1. Description

▶ Chpater3 는 *linear regression* 에 대한 챕터로 선형 회귀는 quantitative response 를 예측하는 유용한 방법이다.

▶ OLS 회귀의 가정으로는 정규성, 독립성, 선형성, 등분산성이 있다. 정규성이란 반응변수가 정규분포를 따른다면 잔차 또한 정규분포하며 평균은 0 임을 말한다. 반응변수는 서로 독립적이어야 한다는 것은 독립성을 말한다. 선형성이란 종속변수와 독립변수가 선형관계에 있다면 잔차와 예측치 사이에 어떤 체계적인 관계가 있으면 안 된다는 것이다. 마지막으로 등분산성은 분산이 일정하다는 가정이다.

▶ 최소 제곱법을 사용하는 선형모형을 만드는 lm() 함수를 통해 OLS 회귀모형을 만들고 summary() 함수로 모형의 회귀계수 및 통계 수치를 볼 수 있지만 이 모형이 적절한 것인지 검증이 필요하다.

▶ 회귀 모형을 만들고, 이상치(outlier), 큰 지레점(high leverage point), 영향관측치(influential observation), 다중공선성 등을 확인해야한다.

- ✓ outlier: 회귀모형으로 잘 예측되지 않는 관측치 (즉, 아주 큰 양수/음수의 residual)
- ✓ high leverage point : 비정상적인 예측변수의 값에 의한 관측치. 예측변수의 이상치로 볼 수 있다.
- ✓ influential observation : 통계 모형 계수 결정에 불균형한 영향을 미치는 관측치로 Cook's distance 라는 통계치로 확인할 수 있다.
- ✓ 독립변수들 간에 강한 상관관계가 있는 것을 다중공선성이라고 한다. 다중공선성이 있는지 진단하는 지표로 분산팽창요인(Variance Inflation Factor) 줄여서 VIF 가 있다. VIF 값이 10 이상이면 독립변수들 간에 다중공선성이 존재한다고 판단한다.

2. Implement

Lab : Linear Regression

- lm() (=model) 함수를 사용해서 단순 선형 회귀 모델을 fitting 시킨 후 summary(model)로 잔차 분포와 회귀계수, F 통계량, R squared 를 알 수 있다.

- lm 함수 내에서 모델을 만들 때, *는 상호작용을 표현하고, I()를 통해 비선형 변형이 가능하다.

- hatvalues(model) 함수로 hat matrix 의 대각행렬 값(hii)를 구해서 레버리지 판단에 사용한다.

- vif(model)은 다중공선성을 진단하는 지표인 VIF 를 계산하는 함수이다.

- contrasts(variable)은 factor 인 변수를 더비 변수로 어떻게 만들어지는 알려준다.

3.8. This question involves the use of simple linear regression on the Auto data set.
(a) Use the lm() function to perform a simple linear regression with mpg as the response and horsepower as the predictor. Use the summary() function to print the results. Comment on the output.
For example:

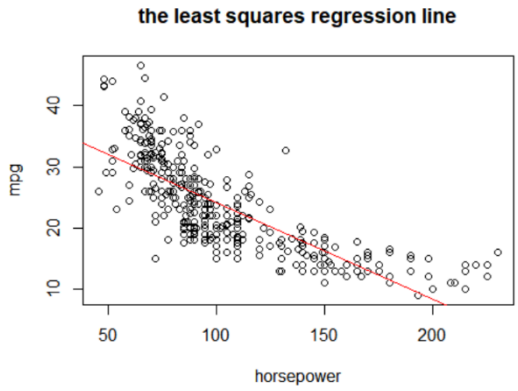
i. Is there a relationship between the predictor and the response?
ii. How strong is the relationship between the predictor and the response?
iii. Is the relationship between the predictor and the response positive or negative?
iv. What is the predicted mpg associated with a horsepower of 98? What are the associated 95% confidence and prediction intervals?

lm(formula = mpg ~ horsepower, data = Auto)				
Coefficients :				
	Estimate	Std.Error	t-value	Pr(>| t |)
(intercept)	39.9359	0.7175	55.66	<2e-16
horsepower	-0.1578	0.0064	-24.49	<2e-16
Residual standard error: 4.906 on 390 degrees of freedom				
Multiple R2: 0.6059, Adjusted R2: 0.6049				
F-statistic: 599.7 on 1 and 390 DF, p-value: < 2.2e-16				

