Linear Regression Modelling Assignment

Mohd Azmi

31/10/2019

Pre-amble

Problem Background

A team of researchers wanted to find relationship between weight and head circumference among newborn. The researchers also wanted to know is there any other factors that can be used to predict the head circumference.

In this dataset, 550 terms babies were sampled from a population. their weight, length and head circumference were recorded at one month of age. The babies also had their gender and parity recorded. (Parity is their birth order in the family). Mother's education level were also recorded.

Dataset Source: Medical Statistics: A Guide to Data Analysis and Critical Appraisal

Loading Library

```
library(foreign)
library(tidyverse)
library(corrplot)
library(knitr)
library(broom)
library(tidyr)
library(psych)
library(rmdformats)
library(car)
```

Loading Dataset

```
Data01 <- read.spss("weights.sav", use.value.label=T, to.data.frame=T)
```

Descriptive Analysis and Data Exploration

In the first step, data exploration was done to get the overview of the data.

Descriptive Analysis

Descriptive analysis was done to the dataset to get the overview of the data, including statistic for each variables.

```
kable(summary(Data01[, 5:6]))
```

gender	educatio

gender	educatio
Male :275	primary :198
Female:275	secondary: 99
NA	tertiary :253

```
kable(describe(Data01))
```

	vars	n	mean	sd	median	trimmed	mad	min	max	range	
id*	1	550	275.500000	158.9155960	275.50	275.500000	203.8575000	1.00	550.00	549.000000	0
weight	2	550	4.366418	0.6018215	4.33	4.356341	0.6226921	2.92	6.49	3.570000	0
length	3	550	54.840909	2.4069078	55.00	54.789773	2.5945500	48.00	62.00	14.000000	0
headc	4	550	37.895273	1.3615643	38.00	37.905909	1.4826000	34.00	41.20	7.200001	-0
gender*	5	550	1.500000	0.5004552	1.50	1.500000	0.7413000	1.00	2.00	1.000000	0
educatio*	6	550	2.100000	0.9008193	2.00	2.125000	1.4826000	1.00	3.00	2.000000	-0
parity*	7	550	2.109091	0.9894291	2.00	2.011364	1.4826000	1.00	4.00	3.000000	0

Based on the descriptive analysis, we can conclude that

- there were 550 samples
- mean and standard deviation of each numerical data - which include weight, length and head circumference
- distribution (count) of each categorical data - which include the baby's gender, mother's education level and parity

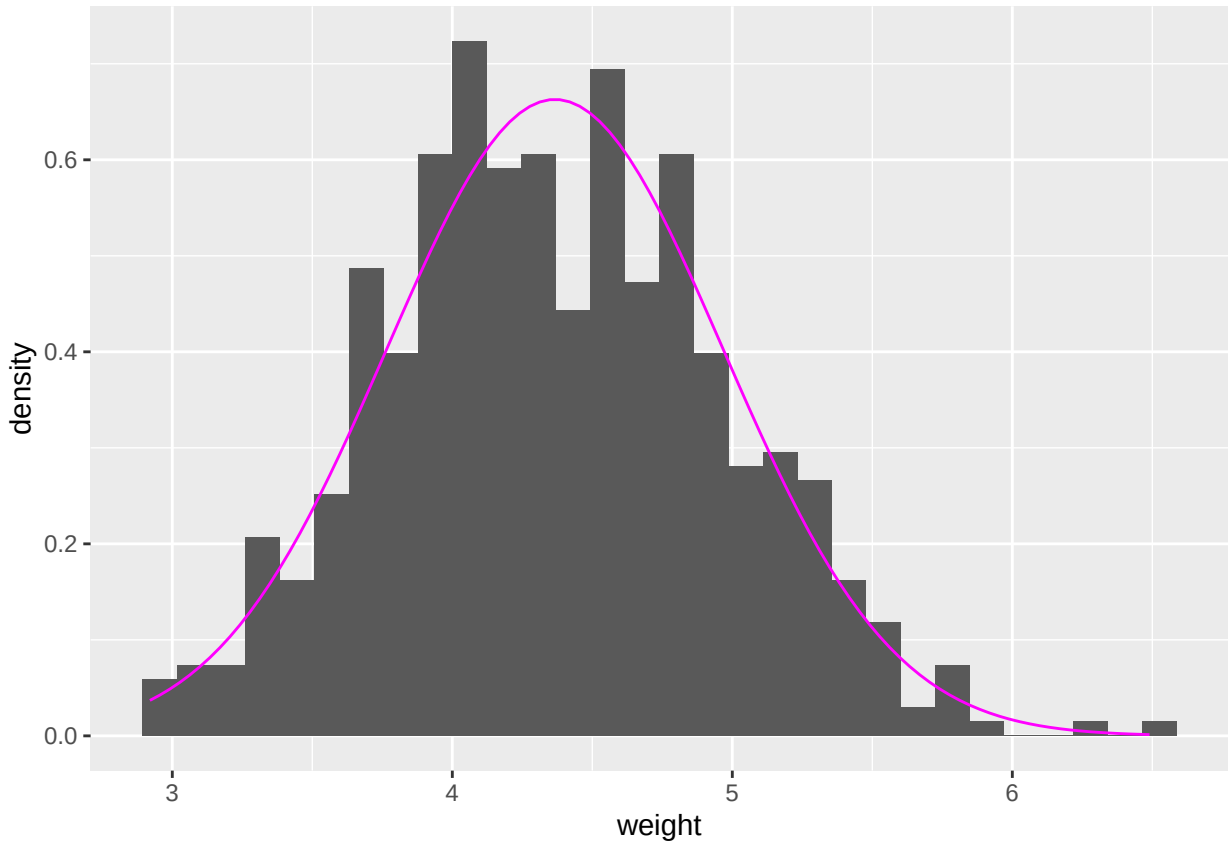
Data Exploration

For data exploration, numerical data can be plotted into histogram, which can then be use to assess (qualitatively / eyeball) whether the data is normally distributed.

Histogram

Histogram for *Weight*

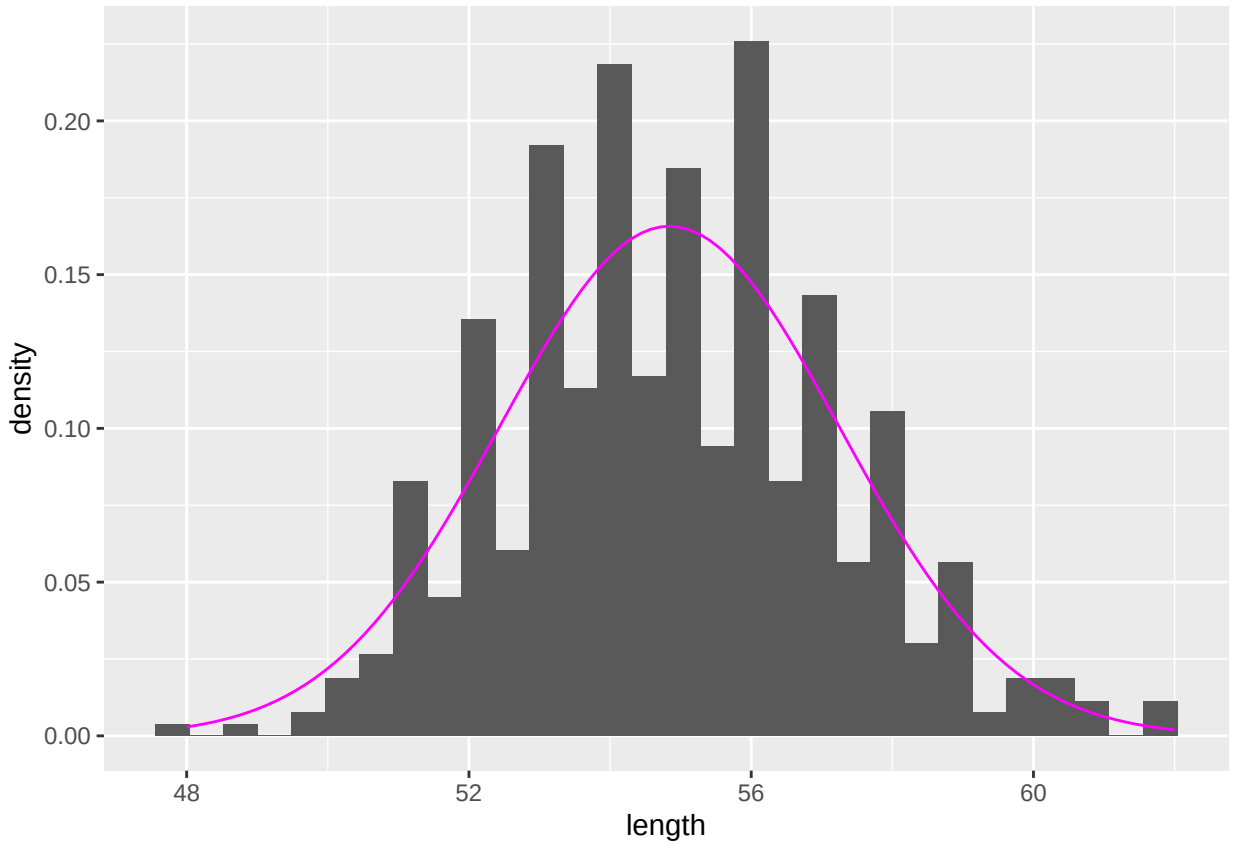
```
ggplot(Data01, aes(weight)) +
  geom_histogram(aes(y=..density..)) +
  stat_function(fun = dnorm, colour = "magenta",
                args = list(mean = mean(Data01$weight),
                            sd = sd(Data01$weight)))
```



The weight distribution look normally distributed

Histogram for *Length*

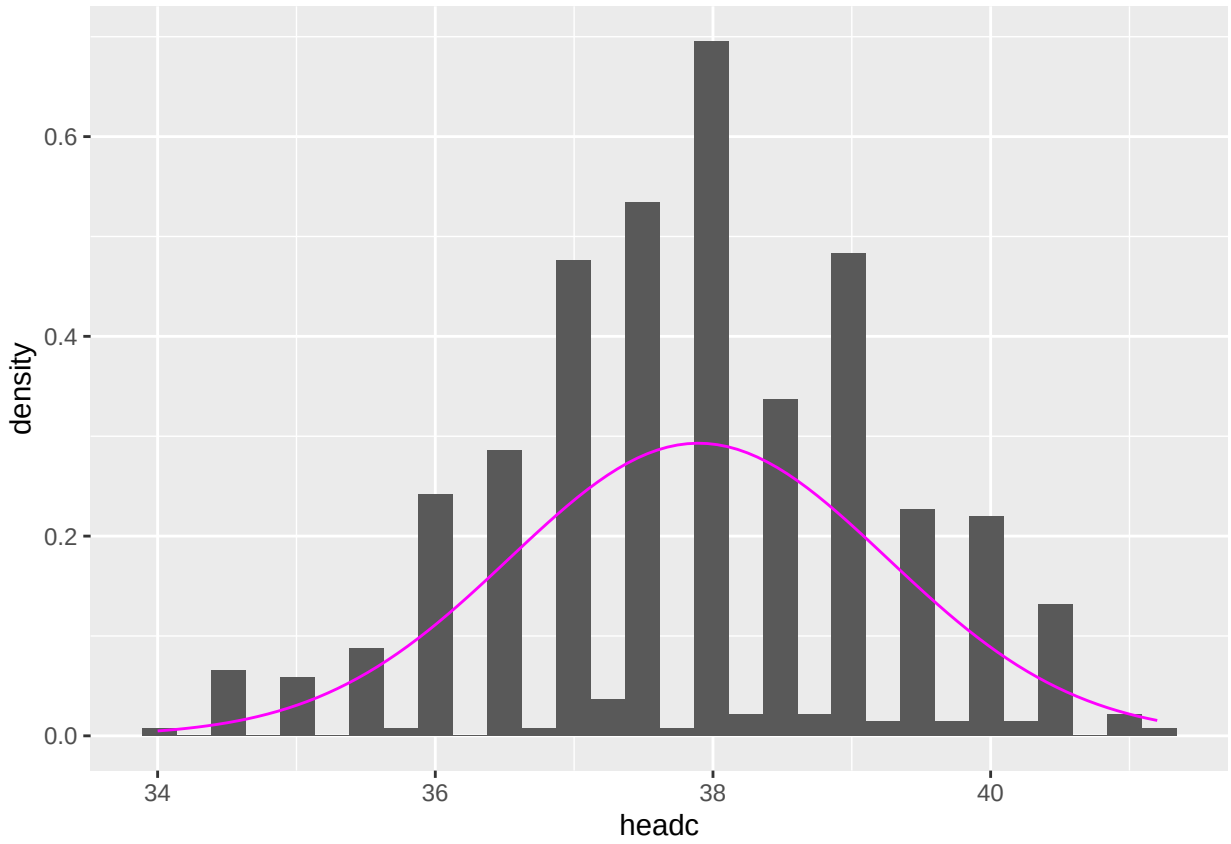
```
ggplot(Data01, aes(length)) +
  geom_histogram(aes(y=..density..)) +
  stat_function(fun = dnorm, colour = "magenta",
                args = list(mean = mean(Data01$length),
                            sd = sd(Data01$length)))
```



