4g12)
  summary(lm4g1 <- lm(y ~ x1, data4g)) ; myplotlm(lm4g1)
  summary(lm4g2 <- lm(y ~ x2, data4g)) ; myplotlm(lm4g2)

  gdat <- gather(data4g, var, x, -y)
  ggplot(gdat, aes(x, y)) + theme_bw() +
    geom_point(col = 'gray') + geom_smooth(method = 'lm', se = F, col = 'pink', size = 2) +
    geom_point(data = gdat[c(101,202),], aes(x, y), col = 'red', size = 2) +
    facet_grid(~var, scales = 'free')
```