STAT 224 Autumn 2022 HW1

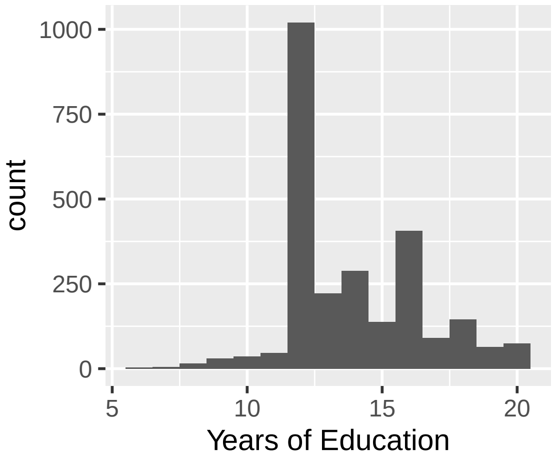
Matthew Zhao

```
NLSY = read.table("http://www.stat.uchicago.edu/~yibi/s224/data/NLSY.txt", header=T)
```

```
library(mosaic)
```

Q1

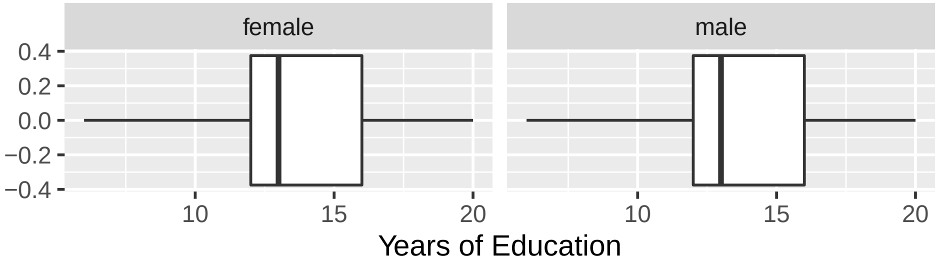
```
library(ggplot2)
ggplot(data = NLSY, aes(x=Edu2006)) +
  geom_histogram(binwidth = 1) +
  xlab('Years of Education')
```



The location of the modes is due to how to the education system in the US works, specifically the divisions between 1-12 (elementary, middle, and high school), 13-16 (college), and beyond. Since education before college is provided to all Americans, vast majority can at least complete up to 12 years, hence the large peak at 12. Then, some portion can afford college (but not all complete it), explaining the peak at 16. Finally, there are also many 2-3 year programs e.g. MBA, MA/MS, JD, etc explaining the small peak after. The ideal bin size here is around 1.

Q2

```
ggplot(data = NLSY, aes(x=Edu2006)) +
  geom_boxplot() +
  facet_wrap(~Gender) +
  xlab('Years of Education')
```



```
library(tidyverse)
male <- NLSY %>% filter(Gender=='male')
female <- NLSY %>% filter(Gender=='female')

print('Male Edu Summary:')
```

```
## [1] "Male Edu Summary:"
```

```
summary(male$Edu2006)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   12.00   13.00   13.81   16.00   20.00
```

```
print('Female Edu Summary:')
```

```
## [1] "Female Edu Summary:"
```

```
summary(female$Edu2006)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    6.00   12.00   13.00   13.97   16.00   20.00
```

```
print('Male Edu Summary:')
```

```
## [1] "Male Edu Summary:"
```

```
sd(male$Edu2006)
```

```
## [1] 2.588275
```

```
print('Female Edu Summary:')
```

```
## [1] "Female Edu Summary:"
```

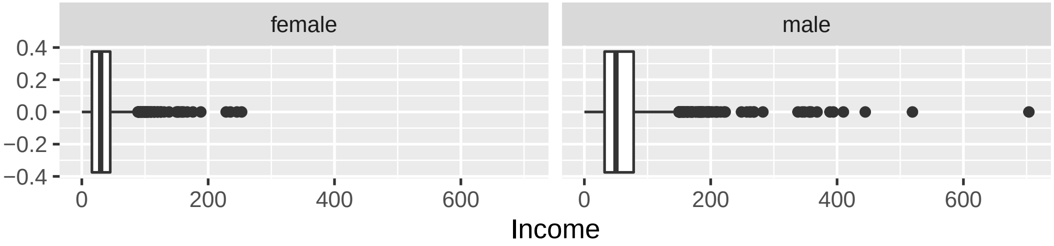
```
sd(female$Edu2006)
```

```
## [1] 2.412262
```