The length distribution look normally distributed

Histogram for *Head Circumference*

```
ggplot(Data01, aes(headc)) +
  geom_histogram(aes(y=..density..)) +
  stat_function(fun = dnorm, colour = "magenta",
                args = list(mean = mean(Data01$headc),
                            sd = sd(Data01$headc)))
```



The weight distribution look normally distributed

Correlation

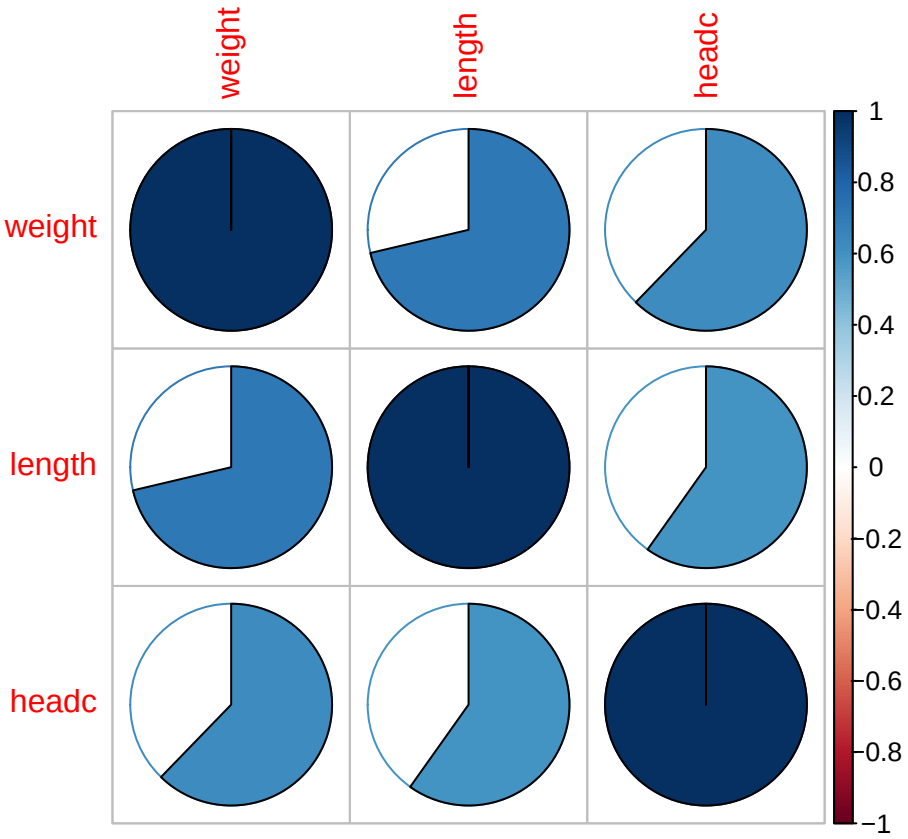
Correlation between numerical data can be calculated using correlation matrix and visualized using correlogram.

- Correlation Matrix and Correlogram

```
Data02 <- Data01 %>% select_if(is.numeric)
CorData02 <- cor(Data02, use = "complete.obs", method = "pearson")
kable(round(CorData02,2))
```

	weight	length	headc
weight	1.00	0.71	0.62
length	0.71	1.00	0.60
headc	0.62	0.60	1.00

```
corrplot(CorData02, method = "pie")
```



Based on the correlation Matrix, we can conclude that

- weight and length had high correlation, with r = 0.71
- other combination had moderate correlation, with $0.3 < r < 0.7$

Modelling - Univariable

Simple linear regression will be done to assess the association of each variable with head circumference.

Univariable Models

ModelWeight

ModelWeight was done using these variables

- outcome: Head Circumference
- covariate: weight

```
ModelWeight <- lm(headc ~ weight, data = Data01)
summary(ModelWeight)
```

```
##
## Call:
## lm(formula = headc ~ weight, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -4.1255 -0.6117  0.0257  0.6825  3.9024
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 31.74618    0.33339   95.22   <2e-16 ***
## weight       1.40827    0.07564   18.62   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.067 on 548 degrees of freedom
## Multiple R-squared:  0.3875, Adjusted R-squared:  0.3863
## F-statistic: 346.6 on 1 and 548 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelWeight, conf.int = T))
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	31.746184	0.3333887	95.22274	0	31.09131	32.401060
weight	1.408268	0.0756391	18.61826	0	1.25969	1.556847

The weight is significantly associated with head circumference, in which for each increase in 1 kg of weight, the head circumference increase by 1.41 cm.

ModelLength

ModelLength was done using these variables

- outcome: Head Circumference
- covariate: length

```
ModelLength <- lm(headc ~ length, data = Data01)
summary(ModelLength)
```

```
##
## Call:
## lm(formula = headc ~ length, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.3029 -0.7627  0.0509  0.7201  2.9124
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 19.33512    1.06292   18.19   <2e-16 ***
## length       0.33844    0.01936   17.48   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.092 on 548 degrees of freedom
## Multiple R-squared:  0.3579, Adjusted R-squared:  0.3568
## F-statistic: 305.5 on 1 and 548 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelLength, conf.int = T))
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	19.3351183	1.0629231	18.19051	0	17.2472159	21.4230206
length	0.3384363	0.0193633	17.47821	0	0.3004009	0.3764717

The length is significantly associated with head circumference, in which for each increase in 1cm of length, the head circumference increase by 0.34 cm.

ModelGender

ModelGender was done using these variables

- outcome: Head Circumference
- covariate: gender

```
ModelGender <- lm(headc ~ gender, data = Data01)
summary(ModelGender)
```

```
##
## Call:
## lm(formula = headc ~ gender, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8462 -0.8462  0.0556  0.6538  3.5556
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  38.34618    0.07753 494.571  < 2e-16 ***
## genderFemale -0.90182    0.10965  -8.225 1.43e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.286 on 548 degrees of freedom
## Multiple R-squared:  0.1099, Adjusted R-squared:  0.1082
## F-statistic: 67.64 on 1 and 548 DF,  p-value: 1.431e-15
```

```
kable(tidy(ModelGender, conf.int = T))
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	38.3461818	0.0775343	494.570826	0	38.193881	38.4984826
genderFemale	-0.9018182	0.1096500	-8.224516	0	-1.117204	-0.6864324

The gender was significantly associated with head circumference.

ModelEducation

ModelEducation was done using these variables

- outcome: Head Circumference
- covariate: Mother's Education

```r
ModelEducation <- lm(headc ~ educatio, data = Data01)
summary(ModelEducation)
```

```
##
## Call:
## lm(formula = headc ~ educatio, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8925 -0.8925  0.1075  1.1075  3.2424
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)       37.86768    0.09691 390.738   <2e-16 ***
## educatiosecondary  0.08990    0.16786   0.536    0.592
## educatiotertiary   0.02481    0.12939   0.192    0.848
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.364 on 547 degrees of freedom
## Multiple R-squared:  0.0005277,  Adjusted R-squared:  -0.003127
## F-statistic: 0.1444 on 2 and 547 DF,  p-value: 0.8656
```

```r
kable(tidy(ModelEducation, conf.int = T))
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	37.8676768	0.0969133	390.7375893	0.0000000	37.6773089	38.0580446
educatiosecondary	0.0898990	0.1678588	0.5355634	0.5924780	-0.2398277	0.4196258
educatiotertiary	0.0248134	0.1293932	0.1917674	0.8479956	-0.2293550	0.2789818

There was no significant association between mother's education level and head circumference.

ModelParity

ModelParity was done using these variables

- outcome: Head Circumference
- covariate: parity

```r
ModelParity <- lm(headc ~ parity, data = Data01)
summary(ModelParity)
```

```
##
## Call:
## lm(formula = headc ~ parity, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.6594 -0.9615 -0.0371  0.9629  3.2385
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               37.6594     0.1010 372.990   <2e-16 ***
```

```
## parityOne sibling          0.3020     0.1405   2.149   0.0321 *
## parity2 siblings           0.4164     0.1613   2.582   0.0101 *
## parity3 or more siblings   0.3777     0.1995   1.893   0.0589 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.355 on 546 degrees of freedom
## Multiple R-squared:  0.0156, Adjusted R-squared:  0.0102
## F-statistic: 2.885 on 3 and 546 DF,  p-value: 0.03519
```

```
kable(tidy(ModelParity, conf.int = T))
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	37.6594445	0.1009663	372.990152	0.0000000	37.4611145	37.8577745
parityOne sibling	0.3020139	0.1405392	2.148966	0.0320760	0.0259502	0.5780776
parity2 siblings	0.4164176	0.1612848	2.581878	0.0100859	0.0996030	0.7332323
parity3 or more siblings	0.3776523	0.1994750	1.893232	0.0588556	-0.0141800	0.7694846

There was significant association between the newborn parity and head circumference.

Univariable Summary

all simple linear regressions were significant, except for variable mother education.

