Biostatistics **Student number: 21626103**

Name: Nicholas Donnelly

Assignment 2 (Longitudinal Data Analysis)

1. BACKGROUND/RATIONALE FOR THE STUDY

Infants that are born to HIV positive mothers would require early HIV diagnostics as well as immediate treatment. The objective of this study was to determine the evolution of the length of infants born to HIV positive mothers and to study the effects of various explanatory variables on their evolution.

2. RESEARCH METHODOLOGY

A Longitudinal data analysis was conducted on this study, focusing on infants that had two or more measurements taken from 571 infants. For significance purposes, two explanatory variables were removed from this study due to insignificant p-values from various general linear models. These variables included the maternal viral load group, as well as whether the mother was started on ARV treatment within four weeks of delivery. Various frequency plots were implemented in order to show the relationship between an infant that was born prematurely, being born below normal body weight, or whether the mother was symptomatic at the time of birth. A summary statistics table was also created to identify various statistics such as the mean, standard deviation as well as the maximum and minimum values between length and age of the infants. After careful consideration of different summary statistics and distribution models, a sample was taken from the population and only infants aged 0-300 days old were used for research purposes as there weren't enough infants born above 300 that would have made an impact to the dataset.

3. Results

It was concluded that a mother with HIV did not in fact contribute to a difference in the infant's length overtime. Instead, we found that a child born prematurely or under the below normal body weight had an increasingly smaller length overtime.

Appendix

Summary Statistics

```
/*creates an output of 2 or more measurements per patient*/

proc sort data= lendata
            out= lent_dat nouniquekey;
by ID;
proc print data=lent_dat;
run;

/*   identify how many patients are in the dataset that have two or more
measurements */

title "Number of infants per two measurements";
proc sql;
   create table new as
     select count(distinct(ID)) as Patient
            from lent_dat;
                  proc print data=new;
quit;
```

Number of infants per two measurements

Obs	Patient
1	571

Figure 1.1

```
/* identify significant p-values */


proc glm data = lent_dat;
      class Preamture;
            model Premature =  TreatLess4Weeks VLGroup Symptomatic
      LowBWeight ;
output out = lent_Pr;
run;
```

The GLM Procedure

Dependent Variable: Premature

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	8.66723702	2.16680925	21.42	<.0001
Error	229	23.16182281	0.10114333		
Corrected Total	233	31.82905983			

R-Square	Coeff Var	Root MSE	Premature Mean
0.272306	195.8398	0.318030	0.162393

Source	DF	Type I SS	Mean Square	F Value	Pr > F
TreatLess4Weeks	1	0.04066697	0.04066697	0.40	0.5267
VLGroup	1	0.05421894	0.05421894	0.54	0.4648
Symptomatic	1	3.29123231	3.29123231	32.54	<.0001
LowBWeight	1	5.28111879	5.28111879	52.21	<.0001

Source	DF	Type III SS	Mean Square	F Value	Pr > F
TreatLess4Weeks	1	0.01333037	0.01333037	0.13	0.7169
VLGroup	1	0.01481317	0.01481317	0.15	0.7023
Symptomatic	1	0.00502236	0.00502236	0.05	0.8239
LowBWeight	1	5.28111879	5.28111879	52.21	<.0001

Parameter	Estimate	Standard Error	t Value	Pr > |t|
Intercept	0.1028868837	0.04724135	2.18	0.0304
TreatLess4Weeks	-.0378085953	0.10414492	-0.36	0.7169
VLGroup	-.0106077049	0.02771827	-0.38	0.7023
Symptomatic	-.0192073265	0.08619491	-0.22	0.8239
LowBWeight	0.5613712150	0.07768833	7.23	<.0001

Figure 1.2

Interpretation:

From the output above, it is clear that we can remove two explanatory variables, these being VLGroup and TreatLess4Weeks. This is evident due to their p-values being greater than 0.05. On the other hand, Premature, Symptomatic and LowBWeight have p-values less than 0.05 making these explanatory variables highly significant.

```
/*drop unwanted variables from the original data set */

data len_main;
            set lent_dat;
      drop Weight HeadCr VLGroup TreatLess4Weeks;
      proc print data= len_main;
      run;

/*   identifying the percentage of babies born prematurely, had a
symptomatic mother as well as being born lower than the normal birth weight
*/

ods graphics on;
      title "Percentage of babies born lower than the normal birth weight";
      proc freq data=len_main order=freq nlevels;
            tables LowBWeight*premature*Symptomatic / nocum
      plots=freqplot(orient=horizontal scale= percent);
run;
```

Frequency Percent Row Pct Col Pct	Table 1 of Premature by Symptomatic		
	Controlling for LowBWeight=0		
	Symptomatic		
Premature	0	1	Total
0	362 89.83 97.05 92.82	11 2.73 2.95 84.62	373 92.56
1	28 6.95 93.33 7.18	2 0.50 6.67 15.38	30 7.44
Total	390 96.77	13 3.23	403 100.00
Frequency Missing = 27			

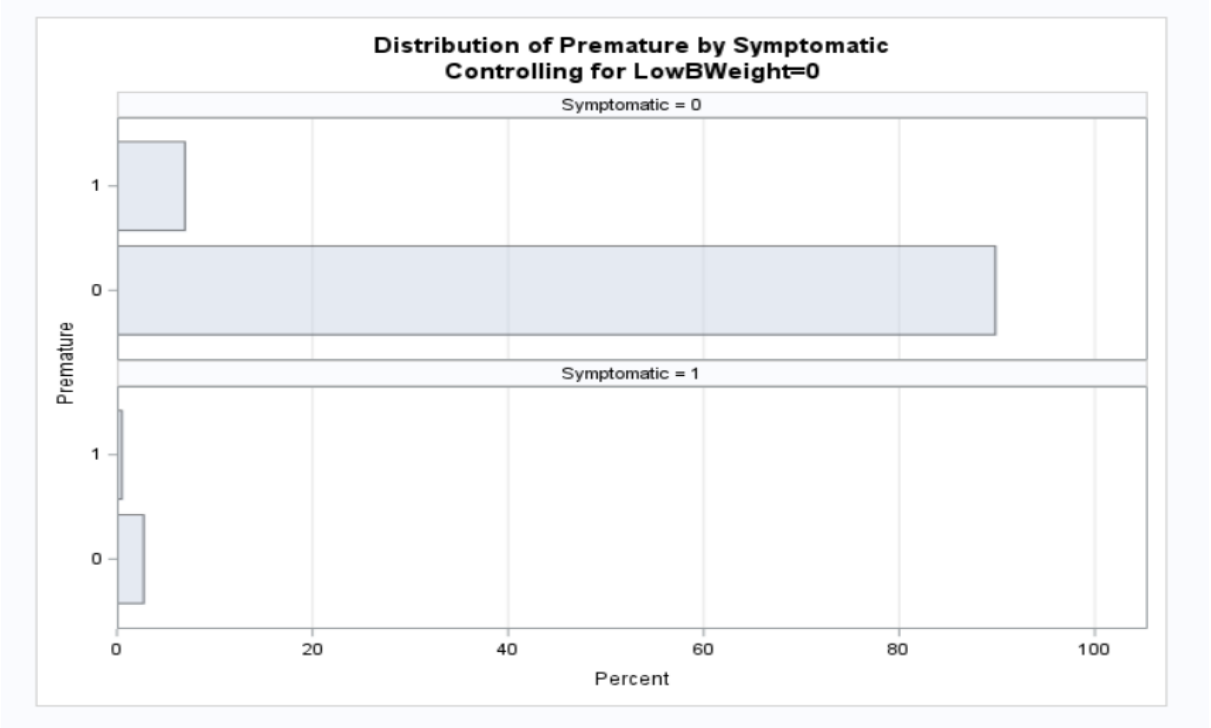


Figure 1.3

Interpretation:

Table 1: Given that the infant was born at a normal birth weight

- 6.95 percent of infants were premature given that the mother was not symptomatic
- 89.83 percent of infants were not premature given that the mother was not symptomatic
- 0.50 percent of infants were premature given that the mother was symptomatic
- 2.73 percent of infants were not premature given that the mother was symptomatic

We can therefore conclude that if the infant was born at a normal birth weight, majority of the infants had a non-symptomatic mother and were not born prematurely

Frequency Percent Row Pct Col Pct	Table 2 of Premature by Symptomatic		
	Controlling for LowBWeight=1		
	Symptomatic		
Premature	0	1	Total
0	17 17.17 37.78 45.95	28 28.28 62.22 45.16	45 45.45
1	20 20.20 37.04 54.05	34 34.34 62.96 54.84	54 54.55
Total	37 37.37	62 62.63	99 100.00
Frequency Missing = 5			

