기말 프로젝트

20173218 김주형, 20187115 조금주

2020-12-01

```
# Galton은 부모와 자식들 간의 키의 상관관계를 분석해 본 결과, 다음과 같은 재미있는 관계를 찾아내었다. 즉, 특이하게 큰 부모의 자식들은 대게 크긴 하되 부모들보다는 대부분 작았고, 특이하게 작은 부모들의 자식들은 대게 작긴 하되 부모들보다는 대부분 크다는 사실이다. 이러한 경향은 사람들의 키가 평균키로 회귀하려는 경향이 있음을 말하는 것인데, 바로 이 연구에서부터 회귀분석이라는 용어가 사용되게 되었다.(즉 부모의 키가 크(작)더라도 그 자식들은 결국 보통키로 회귀(돌아간다)한다는 뜻이다.) 단, 키의 단위는 inch 이다. (1 inch = 2.54 cm)
```

폰트 불러오기

```
library(showtext)
```

```
## Loading required package: sysfonts
```

```
## Loading required package: showtextdb
```

```
showtext_auto()
font_add_google('Libre Baskerville', 'LB')
```

데이터 불러오기

```
library(HistData)

data("GaltonFamilies")

G <- GaltonFamilies
```

아이 성별 구분

```
# 남자아이 키
male <- G[G$gender == "male", ]
str(male)
```

```
## 'data.frame':    481 obs. of  8 variables:
##  $ family         : Factor w/ 205 levels "001","002","003",..: 1 2 2 3 4 4 5 5 5 7 ...
##  $ father         : num  78.5 75.5 75.5 75 75 75 75 75 75 74 ...
##  $ mother         : num  67 66.5 66.5 64 64 64 58.5 58.5 58.5 68 ...
##  $ midparentHeight: num  75.4 73.7 73.7 72.1 72.1 ...
##  $ children       : int  4 4 4 2 5 5 6 6 6 6 ...
##  $ childNum       : int  1 1 2 1 1 2 1 2 3 1 ...
##  $ gender         : Factor w/ 2 levels "female","male": 2 2 2 2 2 2 2 2 2 2 ...
##  $ childHeight    : num  73.2 73.5 72.5 71 70.5 68.5 72 69 68 76.5 ...
```

```
# 여자아이 키
female <- G[G$gender == "female", ]
str(female)
```

```
## 'data.frame':    453 obs. of  8 variables:
##  $ family         : Factor w/ 205 levels "001","002","003",..: 1 1 1 2 2 3 4 4 4 5 ...
##  $ father         : num  78.5 78.5 78.5 75.5 75.5 75 75 75 75 75 ...
##  $ mother         : num  67 67 67 66.5 66.5 64 64 64 64 58.5 ...
##  $ midparentHeight: num  75.4 75.4 75.4 73.7 73.7 ...
##  $ children       : int  4 4 4 4 4 2 5 5 5 6 ...
##  $ childNum       : int  2 3 4 3 4 2 3 4 5 4 ...
##  $ gender         : Factor w/ 2 levels "female","male": 1 1 1 1 1 1 1 1 1 1 ...
##  $ childHeight    : num  69.2 69 69 65.5 65.5 68 67 64.5 63 66.5 ...
```

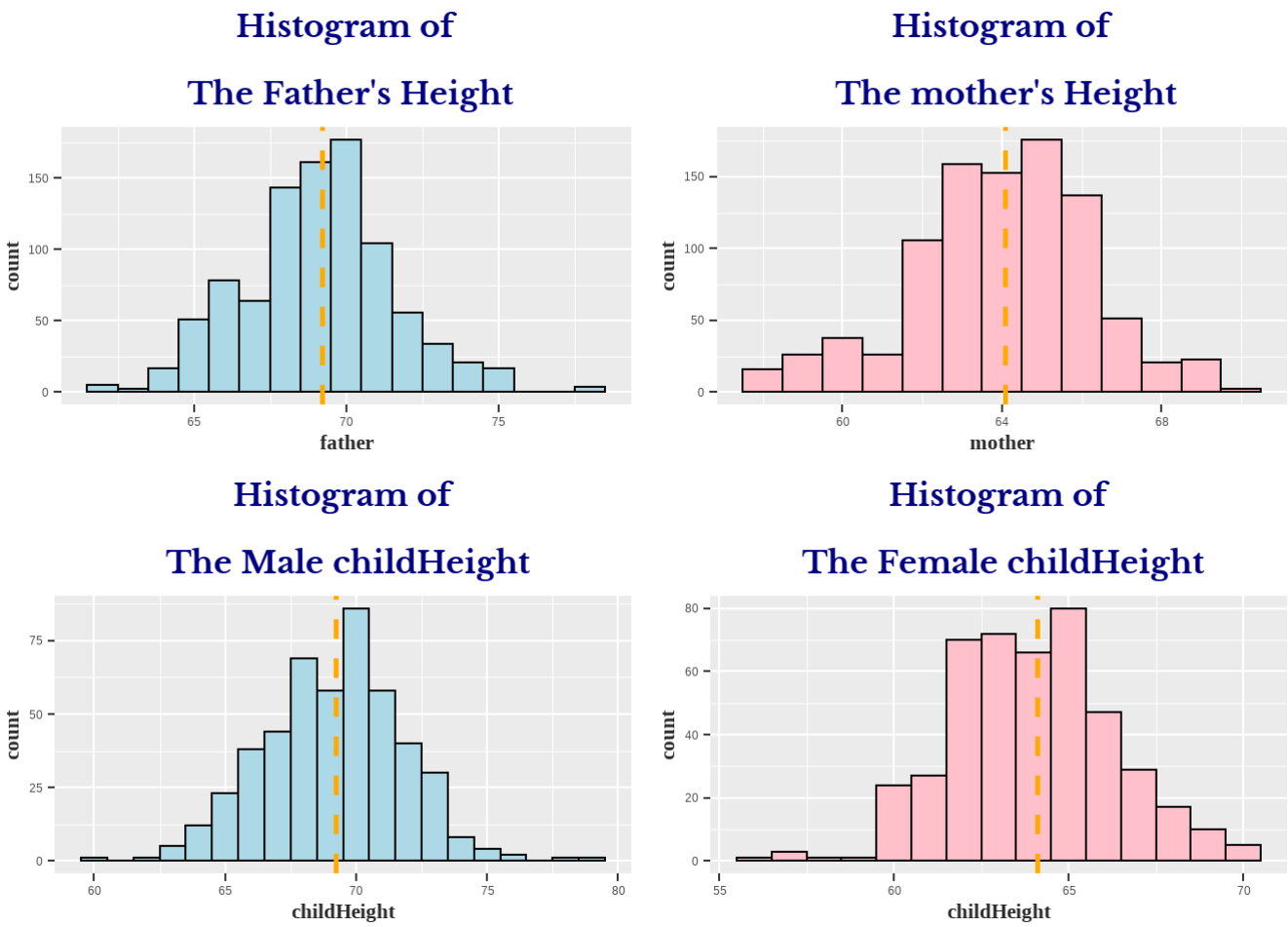
부모님 키와 아이의 키 - Histogram

```
# Histogram
library(ggplot2)
library(gridExtra)
library(ggExtra)
library(car)
```

```
## Loading required package: carData
```

```
# 아버지 키
gg_f_h <- ggplot(data = G, aes(x = father)) +
  geom_histogram(binwidth = 1, fill = "lightblue",
                 colour = "black") +
  geom_vline(xintercept = mean(G$father), linetype = "dashed", color = "#FFAA00", lwd = 1) +
  ggtitle("Histogram of Wn The Father's Height") +
  theme(plot.title = element_text(family = "LB",
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "Navy Blue")) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))

# 어머니 키
gg_m_h <- ggplot(data = G, aes(x = mother)) +
  geom_histogram(binwidth = 1, fill = "pink",
                 colour = "black") +
  geom_vline(xintercept = mean(G$mother), linetype = "dashed", color = "#FFAA00", lwd = 1) +
  ggtitle("Histogram of Wn The mother's Height") +
  theme(plot.title = element_text(family = "LB",
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "Navy Blue")) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))

# 남자아이 키
gg_m_c_mid <- ggplot(data = male, aes(x = childHeight)) +
  geom_histogram(binwidth = 1, fill = "lightblue",
                 colour = "black") +
  geom_vline(xintercept = mean(male$childHeight), linetype = "dashed", color = "#FFAA00", lwd =
1) +
  ggtitle("Histogram of Wn The Male childHeight") +
  theme(plot.title = element_text(family = "LB",
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "Navy Blue")) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))

# 여자아이 키
gg_f_c_mid <- ggplot(data = female, aes(x = childHeight)) +
  geom_histogram(binwidth = 1, fill = "pink",
                 colour = "black") +
  geom_vline(xintercept = mean(female$childHeight), linetype = "dashed", color = "#FFAA00", lwd
= 1) +
  ggtitle("Histogram of Wn The Female childHeight") +
  theme(plot.title = element_text(family = "LB",
```

```
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "Navy Blue")) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))

grid.arrange(gg_f_h, gg_m_h, gg_m_c_mid, gg_f_c_mid, nrow = 2, ncol = 2)
```




```
# 아버지 키 평균
mean(G$father) # 단위는 inch 이다. 69.19711 in = 175.76 cm
```

```
## [1] 69.19711
```

```
# 어머니 키 평균
mean(G$mother) # 64.08929 in = 162.79 cm
```

```
## [1] 64.08929
```

```
# 남자아이 키 평균
mean(male$childHeight) # 69.2341 in = 175.85 cm
```

```
## [1] 69.2341
```

