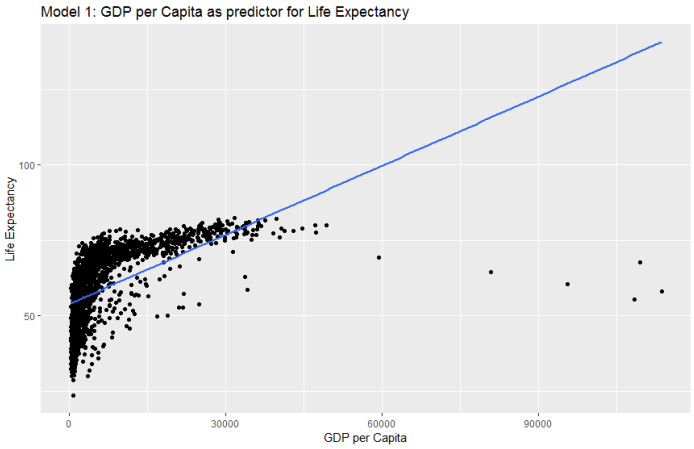
Coursework: Gapminder and Multilevel Modelling
Week 4

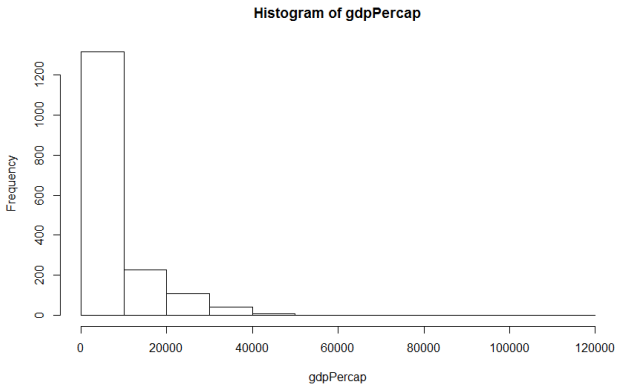
By Ida Johanne Austad

1. **Fit simple linear regression, with 'lifeExp' as dependent variable and 'gdpPercapita' as predictor. Visualise the fitted line and summarize the model. Briefly interpret the results.**

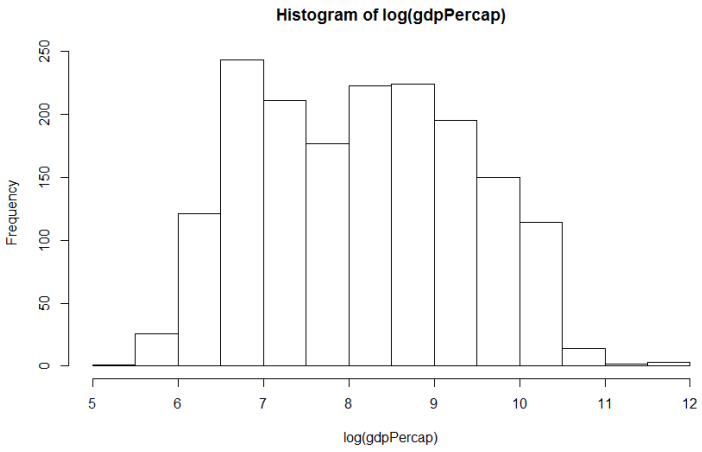
See attached code.