```
SLRResult <- matrix(c("Weight", 1.51, "(1.26, 1.56)", "< 0.001", 0.39,
                      "Length", 0.34, "(0.30, 0.38)", "< 0.001", 0.36,
                      "GenderFemale", -0.90, "(-1.12, -0.69)", "<0.001", 0.11,
                      "Mother's Education Level - Secondary", 0.09, "(-0.24, 0.42)", 0.592, "< 0.01",
                      "Mother's Education Level - Tertiary", 0.02, "(-0.23, 0.28", 0.848, "",
                      "Parity - One Sibling", 0.30, "(0.03, 0.58)", 0.032, 0.02,
                      "Parity - Two Siblings", 0.42, "(0.10, 0.73", 0.010, "",
                      "Parity - Three Siblings", 0.38, "(-0.01, 0.77)", 0.059, ""),  ncol = 5, byrow = T
colnames(SLRResult) <- c("Variables", "beta", "(95% CI)", "p-value", "R^2^")
kable(SLRResult)
```

Variables	beta	(95% CI)	p-value	R^2
Weight	1.51	(1.26, 1.56)	< 0.001	0.39
Length	0.34	(0.30, 0.38)	< 0.001	0.36
GenderFemale	-0.9	(-1.12, -0.69)	<0.001	0.11
Mother's Education Level - Secondary	0.09	(-0.24, 0.42)	0.592	< 0.01
Mother's Education Level - Tertiary	0.02	(-0.23, 0.28	0.848	
Parity - One Sibling	0.3	(0.03, 0.58)	0.032	0.02
Parity - Two Siblings	0.42	(0.10, 0.73	0.01	
Parity - Three Siblings	0.38	(-0.01, 0.77)	0.059	

Multivariable

To explore which covariate is the predictor for newborn's head circumference, several multiple linear regression models are created.

1. Model with weight and length as predictors
2. Model with all predictors included (weight, length, newborn's gender, mother's education level and newborn's parity)
3. Model with newborn's gender, mother's education level and newborn's parity as predictors

Model Exploration

ModelWeightLength

ModelWeightLength was done using these variables

- Outcome: Head Circumference
- Covariate: Weight and Length

```
ModelWeightLength <- lm(headc ~ weight + length, data = Data01)
summary(ModelWeightLength)
```

```
##
## Call:
## lm(formula = headc ~ weight + length, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.4218 -0.5985  0.0380  0.7048  2.9651
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 24.21402    1.14407  21.165  < 2e-16 ***
## weight       0.90147    0.10363   8.699  < 2e-16 ***
## length       0.17770    0.02591   6.858  1.9e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.024 on 547 degrees of freedom
## Multiple R-squared:  0.436,  Adjusted R-squared:  0.4339
## F-statistic: 211.4 on 2 and 547 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelWeightLength, conf.int = T))
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	24.2140161	1.1440663	21.164871	0	21.9667148	26.4613174
weight	0.9014722	0.1036307	8.698894	0	0.6979094	1.1050350
length	0.1776968	0.0259117	6.857770	0	0.1267981	0.2285955

ModelWtLtGenEduPar

ModelWtLtGenEduPar was done using these variables

- Outcome: Head Circumference
- Covariate: Weight, Length + gender + mother education + parity

```
ModelWtLtGenEduPar <- lm(headc ~ weight + length + gender + educatio + parity, data = Data01)
summary(ModelWtLtGenEduPar)
```

```
##
## Call:
## lm(formula = headc ~ weight + length + gender + educatio + parity,
##     data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0987 -0.6383  0.0085  0.6604  3.0413
##
## Coefficients:
##                            Estimate Std. Error t value Pr(>|t|)
## (Intercept)                24.43094    1.15667  21.122  < 2e-16 ***
## weight                      0.75217    0.10885   6.910 1.37e-11 ***
## length                      0.18700    0.02613   7.157 2.70e-12 ***
## genderFemale               -0.33734    0.09369  -3.601 0.000347 ***
## educatiosecondary          -0.07035    0.12843  -0.548 0.584085
## educatiotertiary           -0.03379    0.09930  -0.340 0.733803
## parityOne sibling           0.17507    0.10721   1.633 0.103066
## parity2 siblings            0.23295    0.12358   1.885 0.059950 .
## parity3 or more siblings    0.10108    0.15458   0.654 0.513435
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.014 on 541 degrees of freedom
## Multiple R-squared:  0.4533, Adjusted R-squared:  0.4452
## F-statistic: 56.06 on 8 and 541 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelWtLtGenEduPar, conf.int = T))
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	24.4309365	1.1566692	21.1218000	0.0000000	22.1588233	26.7030498
weight	0.7521740	0.1088512	6.9101129	0.0000000	0.5383512	0.9659968
length	0.1870002	0.0261280	7.1570770	0.0000000	0.1356754	0.2383250
genderFemale	-0.3373389	0.0936908	-3.6005567	0.0003467	-0.5213812	-0.1532967
educatiosecondary	-0.0703502	0.1284333	-0.5477564	0.5840851	-0.3226392	0.1819389
educatiotertiary	-0.0337868	0.0993011	-0.3402460	0.7338033	-0.2288498	0.1612762
parityOne sibling	0.1750702	0.1072124	1.6329288	0.1030658	-0.0355334	0.3856738
parity2 siblings	0.2329539	0.1235755	1.8851145	0.0599501	-0.0097926	0.4757004
parity3 or more siblings	0.1010826	0.1545771	0.6539296	0.5134350	-0.2025624	0.4047275

ModelWtLtGenPar

ModelWtLtGenPar was done using these variables

- Outcome: Head Circumference
- Covariate: Weight, Length + gender + parity

```
ModelWtLtGenPar <- lm(headc ~ weight + length + gender + parity, data = Data01)
summary(ModelWtLtGenPar)
```

```
##
## Call:
## lm(formula = headc ~ weight + length + gender + parity, data = Data01)
##
```

```
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1464 -0.6190  0.0129  0.6439  3.0338
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)              24.49275    1.14952  21.307  < 2e-16 ***
## weight                    0.75538    0.10849   6.963 9.66e-12 ***
## length                    0.18499    0.02582   7.165 2.54e-12 ***
## genderFemale             -0.33821    0.09353  -3.616 0.000327 ***
## parityOne sibling         0.18336    0.10577   1.734 0.083560 .
## parity2 siblings          0.24077    0.12220   1.970 0.049314 *
## parity3 or more siblings  0.11917    0.15008   0.794 0.427487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.013 on 543 degrees of freedom
## Multiple R-squared:  0.453,  Adjusted R-squared:  0.4469
## F-statistic: 74.94 on 6 and 543 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelWtLtGenPar, conf.int = T))
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	24.4927533	1.1495202	21.3069368	0.0000000	22.2347022	26.7508045
weight	0.7553791	0.1084856	6.9629416	0.0000000	0.5422762	0.9684820
length	0.1849914	0.0258173	7.1653972	0.0000000	0.1342774	0.2357055
genderFemale	-0.3382104	0.0935317	-3.6159986	0.0003271	-0.5219386	-0.1544822
parityOne sibling	0.1833599	0.1057694	1.7335817	0.0835600	-0.0244074	0.3911273
parity2 siblings	0.2407736	0.1222030	1.9702758	0.0493144	0.0007251	0.4808220
parity3 or more siblings	0.1191748	0.1500764	0.7940944	0.4274874	-0.1756266	0.4139762

Model Comparison

All three model, ModelWtLt, ModelWtLtGenEduPar and ModelWtLtGenPar were compared.

```
anova(ModelWeightLength, ModelWtLtGenEduPar)
```

```
## Analysis of Variance Table
##
## Model 1: headc ~ weight + length
## Model 2: headc ~ weight + length + gender + educatio + parity
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
## 1    547 574.06
## 2    541 556.44  6    17.621 2.8553 0.009551 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(ModelWeightLength, ModelWtLtGenPar)
```

```
## Analysis of Variance Table
##
## Model 1: headc ~ weight + length
## Model 2: headc ~ weight + length + gender + parity
##   Res.Df    RSS Df Sum of Sq      F   Pr(>F)
```

```
## 1    547 574.06
## 2    543 556.76  4    17.302 4.2187 0.002262 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
anova(ModelWtLtGenEduPar, ModelWtLtGenPar)
```

```
## Analysis of Variance Table
##
## Model 1: headc ~ weight + length + gender + educatio + parity
## Model 2: headc ~ weight + length + gender + parity
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    541 556.44
## 2    543 556.76 -2  -0.31829 0.1547 0.8567
```

For model comparison, between ModelWtLtGenEduPar and ModelWtLtGenPar, there is no significant different. Thus, ModelWtLtGenEduPar was not selected.

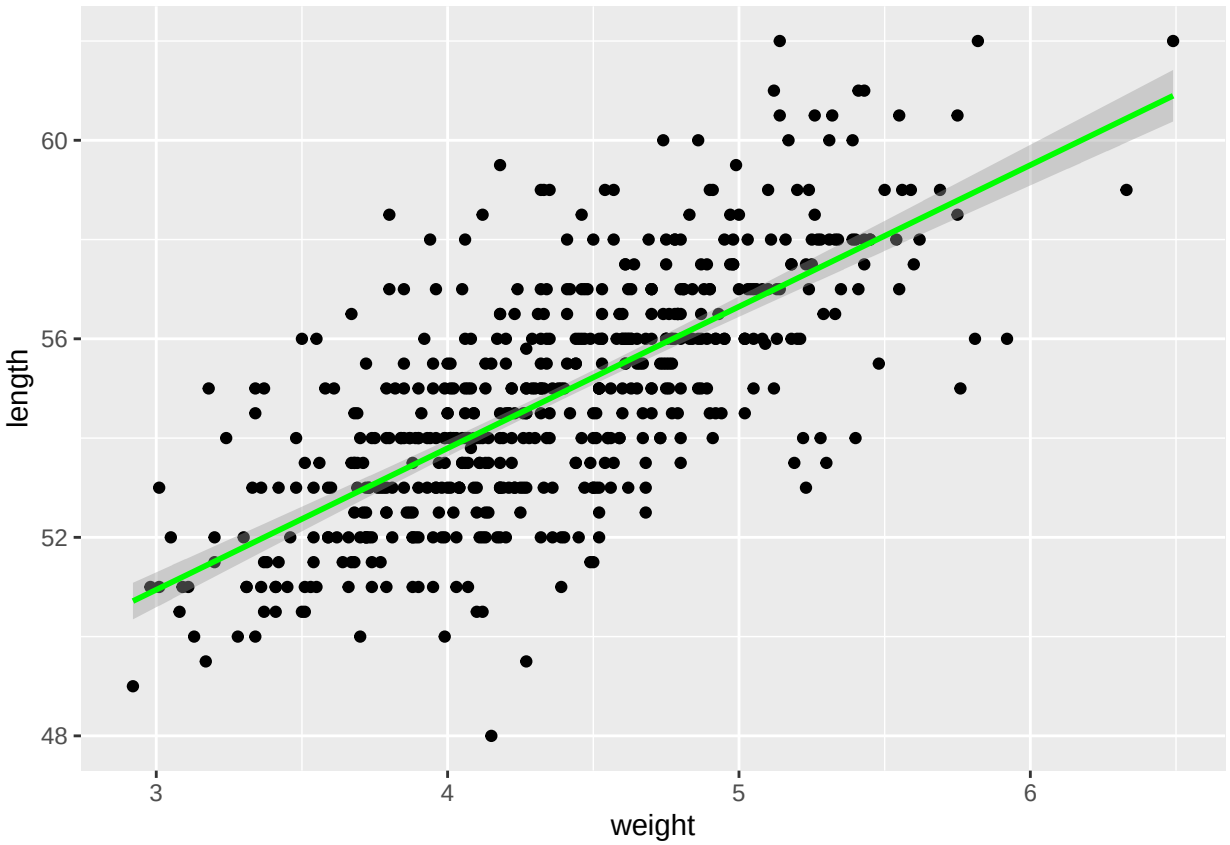
Thus, to choose either ModelWeightLength or ModelWtLtGenPar, adjusted R2 will be use as guidance.

For ModelWeightLength, the adjusted R2 is 0.4339, while ModelWtLtGenPar, adjusted R2is 0.4452. thus ModelWtLtGenPar is chose.

Handling Confounding

There is high correlation between variable weight and length.

```
ggplot(Data01, aes(weight, length)) +
  geom_point() +
  stat_smooth(method = 'lm', col = 'green')
```



Thus need to decide whether to include either one or both variables in the model.

