STAT 22000 Summer 2020 HW2 Solutions

Yibi Huang

First, let's load the data set to R and load the mosaic library.

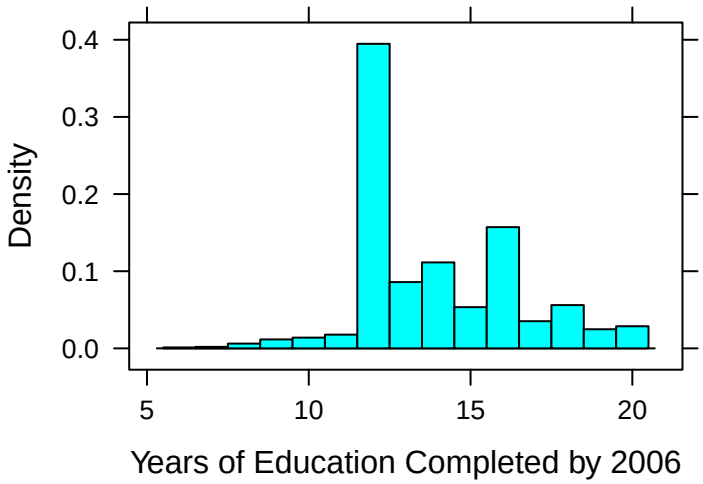
```
NLSY = read.table("NLSY.txt",header=T)
library(mosaic)
```

(a)

[2pts = 1pt for the histogram + 1pt for the comment on the modes of the histogram]

As `Edu2006` is integer-valued, the natural bandwidth for the histogram is 1.

```
histogram(~Edu2006, data=NLSY, width=1, xlab="Years of Education Completed by 2006")
```



The histogram has two peaks at 12 an 16 since most people either finished high school (12 years) or had a college degree (16 years).

(b)

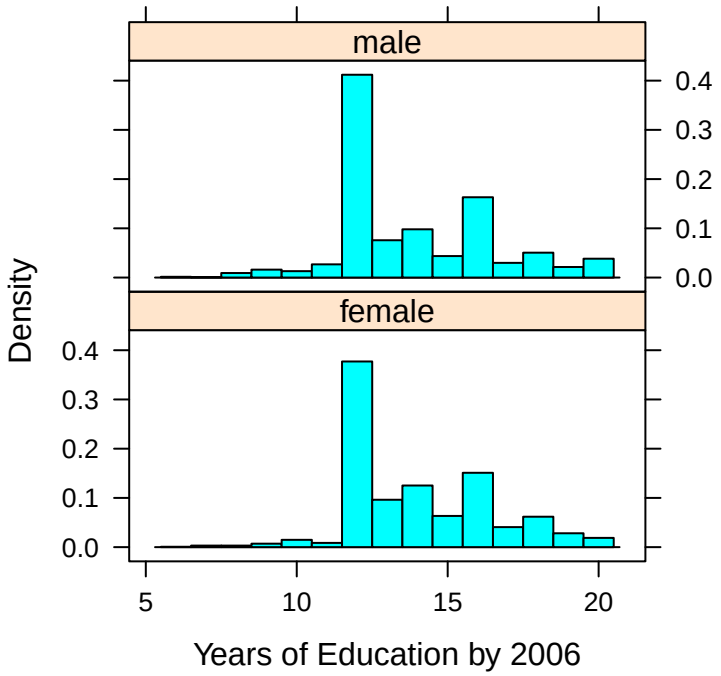
[2pts = 1pt for the plot and the summary + 1pt for the comment]

From the histograms and the data summary below, the education level of males and females were pretty close. They have identical five-number summaries and their means for `Edu2006` were pretty close (13.81 years for men and 13.91 years for women.)

```
favstats(Edu2006 ~ Gender, data=NLSY)
```

```
##   Gender min Q1 median Q3 max     mean       sd    n missing
## 1 female   6 12     13 16  20 13.97027 2.412262 1278       0
## 2   male   6 12     13 16  20 13.81317 2.588275 1306       0
```