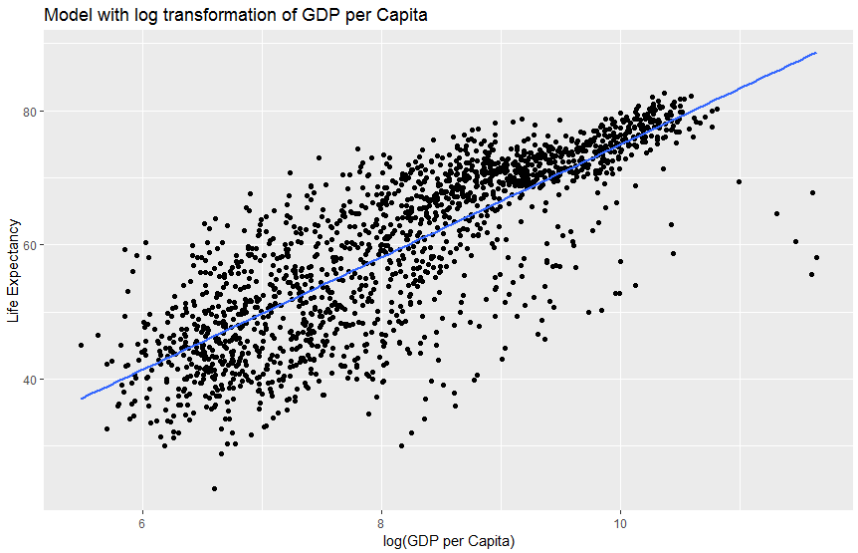
The first attempt to fit a model was done using `lm(lifeExp ~ gdpPercap)`. However, by investigating the model through the plot and the histogram included below, it was clear that the data does not follow a linear path and that GDP per Capita is not normally distributed.





Therefore, we can do a log transformation of the data on GDP per Capita which results in a distribution closer to normal, as shown in the histogram below, and a better fitted model – as shown in the plot on the next page.





Summary of the model using transformed data:

```
> summary(model3)

Call:
lm(formula = lifeExp ~ log(gdpPercap))

Residuals:
    Min      1Q  Median      3Q     Max
-32.778  -4.204   1.212   4.658  19.285

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     -9.1009     1.2277  -7.413 1.93e-13 ***
log(gdpPercap)   8.4051     0.1488  56.500  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 7.62 on 1702 degrees of freedom
Multiple R-squared:  0.6522,      Adjusted R-squared:  0.652
F-statistic:  3192 on 1 and 1702 DF,  p-value: < 2.2e-16
```

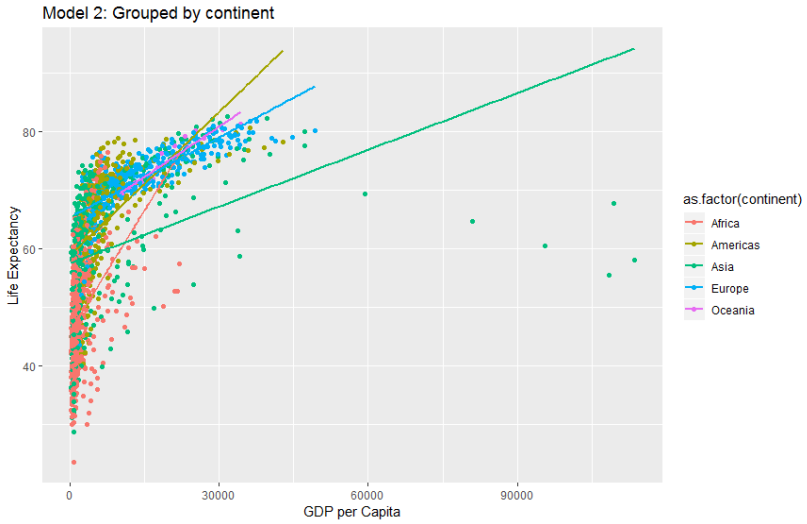
From the summary above we see that the p-values are small, which indicates that the coefficient estimates are significant. This is reflected in the significance scores which are both ***. Furthermore, we see that the estimated coefficient for `log(gdpPercap)` is positive. This means that as GDP per Capita increases in a country, so does the Life Expectancy. This positive relationship is reflected in the plot above.

For the rest of the coursework the transformed data on GDP per Capita will be used, except from when plots are presented and interpreted. This is to make visual interpretation “easier”; for instance when interpreting the scales.

2. **Create a plot where you check if the fitted line would vary if continent information were to be included in the model. Should we include the nested structure of the data into our model specification? Explain your answer.**

See attached code.

In the plot below we have included continent information in our model to see what it looks like if we fit one regression line to each continent.



From the plot above we see that the lines for the different continents have both different slopes and different intercepts. This indicates that the relationship between GDP per Capita and Life Expectancy varies across continents. Based on this we can argue that we should include continent information in our data as we go forward with our analysis.