- ModelGenParWt
- ModelGenParLt
- ModelGenParWtLt

ModelGenParWt

ModelGenParWt was done using these variables

- outcome: head circumference
- covariate: Gender + Parity + **Weight**

```
ModelGenParWt <- lm(headc ~ weight + gender + parity, data = Data01)
summary(ModelGenParWt)
```

```
##
## Call:
## lm(formula = headc ~ weight + gender + parity, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.9288 -0.6261  0.0252  0.6784  3.9327
##
## Coefficients:
##                         Estimate Std. Error t value Pr(>|t|)
## (Intercept)             32.31077    0.37823  85.426   <2e-16 ***
## weight                   1.29267    0.08195  15.775   <2e-16 ***
```

```
## genderFemale             -0.32062    0.09773  -3.281   0.0011 **
## parityOne sibling         0.13451    0.11033   1.219   0.2233
## parity2 siblings          0.16550    0.12726   1.300   0.1940
## parity3 or more siblings  0.16502    0.15672   1.053   0.2928
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.058 on 544 degrees of freedom
## Multiple R-squared:  0.4012, Adjusted R-squared:  0.3957
## F-statistic: 72.91 on 5 and 544 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelGenParWt))
```

term	estimate	std.error	statistic	p.value
(Intercept)	32.3107696	0.3782332	85.425530	0.0000000
weight	1.2926725	0.0819468	15.774542	0.0000000
genderFemale	-0.3206162	0.0977300	-3.280631	0.0011021
parityOne sibling	0.1345078	0.1103253	1.219193	0.2232994
parity2 siblings	0.1654975	0.1272595	1.300473	0.1939895
parity3 or more siblings	0.1650175	0.1567243	1.052915	0.2928472

ModelGenParLt

ModelGenParLt was done using these variables

- outcome: head circumference
- covariate: Gender + Parity + **Length**

```
ModelGenParLt <- lm(headc ~ length + gender + parity, data = Data01)
summary(ModelGenParLt)
```

```
##
## Call:
## lm(formula = headc ~ length + gender + parity, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8906 -0.6971  0.0589  0.6783  2.8103
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               21.00975    1.07920  19.468  < 2e-16 ***
## length                     0.30924    0.01945  15.896  < 2e-16 ***
## genderFemale              -0.52987    0.09321  -5.685 2.14e-08 ***
## parityOne sibling          0.26748    0.10957   2.441  0.01495 *
## parity2 siblings           0.37578    0.12581   2.987  0.00295 **
## parity3 or more siblings   0.16532    0.15634   1.057  0.29077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 544 degrees of freedom
## Multiple R-squared:  0.4041, Adjusted R-squared:  0.3986
## F-statistic: 73.79 on 5 and 544 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelGenParLt))
```

term	estimate	std.error	statistic	p.value
(Intercept)	21.0097523	1.0792004	19.467888	0.0000000
length	0.3092434	0.0194546	15.895606	0.0000000
genderFemale	-0.5298706	0.0932092	-5.684747	0.0000000
parityOne sibling	0.2674846	0.1095670	2.441288	0.0149527
parity2 siblings	0.3757829	0.1258104	2.986899	0.0029452
parity3 or more siblings	0.1653192	0.1563364	1.057458	0.2907718

ModelGenParWtLT

ModelGenParWtLt was done using these variables

- outcome: head circumference
- covariate: Gender + Parity + **Weight** + **Length**

```
ModelGenParWtLt <- lm(headc ~ weight + length + gender + parity, data = Data01)
summary(ModelGenParWtLt)
```

```
##
## Call:
## lm(formula = headc ~ weight + length + gender + parity, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.1464 -0.6190  0.0129  0.6439  3.0338
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               24.49275    1.14952  21.307  < 2e-16 ***
## weight                     0.75538    0.10849   6.963 9.66e-12 ***
## length                     0.18499    0.02582   7.165 2.54e-12 ***
## genderFemale              -0.33821    0.09353  -3.616 0.000327 ***
## parityOne sibling          0.18336    0.10577   1.734 0.083560 .
## parity2 siblings           0.24077    0.12220   1.970 0.049314 *
## parity3 or more siblings   0.11917    0.15008   0.794 0.427487
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.013 on 543 degrees of freedom
## Multiple R-squared:  0.453,  Adjusted R-squared:  0.4469
## F-statistic: 74.94 on 6 and 543 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelGenParWtLt))
```

term	estimate	std.error	statistic	p.value
(Intercept)	24.4927533	1.1495202	21.3069368	0.0000000
weight	0.7553791	0.1084856	6.9629416	0.0000000
length	0.1849914	0.0258173	7.1653972	0.0000000
genderFemale	-0.3382104	0.0935317	-3.6159986	0.0003271
parityOne sibling	0.1833599	0.1057694	1.7335817	0.0835600
parity2 siblings	0.2407736	0.1222030	1.9702758	0.0493144

term	estimate	std.error	statistic	p.value
parity3 or more siblings	0.1191748	0.1500764	0.7940944	0.4274874

Confounder Model Comparison

for Weight

```
ModelGenParWt01 <- tidy(ModelGenParWt)
ModelGenParWtLt01 <-tidy(ModelGenParWtLt)
((ModelGenParWtLt01 [2,2] - ModelGenParWt01 [2,2])/ ModelGenParWt01 [2,2])*100
```

```
##    estimate
## 1 -41.56454
```

for Length

```
ModelGenParLt01 <- tidy(ModelGenParLt)
((ModelGenParWtLt01 [2,2] - ModelGenParLt01 [2,2])/ ModelGenParLt01 [2,2])*100
```

```
##   estimate
## 1 144.2669
```

Since Length have bigger difference, weight will be removed from the model. In addition to that, adjusted R2 for model with length is higher than model with weight variable.

Alternative for multicollinearity

To find multicollinearity between variables, VIF can be calculated.

```
kable(vif(lm(headc ~ weight + length + gender + parity, data = Data01)))
```

	GVIF	Df	GVIF^(1/(2*Df))
weight	2.282350	1	1.510745
length	2.067493	1	1.437878
gender	1.173141	1	1.083116
parity	1.040684	3	1.006668

Based on this result, it's seem that while weight and lenght has higher VIF value compared to other variables, but the VIF value for both variables were still less than 10.

Interaction

Model to check for interaction is

Outcome: Head Circumference Covariate: length, gender and parity

Interaction term - Length and Gender

ModelGenParLt_LtGen is used to examine interaction term between **length** and **gender**.

- Outcome: Head Circumference
- Covariate: length, gender, parity, length*gender

```r
ModelGenParLt_LtGen <- lm(headc ~ length + gender + parity + length:gender, data = Data01)
summary(ModelGenParLt_LtGen)
```

```
##
## Call:
## lm(formula = headc ~ length + gender + parity + length:gender,
##     data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8363 -0.6927  0.0521  0.6727  2.7967
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               20.48514    1.43534  14.272   <2e-16 ***
## length                     0.31866    0.02583  12.338   <2e-16 ***
## genderFemale               0.66047    2.14748   0.308   0.7585
## parityOne sibling          0.27091    0.10981   2.467   0.0139 *
## parity2 siblings           0.37822    0.12597   3.003   0.0028 **
## parity3 or more siblings   0.17239    0.15695   1.098   0.2725
## length:genderFemale       -0.02174    0.03918  -0.555   0.5792
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.057 on 543 degrees of freedom
## Multiple R-squared:  0.4045, Adjusted R-squared:  0.3979
## F-statistic: 61.46 on 6 and 543 DF,  p-value: < 2.2e-16
```

```r
kable(tidy(ModelGenParLt_LtGen))
```

term	estimate	std.error	statistic	p.value
(Intercept)	20.4851443	1.4353430	14.2719507	0.0000000
length	0.3186601	0.0258270	12.3382627	0.0000000
genderFemale	0.6604726	2.1474780	0.3075573	0.7585371
parityOne sibling	0.2709094	0.1098104	2.4670650	0.0139301
parity2 siblings	0.3782180	0.1259670	3.0025166	0.0028008
parity3 or more siblings	0.1723860	0.1569536	1.0983248	0.2725495
length:genderFemale	-0.0217392	0.0391823	-0.5548217	0.5792451

The interaction model was compared with model without interaction term.

```r
anova(ModelGenParLt_LtGen, ModelGenParLt)
```

```
## Analysis of Variance Table
##
## Model 1: headc ~ length + gender + parity + length:gender
## Model 2: headc ~ length + gender + parity
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    543 606.13
## 2    544 606.47 -1  -0.34362 0.3078 0.5792
```

interaction term between length and gender is not significant.

Interaction term - Length and Parity

ModelGenParLt_LtPar is used to examine interaction term between **length** and **parity**.

- Outcome: Head Circumference
- Covariate: length, gender, parity, length*parity

```
ModelGenParLt_LtPar <- lm(headc ~ length + gender + parity + length:parity, data = Data01)
summary(ModelGenParLt_LtPar)
```

```
##
## Call:
## lm(formula = headc ~ length + gender + parity + length:parity,
##     data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.5268 -0.6998  0.0124  0.6838  2.7918
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     18.44617    1.81033  10.189  < 2e-16 ***
## length                           0.35600    0.03290  10.819  < 2e-16 ***
## genderFemale                    -0.51597    0.09311  -5.541  4.7e-08 ***
## parityOne sibling                4.28936    2.49964   1.716   0.0867 .
## parity2 siblings                 1.71836    2.82179   0.609   0.5428
## parity3 or more siblings         8.48474    3.68200   2.304   0.0216 *
## length:parityOne sibling        -0.07350    0.04563  -1.611   0.1078
## length:parity2 siblings         -0.02463    0.05146  -0.479   0.6323
## length:parity3 or more siblings -0.15072    0.06658  -2.264   0.0240 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.053 on 541 degrees of freedom
## Multiple R-squared:  0.4109, Adjusted R-squared:  0.4022
## F-statistic: 47.18 on 8 and 541 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelGenParLt_LtPar))
```

term	estimate	std.error	statistic	p.value
(Intercept)	18.4461678	1.8103299	10.1893961	0.0000000
length	0.3560044	0.0329048	10.8192080	0.0000000
genderFemale	-0.5159703	0.0931148	-5.5412281	0.0000000
parityOne sibling	4.2893630	2.4996420	1.7159909	0.0867362
parity2 siblings	1.7183562	2.8217918	0.6089593	0.5428071
parity3 or more siblings	8.4847446	3.6820009	2.3043842	0.0215790
length:parityOne sibling	-0.0735010	0.0456258	-1.6109525	0.1077733
length:parity2 siblings	-0.0246343	0.0514560	-0.4787454	0.6323132
length:parity3 or more siblings	-0.1507180	0.0665791	-2.2637423	0.0239852

The interaction model was compared with model without interaction term.

```
anova(ModelGenParLt_LtPar, ModelGenParLt)
```

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

```
##
## Model 1: headc ~ length + gender + parity + length:parity
## Model 2: headc ~ length + gender + parity
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    541 599.52
## 2    544 606.47 -3   -6.9493 2.0903 0.1005
```

interaction term between length and parity is not significant

Interaction term - Gender and Parity

ModelGenParLt_GenPar is used to examine interaction term between **parity** and **gender**.

