Data analysis project: Relatioship of family's height

Zhuoyue Wang, Weizhuo Wang, Juhee Chung

August 5, 2017

Contents

STAT 420 Data analysis project: The relationship between the height of parents and kids

We usually assume that if both parents are tall, the children would also be tall. But is the fact like what we expected?

In this project, we want to find out whether this cognition is true, and if it is true, how strong is the effect of parents??? height on their children? If there is a relationship between them, what kind of relationship do they have, linear or logarithmic?

Galton's family heights data has been a preeminent historical dataset in regression analysis, on which the original model and basic results have survived the close scrutiny of statisticians for 132 years. The dataset gives information based on the famous 1885 study of Francis Galton exploring the relationship between the heights of adult children and the heights of their parents. (retrieved from http://www.math.uah.edu/stat/data/Galton.csv).

This file has 898 observations and 6 variables. One of them is categorical variable, and others are numerical variable. It records 898 children's height from 205 families.

The variable information is given below:

Family: The family that the child belongs to, labeled by the numbers from 1 to 204 and 136A

Father: The father's height, in inches

Mother: The mother's height, in inches

Average: The average height between father and mother in a family, in inches

Gender: The gender of the child, male (M) or female (F)

Height: The height of the child, in inches

Kids: The number of kids in the family of the child

```
galton = read.csv("Galton.csv")
summary(galton)
```

```
##      Family        Father          Mother      Gender      Height
##  185    : 15   Min.   :62.00   Min.   :58.00   F:433   Min.   :56.00
##  166    : 11   1st Qu.:68.00   1st Qu.:63.00   M:465   1st Qu.:64.00
##  66     : 11   Median :69.00   Median :64.00           Median :66.50
##  130    : 10   Mean   :69.23   Mean   :64.08           Mean   :66.76
```

```
##  136    : 10   3rd Qu.:71.00   3rd Qu.:65.50           3rd Qu.:69.70
##  140    : 10   Max.   :78.50   Max.   :70.50           Max.   :79.00
##  (Other):831
##       Kids            Average
##  Min.   : 1.000   Min.   :62.00
##  1st Qu.: 4.000   1st Qu.:65.50
##  Median : 6.000   Median :66.80
##  Mean   : 6.136   Mean   :66.67
##  3rd Qu.: 8.000   3rd Qu.:67.50
##  Max.   :15.000   Max.   :72.80
##
```

Simple Linear Regression

```
ParentAverageVsKid = lm(Height ~ Average, data = galton)
summary(ParentAverageVsKid)
```

```
##
## Call:
## lm(formula = Height ~ Average, data = galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -8.9710 -2.6807 -0.1771  2.7789 11.6873
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.15174    4.30266   5.148 3.23e-07 ***
## Average      0.66905    0.06451  10.371  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.387 on 896 degrees of freedom
## Multiple R-squared:  0.1072, Adjusted R-squared:  0.1062
## F-statistic: 107.6 on 1 and 896 DF,  p-value: < 2.2e-16
```

```
anova(ParentAverageVsKid)
```

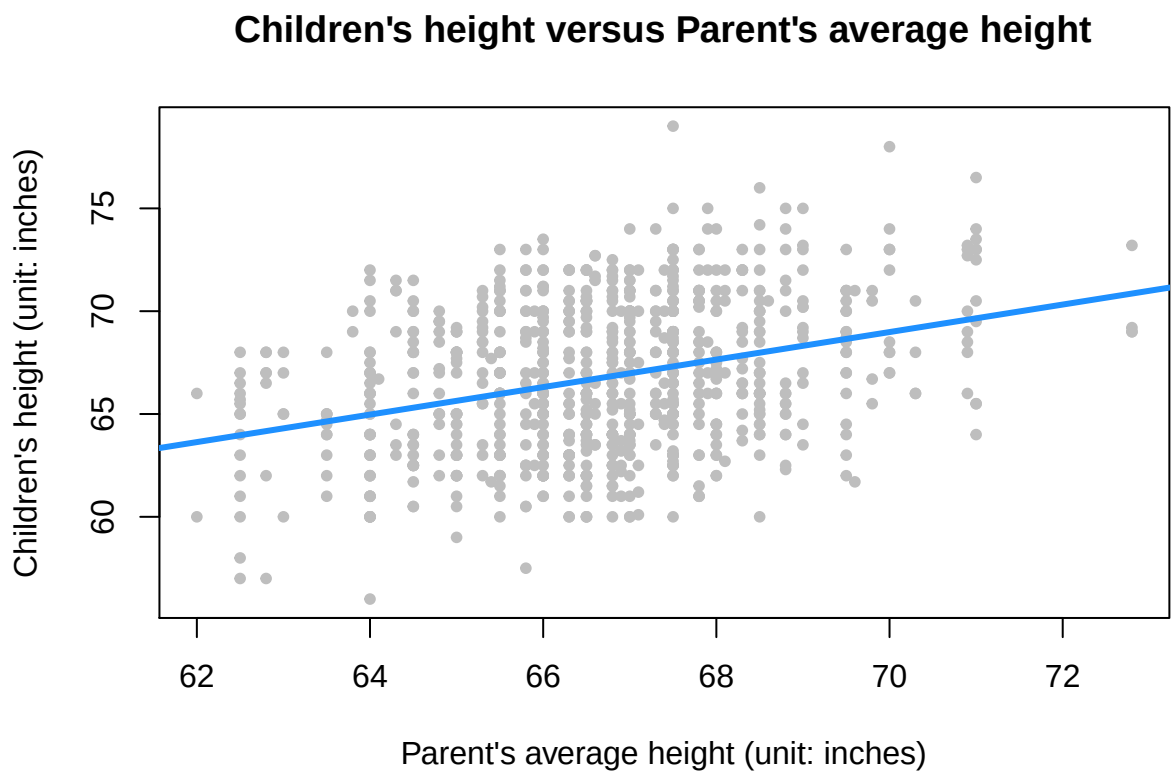
```
## Analysis of Variance Table
##
## Response: Height
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Average     1  1234.2 1234.22  107.56 < 2.2e-16 ***
## Residuals 896 10280.8   11.47
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
ParentAverageVsKidRMSE = sqrt(mean(resid(ParentAverageVsKid)^2))
ParentAverageVsKidRMSE
```

```
## [1] 3.383579
```

We can see that the p-value is nearly equal to zero. So we can have statistical confidence to conclude that there is a linear relationship between parent's average height and childen's height.

```r
plot(Height ~ Average, data = galton,
     xlab = "Parent's average height (unit: inches)",
     ylab = "Children's height (unit: inches)",
     main = "Children's height versus Parent's average height",
     pch  = 20,
     col  = "grey")
abline(ParentAverageVsKid, lwd = 3, col = "dodgerblue")
```




```r
FatherVsKid = lm(Height ~ Father, data = galton)
summary(FatherVsKid)
```

```
##
## Call:
## lm(formula = Height ~ Father, data = galton)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.2683  -2.6689  -0.2092   2.6342  11.9329
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 39.11039    3.22706  12.120   <2e-16 ***
## Father       0.39938    0.04658   8.574   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.446 on 896 degrees of freedom
## Multiple R-squared:  0.07582,    Adjusted R-squared:  0.07479
## F-statistic: 73.51 on 1 and 896 DF,  p-value: < 2.2e-16
```

```r
anova(FatherVsKid)
```

