S631 HW2

Erik Parker

September 2, 2017

```r
library("alr4")

UN <- UN11

UN$fertility <- round(UN$fertility, 0)

fertility <- round(UN$fertility, 0)

lifeExpF <- UN$lifeExpF
```

1. Assume the data is the entire population of interest, $S = \{\text{set of all UN members}\}$. Let female life expectancy, $lifeExpF$, be the response variable and $fertility$ (rounded to the nearest integer) the predictor. Obtain the following results:

a) Find the expected value and the variance of $lifeExpF$

```r
mean(lifeExpF)
```

```
## [1] 72.29319
```

```r
var(lifeExpF)
```

```
## [1] 102.491
```

b) Find the expected value of $lifeExpF$ given that $fertility = i$ where $i = 1, ..., 7$.

```r
mean(UN[fertility == 1, 5])
```

```
## [1] 80.96565
```

```r
mean(UN[fertility == 2, 5])
```

```
## [1] 77.77853
```

```r
mean(UN[fertility == 3, 5])
```

```
## [1] 68.85352
```

```r
mean(UN[fertility == 4, 5])
```

```
## [1] 64.70913
```

```r
mean(UN[fertility == 5, 5])
```

```
## [1] 57.55556
```

```
mean(UN[fertility == 6, 5])
```

```
## [1] 54.38778
```

```
mean(UN[fertility == 7, 5])
```

```
## [1] 55.77
```

So, using the $mean()$ command, to find the expected value of $LifeExpF$ when $fertility$ is equal to 1-7, we find that in general, female life expectancy seems to decrease with increasing fertility.

c) Find the variance of $LifeExpF$ given that $fertility = i$ where $i = 1, ..., 7$.

```
var(UN[fertility == 1, 5])
```

```
## [1] 13.15358
```

```
var(UN[fertility == 2, 5])
```

```
## [1] 22.69346
```

```
var(UN[fertility == 3, 5])
```

```
## [1] 86.26717
```

```
var(UN[fertility == 4, 5])
```

```
## [1] 55.31225
```

```
var(UN[fertility == 5, 5])
```

```
## [1] 38.8089
```

```
var(UN[fertility == 6, 5])
```

```
## [1] 19.76342
```

```
var(UN[fertility == 7, 5])
```

```
## [1] NA
```

Using the $var()$ command to find the conditional variance of $LifeExpF$ for the countries with $fertility$ equal to 1-7, we can see that there is no linear pattern to this statistic, as opposed to what was seen from the expected value.

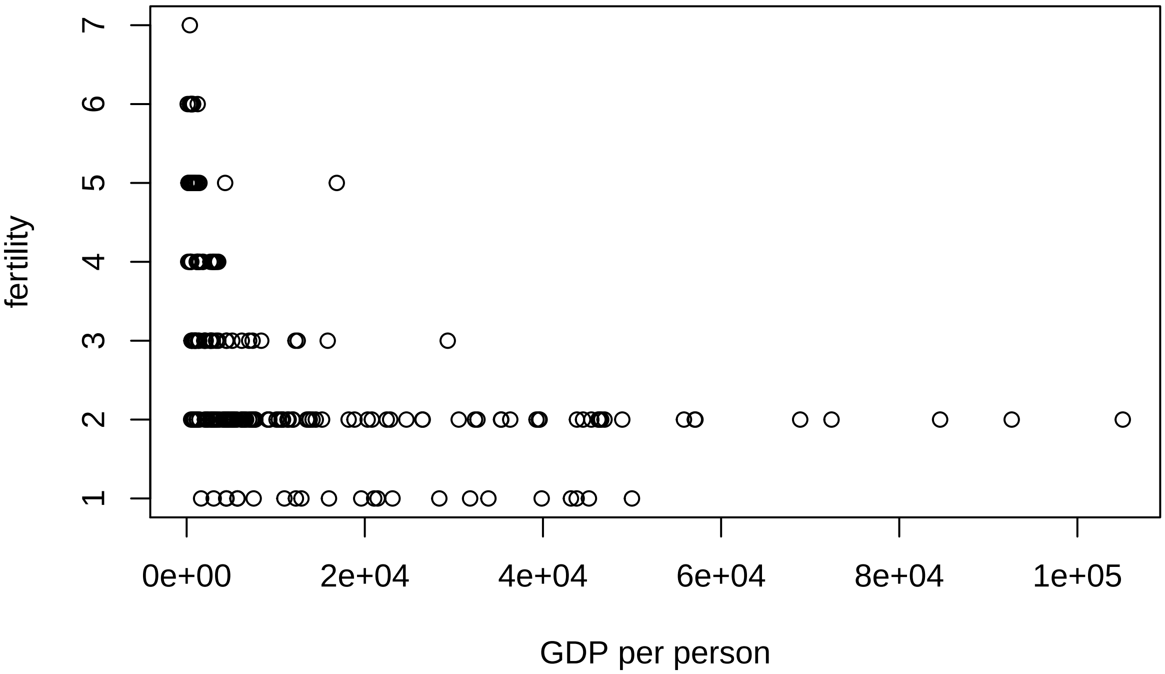
2. Here we will study the dependence of $fertility$ on *ppgdp*.