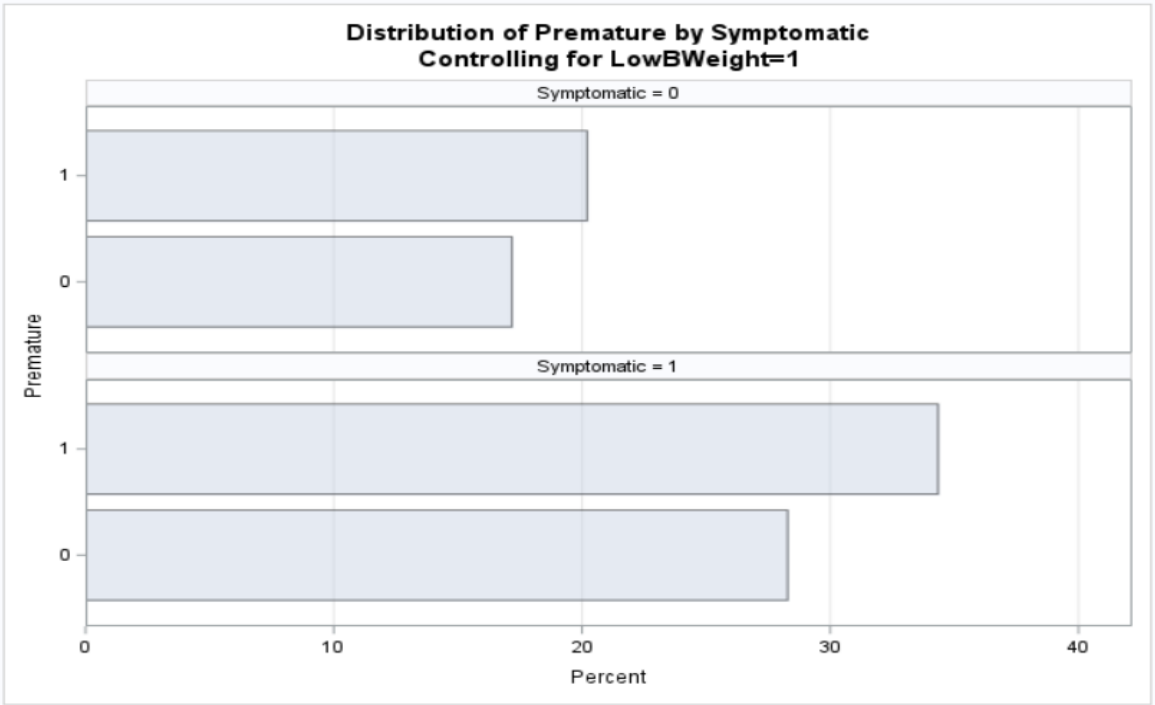


Figure 1.4

Interpretation:

Table 2: Given that the infant was born below the normal birth weight

- 20.20 percent of infants were premature given that the mother was not symptomatic
- 17.17 percent of infants were not premature given that the mother was not symptomatic
- 34.34 percent of infants were premature given that the mother was symptomatic

- 28.28 percent of infants were not premature given that the mother was symptomatic

We can therefore conclude that if the infant was born below the normal birth weight, majority of the infants had a symptomatic mother and were born prematurely.

```
/*  plot summary statistics  */

proc means data=len_main ;
var length age;
run;
```

The MEANS Procedure

Variable	N	Mean	Std Dev	Minimum	Maximum
Length	2378	61.2377628	6.7056544	34.0000000	80.8000000
Age	2935	124.6918789	110.2271074	0	528.1875000

Figure 1.5

Interpretation:

- The average length is 61.24 which gives a sense that that data could be normally distributed since it lies between the maximum and minimum values.
- The average age is 125 days old with a standard deviation of 110.23. Due to the average being significantly lower than the maximum age of 528 days and a low standard deviation, it is required that we extract a portion of age that will provide us with the best results when doing further analysis.

```
  /* extacting which age to use */

data len_new;
      set len_main;
      if age>=300 then delete;
 proc print data=len_new;
 run;


/* Distribution of Age*/

data lent_age;
set len_new;
if age in ("0") then delete;
run;
```

```
title 'Distribution of  Age';
proc sgplot data=lent_age;
 histogram Age;
 density age / type=normal ;
 density age/ type=kernel;
 keylegend / location=inside
 position=topright across=1;
 run;
```

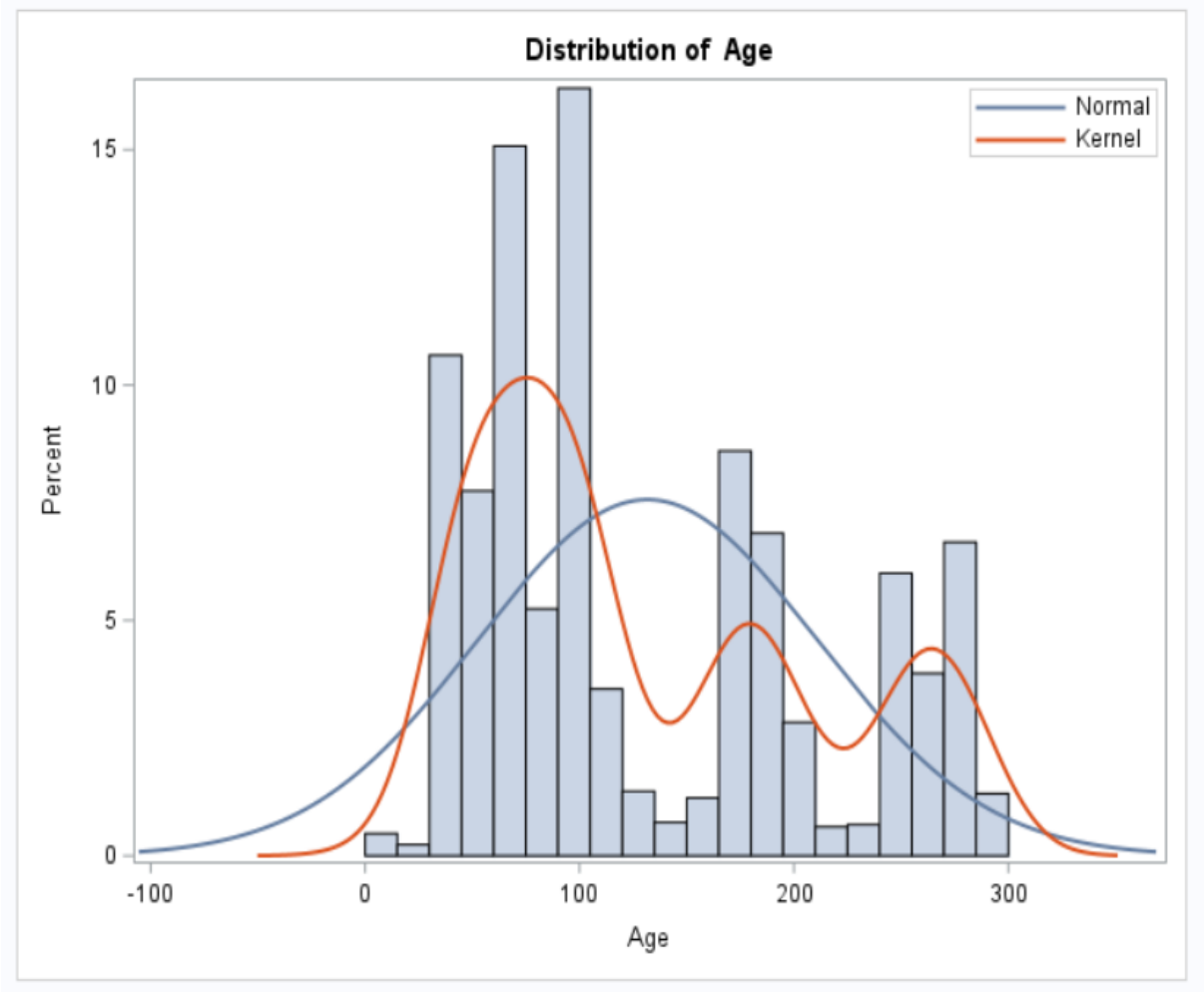


Figure 1.6

```
  /* Age by Sex */

title 'Age by Sex';
proc sgpanel data=len_gen;
 panelby gender / layout=columnlattice
 onepanel novarname;
 histogram age /
 fillattrs=graphdata3;
 run;
```

