STAT-S631

Assignment 2

John Koo

```
un.df <- alr4::UN11
```

```
dp <- loadNamespace('dplyr')
import::from(magrittr, `%>%`, `%<>%`)
library(ggplot2)
theme_set(theme_bw())
import::from(scales, log_trans, pretty_breaks)
import::from(xtable, xtable)
```

Problem 1

Assume that your data set is the entire population of interest, i.e., let $S = \{\text{set of all UN members}\}$ (you can assume all the data entries correspond to all UN members). Let female life expectancy, `lifeExpF`, be the response variable and `fertility` (rounded to the nearest integer) the predictor. Obtain the following results:

```
un.df %<>% dp$mutate(fertility.rounded = round(fertility))
```

Part a

Find the expected value and the variance of `lifExpF`

```
mean(un.df$lifeExpF)
```

```
[1] 72.29319
```

```
var(un.df$lifeExpF)
```

```
[1] 102.491
```

Part b

Find the expected value of `lifExpF` given that `fertility` $= i$ where $i = 1, ..., 7$.

```
un.df %>%
  dp$group_by(fertility.rounded) %>%
  dp$summarise(mean.female.life.exp = mean(lifeExpF)) %>%
  dp$ungroup() %>%
  xtable() %>%
  print(include.rownames = FALSE)
```

fertility.rounded	mean.female.life.exp
1.00	80.97
2.00	77.78
3.00	68.85
4.00	64.71
5.00	57.56
6.00	54.39
7.00	55.77

Part c

Find the variance of `lifExpF` given that `fertility` $= i$ where $i = 1, ..., 7$.

```
un.df %>%
  dp$group_by(fertility.rounded) %>%
  dp$summarise(var.female.life.exp = var(lifeExpF)) %>%
  dp$ungroup() %>%
  xtable() %>%
  print(include.rownames = FALSE)
```

fertility.rounded	var.female.life.exp
1.00	13.15
2.00	22.69
3.00	86.27
4.00	55.31
5.00	38.81
6.00	19.76
7.00	

$var(\text{female life expectancy}|\text{fertility} = 7)$ could not be computed because only one country has a fertility rate of 7, and variance is undefined when $n = 1$.

Problem 2

United Nations (Data file: `UN11`) The data in the file `UN11` contains several variables, including `ppgdp`, the gross national product per person in USD, and `fertility`, the birth rate per 1000 females, both from the year 2009. The data are for 199 localities, mostly UN member countries, but also other areas such as Hong Kong, that are not independent countries. The data were collected from United Nations (2011). We will study the dependence of `fertility` on `ppgdp`.

