R을 이용한 통계 기초와 데이터 분석

Lecture 7

남현진

한성대학교

2020

데이터

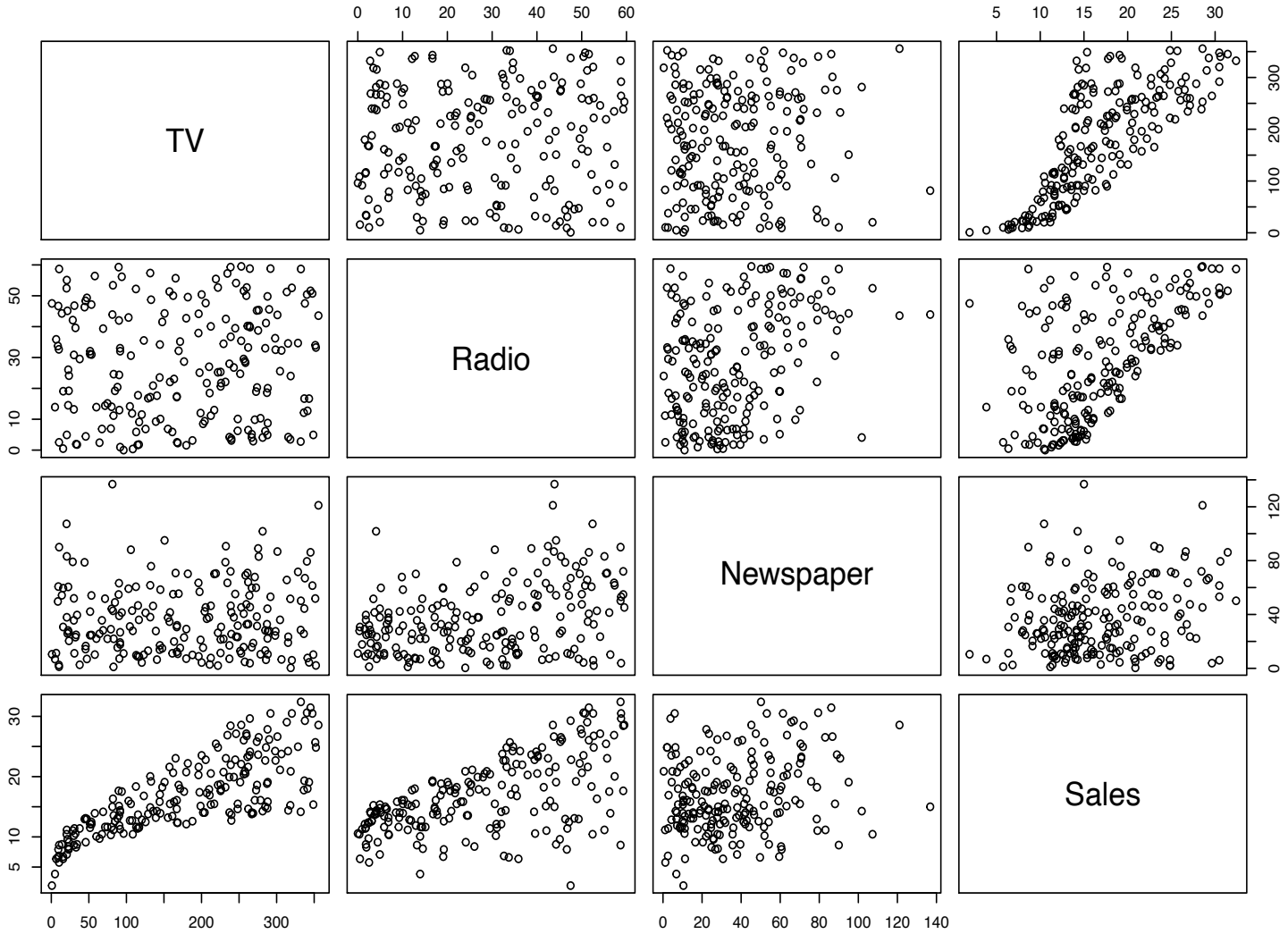
오늘 사용할 데이터는 200개의 세일즈 데이터로 판매 액(천 단위 달러)과 매체 별 광고 지출비가 천단위 달러로 포함되어있다. 매체는 총 세가지로 TV, Radio, 그리고 Newspaper가 있다.

```
data <- read.csv('data.csv')
str(data)
```

```
## 'data.frame':    200 obs. of  4 variables:
##  $ TV       : num  276.1 53.4 20.6 181.8 217 ...
##  $ Radio    : num  45.4 47.2 55.1 49.6 13 ...
##  $ Newspaper: num  83 54.1 83.2 70.2 70.1 ...
##  $ Sales    : num  26.5 12.5 11.2 22.2 15.5 ...
```