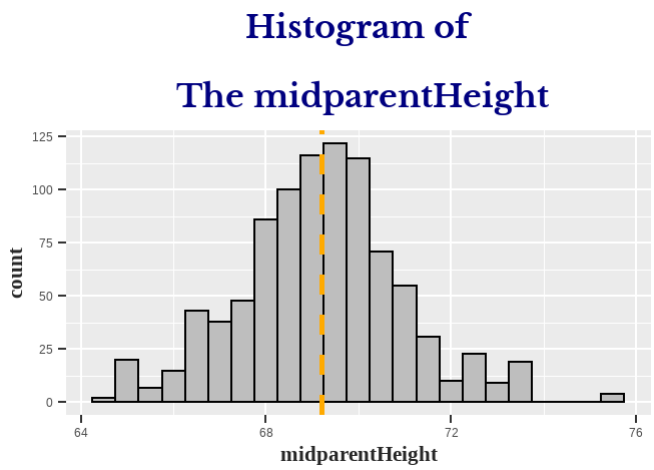
```
# 여자아이 키 평균
mean(female$childHeight) # 64.10397 in = 162.82 cm
```

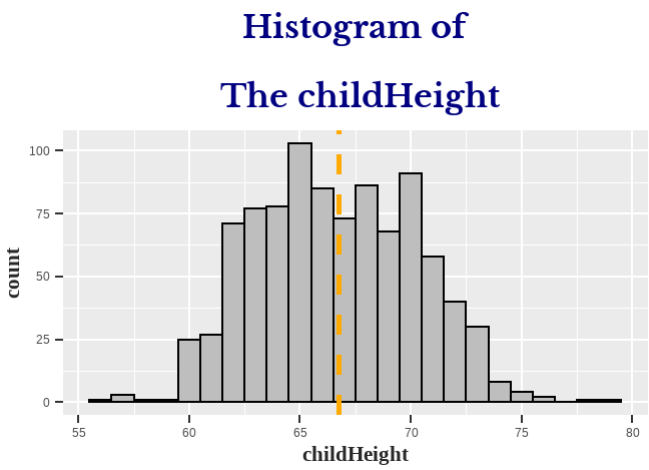
```
## [1] 64.10397
```

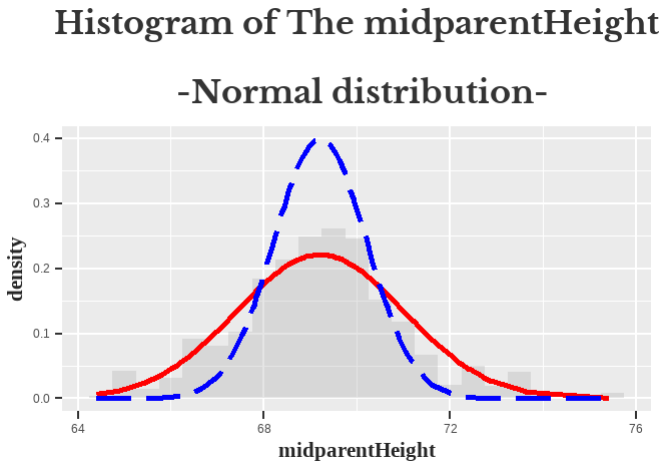
부모님의 평균 키와 아이들의 키

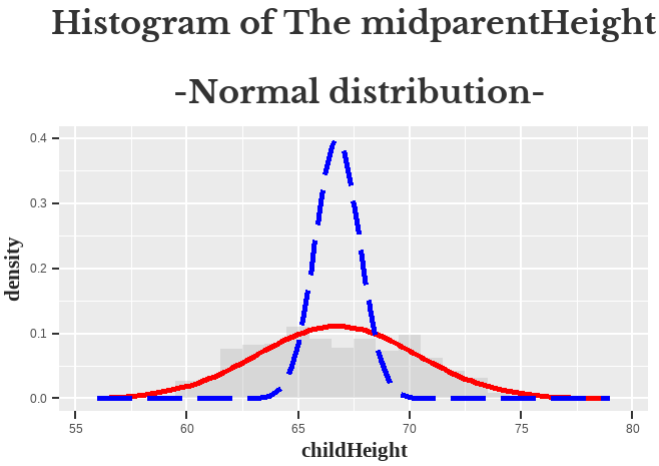
```
# 부모님
gg_p_mid <- ggplot(data = G, aes(x = midparentHeight)) +
  geom_histogram(binwidth = 0.5,
                 fill = "grey",
                 colour = "black") +
  geom_vline(xintercept = mean(G$midparentHeight),
             linetype = "dashed",
             color = "#FFAA00",
             lwd= 1) +
  ggtitle("Histogram of ₩n The midparentHeight") +
  theme(plot.title = element_text(family = "LB",
                                  size = 24,face = "bold",
                                  hjust = 0.5,
                                  color = "Navy Blue")) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))
# 아이들
gg_c_mid <- ggplot(data = G, aes(x = childHeight)) +
  geom_histogram(binwidth = 1,
                 fill = "grey",
                 colour = "black") +
  geom_vline(xintercept = mean(G$childHeight),
             linetype = "dashed",
             color = "#FFAA00",
             lwd = 1) +
  ggtitle("Histogram of ₩n The childHeight") +
  theme(plot.title = element_text(family = "LB",
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "Navy Blue")) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))


# 정규분포
n1 <- ggplot(data = G, aes(x = midparentHeight)) +
  geom_histogram(aes(y = ..density..),
                 binwidth = 0.5,
                 fill = "darkgrey",
                 alpha = 0.3) +
  geom_line(aes( y = dnorm(midparentHeight,
                           mean(midparentHeight),
                           sd(midparentHeight))),
            color = "red",
            linetype=1 ,
            lwd = 1) +
  geom_line(aes(y = dnorm(midparentHeight,
                          mean(midparentHeight), 1)),
            color = "blue",
            linetype=5,
            lwd = 1) +
  ggtitle("Histogram of The midparentHeight ₩n -Normal distribution-")+
```

```
  theme(plot.title = element_text(family = "LB",
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "grey20")) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))

n2<- ggplot(data = G, aes(x = childHeight)) +
  geom_histogram(aes(y =  ..density..),
                 binwidth = 1,
                 fill = "darkgrey",
                 alpha =0.3) +
  geom_line(aes( y = dnorm(childHeight,
                           mean(childHeight),
                           sd(childHeight))),
            color = "red",
            linetype=1 ,
            lwd = 1) +
  geom_line(aes(y = dnorm(childHeight,
                          mean(childHeight),1)),
            color = "blue",
            linetype=5,
            lwd = 1) +
