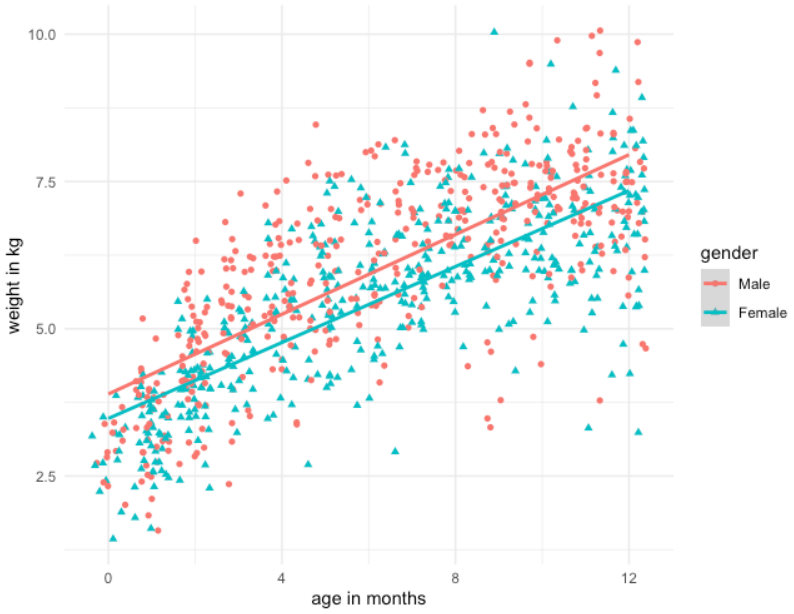
Dehshini Ali

Problem set 2

Scatter Plot of Weight Against Age

1.

2.



Ancova model of weight on age and gender.

Call:
lm(formula = weight ~ age + gender + age:gender, data = nepaldata)

Residuals:

Min	1Q	Median	3Q	Max
-4.1448	-0.6332	0.0229	0.6999	3.6221

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.89218	0.10027	38.817	< 2e-16 ***
age	0.33849	0.01382	24.488	< 2e-16 ***
genderFemale	-0.41504	0.14077 -	2.948	0.00328 **
age:genderFemale	-0.01619	0.01946	-0.832	0.40564

Residual standard error: 1.07 on 894 degrees of freedom
Multiple R-squared: 0.5746, Adjusted R-squared: 0.5732
F-statistic: 402.6 on 3 and 894 DF, p-value: < 2.2e-16

3.

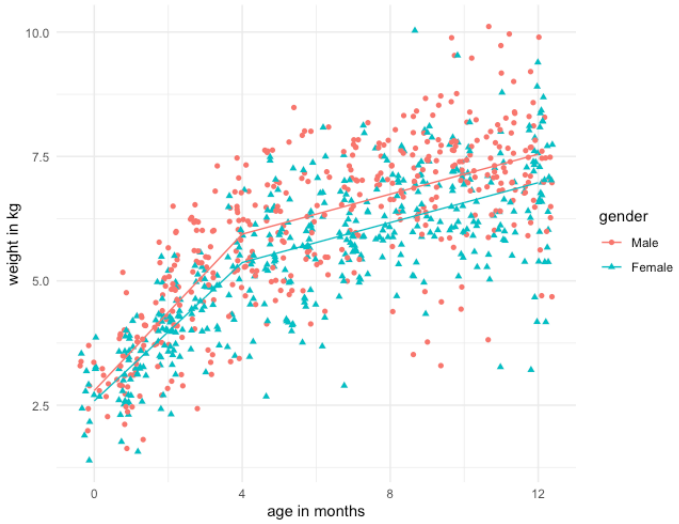
The intercept (B_0) is the weight in kg (3.89) when age is zero months for a male (newborn male)
The coefficient for age (B_1) is the average change in weight per 1-month increase in age for males.
The coefficient for genderFemale (B_2) is the difference in weight for females compared to males at age 0. Thus, females are 0.415kg less heavy at birth compared to males.
The coefficient for the interaction term (B_3) is the difference in slopes for females compared to males. Thus, there is a 0.0162 decrease in the change in weight for every 1 month of age for females compared to males.

The birthweight for males is 3.89kg while that for females is 3.89-0.415=3.475kg. This change in birthweight is statistically significant with a t-value of 2.948 and an associated p-value of 0.00328.
The change in weight with age for boys is 0.338(kg/month) while that for girls is 0.338-0.0162=0.322kg/month. This change in slopes of 0.0162 is not significantly different with p-value of 0.406.

