응용R 팀프로젝트

Team B

2021-06-29

##Francis Galton 은 인간 특성의 변이와 유전을 연구했습니다. 그중 Galton은 유전을 이해하기 위해 가족으로부터 키 데이터를 수집하고 이 과정에서 그는 상관관계와 회귀의 개념, 정규 분포를 따르는 데이터 쌍에 대한 관계를 연구했습니다. 물론, 이 데이터가 수집되었을 당시 우리의 유전학 지식은 오늘날 우리가 알고 있는 것에 비해 상당히 제한적이었지만 Galton은 이 데이터로부터 "부모와 자식 신장 사이에는 선형적인 관계가 있고 신장이 커지거나 작아지는 것보다는 전체 신장 평균으로 회귀하는 경향이 있다"라는 가설을 세웠습니다.또한 이러한 가설로써 대답하고자 했던 부분은 "부모의 키를 기준으로 아이의 키를 얼마나 잘 예측할 수 있습니까?" 입니다.

```
library(HistData)
library(knitr)
#library(magrittr)
library(ggplot2)
library(gridExtra)
library(carData)
library(car)
library(jpeg)
library(png)
library(plotrix)
library(rasterImage)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following object is masked from 'package:car':
##
##     recode
```

```
## The following object is masked from 'package:gridExtra':
##
##     combine
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.0 --
```

```
## √ tibble  3.0.4     √ purrr   0.3.4
## √ tidyr   1.1.2     √ stringr 1.4.0
## √ readr   1.4.0     √ forcats 0.5.0
```

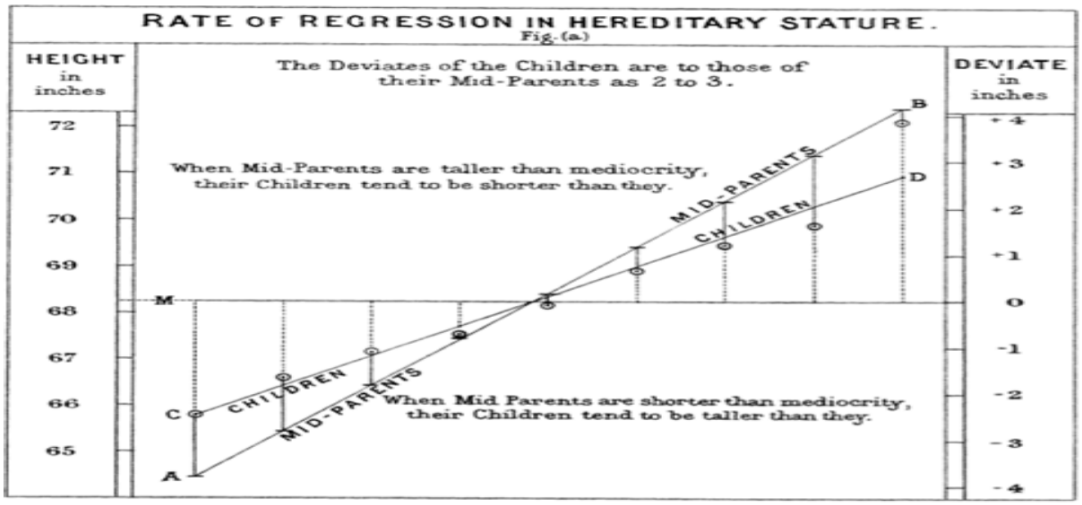
```
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::combine() masks gridExtra::combine()
## x dplyr::filter()  masks stats::filter()
## x dplyr::lag()     masks stats::lag()
## x dplyr::recode()  masks car::recode()
## x purrr::some()    masks car::some()
```

```
data("GaltonFamilies")
str(GaltonFamilies)
```

```
## 'data.frame':    934 obs. of  8 variables:
##  $ family         : Factor w/ 205 levels "001","002","003",..: 1 1 1 1 2 2 2 2 3 3 ...
##  $ father         : num  78.5 78.5 78.5 78.5 75.5 75.5 75.5 75.5 75 75 ...
##  $ mother         : num  67 67 67 67 66.5 66.5 66.5 66.5 64 64 ...
##  $ midparentHeight: num  75.4 75.4 75.4 75.4 73.7 ...
##  $ children       : int  4 4 4 4 4 4 4 4 2 2 ...
##  $ childNum       : int  1 2 3 4 1 2 3 4 1 2 ...
##  $ gender         : Factor w/ 2 levels "female","male": 2 1 1 1 2 2 1 1 2 1 ...
##  $ childHeight    : num  73.2 69.2 69 69 73.5 72.5 65.5 65.5 71 68 ...
```

```
#프란시스 골턴 사진
francisgalton = "https://cdn.britannica.com/s:290x800/13/11613-004-33F53EAF/Francis-Galton-detail-oil-painting-G-Graef-1882.jpg"
download.file(francisgalton, 'francisgalton.jpg', mode = 'wb')
francisgalton <- readJPEG("francisgalton.jpg", native=TRUE)
plot(0:1,0:1,type="n", ann=FALSE, axes=FALSE)
rasterImage(francisgalton,0,0,1,1)
```

```
#프란시스 골턴 연구(우리가 비슷하게 그려보고자 하는 그래프)
galtonstudy = "https://curranbauer.org/wp-content/uploads/2017/05/reg-to-mean-from-galton2.png"
download.file(galtonstudy,'galtonstudy.png', mode = 'wb')
Galtonstudy <- readPNG("galtonstudy.png", native=TRUE)
plot(0:1,0:1,type="n", ann=FALSE, axes=FALSE)
rasterImage(Galtonstudy,0,0,1,1)
```



```
#JPG파일 주소 : "https://www.researchgate.net/profile/Yeming_Ma2/publication/280970132/figure/fig1/AS:284517131669510@1444845578444/Rate-of-regression-in-hereditary-stature-Galton-1886-Plate-IX-fig-a-The-short_Q320.jpg"
```