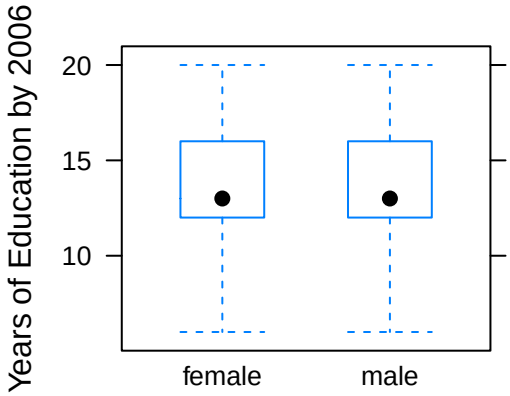
```
histogram(~Edu2006 | Gender, data=NLSY, width=1,
          xlab="Years of Education by 2006",layout=c(1,2))
```



(c)

[2pts = 1pt for the boxplot + 1pt for the missing info] The side-by-side boxplot is as follows. We are not able to observe the modes at 12 and 16 from the boxplot which can be observed in the histogram.

```
bwplot(Edu2006 ~ Gender, data=NLSY, ylab="Years of Education by 2006")
```

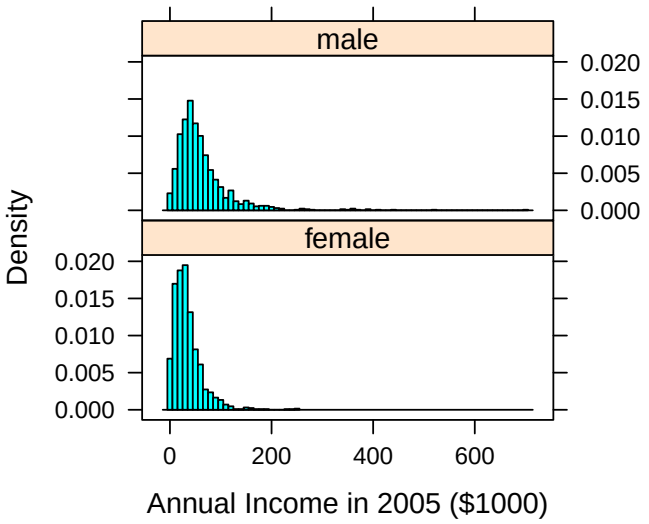


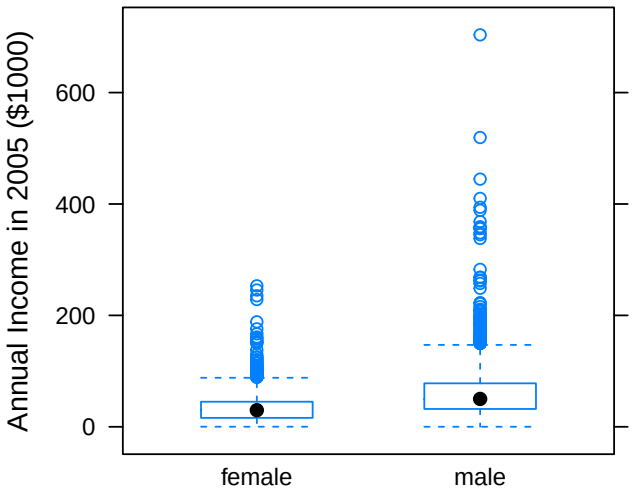
(d)

[4pts in total = 1pt for the plots + 3pts for the comments (1pt each)]

```
histogram(~Income2005 | Gender, data=NLSY, width=10,
          xlab="Annual Income in 2005 ($1000)",layout=c(1,2))
bwplot(Income2005 ~ Gender, data=NLSY, ylab="Annual Income in 2005 ($1000)")
favstats(Income2005 ~ Gender, data=NLSY)
```

```
##   Gender   min Q1  median Q3     max     mean       sd    n missing
## 1 female 0.147 16 29.8105 45 253.043 35.21068 28.77637 1278       0
## 2   male 0.063 32 50.0000 78 703.637 63.31874 55.86107 1306       0
```