4.

The assumption of linearity in the first year is not entirely correct for both genders. The plot of the residuals vs the predictor variable (age) does not look entirely randomly distributed around the mean of 0 for both genders. There are more values below zero at less than 4 months and greater than 8 months of age.

5.



lm(formula = weight ~ age * gender + age_splined * gender, data = nepaldata)

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.79295	0.15280	18.279	<2e-16 ***
age	0.78656	0.05101	15.420	<2e-16 ***
genderFemale	-0.20707	0.21160	-0.979	0.328
age_splined	-0.58579	0.06454	-9.077	<2e-16 ***
age:genderFemale	-0.09096	0.07102	-1.281	0.201
genderFemale:age_splined	0.09081	0.09041	1.004	0.315

The intercept(B_0) of 2.79295 kg is the birthweight of males, thus the weight for males when age is 0.

The coefficient "age" of 0.78656 kg/month is the change in weight per month of age (slope) for males when age is 4months or less.

The coefficient "genderFemale" of -0.20707 is the difference in weight between females and males when age is 0 months. Thus, birthweight for females is 0.20707kg less than for males.

The coefficient for "age_splined" of -0.58579 kg/month is the change in slope for males when age > 4months compared to when age<4months.

The coefficient "age:genderFemale" of -0.09096kg/month is the difference in slopes between females and males when age <= 4 months.

The coefficient "genderFemale:age_splined" of 0.09081 is the difference in slopes between females and males when age > 4 months.

6.

H_0 : coef(age_splined) = coef(genderFemale:age_splined) = 0
Analysis of Variance Table

Model 1: weight ~ age + gender + age:gender
Model 2: weight ~ age * gender + age_splined * gender

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	894	1022.66				
2	892	880.95	2	141.71	71.744	< 2.2e-16

The residual sums of squares (RSS) is lesser for model B vs model A. The F test of the model with the splined-age term and its interaction with gender vs the model with age and its interaction with gender is significant with p-value of 0.00. Reject the null and conclude that the spline term or its interaction with gender is not zero.

7.

The assumption of equal variance is still violated as there is more variability on the right side of the graph compared to the left.

8.

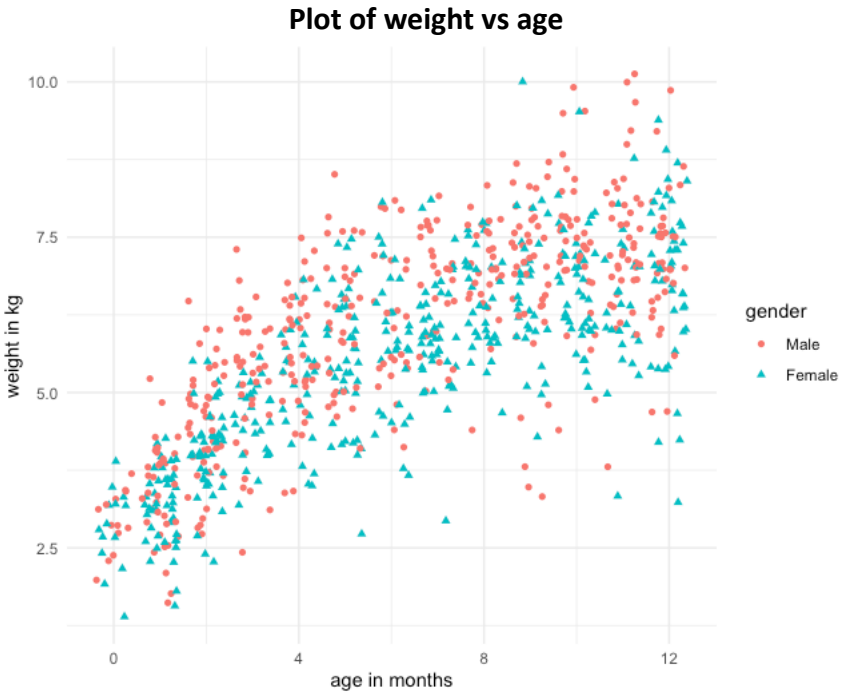
Write a paragraph to summarize your findings regarding growth of boys and girls as if for a public health journal. Be quantitative (use numbers) and avoid statistical jargon.

- The average birthweight for boys is 2.79kg while the average for girls is 2.59kg.
- There is an increase of 0.79kg in weight for every month of age for boys until age 4months when the increase in weight is 0.20kg (0.79-0.59) for every month of age till 12months.
- There is an increase of 0.70kg (0.79-0.09) in weight for every month of age for girls until age 4 months when the increase in weight is 0.20 kg [0.79+(-.0.9)+(-0.59)+0.09] for every month of age till 12 months.