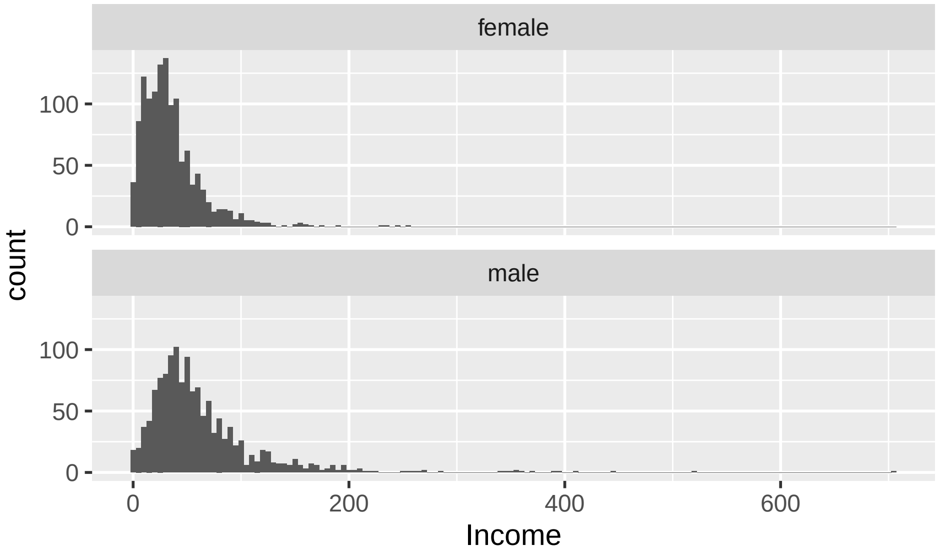
It appears that men and women have similar education levels. The boxplots are identical because the distribution of education is the same for both genders.

Q3

```
ggplot(data = NLSY, aes(x=Income2005)) +
  geom_boxplot() +
  facet_wrap(~Gender) +
  xlab('Income')
```



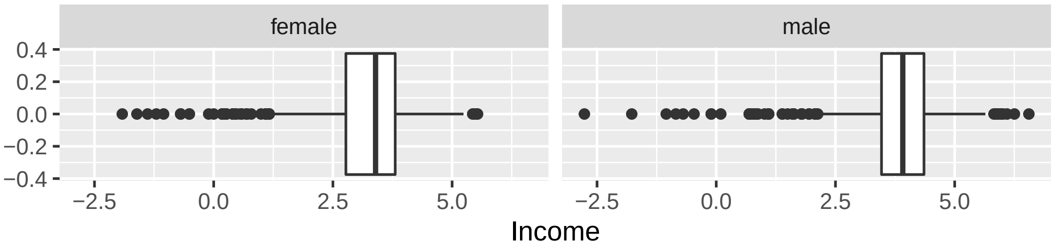
```
ggplot(data = NLSY, aes(x=Income2005)) +
  geom_histogram(binwidth = 5) +
  facet_wrap(~Gender,ncol=1) +
  xlab('Income')
```



