Regression

5. Econometrics with R project

5.1 단순회귀분석

5.1.1 회귀분석의 기초 개념

* 회귀분석이 연속형 변수들에 대해 독립 변수와 종속 변수 사이의 상관관계를 나타내는 것이라면, 단순 회귀 분석은 독립 변수가 단일개일 때의 분석을 의미한다
* 기본적인 회귀모형은, \[y\_{i} = \beta\_0 + \beta\_1X\_{i} + e\_{i}\]

library(AER)

## Loading required package: car

## Loading required package: carData

## Loading required package: lmtest

## Loading required package: zoo

##   
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':  
##   
## as.Date, as.Date.numeric

## Loading required package: sandwich

## Loading required package: survival

library(MASS)

STR <- c(15, 17, 19, 20, 22, 23.5, 25)  
TestScore <- c(680, 640, 670, 660, 630, 660, 635)   
  
# Print out sample data  
STR

## [1] 15.0 17.0 19.0 20.0 22.0 23.5 25.0

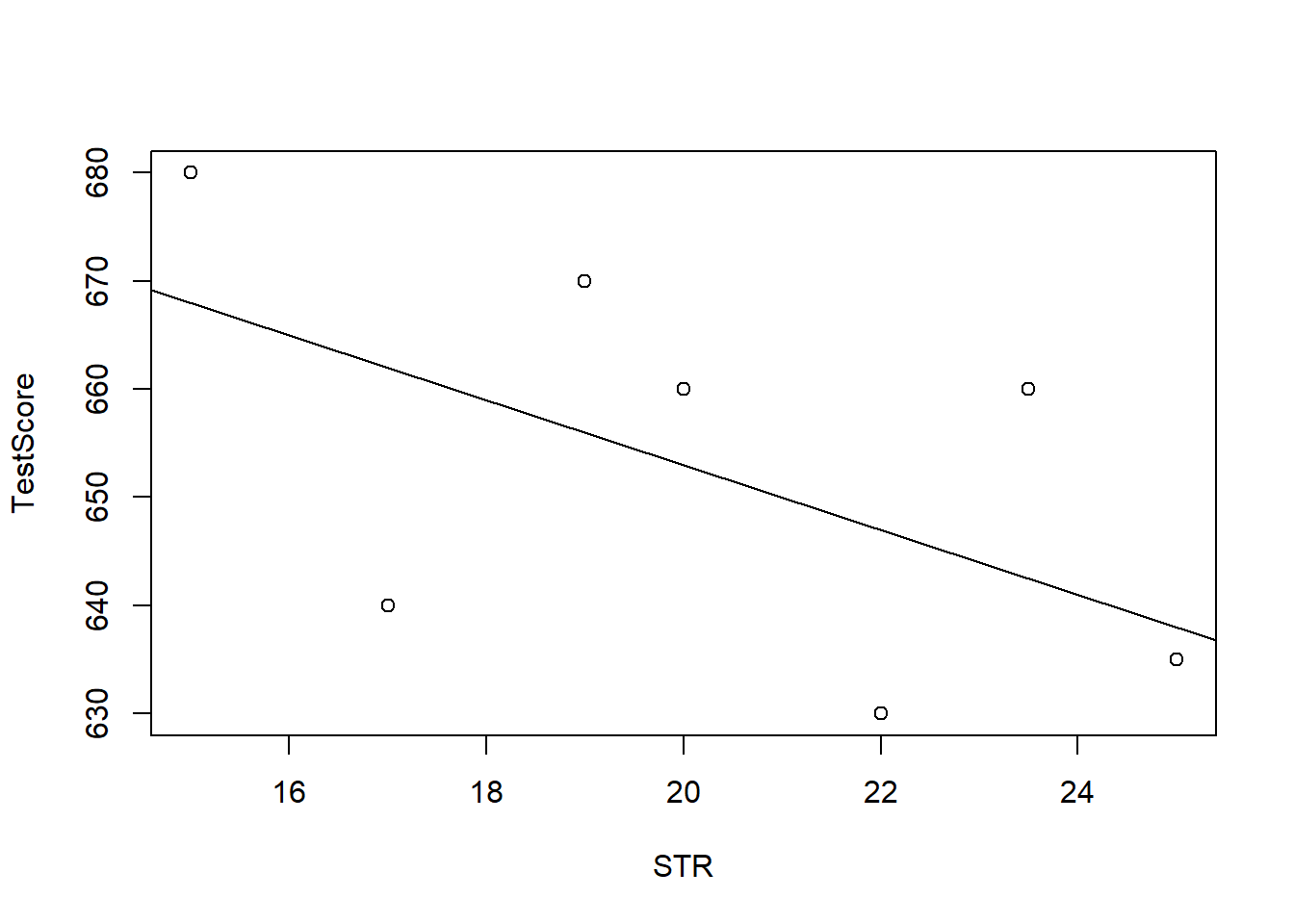
TestScore

## [1] 680 640 670 660 630 660 635

위 데이터를 활용하여 산점도를 만들기

* abline(a,b) : 절편이 a, 기울기가 b인 직선 추가

plot(TestScore ~ STR)  
abline(a = 713, b = -3)



Use CASchools data

data(CASchools)  
head(CASchools)

## district school county grades students teachers  
## 1 75119 Sunol Glen Unified Alameda KK-08 195 10.90  
## 2 61499 Manzanita Elementary Butte KK-08 240 11.15  
## 3 61549 Thermalito Union Elementary Butte KK-08 1550 82.90  
## 4 61457 Golden Feather Union Elementary Butte KK-08 243 14.00  
## 5 61523 Palermo Union Elementary Butte KK-08 1335 71.50  
## 6 62042 Burrel Union Elementary Fresno KK-08 137 6.40  
## calworks lunch computer expenditure income english read math  
## 1 0.5102 2.0408 67 6384.911 22.690001 0.000000 691.6 690.0  
## 2 15.4167 47.9167 101 5099.381 9.824000 4.583333 660.5 661.9  
## 3 55.0323 76.3226 169 5501.955 8.978000 30.000002 636.3 650.9  
## 4 36.4754 77.0492 85 7101.831 8.978000 0.000000 651.9 643.5  
## 5 33.1086 78.4270 171 5235.988 9.080333 13.857677 641.8 639.9  
## 6 12.3188 86.9565 25 5580.147 10.415000 12.408759 605.7 605.4

* STR, TestScore 파생변수 생성

CASchools$STR <- CASchools$students/CASchools$teachers   
CASchools$score <- (CASchools$read + CASchools$math)/2

* STR, score 평균, 표준편차 계산하기

avg\_STR <- mean(CASchools$STR)   
avg\_score <- mean(CASchools$score)  
  
sd\_STR <- sd(CASchools$STR)   
sd\_score <- sd(CASchools$score)

* quantile(data,quantiles) : data를 quantiles 기준으로 나누어줌

quantiles <- c(0.10, 0.25, 0.4, 0.5, 0.6, 0.75, 0.9)  
quant\_STR <- quantile(CASchools$STR, quantiles)  
quant\_score <- quantile(CASchools$score, quantiles)

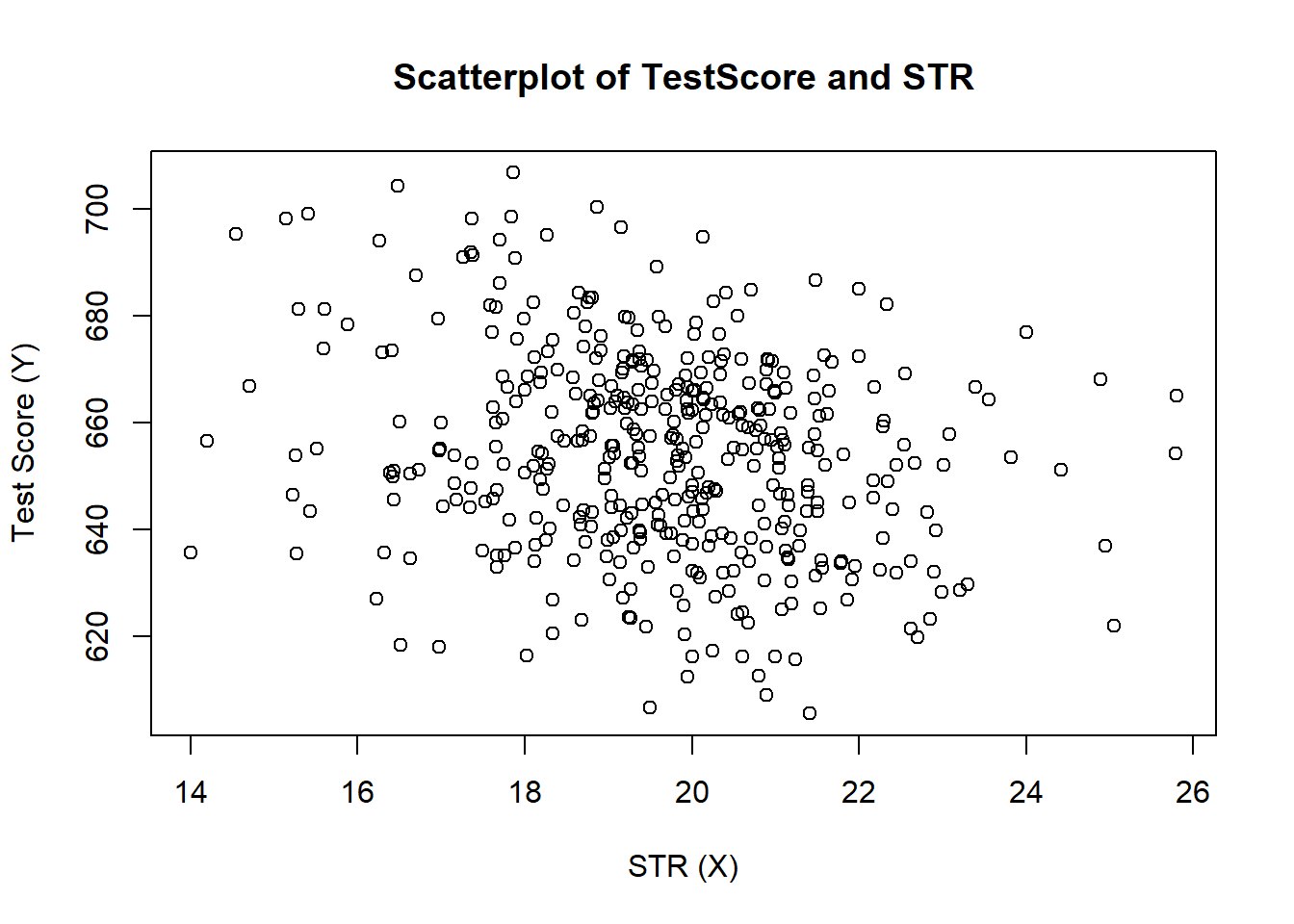
# gather everything in a data.frame   
DistributionSummary <- data.frame(Average = c(avg\_STR, avg\_score),   
 StandardDeviation = c(sd\_STR, sd\_score),   
 quantile = rbind(quant\_STR, quant\_score))