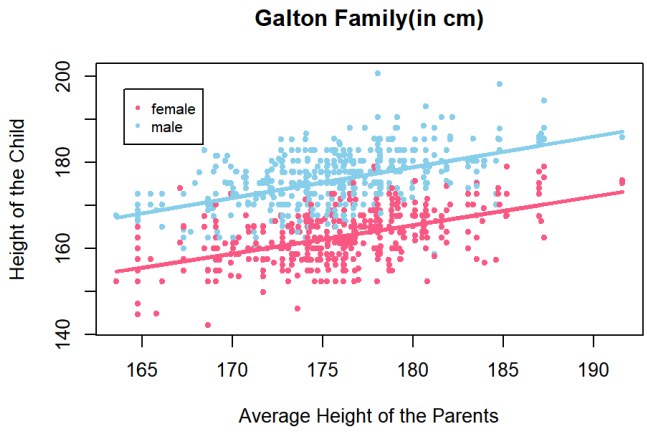
#데이터 정리

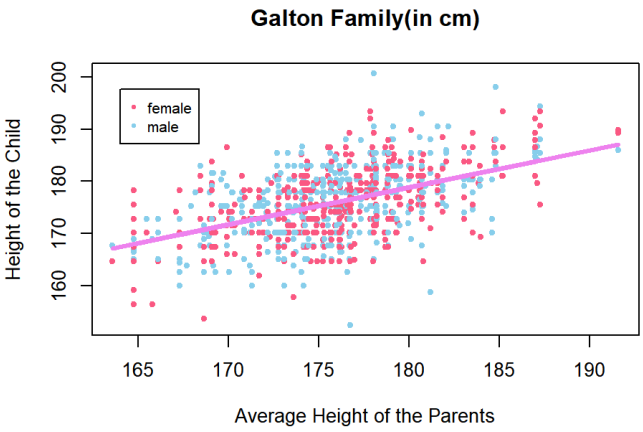
```
#의미있는 데이터만 뽑아내기
fam <- select(GaltonFamilies, father, mother, midparentHeight, gender, childHeight)
#fam <- GaltonFamilies[, c(2, 3, 4, 7, 8)]

#inch 를 cm로 바꾸기
f <- data.frame(round(fam[,-4]*2.54, 2), fam[, "gender"])
colnames(f) <- c("Father", "Mother", "MidParent", "Child", "Childtype")

#MidParent = (Father + 1.08 * Mother) / 2

#mean(fm$Child) / mean(fw$Child) # 아들 키의 평균 / 딸 키의 평균
#1.080
#mean(f$Father) / mean(f$Mother) # 아빠 키의 평균 / 엄마 키의 평균
#1.079
#mean(c(fm$Father, fm$Child)) / mean(c(fw$Mother, fw$Child)) # 아들과 아빠 키의 평균 / 딸과 엄마키의 평균
#1.078

#즉 갈톤 패밀리의 남자와 여자의 키차이는 1.08 배임을 알수있다.
#그러므로 남자와 여자의 키차이를 보완하기위해 Midparent에 엄마키* 1.08인 데이터가 추가된것이다.

#딸의 키에 1.08을 곱한 데이터를 추가
f <- mutate(f, newChild = round(ifelse(f$Childtype == "female",1.08,1) * f$Child, 2))
#f <- data.frame(f, "newChild" = round(ifelse(f$Childtype == "female",1.08,1) * f$Child, 2))

f <- as_tibble(select(f, Father, Mother, MidParent, Child, newChild, Childtype))

#전체 데이터를 아들과 딸의 데이터로 나누기(그래프 그릴때 변수 불러오기를 편하게 하기 위한 작업)
fm <- subset(f, Childtype == "male")
fw <- subset(f, Childtype == "female")

f
```

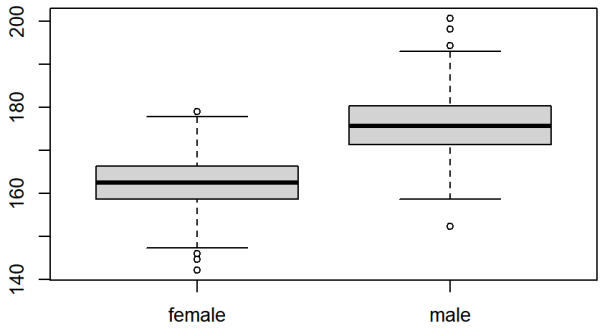
```
## # A tibble: 934 x 6
##    Father Mother MidParent Child newChild Childtype
##     <dbl>  <dbl>     <dbl> <dbl>    <dbl> <fct>
##  1   199.   170.      192.  186.     186. male
##  2   199.   170.      192.  176.     190. female
##  3   199.   170.      192.  175.     189. female
##  4   199.   170.      192.  175.     189. female
##  5   192.   169.      187.  187.     187. male
##  6   192.   169.      187.  184.     184. male
##  7   192.   169.      187.  166.     180. female
##  8   192.   169.      187.  166.     180. female
##  9   190.   163.      183.  180.     180. male
## 10   190.   163.      183.  173.     187. female
## # ... with 924 more rows
```

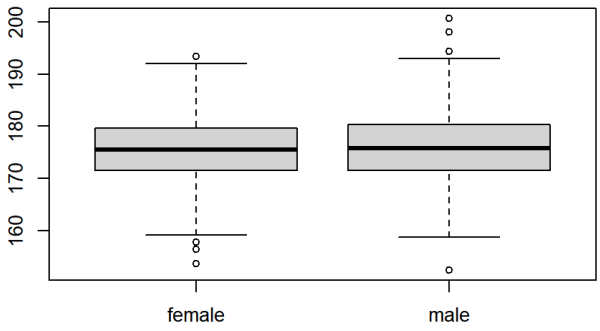
#1.08을 곱한 이유 (시각적 분석) (MidParent = (Father + 1.08 * Mather) / 2) (plot)

```
par(mfrow=c(2,2))

plot(jitter(f$Child) ~ f$MidParent,
     xlab = "Average Height of the Parents",
     ylab = "Height of the Child",
     main = "Galton Family(in cm)",
     pch = 20,
     col = ifelse(f$Childtype == "female", "#FA5882", "skyblue"))
     legend(164, 197, pch = c(20,20), col=c("#FA5882","skyblue"), c("female", "male"), cex = 0.
8)
     lines(fm$MidParent, fitted(lm(Child ~ MidParent, data = fm)), col="skyblue",lwd=2.5)
     lines(fw$MidParent, fitted(lm(Child ~ MidParent, data = fw)), col="#FA5882",lwd=2.5)

plot(jitter(f$newChild) ~ f$MidParent,
     xlab = "Average Height of the Parents",
     ylab = "Height of the Child",
     main = "Galton Family(in cm)",
     pch = 20,
     col = ifelse(f$Childtype == "female", "#FA5882", "skyblue"))
     legend(164, 197.7, pch = c(20,20), col=c("#FA5882","skyblue"), c("female", "male"), cex =
0.8)
     lines(f$MidParent, fitted(lm(newChild ~ MidParent, data = f)), col="violet",lwd=3)

boxplot(f$Child ~ f$Childtype,
        xlab = "",
        ylab = "")

boxplot(f$newChild ~ f$Childtype,
        xlab = "",
        ylab = "")
```