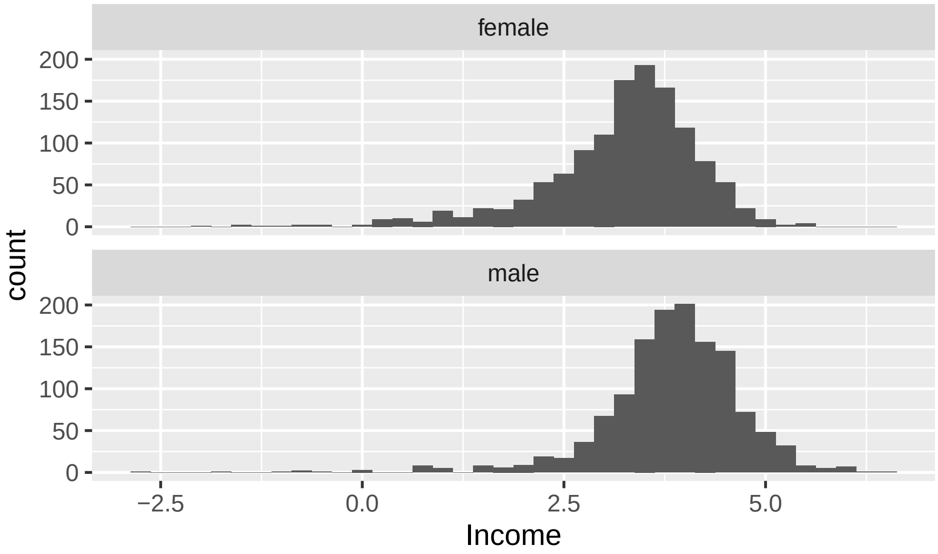
Both histograms are right-skewed and unimodal but the distribution of male income has greater variation (more spread out distribution) and a higher mean and median income as a result.

Q4

```
ggplot(data = NLSY, aes(x=log(Income2005))) +
  geom_boxplot() +
  facet_wrap(~Gender) +
  xlab('Income')
```



```
ggplot(data = NLSY, aes(x=log(Income2005))) +
  geom_histogram(binwidth = 0.25) +
  facet_wrap(~Gender,ncol=1) +
  xlab('Income')
```

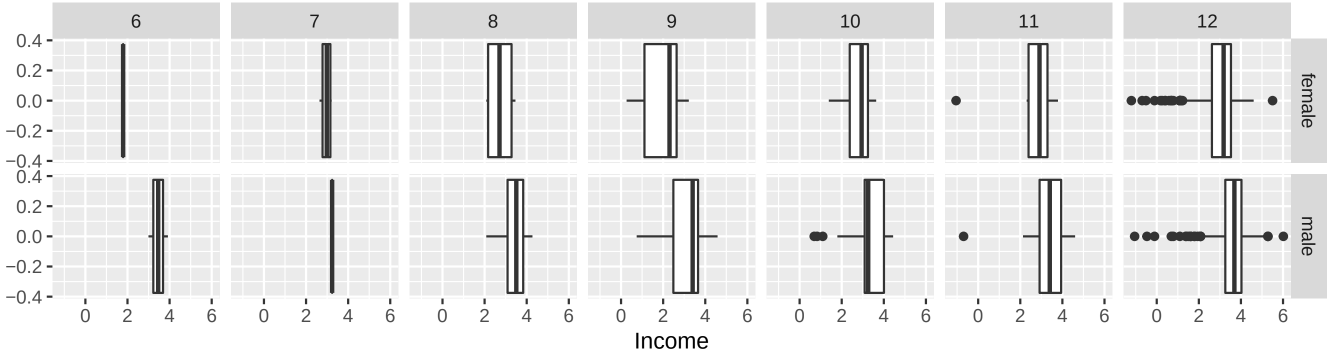


After transformation, the distributions become roughly normal, with a slight left skew.