- Outcome: Head Circumference
- Covariate: length, gender, parity, gender*parity

```
ModelGenParLt_GenPar <- lm(headc ~ length + gender + parity + gender:parity, data = Data01)
summary(ModelGenParLt_GenPar)
```

```
##
## Call:
## lm(formula = headc ~ length + gender + parity + gender:parity,
##     data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.0099 -0.6454  0.0468  0.6698  2.7718
##
## Coefficients:
##                                     Estimate Std. Error t value Pr(>|t|)
## (Intercept)                         20.90810    1.08698  19.235  < 2e-16
## length                               0.30996    0.01949  15.902  < 2e-16
## genderFemale                        -0.40251    0.16055  -2.507  0.01246
## parityOne sibling                    0.43789    0.15322   2.858  0.00443
## parity2 siblings                     0.41749    0.17885   2.334  0.01995
## parity3 or more siblings             0.09413    0.22479   0.419  0.67557
## genderFemale:parityOne sibling      -0.34840    0.21932  -1.589  0.11275
## genderFemale:parity2 siblings       -0.08786    0.25151  -0.349  0.72699
## genderFemale:parity3 or more siblings  0.12234    0.31190   0.392  0.69503
##
## (Intercept)                         ***
## length                              ***
## genderFemale                        *
## parityOne sibling                   **
## parity2 siblings                    *
## parity3 or more siblings
## genderFemale:parityOne sibling
## genderFemale:parity2 siblings
## genderFemale:parity3 or more siblings
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.055 on 541 degrees of freedom
## Multiple R-squared:  0.4081, Adjusted R-squared:  0.3994
## F-statistic: 46.63 on 8 and 541 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelGenParLt_GenPar))
```

term	estimate	std.error	statistic	p.value
(Intercept)	20.9081045	1.0869830	19.2349874	0.0000000
length	0.3099636	0.0194923	15.9018134	0.0000000
genderFemale	-0.4025101	0.1605456	-2.5071388	0.0124633
parityOne sibling	0.4378860	0.1532246	2.8578057	0.0044301
parity2 siblings	0.4174853	0.1788487	2.3342931	0.0199455
parity3 or more siblings	0.0941306	0.2247930	0.4187437	0.6755696
genderFemale:parityOne sibling	-0.3484006	0.2193242	-1.5885190	0.1127532
genderFemale:parity2 siblings	-0.0878565	0.2515103	-0.3493157	0.7269883
genderFemale:parity3 or more siblings	0.1223409	0.3119006	0.3922434	0.6950329

The interaction model was compared with model without interaction term.

```
anova(ModelGenParLt_GenPar, ModelGenParLt)
```

```
## Analysis of Variance Table
##
## Model 1: headc ~ length + gender + parity + gender:parity
## Model 2: headc ~ length + gender + parity
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    541 602.39
## 2    544 606.47 -3   -4.0798 1.2213 0.3012
```

interaction term between gender and parity is not significant.

Model Assessment

Preliminary Final Model need to be assessed by checking the model assumption.

Preliminary Final Model

Preliminary Final Model include these variables

- Outcome: Head Circumference
- Covariate: length, gender, parity

```
ModelPrelim <- lm(headc ~ length + gender + parity, data = Data01)
summary(ModelPrelim)
```

```
##
## Call:
## lm(formula = headc ~ length + gender + parity, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8906 -0.6971  0.0589  0.6783  2.8103
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              21.00975    1.07920  19.468  < 2e-16 ***
## length                    0.30924    0.01945  15.896  < 2e-16 ***
```

```
## genderFemale              -0.52987    0.09321  -5.685 2.14e-08 ***
## parityOne sibling          0.26748    0.10957   2.441  0.01495 *
## parity2 siblings           0.37578    0.12581   2.987  0.00295 **
## parity3 or more siblings   0.16532    0.15634   1.057  0.29077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 544 degrees of freedom
## Multiple R-squared:  0.4041, Adjusted R-squared:  0.3986
## F-statistic: 73.79 on 5 and 544 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelPrelim))
```

term	estimate	std.error	statistic	p.value
(Intercept)	21.0097523	1.0792004	19.467888	0.0000000
length	0.3092434	0.0194546	15.895606	0.0000000
genderFemale	-0.5298706	0.0932092	-5.684747	0.0000000
parityOne sibling	0.2674846	0.1095670	2.441288	0.0149527
parity2 siblings	0.3757829	0.1258104	2.986899	0.0029452
parity3 or more siblings	0.1653192	0.1563364	1.057458	0.2907718

Prediction Value

Using Preliminary Final Model, prediction value was first calculated, by predicting the head circumference using the observed data from predictors variables (length, gender, parity)

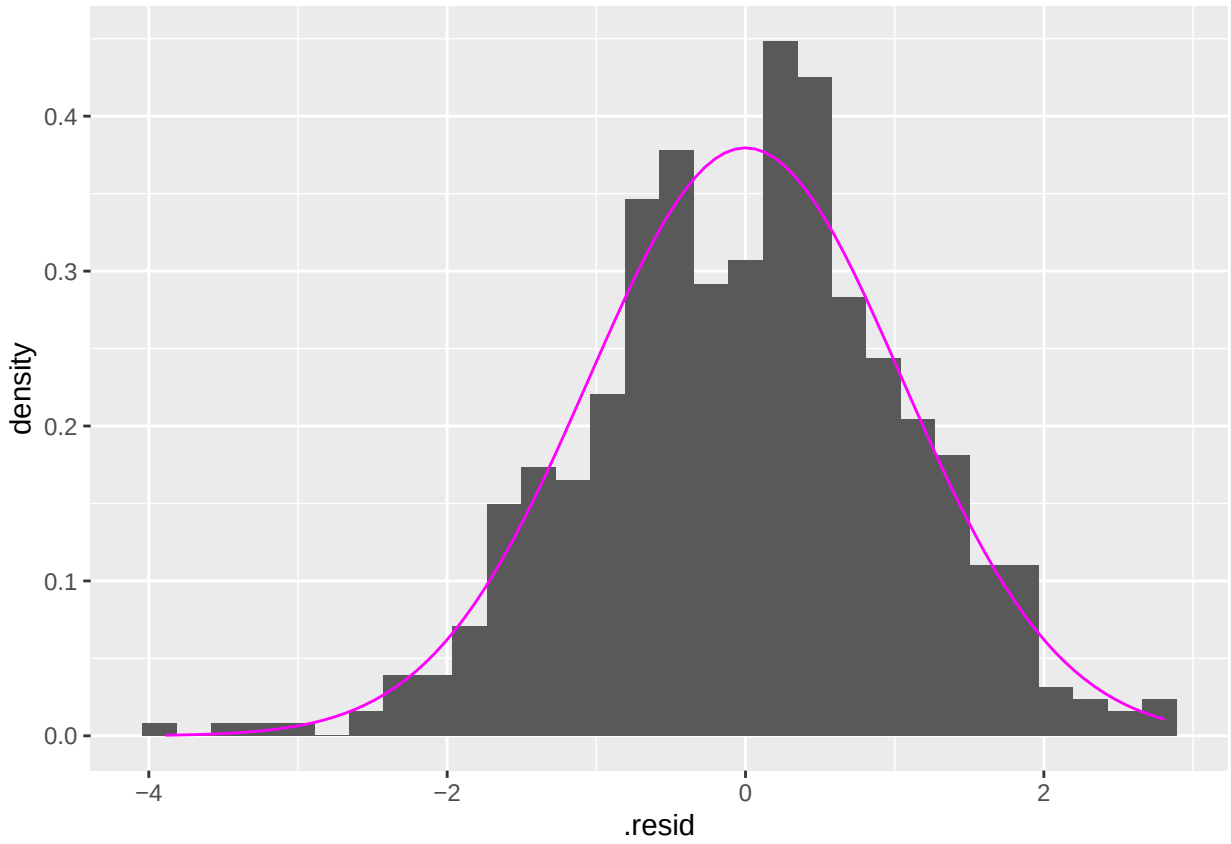
```
Pred.PrelimFinalModel <- augment(ModelPrelim)
kable(head(Pred.PrelimFinalModel))
```

headc	length	gender	parity	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	.s
37.5	55.5	Female	3 or more siblings	37.80821	0.1411034	-0.3082077	0.0178592	1.056748	0.0002629	-0.
38.5	57.0	Female	Singleton	38.10675	0.1077421	0.3932465	0.0104126	1.056696	0.0002458	0.
38.5	56.0	Male	2 siblings	38.70316	0.1089920	-0.2031637	0.0106556	1.056796	0.0000672	-0.
39.0	56.0	Male	One sibling	38.59487	0.0888350	0.4051346	0.0070787	1.056688	0.0001762	0.
39.5	55.0	Male	2 siblings	38.39392	0.1089692	1.1060797	0.0106511	1.055754	0.0019902	1.
34.5	51.5	Female	Singleton	36.40592	0.1037990	-1.9059150	0.0096644	1.053632	0.0053512	-1.

Normality Test for Residual

Histogram for Residual to look for distribution shape.

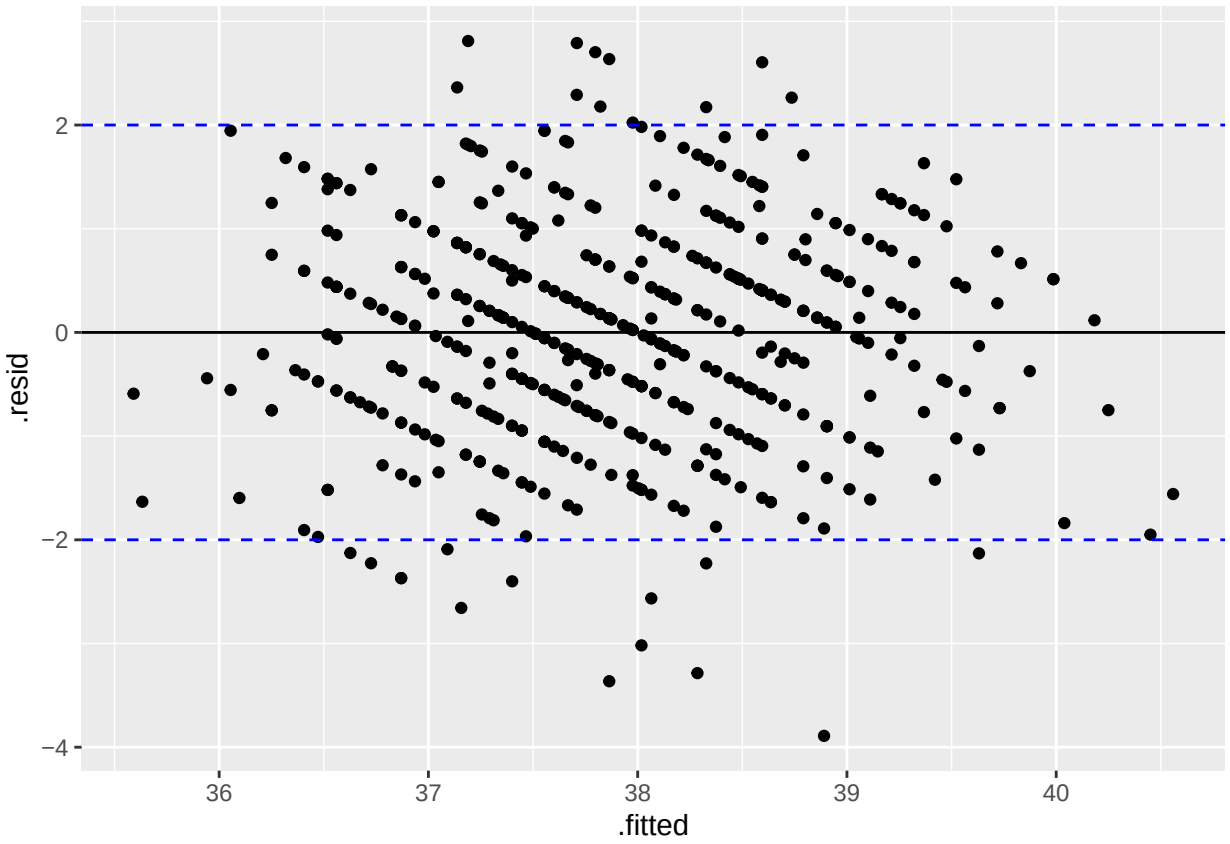
```
ggplot(data = Pred.PrelimFinalModel, aes(x=.resid)) +
  geom_histogram(aes(y=..density..)) +
  stat_function(fun = dnorm, colour = "magenta",
                args = list(mean = mean(Pred.PrelimFinalModel$.resid),
                            sd = sd(Pred.PrelimFinalModel$.resid)))
```



Distribution of model residual look normally distributed.