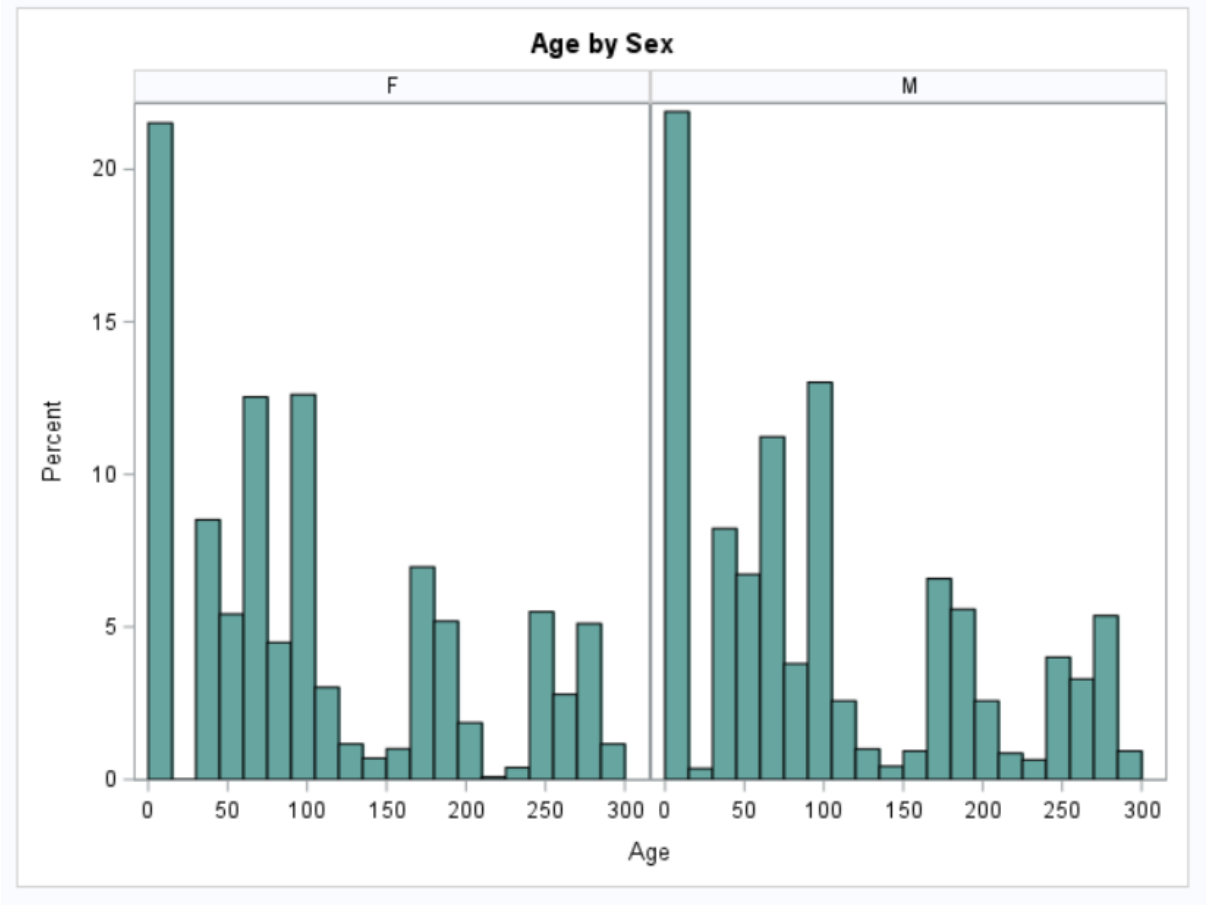


Figure 1.7

Interpretation:

After extracting the age from 0-300 days old we can then plot the new distribution. The distribution of age is also skewed to the right (positively skewed) giving us an indication that most of the infants age is clustered around the left tail of the distribution. This is also evident when we model the distribution per male and female.

```
/* Distribution of Length*/

title 'Distribution of  length';
      proc sgplot data=lent_dat;
            histogram length;
 density length / type=normal ;
 density length/ type=kernel;
      keylegend / location=inside
            position=topright across=1;
 run;
```

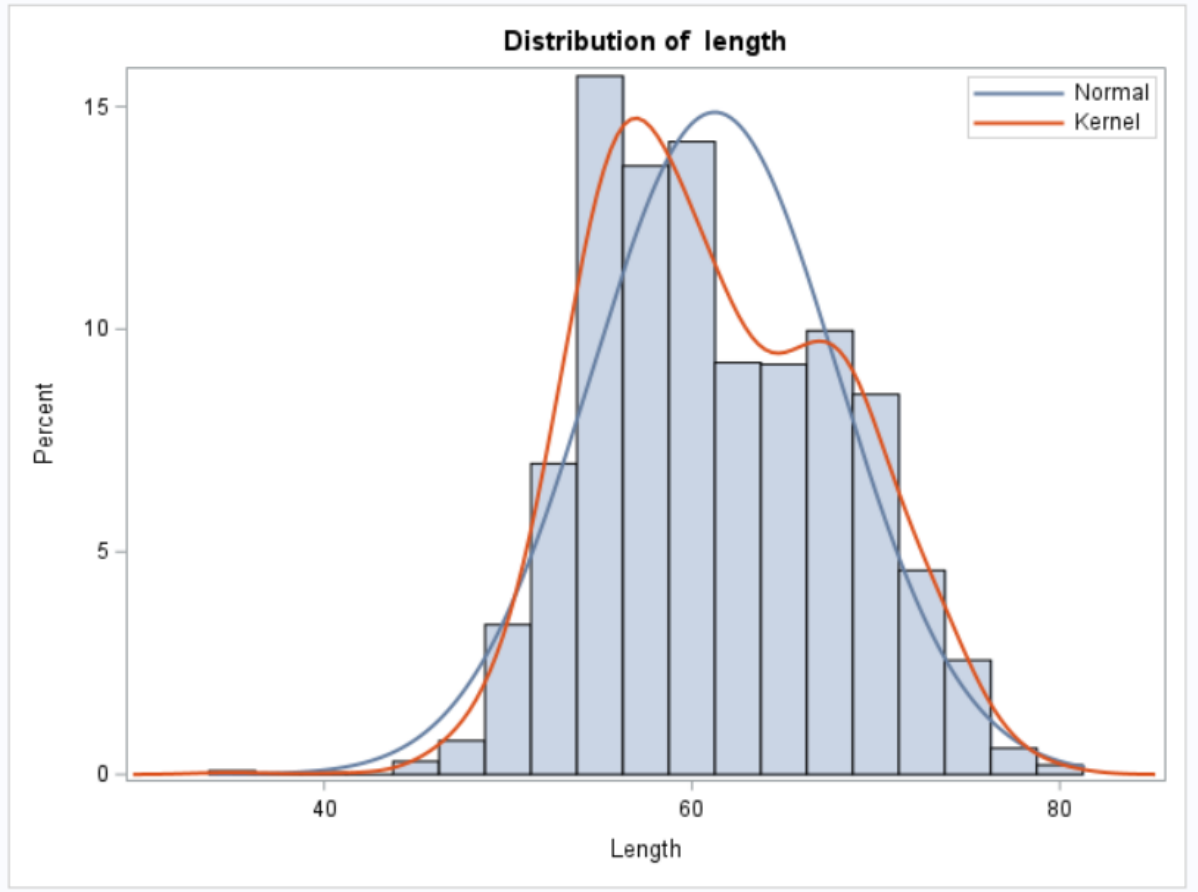


Figure 1.8

```
data len_gen;
      set len_new;
retain gender_new;
by id;
if first.id then gender_new=gender;
else gender=gender_new;
drop gender_new;
```

```
proc print data=len_gen;
run;

/* Length by Sex */

title 'Length by Sex';
proc sgpanel data=len_gen;
 panelby gender / layout=columnlattice
 onepanel novarname;
 histogram length /
 fillattrs=graphdata3;
 run;
```

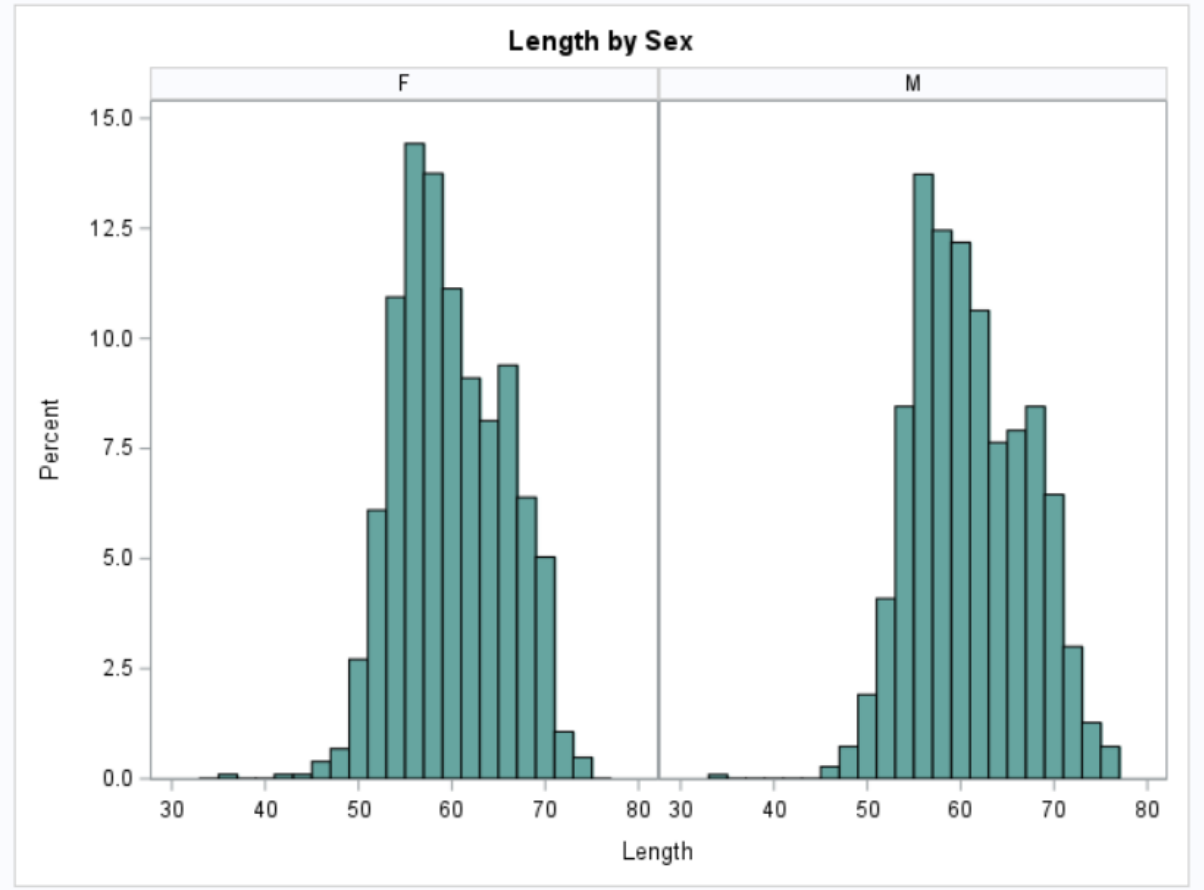


Figure 1.9

Interpretation:

Similar to that of age, the distribution of length is also slightly skewed to the right (positively skewed) giving us an indication that most of the infants have a length clustered around the left tail of the distribution. This is also evident when we model the distribution per male and female.

Formulating a Plausible Random-Effects Model:

Before plotting the model, it is important to identify the relationship between the explanatory variables and the length of each infant given its age.