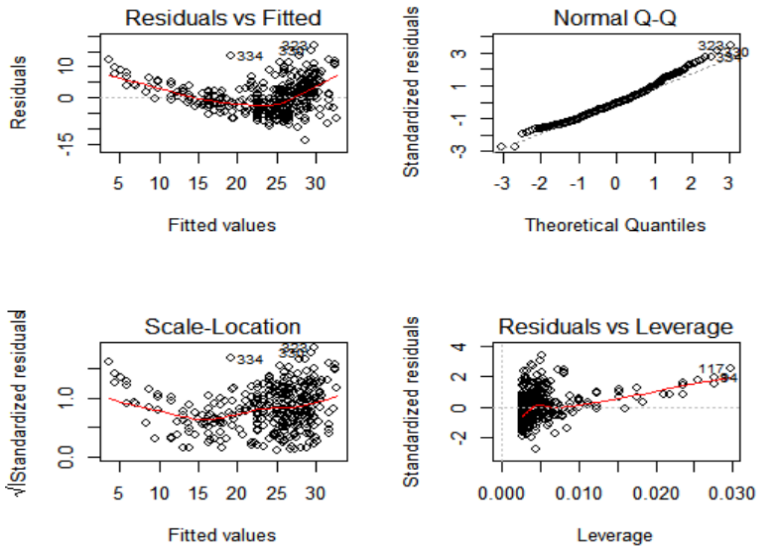
Mpg를 반응변수, horsepower을 예측변수 일 때, horsepower의 회귀계수가 -0.1578로 음수이기 때문에 두변수는 음의 상관관계를 가지고 있다고 할 수 있다. 또한 p-value값이 작아 0에 근사하므로 두 변수사이에 강한 관계가 있다고 생각할 수 있다. Fitting된 모델에서 horsepower이 98일 때, mpg의 예측치와 신뢰구간은 다음과 같다.

	fit	lwr	upr
Confidence intervals	24.467	23.973	24.961
Prediction intervals		14.809	34.125

(b) Plot the response and the predictor. Use the abline() function to display the least squares regression line.



(c) Use the plot() function to produce diagnostic plots of the least squares regression fit. Comment on any problems you see with the fit.

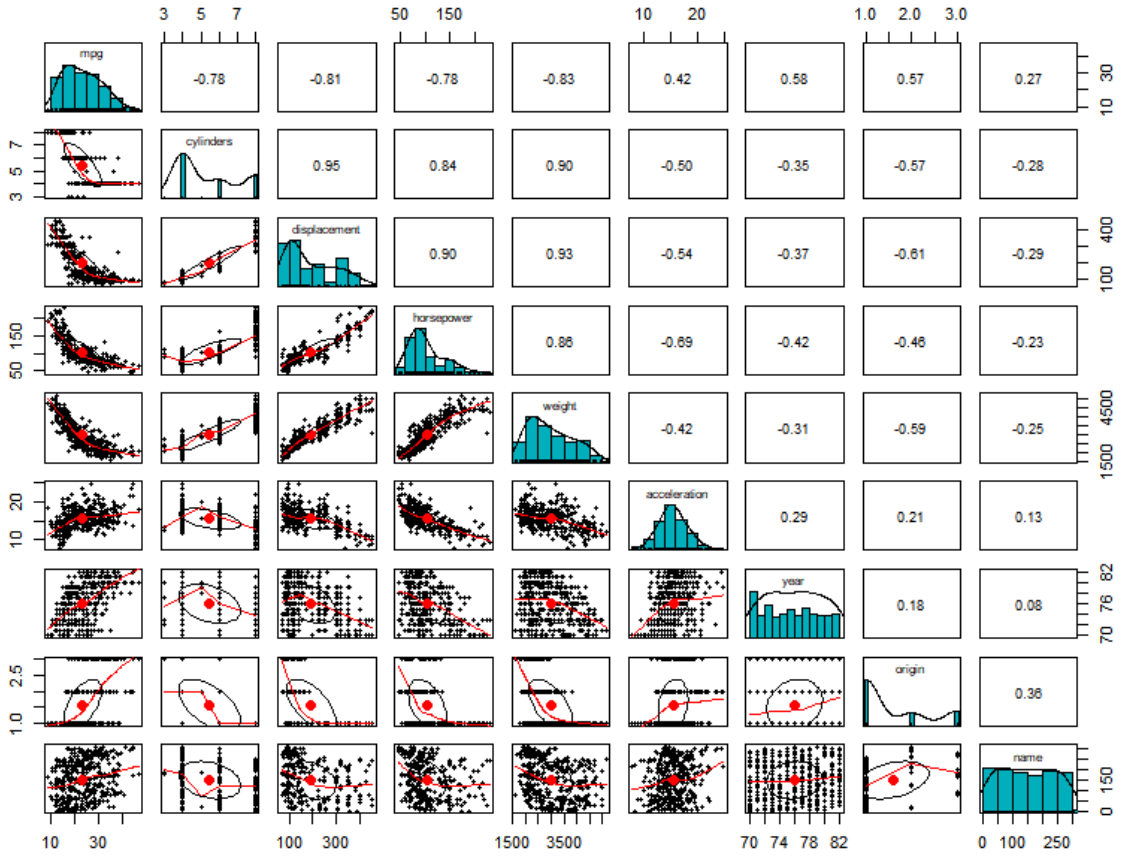


첫번째 plot은 잔차 plot으로, 잔차의 등분산성과 독립성을 검정하기 위한 플롯이다. 두번째 plot은 정규 plot으로, 잔차의 정규성을 검정하기 위함이고, 세번째 plot은 표준화 잔차 plot으로, 잔차 plot과 비슷하다. 네번째 plot은 지레-잔차 plot으로, X값과 Y값의 특이값을 찾아내는데 유용한 플롯이다.