DistributionSummary

## Average StandardDeviation quantile.10. quantile.25. quantile.40.  
## quant\_STR 19.64043 1.891812 17.3486 18.58236 19.26618  
## quant\_score 654.15655 19.053347 630.3950 640.05000 649.06999  
## quantile.50. quantile.60. quantile.75. quantile.90.  
## quant\_STR 19.72321 20.0783 20.87181 21.86741  
## quant\_score 654.45000 659.4000 666.66249 678.85999

* TestScore, STR에 대한 산점도
  + 산점도를 살펴 볼 때 음의 상관관계를 미세하게 확인 할 수 있음

plot(score ~ STR,   
 data = CASchools,  
 main = "Scatterplot of TestScore and STR",   
 xlab = "STR (X)",  
 ylab = "Test Score (Y)")



* TestScore, STR 상관관계
  + cor(): +일경우 양의 상관관계, -일경우 음의 상관관계, 0에 가까울 수록 서로 독립이다

cor(CASchools$STR, CASchools$score)

## [1] -0.2263627

5.1.2 단순선형 회귀모형

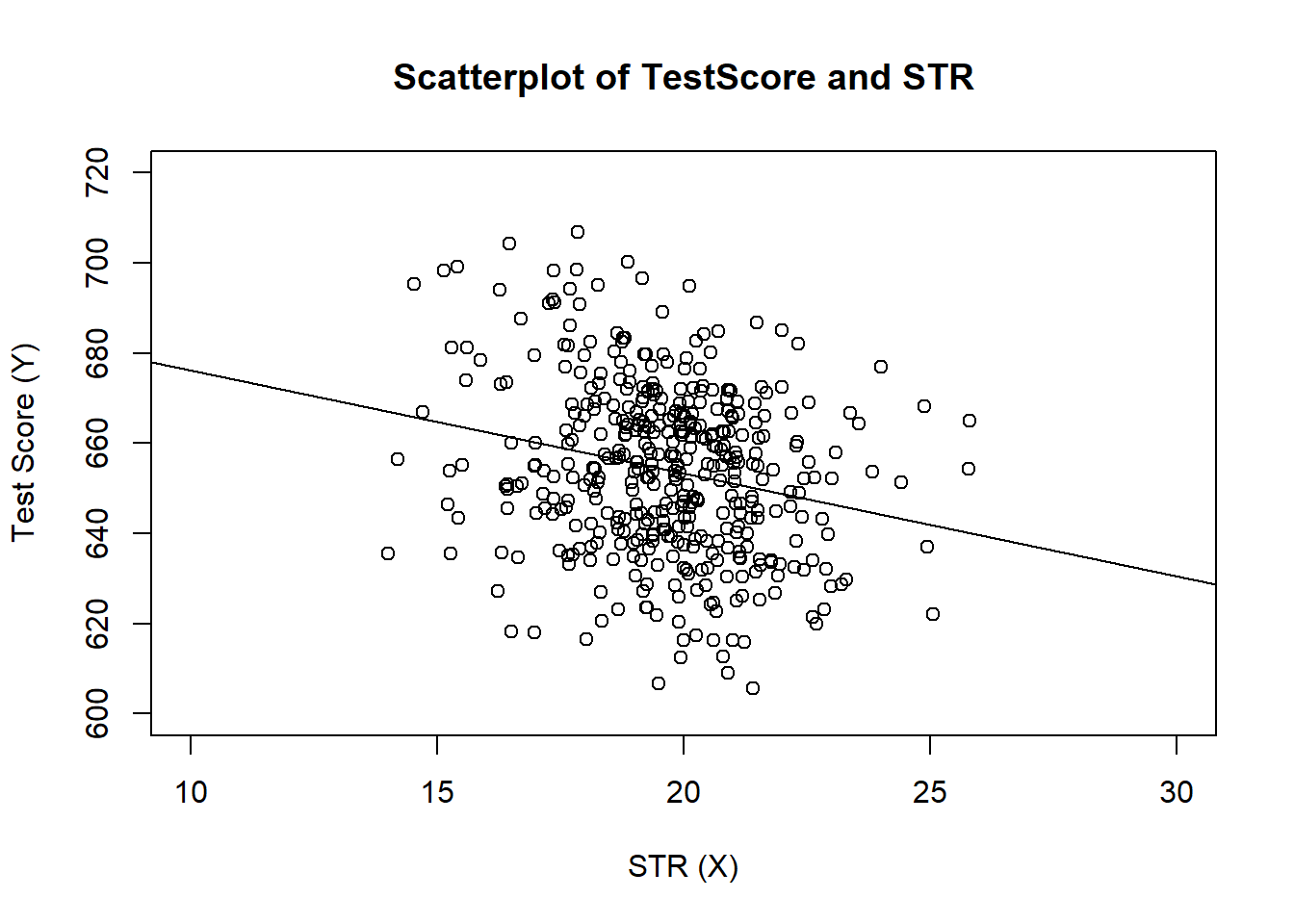
* estimate the model and assign the result to linear\_model
  + \(score = 698.93 -2.28\*STR\)
  + STR이 한단위 증가함에 따라 score은 2.28만큼 감소한다

linear\_model <- lm(score ~ STR, data = CASchools)  
linear\_model

##   
## Call:  
## lm(formula = score ~ STR, data = CASchools)  
##   
## Coefficients:  
## (Intercept) STR   
## 698.93 -2.28

plot the data

plot(score ~ STR,   
 data = CASchools,  
 main = "Scatterplot of TestScore and STR",   
 xlab = "STR (X)",  
 ylab = "Test Score (Y)",  
 xlim = c(10, 30),  
 ylim = c(600, 720))  
  
## add the regression line  
abline(linear\_model)



5.1.3 회귀계수의 추정

* beta\_1\_hat, beta\_0\_hat 구하기

attach(CASchools)

## The following object is masked \_by\_ .GlobalEnv:  
##   
## STR

beta\_1 <- sum((STR - mean(STR)) \* (score - mean(score))) / sum((STR - mean(STR))^2)  
beta\_0 <- mean(score) - beta\_1 \* mean(STR)  
  
beta\_1

## [1] -5.548841

beta\_0

## [1] 766.3224

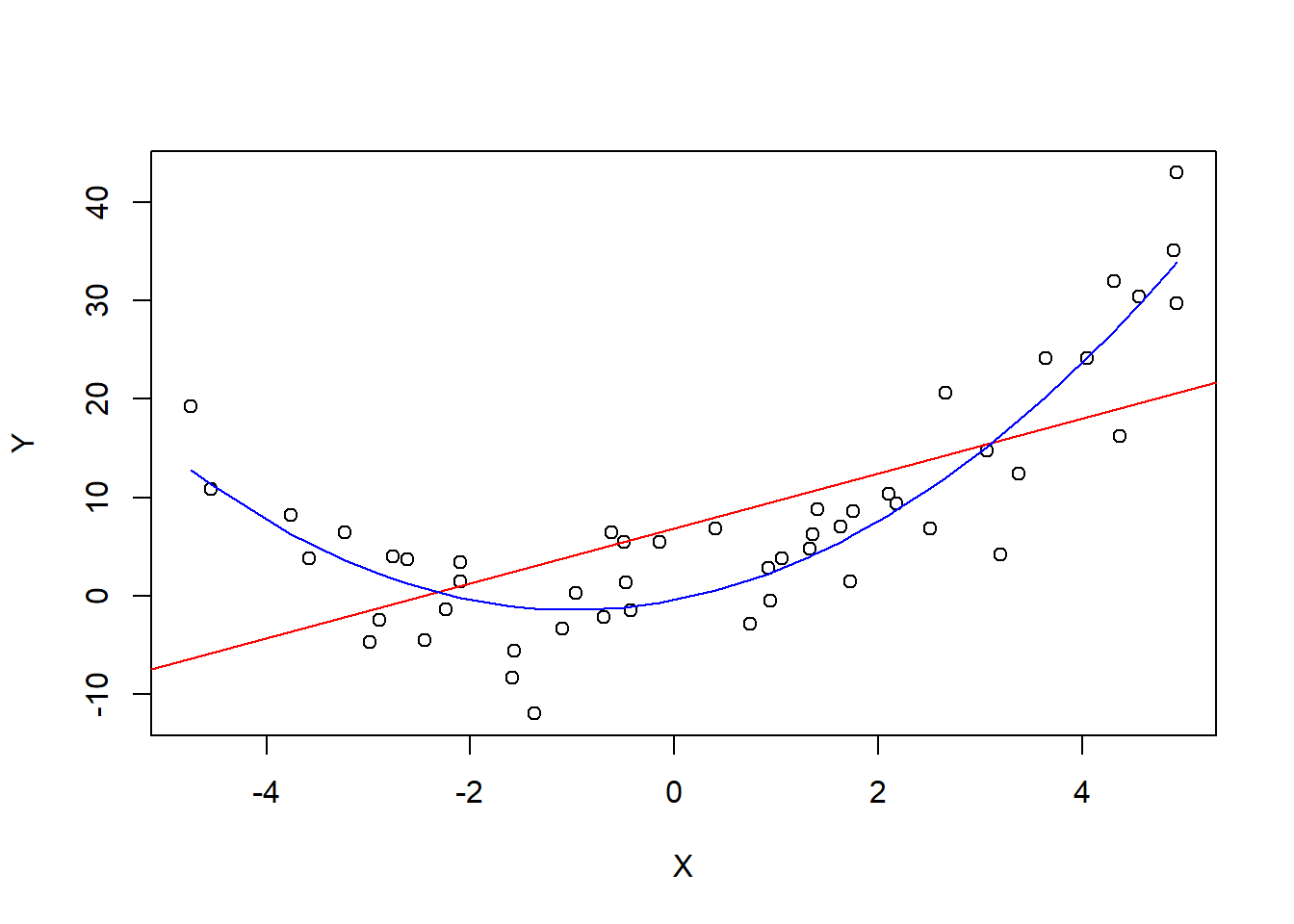
* p-value를 확인하면 매우 작은 값으로 \*\*\*로 나타나며 유의하다고 말 할 수 있다
  + \*\* : <0.001
  + \* : <0.05
  + . : <0.1
* R-squared: 결정계수
  + 회귀식의 적합도를 재는 척도
* Adjusted R-squared: 수정된 결정계수
  + 회귀식의 적합도를 재는 척도 다중회귀분석 시 변수가 추가 됨에 따라 증가하는 r-squared를 보정해준다
* F-statistic: 22.58 on 1 and 418 DF, p-value: 2.783e-06
  + F 통계량을 통해 회귀식의 유의성을 판단

mod\_summary <- summary(linear\_model)  
mod\_summary

##   
## Call:  
## lm(formula = score ~ STR, data = CASchools)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -47.727 -14.251 0.483 12.822 48.540   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 698.9329 9.4675 73.825 < 2e-16 \*\*\*  
## STR -2.2798 0.4798 -4.751 2.78e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.58 on 418 degrees of freedom  
## Multiple R-squared: 0.05124, Adjusted R-squared: 0.04897   
## F-statistic: 22.58 on 1 and 418 DF, p-value: 2.783e-06