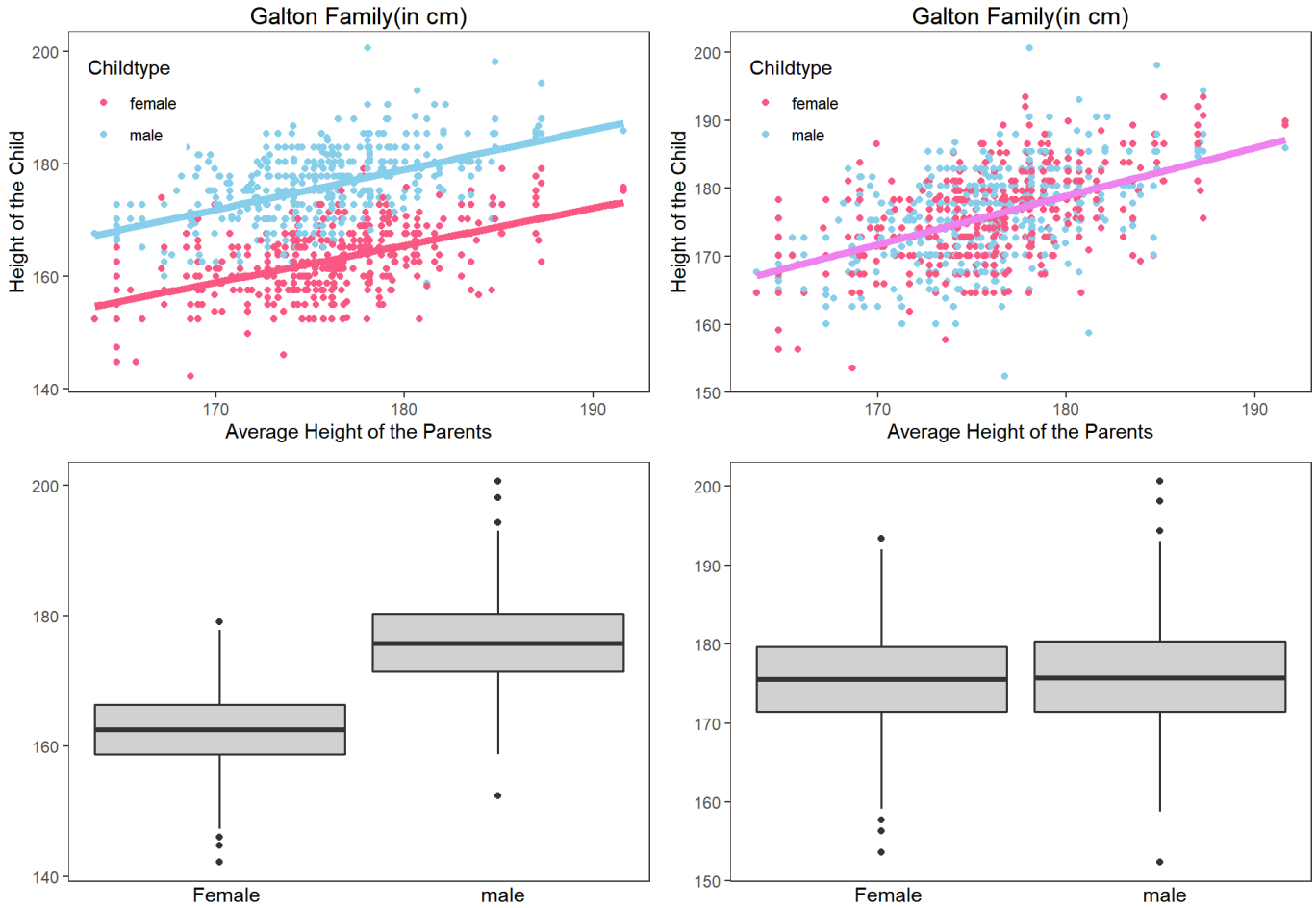




#1.08을 곱한 이유 (시각적 분석) (MidParent = (Father + 1.08 * Mather) / 2) (ggplot)

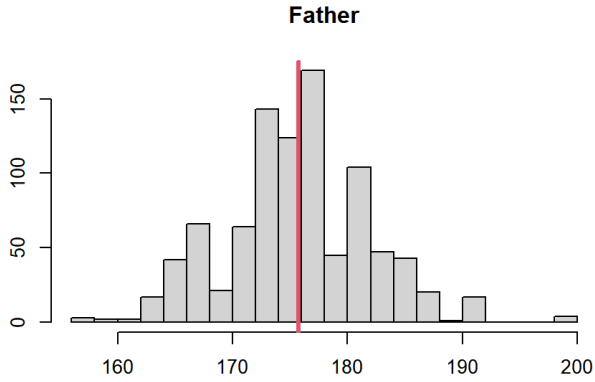
```
aa <- ggplot(data = f, aes(x = MidParent, y = Child, color = Childtype)) +
        geom_point() +
        geom_smooth(method = "lm", se = FALSE , lwd = 2, show.legend = F) +
        scale_colour_manual(values=c("#FA5882","skyblue")) +
        theme_bw() +
        theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
        scale_x_continuous(name = "Average Height of the Parents") +
        scale_y_continuous(name = "Height of the Child") +
        labs(title ="Galton Family(in cm)") +
        theme(legend.position=c(0.11, 0.8),
              plot.title = element_text(hjust = 0.5,
                                        vjust = -0.5,
                                        size = 13))

bb <- ggplot(data = f, aes(x = MidParent, y = newChild, color = Childtype)) +
        geom_point() +
        geom_smooth(method = "lm", se = FALSE, col = "violet", lwd = 2 ) +
        scale_colour_manual(values=c("#FA5882","skyblue")) +
        theme_bw() +
        theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
        scale_x_continuous(name = "Average Height of the Parents") +
        scale_y_continuous(name = "Height of the Child") +
        labs(title ="Galton Family(in cm)") +
        theme(legend.position=c(0.11, 0.8),
              plot.title = element_text(hjust = 0.5,
                                        vjust = -0.5,
                                        size = 13))

aaa <- ggplot(data = f, aes(x = MidParent, y = Child, group = Childtype)) +
        geom_boxplot(fill = "gray82", lwd = 0.7) +
        theme_bw() +
        theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
        scale_x_continuous(name = "Female                                        male", breaks
 = F) +
        scale_y_continuous(name = "")

bbb <- ggplot(data = f, aes(x = MidParent, y = newChild, group = Childtype)) +
        geom_boxplot(fill = "gray82", lwd = 0.7) +
        theme_bw() +
        theme(panel.grid.major = element_blank(), panel.grid.minor = element_blank()) +
        scale_x_continuous(name = "Female                                        male", breaks
 = F) +
        scale_y_continuous(name = "")

grid.arrange(aa, bb, aaa, bbb, ncol = 2, nrow = 2)
```

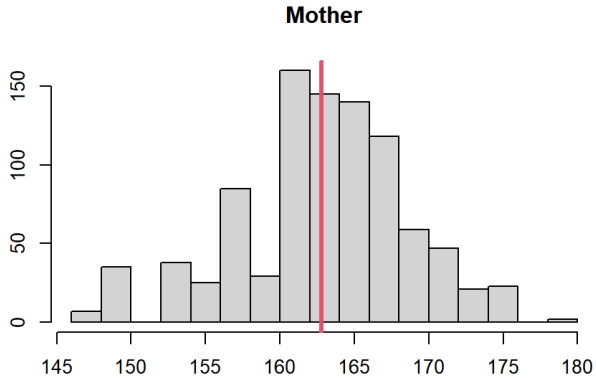
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