3. **Fit a random intercept model with 'lifeExp' as dependent variable and 'gdpPercapita' as predictor:**
   - **write the model down**
   - **show summary of the model fit**
   - **give the estimated regression line**
   - **briefly interpret the results**

See attached code.

The model to be fitted is the following

$$Life\_exp_{ij} = \beta_0 + \beta_1(GDP\ per\ Cap_{ij}) + u_{0j} + \epsilon_{ij}$$

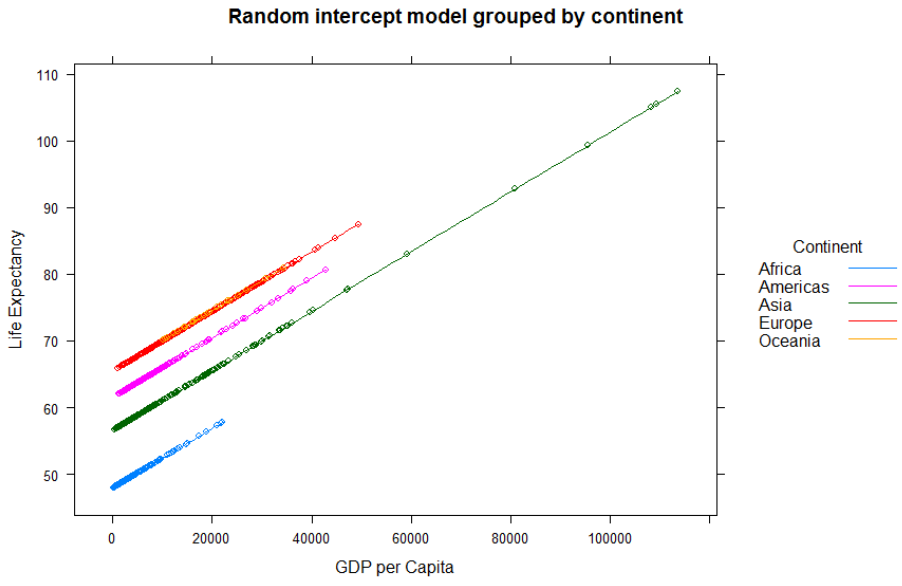
As noted in subtask 1, we use the log transform of the GDP per Capita data – and the summary is thus:

```
rand_int_model <- lmer(lifeExp ~ log(gdpPercap) + (1|continent), REML = FALSE)
> summary(rand_int_model)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: lifeExp ~ log(gdpPercap) + (1 | continent)

     AIC      BIC   logLik deviance df.resid
 11513.0  11534.8  -5752.5  11505.0     1700

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.8746 -0.4938  0.0626  0.6122  2.6343

Random effects:
 Groups    Name        Variance Std.Dev.
 continent (Intercept) 11.50    3.391
 Residual              49.51    7.036
Number of obs: 1704, groups:  continent, 5

Fixed effects:
               Estimate Std. Error t value
(Intercept)      8.2226     2.2002   3.737
log(gdpPercap)   6.4624     0.1824  35.430

Correlation of Fixed Effects:
            (Intr)
lg(gdpPrcp) -0.711
```

Using the estimated coefficients from the model output for fixed effects we get the equation for the average fitted line across continents:

$$Life\_exp_{ij} = 8.22 + 6.46(log(GDP\ per\ Cap_{ij}))$$

This will then vary from continent to continent from the average line's intercept with u_{0j} for each continent j.



From the plot we see that all the continents' line have the same slope (as is our assumption when doing a random intercept model), but that their intercepts vary. It looks as if the slope for Oceania and Europe are almost the same – while the others differ.

4. **Fit a random slope and intercept model with 'lifeExp' as dependent variable and 'gdpPercapita' as predictor:**
   - **write the model down**
   - **show summary of the model fit**
   - **give the estimated regression line**
   - **briefly interpret the results**

See attached code.