* set.seed() : 난수를 고정 시켜 고정된 값이 나오게끔한다
* lm() : 회귀모델 추정 함수
* predict() : 추정된 회귀식을 적용시켜 종속변수를 예측

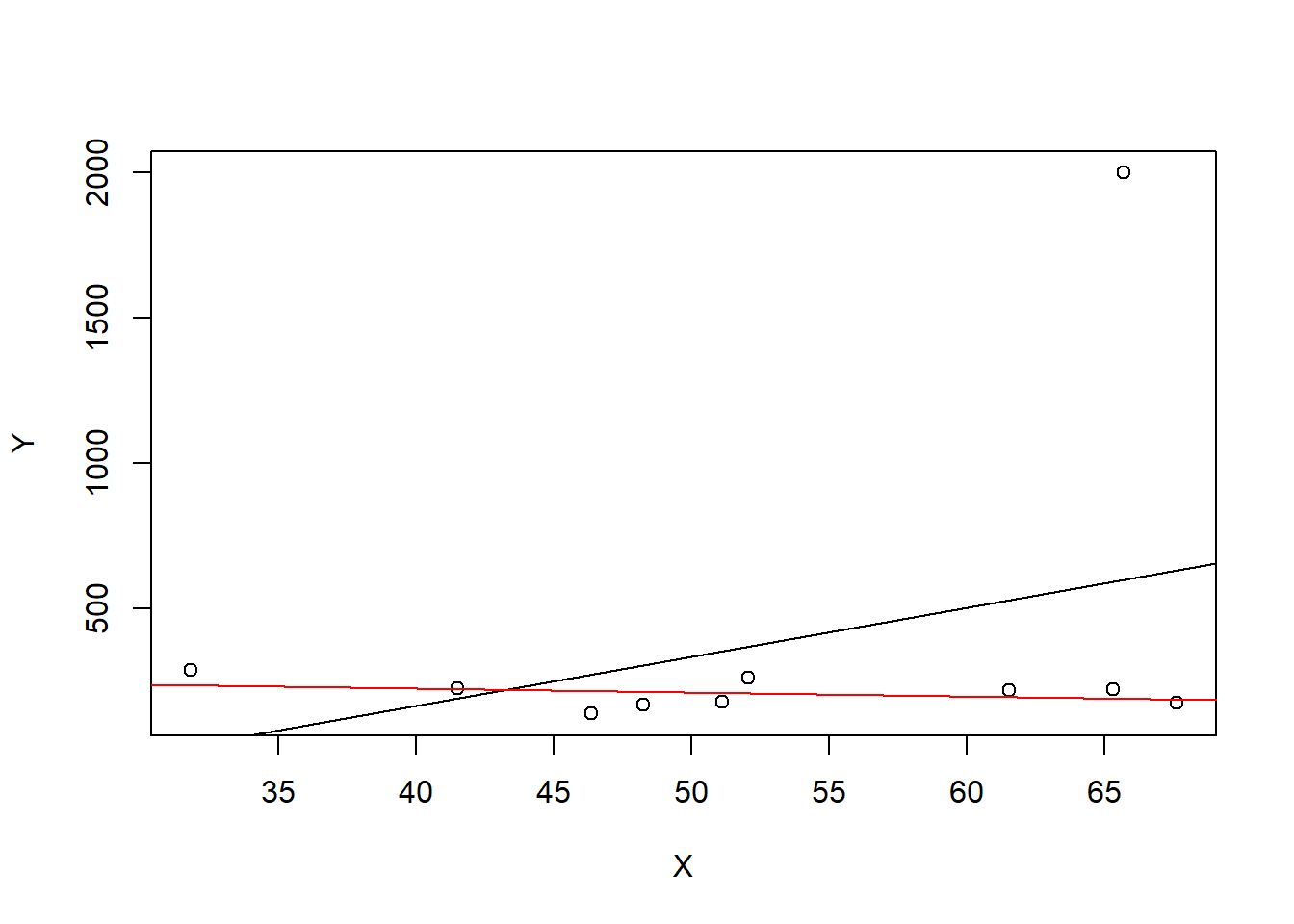
set.seed(321)  
X <- runif(50, min = -5, max = 5)  
u <- rnorm(50, sd = 5)   
  
Y <- X^2 + 2 \* X + u   
mod\_simple <- lm(Y ~ X)  
  
prediction <- predict(lm(Y ~ X + I(X^2)), data.frame(X = sort(X)))  
  
# plot the results  
plot(Y ~ X)  
abline(mod\_simple, col = "red")  
lines(sort(X), prediction,col="blue")



이상값

* outlier : 통계적 자료분석의 결과를 왜곡시키거나, 자료 분석의 적절성을 위협하는 변수값 또는 사례를 말한다
  + outlier에 의해 회귀분석의 결과가 심하게 달라질 수 있음

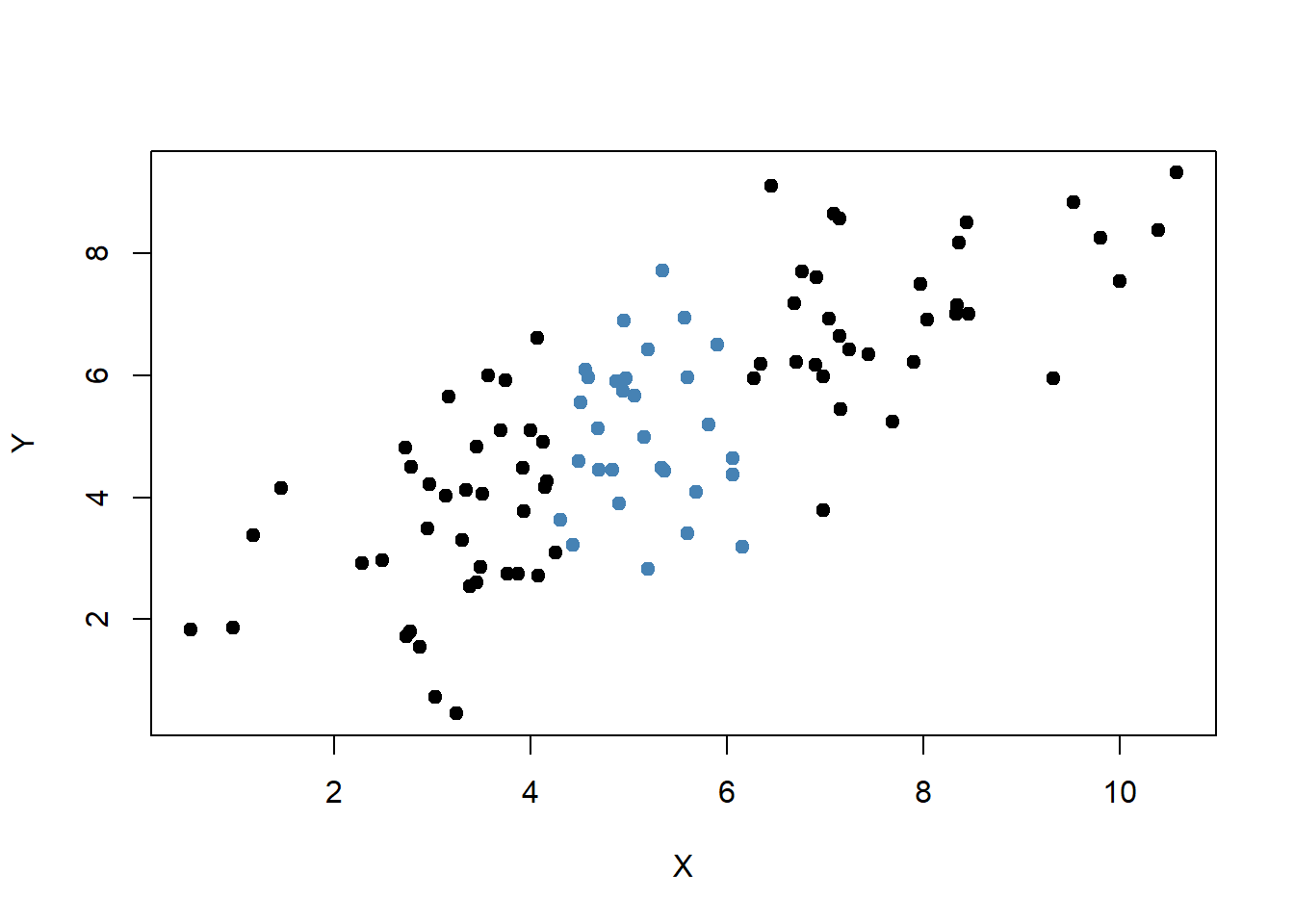
set.seed(123)  
  