첫번째 plot 인 residuals vs fitted plot 의 관계가 선형이 아니고 Q-Q plot 을 보았을 때도 중간부분은 맞지만 오른쪽 끝에서 직선으로 맞지 않는 부분이 보였다.

3.9. This question involves the use of multiple linear regression on the Auto data set.

(a) Produce a scatterplot matrix which includes all of the variables in the data set.



mpg는 대부분의 변수와의 상관관계가 음수이고. displacement와는 -0.81, weight와는 -0.83으로 큰 음의 상관관계를 갖는다. 반대로 cylinders는 displacement와는 0.95, weight와는 0.90으로 큰 양의 상관관계를 갖는다. year은 모든 변수와의 상관계수의 절대값이 0.5이하가 대부분이다.

(b) Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, cor() which is qualitative.

cylinders, displacement, horsepower, weight는 각각 mpg, cylinders, displacement, horsepower, weight와 큰 상관관계를 보인다. 특히 weight와 displacement의 상관계수는 0.93으로 제일 높았다. 또한 mpg와 weight는 -0.83으로 제일 낮은 상관계수를 보였다.

	mpg	cylinders	displacement	horsepower	weight	acceleration	year	origin
mpg	1	-0.78	-0.81	-0.78	-0.83	0.42	0.58	0.57
cylinders	-0.78	1	0.95	0.84	0.90	-0.50	-0.35	-0.57
displacement	-0.81	0.95	1	0.90	0.93	-0.54	-0.37	-0.61
horsepower	-0.78	0.84	0.90	1	0.86	-0.69	-0.42	-0.46
weight	-0.83	0.90	0.93	0.86	1	-0.42	-0.31	-0.59
acceleration	0.42	-0.50	-0.54	-0.69	-0.42	1	0.29	0.21
year	0.58	-0.35	-0.37	-0.42	-0.31	0.29	1	0.18
origin	0.57	-0.57	-0.61	-0.46	-0.59	0.21	0.18	1

(c) Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output. For instance:

i. Is there a relationship between the predictors and the response?

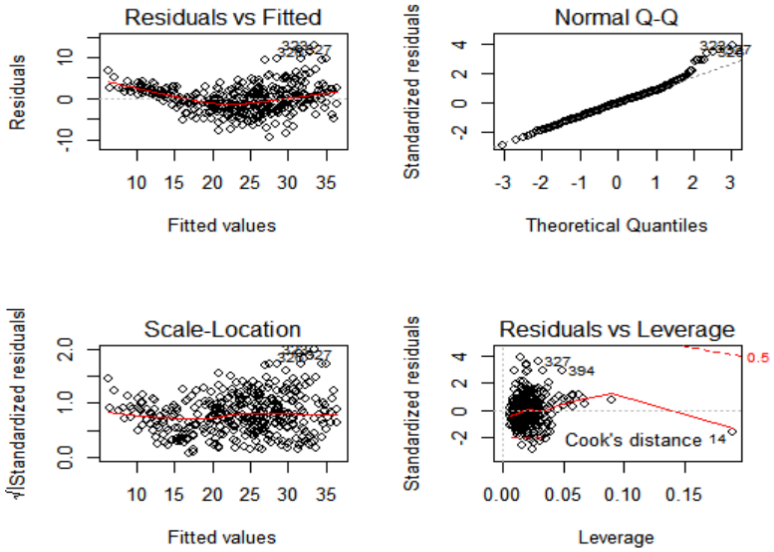
ii. Which predictors appear to have a statistically significant relationship to the response?

iii. What does the coefficient for the year variable suggest?

lm(formula = mpg ~ . - name, data = Auto)					
Coefficients :					
	Estimate	Std.Error	t-value	Pr(>| t |)	
(intercept)	-17.2184	4.6442	-3.707	0.0002	***
cylinders	-0.4933	0.3232	-1.526	0.1278	
displacement	0.0198	0.0075	2.647	0.0084	**
horsepower	-0.169	0.0137	-1.23	0.2196	
weight	-0.0064	0.0006	-9.929	<2e-16	***
acceleration	0.0805	0.0988	0.815	0.4154	
year	0.7507	0.0509	14.729	<2e-16	***
origin	1.4261	0.2781	5.127	4.67E-07	***
Residual standard error: 3.328 on 384 degrees of freedom					
Multiple R2: 0.8215, Adjusted R2: 0.8182					
F-statistic: 252.4 on 7 and 384 DF, p-value: < 2.2e-16					

displacement, weigth, year, orgin의 p-value값이 거의 0에 근사할 정도로 작으므로 통계적으로 유의미한 변수라고 할 수 있다. 회귀계수를 살펴보면 weight의 회귀계수는 음수이므로 음의 관계를 가지고 그 외의 세 변수는 양수의 값을 가지므로 양의 관계를 가진다. 특히 year의 회귀계수가 0.7507로 다른 변수에 비해 월등히 크므로 나중에 나온 자동차일수록 더 높은 mpg 값을 가짐을 의미한다.