Part 1

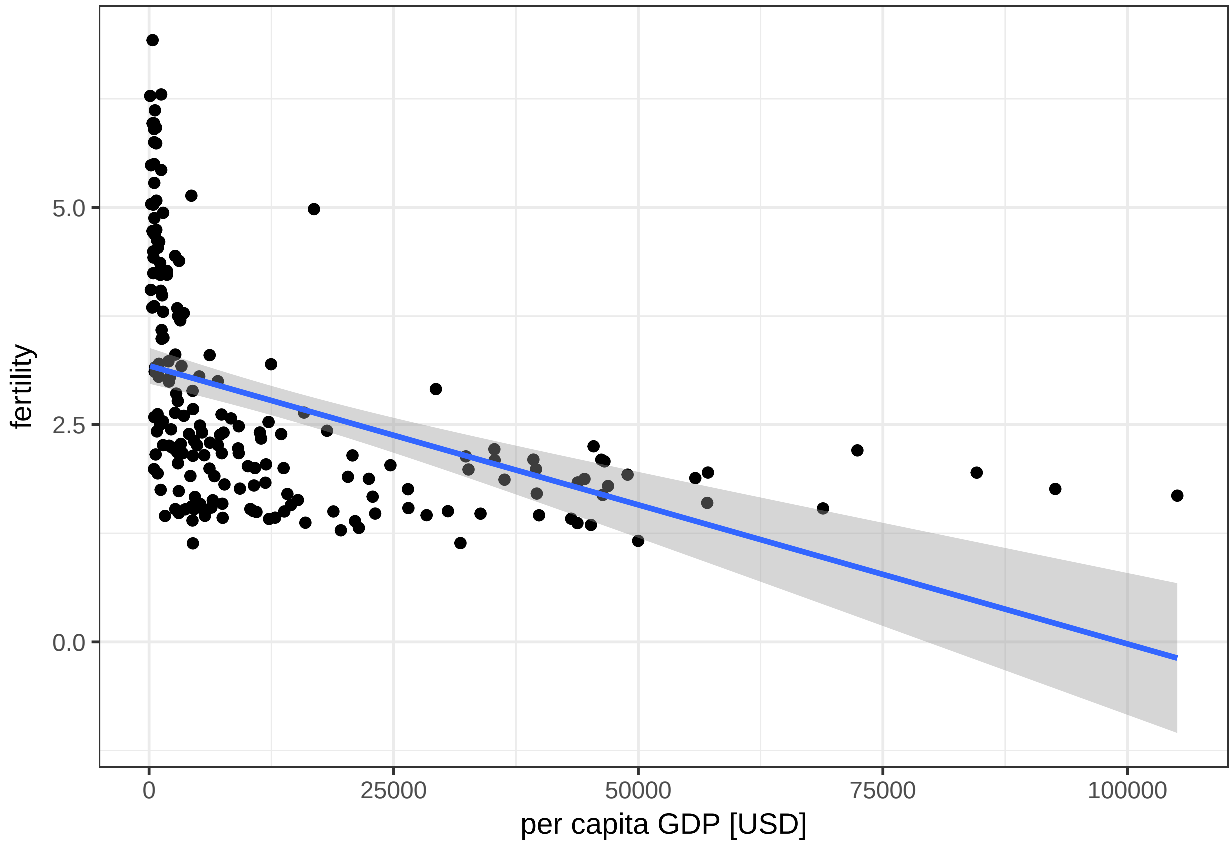
Identify the predictor and the response.

Predictor: GDP per capita
Response: Fertility rate

Part 2

Draw the scatterplot of `fertility` on the vertical axis versus `ppgdp` on the horizontal axis and summarize the information in this graph. Does a