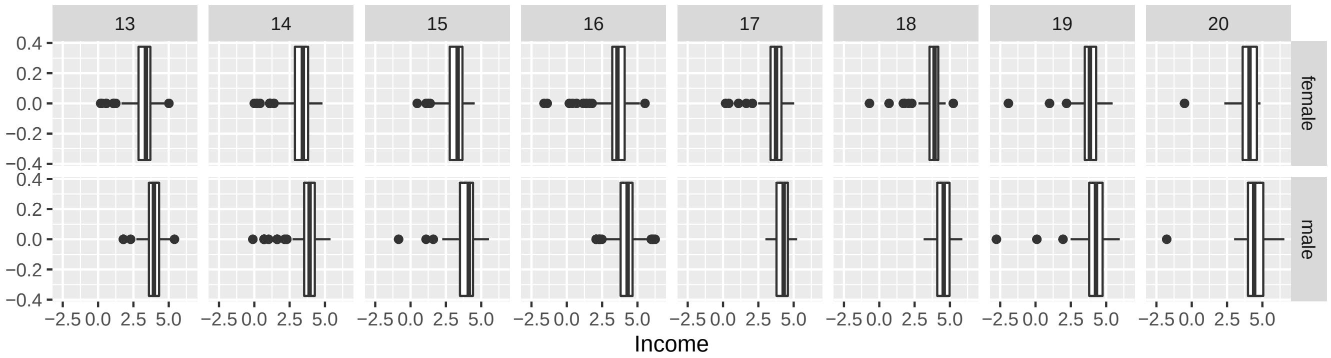
Q5

```
edu_to12 <- NLSY[NLSY['Edu2006']<=12,]
edu_post12 <- NLSY[NLSY['Edu2006']>12,]

ggplot(data = edu_to12, aes(x=log(Income2005))) +
  geom_boxplot() +
  facet_grid(vars(Gender),vars(Edu2006)) +
  xlab('Income')
```



```
ggplot(data = edu_post12, aes(x=log(Income2005))) +
  geom_boxplot() +
  facet_grid(vars(Gender),vars(Edu2006)) +
  xlab('Income')
```

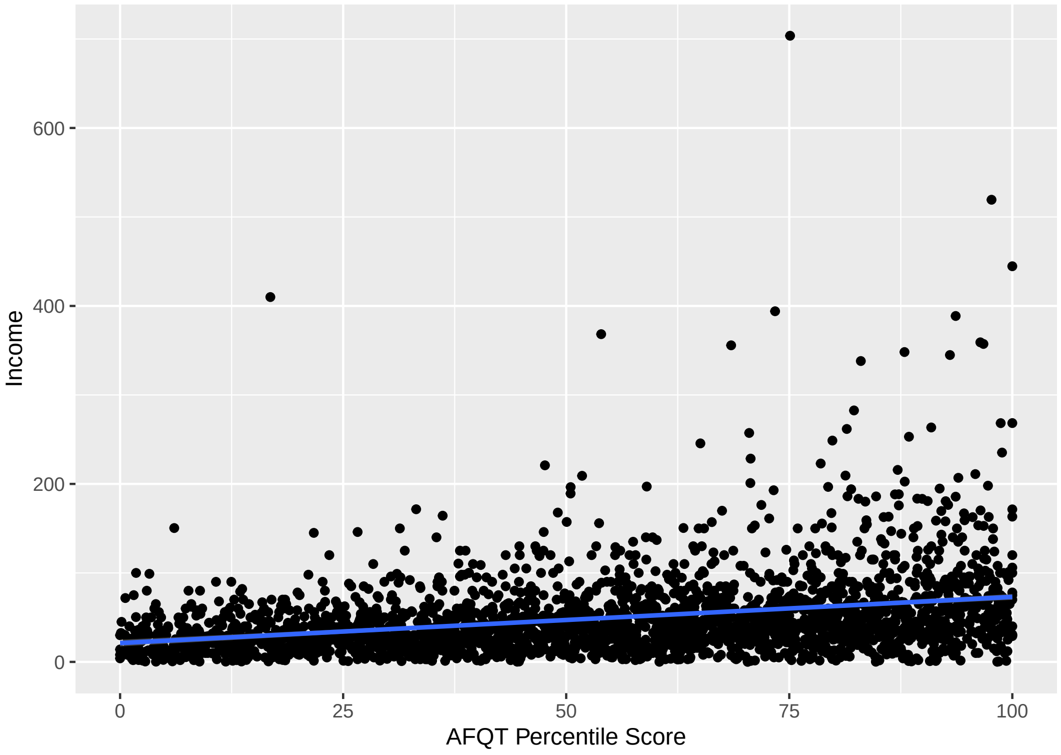


Now we can see that, by comparing the top graphs with the bottom ones, men generally earn more than women with the same education level.