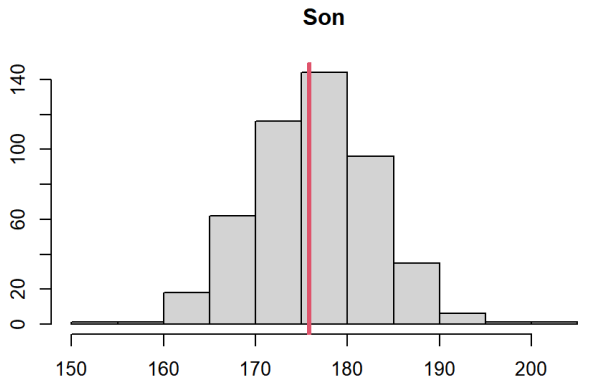


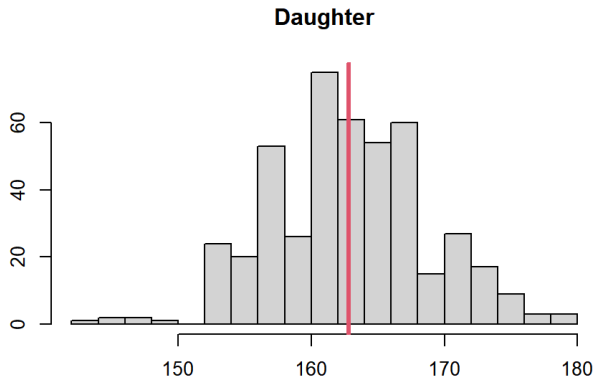
#히스토그램

```
par(mfrow=c(2,2))

hist(f$Father, main = "Father", breaks = 15, xlab = "", ylab = "")
abline(v=mean(f$Father), col=2, lwd=3)

hist(f$Mother, main = "Mother", breaks = 15, xlab = "", ylab = "")
abline(v=mean(f$Mother), col=2, lwd=3)

hist(fm$Child, main = "Son", breaks = 15, xlab = "", ylab = "")
abline(v=mean(fm$Child), col=2, lwd=3)

hist(fw$Child, main = "Daughter", breaks = 15, xlab = "", ylab = "")
abline(v=mean(fw$Child), col=2, lwd=3)
```





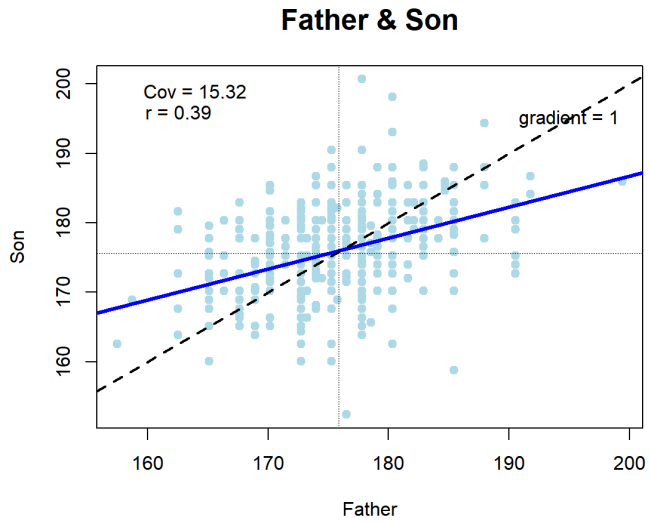


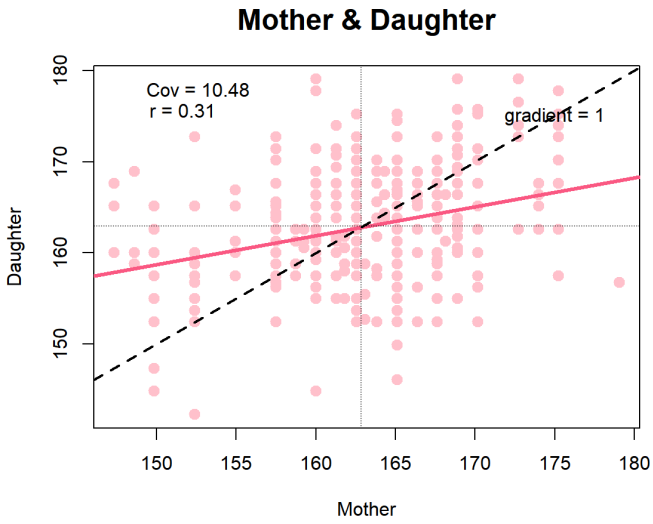



```
#중앙에 빨간선은 평균선을 의미하다.
```

#부모와 자식간의 산점도 1 (아빠, 아들), (엄마, 딸)

```
grad <- "gradient = 1"

cov_m <- cov(fm$Child, fm$Father)
cov_w <- cov(fw$Child, fw$Mother)
cor_m <- cor(fm$Child, fm$Father)
cor_w <- cor(fw$Child, fw$Mother)

par(mfrow=c(1,2))

plot(fm$Child ~ fm$Father, pch = 19, col = "light blue", xlab = "Father", ylab = "Son")
abline(lm(fm$Child ~ fm$Father, data = fm), col = "blue", lwd = 3)
abline(a = 0, b = 1, col="black", lty = 2, lwd =2)
abline(h = mean(fm$Father), v = mean(fm$Child), lwd = 0.1, lty = 3)
text(x = 164, y = 199, labels = paste("Cov =", round(cov_m, digits = 2)))
text(x = 162.6, y = 196, labels = paste("r =", round(cor_m, digits = 2)))
text(x = 195, y = 195, labels = grad)
title(main = "Father & Son", cex.main = 1.5)
### 축설정 못하겠음 axis(side = 1, at = c(160, 200))

plot(fw$Child ~ fw$Mother, pch = 19, col = "pink", lwd = 3, xlab = "Mother", ylab = "Daughter")
abline(lm(fw$Child ~ fw$Mother, data = fw), col = "#FA5882", lwd = 3)
abline(a = 0, b = 1, col="black", lty = 2, lwd =2)
abline(h = mean(fw$Mother), v = mean(fw$Child), lwd = 0.1, lty = 3)
text(x = 152.6, y = 178, labels = paste("Cov =", round(cov_w, digits = 2)))
text(x = 151.65, y = 175.7, labels = paste("r =", round(cor_w, digits = 2)))
text(x = 175, y = 175, labels = grad)
title(main = "Mother & Daughter", cex.main = 1.5)
```