a) Identify the predictor and the response.

In this problem, we are studying the dependence of $fertility$ on *ppgdp*, meaning that *ppgdp* is the predictor and $fertility$ is the response variable.

b) Draw the scatterplot of $fertility$ on the y axis versus *ppgdp* on the x axis. Does a straight-line mean function seem to be plausible for a summary of this graph?

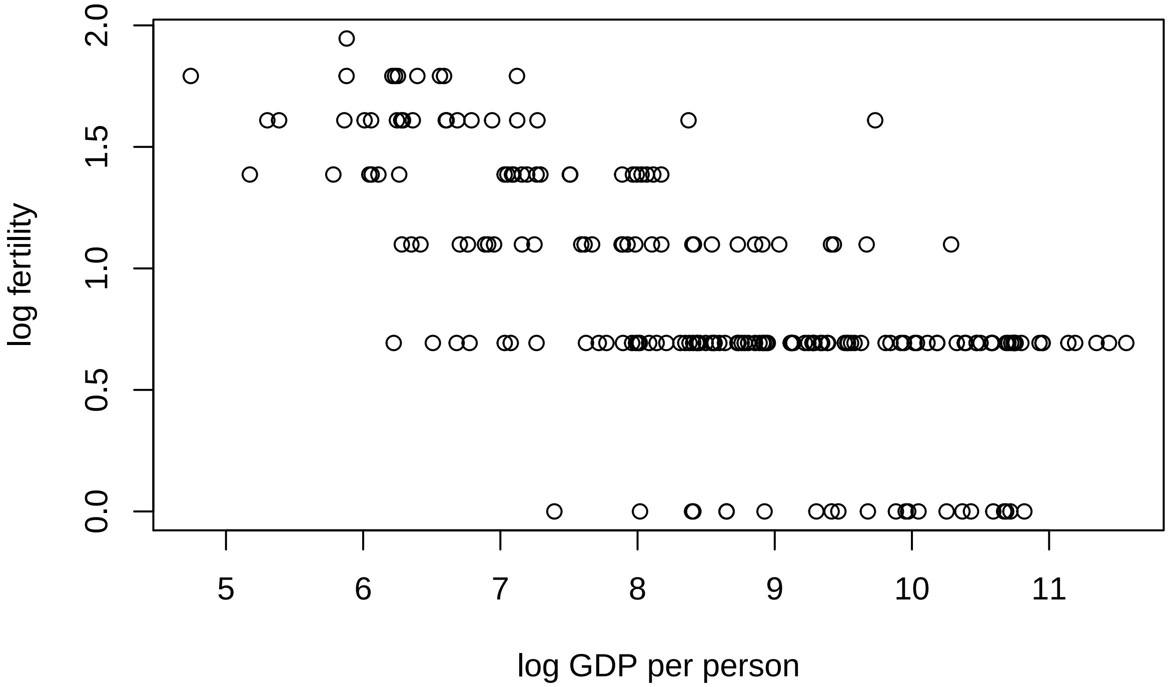
```
plot(UN$ppgdp, UN$fertility, xlab = "GDP per person", ylab = "fertility")
```



The large variance in values of *ppgdp* (from 0 to 100,000) makes this plot quite hard to accurately interpret. But that said, it does seem as though there is an overall decrease in fertility as gross national product per person decreases, though we can not say which of these variables, if either, is causal to this relationship.
A straight-line mean function seems to not be the best summary of the data shown in this graph, but could be plausible.