# generate the data  
X <- sort(runif(10, min = 30, max = 70))  
Y <- rnorm(10 , mean = 200, sd = 50)  
Y[9] <- 2000  
  
# fit model with outlier  
fit <- lm(Y ~ X)  
  
# fit model without outlier  
fitWithoutOutlier <- lm(Y[-9] ~ X[-9])  
  
# plot the results  
plot(Y ~ X)  
abline(fit)  
abline(fitWithoutOutlier, col = "red")



실습1

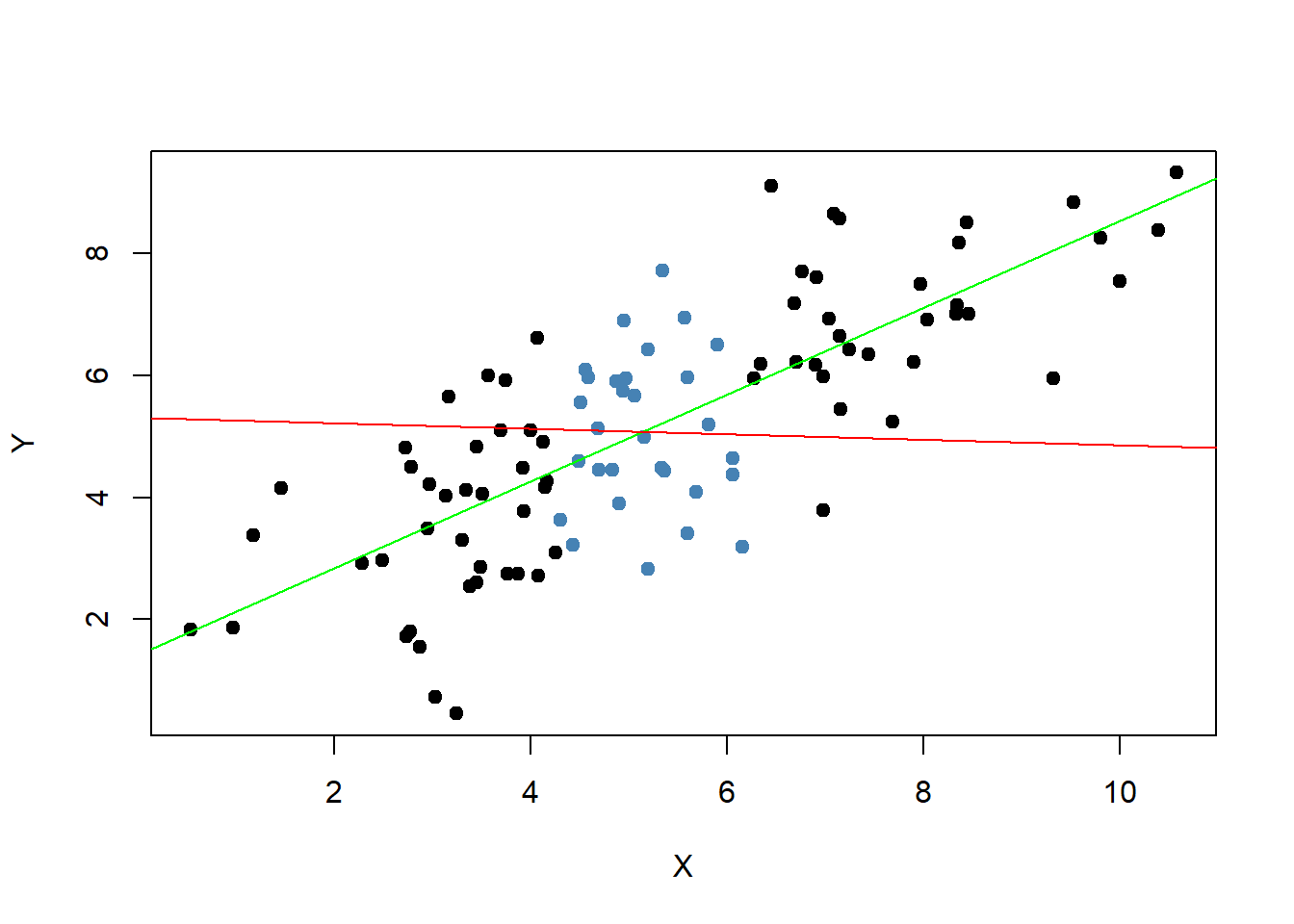
* 임의의 데이터 bvndata를 만들어 평균 기준으로 데이터를 나눈뒤 회귀분석을 실시해보자

# load the MASS package  
library(MASS)  
  
# set seed for reproducibility  
set.seed(4)  
  
# simulate bivarite normal data  
bvndata <- mvrnorm(100,   
 mu = c(5, 5),   
 Sigma = cbind(c(5, 4), c(4, 5)))   
  
# assign column names / convert to data.frame  
colnames(bvndata) <- c("X", "Y")  
bvndata <- as.data.frame(bvndata)  
  
# subset the data  
set1 <- subset(bvndata, abs(mean(X) - X) > 1)  
set2 <- subset(bvndata, abs(mean(X) - X) <= 1)  
  
# plot both data sets  
plot(set1,   
 xlab = "X",   
 ylab = "Y",   
 pch = 19)  
  
points(set2,   
 col = "steelblue",   
 pch = 19)



plot observations

# estimate both regression lines  
lm.set1 <- lm(Y ~ X, data = set1)  
lm.set2 <- lm(Y ~ X, data = set2)  
  
  
plot(set1, xlab = "X", ylab = "Y", pch = 19)  
points(set2, col = "steelblue", pch = 19)  
  
abline(lm.set1, col = "green")  
abline(lm.set2, col = "red")



5.1.4 회귀직선의 적합도

회귀직선의 적합도

* 가정된 회귀직선을 추정한 다음에는 그 회귀식이 얼마나 타당한가를 조사하여야 한다. 이를테면, 회귀분석의 목적은 종속변량을 독립변량의 함수로 설명하고자 함이므로 과연 그 설명의 정도가 어느 정도인지를 알아 볼 필요가 있다. 이와 같은 타당성 조사에는 잔차표준오차(residual standard error)와 결정계수(coefficient ofdetermination)가 사용된다

library(AER)  
library(scales)

* summary()$coefficients을 할 경우 회귀계수 추정값에 대한 정보만을 보여준다

# Testing Two-Sided Hypotheses Concerning the Slope Coefficient  
# load the `CASchools` dataset  
data(CASchools)  
  
# add student-teacher ratio  
CASchools$STR <- CASchools$students/CASchools$teachers  
  
# add average test-score  
CASchools$score <- (CASchools$read + CASchools$math)/2  
  
# estimate the model  
linear\_model <- lm(score ~ STR, data = CASchools)

* 귀무가설 : 독립변수와 종속변수간의 유의한 상관관계가 존재하지않는다
* 대립가설 : 독립변수와 종속변수간의 유의한 상관관계가 존재한다
  + p-value값을 살펴보면 < 0.05이므로 유의수준 5%하에서 귀무가설을 기각할만한 충분한 근거가된다

# print the summary of the coefficients to the console  
summary(linear\_model)$coefficients

## Estimate Std. Error t value Pr(>|t|)  
## (Intercept) 698.932949 9.4674911 73.824516 6.569846e-242  
## STR -2.279808 0.4798255 -4.751327 2.783308e-06

* Residual standard error: 18.58
* Adjusted R-squared: 0.04897

summary(linear\_model)

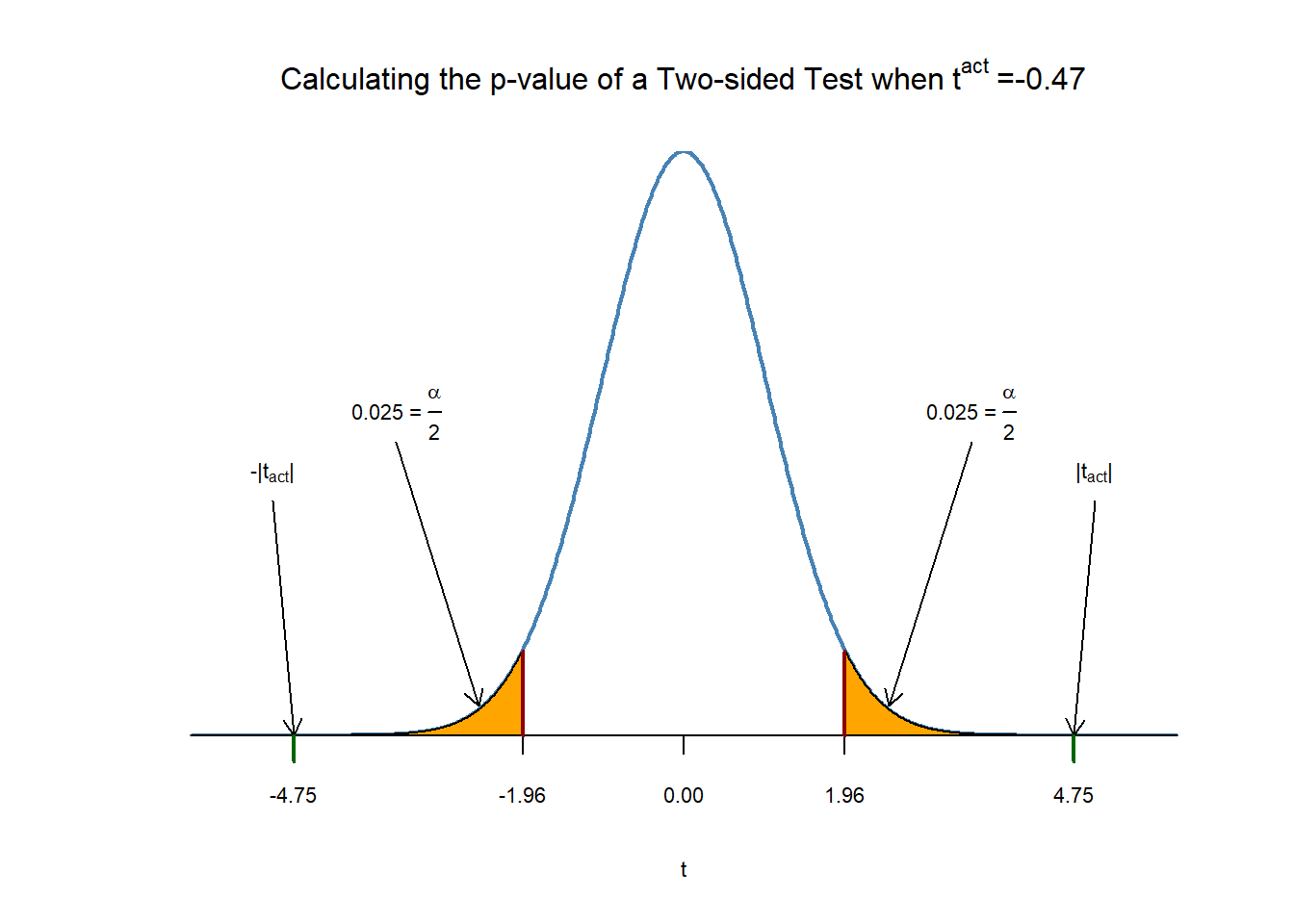
##   
## Call:  
## lm(formula = score ~ STR, data = CASchools)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -47.727 -14.251 0.483 12.822 48.540   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 698.9329 9.4675 73.825 < 2e-16 \*\*\*  
## STR -2.2798 0.4798 -4.751 2.78e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.58 on 418 degrees of freedom  
## Multiple R-squared: 0.05124, Adjusted R-squared: 0.04897   
## F-statistic: 22.58 on 1 and 418 DF, p-value: 2.783e-06