Code:

```
data len_pr;
      set len_gen;
retain pr_new;
by id;
if first.id then pr_new=premature;
else premature=pr_new;
drop pr_new;
run;

data len_sym;
      set len_pr;
retain sym_new;
by id;
if first.id then sym_new=Symptomatic;
else Symptomatic=sym_new;
drop sym_new;
run;

data len_exp;
      set len_sym;
retain Lowbw_new;
by id;
if first.id then Lowbw_new=LowBWeight;
else LowBWeight=Lowbw_new;
drop Lowbw_new;
proc print data=len_exp;
run;

data loes_main_1;
set len_exp;
where Premature=1 and Symptomatic=1 and LowBWeight=1;
proc print data=loes_main_1;
run;
```

```sas
data loes_main_2;
set len_exp;
where Premature=0 and Symptomatic=1 and LowBWeight=1;
proc print data=loes_main_2;
run;


data loes_main_3;
set len_exp;
where Premature=1 and Symptomatic=0 and LowBWeight=1 ;
proc print data=loes_main_3;
run;

data loes_main_5;
set len_exp;
where Premature=0 and Symptomatic=0 and LowBWeight=1;
proc print data=loes_main_5;
run;


data loes_main_6;
set len_exp;
where Premature=1 and Symptomatic=0 and LowBWeight=0 ;
proc print data=loes_main_6;
run;


data loes_main_7;
set len_exp;
where Premature=0 and Symptomatic=1 and LowBWeight=0 ;
proc print data=loes_main_7;
run;


data loes_main_8;
set len_exp;
where Premature=0 and Symptomatic=0 and LowBWeight=0 ;
proc print data=loes_main_8;
run;
```

```
* Plot individual profiles(1);
proc sort data = loes_main_1;
by group;
run;
goptions reset=all ftext=simplex rotate=landscape i=join;
option nobyline;
proc gplot data = loes_main_1;
plot length*age=id/haxis=axis1 vaxis=axis2 nolegend;
axis1 label=(h=2 'Age (days)');
axis2 label=(h=2 A=90 'Length');
title1 h=2 'Premature=1 Symptomatic=1 LowBWeight=1';
run; quit;


* Plot individual profiles(2);
proc sort data = loes_main_2;
by group;
run;
goptions reset=all ftext=simplex rotate=landscape i=join;
option nobyline;
proc gplot data = loes_main_2;
plot length*age=id/haxis=axis1 vaxis=axis2 nolegend;
axis1 label=(h=2 'Age (days)');
axis2 label=(h=2 A=90 'Length');
title1 h=2 'Premature=0 Symptomatic=1 LowBWeight=1';
run; quit;


* Plot individual profiles(3);
proc sort data = loes_main_3;
by group;
run;
goptions reset=all ftext=simplex rotate=landscape i=join;
option nobyline;
proc gplot data = loes_main_3;
plot length*age=id/haxis=axis1 vaxis=axis2 nolegend;
axis1 label=(h=2 'Age (days)');
axis2 label=(h=2 A=90 'Length');
title1 h=2 'Premature=1 Symptomatic=0 LowBWeight=1';
```

```
run; quit;

* Plot individual profiles(5);
proc sort data = loes_main_5;
by group;
run;
goptions reset=all ftext=simplex rotate=landscape i=join;
option nobyline;
proc gplot data = loes_main_5;
plot length*age=id/haxis=axis1 vaxis=axis2 nolegend;
axis1 label=(h=2 'Age (days)');
axis2 label=(h=2 A=90 'Length');
title1 h=2 'Premature=0 Symptomatic=0 LowBWeight=1';
run; quit;

* Plot individual profiles(6);
proc sort data = loes_main_6;
by group;
run;
goptions reset=all ftext=simplex rotate=landscape i=join;
option nobyline;
proc gplot data = loes_main_6;
plot length*age=id/haxis=axis1 vaxis=axis2 nolegend;
axis1 label=(h=2 'Age (days)');
axis2 label=(h=2 A=90 'Length');
title1 h=2 'Premature=1 Symptomatic=0 LowBWeight=0';
run; quit;


* Plot individual profiles(7);
proc sort data = loes_main_7;
by group;
run;
goptions reset=all ftext=simplex rotate=landscape i=join;
option nobyline;
proc gplot data = loes_main_7;
plot length*age=id/haxis=axis1 vaxis=axis2 nolegend;
axis1 label=(h=2 'Age (days)');
axis2 label=(h=2 A=90 'Length');
title1 h=2 'Premature=0 Symptomatic=1 LowBWeight=0';
run; quit;
```

Output:

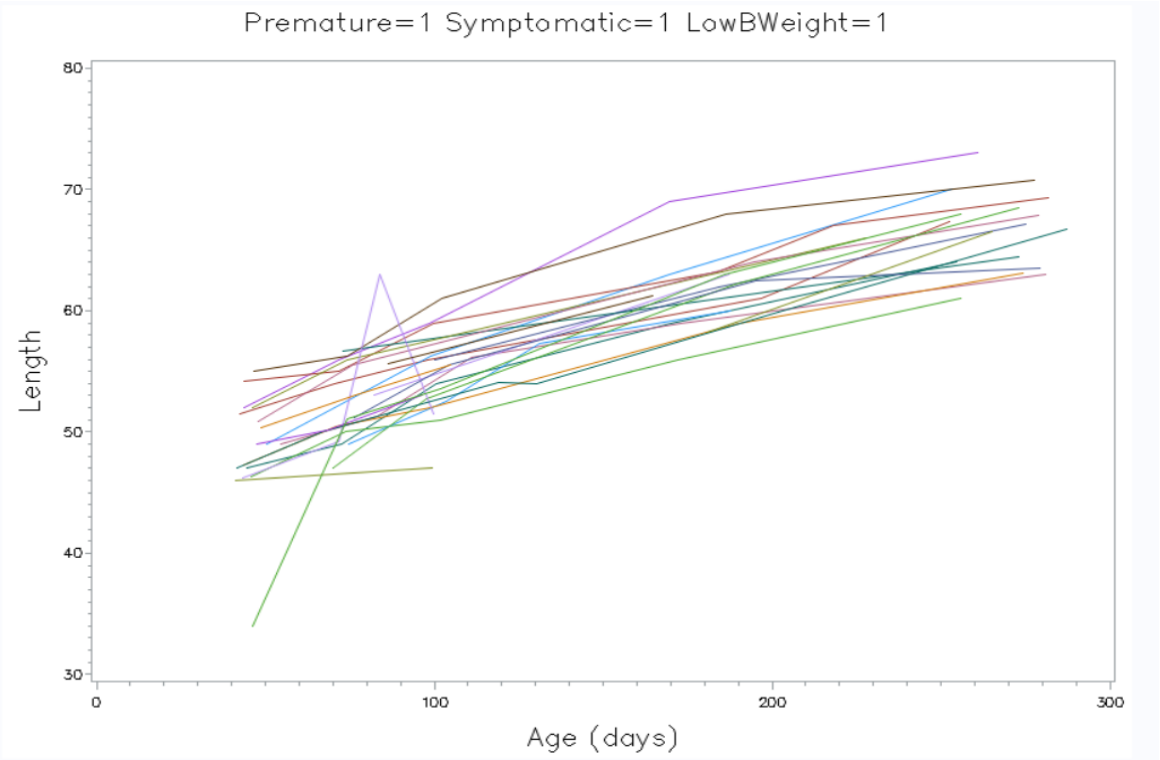


Figure 2.1

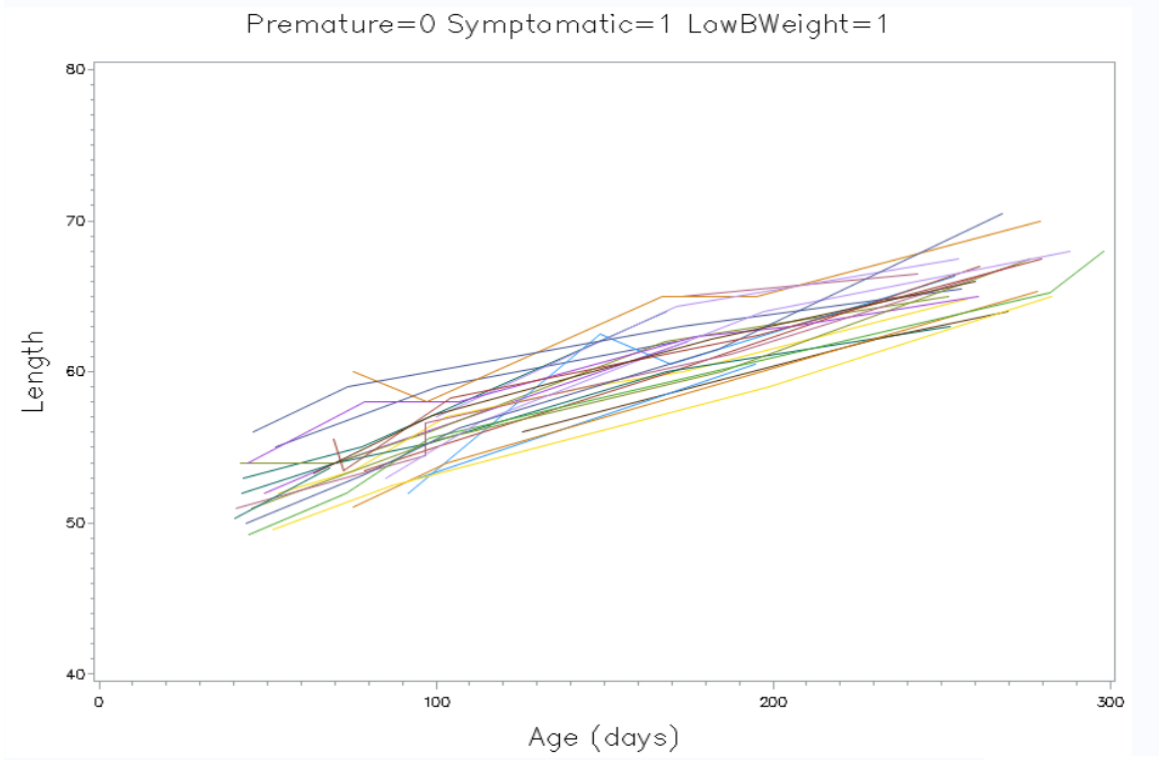


Figure 2.2

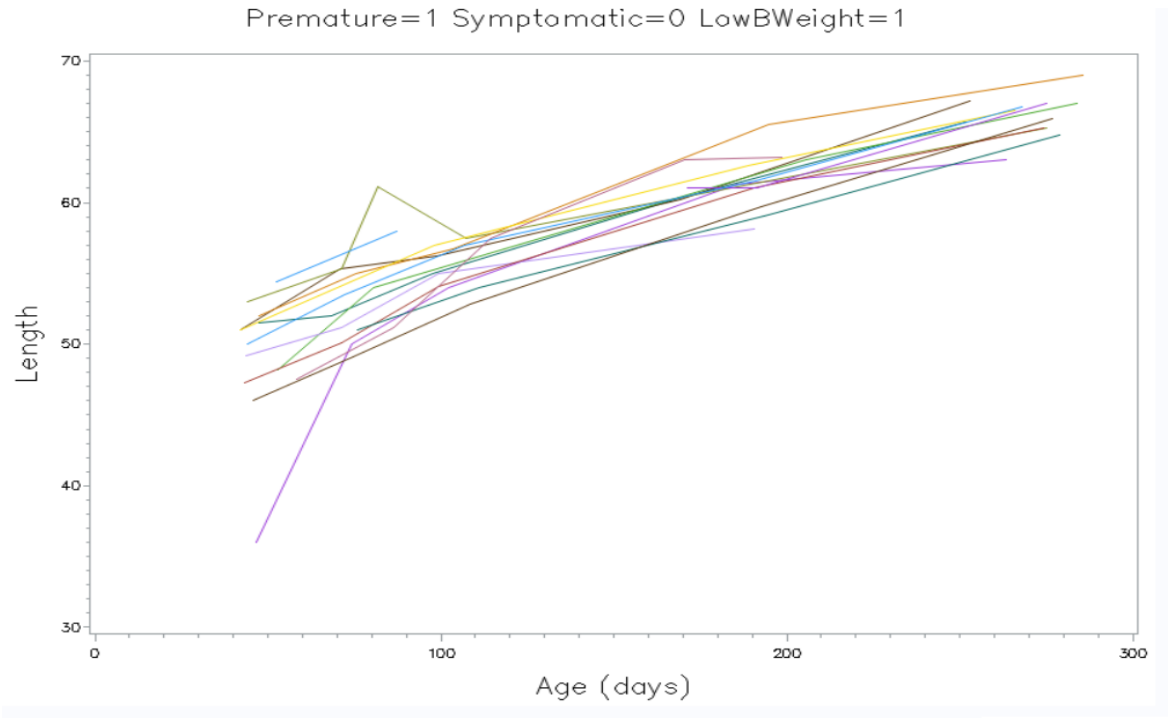


Figure 2.3

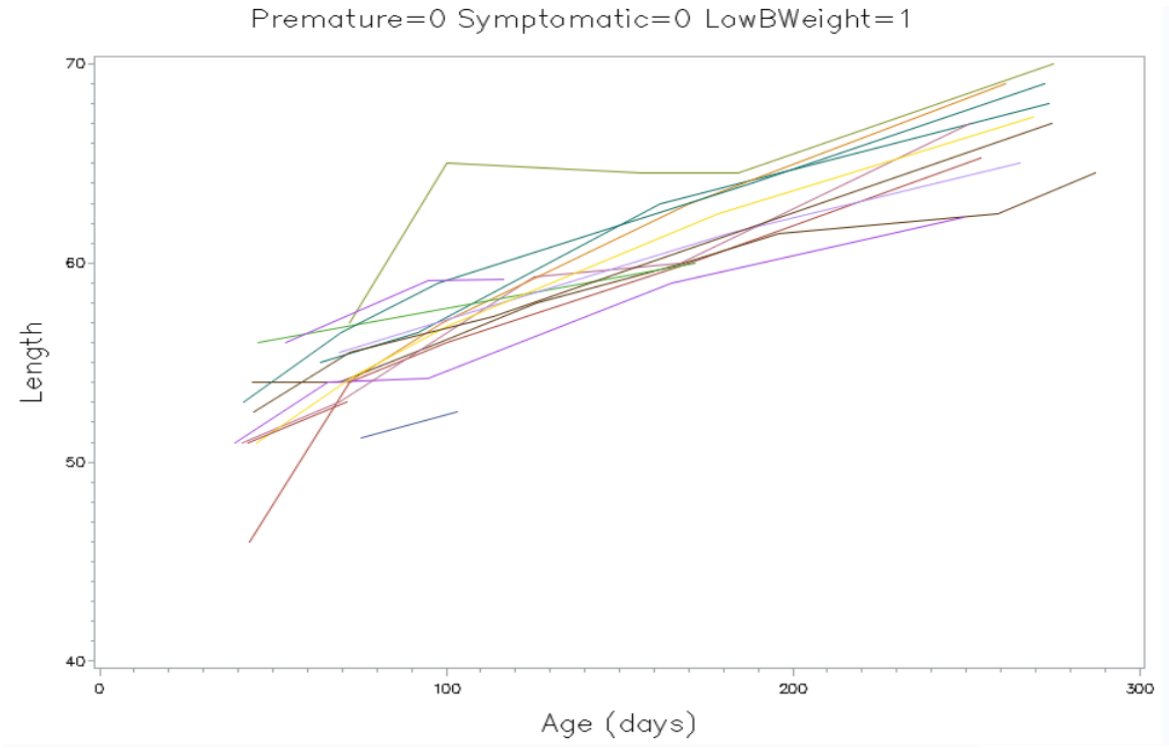


Figure 2.4

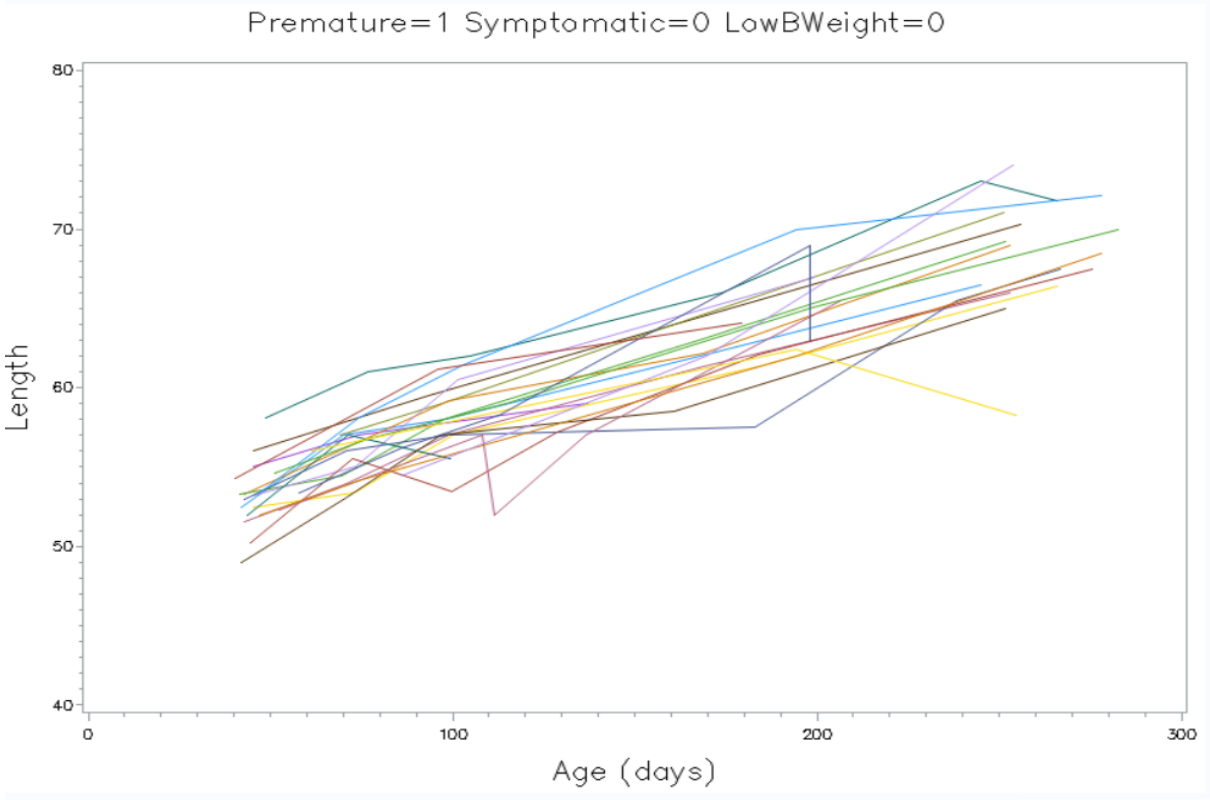


Figure 2.5

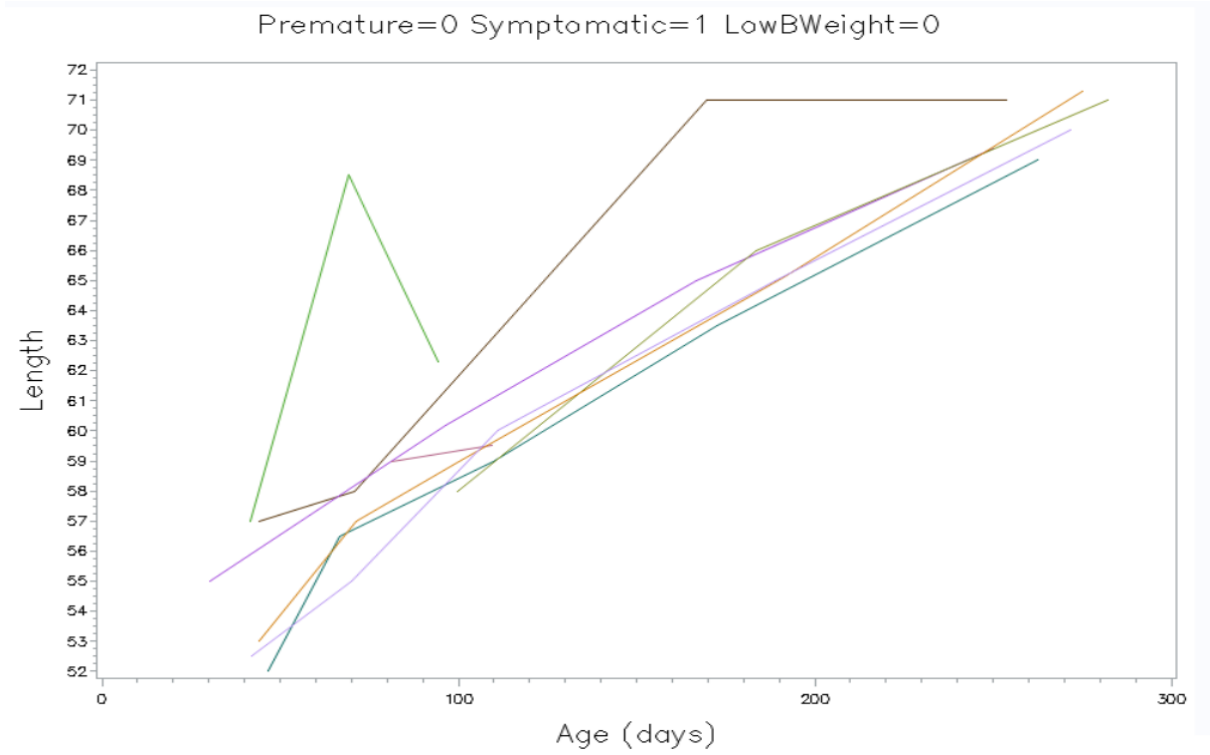


Figure 2.6

Interpretation:

From above, it is evident that the explanatory variables can impact the infant's growth in terms of length over time. Majority of the results show that as the infant gets older, the length of the infant gets bigger. *Figure 2.4* is the only figure above that does not show a drop in length of a single infant over time whereas in the other figures there is at least one infant that has a drop in length. We can therefore say that an infant born prematurely to a symptomatic mother will not necessarily have a decrease in its length over time.

Another important observation to take notice of is the length range of the infants over the days. The two biggest differences were that of *figure 2.3 and 2.6*. In *figure 2.3*, the length of the infants are fairly low in comparison to that of the other graphs with a minimum of around 35cm and a maximum of 68cm. In *figure 2.6*, the length of the infants are fairly high with