데이터

```
plot(data)
```



데이터

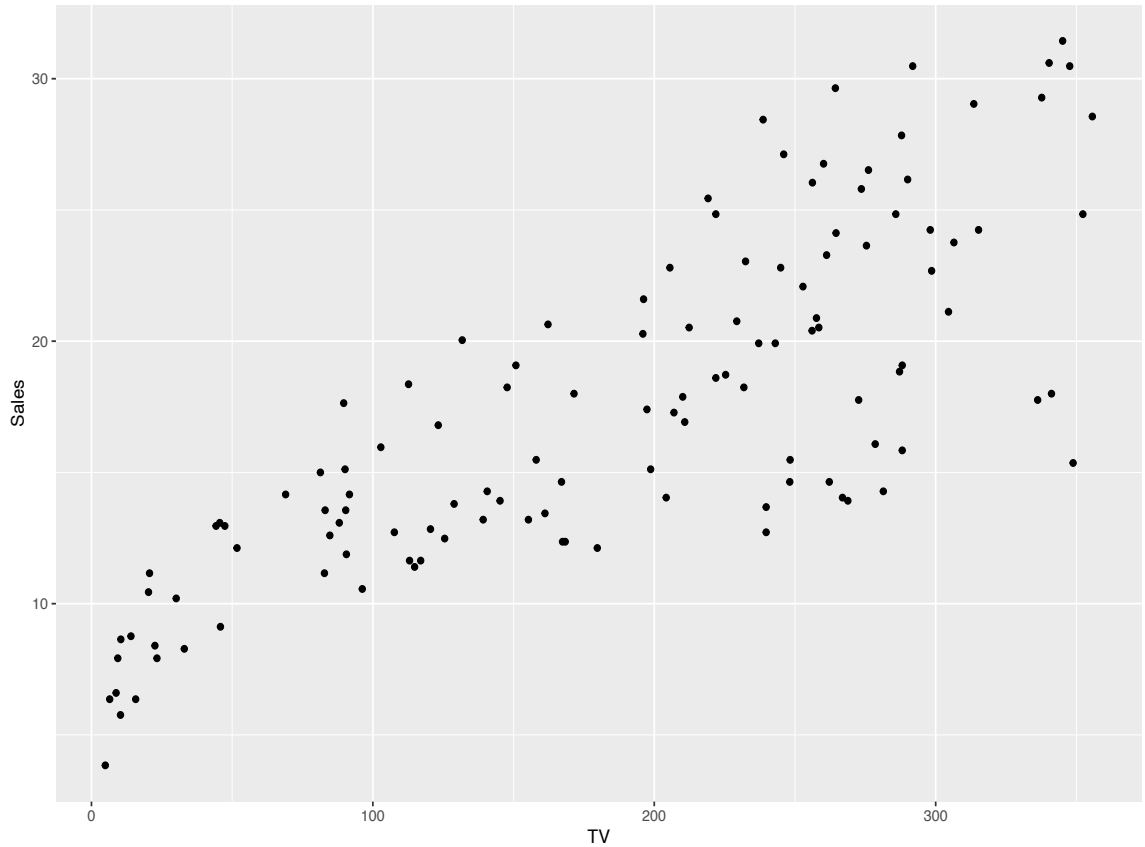
데이터를 6:3 비율로 train 데이터와 test 데이터로 나누어준다.

```
set.seed(123)
sample <- sample(c(TRUE, FALSE), nrow(data), replace = T, prob = c(0.6,0.4))
train <- data[sample, ]
test <- data[!sample, ]
```

단순 선형 회귀

우선 train 데이터를 가지고 TV 광고 지출액과 매출간의 관계를 알아보자.

```
ggplot(train, aes(TV, Sales)) +
  geom_point()
```



단순 선형 회귀

TV 광고 지출액과 매출간의 관계를 보여줄 수 있는 단순 선형 회귀 모형을 만들어보자.

```
library(broom)
library(modelr)
library(ggplot2)
model1 <- lm(Sales ~ TV, data = train)
model1_results <- augment(model1, train)
```

단순 선형 회귀

```
summary(model1)
```

```
##
## Call:
## lm(formula = Sales ~ TV, data = train)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2979  -2.1414  -0.3039   2.6058   8.3214
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 8.116917   0.729110   11.13   <2e-16 ***
## TV          0.050284   0.003463   14.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.845 on 120 degrees of freedom
## Multiple R-squared:  0.6373, Adjusted R-squared:  0.6342
## F-statistic: 210.8 on 1 and 120 DF,  p-value: < 2.2e-16
```

모형 해석

회귀식은 다음과 같이 정의할 수 있다.

$$Sales = 8.12 + 0.05\,TV$$

- 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.