(d) Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?



residuals vs fitted plot가 선형이 아니므로 outlier가 있다고 생각할 수 있다. 아래 왼쪽의 plot을 보면 통계 모형 계수 결정에 불균형한 영향을 미치는 관측치인 Cook's distance을 통해 14번째 관측치가 강한 leverage를 갖고 있다고 할 수 있다.

(e) Use the * and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?

회귀적합에서 ':' 기호를 이용하면 상호작용항을 나타낼 수 있고 '*'를 사용하면 상호작용항을 고려하되 각 개별항과 상호작용항이 될 수 있는 모든 조합을 알아서 배정시키는 기호이다. 예를 들어 displacement * weight로 지정해 주었을 때는 displacement, weight의 각각의 항과 두 상호작용항이 생성되었고, displacement : weight로 지정해 주었을 때는 상호작용항만 존재함을 볼 수 있다. 두 모델 모두 상호작용항이 통계적으로 유의미하다.

lm(formula = mpg ~ displacement * weight, data = Auto)

Coefficients :

	Estimate	Std.Error	t-value	p-value	
(intercept)	5.37E+01	1.94	27.697	<2e-16	***
displacement	-0.0783	1.13E-02	-6.922	0.1278	***
weight	-0.0089	8.47E-04	-10.539	<2e-16	***
displacement:weight	1.74E-05	2.79E-06	6.253	0.2196	***

lm(formula = mpg ~ displacement : weight, data = Auto)

Coefficients :

	Estimate	Std.Error	t-value	p-value	
(intercept)	3.12E+01	0.387	80.59	<2e-16	***
displacement:weight	-1.18E-05	4.6E-07	-25.69	<2e-16	***

(f) Try a few different transformations of the variables, such as log(*X*), $\sqrt{X}$, X^2. Comment on your findings.

lm(formula = mpg ~ displacement + I(log(weight)), data = Auto)

Coefficients :

	Estimate	Std.Error	t-value	p-value	
(intercept)	174.627	14.216	12.284	<2e-16	***
displacement	-0.013	0.005	-2.735	0.0065	**
I(log(weight))	-18.653	1.899	-9.818	<2e-16	***

lm(formula = mpg ~ displacement + I(weight^(1/2)), data = Auto)

Coefficients :

	Estimate	Std.Error	t-value	p-value	
(intercept)	63.009	3.086	20.413	<2e-16	***
displacement	-0.013	0.005	-2.478	0.0136	*
I(log(weight))	-0.683	0.074	-9.132	<2e-16	***

lm(formula = mpg ~ displacement + I(weight^(2)), data = Auto)

Coefficients :

	Estimate	Std.Error	t-value	p-value	
(intercept)	35.21	0.477	73.773	<2e-16	***
displacement	-0.035	0.058	-5.206	3.14e-07	***
I(log(weight))	-6.112e-07	1.114e-07	-5.485	7.46e-08	***

weight 를 로그를 취하거나 루트를 취하여 모델에 fitting 시켜보았다. 세 경우 모두 변환시킨 항이 유의미하게 나왔다. R squared 값을 봤을 때, 로그를 취한 모델이 71.8%로 가장 높았기 때문에 이 모델에서는 weight 에 로그를 취하는 것이 적절하다고 생각한다.

3.13. In this exercise you will create some simulated data and will fit simple linear regression models to it. Make sure to use set.seed(1) prior to starting part (a) to ensure consistent results.

(a) Using the rnorm() function, create a vector, x, containing 100 observations drawn from a $N(0, 1)$ distribution. This represents a feature, X.

rnorm()은 정규분포로 random number을 생성하는 함수이다. parameters은 rnorm(n, mean, sd)으로 mean, sd 로 표준편차를 설정해주고, n 개의 숫자를 생성하겠다는 의미이다.

디폴트는 mean=0, sd=1 이므로 x <- rnorm(100) 로 x 를 생성해준다.

(b) Using the rnorm() function, create a vector, eps, containing 100 observations drawn from a $N(0, 0.25)$ distribution i.e. a normal distribution with mean zero and variance 0.25.