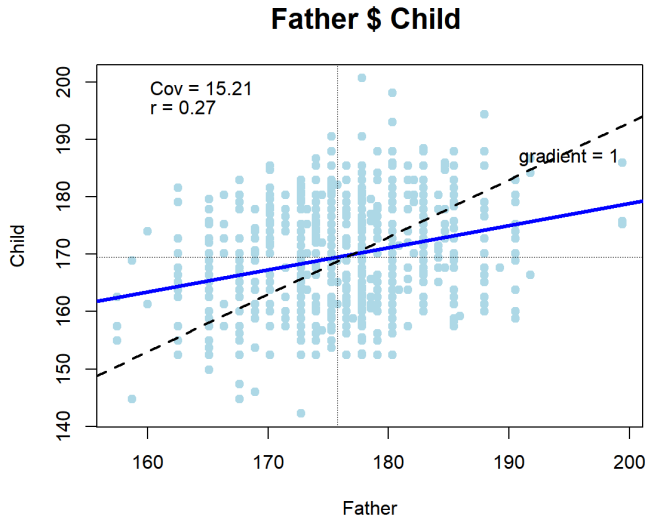


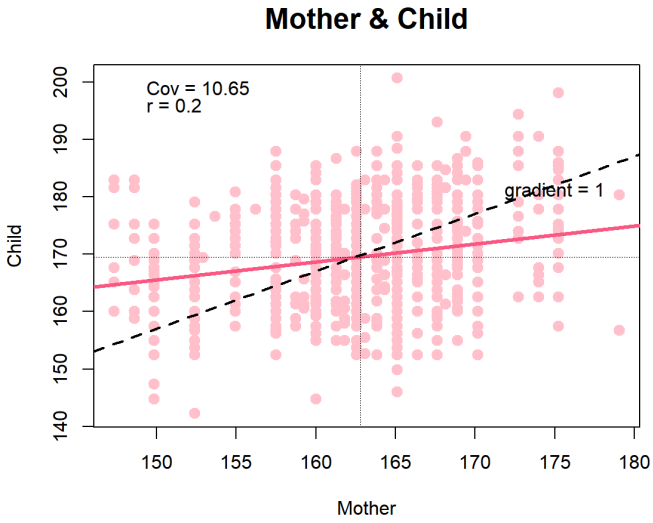


#아빠와 아들, 엄마와 딸의 산점도를 비교해본 결과 두 그래프 모두 공분산이 양수가 나와 아빠와 엄마의 키가 증가할때 아들과 딸의 키가 증가함을 보인다. 또한 상관관계는 그리 높지않은걸로 보이지만 우리가 보여주고싶은것은 선형선의 기울기가 1보다 작아 평균으로 회귀함 이다. 두그래프의 선형선의 기울기는 1보다 작아 자식들의 키(신장)이 평균으로 회귀하는것이 시각적으로 확인된다.

#부모와 자식간의 산점도 2 (아빠,자식) (엄마,자식)

```
cov_fc <- cov(f$Father, f$Child)
cov_mc <- cov(f$Mother, f$Child)
cor_fc <- cor(f$Father, f$Child)
cor_mc <- cor(f$Mother, f$Child)

par(mfrow=c(1,2))

plot(f$Child ~ f$Father, pch = 19, col = "light blue", xlab = "Father", ylab = "Child")
abline(lm(f$Child ~ f$Father, data = f), col = "blue", lwd = 3)
abline(a = -7, b = 1, col="black", lty = 2, lwd =2)
abline(h = mean(f$Child), v = mean(f$Father), lwd = 0.1, lty = 3)
text(x = 164.5, y = 199, labels = paste("Cov =", round(cov_fc, digits = 2)))
text(x = 163.0, y = 195.7, labels = paste("r =", round(cor_fc, digits = 2)))
text(x = 195, y = 187, labels = grad)
title(main = "Father $ Child ", cex.main = 1.5)

plot(f$Child ~ f$Mother, pch = 19, col = "pink", lwd = 3, xlab = "Mother", ylab = "Child")
abline(lm(f$Child ~ f$Mother, data = f), col = "#FA5882", lwd = 3)
abline(a = 7, b = 1, col="black", lty = 2, lwd =2)
abline(h = mean(f$Child), v = mean(f$Mother), lwd = 0.1, lty = 3)
text(x = 152.6, y = 199, labels = paste("Cov =", round(cov_mc, digits = 2)))
text(x = 151.1, y = 196, labels = paste("r =", round(cor_mc, digits = 2)))
text(x = 175, y = 181, labels = grad)
title(main = "Mother & Child", cex.main = 1.5)
```

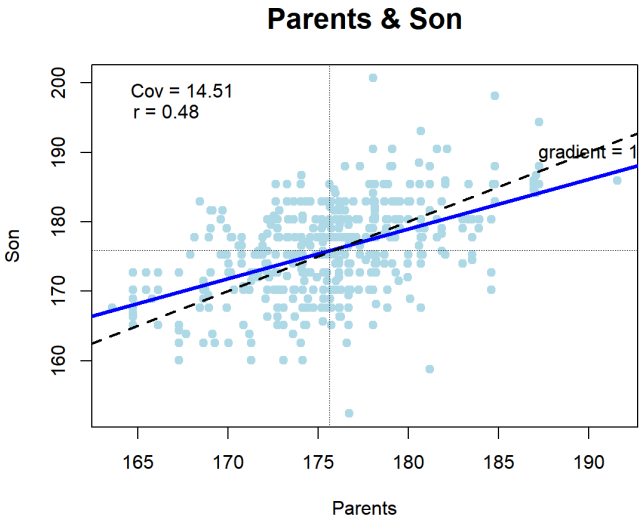


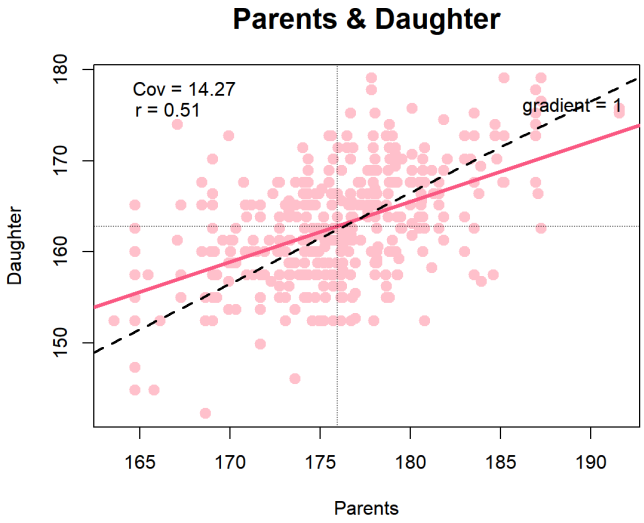


#아빠와 자식, 엄마와 자식의 산점도를 비교해본 결과 두 그래프 모두 공분산이 양수가 나와 아빠와 엄마의 키가 증가할때 자식 키가 증가함을 보인다. 또한 상관관계는 낮은걸로 보이지만 선형선의 기울기가 1보다 작아 자식들의 키가 평균으로 회귀함을 보인다.

#부모와 자식간의 산점도 3 (부모, 아들), (부모, 딸)

```
cov_pm <- cov(fm$MidParent, fm$Child)
cov_pw <- cov(fw$MidParent, fw$Child)
cor_pm <- cor(fm$MidParent, fm$Child)
cor_pw <- cor(fw$MidParent, fw$Child)

par(mfrow=c(1,2))

plot(fm$Child ~ fm$MidParent, pch = 19, col = "light blue", xlab = "Parents", ylab = "Son")
abline(lm(fm$Child ~ fm$MidParent, data = fm), col = "blue", lwd = 3)
abline(a = 0, b = 1, col="black", lty = 2, lwd =2)
abline(h = mean(fm$Child), v = mean(fm$MidParent), lwd = 0.1, lty = 3)
text(x = 167.5, y = 199, labels = paste("Cov =", round(cov_pm, digits = 2)))
text(x = 166.6, y = 196, labels = paste("r =", round(cor_pm, digits = 2)))
text(x = 190, y = 190, labels = grad)
title(main = "Parents & Son", cex.main = 1.5)

plot(fw$Child ~ fw$MidParent, pch = 19, col = "pink", lwd = 3, xlab = "Parents", ylab = "Daught
er")
abline(lm(fw$Child ~ fw$MidParent, data = fw), col = "#FA5882", lwd = 3)
abline(a = -13.5, b = 1, col="black", lty = 2, lwd =2)
abline(h = mean(fw$Child), v = mean(fw$MidParent), lwd = 0.1, lty = 3)
text(x = 167.5, y = 178, labels = paste("Cov =", round(cov_pw, digits = 2)))
text(x = 166.6, y = 175.7, labels = paste("r =", round(cor_pw, digits = 2)))
text(x = 189, y = 176, labels = grad)
title(main = "Parents & Daughter", cex.main = 1.5)
```