5.1.5 단순회귀 분석에서의 추론

T분포

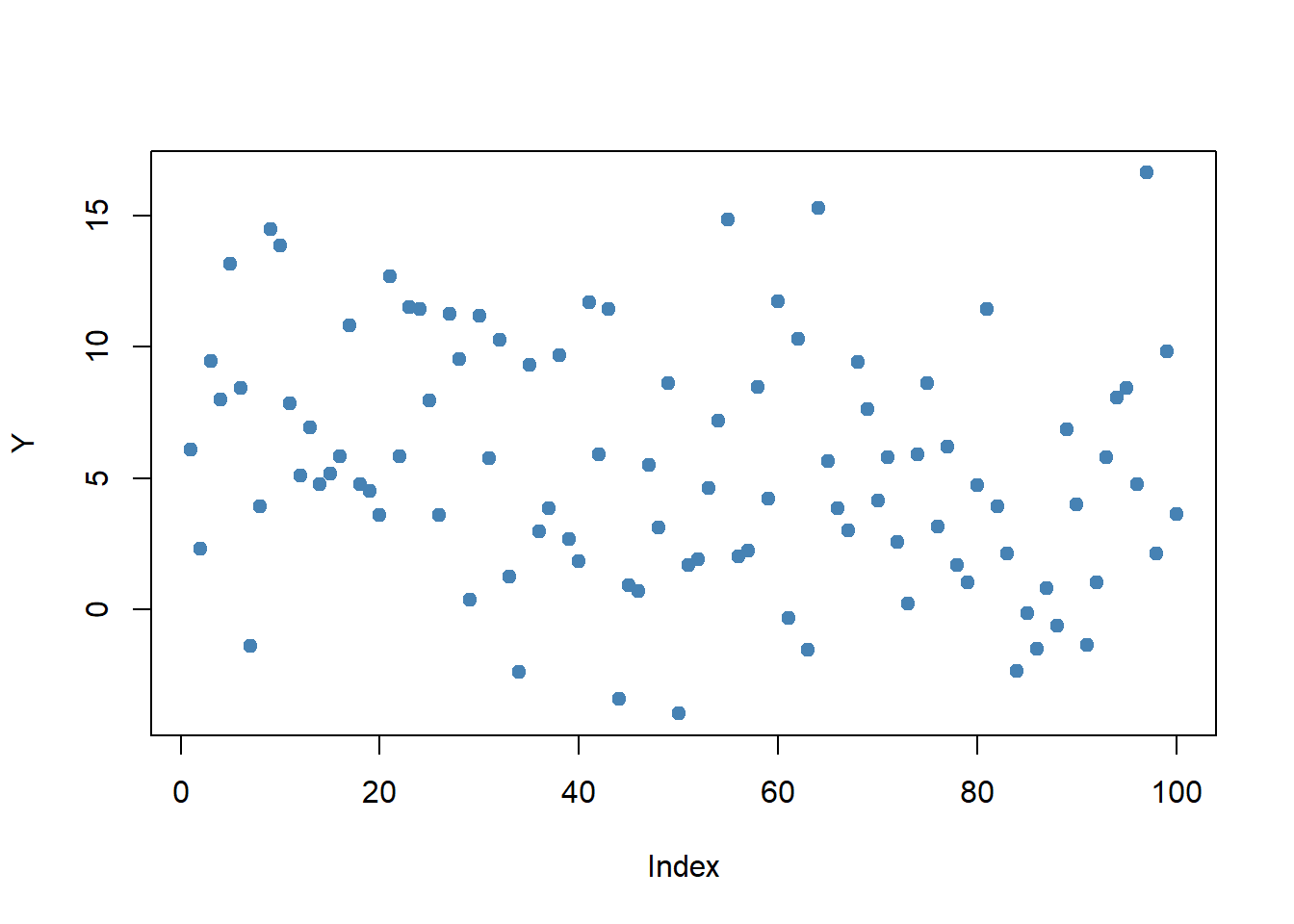
* p-value < 0.05 기준으로 귀무가설을 기각할 확률은 아래 보이는 그래프의 색칠된 부분이다

# Plot the standard normal on the support [-6,6]  
t <- seq(-6, 6, 0.01)  
  
plot(x = t,   
 y = dnorm(t, 0, 1),   
 type = "l",   
 col = "steelblue",   
 lwd = 2,   
 yaxs = "i",   
 axes = F,   
 ylab = "",   
 main = expression("Calculating the p-value of a Two-sided Test when" ~ t^act ~ "=-0.47"),   
 cex.lab = 0.7,  
 cex.main = 1)  
  
tact <- -4.75  
  
axis(1, at = c(0, -1.96, 1.96, -tact, tact), cex.axis = 0.7)  
  
# Shade the critical regions using polygon():  
  
# critical region in left tail  
polygon(x = c(-6, seq(-6, -1.96, 0.01), -1.96),  
 y = c(0, dnorm(seq(-6, -1.96, 0.01)), 0),   
 col = 'orange')  
  
# critical region in right tail  
  
polygon(x = c(1.96, seq(1.96, 6, 0.01), 6),  
 y = c(0, dnorm(seq(1.96, 6, 0.01)), 0),   
 col = 'orange')  
  
# Add arrows and texts indicating critical regions and the p-value  
arrows(-3.5, 0.2, -2.5, 0.02, length = 0.1)  
arrows(3.5, 0.2, 2.5, 0.02, length = 0.1)  
  
arrows(-5, 0.16, -4.75, 0, length = 0.1)  
arrows(5, 0.16, 4.75, 0, length = 0.1)  
  
text(-3.5, 0.22,   
 labels = expression("0.025"~"="~over(alpha, 2)),  
 cex = 0.7)  
text(3.5, 0.22,   
 labels = expression("0.025"~"="~over(alpha, 2)),  
 cex = 0.7)  
  
text(-5, 0.18,   
 labels = expression(paste("-|",t[act],"|")),   
 cex = 0.7)  
text(5, 0.18,   
 labels = expression(paste("|",t[act],"|")),   
 cex = 0.7)  
  
# Add ticks indicating critical values at the 0.05-level, t^act and -t^act   
rug(c(-1.96, 1.96), ticksize = 0.145, lwd = 2, col = "darkred")  
rug(c(-tact, tact), ticksize = -0.0451, lwd = 2, col = "darkgreen")



신뢰구간

# set seed for reproducibility  
set.seed(4)  
  
# generate and plot the sample data  
Y <- rnorm(n = 100,   
 mean = 5,   
 sd = 5)  
  
plot(Y,   
 pch = 19,   
 col = "steelblue")



* 신뢰구간 하한, 상한 구하기

cbind(CIlower = mean(Y) - 1.96 \* 5 / 10, CIupper = mean(Y) + 1.96 \* 5 / 10)

## CIlower CIupper  
## [1,] 4.502625 6.462625

# set seed  
set.seed(1)  
  
# initialize vectors of lower and upper interval boundaries  
lower <- numeric(10000)  
upper <- numeric(10000)  
  
# loop sampling / estimation / CI  
for(i in 1:10000) {  
   
 Y <- rnorm(100, mean = 5, sd = 5)  
 lower[i] <- mean(Y) - 1.96 \* 5 / 10  
 upper[i] <- mean(Y) + 1.96 \* 5 / 10  
   
}  
  
# join vectors of interval bounds in a matrix  
CIs <- cbind(lower, upper)

mean(CIs[, 1] <= 5 & 5 <= CIs[, 2])

## [1] 0.9487

가변수 만들기

CASchools$D <- CASchools$STR < 20

가변수를 이용한 회귀분석

dummy\_model <- lm(score ~ D, data = CASchools)  
summary(dummy\_model)

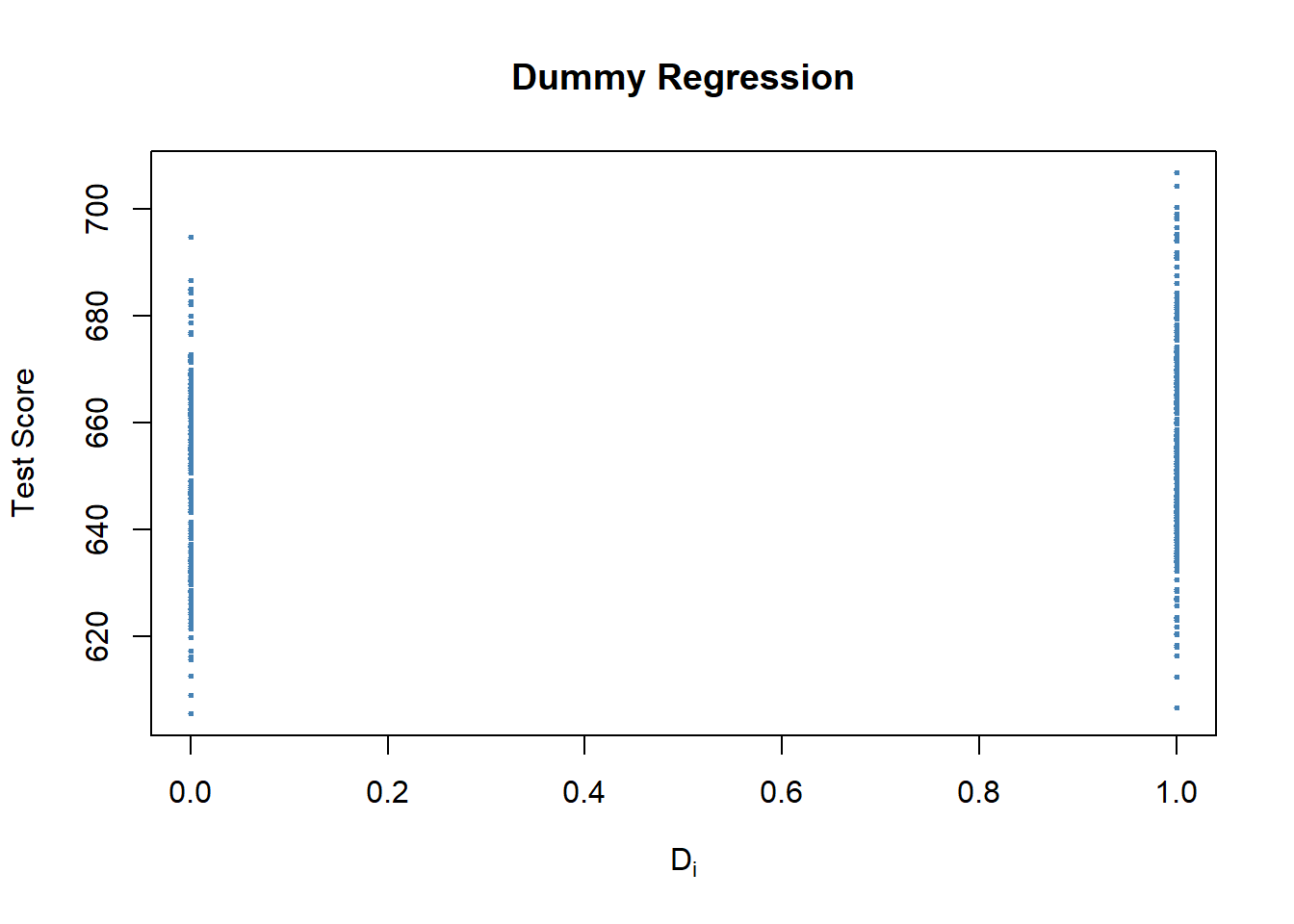
##   
## Call:  
## lm(formula = score ~ D, data = CASchools)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -50.496 -14.029 -0.346 12.884 49.504   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 650.077 1.393 466.666 < 2e-16 \*\*\*  
## DTRUE 7.169 1.847 3.882 0.00012 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 18.74 on 418 degrees of freedom  
## Multiple R-squared: 0.0348, Adjusted R-squared: 0.0325   
## F-statistic: 15.07 on 1 and 418 DF, p-value: 0.0001202