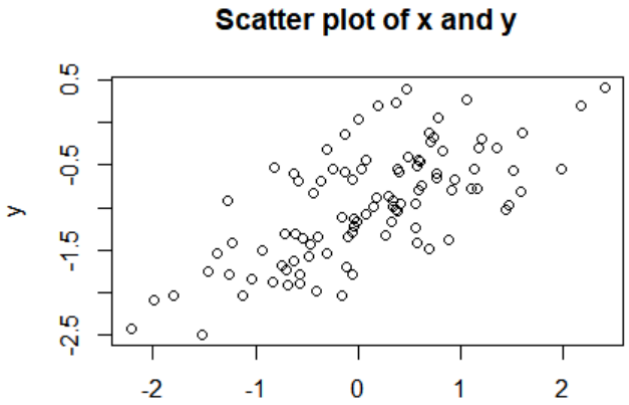
위와 같은 방법으로 eps를 생성하면, eps <- rnorm(100,sd=0.25^0.5)이다.

(c) Using x and eps, generate a vector y according to the model $Y = -1 + 0.5X + \epsilon$. (3.39)

What is the length of the vector y? What are the values of $\beta 0$ and $\beta 1$ in this linear model?

X와 eps를 100개 생성했기 때문에 y의 길이 또한 100이다. (3.39) 선형 모델에서 $\beta 0$은 y절편이므로 -1, $\beta 1$은 기울기를 나타내므로 0.5이다.

(d) Create a scatterplot displaying the relationship between x and y. Comment on what you observe.



x와 y의 산점도를 살펴보면 양의 관계임을 볼 수 있다.

(e) Fit a least squares linear model to predict y using x. Comment on the model obtained. How do $\widehat{\beta 0}$ and $\widehat{\beta 1}$ compare to $\beta 0$ and $\beta 1$?

추정된 $\beta 0$은 -1.019, $\beta 1$은 0.499로 실제 값과 거의 비슷하다.

lm(formula = y ~ x)

Coefficients :

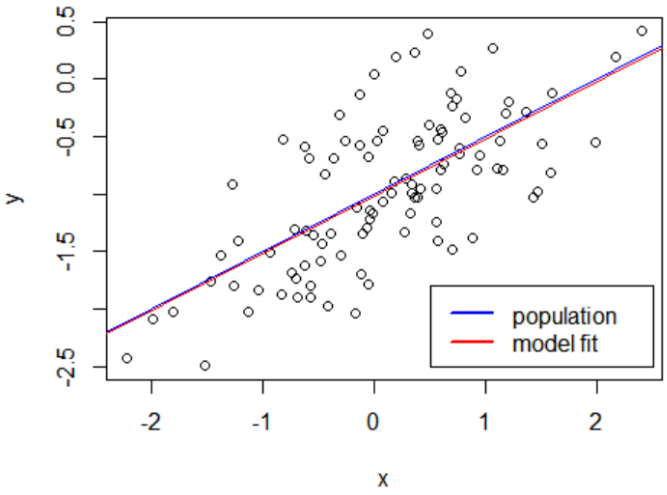
	Estimate	Std.Error	t-value	p-value	
(intercept)	-1.019	0.048	-21.010	<2e-16	***
x	0.499	0.054	9.273	2.58e-15	***

Residual standard error: 0.4814 on 98 degrees of freedom

Multiple R2: 0.4674, Adjusted R2: 0.4619

F-statistic: 85.99 on 1 and 98 DF, p-value: 4.583e-15

(f) Display the least squares line on the scatterplot obtained in (d). Draw the population regression line on the plot, in a different color. Use the legend() command to create an appropriate legend.



파란색이 실제 회귀선이고 빨간색이 fitting된 모델의 회귀선이다. 두 선의 차이가 거의 없으므로 모델이 잘 생성되었다고 할 수 있다.

(g) Now fit a polynomial regression model that predicts y using x and x^2. Is there evidence that the quadratic term improves the model fit? Explain your answer.

lm(formula = y ~ I(x^2))

Coefficients :

	Estimate	Std.Error	t-value	p-value	
(intercept)	-0.971	0.058	-16.517	<2e-16	***
x	0.508	0.053	9.420	2.4e-15	***
I(x^2)	-0.059	0.042	-1.403	0.164	

Residual standard error: 0.479 on 97 degrees of freedom

Multiple R2: 0.4779, Adjusted R2: 0.4672

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

x^2 항의 p-value의 값이 크므로 이 모델에서는 유의미하지 않는다고 생각할 수 있지만, R squared 값은 모델에 x만 존재했을 때 보다 높아지기 때문에 이차항이 모델 성능을 향상시킨다고 확신할 수 없다.