Scatter plot between predicted and residual values.

```
ggplot(data = Pred.PrelimFinalModel, aes(x=.fitted, y=.resid)) +
  geom_point() +
  geom_hline (yintercept = 0) +
  geom_hline (yintercept = 2, linetype=2, color="blue") +
  geom_hline (yintercept = -2, linetype=2, color="blue")
```



Scatter plot seem linear and constant

Thus, these two assumption for linear regression is fulfilled

1. for each set of Xi values, there is a sub-population of Y values which are normally distributed.
2. the variance of the sub-population of Y are all equal

Outlier and Influence

Outlier and influence may affect the model. thus samples with high (or large) influence need to be removed, and the final model should be run again. Comparison between model with all samples and without influence can be done.

Explore Influence

Measure influential in the model

1. cooks distance: value above 1 or above 4/n
2. standardized residual: values above 2 or lower than -2

```
(2*4 +2)/550
```

```
## [1] 0.01818182
```

3. leverage above (2k +2)/n = 0.0182

Select Influence

Sample with influence is selected using filter function

```
influen.obs <- Pred.PrelimFinalModel %>% filter(.std.resid > 2 | .std.resid < -2 | .hat > 0.0182)
summary(influen.obs)
```

```
##      headc           length         gender                   parity
##  Min.   :34.50   Min.   :48.00   Male  :41   Singleton          : 9
##  1st Qu.:36.50   1st Qu.:53.50   Female:28   One sibling        : 8
##  Median :38.20   Median :55.00               2 siblings         :10
##  Mean   :37.93   Mean   :55.59               3 or more siblings:42
##  3rd Qu.:39.50   3rd Qu.:58.00
##  Max.   :41.20   Max.   :62.00
##     .fitted         .se.fit            .resid             .hat
##  Min.   :35.59   Min.   :0.08837   Min.   :-3.8906   Min.   :0.007006
##  1st Qu.:37.26   1st Qu.:0.12820   1st Qu.:-1.6112   1st Qu.:0.014743
##  Median :38.06   Median :0.14574   Median :-0.1835   Median :0.019053
##  Mean   :38.17   Mean   :0.13801   Mean   :-0.2392   Mean   :0.017619
##  3rd Qu.:38.89   3rd Qu.:0.15323   3rd Qu.: 1.2187   3rd Qu.:0.021061
##  Max.   :40.56   Max.   :0.17145   Max.   : 2.8103   Max.   :0.026367
##      .sigma          .cooksd            .std.resid
##  Min.   :1.043   Min.   :2.460e-06   Min.   :-3.7289
##  1st Qu.:1.052   1st Qu.:1.088e-03   1st Qu.:-1.5410
##  Median :1.055   Median :5.801e-03   Median :-0.1754
##  Mean   :1.054   Mean   :6.745e-03   Mean   :-0.2288
##  3rd Qu.:1.057   3rd Qu.:9.694e-03   3rd Qu.: 1.1666
##  Max.   :1.057   Max.   :5.584e-02   Max.   : 2.6864
```

```
kable(head(influen.obs))
```

headc	length	gender	parity	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	.s
39.7	57.0	Male	3 or more siblings	38.80194	0.1434996	0.8980574	0.0184709	1.056116	0.0023117	0.
38.5	58.0	Male	3 or more siblings	39.11119	0.1471642	-0.6111867	0.0194264	1.056500	0.0011283	-0.
40.5	54.5	Female	2 siblings	37.70943	0.1073464	2.7905720	0.0103362	1.049954	0.0122860	2.
41.2	56.0	Male	One sibling	38.59487	0.0888350	2.6051354	0.0070787	1.050860	0.0072849	2.
37.5	58.0	Male	3 or more siblings	39.11119	0.1471642	-1.6111867	0.0194264	1.054523	0.0078408	-1.
38.3	52.0	Female	3 or more siblings	36.72586	0.1511994	1.5741433	0.0205063	1.054626	0.0079179	1.

There are 69 samples in the original dataset which were outlier and have high leverage.

Select Dataset without Influence Samples

Sample without influence is selected using filter function

```
NonInfluenObs <- Pred.PrelimFinalModel %>% filter(.std.resid < 2 & .std.resid > -2 & .hat < 0.0182)
summary(NonInfluenObs)
```

```
##      headc           length         gender                   parity
##  Min.   :34.00   Min.   :49.00   Male  :234   Singleton          :171
##  1st Qu.:37.00   1st Qu.:53.00   Female:247   One sibling        :184
##  Median :38.00   Median :54.50                2 siblings         :106
##  Mean   :37.89   Mean   :54.73                3 or more siblings: 20
##  3rd Qu.:39.00   3rd Qu.:56.00
```

```
##  Max.   :41.00   Max.   :60.50
##     .fitted         .se.fit            .resid              .hat
##  Min.   :35.63   Min.   :0.08807   Min.   :-2.09094   Min.   :0.006957
##  1st Qu.:37.29   1st Qu.:0.09183   1st Qu.:-0.62081   1st Qu.:0.007564
##  Median :37.80   Median :0.09994   Median : 0.09589   Median :0.008959
##  Mean   :37.86   Mean   :0.10430   Mean   : 0.03431   Mean   :0.009947
##  3rd Qu.:38.48   3rd Qu.:0.11030   3rd Qu.: 0.67262   3rd Qu.:0.010912
##  Max.   :39.99   Max.   :0.14242   Max.   : 2.02362   Max.   :0.018194
##      .sigma          .cooksd            .std.resid
##  Min.   :1.053   Min.   :1.620e-07   Min.   :-1.99168
##  1st Qu.:1.056   1st Qu.:1.511e-04   1st Qu.:-0.59201
##  Median :1.056   Median :6.299e-04   Median : 0.09118
##  Mean   :1.056   Mean   :1.291e-03   Mean   : 0.03269
##  3rd Qu.:1.057   3rd Qu.:1.808e-03   3rd Qu.: 0.63943
##  Max.   :1.057   Max.   :9.365e-03   Max.   : 1.92390
```

```
kable(head(NonInfluenObs))
```

headc	length	gender	parity	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	.s
37.5	55.5	Female	3 or more siblings	37.80821	0.1411034	-0.3082077	0.0178592	1.056748	0.0002629	-0.
38.5	57.0	Female	Singleton	38.10675	0.1077421	0.3932465	0.0104126	1.056696	0.0002458	0.
38.5	56.0	Male	2 siblings	38.70316	0.1089920	-0.2031637	0.0106556	1.056796	0.0000672	-0.
39.0	56.0	Male	One sibling	38.59487	0.0888350	0.4051346	0.0070787	1.056688	0.0001762	0.
39.5	55.0	Male	2 siblings	38.39392	0.1089692	1.1060797	0.0106511	1.055754	0.0019902	1.
34.5	51.5	Female	Singleton	36.40592	0.1037990	-1.9059150	0.0096644	1.053632	0.0053512	-1.

When influence samples were removed, there were 481 samples.

the model is re-run using dataset without influence

```
FinalModelNoInfluential <- lm(ModelPrelim, data = NonInfluenObs)
summary(FinalModelNoInfluential)
```

```
##
## Call:
## lm(formula = ModelPrelim, data = NonInfluenObs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12254 -0.64978  0.03099  0.64903  2.05179
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               19.95790    1.06077  18.815  < 2e-16 ***
## length                     0.32905    0.01916  17.171  < 2e-16 ***
## genderFemale              -0.55294    0.08704  -6.353 4.95e-10 ***
## parityOne sibling          0.25534    0.09680   2.638  0.00862 **
## parity2 siblings           0.44252    0.11273   3.926 9.93e-05 ***
## parity3 or more siblings   0.26013    0.21813   1.193  0.23364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9109 on 475 degrees of freedom
## Multiple R-squared:  0.4833, Adjusted R-squared:  0.4779
```

```
## F-statistic: 88.87 on 5 and 475 DF,  p-value: < 2.2e-16
```

```
kable(tidy(FinalModelNoInfluential))
```

term	estimate	std.error	statistic	p.value
(Intercept)	19.9579008	1.0607695	18.814550	0.0000000
length	0.3290487	0.0191630	17.171016	0.0000000
genderFemale	-0.5529428	0.0870370	-6.352961	0.0000000
parityOne sibling	0.2553355	0.0968039	2.637656	0.0086219
parity2 siblings	0.4425243	0.1127273	3.925617	0.0000993
parity3 or more siblings	0.2601264	0.2181259	1.192551	0.2336404

Comparing Final Model with Final Model without Influence

Final Model with Original Dataset (Without Influence removed)

Final Model with Original Dataset (Without Influence removed)

```
ModelFinal <- ModelPrelim
summary(ModelFinal)
```

```
##
## Call:
## lm(formula = headc ~ length + gender + parity, data = Data01)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -3.8906 -0.6971  0.0589  0.6783  2.8103
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               21.00975    1.07920  19.468  < 2e-16 ***
## length                     0.30924    0.01945  15.896  < 2e-16 ***
## genderFemale              -0.52987    0.09321  -5.685 2.14e-08 ***
## parityOne sibling          0.26748    0.10957   2.441  0.01495 *
## parity2 siblings           0.37578    0.12581   2.987  0.00295 **
## parity3 or more siblings   0.16532    0.15634   1.057  0.29077
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.056 on 544 degrees of freedom
## Multiple R-squared:  0.4041, Adjusted R-squared:  0.3986
## F-statistic: 73.79 on 5 and 544 DF,  p-value: < 2.2e-16
```

```
kable(tidy(ModelFinal, conf.int = T))
```