  We conclude that the rate of growth (in weight) for boys is faster in the first 4 months of life compared to girls. After 4 months, the rate of growth is similar for boys and girls.

II.

2.

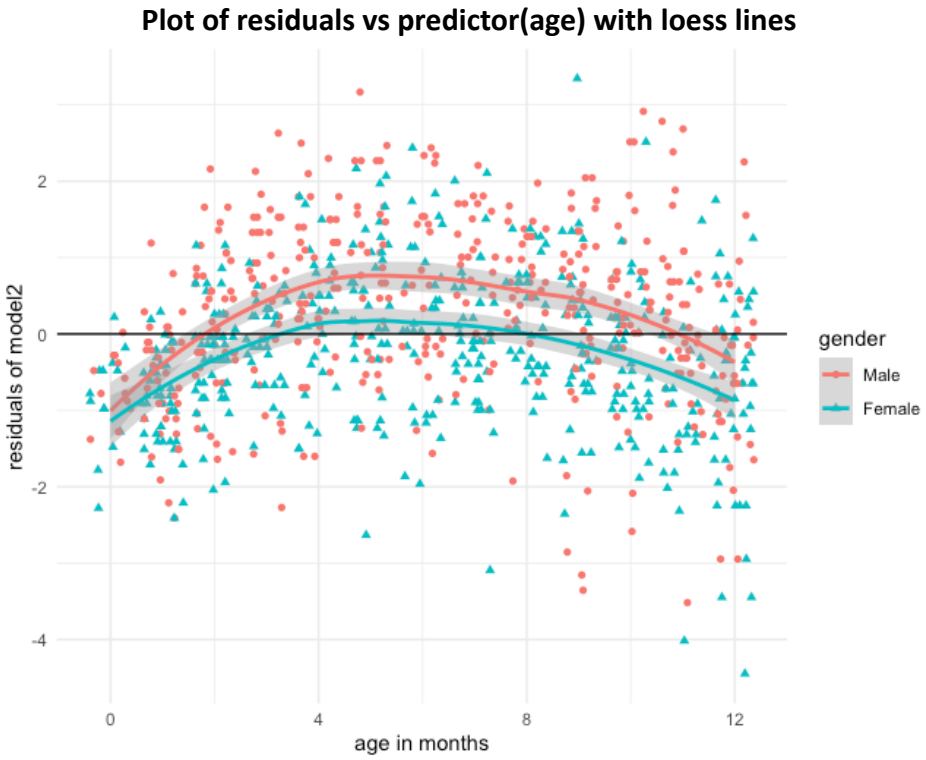

Plot of weight vs age
10.0
7.5
5.0
2.5
weight in kg
0
4
8
12
age in months
gender
Male
Female

3.


Plot of weight vs age with line through means at binned ages
10.0
7.5
5.0
2.5
weight in kg
0
4
8
12
age in months
gender
Male
Female

4.

5.



The plot of the residuals vs predictor variable shows the residuals aren't evenly distributed around the mean of 0, so the assumption of growth being linear in the first 12 months is not entirely correct.

6.

Call:
lm(formula = weight ~ as.factor(age), data = nepaldata)

Residuals:

Min	1Q	Median	3Q	Max
-3.7899	-0.6267	0.0279	0.6405	3.2405

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.8667	0.1965	14.589	< 2e-16 ***
as.factor(age)1	0.3928	0.2276	1.726	0.0847 .
as.factor(age)2	1.4007	0.2235	6.268	5.71e-10 ***
as.factor(age)3	2.0727	0.2333	8.886	< 2e-16 ***
as.factor(age)4	2.6054	0.2323	11.218	< 2e-16 ***
as.factor(age)5	2.9600	0.2292	12.917	< 2e-16 ***
as.factor(age)6	3.1075	0.2379	13.063	< 2e-16 ***
as.factor(age)7	3.3869	0.2309	14.671	< 2e-16 ***
as.factor(age)8	3.8397	0.2349	16.349	< 2e-16 ***
as.factor(age)9	3.8928	0.2296	16.957	< 2e-16 ***
as.factor(age)10	4.1599	0.2276	18.276	< 2e-16 ***
as.factor(age)11	4.1751	0.2327	17.939	< 2e-16 ***
as.factor(age)12	4.1232	0.2276	18.115	< 2e-16 ***