(h) Repeat (a)–(f) after modifying the data generation process in such a way that there is *less* noise in the data. The model (3.39) should remain the same. You can do this by decreasing the variance of the normal distribution used to generate the error term ϵ in (b). Describe your results.

lm(formula = y ~ x)

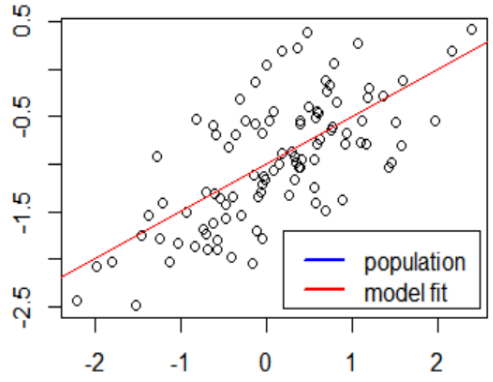
Coefficients :

	Estimate	StdError	t-value	p-value	
(intercept)	-0.9999	0.001	-909.5	<2e-16	***
x	0.5008	0.001	410.1	<2e-16	***

Residual standard error: 0.0109 on 98 degrees of freedom

Multiple R2: 0.9994, Adjusted R2: 0.9994

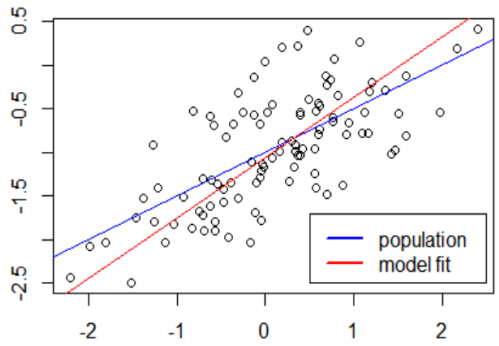
F-statistic: 1.682e+05 on 1 and 98 DF, p-value: < 2.2e-16



epsilon을 발생시킬 때, sd를 0.01로 설정하여 (e) 모델보다 noise를 작게 하였다. 추정된 $\beta 0$은 -0.999, $\beta 1$은 0.5008로 실제 값과 거의 비슷하고 R squared도 99.9%로 fitting시킨 모델과 실제 모델의 격차가 거의 없다. 그래프로 확인해 보았을 때도, 실제 회귀선인 파란선과 fitting 시킨 모델의 회귀선이 일치해 보였다.

(i) Repeat (a)–(f) after modifying the data generation process in such a way that there is *more* noise in the data. The model (3.39) should remain the same. You can do this by increasing the variance of the normal distribution used to generate the error term _ in (b). Describe your results.

lm(formula = y ~ x)					
Coefficients :					
	Estimate	Std.Error	t-value	p-value	
(intercept)	-1.060	0.117	-9.049	1.40e-14	***
x	0.691	0.130	5.315	6.71e-07	***
Residual standard error: 1.163 on 98 degrees of freedom					
Multiple R2: 0.2237, Adjusted R2: 0.2158					
F-statistic: 28.25 on 1 and 98 DF, p-value: 6.713e-07					



epsilon을 발생시킬 때, sd를 0.01로 설정하여 (e) 모델보다 noise를 크게 하였다. 추정된 $\beta 0$은 -1.060로 실제 값과 거의 비슷하지만 $\beta 1$은 0.691로 실제 값과 약간의 차이가 있다. R squared은 22.3%로 fitting시킨 모델과 실제 모델의 차이가 많음을 알 수 있다. 그래프로 확인해 보았을 때도, 실제 회귀선인 파란선과 fitting 시킨 모델의 회귀선의 차이가 확연하게 나타났다.

(j) What are the confidence intervals for $\beta 0$ and $\beta 1$ based on the original data set, the noisier data set, and the less noisy data set? Comment on your results.

	less noisy			original			noisier	
	2.50%	97.50%		2.50%	97.50%		2.50%	97.50%
$\beta 0$	-1.0020	-0.9979		-1.1151	-0.9226		-1.2183	-0.7960
$\beta 1$	0.4961	0.5007		0.3926	0.6064		0.3686	0.8377

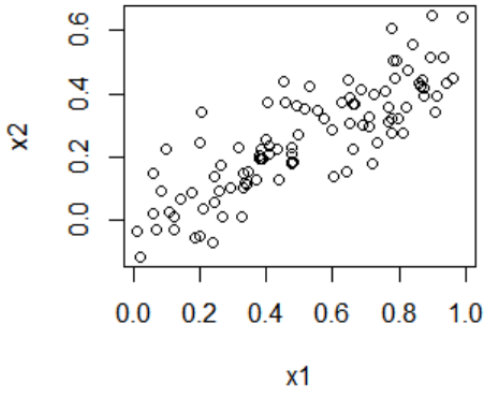
noise 를 작게 생성한 데이터일수록 $\beta 0$, $\beta 1$ 의 신뢰구간이 작아진다.

3.14. This problem focuses on the *collinearity* problem.

(a) The last line corresponds to creating a linear model in which y is a function of x1 and x2. Write out the form of the linear model. What are the regression coefficients?

$y = \beta 0 + \beta 1 \cdot x1 + \beta 2 \cdot x2 + \epsilon$ 의 형식으로 나타낼 수 있고, 이 경우 회귀계수는 $\beta 0 = 2$, $\beta 1 = 2$, $\beta 2 = 0.3$ 이다.

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.



x1과 x2의 상관계수는 0.8351로
강한 양의 상관관계를 가지고 있다.
산점도를 통해서도 두 변수가 얼마나 강한
양의 관계가 있는지 확인할 수 있다.