Q6

a)

```
ggplot(data = NLSY, aes(x=AFQT, y=Income2005)) +
  geom_point() +
  geom_smooth(method='lm') +
  xlab('AFQT Percentile Score') + ylab('Income')
```



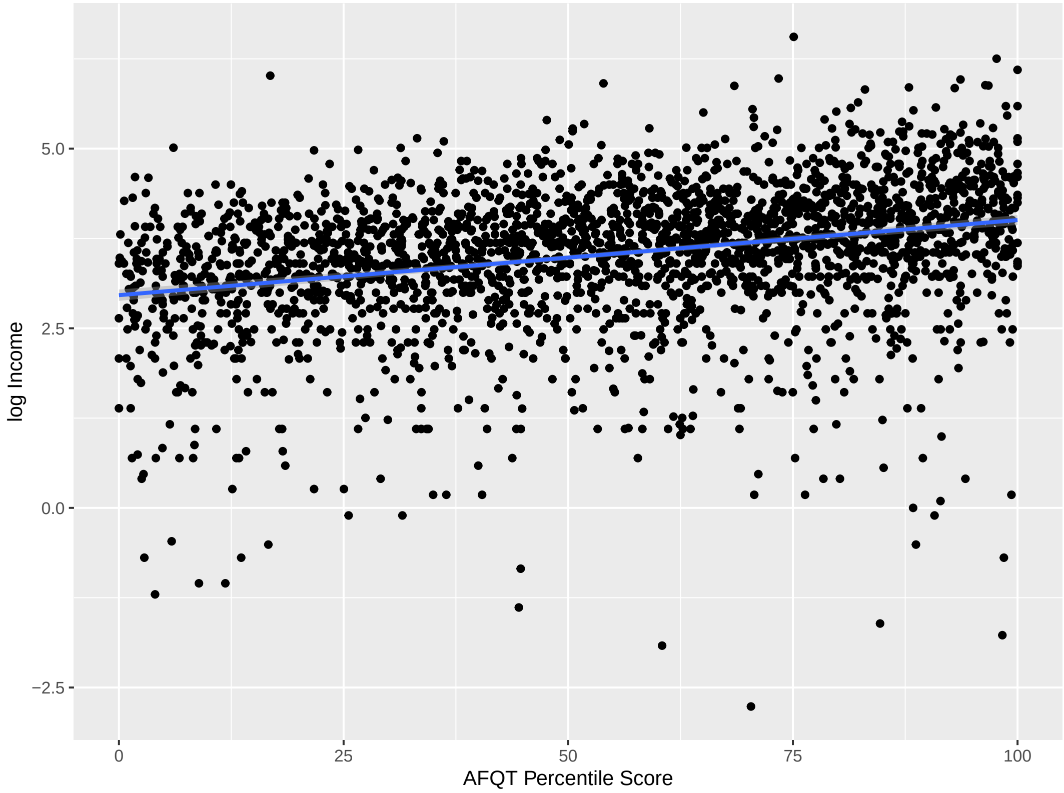
Income appears to slightly increase with AFQT Percentile Score.

b)

Variation in income increases with AFQT Percentile Score.

c)

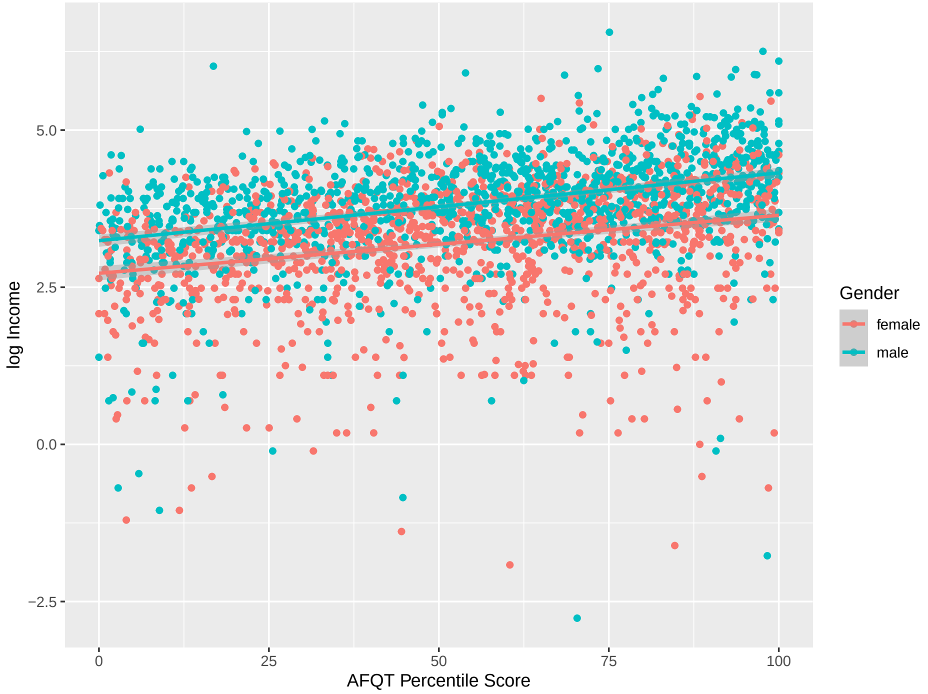
```
ggplot(data = NLSY, aes(x=AFQT, y=log(Income2005))) +
  geom_point() +
  geom_smooth(method='lm') +
  xlab('AFQT Percentile Score') + ylab('log Income')
```