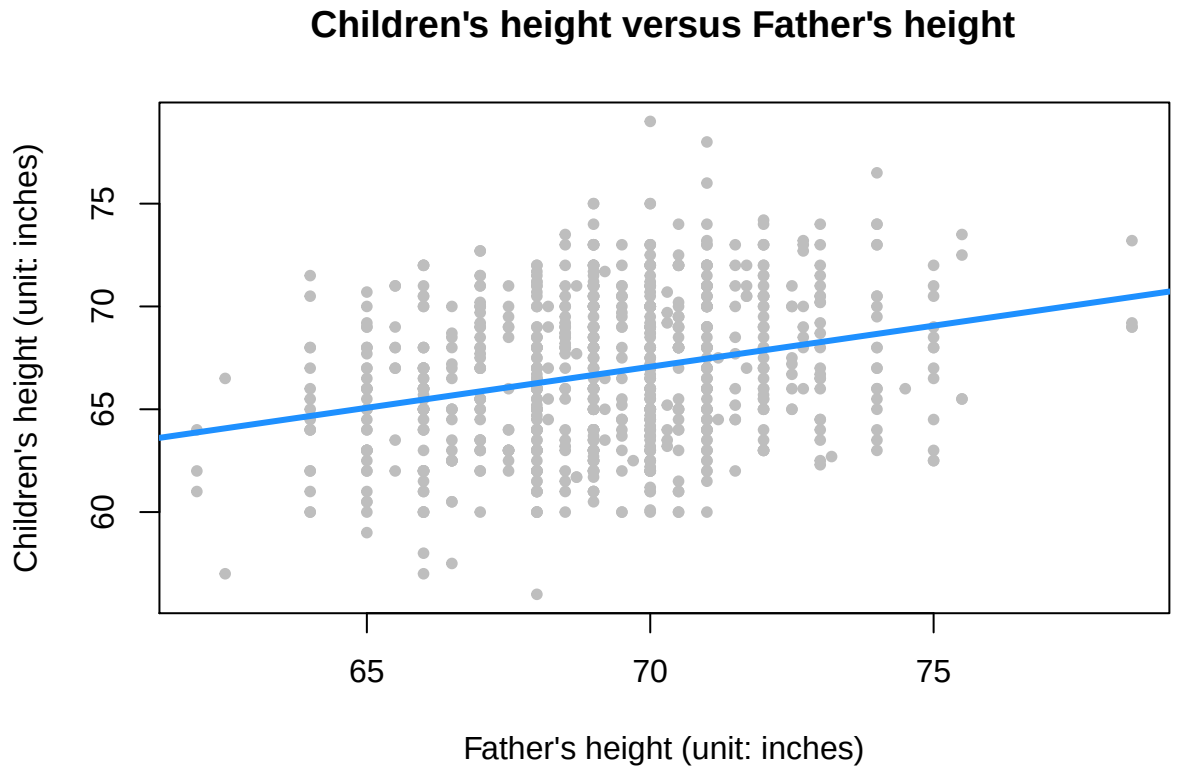
```
## Analysis of Variance Table
##
## Response: Height
##             Df  Sum Sq Mean Sq F value    Pr(>F)
## Father       1   873.1  873.08  73.508 < 2.2e-16 ***
## Residuals 896 10642.0   11.88
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
FatherVsKidRMSE = sqrt(mean(resid(FatherVsKid)^2))
FatherVsKidRMSE
```

```
## [1] 3.442494
```

We can see that the p-value is nearly equal to zero. So we can have statistical confidence to conclude that there is a linear relationship between father's height and childen's height.

```r
plot(Height ~ Father, data = galton,
     xlab = "Father's height (unit: inches)",
     ylab = "Children's height (unit: inches)",
     main = "Children's height versus Father's height",
     pch  = 20,
     col  = "grey")
abline(FatherVsKid, lwd = 3, col = "dodgerblue")
```




```r
MotherVsKid = lm(Height ~ Mother, data = galton)
summary(MotherVsKid)
```

```
##
## Call:
```

```
## lm(formula = Height ~ Mother, data = galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.5474 -2.6346 -0.1079  2.8688 11.9526
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 46.69077    3.25874  14.328  < 2e-16 ***
## Mother       0.31318    0.05082   6.163 1.08e-09 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.511 on 896 degrees of freedom
## Multiple R-squared:  0.04066,    Adjusted R-squared:  0.03959
## F-statistic: 37.98 on 1 and 896 DF,  p-value: 1.079e-09
```

```
anova(MotherVsKid)
```

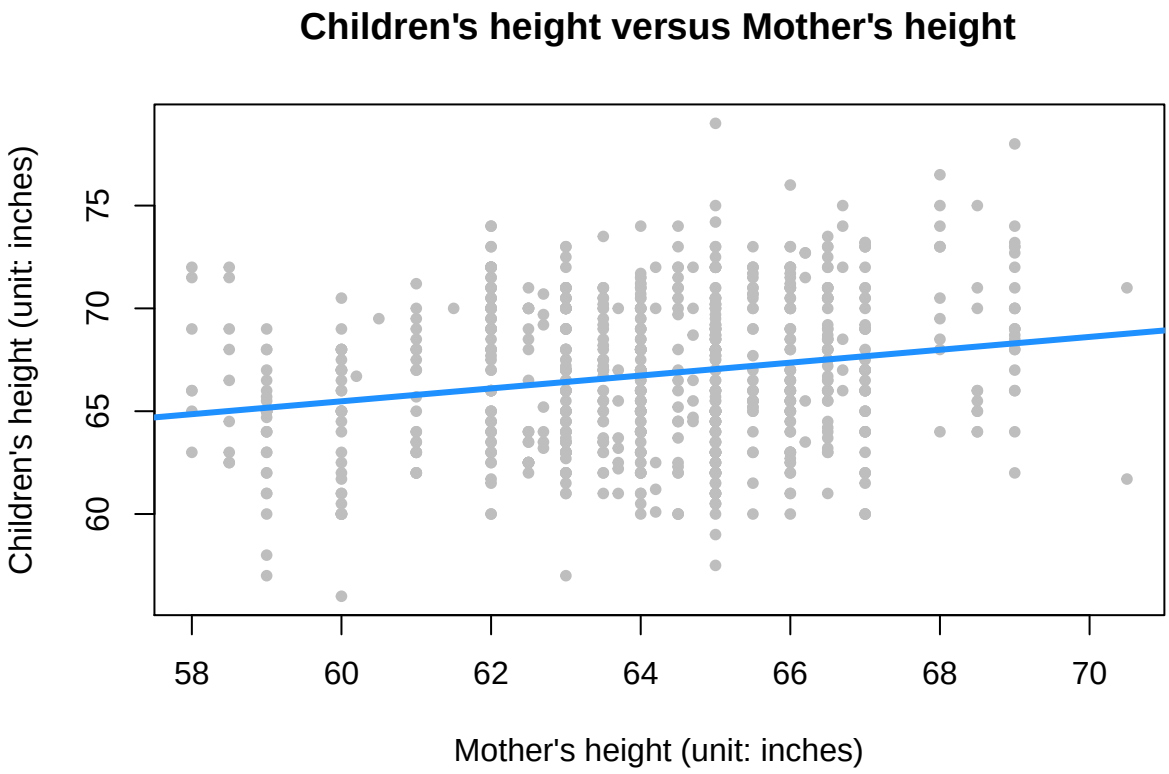
```
## Analysis of Variance Table
##
## Response: Height
##            Df  Sum Sq Mean Sq F value    Pr(>F)
## Mother      1   468.3  468.26   37.98 1.079e-09 ***
## Residuals 896 11046.8   12.33
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
MotherVsKidRMSE = sqrt(mean(resid(MotherVsKid)^2))
MotherVsKidRMSE
```

```
## [1] 3.507359
```

We can see that the p-value is nearly equal to zero. So we can have statistical confidence to conclude that there is a linear relationship between mother's height and childen's height.

```
plot(Height ~ Mother, data = galton,
     xlab = "Mother's height (unit: inches)",
     ylab = "Children's height (unit: inches)",
     main = "Children's height versus Mother's height",
     pch  = 20,
     col  = "grey")
abline(MotherVsKid, lwd = 3, col = "dodgerblue")
```



RMSE represents the square root of the average of the error. In these three models, we can see that the first model(Children's height versus Parent's average height) has the lowest RMSE, which means it is the most precise one. Then the second(Children's height versus Father's height), and the third(Children's height versus Mother's height).