가변수 회귀분석에 대한 신뢰구간

confint(dummy\_model)

## 2.5 % 97.5 %  
## (Intercept) 647.338594 652.81500  
## DTRUE 3.539562 10.79931

# Plot the data  
plot(CASchools$D, CASchools$score, # provide the data to be plotted  
 pch = 20, # use filled circles as plot symbols  
 cex = 0.5, # set size of plot symbols to 0.5  
 col = "Steelblue", # set the symbols' color to "Steelblue"  
 xlab = expression(D[i]), # Set title and axis names  
 ylab = "Test Score",  
 main = "Dummy Regression")

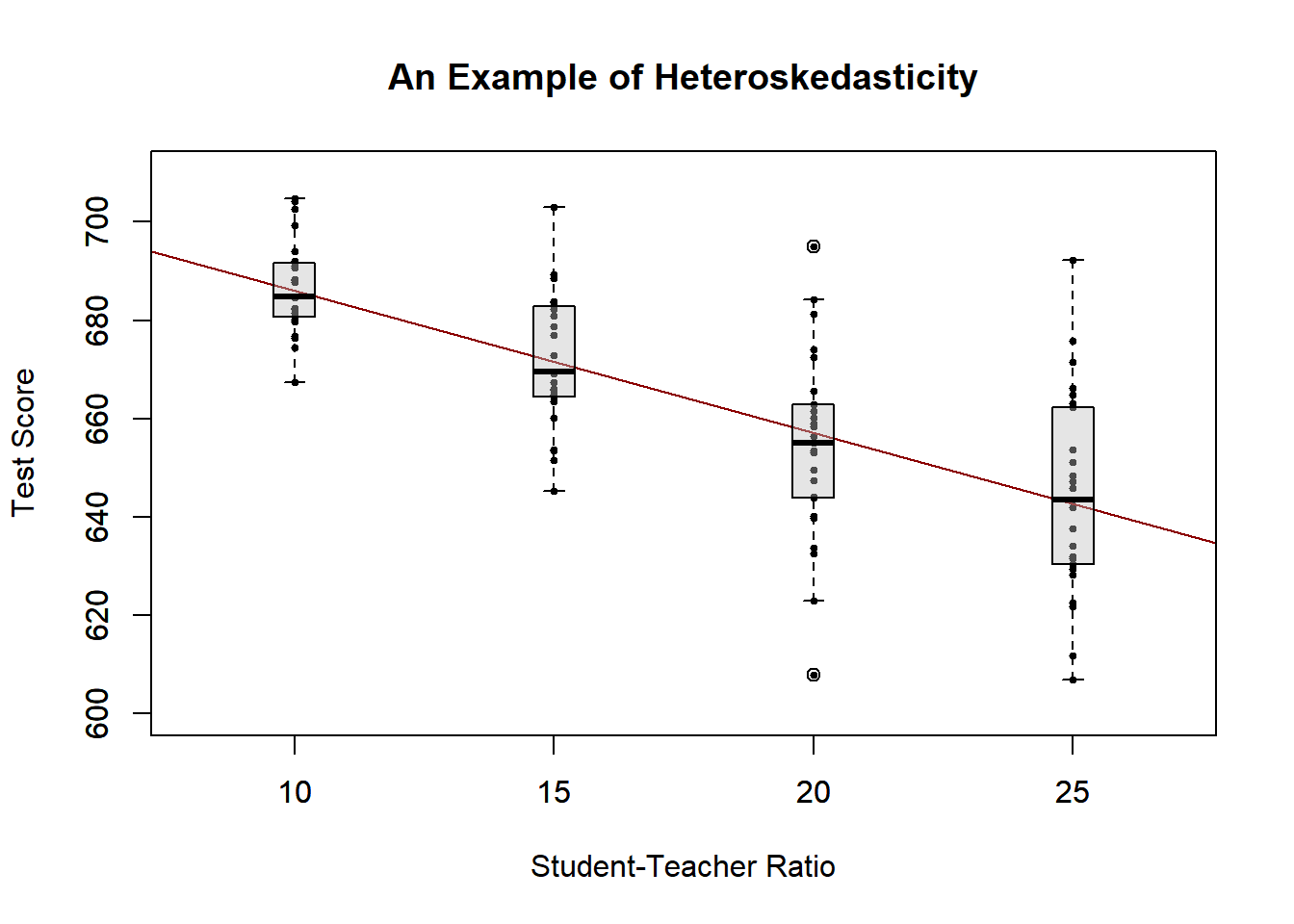


5.2 회귀모형의 진단과 수정

5.2.1 부분 F검정

* 이분산성(Heteroskedasticity) 이분산성은, 동분산성의 가정을 충족시키지 못하는 경우를 말한다. 동분산성의 가정이란, 회귀분석에서 오차항들의 분산이 일정하다는 가정이다

library(scales)  
  
set.seed(123)   
# set up vector of x coordinates  
x <- rep(c(10, 15, 20, 25), each = 25)  
  
# initialize vector of errors  
e <- c()  
  
# sample 100 errors such that the variance increases with x  
e[1:25] <- rnorm(25, sd = 10)  
e[26:50] <- rnorm(25, sd = 15)  
e[51:75] <- rnorm(25, sd = 20)  
e[76:100] <- rnorm(25, sd = 25)  
  
# set up y  
y <- 720 - 3.3 \* x + e  
  
# Estimate the model   
mod <- lm(y ~ x)  
  
# Plot the data  
plot(x = x,   
 y = y,   
 main = "An Example of Heteroskedasticity",  
 xlab = "Student-Teacher Ratio",  
 ylab = "Test Score",  
 cex = 0.5,   
 pch = 19,   
 xlim = c(8, 27),   
 ylim = c(600, 710))  
  
# Add the regression line to the plot  
abline(mod, col = "darkred")  
  
# Add boxplots to the plot  
boxplot(formula = y ~ x,   
 add = TRUE,   
 at = c(10, 15, 20, 25),   
 col = alpha("gray", 0.4),   
 border = "black")



A Real-World Example for Heteroskedasticity

* 데이터에 summary를 할경우 각각의 기초통계량 제공

library(AER)  
data("CPSSWEducation")  
attach(CPSSWEducation)  
  
summary(CPSSWEducation)

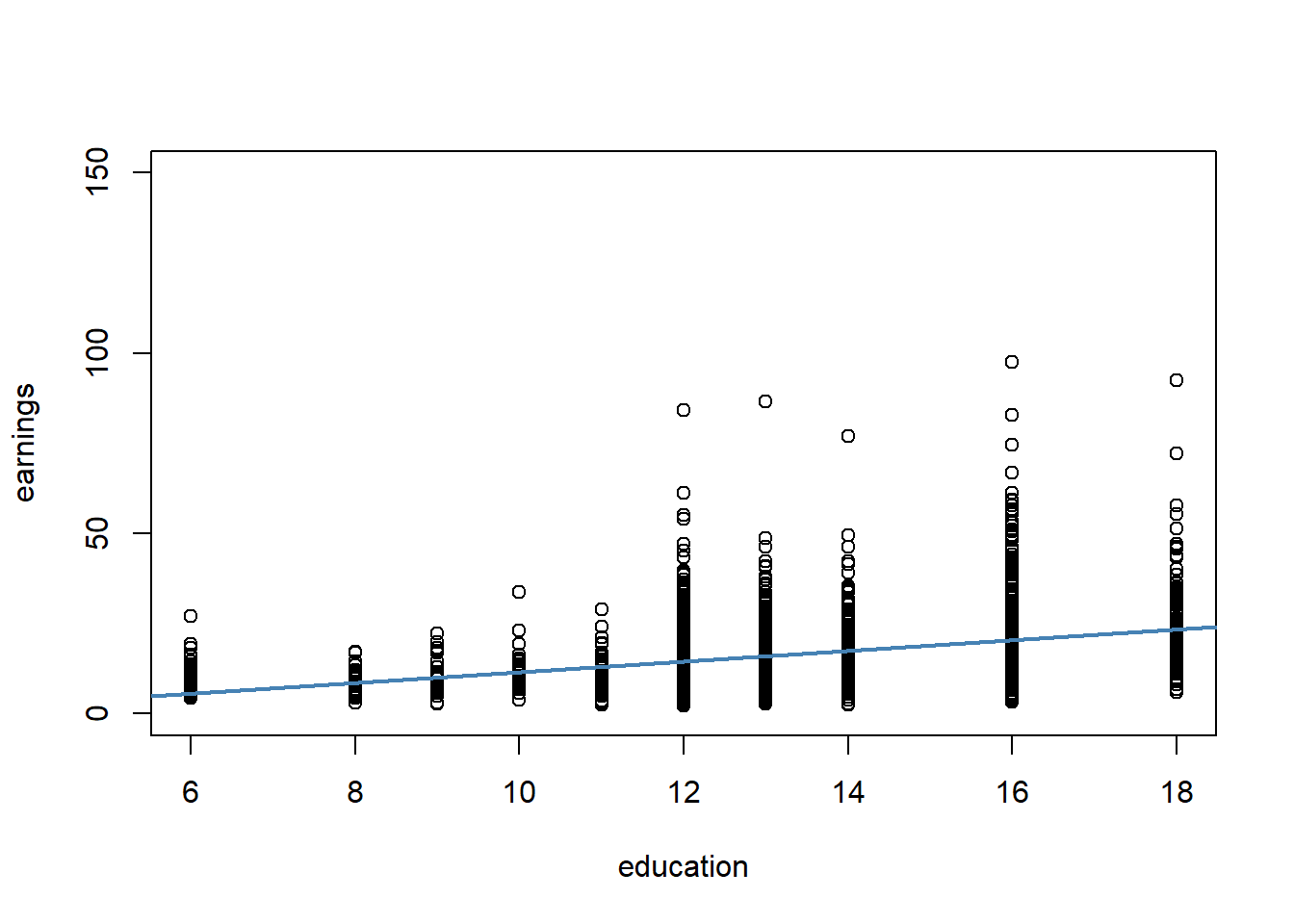
## age gender earnings education   
## Min. :29.0 female:1202 Min. : 2.137 Min. : 6.00   
## 1st Qu.:29.0 male :1748 1st Qu.:10.577 1st Qu.:12.00   
## Median :29.0 Median :14.615 Median :13.00   
## Mean :29.5 Mean :16.743 Mean :13.55   
## 3rd Qu.:30.0 3rd Qu.:20.192 3rd Qu.:16.00   
## Max. :30.0 Max. :97.500 Max. :18.00