straight-line mean function seem to be plausible for a summary of this graph?

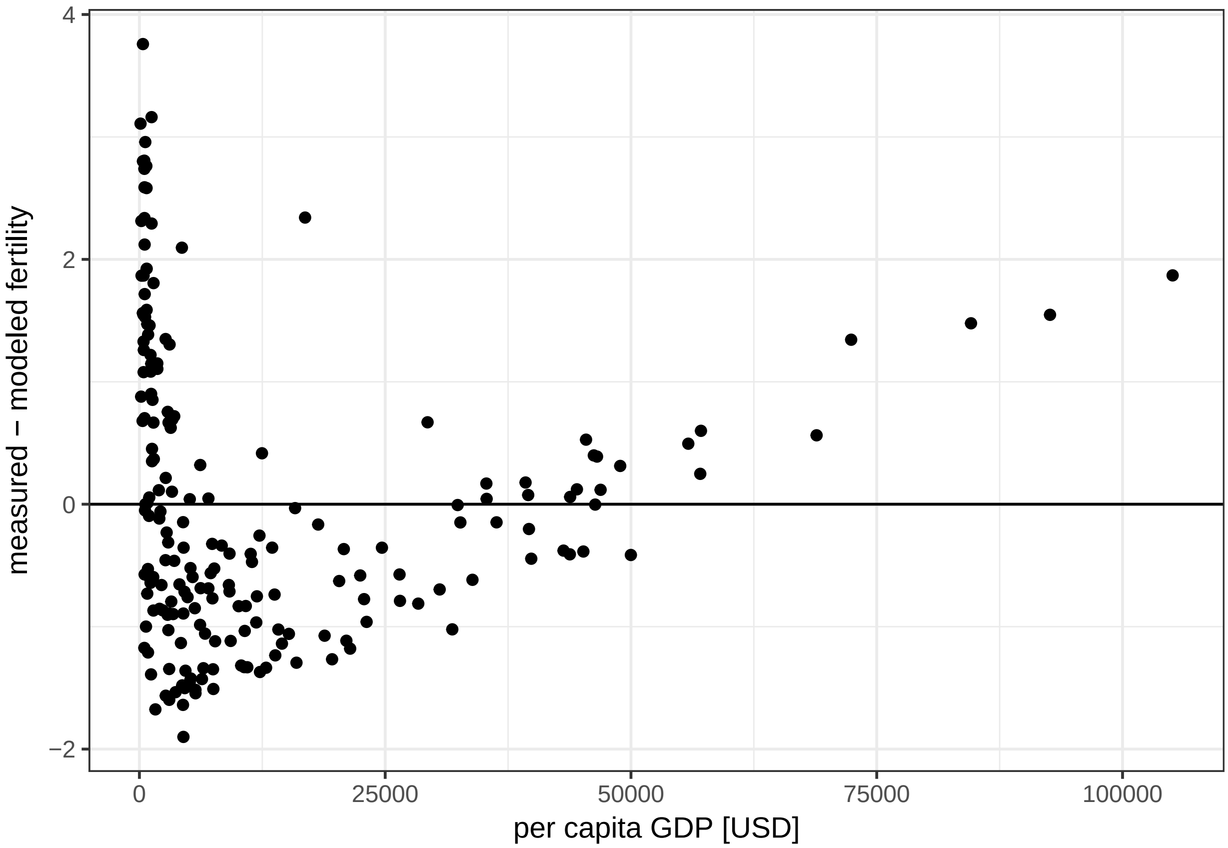
```
ggplot(un.df) +
  geom_point(aes(x = ppgdp, y = fertility)) +
  labs(x = 'per capita GDP [USD]') +
  stat_smooth(aes(x = ppgdp, y = fertility),
              method = 'lm')
```