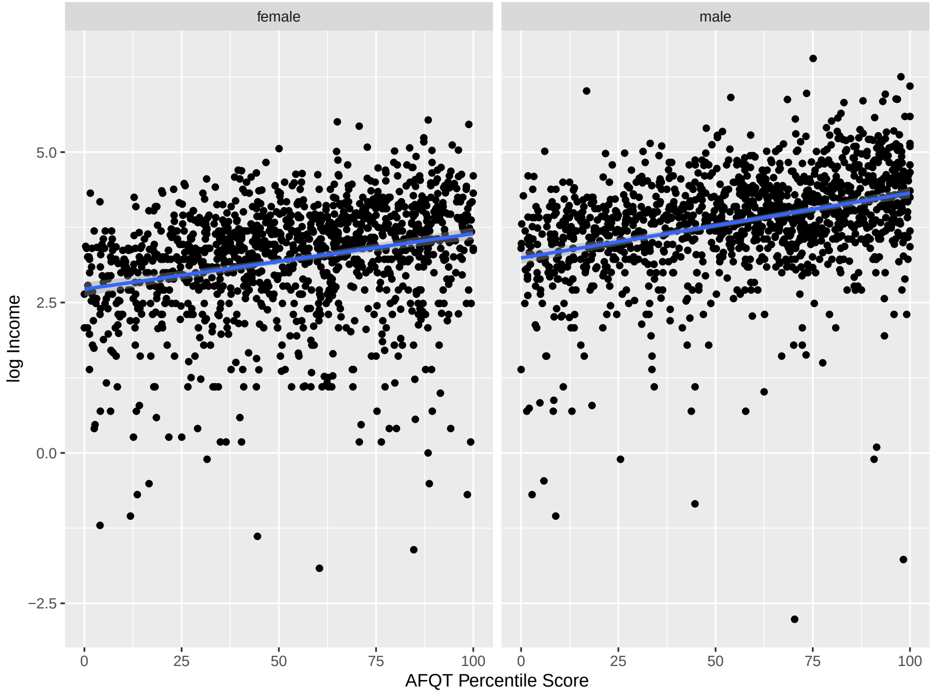
Variability in log income does not appear to change with AFQT.

Q7

```
ggplot(data = NLSY, aes(x=AFQT, y=log(Income2005), color = Gender)) +
  geom_point() +
  geom_smooth(method='lm') +
  xlab('AFQT Percentile Score') + ylab('log Income')
```



```
ggplot(data = NLSY, aes(x=AFQT, y=log(Income2005))) +
  geom_point() +
  facet_wrap(~Gender)+
  geom_smooth(method='lm') +
  xlab('AFQT Percentile Score') + ylab('log Income')
```



a)