a minimum of around 52cm and a maximum of 71cm. We can therefore say that infants born prematurely and under the below normal body weight will have a smaller length and growth over time, whereas an infant that had a symptomatic mother but no other defects will grow faster in length over time.

Exploring the mean structure of the dataset:

Once the data has been manipulated, we need to determine whether the data is balanced or unbalanced. If the data is balanced, averages can be calculated for each occasion separately and standard errors for the means can be added.

```
* Average evolution with standard error bars;
goptions reset=all ftext=simplex rotate=landscape;
proc gplot data=len_exp;
plot length*age/ haxis=axis1 vaxis=axis2;
symbol c=blue i =std1mjt w=2 mode=include;
axis1 label=(h=2 'Age (Days)') order=(0 to 300 by 60) minor=none;
axis2 label=(h=2 A=90 'Length') minor=none;
title h=2 'Average evolution, with standard errors of means';
run; quit;
```

Figure 2.7

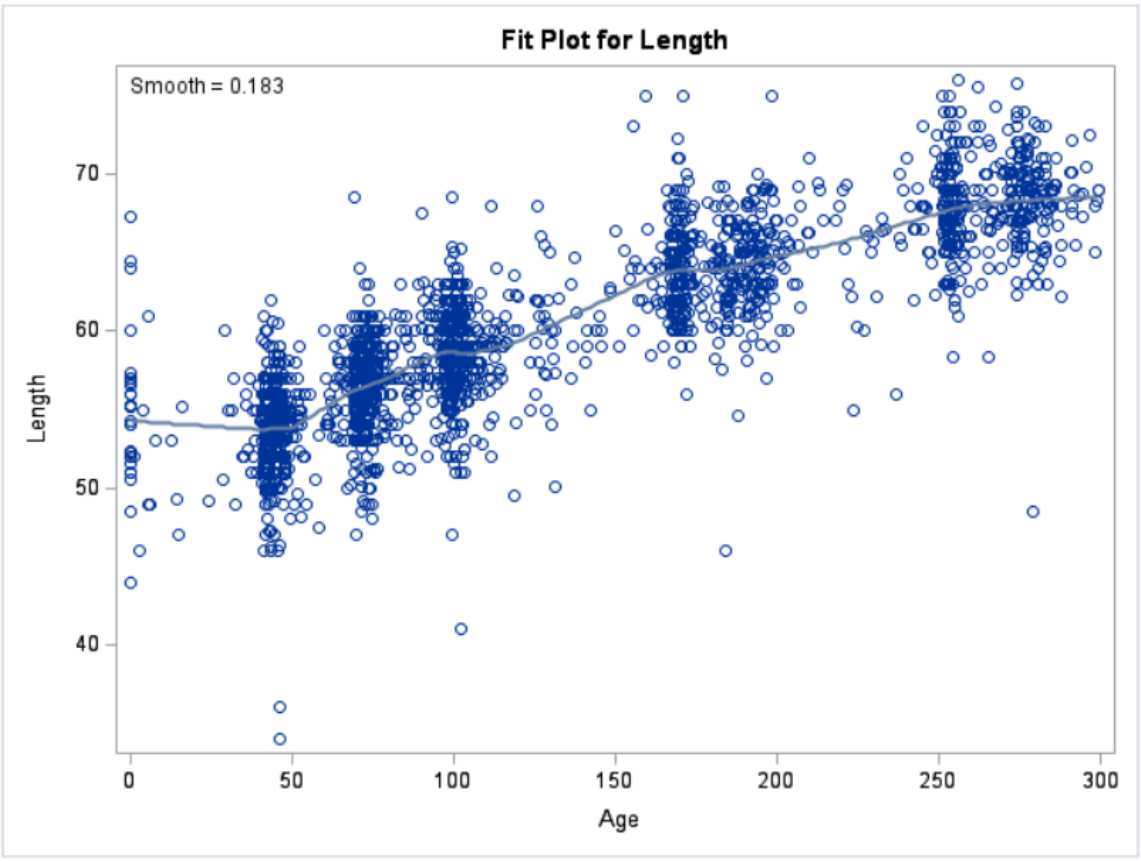
Interpretation: It is clear that the data is unbalanced. Therefore we need discretize the age scale and use sample averaging within intervals. This can be done using smoothing techniques to estimate the average evolution non-parametrically.