(c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are $\widehat{\beta 0}$, $\widehat{\beta 1}$, and $\widehat{\beta 2}$? How do these relate to the true $\beta 0$, $\beta 1$, and $\beta 2$? Can you reject the null hypothesis $H0$: $\beta 1 = 0$? How about the null hypothesis $H0$: $\beta 2 = 0$?

lm(formula = y ~ x1 + x2)					
Coefficients :					
	Estimate	Std.Error	t-value	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	

추정된 회귀계수는 $\widehat{\beta 0} = 2.1305$, $\widehat{\beta 1} = 1.4396$, $\widehat{\beta 2} = 1.0097$로 $\beta 0$은 실제 값과 비슷하지만 $\beta 1$과 $\beta 2$는 차이가 있고 standard error 또한 크다. 따라서 $\beta 1 = 0$인 귀무가설에 대해서는 기각하지만, $\beta 2 = 0$인 귀무가설은 기각할 수 없다.

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis $H0$: $\beta 1 = 0$?

p-value가 0에 근접하기 때문에 $H0$: $\beta 1 = 0$를 기각할 수 있다.

lm(formula = y ~ x1)					
Coefficients :					
	Estimate	Std.Error	t-value	p-value	
(intercept)	2.1124	0.2307	9.155	8.27e-15	***
x1	1.9759	0.3963	4.986	2.66e-06	***

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis $H0$: $\beta 1 = 0$?

p-value가 0에 근접하기 때문에 $H0$: $\beta 1 = 0$를 기각할 수 있다.

lm(formula = y ~ x2)					
Coefficients :					
	Estimate	Std.Error	t-value	p-value	
(intercept)	2.3899	0.1949	12.26	< 2e-16	***
x2	2.8996	0.6330	4.58	1.37e-05	***

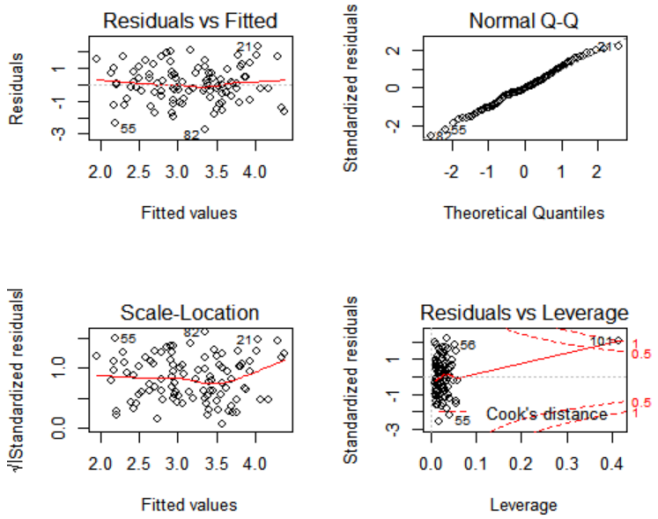
(f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

x1, x2를 가지고 있는 모델에서는 x1가 미미하게 유의하고 x2가 유의미하지 않은 변수라고 보여줬다. 하지만 x1, x2만 가지고 있는 각각의 모델에서는 매우 유의미하다고 할 수 있으므로 (c)-(e)에서의 결과가 서로를 부정할 수는 없을 것이다.

(g) Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.

1)

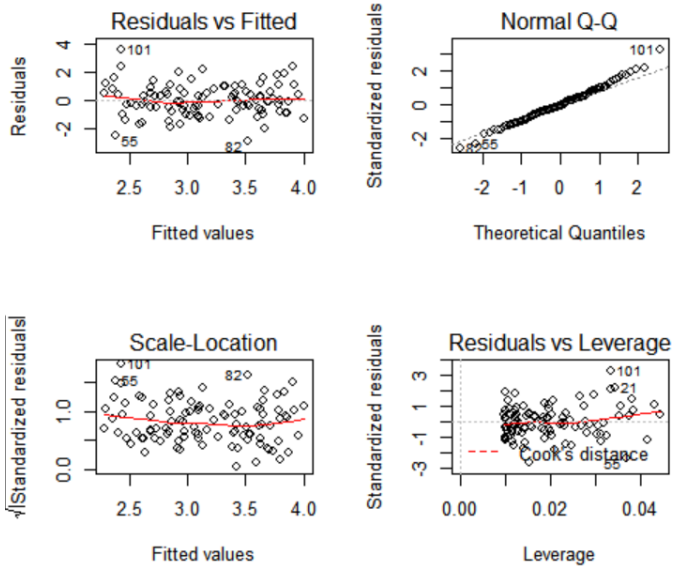
lm(formula = y ~ x1 + x2)					
Coefficients :					
	Estimate	Std.Error	t-value	p-value	
(intercept)	2.2267	0.2314	9.624	7.91e-16	***
x1	0.5394	0.5922	0.911	0.36458	
x2	2.5146	0.8977	2.801	0.00614	**



x2 의 새로운 점이 outlier, x1 과 x2 모두 high leverage point 로 보인다. residuals vs. leverage plot 를 보면 101 번째 관측치가 눈에 띈다. 빨간 실선이 점들과 가깝게 위치해야 하는데 새로운 데이터로 인해 문제가 생긴 것을 알 수 있다.