  ggtitle("Histogram of The midparentHeight ₩n -Normal distribution-")+
  theme(plot.title = element_text(family = "LB",
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "grey20")) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))

grid.arrange(gg_p_mid, gg_c_mid, n1, n2 ,nrow = 2, ncol = 2)
```










```
# 통계량 분석 후, 그래프 분석을 실시.
```

```
# 두 연속형 변수 상관관계

# 통계량 분석
# 1.공분산(covaiance)
# 2.상관계수(correlation coefficient)

# 그래프 분석
# 1.산점도(scatter plot)
# 2.산점도 행렬(scatter matrix plot)
# 3.상관계수행렬(correlation coefficient plot)
```

부모님의 키와 아이의 키의 관계 - 통계량 분석 1

```
# 1.공분산
# 아버지와 아들
cov_f_m <- cov(male$father, male$childHeight)
cov_f_m
```

```
## [1] 2.373964
```

```
# 아버지와 딸
cov_f_f <- cov(female$father, female$childHeight)
cov_f_f
```

```
## [1] 2.671287
```

```
# 어머니와 아들
cov_m_m <- cov(male$mother, male$childHeight)
cov_m_m
```

```
## [1] 1.967655
```

```
# 어머니와 딸
cov_m_f <- cov(female$mother, female$childHeight)
cov_m_f
```

```
## [1] 1.623788
```

```
# 부모님의 평균과 자식들
cov_p_c <- cov(G$midparentHeight, G$childHeight)
cov_p_c
```

```
## [1] 2.070491
```

```
# 공분산이 모두 양수이므로 두 변수 간의 상관관계는 상승하는 경향이 있다.
```

부모님의 키와 아이의 키의 관계 - 통계량 분석 2

```
# 상관계수 - 공분산의 표준화
cor.test(male$father, male$childHeight) # 아버지와 아들
```

```
##
##  Pearson's product-moment correlation
##
## data:  male$father and male$childHeight
## t = 9.3365, df = 479, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3139916 0.4654615
## sample estimates:
##       cor
## 0.3923835
```

```
cor.test(female$father, female$childHeight) # 아버지와 딸
```

```
##
##  Pearson's product-moment correlation
##
## data:  female$father and female$childHeight
## t = 10.069, df = 451, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.3501215 0.5007970
## sample estimates:
##      cor
## 0.428433
```

```
cor.test(male$mother, male$childHeight) # 어머니와 아들
```

```
##
##  Pearson's product-moment correlation
##
## data:  male$mother and male$childHeight
## t = 7.4697, df = 479, p-value = 3.838e-13
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2405444 0.4008365
## sample estimates:
##      cor
## 0.323005
```

```
cor.test(female$mother, female$childHeight) # 어머니와 딸
```

```
##
##  Pearson's product-moment correlation
##
## data:  female$mother and female$childHeight
## t = 6.8053, df = 451, p-value = 3.222e-11
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2191956 0.3864314
## sample estimates:
##       cor
## 0.3051645
```

```
cor.test(G$midparentHeight, G$childHeight) # 부모님의 평균과 자식들
```

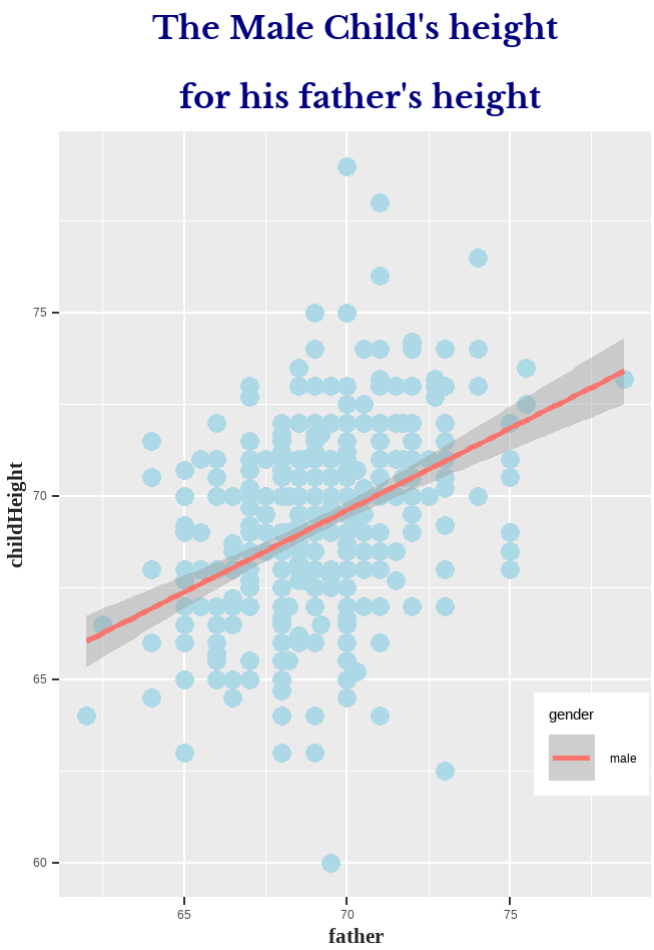
```
##
##  Pearson's product-moment correlation
##
## data:  G$midparentHeight and G$childHeight
## t = 10.345, df = 932, p-value < 2.2e-16
## alternative hypothesis: true correlation is not equal to 0
## 95 percent confidence interval:
##  0.2622011 0.3773285
## sample estimates:
##       cor
## 0.3209499
```