C) Draw a ln transformed scatterplot of the two variables in the previous plot. Does the simple linear regression model seem plausible as a summary?

```
plot(log(UN$ppgdp), log(UN$fertility), xlab = "log GDP per person", ylab = "log fertility")
```



When we log transform *fertility* and *ppgdp*, it becomes much more clear that a simple, decreasing, linear regression between the two variables would serve as a plausible summary for the relationship between these data.

3. Using the data file *wblake*, compute the means and variances for each of the eight subpopulations. Draw a graph of the average length versus *Age* and compare with Figure 1.5. Draw a graph of the standard deviations versus age. Summarize the information.

a) Means and variances of each of eight subpopulations.

```
bass <- wblake

age1 <- bass[bass$Age == 1, ]
age2 <- bass[bass$Age == 2, ]
age3 <- bass[bass$Age == 3, ]
age4 <- bass[bass$Age == 4, ]
age5 <- bass[bass$Age == 5, ]
age6 <- bass[bass$Age == 6, ]
age7 <- bass[bass$Age == 7, ]
age8 <- bass[bass$Age == 8, ]


mean(age1$Length)
```

```
## [1] 98.34211
```

```
mean(age1$Scale)
```

```
## [1] 2.386498
```

```r
var(age1$Length)
```

```
## [1] 808.2312
```

```r
var(age1$Scale)
```

```
## [1] 0.7164305
```

```r
mean(age2$Length)
```

```
## [1] 124.8472
```

```r
mean(age2$Scale)
```

```
## [1] 3.132991
```

```r
var(age2$Length)
```

```
## [1] 697.2862
```

```r
var(age2$Scale)
```

```
## [1] 0.7368941
```

```r
mean(age3$Length)
```

```
## [1] 152.5638
```

```r
mean(age3$Scale)
```

```
## [1] 4.078154
```

```r
var(age3$Length)
```

```
## [1] 411.6679
```

```r
var(age3$Scale)
```

```
## [1] 0.7190008
```

```r
mean(age4$Length)
```

```
## [1] 193.8
```

```r
mean(age4$Scale)
```

```
## [1] 6.209907
```

```r
var(age4$Length)
```

```
## [1] 867.4571
```

```r
var(age4$Scale)
```

```
## [1] 2.328682
```

```r
mean(age5$Length)
```

```
## [1] 221.7206
```

```r
mean(age5$Scale)
```

```
## [1] 8.105592
```

```r
var(age5$Length)
```

```
## [1] 985.6969
```

```r
var(age5$Scale)
```

```
## [1] 2.404859
```

```r
mean(age6$Length)
```

```
## [1] 252.5977
```

```r
mean(age6$Scale)
```

```
## [1] 7.700918
```

```r
var(age6$Length)
```

```
## [1] 1105.08
```

```r
var(age6$Scale)
```

```
## [1] 2.404934
```

```r
mean(age7$Length)
```

```
## [1] 269.8689
```

```r
mean(age7$Scale)
```

```
## [1] 8.517935
```

```r
var(age7$Length)
```

```
## [1] 869.3825
```

```r
var(age7$Scale)
```

```
## [1] 3.012746
```

```r
mean(age8$Length)
```

```
## [1] 306.25
```

```r
mean(age8$Scale)
```

```
## [1] 10.19852
```

```r
var(age8$Length)
```

```
## [1] 1802.917
```

```r
var(age8$Scale)
```

```
## [1] 1.348811
```