So we can conclude that both parents' height affect the kid's height, and father's height is more influential than mother's height.

Multiple Linear Regression

Choose the "best" model

```
galton=read.csv("Galton.csv")
Gender=as.factor(ifelse(as.character(galton$Gender)=="M", 1,0))
```

```
# full additive model
height_add=lm(Height ~ .-Family, data=galton)

# backward AIC selection
height_add_aic=step(height_add, direction="backward", trace=0)

# backward BIC selection
n=length(resid(height_add))
height_add_bic=step(height_add, direction="backward", k=log(n), trace=0)

# choose between AIC and BIC
anova(height_add_aic, height_add_bic)
```

```
## Analysis of Variance Table
##
```

```
## Model 1: Height ~ Mother + Gender + Kids + Average
## Model 2: Height ~ Gender + Average
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    893 4135.4
## 2    895 4164.6 -2   -29.237 3.1567 0.04304 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Fail to reject H0, we prefer BIC selected model.

```
# choose between BIc selected model and original additive model
anova(height_add_bic, height_add)
```

```
## Analysis of Variance Table
##
## Model 1: Height ~ Gender + Average
## Model 2: Height ~ (Family + Father + Mother + Gender + Kids + Average) -
##     Family
##   Res.Df    RSS Df Sum of Sq      F  Pr(>F)
## 1    895 4164.6
## 2    892 4133.8  3    30.854 2.2192 0.08443 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- Fail to reject H0, we still prefer BIC selected model.

```
# full interaction model
height_int=lm(Height ~ Father*Mother*Gender*Kids, data=galton)
summary(height_int)
```

```
##
## Call:
## lm(formula = Height ~ Father * Mother * Gender * Kids, data = galton)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -9.4458 -1.4432  0.1172  1.4473  9.1189
##
## Coefficients:
##                               Estimate Std. Error t value Pr(>|t|)
## (Intercept)                  1.365e+01  1.702e+02   0.080    0.936
## Father                       4.613e-01  2.467e+00   0.187    0.852
## Mother                       4.426e-01  2.644e+00   0.167    0.867
## GenderM                     -3.507e+02  2.564e+02  -1.367    0.172
## Kids                         1.478e+01  2.491e+01   0.593    0.553
## Father:Mother               -2.206e-03  3.831e-02  -0.058    0.954
## Father:GenderM               5.354e+00  3.730e+00   1.435    0.152
## Mother:GenderM               5.466e+00  4.003e+00   1.365    0.172
## Father:Kids                 -2.162e-01  3.668e-01  -0.590    0.556
## Mother:Kids                 -2.456e-01  3.898e-01  -0.630    0.529
## GenderM:Kids                 6.036e+01  3.877e+01   1.557    0.120
## Father:Mother:GenderM       -8.218e-02  5.819e-02  -1.412    0.158
## Father:Mother:Kids           3.588e-03  5.735e-03   0.626    0.532
## Father:GenderM:Kids         -9.163e-01  5.711e-01  -1.604    0.109
## Mother:GenderM:Kids         -9.291e-01  6.085e-01  -1.527    0.127
## Father:Mother:GenderM:Kids   1.409e-02  8.953e-03   1.574    0.116
##
```

```
## Residual standard error: 2.148 on 882 degrees of freedom
## Multiple R-squared:  0.6465, Adjusted R-squared:  0.6405
## F-statistic: 107.5 on 15 and 882 DF,  p-value: < 2.2e-16
```

```
# backward AIC selection
height_int_aic=step(height_int, direction="backward", trace=0)

# backward BIC selection
n=length(resid(height_int))
height_int_bic=step(height_int, direction="backward", k=log(n), trace=0)

# choose between AIC and BIC
anova(height_int_aic, height_int_bic)
```

```
## Analysis of Variance Table
##
## Model 1: Height ~ Father * Mother * Gender * Kids
## Model 2: Height ~ Father + Mother + Gender
##   Res.Df    RSS  Df Sum of Sq      F Pr(>F)
## 1    882 4070.9
## 2    894 4149.2 -12   -78.281 1.4134 0.1536
```

- Fail to reject H0, we prefer BIC selected model.

```
# choose between BIc selected model and original interaction model
anova(height_int_bic, height_int)
```

```
## Analysis of Variance Table
##
## Model 1: Height ~ Father + Mother + Gender
## Model 2: Height ~ Father * Mother * Gender * Kids
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    894 4149.2
## 2    882 4070.9 12    78.281 1.4134 0.1536
```

- Fail to reject H0, we prefer BIC selected model. (This model is same as additive BIC selected model)

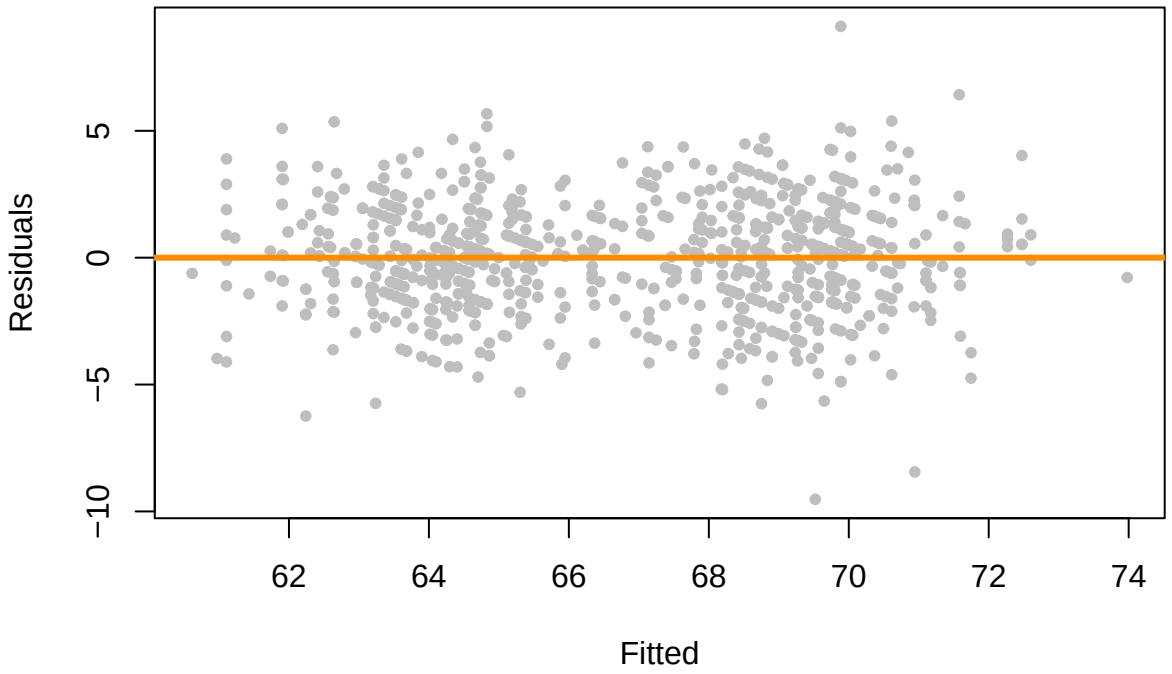
```
# rename the chosen model for future use
height_model=height_int_bic
summary(height_model)
```

```
##
## Call:
## lm(formula = Height ~ Father + Mother + Gender, data = galton)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -9.523 -1.440  0.117  1.473  9.114
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 15.34476    2.74696   5.586 3.08e-08 ***
## Father       0.40598    0.02921  13.900  < 2e-16 ***
## Mother       0.32150    0.03128  10.277  < 2e-16 ***
## GenderM      5.22595    0.14401  36.289  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 2.154 on 894 degrees of freedom
## Multiple R-squared:  0.6397, Adjusted R-squared:  0.6385
## F-statistic:   529 on 3 and 894 DF,  p-value: < 2.2e-16
```