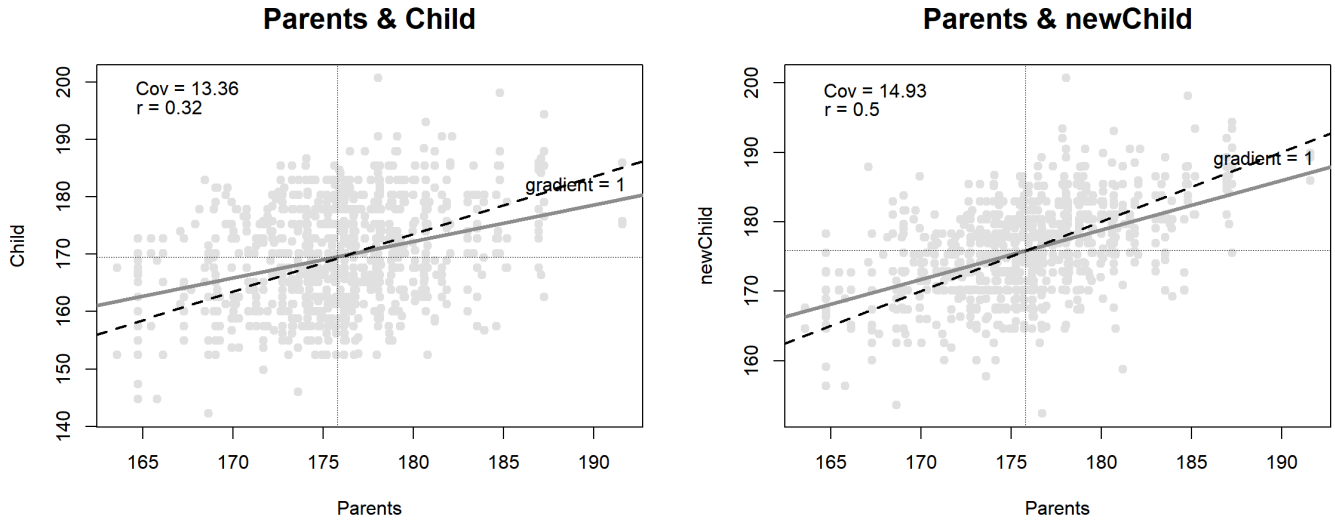




#부모와 아들, 부모와 딸의 산점도를 비교해본 결과 두 그래프 모두 공분산이 양수가 나와 부모의 키가 증가할때 아들과 딸의 키가 증가함을 보인다. 또한 상관관계 또한 조금있는걸로 보이고 선형선의 기울기가 1보다 작아 아들과 딸의 키가 평균으로 회귀함을 보인다.

#부모와 자식간의 산점도 4 (부모, 자식), (부모, 자식 (딸의 키 * 1.08))

```
cov_pc <- cov(f$MidParent, f$Child)
cor_pc <- cor(f$MidParent, f$Child)
cov_pnc <- cov(f$MidParent, f$newChild)
cor_pnc <- cor(f$MidParent, f$newChild)

par(mfrow=c(1,2))

plot(f$Child ~ f$MidParent, pch = 19, col = "gray88", xlab = "Parents", ylab = "Child")
abline(lm(f$Child ~ f$MidParent, data = f), col = "gray55 ", lwd = 3)
abline(a = -6.5, b = 1, col="black", lty = 2, lwd =2)
abline(h = mean(f$Child), v = mean(f$MidParent), lwd = 0.1, lty = 3)
text(x = 167.5, y = 199, labels = paste("Cov =", round(cov_pc, digits = 2)), col = "black")
text(x = 166.5, y = 195.7, labels = paste("r =", round(cor_pc, digits = 2)), col = "black")
text(x = 189, y = 182, labels = grad, col = "black")
title(main = " Parents & Child ", cex.main = 1.5)

plot(f$newChild ~ f$MidParent, pch = 19, col = "gray88", xlab = "Parents", ylab = "newChild")
abline(lm(f$newChild ~ f$MidParent, data = f), col = "gray55 ", lwd = 3)
abline(a = 0, b = 1, col="black", lty = 2, lwd =2)
abline(h = mean(f$newChild), v = mean(f$MidParent), lwd = 0.1, lty = 3)
text(x = 167.5, y = 199, labels = paste("Cov =", round(cov_pnc, digits = 2)), col = "black")
text(x = 166.2, y = 196.3, labels = paste("r =", round(cor_pnc, digits = 2)), col = "black")
text(x = 189, y = 189, labels = grad, col = "black")
title(main = " Parents & newChild ", cex.main = 1.5)
```



#마지막으로 부모와 자식간의 산점도를 비교해본결과 양의 선형관계를 가지고있고 선형선의 기울기가 1보다 작아 자식의 키가 평균으로 회귀함을 볼수있다. 수정된 자식 데이터로 비교한 결과도 같은 해석이 가능하고 다른점은 상관관계가 수정하지않은 자식 데이터보다 높게 나왔다는 점이다.(당연한 결과)

#검정

```
summary(f)
```

```
##      Father          Mother        MidParent         Child      
##  Min.   :157.5   Min.   :147.3   Min.   :163.6   Min.   :142.2  
##  1st Qu.:172.7   1st Qu.:160.0   1st Qu.:173.1   1st Qu.:162.6  
##  Median :175.3   Median :162.6   Median :175.9   Median :168.9  
##  Mean   :175.8   Mean   :162.8   Mean   :175.8   Mean   :169.5  
##  3rd Qu.:180.3   3rd Qu.:167.3   3rd Qu.:178.2   3rd Qu.:177.0  
##  Max.   :199.4   Max.   :179.1   Max.   :191.6   Max.   :200.7  
##     newChild      Childtype  
##  Min.   :152.4   female:453  
##  1st Qu.:171.4   male  :481  
##  Median :175.6               
##  Mean   :175.9               
##  3rd Qu.:180.3               
##  Max.   :200.7
```

```
var.test(f$Child ~ f$Childtype)
```

```
## 
##  F test to compare two variances
## 
## data:  f$Child by f$Childtype
## F = 0.80599, num df = 452, denom df = 480, p-value = 0.02034
## alternative hypothesis: true ratio of variances is not equal to 1
## 95 percent confidence interval:
##  0.6721144 0.9670135
## sample estimates:
## ratio of variances
##          0.8059879
```