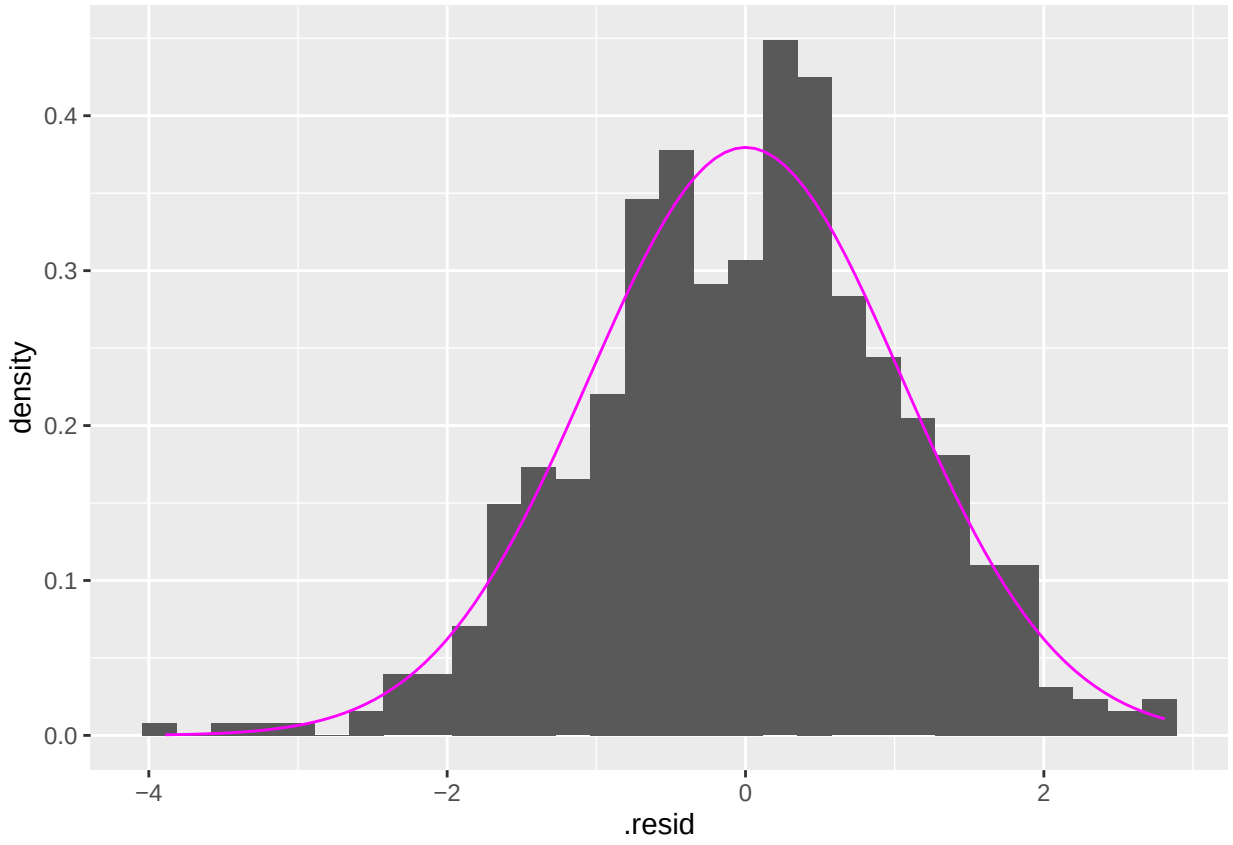
term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	21.0097523	1.0792004	19.467888	0.0000000	18.8898419	23.1296627
length	0.3092434	0.0194546	15.895606	0.0000000	0.2710279	0.3474588
genderFemale	-0.5298706	0.0932092	-5.684747	0.0000000	-0.7129646	-0.3467766
parityOne sibling	0.2674846	0.1095670	2.441288	0.0149527	0.0522584	0.4827108
parity2 siblings	0.3757829	0.1258104	2.986899	0.0029452	0.1286492	0.6229166
parity3 or more siblings	0.1653192	0.1563364	1.057458	0.2907718	-0.1417777	0.4724161

Residual Histogram

```
Pred.FinalModelOriginal <- augment(ModelFinal)
kable(head(Pred.FinalModelOriginal))
```

headc	length	gender	parity	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	.s
37.5	55.5	Female	3 or more siblings	37.80821	0.1411034	-0.3082077	0.0178592	1.056748	0.0002629	-0.
38.5	57.0	Female	Singleton	38.10675	0.1077421	0.3932465	0.0104126	1.056696	0.0002458	0.
38.5	56.0	Male	2 siblings	38.70316	0.1089920	-0.2031637	0.0106556	1.056796	0.0000672	-0.
39.0	56.0	Male	One sibling	38.59487	0.0888350	0.4051346	0.0070787	1.056688	0.0001762	0.
39.5	55.0	Male	2 siblings	38.39392	0.1089692	1.1060797	0.0106511	1.055754	0.0019902	1.
34.5	51.5	Female	Singleton	36.40592	0.1037990	-1.9059150	0.0096644	1.053632	0.0053512	-1.

```
ggplot(data = Pred.FinalModelOriginal, aes(x=.resid)) +
  geom_histogram(aes(y=..density..)) +
  stat_function(fun = dnorm, colour = "magenta",
                args = list(mean = mean(Pred.FinalModelOriginal$.resid),
                            sd = sd(Pred.FinalModelOriginal$.resid)))
```

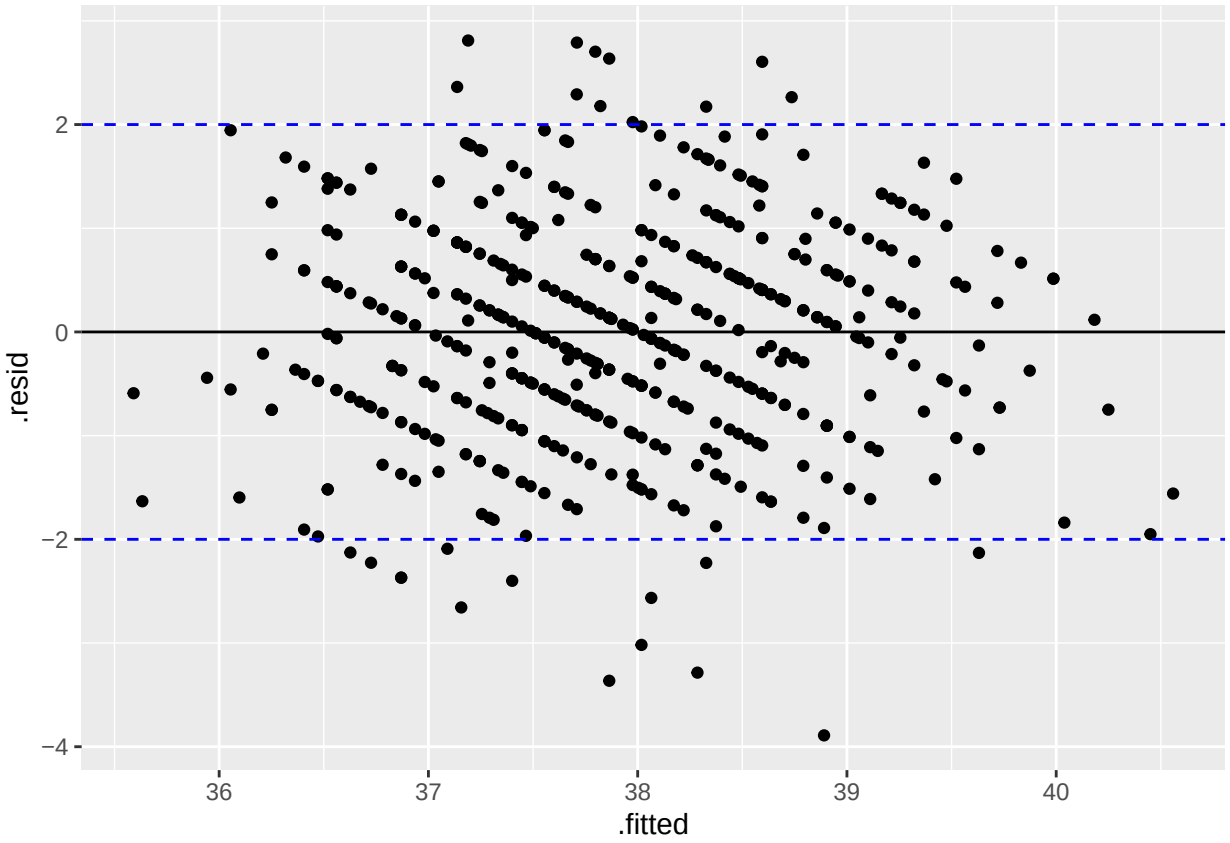


Residual Histogram look normally distributed

Scatter plot between Predicted and Residual value

```
ggplot(data = Pred.FinalModelOriginal, aes(x=.fitted, y=.resid)) +
  geom_point() +
  geom_hline (yintercept = 0) +
  geom_hline (yintercept = 2, linetype=2, color="blue") +
```

```
  geom_hline (yintercept = -2, linetype=2, color="blue")
```



Scatter plot seem linear and constant

Thus, these two assumption for linear regression is fulfilled

1. for each set of Xi values, there is a sub-population of Y values which are normally distributed.
2. the variance of the sub-population of Y are all equal

Final Model with Influence Removed

Final Model with Influence Removed

```
summary(FinalModelNoInfluential)
```

```
##
## Call:
## lm(formula = ModelPrelim, data = NonInfluenObs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12254 -0.64978  0.03099  0.64903  2.05179
##
## Coefficients:
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)              19.95790    1.06077  18.815  < 2e-16 ***
## length                    0.32905    0.01916  17.171  < 2e-16 ***
## genderFemale             -0.55294    0.08704  -6.353 4.95e-10 ***
## parityOne sibling         0.25534    0.09680   2.638  0.00862 **
```

```
## parity2 siblings          0.44252    0.11273   3.926 9.93e-05 ***
## parity3 or more siblings  0.26013    0.21813   1.193  0.23364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9109 on 475 degrees of freedom
## Multiple R-squared:  0.4833, Adjusted R-squared:  0.4779
## F-statistic: 88.87 on 5 and 475 DF,  p-value: < 2.2e-16
```

```
kable(tidy(FinalModelNoInfluential))
```

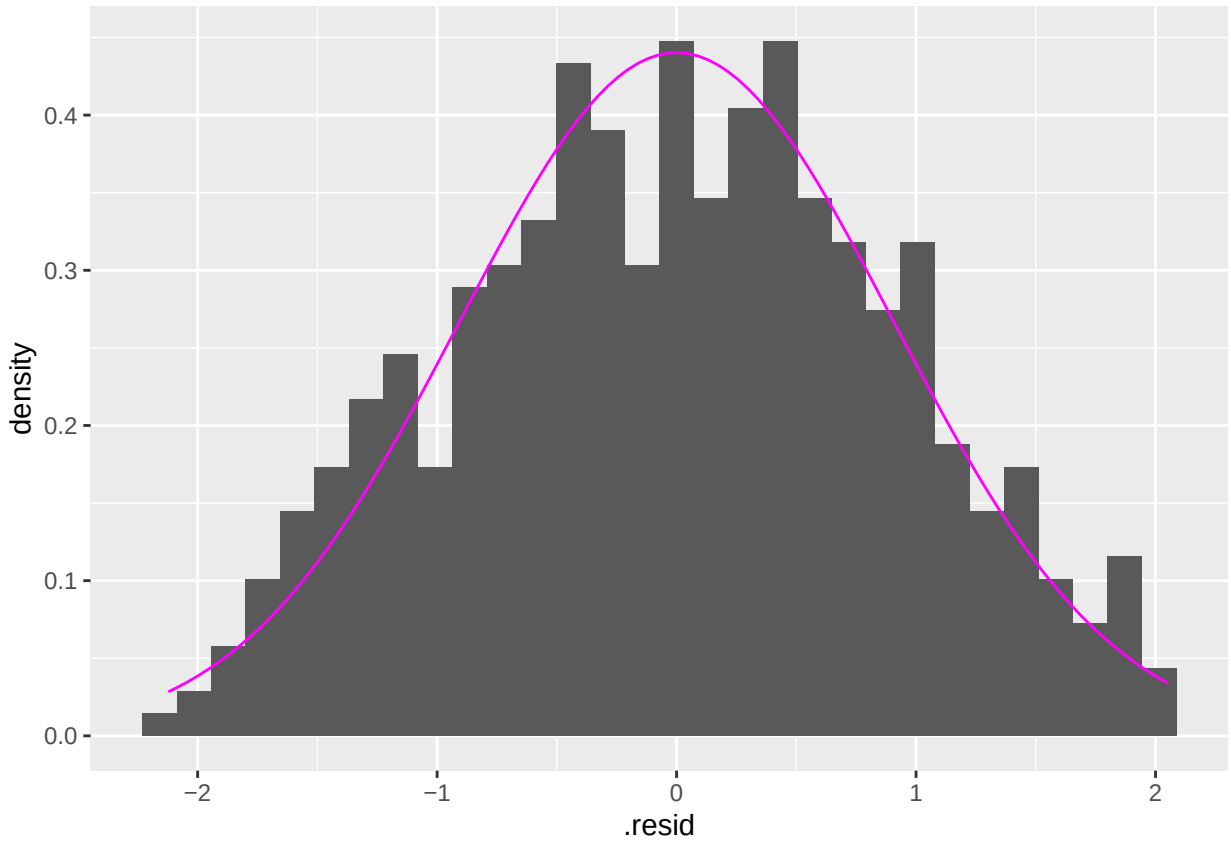
term	estimate	std.error	statistic	p.value
(Intercept)	19.9579008	1.0607695	18.814550	0.0000000
length	0.3290487	0.0191630	17.171016	0.0000000
genderFemale	-0.5529428	0.0870370	-6.352961	0.0000000
parityOne sibling	0.2553355	0.0968039	2.637656	0.0086219
parity2 siblings	0.4425243	0.1127273	3.925617	0.0000993
parity3 or more siblings	0.2601264	0.2181259	1.192551	0.2336404