* 그래프를 통해 이분산성 확인
  + summary() : F-statistic: 441.9 on 1 and 2948 DF, p-value: < 2.2e-16 이므로 이분상성을 띄고있다

# estimate a simple regression model  
labor\_model <- lm(earnings ~ education)  
summary(labor\_model)

##   
## Call:  
## lm(formula = earnings ~ education)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -17.270 -5.355 -1.513 3.194 77.164   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -3.13437 0.95925 -3.268 0.0011 \*\*   
## education 1.46693 0.06978 21.021 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.769 on 2948 degrees of freedom  
## Multiple R-squared: 0.1304, Adjusted R-squared: 0.1301   
## F-statistic: 441.9 on 1 and 2948 DF, p-value: < 2.2e-16

# plot observations and add the regression line  
plot(education,   
 earnings,   
 ylim = c(0, 150))  
  
abline(labor\_model,   
 col = "steelblue",   
 lwd = 2)



단순회귀 분석에서의 추론

* earnings = -3.134 + 1.467\*education
  + education이 한 단위 증가 함에 따라 earnings은 1.467만큼 증가한다

# print the contents of labor\_model to the console  
labor\_model

##   
## Call:  
## lm(formula = earnings ~ education)  
##   
## Coefficients:  
## (Intercept) education   
## -3.134 1.467

95% 신뢰구간

* 절편과 회귀계수에 대한 상한,하한 95% 신뢰구간

confint(labor\_model)

## 2.5 % 97.5 %  
## (Intercept) -5.015248 -1.253495  
## education 1.330098 1.603753

5.2.2 변수변환

library(AER)  
library(stargazer)

##   
## Please cite as:

## Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

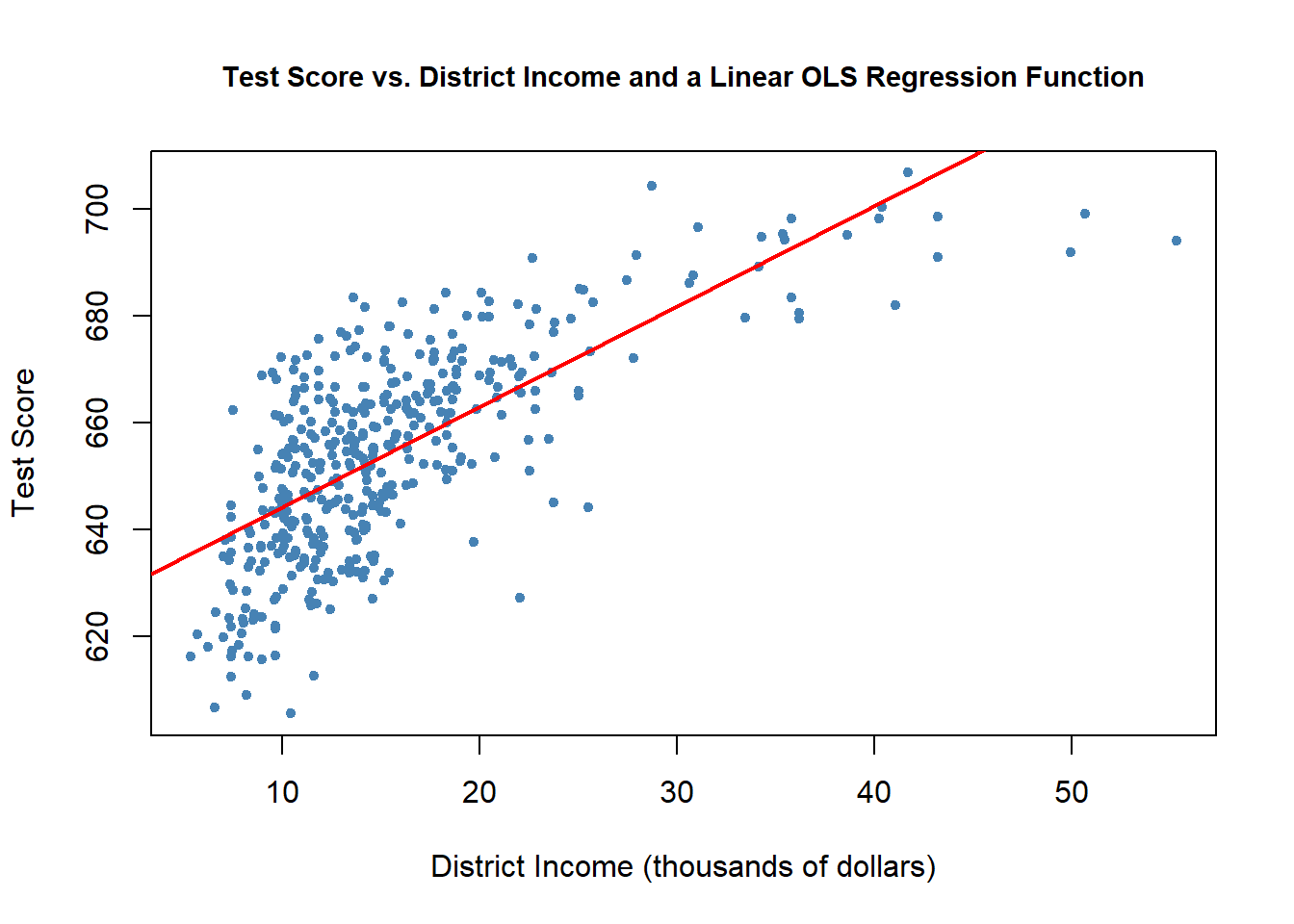
## R package version 5.2.2. https://CRAN.R-project.org/package=stargazer

# prepare the data  
library(AER)   
data(CASchools)  
CASchools$size <- CASchools$students/CASchools$teachers  
CASchools$score <- (CASchools$read + CASchools$math) / 2

cor(CASchools$income, CASchools$score)

## [1] 0.7124308

# A General Strategy for Modelling Nonlinear Regression Functions  
# fit a simple linear model  
linear\_model<- lm(score ~ income, data = CASchools)  
  
# plot the observations  
plot(CASchools$income, CASchools$score,  
 col = "steelblue",  
 pch = 20,  
 xlab = "District Income (thousands of dollars)",   
 ylab = "Test Score",  
 cex.main = 0.9,  
 main = "Test Score vs. District Income and a Linear OLS Regression Function")  
  