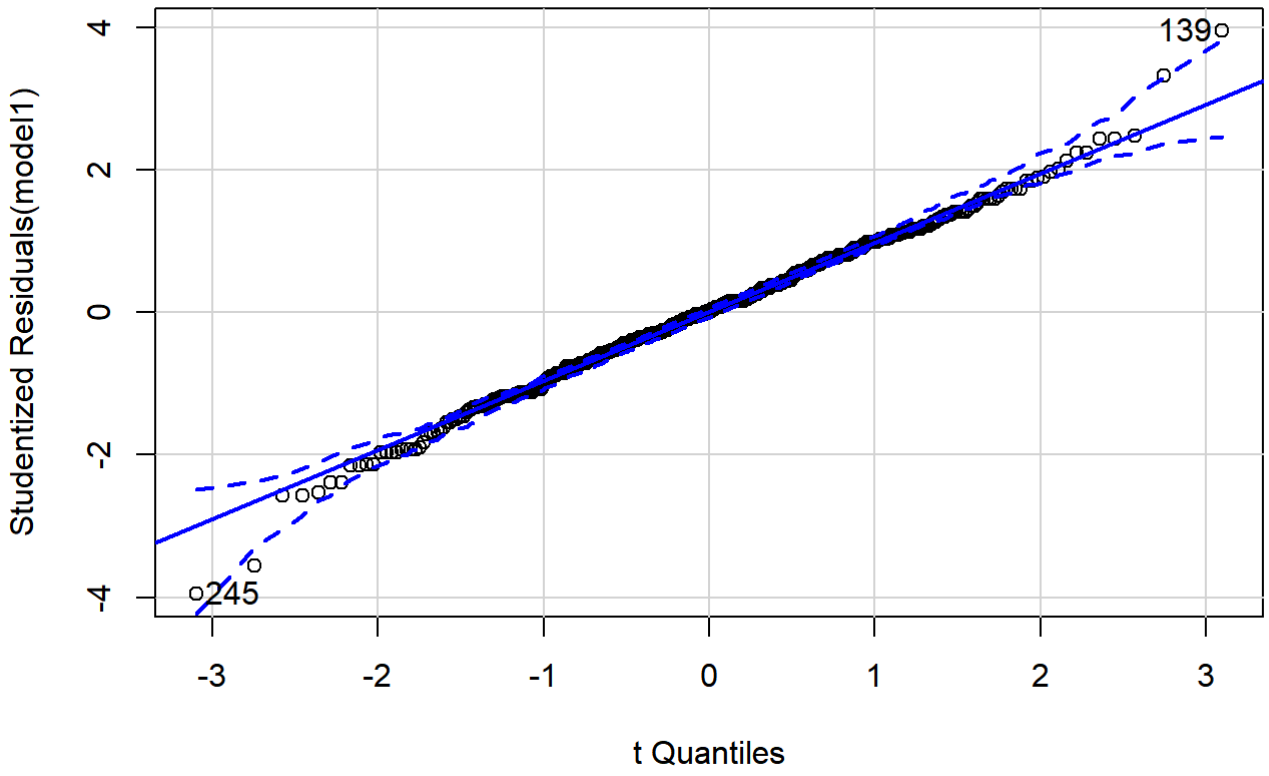
```
t.test(f$Child ~ f$Childtype, var.equal = FALSE)
```

```
## 
##  Welch Two Sample t-test
## 
## data:  f$Child by f$Childtype
## t = -31.476, df = 929.89, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -13.84303 -12.21815
## sample estimates:
## mean in group female   mean in group male 
##             162.8242             175.8548
```

```
#등분산검정과 t검정 결과로 추측했을때 성별의 따른 키차이가 있음이 보인다.
```

#키 예측 (아빠의 키로부터 아들의 키)

```
model1 <- lm(fm$Child ~ fm$Father, data = fm)

#정규성 검정(직관적(시각적)해석을 위한 qqPlot, shapiro-wilk nomality test 의 p-value > 0.05 이상이면 정규분포)
qqPlot(model1)
```




```
## [1] 139 245
```

```
shapiro.test(resid(model1))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model1)
## W = 0.99394, p-value = 0.05214
```

```
# 이상치를 처리하기위해 car패키지 이용
outlierTest(model1)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 245 -3.952172         8.9139e-05     0.042876
## 139  3.946057         9.1371e-05     0.043949
```

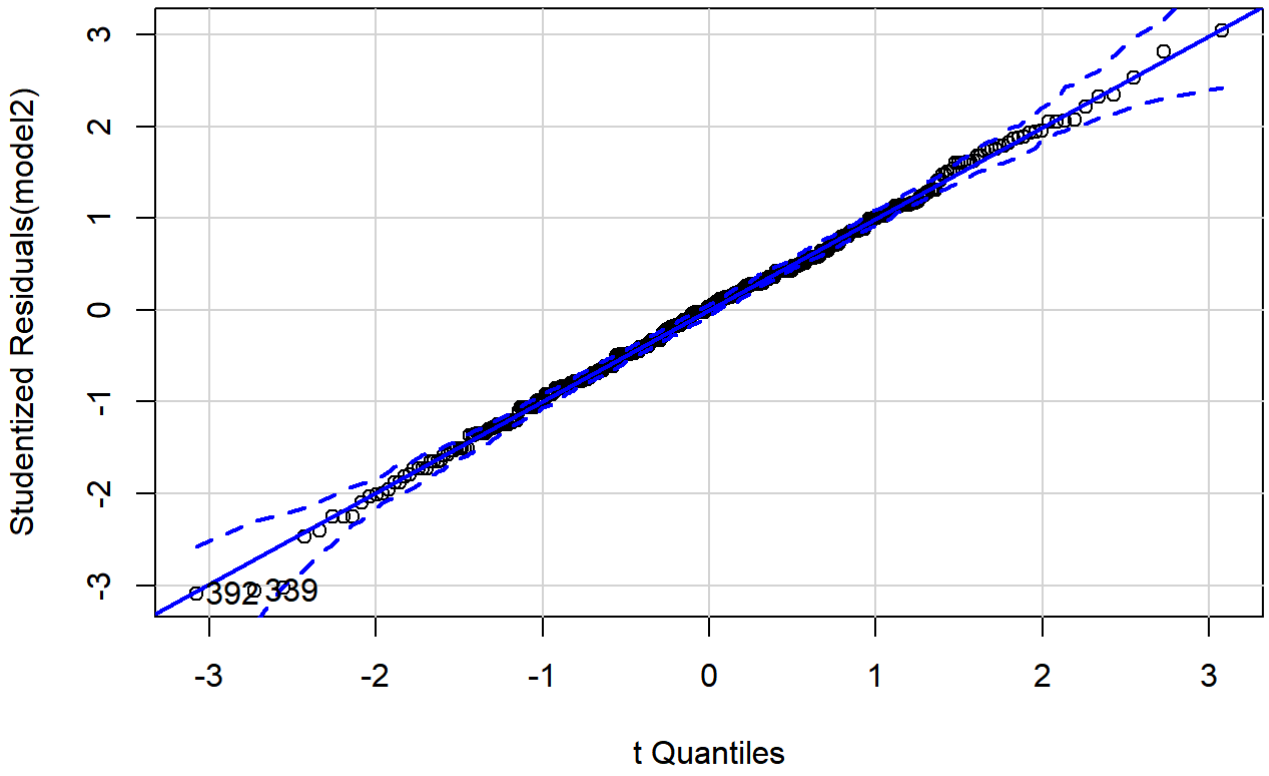
```
fm2 <- subset(fm, rownames(fm) != "293" & rownames(fm) != "487")

model1 <- lm(fm2$Child ~ fm2$Father, data = fm2)

summary(model1)
```

```
##
## Call:
## lm(formula = fm2$Child ~ fm2$Father, data = fm2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -23.8670  -3.8421   0.1155   4.1742  23.8260
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 97.45324    8.41282  11.584   <2e-16 ***
## fm2$Father   0.44646    0.04788   9.325   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.143 on 478 degrees of freedom
## Multiple R-squared:  0.1539, Adjusted R-squared:  0.1521
## F-statistic: 86.95 on 1 and 478 DF,  p-value: < 2.2e-16
```

```
#회귀분석 결과 P-value(유의확률) < 0.05 이므로 회귀식은 통계적으로 유의하다.
#분석결과 : 회귀식 (아들의 키 = 0.446 * 아빠의 키 + 97.453) 이므로 아빠의 키로써 아들의 키를 예측할수있다.
```