Residual standard error: 1.021 on 885 degrees of freedom
Multiple R-squared: 0.6163, Adjusted R-squared: 0.6111
F-statistic: 118.4 on 12 and 885 DF, p-value: < 2.2e-16

	2.5 %	97.5 %
(Intercept)	2.48101330	3.2523201
as.factor(age)1	-0.05389422	0.8395482
as.factor(age)2	0.96211642	1.8393328
as.factor(age)3	1.61493685	2.5305176
as.factor(age)4	2.14956045	3.0612238
as.factor(age)5	2.51025469	3.4097453
as.factor(age)6	2.64060483	3.5743377
as.factor(age)7	2.93376802	3.8399408
as.factor(age)8	3.37873854	4.3006265
as.factor(age)9	3.44224382	4.3433417
as.factor(age)10	3.71319438	4.6066368
as.factor(age)11	3.71832689	4.6319218
as.factor(age)12	3.67648551	4.5699279

7.

Call:
lm(formula = weight ~ age + age_sp1 + age_sp2 + age_sp3, data = nepaldata)
Residuals:
   Min   1Q Median   3Q   Max
-3.9965 -0.6318 0.0424 0.6657 3.3202

Coefficients:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.62533	0.15516	16.920	< 2e-16 ***
age	0.79912	0.10076	7.931	6.49e-15 ***
age_sp1	-0.15861	0.15232	-1.041	0.2981
age_sp2	-0.31121	0.13191	-2.359	0.0185 *
age_sp3	-0.15709	0.08571	-1.833	0.0672 .

Residual standard error: 1.025 on 893 degrees of freedom
Multiple R-squared: 0.6097, Adjusted R-squared: 0.608
F-statistic: 348.8 on 4 and 893 DF, p-value: < 2.2e-16

c.

d.

the coefficient “age_sp1”= -0.15861 is the difference in slopes (weight change per month) comparing age > 2 months to age <= 2 months.

The coefficient “age_sp2” = -0.31121 is the difference in slopes (weight change per month) comparing age >4months to age 2-4months.

The coefficient “age_sp3” = -0.15709 is the difference in slopes comparing age >6months to age 4-6months.

e.

Analysis of Variance Table

Model 1: weight ~ age
Model 2: weight ~ age + age_sp1 + age_sp2 + age_sp3
  Res.Df RSS Df Sum of Sq F Pr(>F)
1 896 1083.20
2 893 938.26 3 144.94 45.984 < 2.2e-16 ***

f.

H_0 : age_sp1 = age_sp2 = age_sp3 = 0
The anova test has an F of 45.98 with p-value of 0.00. the RSS for the simple linear model is 1083.20 while the RSS for the spline model is 938.26 which is less than the former.
Reject the null hypothesis and conclude that at least one of the spline terms is not 0.
The spline model is a better fit for the data compared to the simple linear model of weight on age.

8.

	Using regress			Using glm	
Model	Residual.SS	MSE	AIC	Log-likelihood	AIC
Linear (2)	1083.2	1.21	2720.797	-1358.399	2720.797
Monthly means (13)	922.61	1.042	2598.689	-1286.344	2598.689
Linear spline(5)	938.26	1.05	2597.798	-1293.899	2597.798

9.

Looking at the plots, the model with the monthly means seems to fit the data appropriately and shows the fastest growth in weight around the first 2 months and a flat slope in the 10-12 months.
The model with the monthly means has the lowest residual sums of squares and lowest residual mean square. However, the multiple splines model has the lowest (best) AIC value among the three even though not very different from that of the monthly means model.
It is known that children grow the fastest in weight in the first 4 months after birth.

The AIC value is a measure of how well the model fits the data we have. Compared to values from the other models, the lower the AIC, the better the fit. Based on this, the linear splines model has the best fit but is not too different from the monthly means model.
The MSE is a measure of the variability of the residuals (errors). The lower the value, the less the error and the better the prediction.