2)

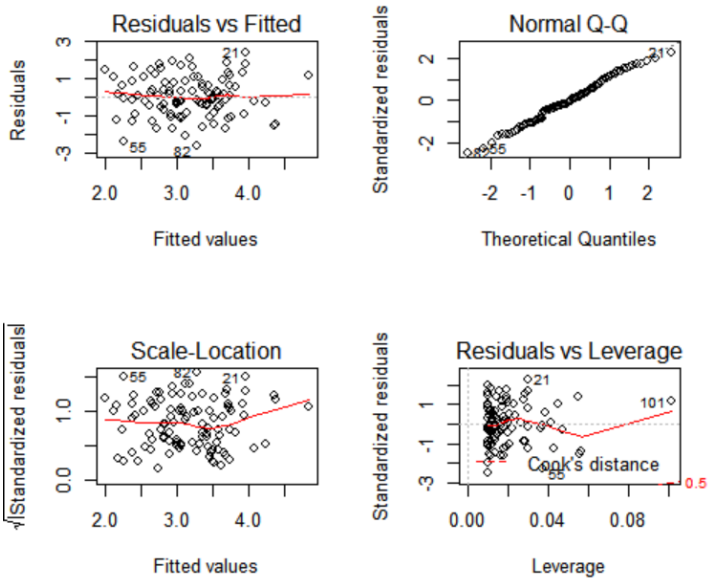
lm(formula = y ~ x1)					
Coefficients :					
	Estimate	Std.Error	t-value	p-value	
(intercept)	2.2569	0.2390	9.445	1.78e-15	***
x1	1.7657	0.4124	4.282	4.29e-05	***



x1 만 존재하는 모델에서는 새로운 점이 high leverage point 이지만 outlier 가 아니므로 큰 문제를 일으키지 않는다.

3)

lm(formula = y ~ x2)					
Coefficients :					
	Estimate	Std.Error	t-value	p-value	
(intercept)	2.3451	0.1912	12.264	< 2e-16	***
x2	3.1190	0.6040	5.164	1.23e-06	***



x2 만 존재하는 모델에서는 새로운 점인 101 번째 관측치가 high leverage point 이지만 Normal Q-Q plot 을 보았을 때 모든 점들이 직선에 가까워 보이므로 큰 문제가 되지 않아 보인다.

3. Discussion

▶ 먼저 간단한 예제로 R 코드를 공부하고 그 코드를 바탕으로 데이터에 적용시켜 보았다.

▶ Exercise 3.8 에서는 자동차에 대한 Auto data 를 이용하여 단순 선형 회귀를 이용하였다. 예측 변수와 반응 변수간의 관계를 살펴보고 plot 을 그려서 시각화로도 관계를 파악할 수 있었다. 4 가지의 잔차 그래프를 통해 잔차의 등분산성과 독립성을 검정할 수 있었고 몇 번째 관측치에서 outlier 와 high leverage point 가 존재하는지 파악할 수 있었다.

▶ Exercise 3.9 에서는 마찬가지로 Auto data 를 이용하여 다중 선형 회귀를 사용하였다. scatterplot matrix 와 상관계수 행렬을 통해 변수간의 관계를 살펴보았다. model 에 '*', ':'를 이용하여 상호작용항을 넣기도 하고 변수를 변형시켜가며 모델의 성능이 얼마나 증가했는지도 볼 수 있었다.

▶ Exercise 3.13 에서는 데이터를 임의로 생성시켜 단순 선형 회귀에 fitting 시켜 각각의 값과 실제 값을 비교하는 과정을 거쳤다. 같은 x 와 y 의 데이터로 noise 의 분산을 바꿔가며 진행하였는데, noise 가 작을수록 추정된 $\beta 0$ 과 $\beta 1$이 실제 $\beta 0$, $\beta 1$과 근사했고 예측력이 높아지며 신뢰구간이 좁아지는 것을 알 수 있었다.

▶ Exercise 3.14 에서는 공선성 문제에 대해 살펴보았다. 강한 양의 상관관계를 가지고 있는 x1 과 x2 를 generate 하여 model 을 fitting 시켰다. 독립변수 간의 강한 상관관계를 보이면 다중공선성 문제가 발생하고 그렇게 되면 회귀계수가 해당 변수의 종속변수에 미치는 영향력을 정확하게 설명하지 못하게 된다. 따라서 다중공선성을 고려하지 않고 회귀분석을 수행한 후 그 결과를 해석하면 잘못된 결론을 내리게 되는 문제가 발생한다.