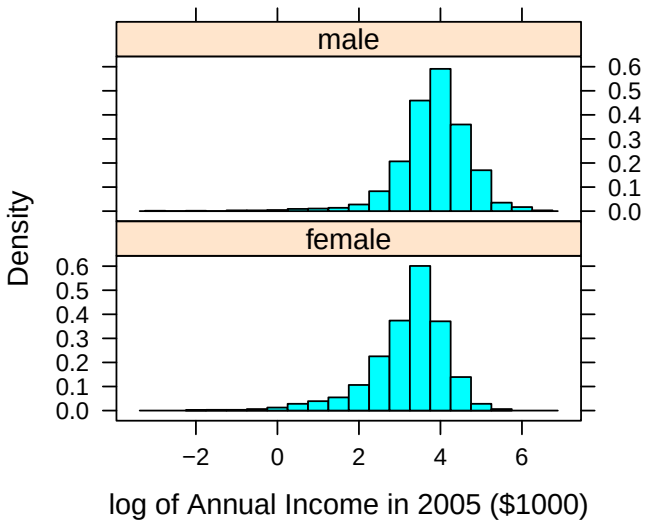
- [1pt] The distributions of `Income2005` are **right-skewed** for both gender
- [1pt] Males generally had a higher income. The mean, Q1, median, and Q3 for males were all higher than those for females.
- [1pt] Males' distribution of income also had a higher variability than females' as SD was 55.9 thousands of dollars for males and 28.8 thousands of dollars for females. One can also argue that males' histogram had a larger spread or the height of the box (IQR) for males' boxplot is higher than that for females.

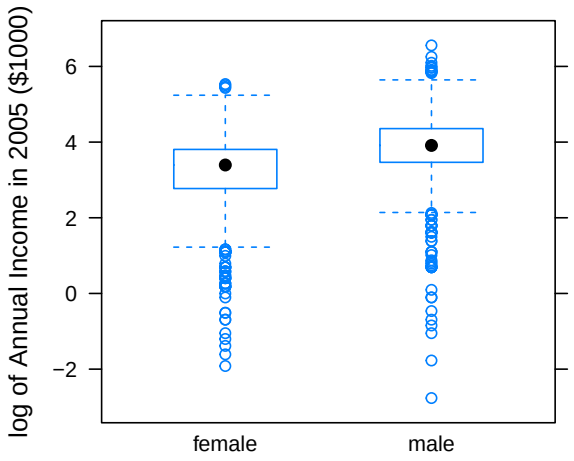
(e)

[2pts = 1pt for the plots + 1pt for the left-skewness]

After the log-transformation, the distribution of `Income2005` become **left-skewed** (the left tail is longer).

```
histogram(~log(Income2005) | Gender, data=NLSY, width=0.5,
          xlab="log of Annual Income in 2005 ($1000)",layout=c(1,2))
bwplot(log(Income2005) ~ Gender, data=NLSY,
       ylab="log of Annual Income in 2005 ($1000)")
```

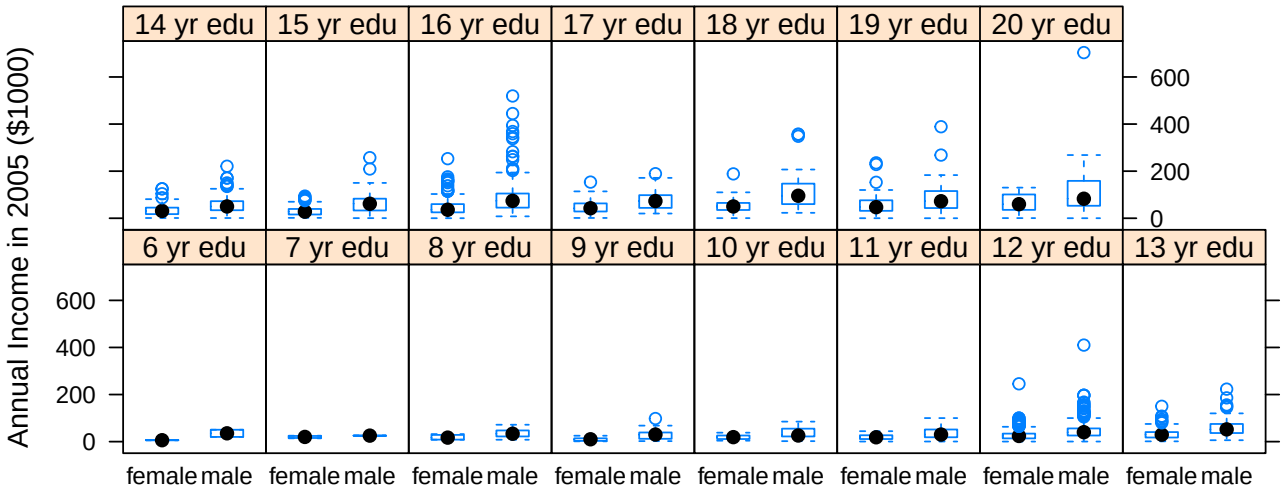




(f)