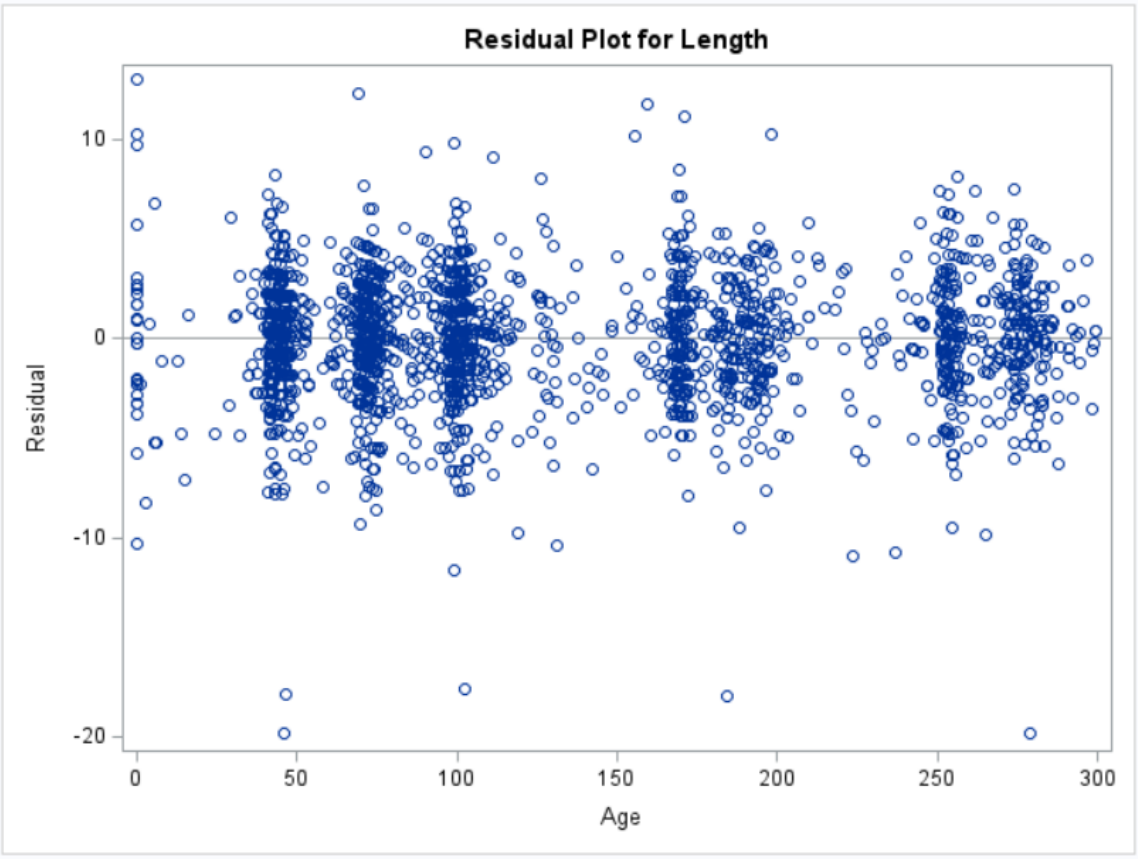
```
* Loess smooth overall;
proc loess data = len_exp;
ods output scoreresults=out;
model length=age;
score data=len_exp;
run;
proc sort data=out;
by age;
run;
goptions reset=all ftext=simplex rotate=landscape;
proc gplot data=out;;
plot p_length*age / overlay haxis = axis1 vaxis = axis2;
symbol1 c=red i=join w=2;
axis1 label=(h=2 'Age') minor=none;
axis2 label=(h=2 a=90 'length') minor=none;
title h=3 'Loess Smoothing';
run; quit;
```

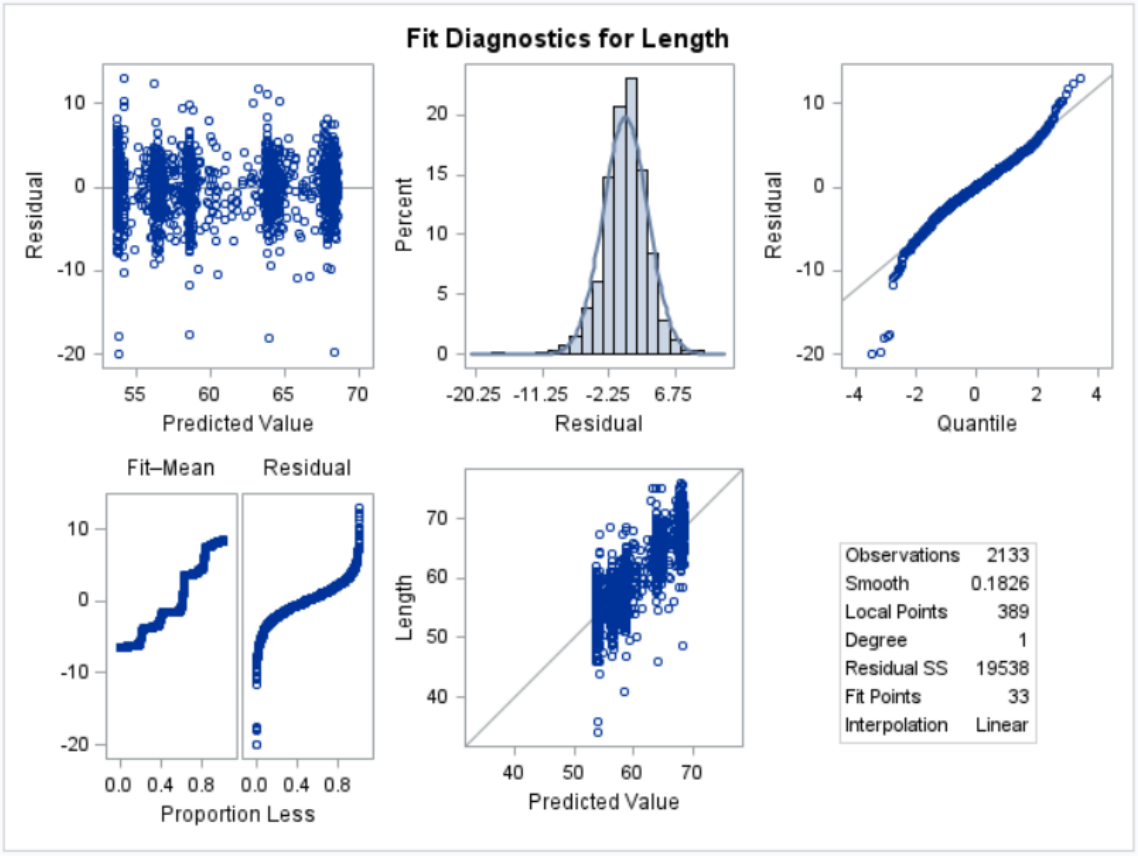
Loess Smoothing

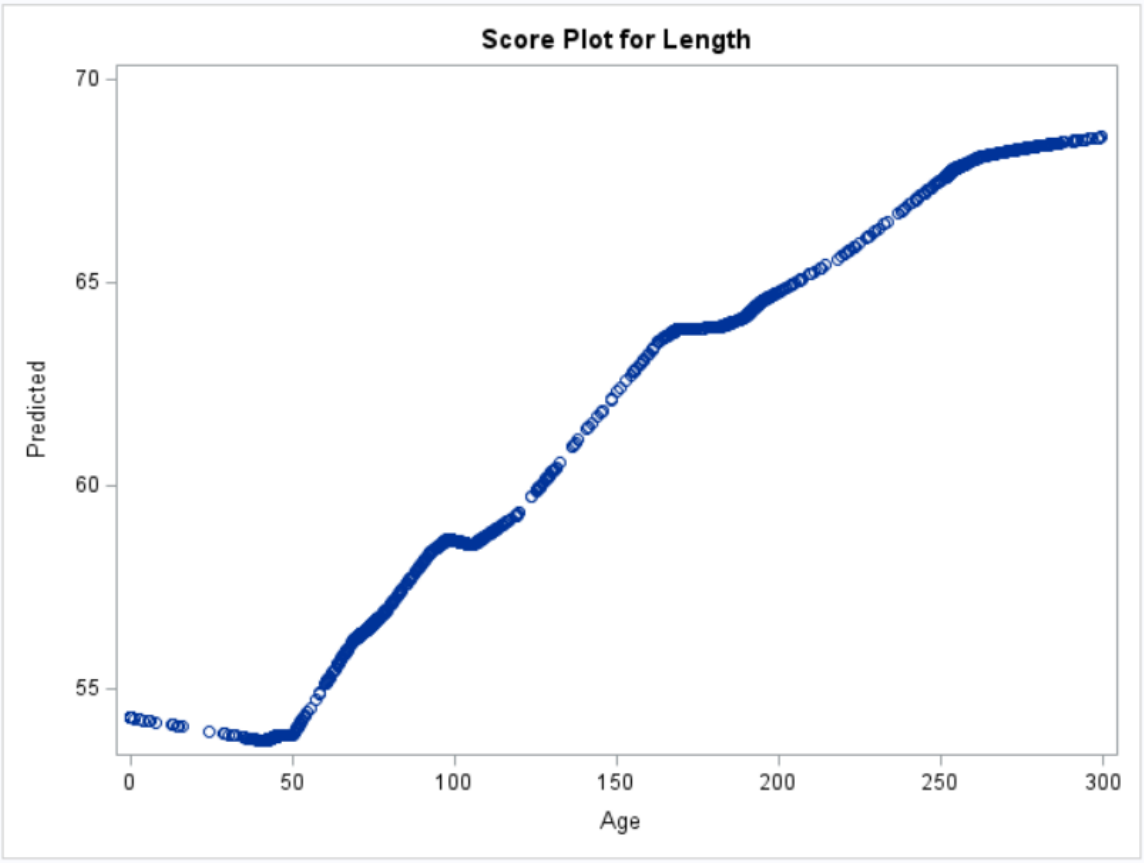
The LOESS Procedure
Selected Smoothing Parameter: 0.183
Dependent Variable: Length

Fit Summary	
Fit Method	kd Tree
Blending	Linear
Number of Observations	2133
Number of Fitting Points	33
kd Tree Bucket Size	77
Degree of Local Polynomials	1
Smoothing Parameter	0.18261
Points in Local Neighborhood	389
Residual Sum of Squares	19538
Trace[L]	11.89560
GCV	0.00434
AICC	3.22702


Fit Plot for Length
Smooth = 0.183
Length
70
60
50
40
0
50
100
150
200
250
300
Age


Residual Plot for Length
Residual
10
0
-10
-20
0
50
100
150
200
250
300
Age


Fit Diagnostics for Length
Residual
Predicted Value
Percent
Residual
Residual
Quantile
Fit–Mean
Residual
Proportion Less
Length
Predicted Value
Observations 2133
Smooth 0.1826
Local Points 389
Degree 1
Residual SS 19538
Fit Points 33
Interpolation Linear