** Check if the model violates some assumptions **

```
plot(fitted(height_model), resid(height_model), col = "grey", pch = 20, xlab = "Fitted", ylab = "Residua
abline(h = 0, col = "darkorange", lwd = 3)
```



- The plot does not seem "normal".

```
# homoscedasticity
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

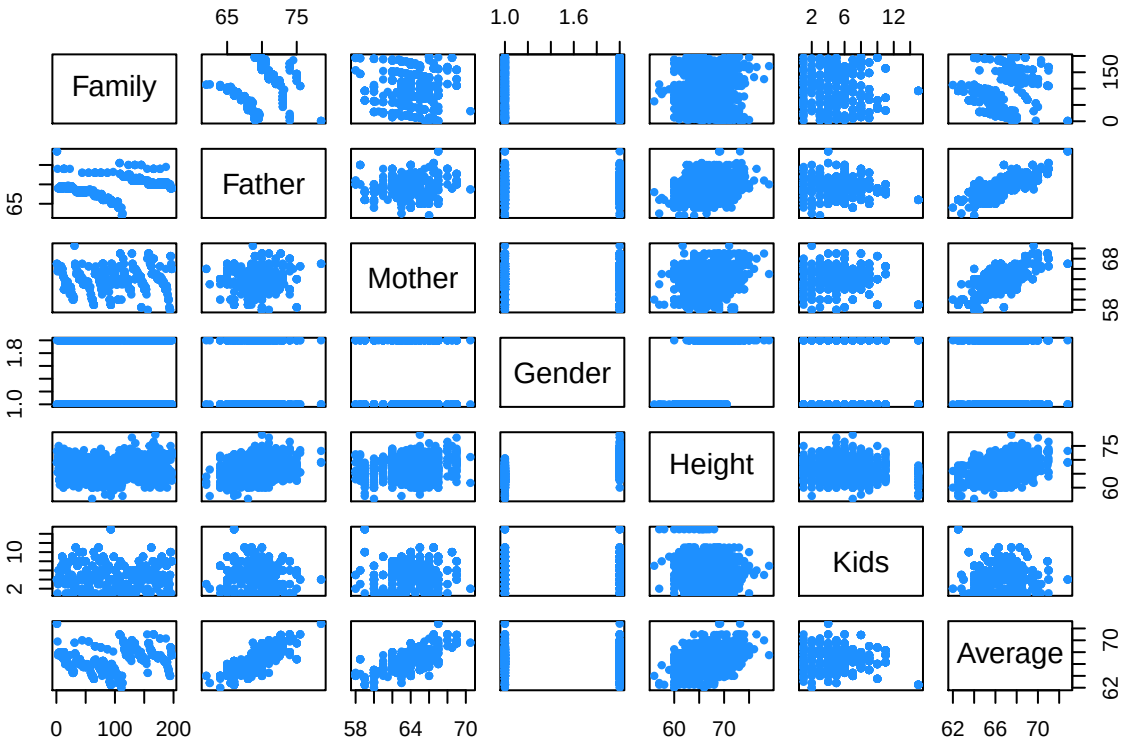
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(height_model)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  height_model
## BP = 7.5002, df = 3, p-value = 0.05755
```

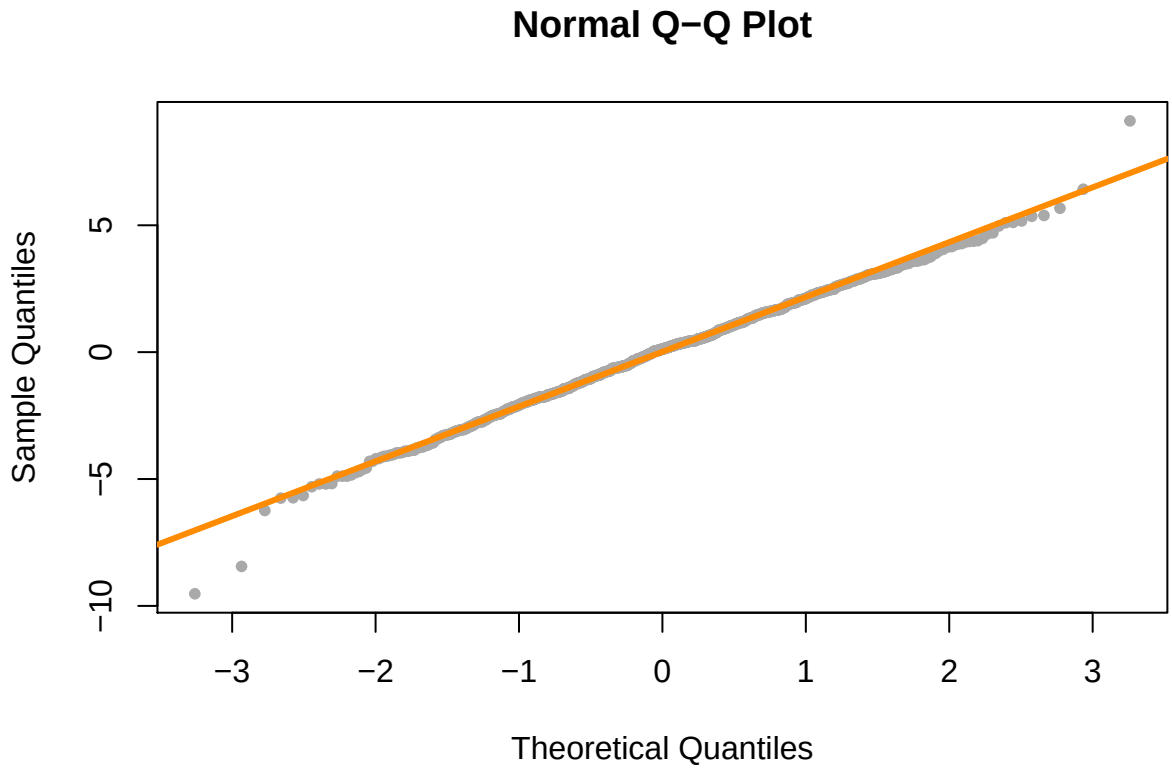
- p-value > 0.05
- We do not reject H0, the model is good with homoscedasticity assumption.

```
# multicolinearity
pairs(galton, col = "dodgerblue", pch=20)
```



- There is no serious collinearity problem.

```
# normal distribution
qqnorm(resid(height_model), col = "darkgrey", pch=20)
qqline(resid(height_model), col = "darkorange", lwd = 3)
```



```r
shapiro.test(resid(height_model))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(height_model)
## W = 0.99569, p-value = 0.01305
```

- 0.01 < p-value < 0.05
- Reject H0 at 0.05 level, but do not reject at 0.01 level.
- The residuals approximately follow a normal distribution.

** Make prediction **

```r
# A boy kid in a 4-kids family with mother 64 inches tall and father 73 inches tall

pred=data.frame(Father=73, Mother=62, Gender="M", Kids=4)
predict(height_model, pred, interval="prediction", level=0.95)
```

```
##        fit      lwr      upr
## 1 70.13981 65.89919 74.38042
```

```r
conf=data.frame(Father=73, Mother=62, Gender="M", Kids=4)
predict(height_model, conf, interval="confidence", level=0.95)
```

```
##        fit      lwr      upr
## 1 70.13981 69.81457 70.46505
```

- The predicted height in this case is 70.13981
- The prediction interval is [65.89919, 74.38042], the confidence interval is [69.81457, 70.46505].