# add the regression line to the plot  
abline(linear\_model,   
 col = "red",   
 lwd = 2)

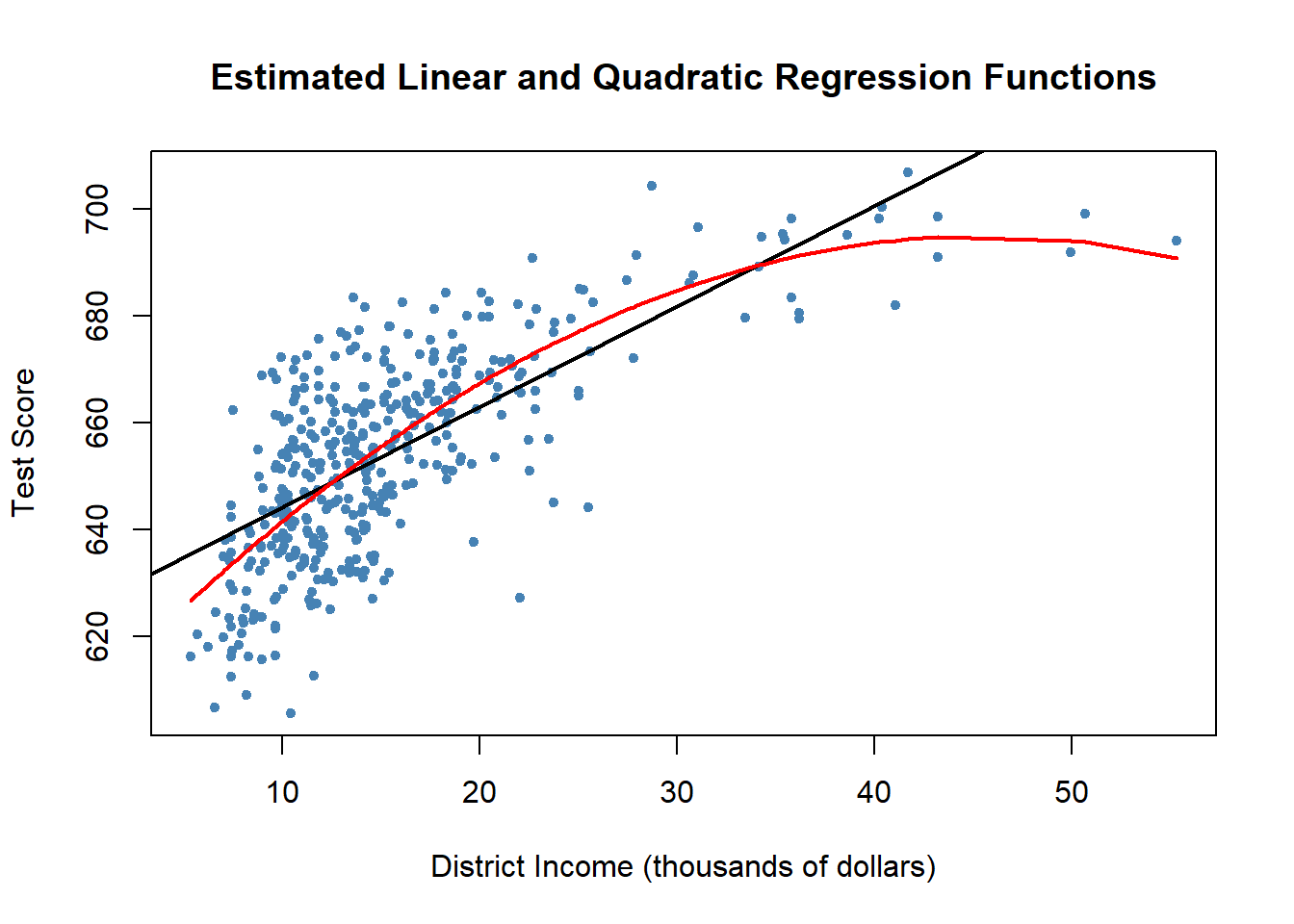


* income에 제곱항을 넣어 비선형모형 확인

# fit the quadratic Model  
quadratic\_model <- lm(score ~ income + I(income^2), data = CASchools)  
  
# obtain the model summary  
coeftest(quadratic\_model, vcov. = vcovHC, type = "HC1")

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 607.3017435 2.9017544 209.2878 < 2.2e-16 \*\*\*  
## income 3.8509939 0.2680942 14.3643 < 2.2e-16 \*\*\*  
## I(income^2) -0.0423084 0.0047803 -8.8505 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# draw a scatterplot of the observations for income and test score  
plot(CASchools$income, CASchools$score,  
 col = "steelblue",  
 pch = 20,  
 xlab = "District Income (thousands of dollars)",  
 ylab = "Test Score",  
 main = "Estimated Linear and Quadratic Regression Functions")  
  
# add a linear function to the plot  
abline(linear\_model, col = "black", lwd = 2)  
  
# add quatratic function to the plot  
order\_id <- order(CASchools$income)  
  
lines(x = CASchools$income[order\_id],   
 y = fitted(quadratic\_model)[order\_id],  
 col = "red",   
 lwd = 2)

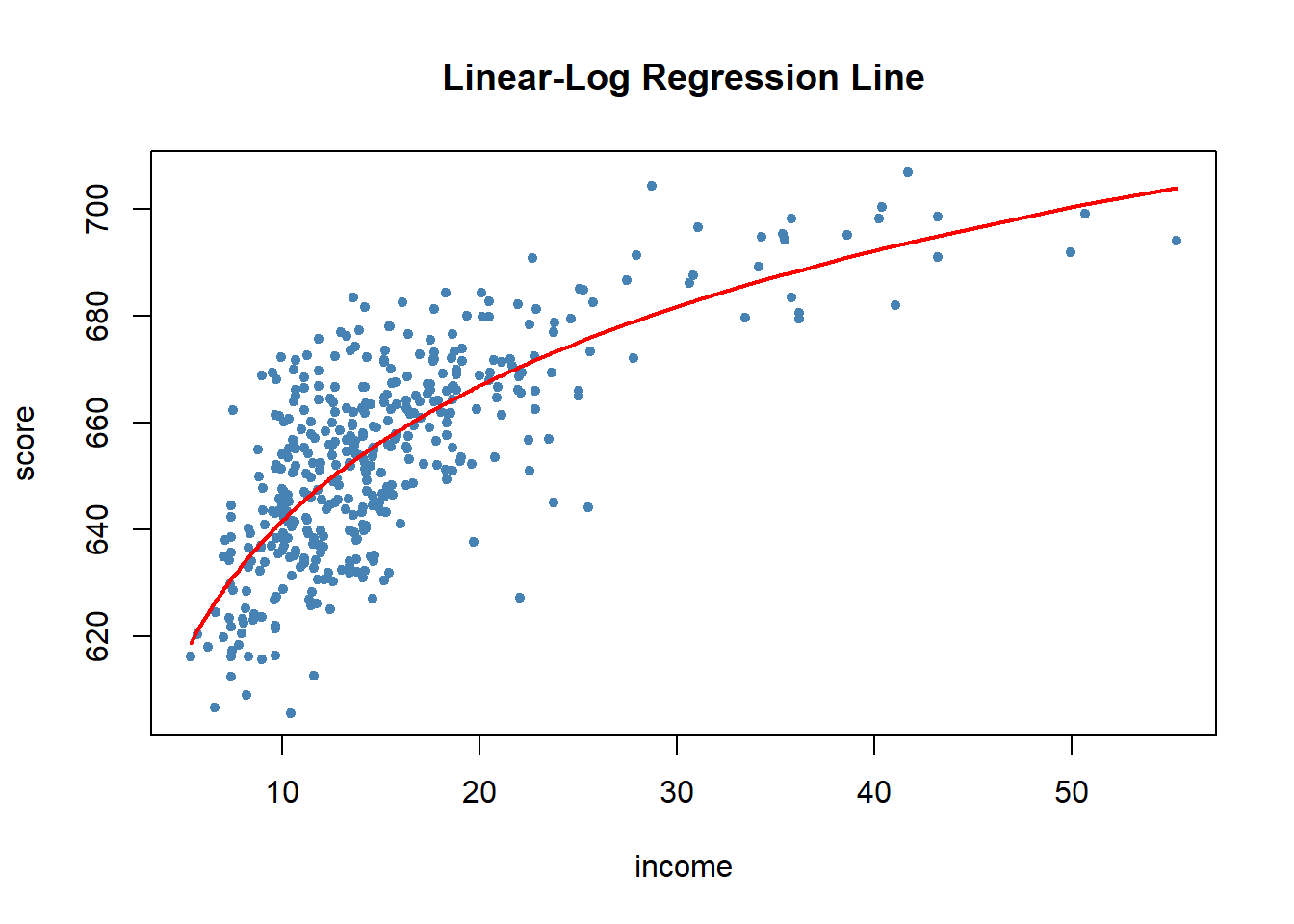


Logarithms

# estimate a level-log model  
LinearLog\_model <- lm(score ~ log(income), data = CASchools)  
  
# compute robust summary  
coeftest(LinearLog\_model,   
 vcov = vcovHC, type = "HC1")

##   
## t test of coefficients:  
##   
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 557.8323 3.8399 145.271 < 2.2e-16 \*\*\*  
## log(income) 36.4197 1.3969 26.071 < 2.2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

# draw a scatterplot  
plot(score ~ income,   
 col = "steelblue",  
 pch = 20,  
 data = CASchools,  
 main = "Linear-Log Regression Line")  
  
# add the linear-log regression line  
order\_id <- order(CASchools$income)  
  
lines(CASchools$income[order\_id],  
 fitted(LinearLog\_model)[order\_id],   
 col = "red",   
 lwd = 2)



5.2.3 표준화 회귀계수

* 단위가 다를 경우 한단위 증가에 따른 표준화 된 값의 증가분을 알고싶을 때 scale()함수를 사용
  + scale()을 사용할 경우 Estimate의 값이 클 수록 한단위 증가분에 대한 증가량이 크다는 것을 알 수 있다

w = c(65,83,72,45,42,56,89,72,47,57,55)  
h = c(160, 188,180,154,153,160,185,182,163,171,167)  
age = c(12,24,27,11,14,15,19,30,15,17,16)  
  
df = data.frame(age,h,w,stringsAsFactors = F)  
  
l1 = lm(h~ age+w,df)  
summary(l1)

##   
## Call:  
## lm(formula = h ~ age + w, data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -5.4506 -2.6068 -0.4169 3.0404 5.2536   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 121.9604 5.1613 23.63 1.09e-08 \*\*\*  
## age 0.8721 0.2642 3.30 0.01085 \*   
## w 0.5081 0.1059 4.80 0.00136 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.943 on 8 degrees of freedom  
## Multiple R-squared: 0.9219, Adjusted R-squared: 0.9024   
## F-statistic: 47.22 on 2 and 8 DF, p-value: 3.72e-05

l2 = lm(scale(h)~ scale(age)+scale(w),df)  
summary(l2)

##   
## Call:  
## lm(formula = scale(h) ~ scale(age) + scale(w), data = df)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -0.43192 -0.20657 -0.03304 0.24093 0.41630   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -6.457e-16 9.421e-02 0.0 1.00000   
## scale(age) 4.291e-01 1.300e-01 3.3 0.01085 \*   
## scale(w) 6.241e-01 1.300e-01 4.8 0.00136 \*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 0.3124 on 8 degrees of freedom  
## Multiple R-squared: 0.9219, Adjusted R-squared: 0.9024   
## F-statistic: 47.22 on 2 and 8 DF, p-value: 3.72e-05

5.2.4 다중공선성의 탐색

* 다중공선성, 통계학의 회귀분석에서 독립변수들 간에 강한 상관관계가 나타나는 문제이다. 독립변수들간에 정확한 선형관계가 존재하는 완전공선성의 경우와 독립변수들간에 높은 선형관계가 존재하는 다중공선성으로 구분하기도 한다. 이는 회귀분석의 전제 가정을 위배하는 것이므로 적절한 회귀분석을 위해 해결해야 하는 문제가 된다
* 다중공선성 문제가 있다면 독립변수와 종속변수의 관계가 양의 상관관계(+)에서 음의상관관계(-)로 잘못 판단 되어지는 경우가 있다
* 해결법
  + 상관관계가 높은 독립변수중 하나 혹은 일부를 제거
  + 변수를 변형시키거나 새로운 관측치를 이용
  + 자료를 수집하는 현장의 상황을 보아 상관관계의 이유를 파악하여 해결
* 판단
  + vif : 분산팽창인자가 10상일 경우 다중공선성이 존재한다고 판단한다

다중공선성 판단

* 위 데이터 프레임에 대한 다중공선성 확인
  + vif값이 10보다 작으므로 다중공선성은 존재하지 않는다
  + 만약 vif값이 10보다 큰 변수가 있다면 그 변수를 제거 한 후 회귀모델을 다시 적용한다

library(car) # vif 함수 사용하기 위한 라이브러리  
vif(l1)

## age w   
## 1.732052 1.732052