단순 선형 회귀

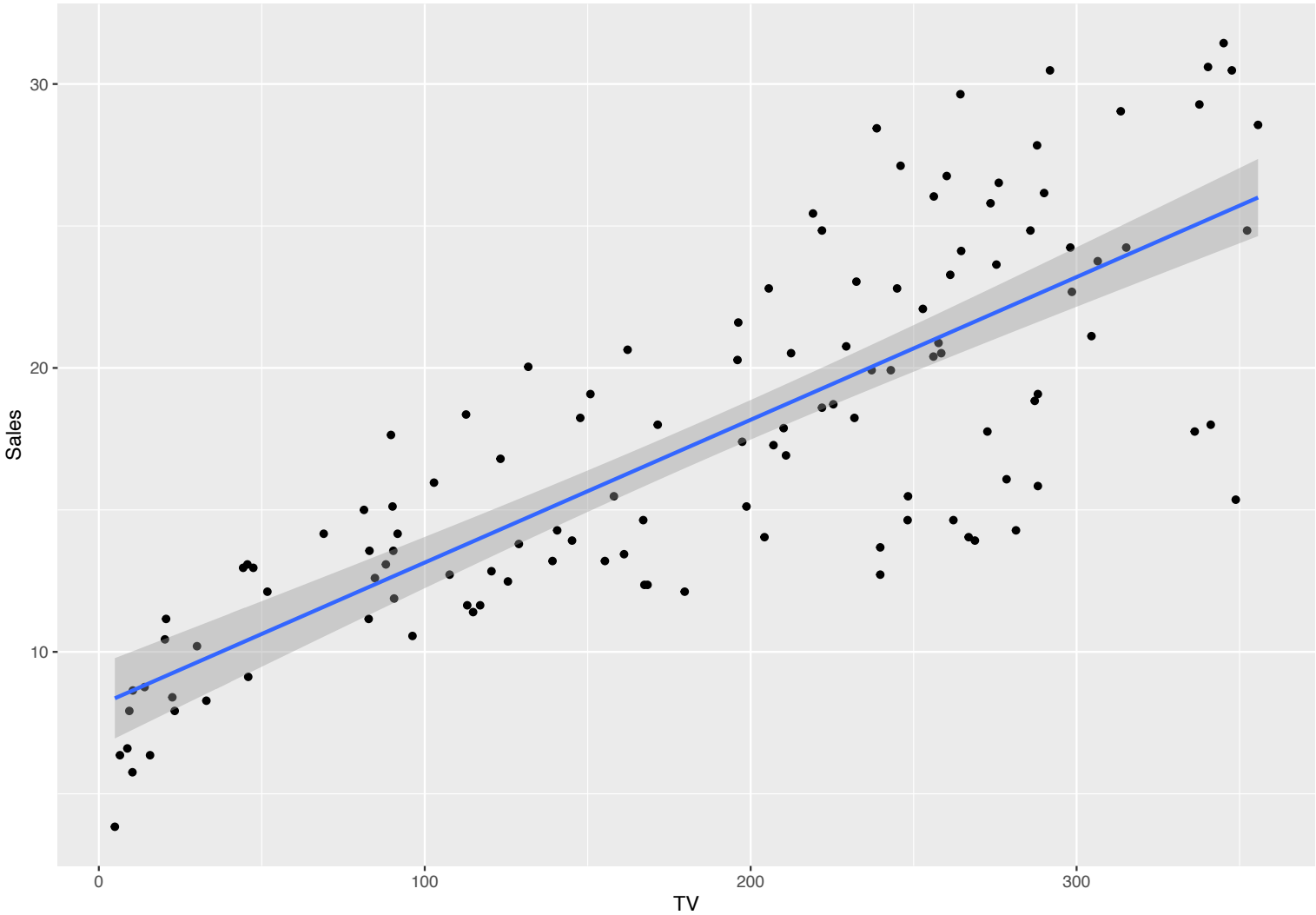
단순 선형 회귀 모형 Model1을 산점도에 올려보면 다음과 같다.

```
ggplot(train, aes(TV, Sales)) +
  geom_point() +
  geom_smooth(method = "lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```

단순 선형 회귀