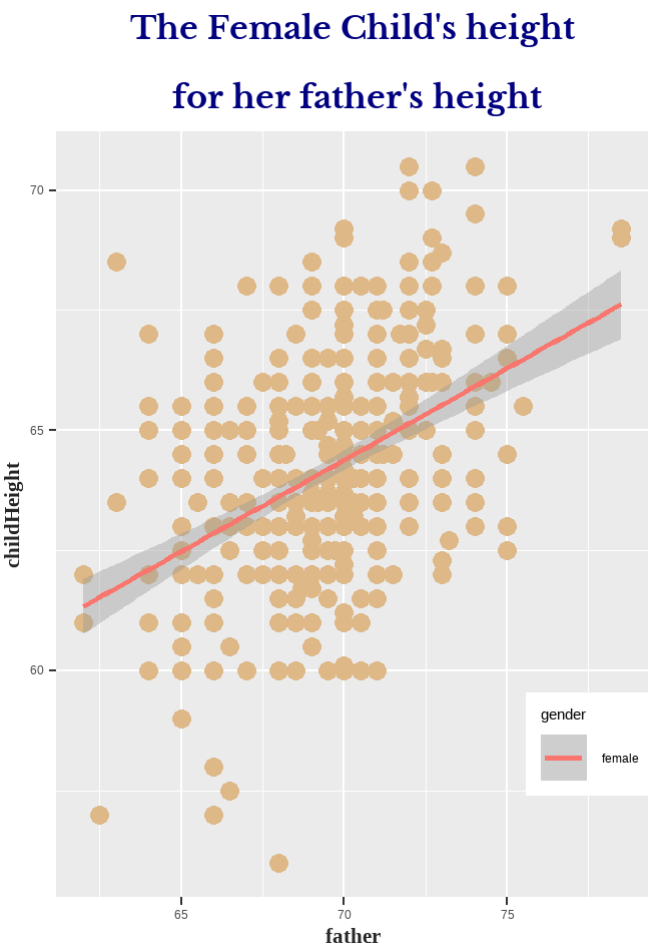
부모님의 키와 아이의 키와 관계 - 그래프 분석 1

```
# 아버지와 아들, 어머니와 딸, 부모님의 평균과 자식들의 키에 대한 산점도(scatter plot) 그리기
```

```
# 아버지의 키(father)와 아들의 키의 관계
gg_f_mc <- ggplot(data = male, aes(x = father, y = childHeight,
                        color = gender)) +
  geom_point(size = 3, color = "lightblue") + stat_smooth(method = "lm") +
  ggtitle("The Male Child's height Wn for his father's height") +
  theme(plot.title = element_text(family = "LB",
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "Navy Blue"),
         legend.position = c(0.9, 0.2)) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))

# 아버지의 키(father)와 딸의 키의 관계
gg_f_fc <- ggplot(data = female, aes(x = father, y = childHeight,
                        color = gender)) +
  geom_point(size = 3, color = "burlywood") + stat_smooth(method = "lm") +
  ggtitle("The Female Child's height Wn for her father's height")+
  theme(plot.title = element_text(family = "LB",
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "Navy Blue"),
        legend.position = c(0.9, 0.2)) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))

grid.arrange(gg_f_mc, gg_f_fc, nrow = 1, ncol = 2)
```

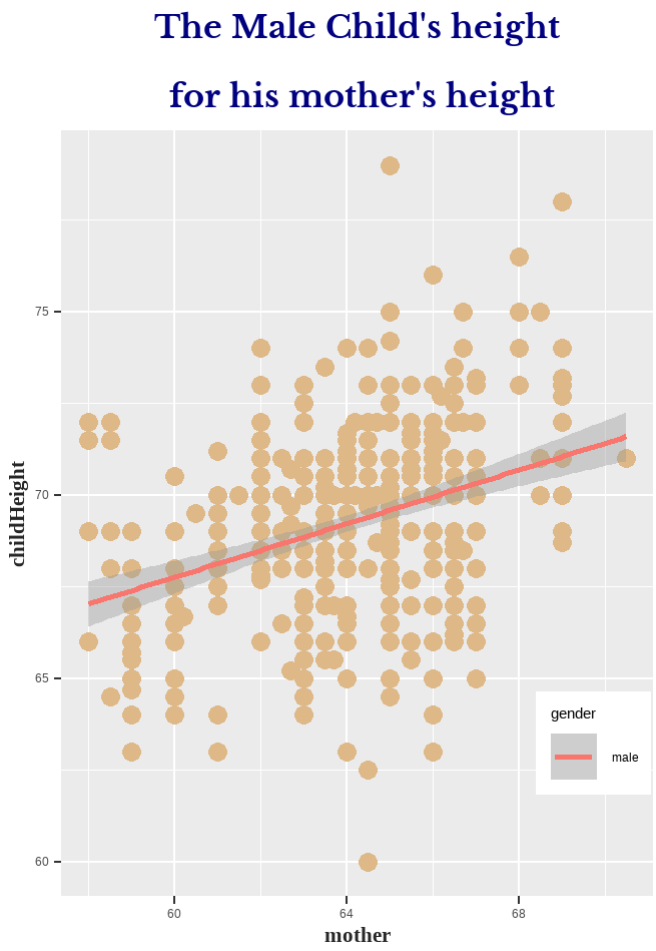
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```

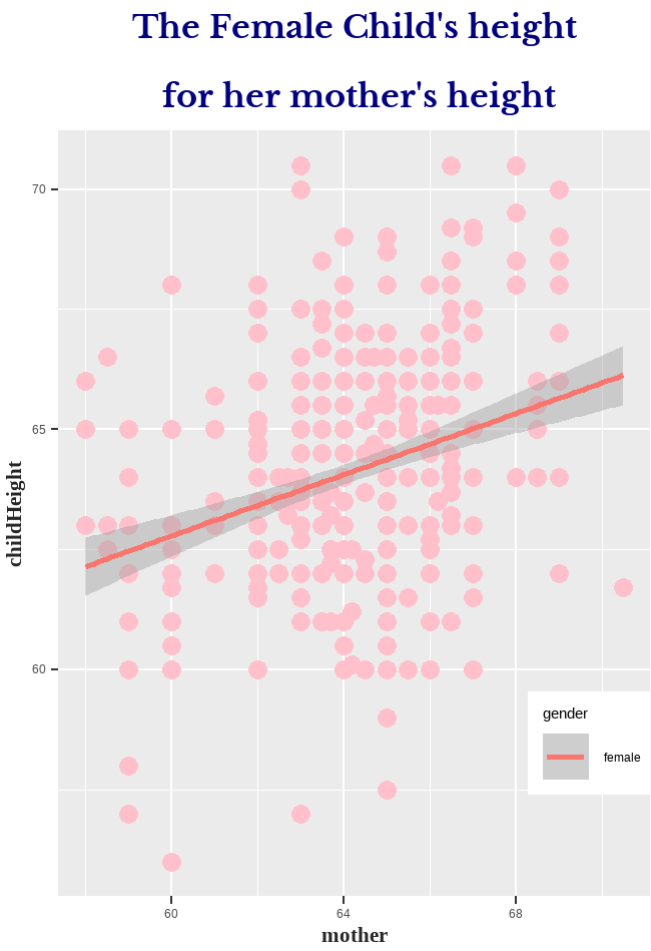




```
# 어머니의 키(mother)와 아들의 키의 관계
gg_m_mc <- ggplot(data = male, aes(x = mother, y = childHeight,
                        color = gender)) +
  geom_point(size = 3, color = "burlywood") + stat_smooth(method = "lm") +
  ggtitle("The Male Child's height Wn for his mother's height") +
  theme(plot.title = element_text(family = "LB",
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "Navy Blue"),
        legend.position = c(0.9, 0.2)) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))

# 어머니의 키(mother)와 딸의 키의 관계
gg_m_fc <- ggplot(data = female, aes(x = mother, y = childHeight,
                        color = gender)) +
  geom_point(size = 3, color = "pink") + stat_smooth(method = "lm") +
  ggtitle("The Female Child's height Wn for her mother's height") +
  theme(plot.title = element_text(family = "LB",
                                  size = 24,
                                  face = "bold",
                                  hjust = 0.5,
                                  color = "Navy Blue"),
        legend.position = c(0.9, 0.2)) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))

grid.arrange(gg_m_mc, gg_m_fc, nrow = 1, ncol = 2)
```