The model we want to fit:

$$Life\_exp_{ij} = \beta_0 + \beta_1(GDP\ per\ Cap_{ij}) + u_{0j} + u_{1j}(GDP\ per\ Cap_{ij}) + \epsilon_{ij}$$

As noted in subtask 1, we use the log transform of the GDP per Capita data – and the summary for our model is thus:

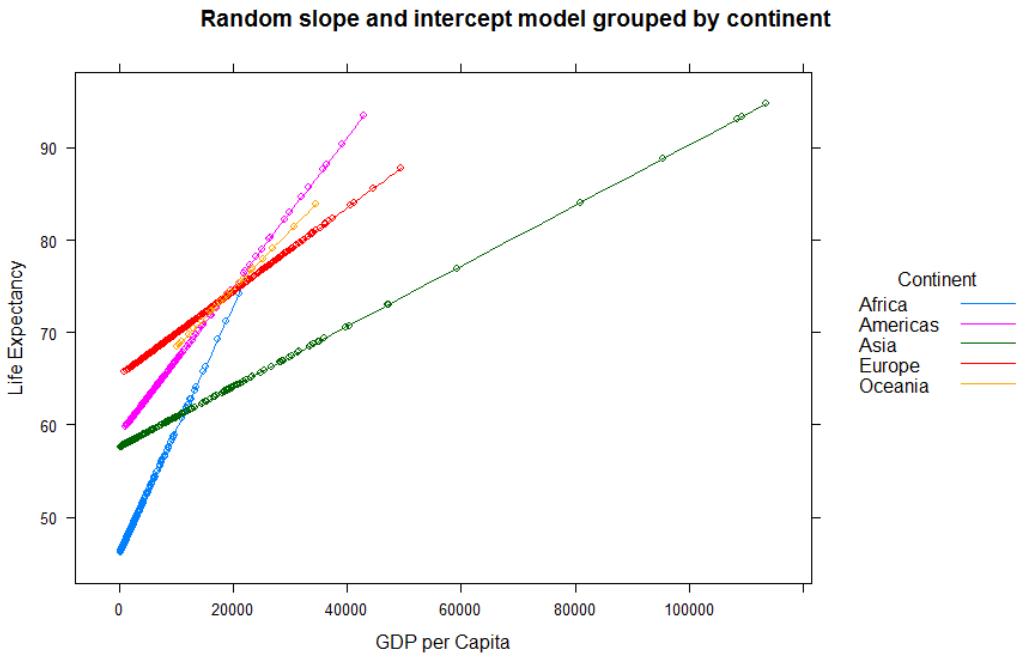
```
rand_slopint_model<-lmer(lifeExp ~ log(gdpPercap) + (1 + log(gdpPercap)|continent), REML = F
ALSE)
> summary(rand_slopint_model)
Linear mixed model fit by maximum likelihood  ['lmerMod']
Formula: lifeExp ~ log(gdpPercap) + (1 + log(gdpPercap) | continent)

     AIC      BIC   logLik deviance df.resid
 11495.3  11528.0  -5741.7  11483.3     1698

Scaled residuals:
    Min      1Q  Median      3Q     Max
-3.8220 -0.5322  0.0545  0.6168  2.6260

Random effects:
 Groups    Name           Variance Std.Dev. Corr
 continent (Intercept)    125.903  11.221
           log(gdpPercap)   1.974   1.405   -0.96
 Residual                  48.567   6.969
Number of obs: 1704, groups:  continent, 5

Fixed effects:
              Estimate Std. Error t value
(Intercept)      1.7693     5.7682   0.307
log(gdpPercap)   7.1552     0.6994  10.231

Correlation of Fixed Effects:
            (Intr)
lg(gdpPrcp) -0.962
```

Using the estimated coefficients from the model output for fixed effects we get the equation for the fitted regression line for continent j as:

$$Life\_exp_{ij} = 1.77 + \hat{u}_{0j} + (7.16 + \hat{u}_{1j})\log(GDP\ per\ Cap_{ij})$$

The effect of the log transform of GDP per capita for continent j is estimated to$(7.16 + \hat{u}_{1j})$, while the variance between the continents' slope is 1.97. From the summary output we can also see that the average increase in Life Expectancy across continents per unit increase in `log(gdpPercap)` is 7.16.