```
## `geom_smooth()` using formula 'y ~ x'
```



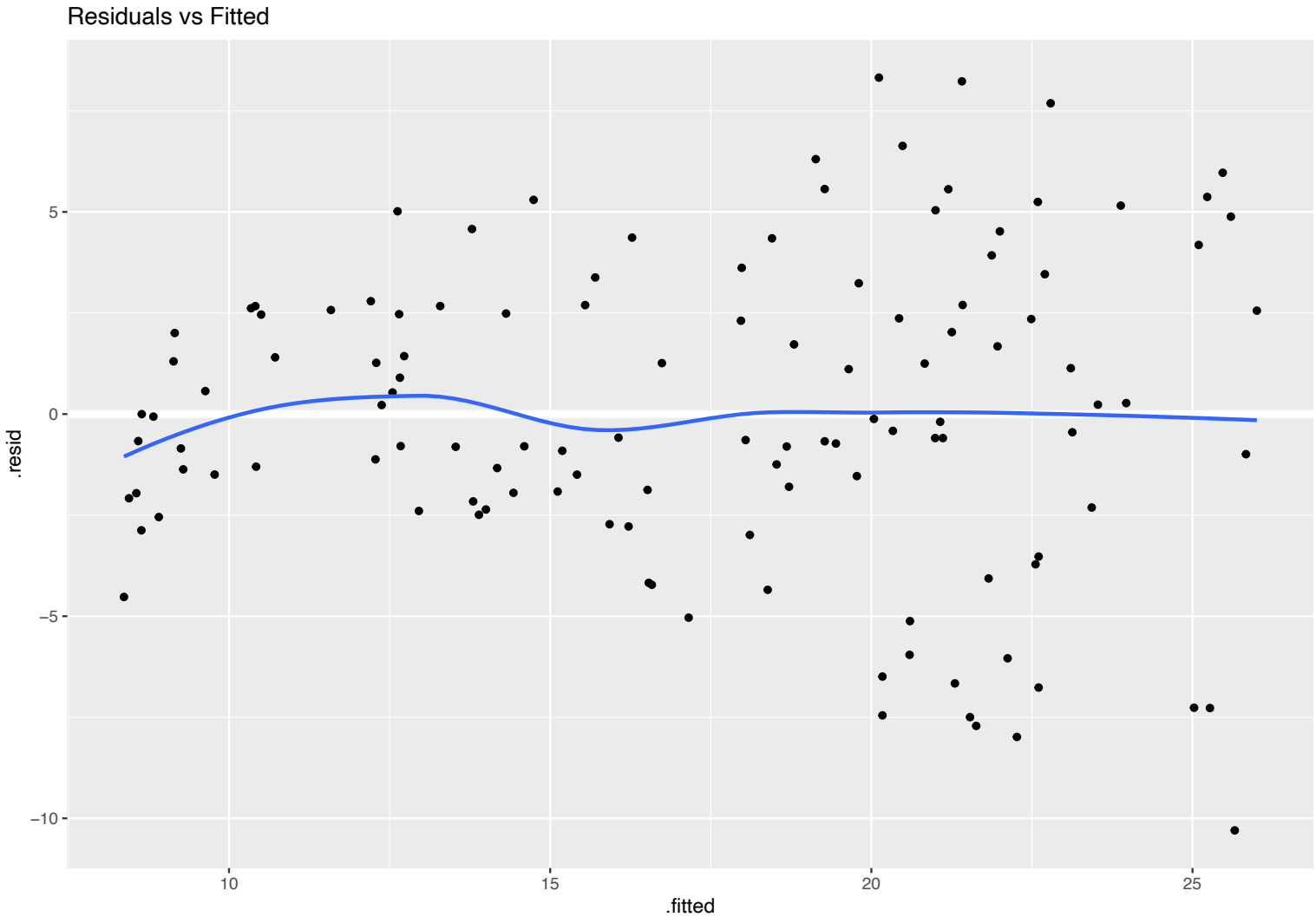
단순 선형 회귀

Residuals vs Fitted 플랏은 X 축이 선형 회귀로 예측된 Y 값이며 Y 축에는 잔차를 보여준다. 선형 회귀에서 오차는 평균이 0이고 분산이 일정한 정규 분포를 가정하였으므로, 예측된 Y 값과 무관하게 잔차의 평균은 0이고 분산은 일정해야 한다. 따라서 이 그래프에서는 기울기 0인 직선이 관측되는 것이 이상적이다.