Logarithmic Regression

```r
model_1=lm(Height~Father +Mother + Gender, data = galton)
model_log = lm(log(Height)~ log(Father) + log(Mother) + Gender + log(Kids), data =galton)

anova(model_log)
```

```
## Analysis of Variance Table
##
## Response: log(Height)
##              Df  Sum Sq Mean Sq   F value Pr(>F)
## log(Father)   1 0.19869 0.19869  190.5477 <2e-16 ***
## log(Mother)   1 0.08402 0.08402   80.5773 <2e-16 ***
## Gender        1 1.37263 1.37263 1316.3611 <2e-16 ***
## log(Kids)     1 0.00181 0.00181    1.7386 0.1877
## Residuals   893 0.93117 0.00104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#exclude Kids
#changer Gender variable to numeric

Gender = factor(ifelse(as.character(galton$Gender)=="M",1,0))
Gender
```

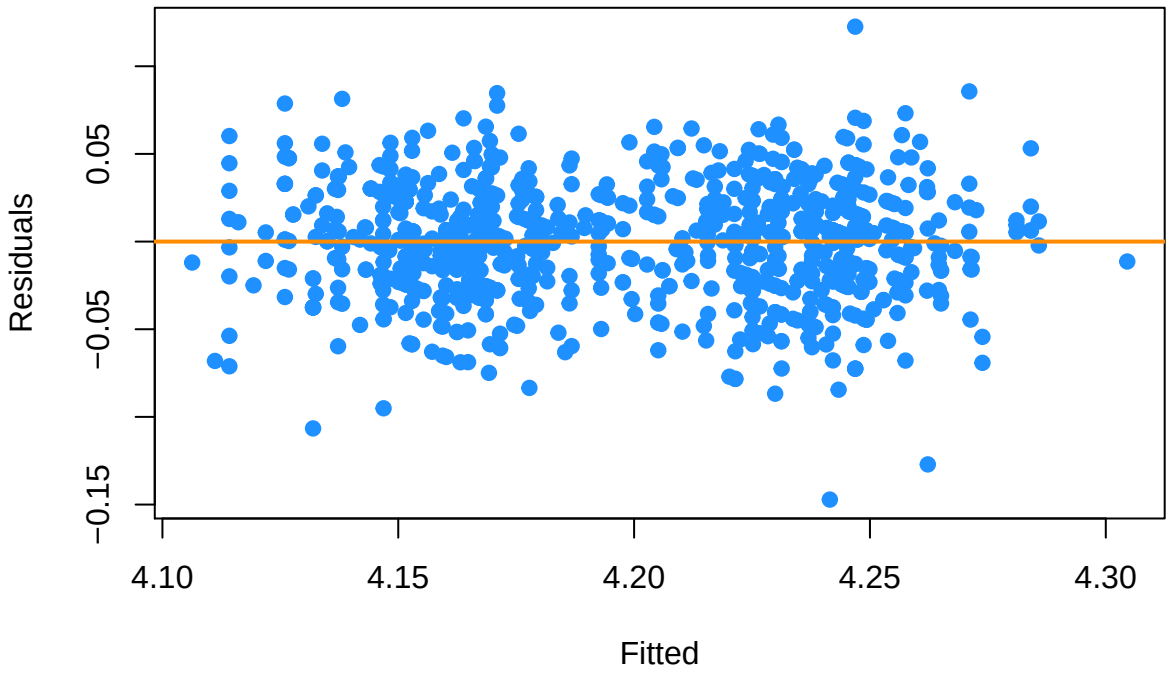
```
##   [1] 1 0 0 0 1 1 0 0 1 0 1 1 0 0 0 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 1 1
```

```
##  [36] 0 0 0 0 0 0 0 1 1 1 1 0 1 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0 0 0 0 1 1 1 0
##  [71] 0 0 0 0 1 0 0 1 1 0 1 1 1 0 0 0 0 0 0 0 1 1 1 0 0 1 1 0 1 1 0 0 0 0 0
## [106] 0 0 0 1 1 1 0 0 0 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 1 1 1 0 1 0 0 0 1 1
## [141] 1 0 0 0 1 0 1 1 1 1 0 1 1 1 1 1 0 0 1 0 1 1 1 1 0 0 0 0 0 0 0 0 0 1 1
## [176] 1 0 1 1 1 1 1 1 0 0 0 0 0 0 1 1 0 0 0 1 1 1 1 0 0 0 0 0 1 1 1 0 1 1 0
## [211] 0 0 1 1 1 0 0 1 0 0 0 0 1 1 1 1 1 0 0 0 1 1 1 1 1 0 1 1 0 0 0 0 1 1 1
## [246] 0 0 0 0 1 1 1 1 0 0 0 0 0 0 0 1 1 1 1 0 0 0 0 0 1 1 1 1 1 0 0 0 1 1 0
## [281] 0 0 1 1 0 0 0 0 1 1 1 0 0 0 0 1 1 0 1 1 1 1 1 0 0 0 0 1 1 1 1 1 1 0 1
## [316] 1 1 0 1 1 0 0 0 1 1 1 1 1 1 0 0 0 1 0 0 0 1 1 1 1 1 1 0 0 0 1 1 1 0 0
## [351] 0 0 0 1 1 1 0 0 1 1 0 0 1 1 0 0 1 1 0 0 1 1 1 1 1 1 0 0 0 1 1 1 1 0 0 0
## [386] 1 1 0 1 1 1 1 1 0 0 0 0 1 1 0 1 1 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 1 1 1 0
## [421] 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0 1 1 1 1 0 1 1 0 0 1 0 0 0 0 0 1 1
## [456] 1 1 0 0 0 1 1 1 1 0 0 0 0 0 1 1 1 0 0 0 0 1 1 1 0 0 0 0 1 1 0 0 1 0 0
## [491] 1 1 1 1 1 0 0 1 1 1 0 0 0 0 1 0 0 0 1 1 1 1 1 1 1 0 1 1 1 1 0 0 0 0 1
## [526] 1 0 0 1 1 1 0 0 1 1 1 1 1 1 0 0 0 1 0 0 1 1 0 0 1 1 0 1 1 0 1 1 1 1 1
## [561] 1 1 1 1 0 1 1 1 0 1 1 1 1 1 0 0 0 1 1 0 0 1 1 1 1 0 0 0 0 1 1 1 1 0 0 0
## [596] 0 0 0 1 1 0 0 1 1 1 1 0 0 1 1 1 0 0 0 0 0 0 0 1 1 1 1 1 0 0 0 1 1 1 0
## [631] 1 1 1 0 0 1 1 1 1 1 1 0 0 0 1 1 1 0 0 0 1 1 1 1 1 1 0 1 1 0 1 1 1 1 0 0
## [666] 1 1 0 0 0 0 0 0 0 1 1 1 0 1 1 1 1 1 1 1 0 0 0 0 0 1 1 1 0 0 1 1 1 0 0 0 0
## [701] 1 1 0 0 0 1 1 0 0 1 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 0
## [736] 1 1 0 1 1 1 1 1 0 1 1 1 1 1 1 0 0 0 1 1 1 0 0 0 0 0 0 1 1 0 0 0 1 1 0 0
## [771] 0 0 1 1 1 1 1 0 0 0 0 1 1 1 0 0 1 0 0 1 1 1 0 0 0 1 1 0 0 0 0 0 1 1 1 1
## [806] 0 1 1 1 1 1 1 1 1 1 0 0 0 0 0 0 0 0 0 1 1 1 0 0 1 0 0 0 1 1 1 0 0 0 0 0 0
## [841] 1 0 1 1 1 1 1 0 0 1 1 0 0 0 0 1 0 1 1 0 1 1 1 0 1 1 1 1 0 1 1 0 0 0 0 0
## [876] 1 1 0 0 0 0 0 0 1 1 0 1 0 0 1 0 1 1 1 1 1 1 0 0 0
## Levels: 0 1
```