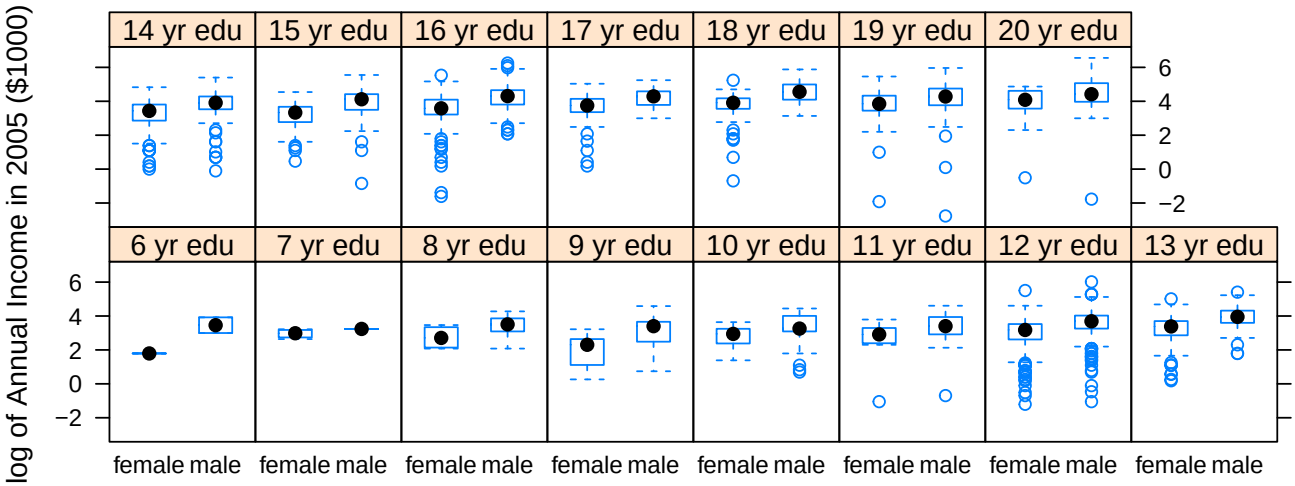
[2pts = 1pt for the plots + 1pt for the comparison]

```
NLSY$Edu2006.fac = factor(NLSY$Edu2006, labels = paste(6:20,"yr edu"))
bwplot(Income2005 ~ Gender | Edu2006.fac, data=NLSY, layout=c(8,2),
       ylab="Annual Income in 2005 ($1000)")
```



Taking logarithm might make it easier to tell which gender had a higher income. We can see that men had a higher median income than women in all the plots. Hence, we can conclude men earned more than women in general, even if they received the same number of years of education.

```
bwplot(log(Income2005) ~ Gender | Edu2006.fac, data=NLSY,
      layout=c(8,2), ylab="log of Annual Income in 2005 ($1000)")
```



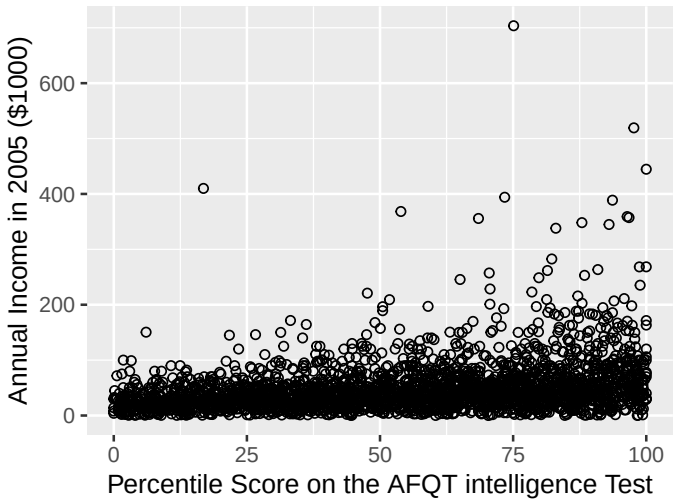
(g)