```
model1_plot1 <- ggplot(model1_results, aes(.fitted, .resid)) +
  geom_ref_line(h = 0) +
  geom_point() +
  geom_smooth(se = FALSE) +
  ggtitle("Residuals vs Fitted")
```

단순 선형 회귀

```
## `geom_smooth()` using method = 'loess' and formula 'y ~ x'
```

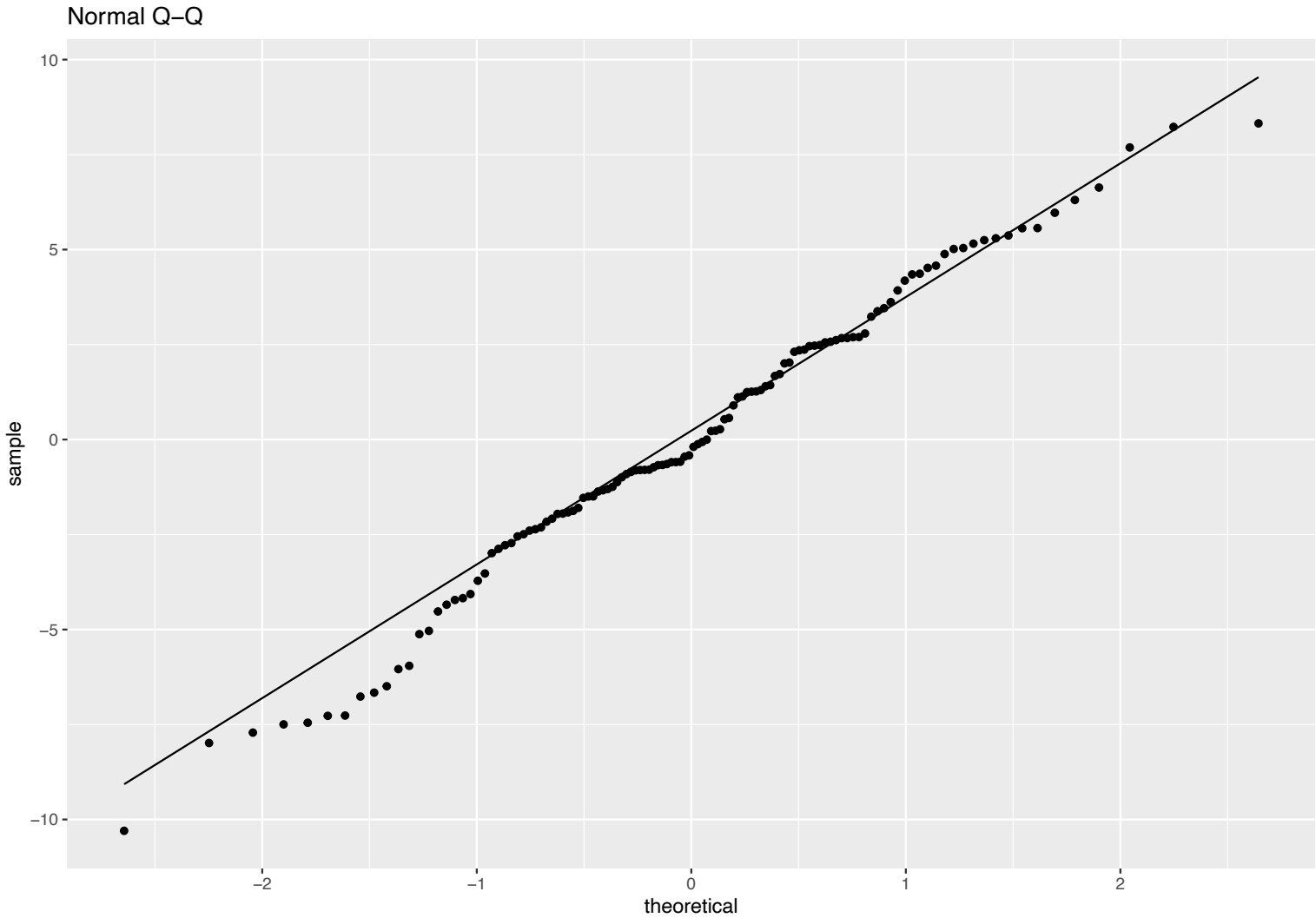


단순 선형 회귀

Normal Q-Q는 잔차가 정규 분포를 따르는지 확인하기 위한 Q-Q Plot을 보여준다. 즉 잔차의 분포의 quantile 값과 정규 분포의 quantile 값들을 비교해주는 플랏인데 만약 점들이 선형의 직선위에 올라와 있다면 잔차가 정규성을 따른다고 말할 수 있다.