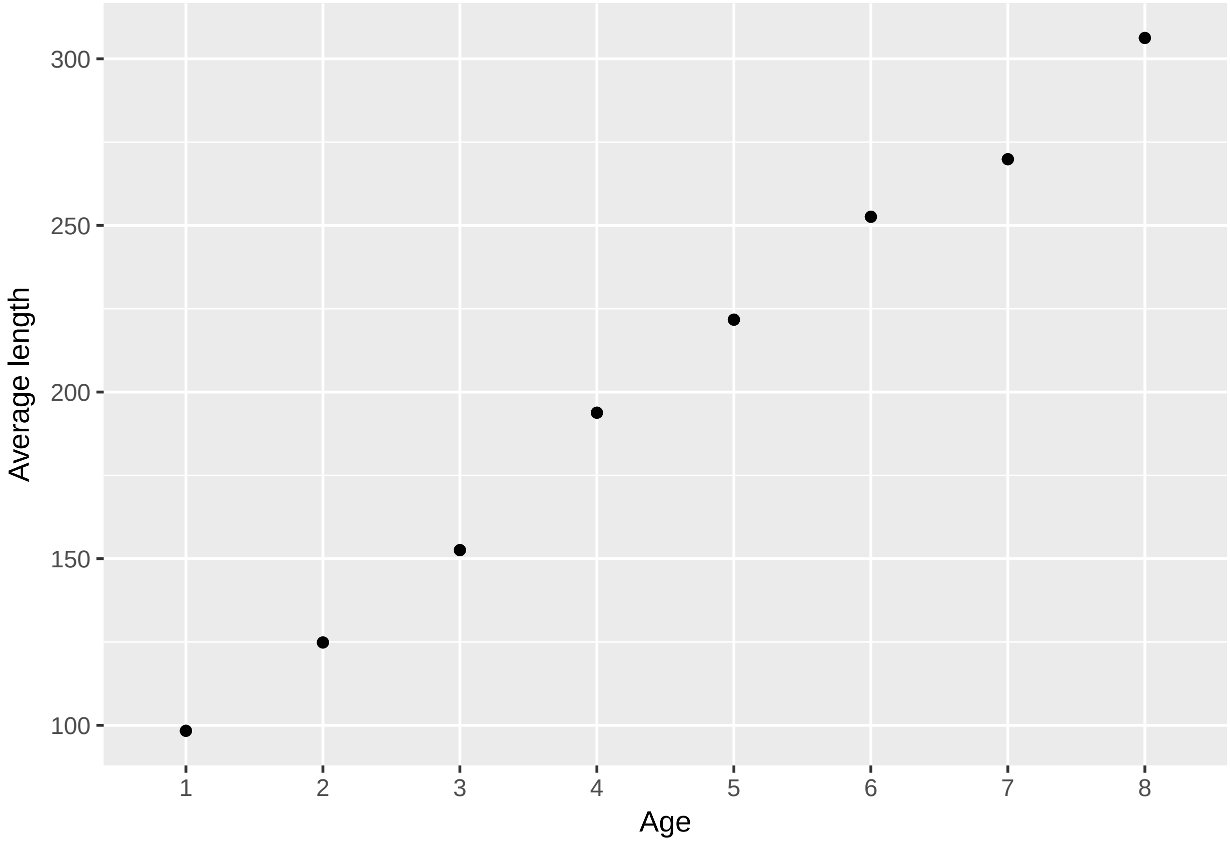
Graph of average length versus *Age* and compare with Figure 1.5.

```r
library(ggplot2)

mean.lengths <- c(mean(age1$Length), mean(age2$Length), mean(age3$Length), mean(age4$Length),
    mean(age5$Length), mean(age6$Length), mean(age7$Length), mean(age8$Length))
```

```
mean.lengths <- as.data.frame(mean.lengths)
mean.lengths$Age <- c("1", "2", "3", "4", "5", "6", "7", "8")


ggplot(mean.lengths, aes(x = Age, y = mean.lengths)) + geom_point() + labs(y = "Average length")
```