Below we can see the plot for the random intercept model, which shows that all the continents have a different relationship between Life Expectancy and GDP per Capita. Some continents have intercepts that are very close, such as the Americas and Asia, but we can see that the slope for the Americas is much steeper.



Attachment: R code

```
#coursework4

#installing and taking a look at the data
library(ggplot2)
library(dplyr)
library(lattice)
library(moderndive)
library(skimr)
library(lme4)
install.packages("gapminder")
library(gapminder)
attach(gapminder)
str(gapminder)                              #taking a look at the data

#1 Simple linear regression, lifeExp ~ gdpPercap
model1 <- lm(lifeExp ~ gdpPercap)
summary(model1)

#plotting model - see that the data does not follow a linear path
d <- gapminder %>%
  ggplot(aes(x = gdpPercap, y = lifeExp)) +
  geom_point() +
  geom_smooth(method = "lm", se= FALSE)
d + labs(x = "GDP per Capita", y = "Life Expectancy", title = "Model 1: GDP
per Capita as predictor for Life Expectancy")

#investigating the data - see that gdp per capita is not normally distribut
ed
hist(gdpPercap)                            #only this is included in report
par(mfrow=c(2,2))
plot(model1)

#do a logtrans of the function to see if it improves - looks better
hist(log(gdpPercap))                       #included in report
gapminder <- gapminder %>%
  mutate(log_gdp = log(gdpPercap))
model3 <- lm(formula = lifeExp ~ log(gdpPercap))
summary(model3)
par(mfrow=c(2,2))
plot(model3)                                          #to check the model wit
h transformed data further

#plotting model with log transformation
d <- gapminder %>%
  ggplot(aes(x = log_gdp, y = lifeExp)) +
  geom_point() +
  geom_smooth(method = "lm", se= FALSE)
d + labs(x = "log(GDP per Capita)", y = "Life Expectancy", title = "Model w
ith log transformation of GDP per Capita")


#2 Plotting to check if the fitted line will vary if continent is included
p <- gapminder %>%                              #separating by continent
  ggplot(aes(x = gdpPercap, y = lifeExp, colour = as.factor(continent))) +
  geom_point() +
  geom_smooth(method = "lm", se= FALSE)
p + labs(x = "GDP per Capita", y = "Life Expectancy", title = "Model 2: Gro
uped by continent")

#3 Fit a random intercept model with 'lifeExp' as dependent variable and 'l
og(gdpPercap)' as predictor
```

```
rand_int_model <- lmer(lifeExp ~ log(gdpPercap) + (1|continent), REML = FAL
SE)
summary(rand_int_model)

rand_int_model2 <- lmer(lifeExp ~ gdpPercap + (1|continent), REML = FALSE)
#without logtrans for figure
predlifeexp <-fitted(rand_int_model2)
xyplot(predlifeexp ~ gdpPercap , gapminder,
       groups=continent, main="Random intercept model grouped by continent"
,
       xlab="GDP per Capita",
       ylab="Life Expectancy",
       auto.key=list(space= 'right', title='Continent', cex.title=1,
                     lines=TRUE, points=FALSE),
       type = c("p", "smooth"))

#4 Fit a random slope and intercept model with 'lifeExp' as dependent varia
ble and 'log(gdpPercapita)' as predictor
rand_slopint_model<-lmer(lifeExp ~ log(gdpPercap) + (1 + log(gdpPercap)|con
tinent), REML = FALSE)
summary(rand_slopint_model)

rand_slopint_model2<-lmer(lifeExp ~ gdpPercap + (1 + gdpPercap|continent),
REML = FALSE) #without logtrans for figure
predlifeexp2 <-fitted(rand_slopint_model2)
xyplot(predlifeexp2 ~ gdpPercap , gapminder,
       groups=continent, main="Random slope and intercept model grouped by
continent",
       xlab="GDP per Capita",
       ylab="Life Expectancy",
       auto.key=list(space= 'right', title='Continent', cex.title=1,
                     lines=TRUE, points=FALSE),
       type = c("p", "smooth"))
```