```
model1_plot2 <- ggplot( model1_results, aes( sample = .resid)) +
  stat_qq()+
  stat_qq_line() +
  ggtitle("Normal Q-Q")
```

단순 선형 회귀



단순 선형 회귀

TV 광고 지출액을 가지고 만든 단순 회귀 모형을 이용해 테스트 데이터의 매출액을 예측해보자. MSE는 예측한 값의 오차 제곱 합을 나타내는데 이는 회귀 모형이 테스트에 얼마나 잘 적합되었는지를 보여주는 비용함수이다. MSE 값이 작으면 작을수록 모델이 잘 적합도었다고 말할 수 있다.

```
pred1 <- predict(model1, test)
test$Pred1 <- pred1
mse1 <- mean((test$Sales - test$Pred1)^2)
mse1
```

```
## [1] 16.34391
```

다중 선형 회귀

앞에서 만든 단순 선형 회귀 분석에 다른 변수들 또한 추가해서 다중 선형 회귀 모형을 만들어 보자.

```
model2 <- lm(Sales ~ ., data = train)
model2 <- lm(Sales ~ TV + Radio + Newspaper + Sales, data = train)
```

```
## Warning in model.matrix.default(mt, mf, contrasts): the response appeared on the
## right-hand side and was dropped
```

```
## Warning in model.matrix.default(mt, mf, contrasts): problem with term 4 in
## model.matrix: no columns are assigned
```

다중 선형 회귀

```
summary(model2)
##
## Call:
## lm(formula = Sales ~ TV + Radio + Newspaper + Sales, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -5.8111 -0.7760  0.2598  1.2768  3.2165
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.386647   0.443243   7.641 6.29e-12 ***
## TV           0.047362   0.001657  28.577  < 2e-16 ***
## Radio        0.196375   0.010347  18.979  < 2e-16 ***
## Newspaper   -0.010593   0.006460  -1.640    0.104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.833 on 118 degrees of freedom
## Multiple R-squared:  0.9189, Adjusted R-squared:  0.9169
## F-statistic: 445.9 on 3 and 118 DF,  p-value: < 2.2e-16
```

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.

다중 선형 회귀

앞에서 만든 다중 선형 회귀 모형 model2에서 유의하지 않다고 나온 Newspaper를 변수에서 제거한 새로운 다중 선형 회귀 모형을 만들어 보자.

```
model3 <- lm(Sales ~ TV + Radio , data = train)
model3_results <- augment(model3, train)
```

다중 선형 회귀