From the scatterplot, it appears that a straight line would be a poor model for these data. The rate at which `fertility` changes per `ppgdp` is not constant.

To test this, we can try constructing an OLS model and checking the residuals. If the residuals appear to depend on the predictor, then a linear model would not describe the data very well.

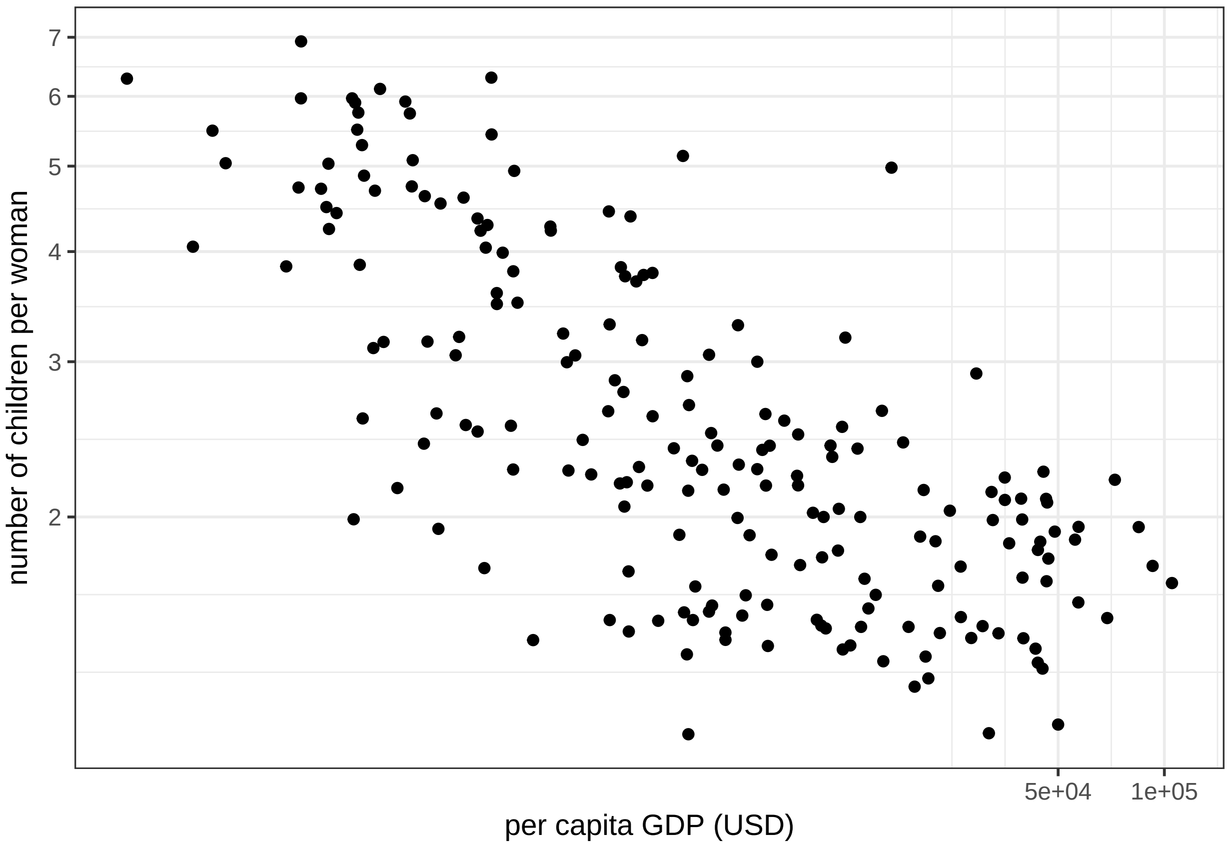
```
linear.mod.q2p2 <- lm(fertility ~ ppgdp, data = un.df)
un.df %>%
  dp$mutate(fertility.hat = predict(linear.mod.q2p2,
                                    newdata = .)) %>%
  dp$mutate(resid = fertility - fertility.hat) %>%
  ggplot() +
  geom_point(aes(x = ppgdp, y = resid)) +
  geom_hline(yintercept = 0) +
  labs(x = 'per capita GDP [USD]',
       y = 'measured - modeled fertility')
```