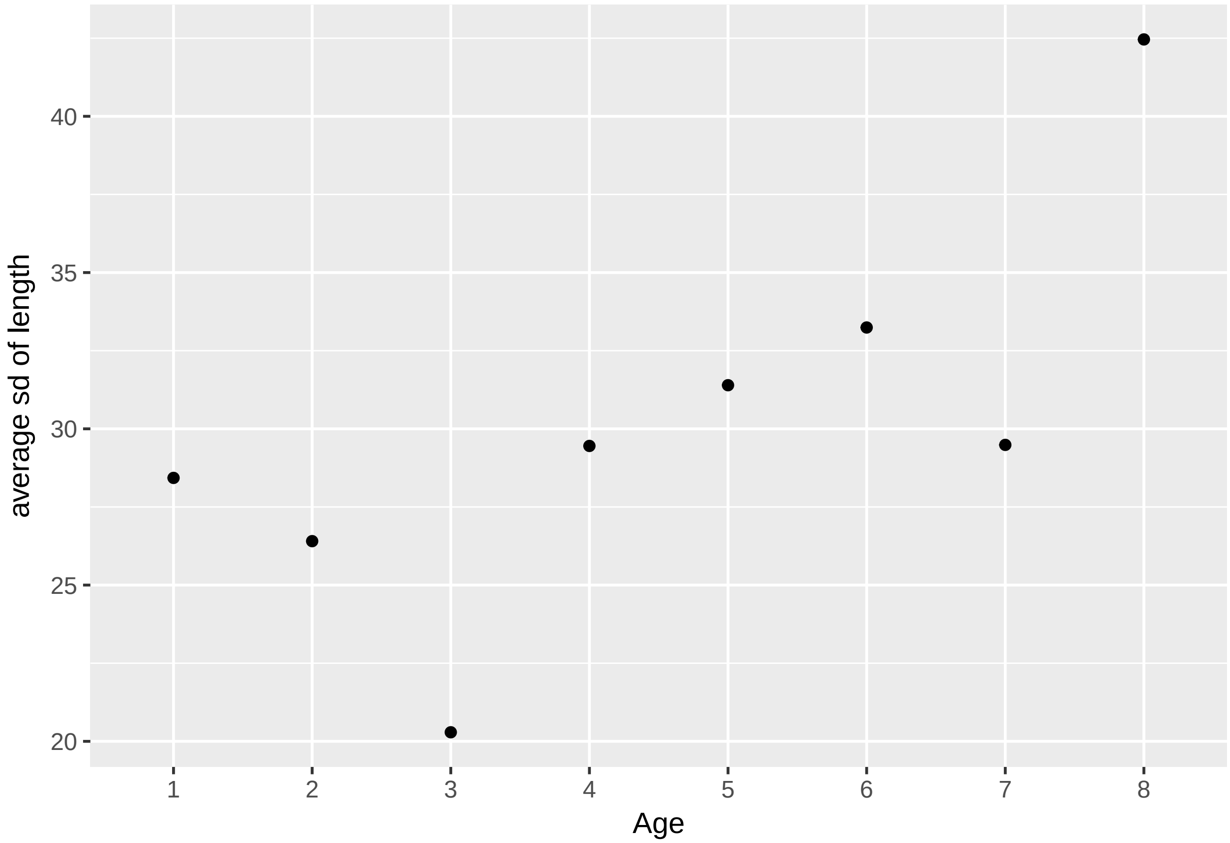


Though I chose to represent the average lengths of the fish of different ages with dots instead of a line, my plot closely resembles the relationship shown in figure 1.5.

Graph of standard deviations versus age.

```
sd.lengths <- c(sqrt(var(age1$Length)), sqrt(var(age2$Length)), sqrt(var(age3$Length)),
    sqrt(var(age4$Length)), sqrt(var(age5$Length)), sqrt(var(age6$Length)),
    sqrt(var(age7$Length)), sqrt(var((age8$Length))))
sd.lengths <- as.data.frame(sd.lengths)
sd.lengths$Age <- c("1", "2", "3", "4", "5", "6", "7", "8")

ggplot(sd.lengths, aes(x = Age, y = sd.lengths)) + geom_point() + labs(y = "average sd of length")
```



Overall, these data show that in general the average length of smallmouth bass from West Bearskin lake increases linearally with age. However, unsuprisingly, the variance and standard deviations of these fish lengths do not show the same relationship. The standard deviation plot does hint at a general pattern of higher deviations in age subpopulations with fewer observations, and lower values in populations with more samples - following the prediction of the law of large numbers whereby a sample mean approaches the population mean as the sample size increases, and thus the standard deviation will also decrease.