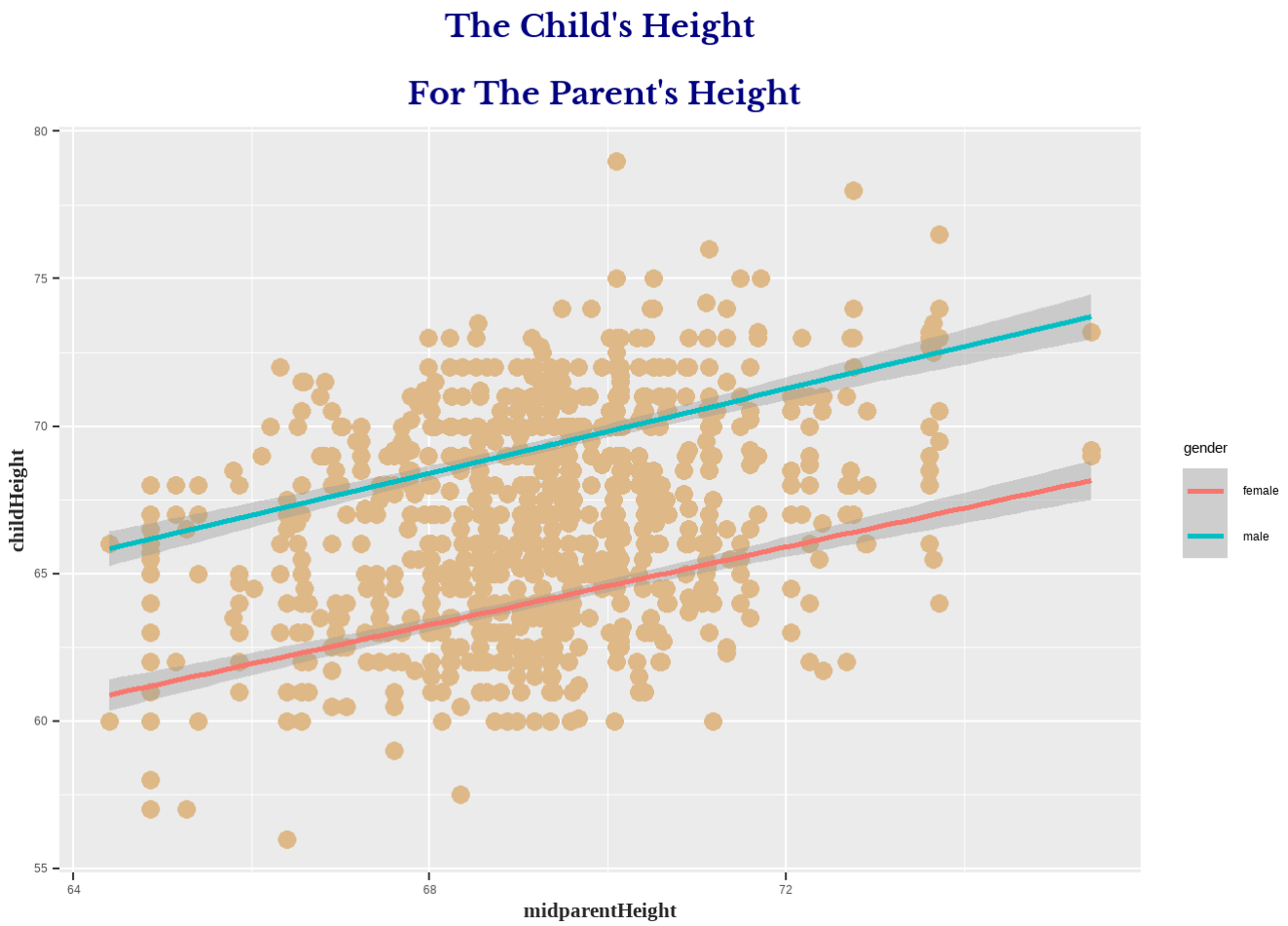
```
## `geom_smooth()` using formula 'y ~ x'
## `geom_smooth()` using formula 'y ~ x'
```






```
# 남, 여자아이 합치기
ggplot(data = G, aes(x = midparentHeight, y = childHeight,
                     color = gender)) +
  geom_point(size = 3, color = "burlywood") + stat_smooth(method = "lm") +
  ggtitle("The Child's Height Wn For The Parent's Height")+
  theme(plot.title = element_text(family = "LB", size = 24, face = "bold",
                                  hjust = 0.5,color="Navy Blue")) +
  theme(axis.title = element_text(family = "serif",
                                  face = "bold",
                                  size = 16,
                                  color = "grey15"))
```

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



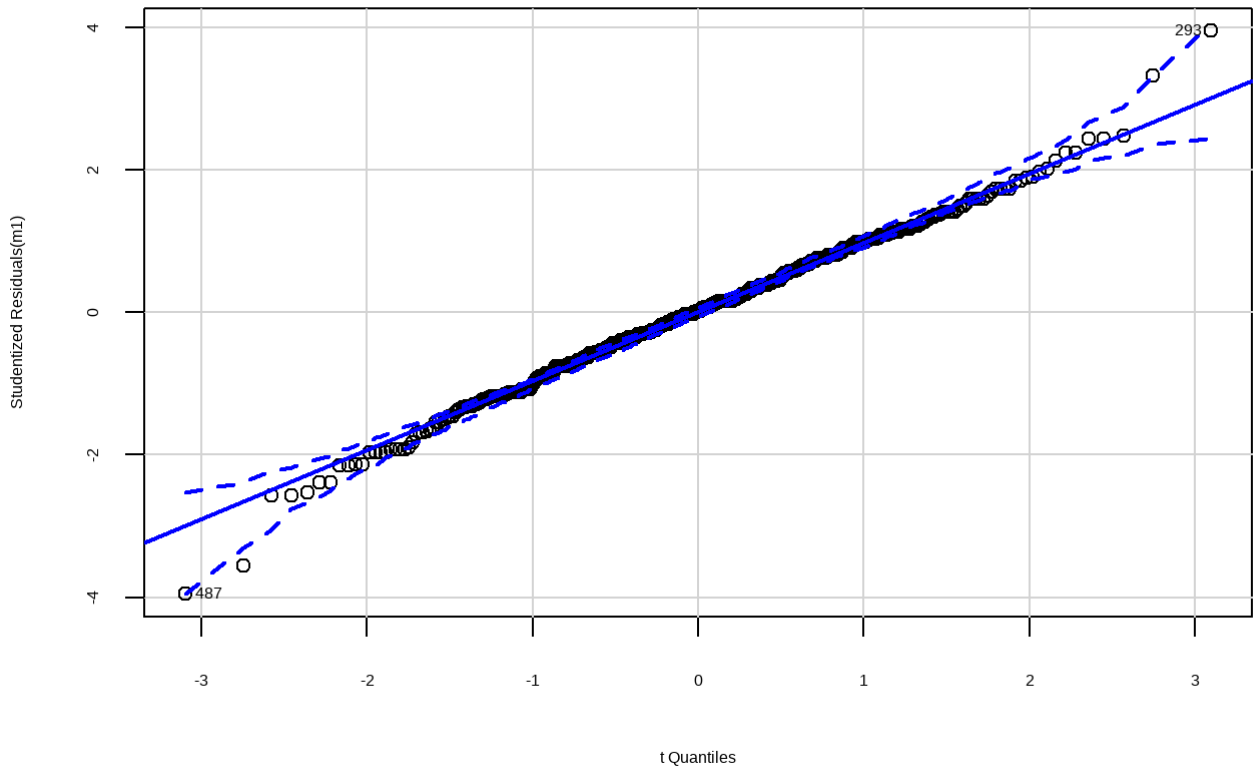
부모에 의한 자식의 키 예측(회귀식 추정)

```
# 아버지와 아들
```

```
# 회귀선
m1 <- lm(childHeight ~ father, male) # 아버지와 아들

# 이상치 제거
outlierTest(m1)
```

```
##      rstudent unadjusted p-value Bonferroni p
## 487 -3.952163         8.9142e-05     0.042877
## 293  3.946110         9.1351e-05     0.043940
```

```
out1 <- subset(male, rownames(male) != "293" & rownames(male) != "487")

# 정규성 확인
qqPlot(m1) # 그래프 분석
```




```
## 293 487
## 139 245
```

```
# 회귀분석
summary(m1)
```

```
##
## Call:
## lm(formula = childHeight ~ father, data = male)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.3959 -1.5122  0.0413  1.6217  9.3808
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 38.36258    3.30837  11.596   <2e-16 ***
## father       0.44652    0.04783   9.337   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.416 on 479 degrees of freedom
## Multiple R-squared:  0.154,  Adjusted R-squared:  0.1522
## F-statistic: 87.17 on 1 and 479 DF,  p-value: < 2.2e-16
```