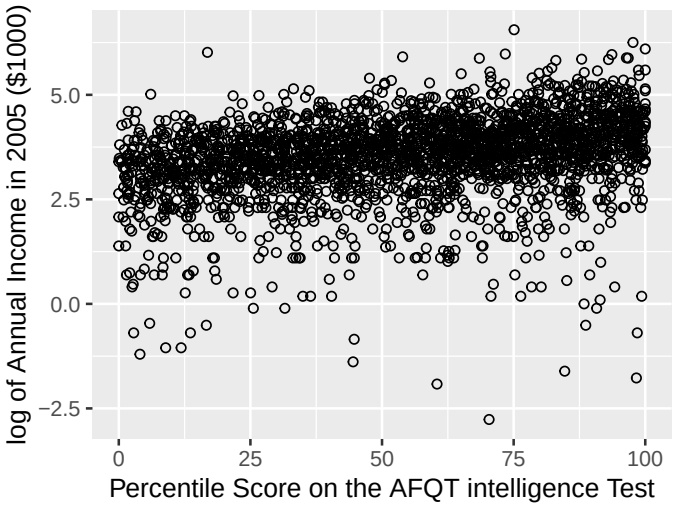
[4pts in total = 1pt for the plots + 3pts for the comments (1pt each)]

1. [1pt] We can see that `Income2005`in general **increases** with `AQFT`.
2. [1pt] The variability of `Income2005` **increases** with `AQFT`. As `AQFT` gets higher, the spread of `Income2005` also gets higher.
3. [1pt] After taking logarithm, the variability of the **logarithm** of `Income2005` doesn't seem to change with `AQFT` as the points scatter in a band of constant width.

The scatter plots are shown below.

```
qplot(AFQT, Income2005, data=NLSY, shape=I(1),
      xlab="Percentile Score on the AFQT intelligence Test",
      ylab="Annual Income in 2005 ($1000)")
qplot(AFQT, log(Income2005), data=NLSY, shape=I(1),
      xlab="Percentile Score on the AFQT intelligence Test",
      ylab="log of Annual Income in 2005 ($1000)")
```



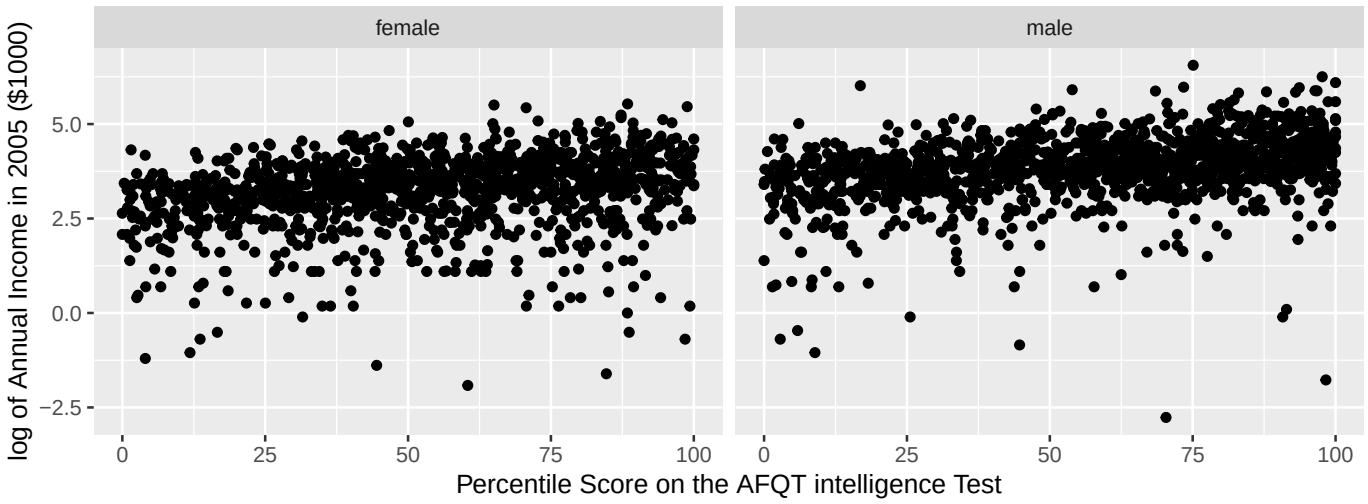


(h)

[4pts in total = 1pt for the plots + 3pts for the comments (see below)]