#키 예측 (엄마의 키로 부터 딸의 키)

```
model2 <- lm(fw$Child ~ fw$Mother, data = fw)

#정규성 검정
qqPlot(model2)
```




```
## [1] 339 392
```

```
shapiro.test(resid(model2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model2)
## W = 0.99797, p-value = 0.8704
```

```
#이상치 처리
outlierTest(model2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 392 -3.094101          0.0020971      0.94999
```

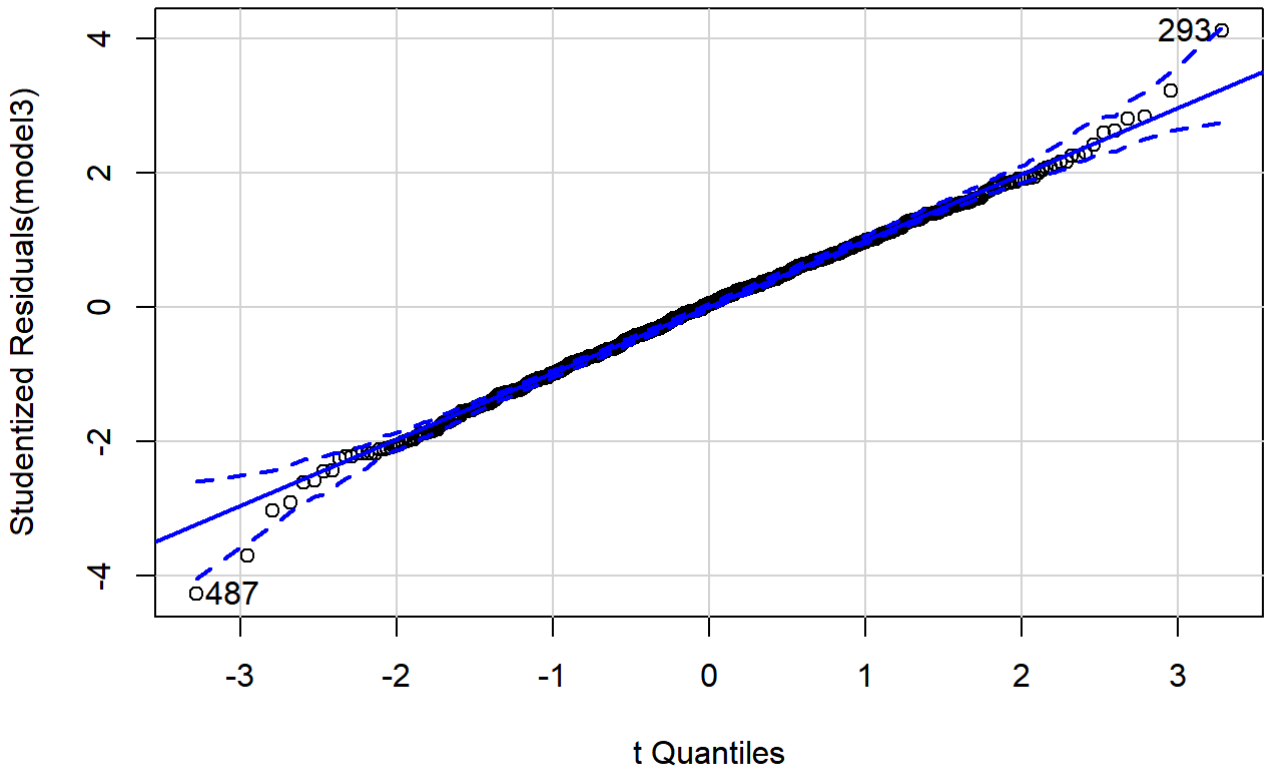
```
fw2 <- subset(fw, rownames(fw) != "392")

model2 <- lm(fw2$Child ~ fw2$Mother, data = fw2)

summary(model2)
```

```
##
## Call:
## lm(formula = fw2$Child ~ fw2$Mother, data = fw2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.242  -3.872   0.178   3.628  17.143
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 110.59440    7.55395  14.641  < 2e-16 ***
## fw2$Mother    0.32079    0.04633   6.923 1.53e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.651 on 450 degrees of freedom
## Multiple R-squared:  0.09627,    Adjusted R-squared:  0.09426
## F-statistic: 47.93 on 1 and 450 DF,  p-value: 1.527e-11
```

```
#회귀분석 결과 P-value(유의확률) < 0.05 이므로 회귀식은 통계적으로 유의하다.
#분석결과 : 회귀식 (딸의 키 = 0.321 * 엄마의 키 + 110.594) 이므로 엄마의 키로써 딸의 키를 예측할수있다.
```

#키 예측 (부모님의 키로 부터 자식의 키)

```
#딸의키는 기존의 데이터에 1.08을 곱한 newChild의 데이터를 사용하기로한다.

model3 <- lm(f$newChild ~ f$MidParent, data =f)

#정규성 검정
qqPlot(model3)
```




```
## [1] 293 487
```

```
shapiro.test(resid(model3))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model3)
## W = 0.99687, p-value = 0.06325
```

```
#이상치 처리
outlierTest(model3)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 487 -4.274910         2.1091e-05     0.019699
## 293  4.107187         4.3573e-05     0.040698
```

```
f2 <- subset(f, rownames(f) != "293" & rownames(f) != "487")


model3 <- lm(f2$newChild ~ f2$MidParent, data =f2)

summary(model3)
```

```
##
## Call:
## lm(formula = f2$newChild ~ f2$MidParent, data = f2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.9391  -3.7786   0.2392   3.8681  18.2244
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  50.87359    7.04074   7.226 1.04e-12 ***
## f2$MidParent  0.71098    0.04004  17.757  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.598 on 930 degrees of freedom
## Multiple R-squared:  0.2532, Adjusted R-squared:  0.2524
## F-statistic: 315.3 on 1 and 930 DF,  p-value: < 2.2e-16
```

```
#회귀분석 결과 P-value(유의확률) < 0.05 이므로 회귀식은 통계적으로 유의하다.
#분석결과 : 회귀식 ( 자식의 키 = 0.711 * 부모의 키 + 50.873 ) 이므로 부모의 키로써 자식의 키를
 예측할수있다.
#만약 딸의 키를 예측하고 싶으면 예측된 자식의 키 나누기 1.08을 해주면 된다.
```