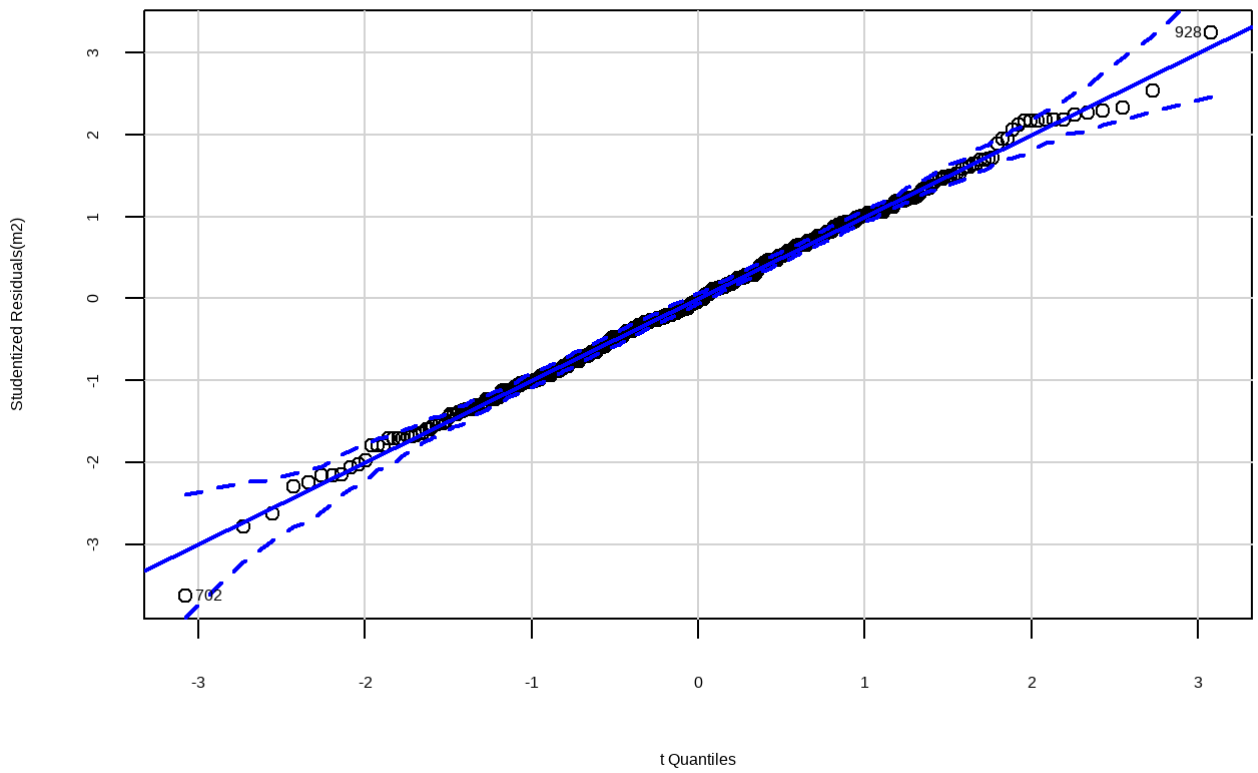
```
# yhat = 38.36258 + 0.44652 * x   # <- 회귀식 (단위: inch)
# |t|통계량 Pr(>|t|) 값(2e-16)이 유의수준 0.05보다 작기 때문에 통계적으로 유의함.
# 잔차 표준오차(Residual standard error): 2.416
# 결정계수(Multiple R-Squared): 0.154 아버지의 키가 아들의 키에 대하여 15.4% 설명할 수 있다.
# 모양이 적합한지를 보기 위해 F-검정통계량(F-statistic)의 값(87.17)을 보면 유의수준 0.05보다 작기 때문에 모양이 유의하다고 할 수 있다.
# 즉, 아버지의 키는 아들의 키의 약한 영향을 미친다고 할 수 있다.
```

```
# 아버지와 딸
```

```
# 회귀선
m2 <- lm(childHeight ~ father, female) # 아버지와 딸

# 이상치 제거
outlierTest(m2)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 702 -3.630538         0.00031533      0.14284
```

```
out2 <- subset(female, rownames(female) != "702")

# 정규성 확인
qqPlot(m2) # 그래프 분석
```



```
## 702 928
## 339 449
```

```
# 회귀분석
summary(m2)
```

```
##
## Call:
## lm(formula = childHeight ~ father, data = female)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6234 -1.5047 -0.0767  1.4953  6.7831
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 37.69497    2.62458   14.36   <2e-16 ***
## father       0.38130    0.03787   10.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.131 on 451 degrees of freedom
## Multiple R-squared:  0.1836, Adjusted R-squared:  0.1817
## F-statistic: 101.4 on 1 and 451 DF,  p-value: < 2.2e-16
```