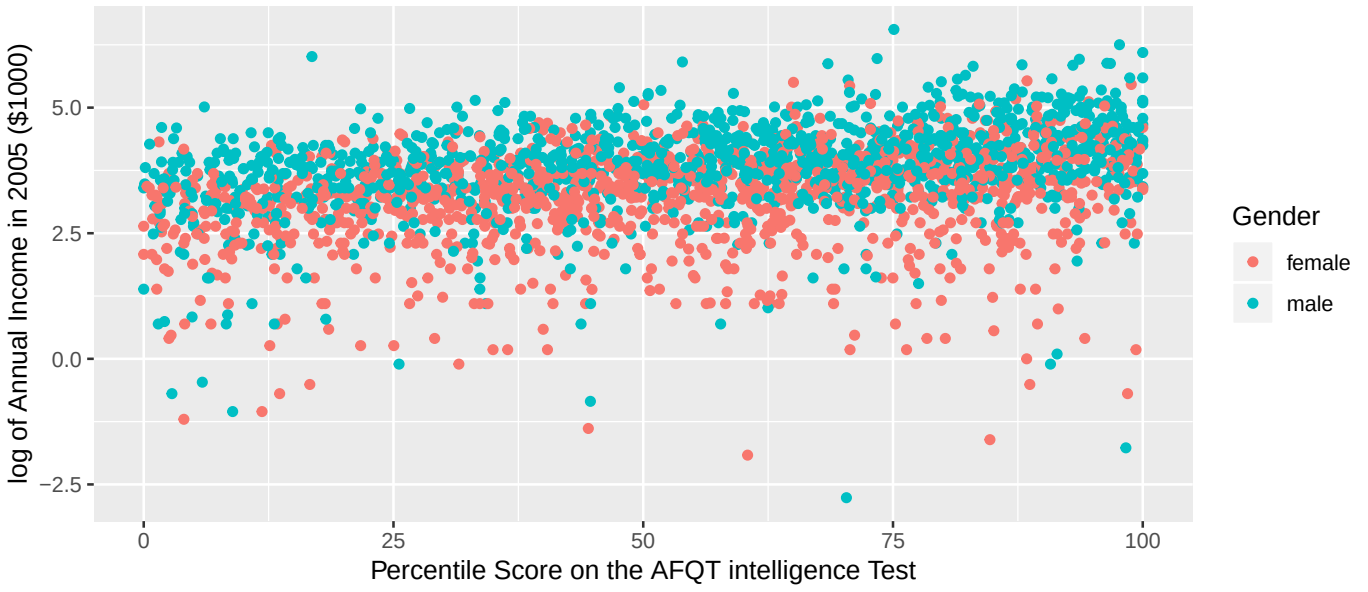
1. [2pts] From the plot below, the logarithm of `Income2005` increase with `AQFT` for each gender. However, the variability of the logarithm of `Income2005` doesn't seem th change with `AQFT` as the points just scatter in a band of a constant width.

```
qplot(AFQT, log(Income2005), data=NLSY,facets = ~Gender,
      xlab="Percentile Score on the AFQT intelligence Test",
      ylab="log of Annual Income in 2005 ($1000)")
```



2. [1pt] Comparing points with similar x-values in the plot below, we see that blue points (males) tend to have higher y-values than points (females). Men generally earned more than women, even if they have similar intellegence test score percentiles.

```
qplot(AFQT, log(Income2005), data=NLSY, color=Gender,
      xlab="Percentile Score on the AFQT intelligence Test",
      ylab="log of Annual Income in 2005 ($1000)")
```