```
summary(model3)
##
## Call:
## lm(formula = Sales ~ TV + Radio, data = train)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.1372 -0.8006  0.2107  1.2934  3.1718
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3.180344   0.428014    7.43  1.8e-11 ***
## TV          0.047413   0.001669   28.41  < 2e-16 ***
## Radio       0.189193   0.009440   20.04  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.846 on 119 degrees of freedom
## Multiple R-squared:  0.9171, Adjusted R-squared:  0.9157
## F-statistic: 658.2 on 2 and 119 DF,  p-value: < 2.2e-16
```

모형 해석

회귀식은 다음과 같이 정의할 수 있다.

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

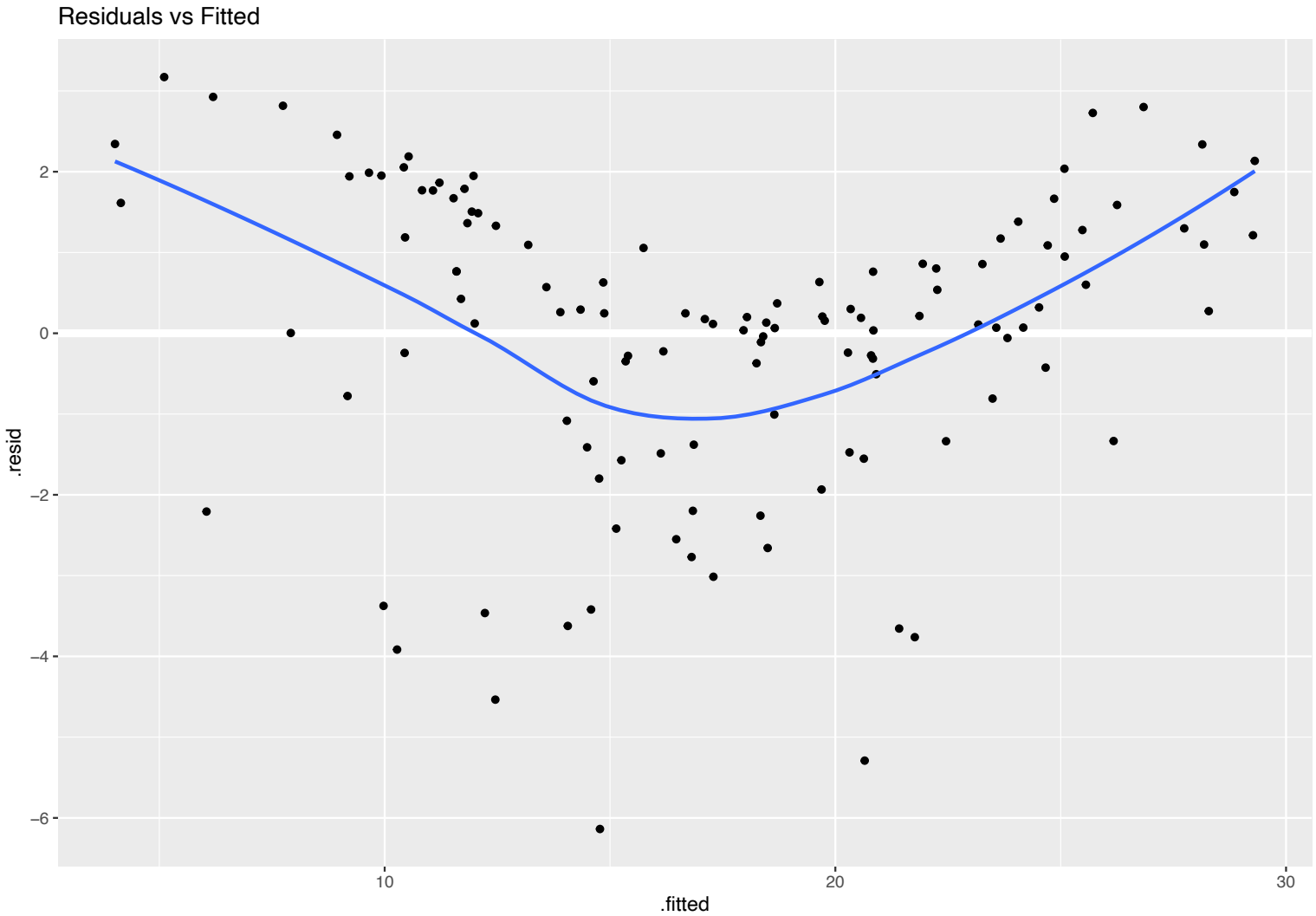
- 상수항은 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 상수항의 계수가 0이 아니라고 말할 수 있다.
- TV, Radio는 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 TV의 계수가 0이 아니라고 말할 수 있다.
- F 통계량은 658.2로 p-value가 0.05보다 작다. 따라서 귀무가설을 기각하고 $\beta_0, \beta_1, \beta_2$ 중 적어도 하나는 0 이 아니라고 결론을 내릴 수 있다.
- R^2 는 0.9179이다. 즉 모델은 자료의 91.79 퍼센트의 변동성을 설명할 수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

다중 선형 회귀

Model3을 이용해서 Residuals vs Fitted 플랏을 그려보자. Model3의 결과는 Model1 보다 F통계량과 R^2 측면에서는 좋아보였지만 단차의 측면에서는 Model1이 조금 더 이상적인 모습을 보여준다. 잔차의 분산은 일정하지 않은 모습을 보이고 있다.