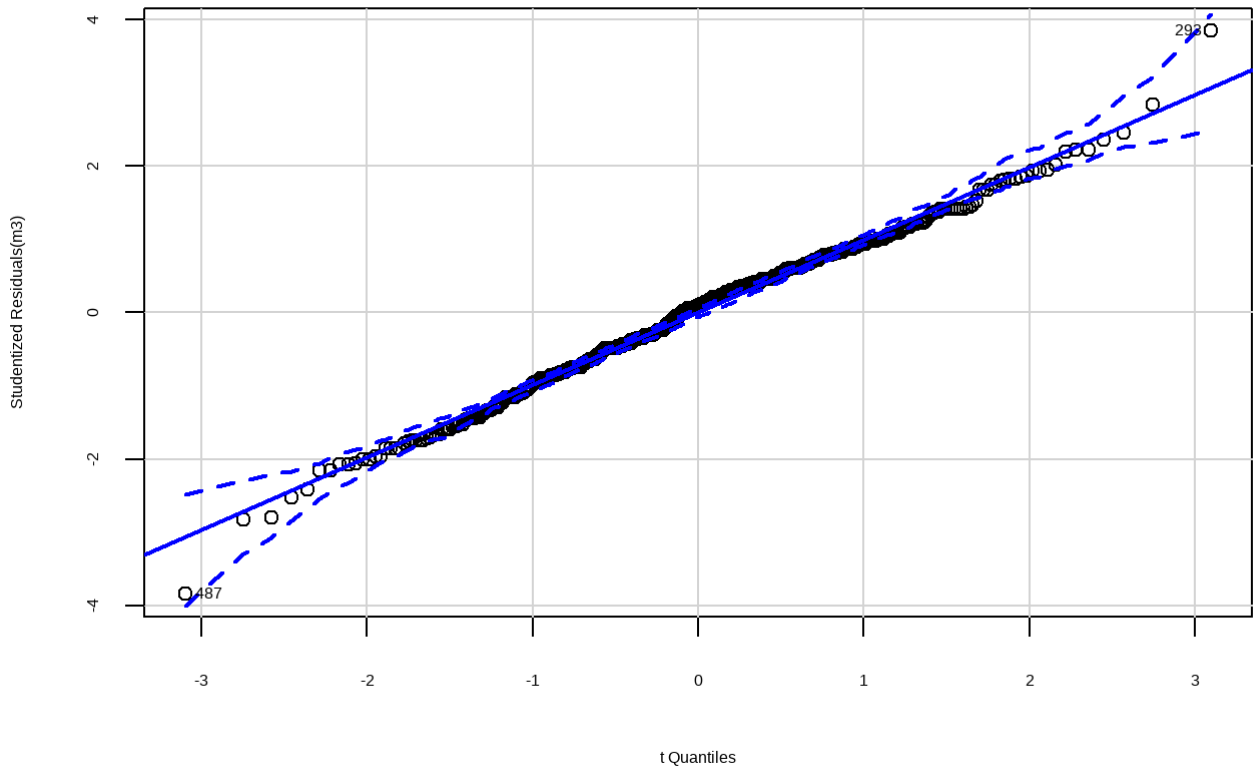
```
# yhat = 37.69497 + 0.3813 * x   # <- 회귀식 (단위: inch)
# |t|통계량 Pr(>|t|) 값(2e-16)이 유의수준 0.05보다 작기 때문에 통계적으로 유의함.
# 잔차 표준오차(Residual standard error): 2.131
# 결정계수(Multiple R-Squared): 0.1836 아버지의 키가 딸의 키에 대하여 18.36% 설명할 수 있다.
# 모양이 적합한지를 보기 위해 F-검정통계량(F-statistic)의 값(101.4)을 보면 유의수준 0.05보다 작기 때문에 모양이 유의하다고 할 수 있다.
# 즉, 아버지의 키는 딸의 키의 약한 영향을 끼친다고 할 수 있다.
```

```
# 어머니와 아들
```

```
# 회귀선
m3 <- lm(childHeight ~ mother, male) # 어머니와 아들

# 이상치 제거
outlierTest(m3)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 293 3.845436         0.00013663     0.065718
```

```
out3 <- subset(male, rownames(male) != "293")

# 정규성 확인
qqPlot(m3) # 그래프 분석
```




```
## 293 487
## 139 245
```

```
# 회귀분석
summary(m3)
```

```
##
## Call:
## lm(formula = childHeight ~ mother, data = male)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4045 -1.6569  0.2305  1.6829  9.4130
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.85804    3.13150   14.64  < 2e-16 ***
## mother       0.36506    0.04887    7.47 3.84e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.486 on 479 degrees of freedom
## Multiple R-squared:  0.1043, Adjusted R-squared:  0.1025
## F-statistic:  55.8 on 1 and 479 DF,  p-value: 3.838e-13
```

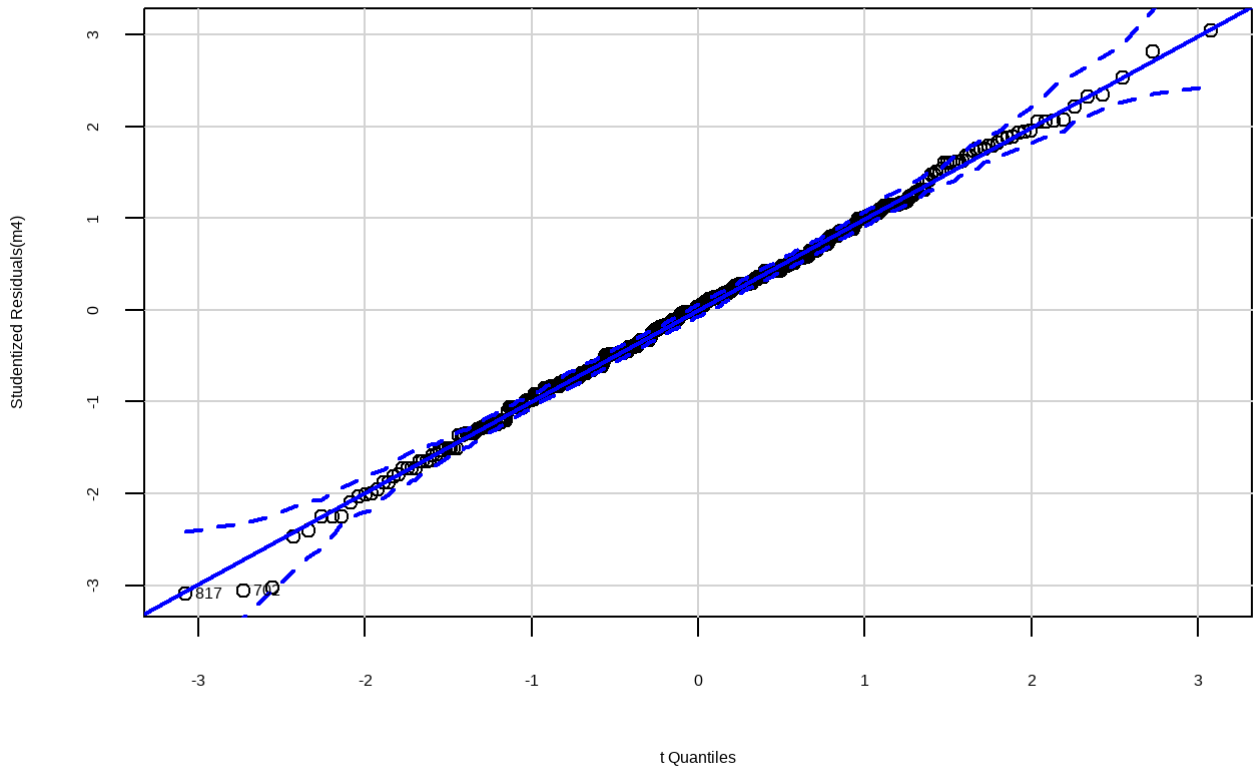
```
# yhat = 45.85804 + 0.36506 * x   # <- 회귀식 (단위: inch)
# |t|통계량 Pr(>|t|) 값(2e-16, 3.84e-13)이 유의수준 0.05보다 작기 때문에 통계적으로 유의함.
# 잔차 표준오차(Residual standard error): 2.486
# 결정계수(Multiple R-Squared): 10.43 어머니의 키가 아들의 키에 대하여 10.43% 설명할 수 있다.
# 모양이 적합한지를 보기 위해 F-검정통계량(F-statistic)의 값(55.8)을 보면 유의수준 0.05보다 작기 때문에 모양이 유의하다고 할 수 있다.
# 즉, 어머니의 키는 아들의 키의 약한 영향을 끼친다고 할 수 있다.
```

```
# 어머니와 딸
```

```
# 회귀선
m4 <- lm(childHeight ~ mother, female) # 어머니와 딸

# 이상치 제거
outlierTest(m4)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##      rstudent unadjusted p-value Bonferroni p
## 817 -3.094112           0.002097      0.94995
```

```
out4 <- subset(female, rownames(female) != "817")

# 정규성 확인
qqPlot(m4) # 그래프 분석
```



```
## 702 817
## 339 392
```

```
# 회귀분석
summary(m4)
```

```
##
## Call:
## lm(formula = childHeight ~ mother, data = female)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.8749 -1.5340  0.0799  1.4434  6.7616
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 43.68897    3.00171  14.555  < 2e-16 ***
## mother       0.31824    0.04676   6.805 3.22e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.246 on 451 degrees of freedom
## Multiple R-squared:  0.09313,    Adjusted R-squared:  0.09111
## F-statistic: 46.31 on 1 and 451 DF,  p-value: 3.222e-11
```