Part 3

Draw the scatterplot of log(`fertility`) versus log(`ppgdp`) using natural logarithms. Does the simple linear regression model seem plausible for a summary of this graph? If you use a different base of logarithms, the *shape* of the graph won't change, but the *values on the axes* will change.

```
ggplot(un.df) +
  geom_point(aes(x = ppgdp, y = fertility)) +
  labs(x = 'per capita GDP (USD)',
       y = 'number of children per woman') +
  scale_x_continuous(trans = log_trans(), breaks = pretty_breaks()) +
  scale_y_continuous(trans = log_trans(), breaks = pretty_breaks())
```



From the log-log plot, the rate at which log(`fertility`) changes per log(`ppgdp`) appears to be more constant than the rate at which `fertility` changes per `ppgdp`. So a linear regression appears to be more appropriate after taking the log transformations of each.

Problem 3

Smallmouth bass data (Data file: `wblake`) Compute the means and the variances for each of the eight subpopulations in the smallmouth bass data. Draw a graph of average length versus `Age` and compare with Figure 1.5. Draw a graph of the standard deviations versus age. If the variance function is constant, then a plot of standard deviation versus `Age` should be a null plot. Summarize the information.

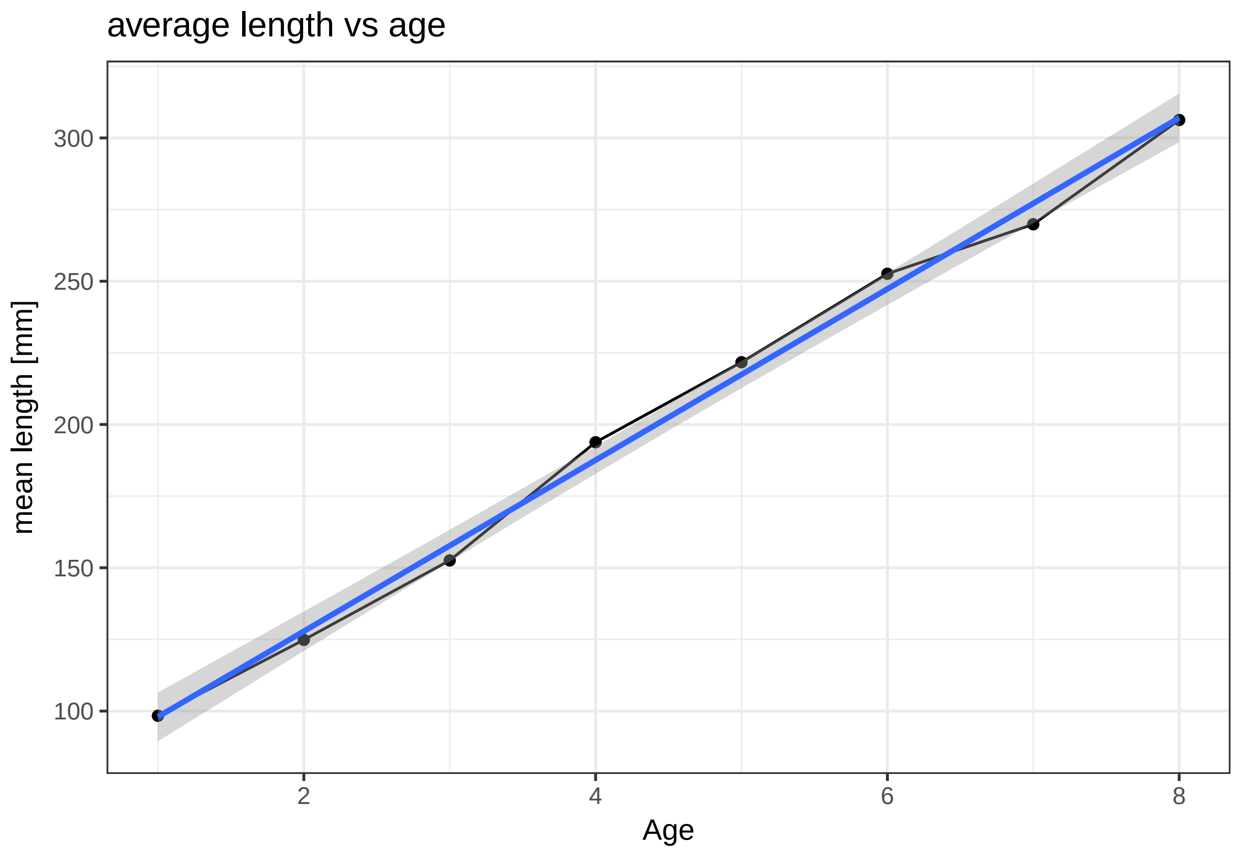
```
wblake.df <- alr4::wblake
```

Means and variances for each age:

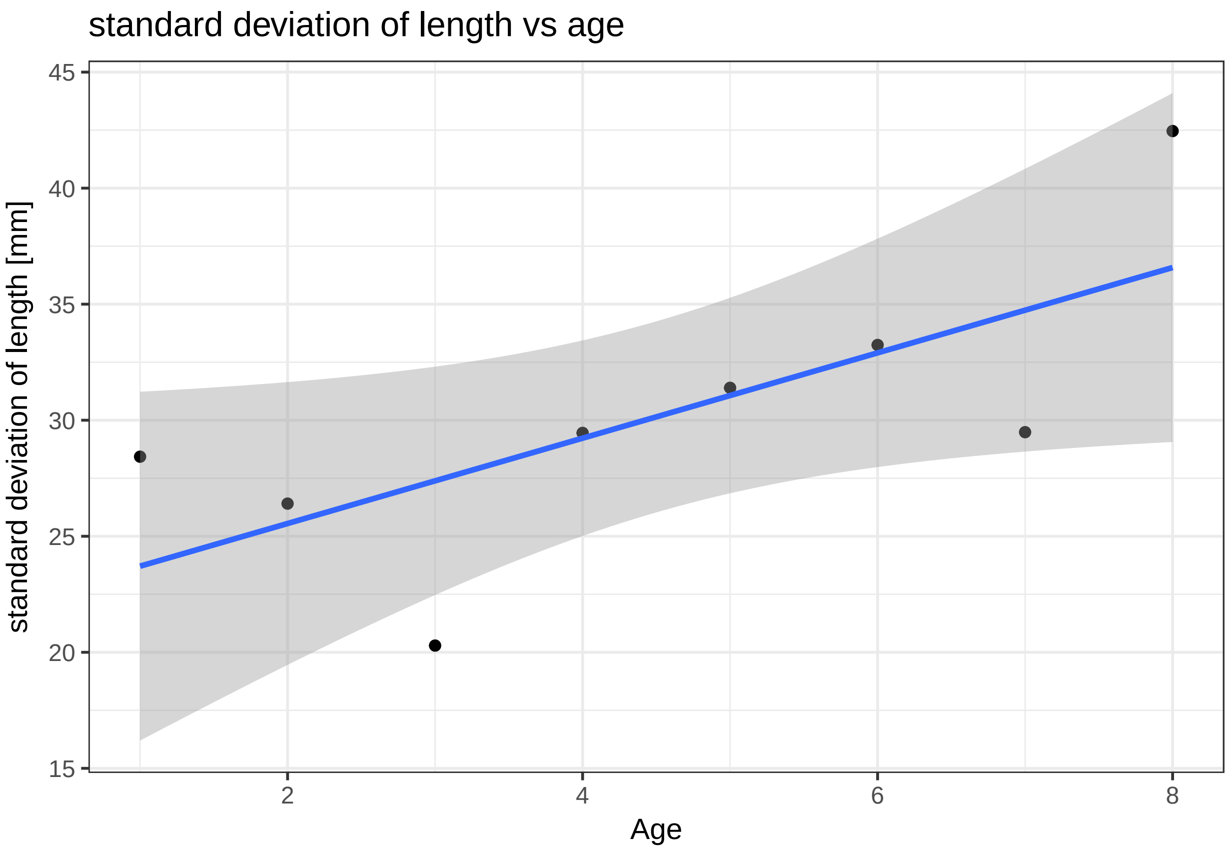
```
bass.summary.df <- wblake.df %>%
  dp$group_by(Age) %>%
  dp$summarise_all(c('mean', 'sd')) %>%
  dp$ungroup()

bass.summary.df %>%
  xtable() %>%
  print(include.rownames = FALSE)
```

Age	Length_mean	Scale_mean	Length_sd	Scale_sd
1	98.34	2.39	28.43	0.85
2	124.85	3.13	26.41	0.86
3	152.56	4.08	20.29	0.85
4	193.80	6.21	29.45	1.53
5	221.72	8.11	31.40	1.55
6	252.60	7.70	33.24	1.55
7	269.87	8.52	29.49	1.74
8	306.25	10.20	42.46	1.16

```r
ggplot(bass.summary.df) +
  geom_point(aes(x = Age, y = Length_mean)) +
  labs(title = 'average length vs age',
       y = 'mean length [mm]') +
  geom_line(aes(x = Age, y = Length_mean)) +
  stat_smooth(aes(x = Age, y = Length_mean), method = 'lm')
```




```r
ggplot(bass.summary.df) +
  geom_point(aes(x = Age, y = Length_sd)) +
  labs(title = 'standard deviation of length vs age',
       y = 'standard deviation of length [mm]') +
  stat_smooth(aes(x = Age, y = Length_sd), method = 'lm')
```



It looks like there *might* be a linear trend, but we can check for sure using OLS:

```
linear.mod.q3 <- lm(Length_sd ~ Age, data = bass.summary.df)
summary(linear.mod.q3)
```

```
Call:
lm(formula = Length_sd ~ Age, data = bass.summary.df)

Residuals:
    Min      1Q  Median      3Q     Max 
-7.0982 -1.1441  0.3357  1.8220  5.8814 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)   
(Intercept)  21.8729     3.7079   5.899  0.00105 **
Age           1.8383     0.7343   2.504  0.04631 * 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.759 on 6 degrees of freedom
Multiple R-squared:  0.5109,	Adjusted R-squared:  0.4294 
F-statistic: 6.268 on 1 and 6 DF,  p-value: 0.04631
```

The t-test on $\hat{\beta}_1$ suggests that maybe there is a significant lienar trend. This depends on the α threshold chosen at the start as well as the fact that there aren't very many data points here to create a linear model.