```
model3_plot1 <- ggplot(model3_results, aes(.fitted, .resid)) +
  geom_ref_line(h = 0) +
  geom_point() +
  geom_smooth(se = FALSE) +
  ggtitle("Residuals vs Fitted")
```

다중 선형 회귀

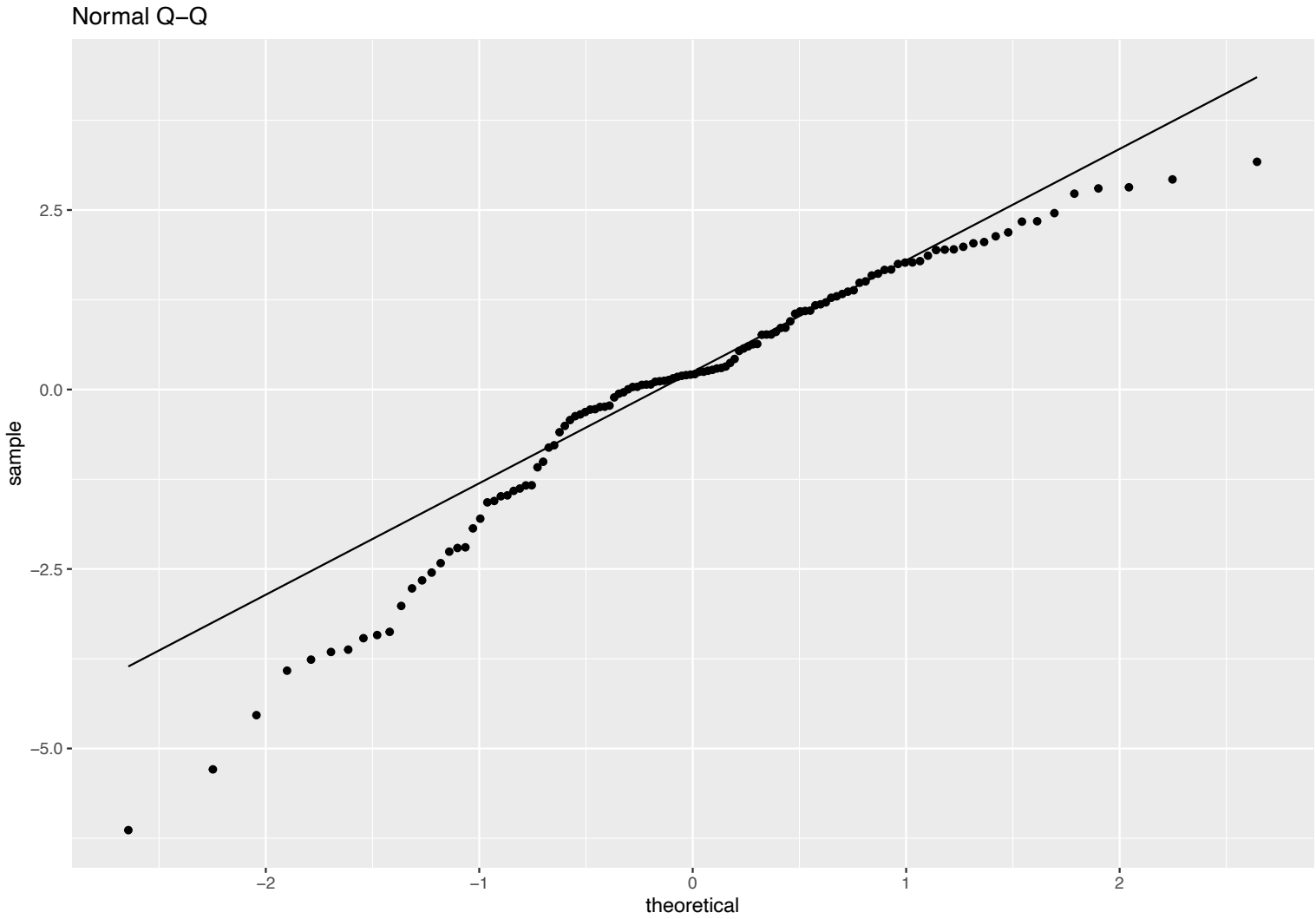


다중 선형 회귀

Normal Q-Q는 잔차가 정규 분포를 따르는지 확인하기 위한 Q-Q Plot을 보여준다. Model3의 결과는 Model1 보다 정규성이 조금 덜 만족된다고 볼 수 있다.

```
model3_plot2 <- ggplot( model3_results, aes( sample = .resid)) +
  stat_qq()+
  stat_qq_line() +
  ggtitle("Normal Q-Q")
```

다중 선형 회귀



다중 선형 회귀

TV와 Radio 광고 지출액과 가지고 만든 다중 회귀 모형을 이용해 테스트 데이터의 매출액을 예측해보자.

```
pred3 <- predict(model3, test)
test$Pred3 <- pred3
mse3 <- mean((test$Sales - test$Pred3)^2)
```

다중 선형 회귀

Model1과 Model3의 MSE를 비교해보면 MSE1는 16.34391이고 MSE3는 5.159816이다. 즉 선형 회귀 모형에서 여러가지 변수를 추가함으로써 데이터의 변동성을 더 잘 설명할 수 있다고 말할 수 있다.

```
mse1
```

```
## [1] 16.34391
```

```
mse3
```

```
## [1] 5.159816
```