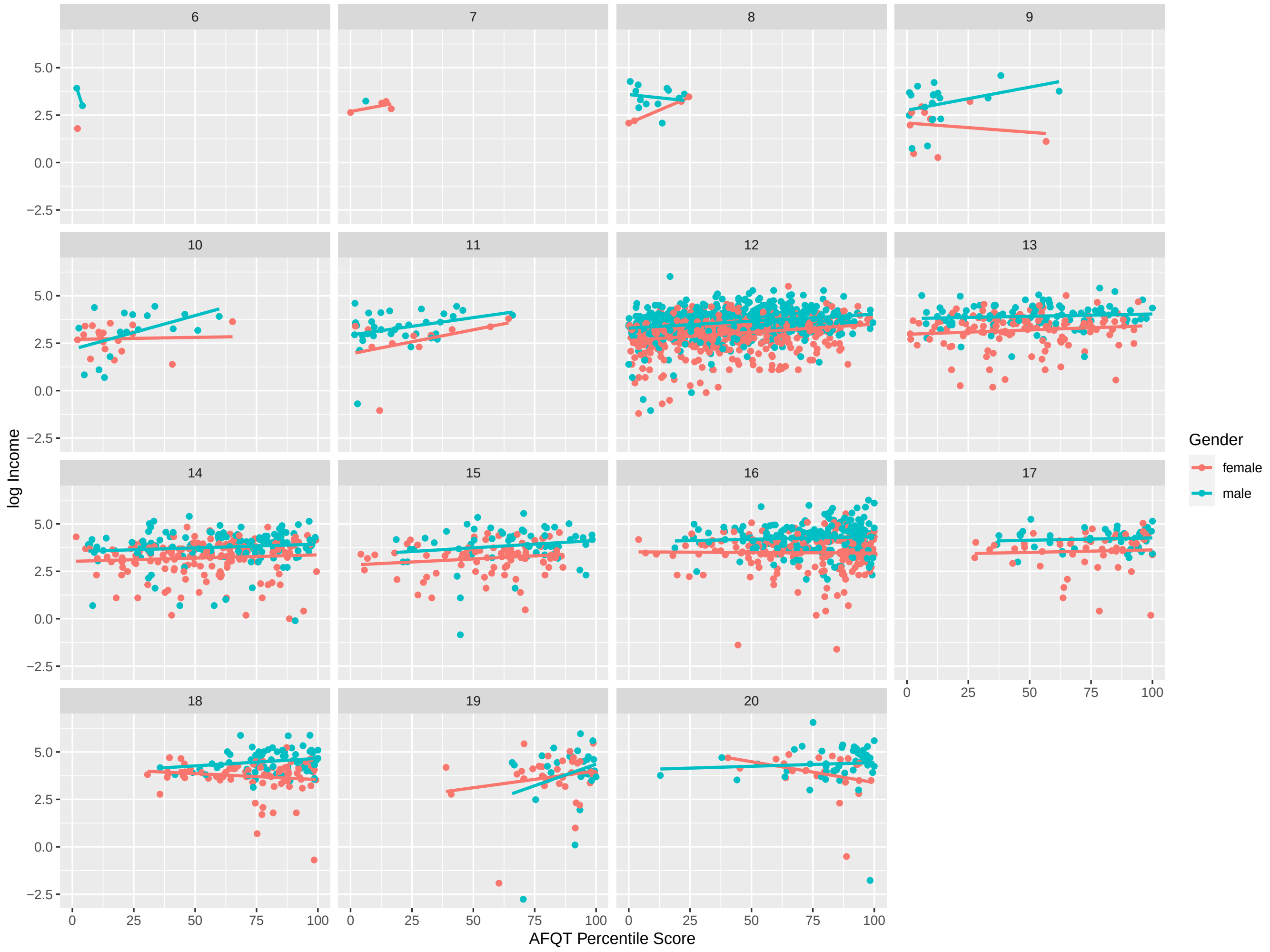
For both genders, log income increases with AFQT Percentile Score and variability stays the same.

b)

Men generally do earn more than women at each AFQT Percentile Score.

Q8

```
ggplot(data = NLSY, aes(x=AFQT, y=log(Income2005), color = Gender)) +
  geom_point() +
  facet_wrap(~Edu2006) +
  geom_smooth(method='lm',se=F) +
  xlab('AFQT Percentile Score') + ylab('log Income')
```



While it is difficult for some education levels, on average, men generally earn more than women even if they have the same AFQT Percentile and education level.