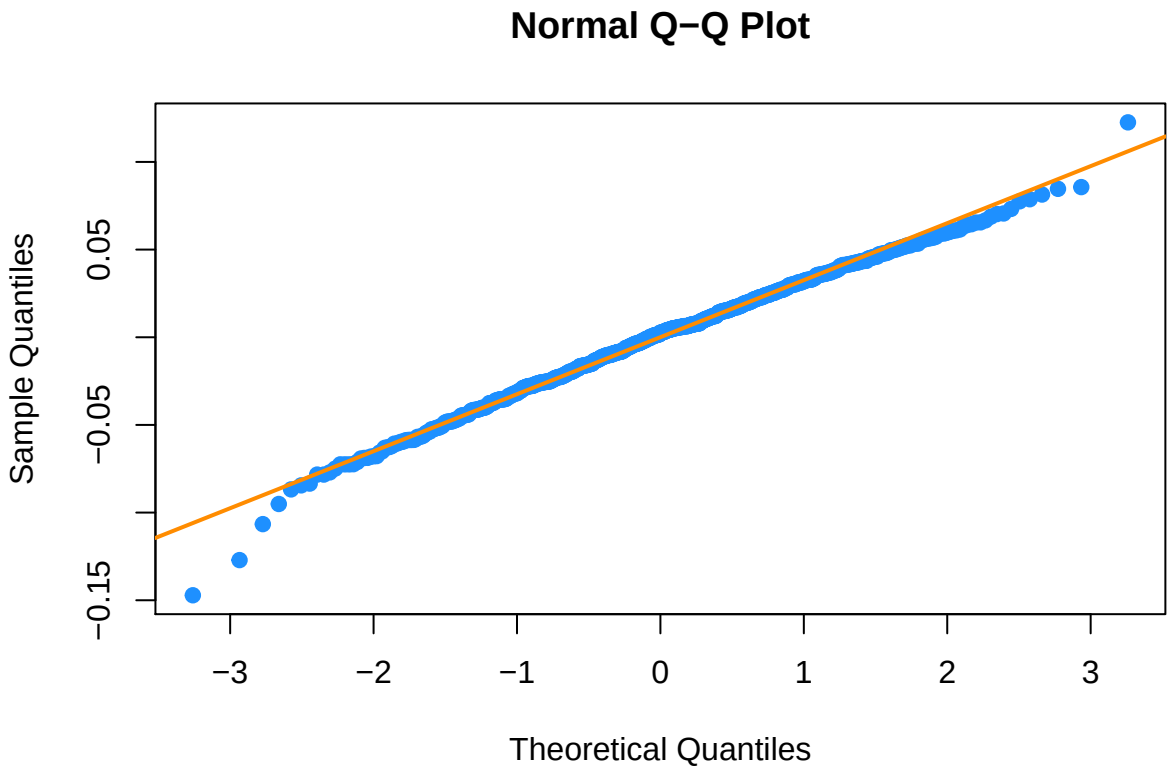
```
plot_fitted_resid = function(model, pointcol = "dodgerblue", linecol = "darkorange") {
  plot(fitted(model), resid(model),
       col = pointcol, pch = 20, cex = 1.5,
       xlab = "Fitted", ylab = "Residuals")
  abline(h = 0, col = linecol, lwd = 2)
}
plot_qq = function(model, pointcol = "dodgerblue", linecol = "darkorange") {
  qqnorm(resid(model), col = pointcol, pch = 20, cex = 1.5)
  qqline(resid(model), col = linecol, lwd = 2)
}
```

```
model_2 = lm(log(Height)~ log(Father) + log(Mother) + Gender, data=galton)

plot_fitted_resid(model_2)
```




```
plot_qq(model_2)
```