Score Plot for Length
Predicted
Age

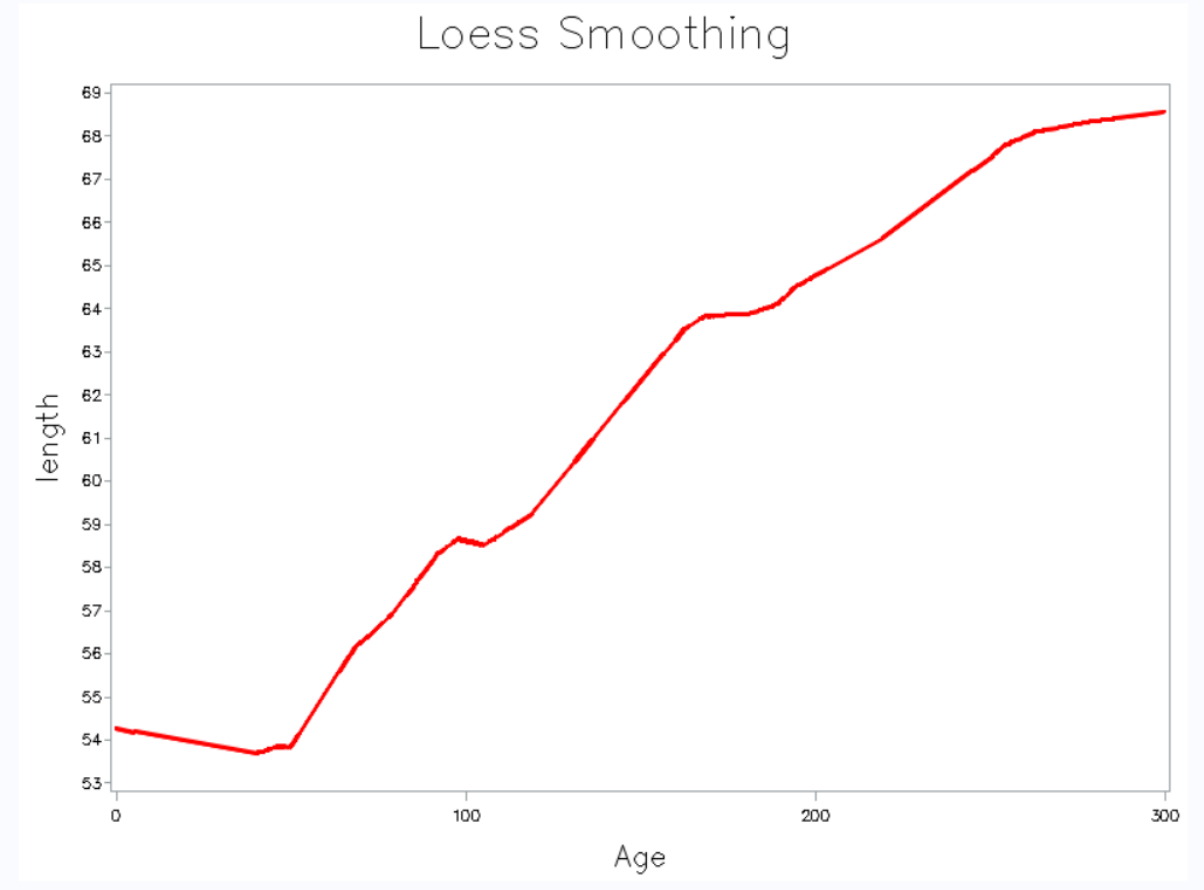


Figure 2.8

Interpretation:

In this figure, we can identify a positive correlation between both age and length and we can see a continuously increasing pattern over time even though there was evidence of multiple outliers when observing the residual plot.

We will now fit linear mixed models in order to identify which model is the best model to use with the given data set:

```
proc mixed data=len_exp noclprint;
class id;
model length=age premature age*premature symptomatic age*symptomatic
LowBWeight age*LowBWeight / s;
random intercept age/ type=un subject=id g v;
repeated / type=simple subject=id r rcorr;
title 'Kidney: Random Intercept + Slope, Simple';
run;
```

Solution for Fixed Effects

Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	52.3780	0.1383	493	378.72	<.0001
Age	0.06370	0.000764	419	83.37	<.0001
Premature	-2.5547	0.3883	934	-6.58	<.0001
Age*Premature	0.008648	0.002117	934	4.09	<.0001
Symptomatic	-0.1574	0.4709	934	-0.33	0.7382
Age*Symptomatic	0.004686	0.002651	934	1.77	0.0774
LowBWeight	-2.8088	0.4464	934	-6.29	<.0001
Age*LowBWeight	-0.00097	0.002510	934	-0.39	0.6998

Type 3 Tests of Fixed Effects

Effect	Num DF	Den DF	F Value	Pr > F
Age	1	419	6950.07	<.0001
Premature	1	934	43.29	<.0001
Age*Premature	1	934	16.69	<.0001
Symptomatic	1	934	0.11	0.7382
Age*Symptomatic	1	934	3.12	0.0774
LowBWeight	1	934	39.59	<.0001
Age*LowBWeight	1	934	0.15	0.6998

Figure 2.9

Comment: Since certain p-values are greater than 0.05 we need to remove them in order to achieve the best model.

```
proc mixed data=len_exp noclprint;
class id;
model length=age premature age*premature  LowBWeight  / s;
random intercept age/ type=un subject=id g v;
repeated / type=simple subject=id r rcorr;
title 'Kidney: Random Intercept + Slope, Simple';
run;
```

Estimated V Matrix for ID 3			
Row	Col1	Col2	Col3
1	6.8725	3.5170	3.4986
2	3.5170	7.9493	4.3228
3	3.4986	4.3228	7.6525

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
UN(1,1)	ID	3.4121
UN(2,1)	ID	-0.00093
UN(2,2)	ID	0.000042
Residual	ID	3.4663

Fit Statistics	
-2 Res Log Likelihood	8534.2
AIC (Smaller is Better)	8542.2
AICC (Smaller is Better)	8542.3
BIC (Smaller is Better)	8559.1

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	52.3589	0.1360	497	384.90	<.0001
Age	0.06402	0.000735	424	87.12	<.0001
Premature	-2.5739	0.3677	944	-7.00	<.0001
Age*Premature	0.008866	0.001826	944	4.85	<.0001
LowBWeight	-2.6828	0.2917	944	-9.20	<.0001

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Age	1	424	7589.67	<.0001
Premature	1	944	49.01	<.0001
Age*Premature	1	944	23.57	<.0001
LowBWeight	1	944	84.59	<.0001

Figure 2.10

Interpretation:

Although this model looks fairly good due to significant p-values, it is important to test other methods of modelling and then choose the best model from this. This can be done by looking at the Akaike information criterion (AIC) and the Bayesian information criterion (BIC) of each model.

Comparisons between different models were made and one model in particular had a smaller AIC and BIC than the model above. This was achieved by taking an AR(1).

```
proc mixed data=len_exp noclprint;
class id;
model length=age premature age*premature  LowBWeight  / s;
random intercept age/ type=un subject=id g v;
repeated / type=AR(1) subject=id r rcorr;
title 'Kidney: Random Intercept + Slope, Simple';
run;
```

Estimated R Matrix for ID 3

Row	Col1	Col2	Col3
1	4.1520	0.8295	0.1657
2	0.8295	4.1520	0.8295
3	0.1657	0.8295	4.1520

Estimated R Correlation Matrix for ID 3

Row	Col1	Col2	Col3
1	1.0000	0.1998	0.03991
2	0.1998	1.0000	0.1998
3	0.03991	0.1998	1.0000

Estimated G Matrix

Row	Effect	study_id	Col1	Col2
1	Intercept	3	2.3780	0.003543
2	Age	3	0.003543	6.197E-6

Estimated V Matrix for ID 3

Row	Col1	Col2	Col3
1	6.8271	4.0455	3.2920
2	4.0455	8.0349	4.6018
3	3.2920	4.6018	7.8172

Estimated V Matrix for ID 3			
Row	Col1	Col2	Col3
1	6.8271	4.0455	3.2920
2	4.0455	8.0349	4.6018
3	3.2920	4.6018	7.8172

Covariance Parameter Estimates		
Cov Parm	Subject	Estimate
UN(1,1)	ID	2.3780
UN(2,1)	ID	0.003543
UN(2,2)	ID	6.197E-6
AR(1)	ID	0.1998
Residual		4.1520

Fit Statistics	
-2 Res Log Likelihood	8519.8
AIC (Smaller is Better)	8529.8
AICC (Smaller is Better)	8529.8
BIC (Smaller is Better)	8550.9

Null Model Likelihood Ratio Test		
DF	Chi-Square	Pr > ChiSq
4	550.10	<.0001

Solution for Fixed Effects					
Effect	Estimate	Standard Error	DF	t Value	Pr > |t|
Intercept	52.3674	0.1372	497	381.62	<.0001
Age	0.06356	0.000731	424	86.91	<.0001
Premature	-2.5665	0.3710	944	-6.92	<.0001
Age*Premature	0.008628	0.001819	944	4.74	<.0001
LowBWeight	-2.6742	0.2925	944	-9.14	<.0001

Type 3 Tests of Fixed Effects				
Effect	Num DF	Den DF	F Value	Pr > F
Age	1	424	7552.85	<.0001
Premature	1	944	47.86	<.0001
Age*Premature	1	944	22.50	<.0001
LowBWeight	1	944	83.61	<.0001

Figure 2.11

Interpretation:

It is clear that the AIC and BIC are smaller in the model above than that of the previous model making this the best model of our choice.

Model in equation format:

β0 = intercept

β1Ai = intercept for age

β2Pj = intercept for premature

β3AiPj = common slope for age and premature

β4Lt = intercept for LowBWeight

$$Yijt = \beta 0 + \beta 1Ai + \beta 2Pj + \beta 3AiPj + \beta 4Lt + \varepsilon ijt$$

$$Yijt = 52.3674 + 0.06356Ai - 2.5665Pj + 0.008628AiPj + -2.6742Lt$$