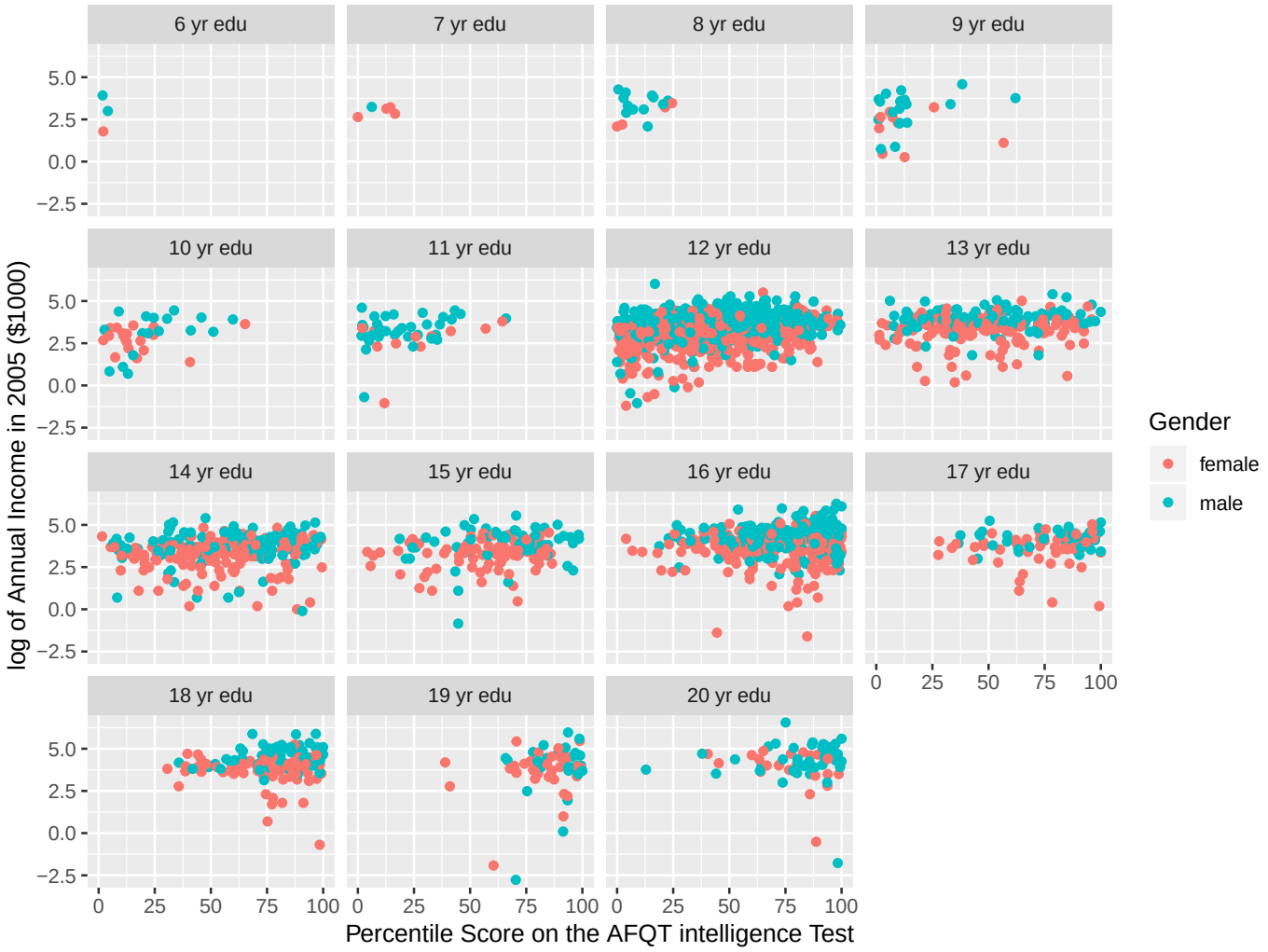


(i)

[3pts] In the split-panel scatter plot below, note that points in a single subplot were of the same years of education. We can compare the annual incomes of males and females controlling for `Edu2006` and `AFQT` by comparing the y-values of blue points (males) and red points (females) with similar x-values in a single subplots as such cases had identical `Edu2006` and similar `AFQT` percentiles.

We see that in nearly all of the subplots, blue points (males) tend to have higher y-values than points (females). We can then conclude that men generally earned more than women, even if they received the same years of education and had the same intellegence test score percentiles.

```
qplot(AFQT, log(Income2005), data=NLSY,
      color=Gender, facets=~Edu2006.fac,
      xlab="Percentile Score on the AFQT intelligence Test",
      ylab="log of Annual Income in 2005 ($1000)")
```



(Not required.) Adding separate linear regression lines between `log(Income2005)` and `AFQT` for men and women in each subplot makes it easier to compare the `log(Income2005)` of the two gender. (A linear regression line between x and y is the straight line that best depicts the linear relationship between x and y in a scatterplot.) We can see that the blue lines are almost always above the red lines, meaning that males generally had higher `log(Income2005)` than females, after adjusting for `Edu2006` and `AFQT`.

```
qplot(AFQT, log(Income2005), data=NLSY,
      color=Gender, facets=~Edu2006.fac, shape=I(1),
```

```
    xlab="Percentile Score on the AFQT intelligence Test",
    ylab="log of Annual Income in 2005 ($1000)")+
geom_smooth(method="lm")
```