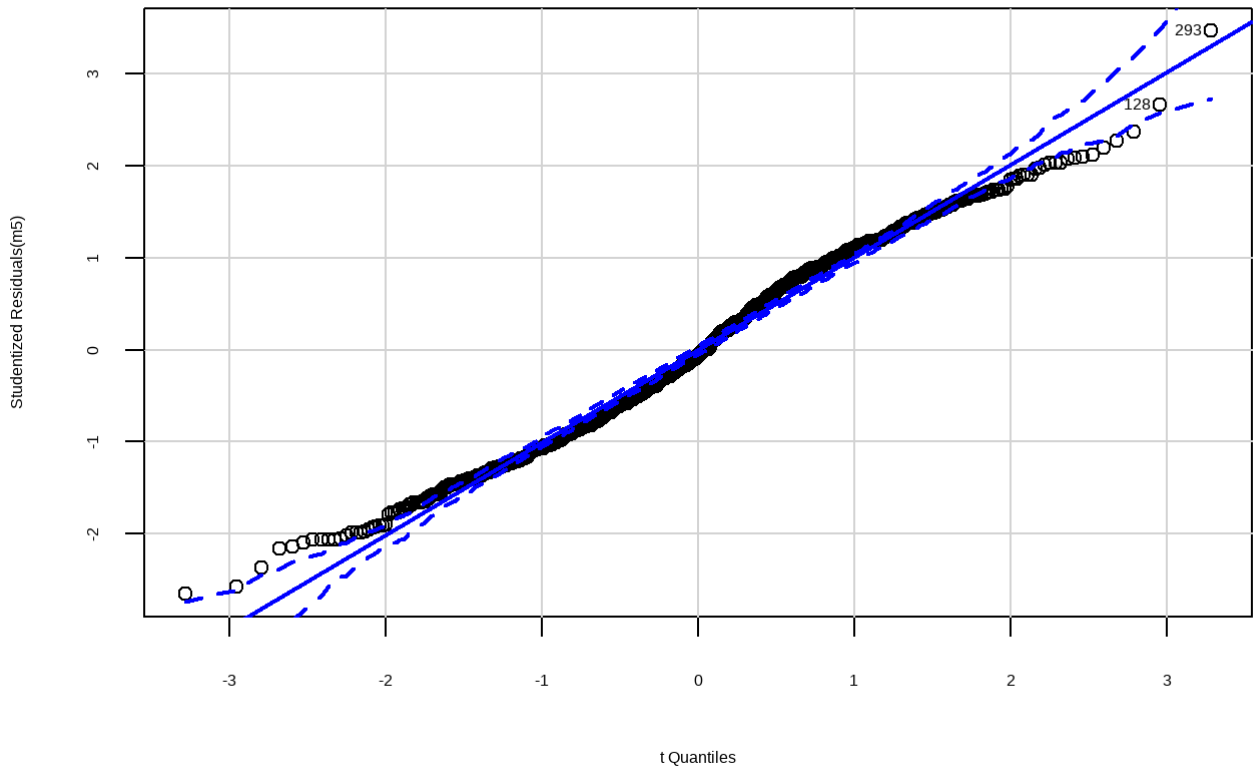
```
# yhat = 43.68897 + 0.31824 * x   # <- 회귀식 (단위: inch)
# |t|통계량 Pr(>|t|) 값(2e-16, 3.22e-11)이 유의수준 0.05보다 작기 때문에 통계적으로 유의함.
# 잔차 표준오차(Residual standard error): 2.246
# 결정계수(Multiple R-Squared): 0.09313 아버지의 키가 딸의 키에 대하여 9.313% 설명할 수 있다.
# 모양이 적합한지를 보기 위해 F-검정통계량(F-statistic)의 값(46.31)을 보면 유의수준 0.05보다 작기 때문에 모양이 유의하다고 할 수 있다.
# 즉, 어머니의 키는 딸의 키의 약한 영향을 끼친다고 할 수 있다.
```

```
# 부모님의 평균과 자식들
```

```
# 회귀선
m5 <- lm(childHeight ~ midparentHeight, G) # 부모님의 평균과 자식들

# 이상치 제거
outlierTest(m5)
```

```
## No Studentized residuals with Bonferroni p < 0.05
## Largest |rstudent|:
##     rstudent unadjusted p-value Bonferroni p
## 293 3.467721         0.00054892      0.51269
```

```
out5 <- subset(G, rownames(G) != "293")

# 정규성 확인
qqPlot(m5) # 그래프 분석
```




```
## [1] 128 293
```

```
# 회귀 분석
summary(m5)
```

```
##
## Call:
## lm(formula = childHeight ~ midparentHeight, data = G)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9570 -2.6989 -0.2155  2.7961 11.6848
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     22.63624    4.26511   5.307 1.39e-07 ***
## midparentHeight  0.63736    0.06161  10.345  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.392 on 932 degrees of freedom
## Multiple R-squared:  0.103,  Adjusted R-squared:  0.102
## F-statistic:   107 on 1 and 932 DF,  p-value: < 2.2e-16
```

```
# yhat = 22.63624 + 0.63736 * x   # <- 회귀식 (단위: inch)
# |t|통계량 Pr(>|t|) 값(1.39e-07, 2e-16)이 유의수준 0.05보다 작기 때문에 통계적으로 유의함.
# 잔차 표준오차(Residual standard error): 3.392
# 결정계수(Multiple R-Squared): 0.103 아버지의 키가 딸의 키에 대하여 10.3% 설명할 수 있다.
# 모양이 적합한지를 보기 위해 F-검정통계량(F-statistic)의 값(107)을 보면 유의수준 0.05보다 작기
때문에 모양이 유의하다고 할 수 있다.
# 즉, 부모님의 키는 자식의 키의 약한 영향을 끼친다고 할 수 있다.
```