Script file.

```
#load the dataset as nepalA

#quick view the dataset
head(nepalA)

#filter the dataset and select age <= 12 and omit NAs of height, weight, armcir
nepaldata = nepalA %>%
  filter(age <= 12) %>%
  filter(!is.na(height), !is.na(weight), !is.na(armcirc))

#create gender variable, replace 1 and 2 by male and female.
nepaldata = nepaldata %>%
  mutate(gender = ifelse(sex==1, "Male", "Female"))

#scatter-plot of weight against age, color it by gender
#use ggplot
g = ggplot(nepaldata, aes(x=age, y=weight))
g = g + geom_point(aes(shape=gender, colour=gender), position = "jitter") +
  xlab("age in months") +
  ylab("weight in kg") +
  theme_minimal() +
  geom_smooth(method=lm, colour="black") +
  geom_smooth(method=lm, aes(x=age, y=modelA$fitted.values, colour=gender))

#create a multiple regression of weight on age and gender
model1 = lm(weight ~ age + gender, data=nepaldata)
summary(model1)
confint(model1)

#create an ancova model for weight on age and gender(male reference)
modelA = lm(weight ~ age + gender + age:gender, data=nepaldata)
summary(modelA)

#fit a line for the fitted values for model A(there will be 2, one for males and females)
#plot the residuals on y and predictor on x
#use gg
res_v_x = ggplot(data = nepaldata)
res_v_x = res_v_x +
  geom_point(position = "jitter",
             alpha=1,
             size=1.5,
             aes(x=age, y=modelA$residuals, colour=gender)) +
  xlab("age in months") +
  ylab("residuals") +
```

```
  theme_minimal()

#create a new variable for the linear spline with the age knot at 4months
nepaldata = nepaldata %>%
  mutate(age_splined = ifelse(age > 4, age-4, 0))

#create for new. model for the splined age variable.
modelB = lm(weight ~ age*gender + age_splined*gender, data=nepaldata)
summary(modelB)
confint(modelB)

#plot the fitted values from modelB against age
#use ggplot
gB = ggplot(nepaldata, aes(x=age, y=weight))
gB = gB + geom_point(aes(shape=gender, colour=gender), position = "jitter") +
  xlab("age in months") +
  ylab("weight in kg") +
  theme_minimal()
gB = gB + geom_line(aes(x = age, y=modelB$fitted.values, color=gender))

#anova test
anova(modelA, modelB)

#plot the residuals from modelB against age
gB_res = ggplot(data=nepaldata)
gB_res = gB_res + geom_point(aes(x=age, y=modelB$residuals, colour=gender,
shape=gender), position = "jitter") +
  xlab("age in months") +
  ylab("residuals") +
  theme_minimal()
gB_res = gB_res + geom_hline(yintercept=0, color="red")

########################
#part 2 of problem set#
########################
#use same dataset

#sample mean weight for each month of age
nepaldata %>% group_by(age) %>% summarise(mean=mean(weight))

#ggplot it. #add means to plot "g"
g + stat_summary(aes(x=age, y=weight), fun = mean, colour="red", linewidth=1.2,
geom="line")

#regress weight on age
model2 = lm(weight ~ age, data=nepaldata)
summary(model2)
```

```
#gglot it. # add regression line to plot "g"
g + geom_smooth(method=lm, colour="black")

#plot residuals vs age for model2
g2 = ggplot(data=nepaldata, aes(x=age, y=model2$residuals, colour=gender, shape=gender))
g2 + xlab("age in months") +
  ylab("residuals of model2") +
  geom_point(aes(), position="jitter") +
  geom_smooth(aes(x=age, y=model2$residuals), method=loess) +
  theme_minimal() +
  geom_hline(yintercept = 0, colour="black")

#regress weight on the different ages
model3 = lm(weight ~ as.factor(age), data=nepaldata)
summary(model3)

#lets do a multi-spline model.
#create new variables
nepaldata = nepaldata %>%
  mutate(age_sp1 = ifelse(age > 2, age-2, 0)) %>%
  mutate(age_sp2 = ifelse(age > 4, age-4, 0)) %>%
  mutate(age_sp3 = ifelse(age > 6, age-6, 0))

#regress weight on the new variables
model4 = lm(weight ~ age + age_sp1 + age_sp2 + age_sp3, data=nepaldata)
summary(model4)

#add the predicted values from model4 to the first graph "g"
g + geom_line(aes(x = age, y=model4$fitted.values), color="red", linewidth=1.4)

#compare this multi-splined model to the linear model
anova(model2, model4)

#do AIC comparison for the 3 models.
AIC(model2, model3, model4)

#use glm to create the 3 models
model2g = glm(weight~age, data=nepaldata)
model3g = glm(weight~as.factor(age), data=nepaldata)
model4g = glm(weight~age+age_sp1+age_sp2+age_sp3, data=nepaldata)

#get the log-likelihoods from the glm models
logLik(model2g)
logLik(model3g)
logLik(model4g)
```

```
#AIC
AIC(model2g)-2
AIC(model3g)-2
AIC(model4g)-2
```