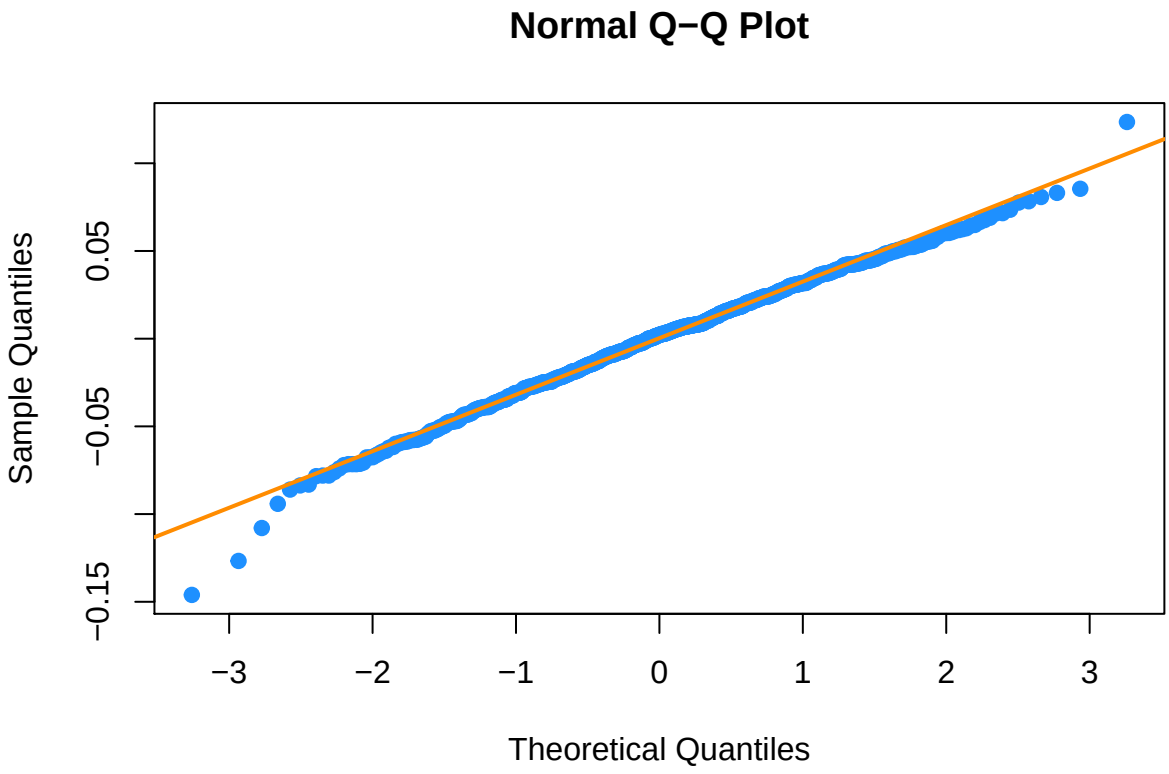



```
anova(model_2)
```

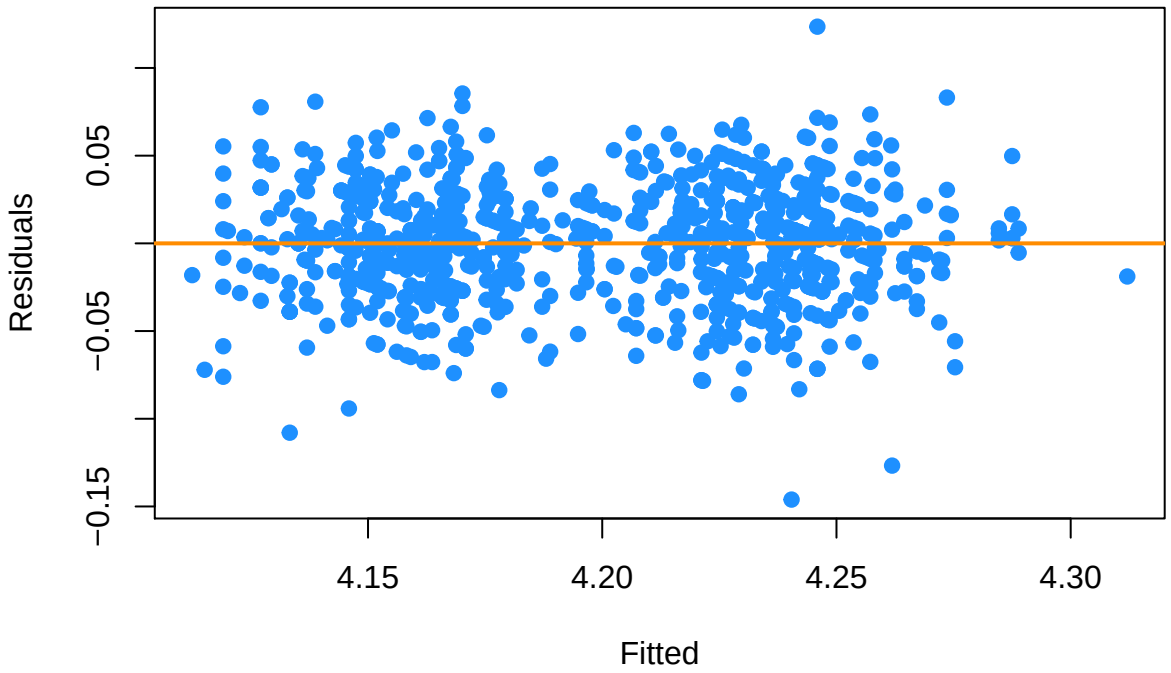
```
## Analysis of Variance Table
##
## Response: log(Height)
##              Df  Sum Sq Mean Sq  F value    Pr(>F)
## log(Father)   1 0.19869 0.19869  190.390 < 2.2e-16 ***
```

```
## log(Mother)    1 0.08402 0.08402   80.511 < 2.2e-16 ***
## Gender         1 1.37263 1.37263 1315.274 < 2.2e-16 ***
## Residuals    894 0.93298 0.00104
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
model_3 = lm(log(Height)~ log(Father) + log(Mother) + Gender + Father*Mother*Gender, data=galton)
plot_qq(model_3)
```




```
plot_fitted_resid(model_3)
```



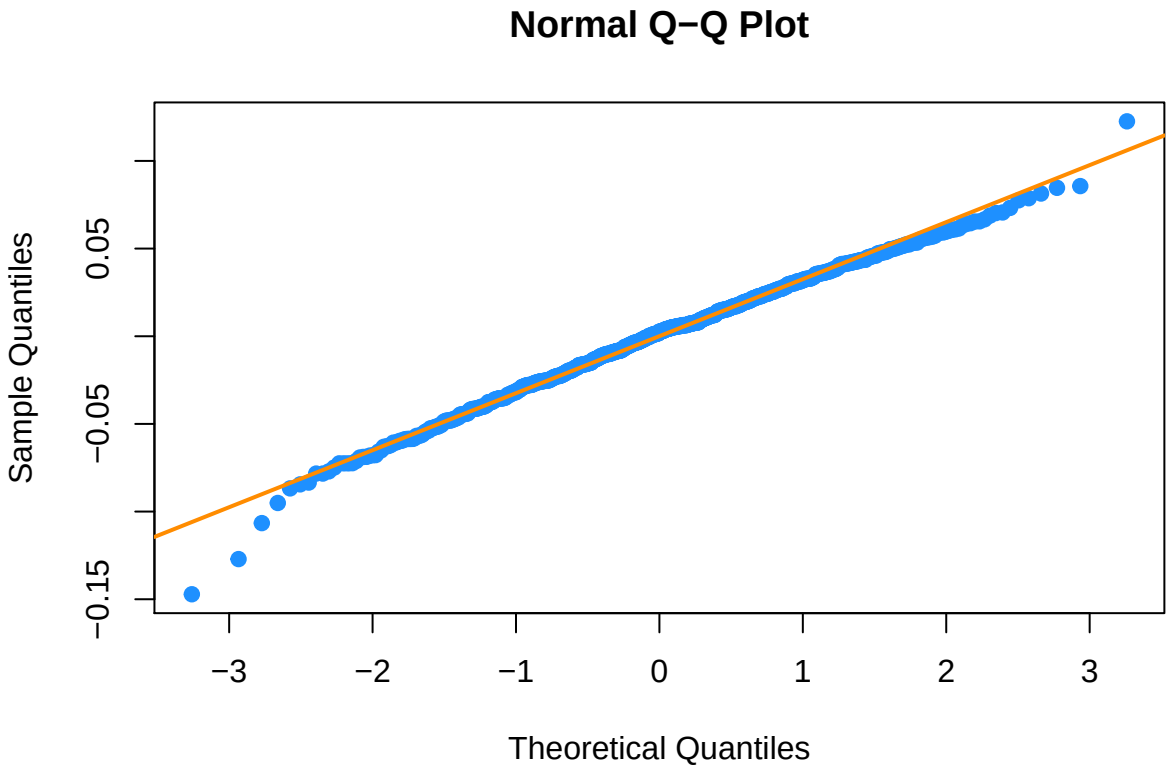

```
anova(model_3)
```

```
## Analysis of Variance Table
##
## Response: log(Height)
##                        Df  Sum Sq Mean Sq   F value Pr(>F)
## log(Father)             1 0.19869 0.19869  189.5888 <2e-16 ***
## log(Mother)             1 0.08402 0.08402   80.1718 <2e-16 ***
## Gender                  1 1.37263 1.37263 1309.7368 <2e-16 ***
## Father                  1 0.00093 0.00093    0.8869 0.3466
## Mother                  1 0.00082 0.00082    0.7868 0.3753
## Father:Mother           1 0.00045 0.00045    0.4300 0.5122
## Gender:Father           1 0.00008 0.00008    0.0771 0.7813
## Gender:Mother           1 0.00000 0.00000    0.0037 0.9517
## Gender:Father:Mother    1 0.00005 0.00005    0.0514 0.8206
## Residuals             888 0.93064 0.00105
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(model_2,model_3)
```

```
## Analysis of Variance Table
##
## Model 1: log(Height) ~ log(Father) + log(Mother) + Gender
## Model 2: log(Height) ~ log(Father) + log(Mother) + Gender + Father * Mother *
##     Gender
##   Res.Df     RSS Df Sum of Sq      F Pr(>F)
## 1    894 0.93298
## 2    888 0.93064  6 0.0023434 0.3727 0.8965
```

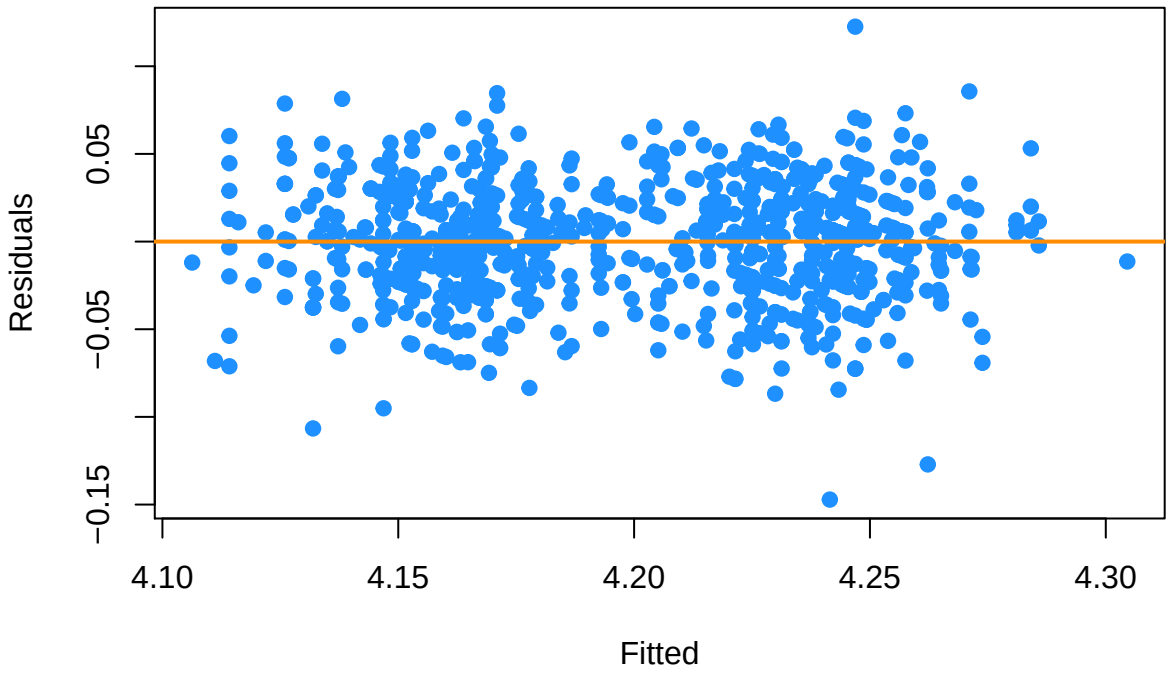
```
#I choose model_2,
plot_qq(model_2)
```