Residual Histogram

```
Pred.FinalModelNoInfluence <- augment(FinalModelNoInfluential)
kable(head(Pred.FinalModelNoInfluence))
```

headc	length	gender	parity	.fitted	.se.fit	.resid	.hat	.sigma	.cooksd	
37.5	55.5	Female	3 or more siblings	37.92729	0.2043123	-0.4272899	0.0503084	0.9116454	0.0020456	-0
38.5	57.0	Female	Singleton	38.16074	0.0997718	0.3392634	0.0119968	0.9117330	0.0002841	0
38.5	56.0	Male	2 siblings	38.82715	0.0991827	-0.3271550	0.0118556	0.9117425	0.0002610	-0
39.0	56.0	Male	One sibling	38.63997	0.0797026	0.3600338	0.0076559	0.9117167	0.0002024	0
39.5	55.0	Male	2 siblings	38.49811	0.0989888	1.0018938	0.0118093	0.9106920	0.0024383	1
34.5	51.5	Female	Singleton	36.35097	0.0944347	-1.8509685	0.0107477	0.9078526	0.0075579	-2

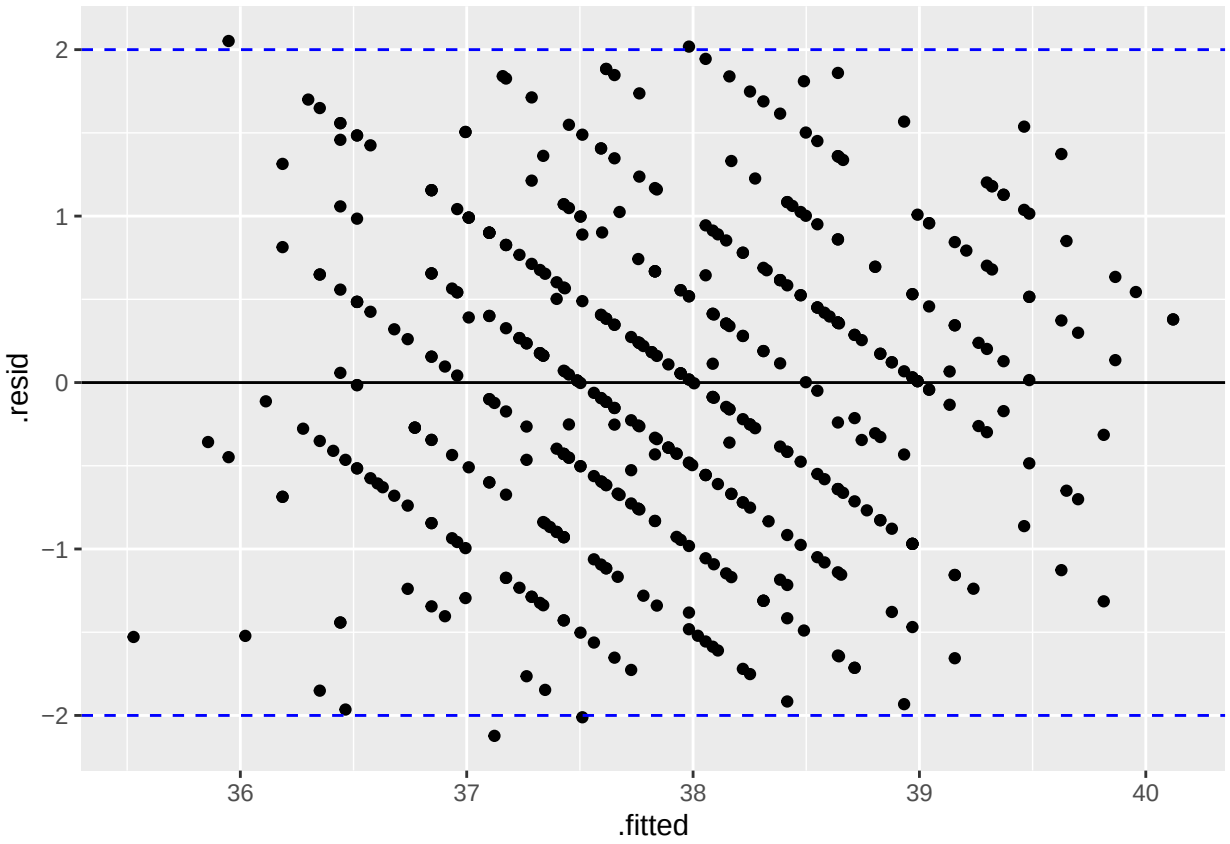
```
ggplot(data = Pred.FinalModelNoInfluence, aes(x=.resid)) +
  geom_histogram(aes(y=..density..)) +
  stat_function(fun = dnorm, colour = "magenta",
                args = list(mean = mean(Pred.FinalModelNoInfluence$.resid),
                            sd = sd(Pred.FinalModelNoInfluence$.resid)))
```



Residual Histogram look normally distributed

Scatter plot between Predicted and Residual value

```
ggplot(data = Pred.FinalModelNoInfluence, aes(x=.fitted, y=.resid)) +
  geom_point() +
  geom_hline (yintercept = 0) +
  geom_hline (yintercept = 2, linetype=2, color="blue") +
  geom_hline (yintercept = -2, linetype=2, color="blue")
```



Scatter plot seem linear and constant

Thus, these two assumption for linear regression is fulfilled

1. for each set of Xi values, there is a sub-population of Y values which are normally distributed.
2. the variance of the sub-population of Y are all equal

Qualitatively, diagnostic plot for Final Model without Influence look *leaner*. Otherwise adjusted R^2 for Final Model without Influence was higher.

Thus the final model will based on new model without influence samples.

Final Model & Interpretation

Final Model

Model

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3$

- X_1 - length
- X_2 - gender
- X_3 - parity

```
summary(FinalModelNoInfluential)
```

```
##
## Call:
## lm(formula = ModelPrelim, data = NonInfluenObs)
##
```

```
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.12254 -0.64978  0.03099  0.64903  2.05179
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)               19.95790    1.06077  18.815  < 2e-16 ***
## length                     0.32905    0.01916  17.171  < 2e-16 ***
## genderFemale              -0.55294    0.08704  -6.353 4.95e-10 ***
## parityOne sibling          0.25534    0.09680   2.638  0.00862 **
## parity2 siblings           0.44252    0.11273   3.926 9.93e-05 ***
## parity3 or more siblings   0.26013    0.21813   1.193  0.23364
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9109 on 475 degrees of freedom
## Multiple R-squared:  0.4833, Adjusted R-squared:  0.4779
## F-statistic: 88.87 on 5 and 475 DF,  p-value: < 2.2e-16
```

```
kable(tidy(FinalModelNoInfluential, conf.int = T))
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	19.9579008	1.0607695	18.814550	0.0000000	17.8735197	22.0422819
length	0.3290487	0.0191630	17.171016	0.0000000	0.2913939	0.3667035
genderFemale	-0.5529428	0.0870370	-6.352961	0.0000000	-0.7239680	-0.3819176
parityOne sibling	0.2553355	0.0968039	2.637656	0.0086219	0.0651186	0.4455525
parity2 siblings	0.4425243	0.1127273	3.925617	0.0000993	0.2210184	0.6640302
parity3 or more siblings	0.2601264	0.2181259	1.192551	0.2336404	-0.1684847	0.6887375

head circumference = 19.96 + 0.33(length) - 0.55(female = 1) + 0.26(One sibling = 1) + 0.44(2 siblings = 1) + 0.26(3 or more siblings = 1)

Model Interpretation

Multiple Linear Regression Analysis show that length, gender and parity had singificant linear relationship to head circumference. The model explains 48.8% of variation of head circumference in the study sample.

The relationship can be explained as below

- those with 1 cm length increment have increment of 0.33cm (95% CI = 0.29, 0.37) head circumference when adjusted to gender and parity
- When comparing female against male, female baby will have 0.55cm (95% CI = 0.38, 0.72) lower head circumference, when adjusted to length and parity
- when comparing one sibling to singleton baby, one sibling will have 0.26cm (95% CI = 0.07, 0.46) higher head circumference, when adjusted to length and gender
- when comparing two sibling to singleton baby, two sibling will have 0.44cm (95% CI = 0.22, 0.66) higher head circumference, when adjusted to length and gender
- when comparing 3 or more sibling to singleton baby, 3 or more sibling will have 0.17cm (95% CI = -0.17, 0.69) higher head circumference, when adjusted to length and gender

Model assumptions were met. There were no interaction nor multicollinearity between independent variables.

Low coefficient of determination, R^2 is most likely due to underfitting. There might be unexplained variation, which can be exlained by other variables outside of the scope of this study. But for this research purpose,

with available variables, the model stated above is the best model available.

Prediction

Prediction with New Dataset

Create New Dataset

new data was created using these parameter

- length - 45cm, mean newborn's length, 65cm
- gender - both male and female
- parity - singleton, one sibling, 2 siblings and 3 or more siblings

```
DS_cons <- expand.grid(length = c(45, mean(NonInfluenObs$length), 65),
                       gender = c('Male', 'Female'),
                       parity = c('Singleton', 'One sibling','2 siblings', '3 or more siblings'))
kable(head(DS_cons))
```

length	gender	parity
45.00000	Male	Singleton
54.73306	Male	Singleton
65.00000	Male	Singleton
45.00000	Female	Singleton
54.73306	Female	Singleton
65.00000	Female	Singleton

```
kable(tail(DS_cons))
```

	length	gender	parity
19	45.00000	Male	3 or more siblings
20	54.73306	Male	3 or more siblings
21	65.00000	Male	3 or more siblings
22	45.00000	Female	3 or more siblings
23	54.73306	Female	3 or more siblings
24	65.00000	Female	3 or more siblings

Calculate Predicted Values with New Dataset

```
DataSetPred.NewDS <- augment(FinalModelNoInfluential, newdata = DS_cons)
kable(head(DataSetPred.NewDS))
```

length	gender	parity	.fitted	.se.fit
45.00000	Male	Singleton	34.76509	0.2115272
54.73306	Male	Singleton	37.96774	0.0816774
65.00000	Male	Singleton	41.34607	0.2047319
45.00000	Female	Singleton	34.21215	0.1906207
54.73306	Female	Singleton	37.41480	0.0826787
65.00000	Female	Singleton	40.79313	0.2261018

length	gender	parity	.fitted	.se.fit

```
kable(tail(DataSetPred.NewDS))
```

length	gender	parity	.fitted	.se.fit
45.00000	Male	3 or more siblings	35.02522	0.3041620
54.73306	Male	3 or more siblings	38.22787	0.2176069
65.00000	Male	3 or more siblings	41.60620	0.2740778
45.00000	Female	3 or more siblings	34.47228	0.2797485
54.73306	Female	3 or more siblings	37.67493	0.2041311
65.00000	Female	3 or more siblings	41.05325	0.2801387