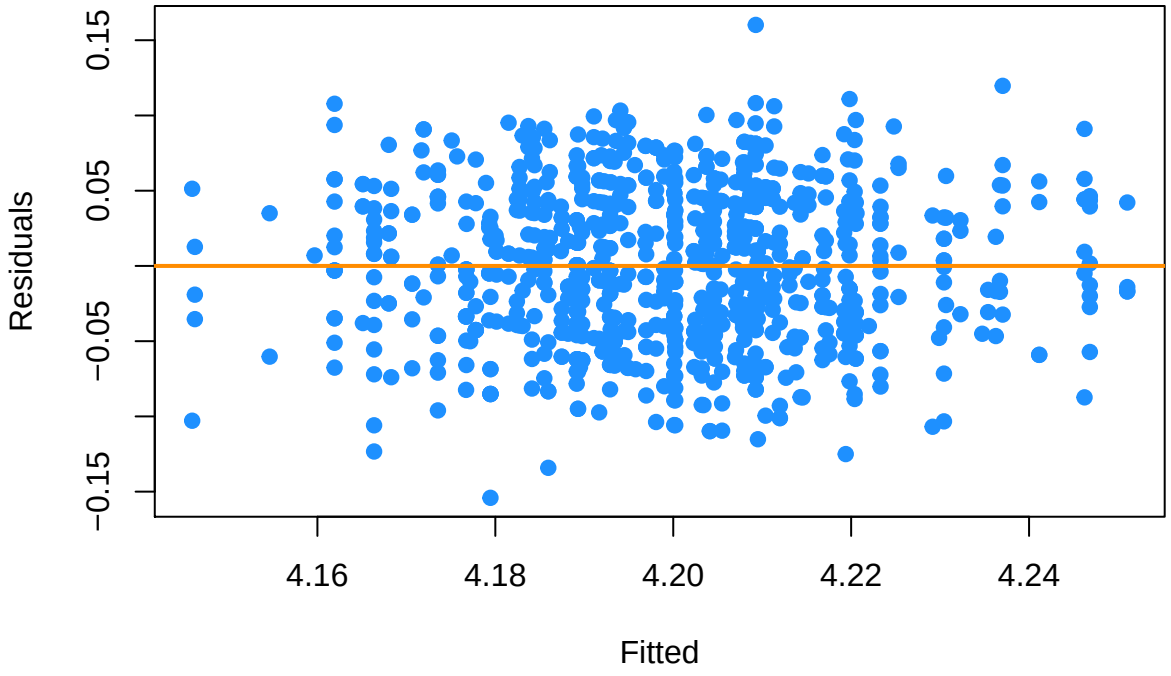

```
shapiro.test(resid(model_2))
```

```
##
##  Shapiro-Wilk normality test
##
## data:  resid(model_2)
## W = 0.99439, p-value = 0.002018
```

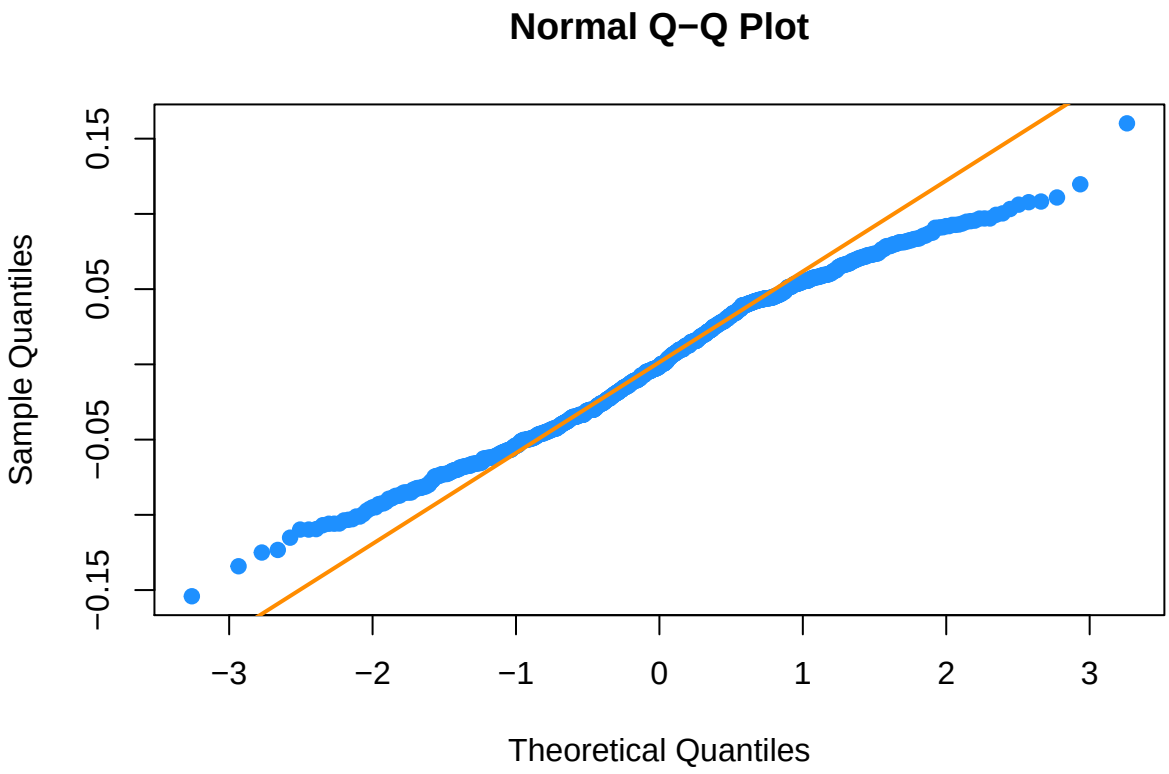
```
#roughly the points fall very close to the line in the QQ plot, and Shapiro test did not reject, I beli
plot_fitted_resid(model_2)
```




```
#it seems that linear assumption is violated, constant variance assumption also violated
model_4 = lm(log(Height)~log(Father)* log(Mother) +  I(Father ^ 2)+I(Mother ^ 2) , data = galton )
plot_fitted_resid(model_4)
```




```
plot_qq(model_4)
```




```
#removed non constant variance bit.
```

Conclusion

Because in Logarithmic regression we use log(height) as the response variable, we can't use anova to compare these models created by different types of regression.

Instead of anova, we use adjusted R square to compare the useness of different models.

```
summary(ParentAverageVsKid)$adj.r.squared
```

```
## [1] 0.1061863
```

```
summary(height_model)$adj.r.squared
```

```
## [1] 0.6384661
```

```
summary(model_2)$adj.r.squared
```

```
## [1] 0.6383321
```

We can see that the results got from Multiple Linear Regression and Logarthemic Regression are much higher than that from Simple linear Regression. In additon, these two values are nearly same. So We believe that Multiple Linear Regression and Logarthemic Regression both can explain the relationship between the height of parents and that of kids.