Applied Regression Analysis
Homework 1

Mitchell Matheny

June 14, 2018

For problems 6 through 10. I have typed out the short version of the work with solutions. The hand written work is attached at the end of this document.

1 Exercise 1.1

The study used in this exercise looks at a study of teenage gambling in Britain. Make a numerical and graphical summary of the data, commenting on what you notice.
When initially looking at the summary of the data. One notices that the number male and female entries are treated as numerical data. We therefor need to convert them into what is called a factor. This is accomplished by using the following command:

teengamb$sex ¡ - factor(teengamb$sex)

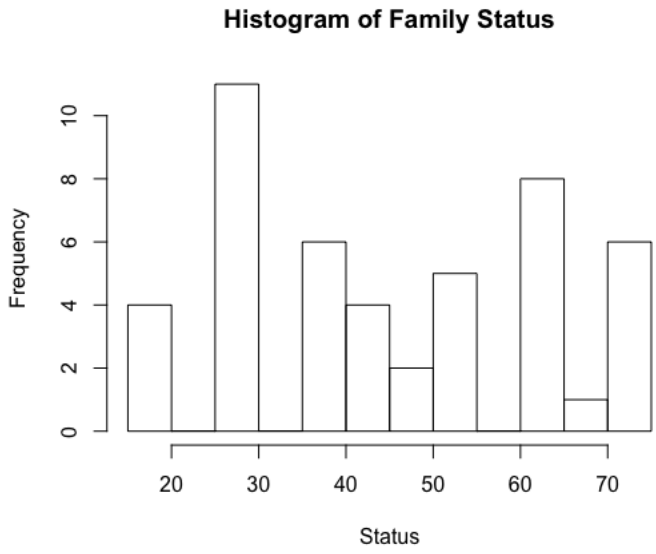
After doing so the summary looks like this:

```
> summary(teengamb)
    sex         status          income          verbal          gamble
 Male  :28   Min.   :18.00   Min.   : 0.600   Min.   : 1.00   Min.   :  0.0
 Female:19   1st Qu.:28.00   1st Qu.: 2.000   1st Qu.: 6.00   1st Qu.:  1.1
             Median :43.00   Median : 3.250   Median : 7.00   Median :  6.0
             Mean   :45.23   Mean   : 4.642   Mean   : 6.66   Mean   : 19.3
             3rd Qu.:61.50   3rd Qu.: 6.210   3rd Qu.: 8.00   3rd Qu.: 19.4
             Max.   :75.00   Max.   :15.000   Max.   :10.00   Max.   :156.0
```

We can now easily see how many males and females where included in this study. This will also make it much easier to differentiate the sex on the various plots to follow.

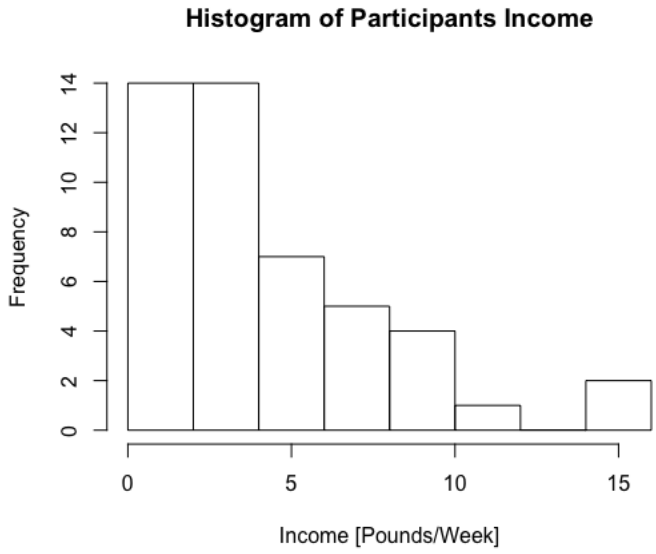
Comments on Numerical Summary and Univariate Plots:

Status The numerical values for family status seem a bit arbitrary. The documentation doesnt really specify how these numbers were calculated, this would take a bit more research into the actual study. With numerical values ranging from 18 to 75, both the mean and median are above the middle of the range of 40. Indicating that the societal status of the parents of the teenagers in this study are on the higher end of the spectrum. We can see this more clearly by looking at a histogram of the data:



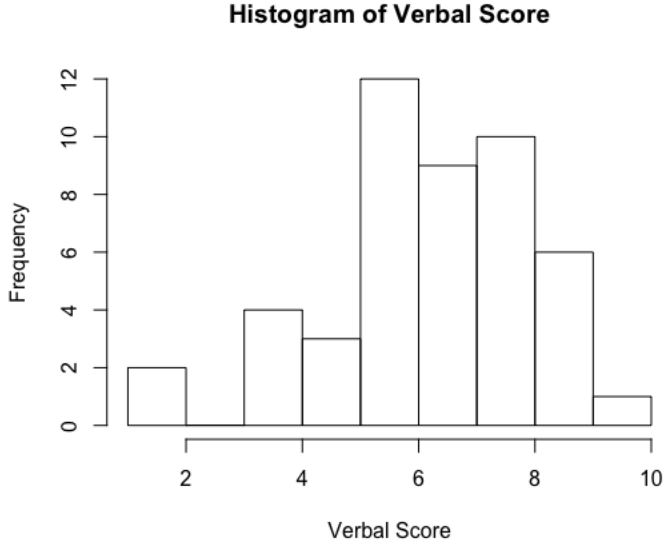
As we can see, the majority of the participants have a status above 40.

Income: The numerical values for income are given in pounds per week. The average income is 4.6 pounds per week, which is equivalent to only $6.16 U.S. This seems like a very low value compared to the average of 500 pounds per week ($670 U.S), but considering that these are teenagers, the values seem acceptable. Looking at the histogram:



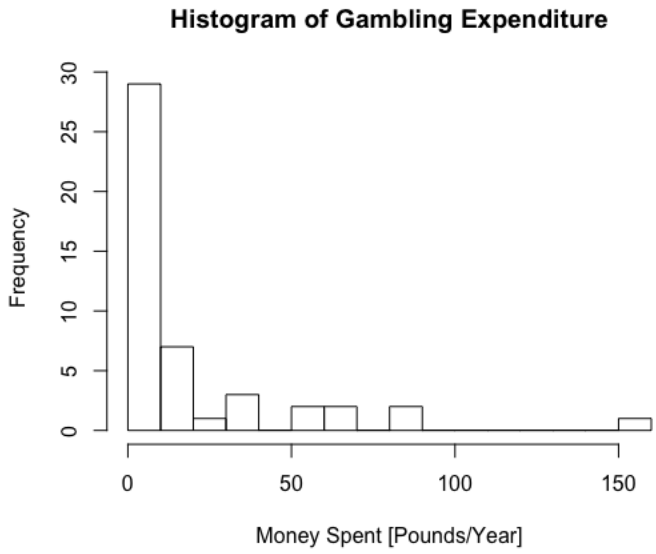
We can see the the 60% of the participants make less than 5 pounds per week ($6.69 U.S), and the average is screwed by the 2 participants making 15 pounds per week ($20.00 U.S)

Verbal: The verbal tests were scored on a scale of 0 to 12. Looking at the summary, we can see that the average value of 6.6 is just above the 50% correct mark, possibly indicating that the scores are screwed towards the higher end of the scale. Looking at the histogram:



We can clearly see that the majority of the participants scored 50% or better.

Gamble: The participants reported the amount of money the spent on gambling in pounds per year. Looking at the summary, we can see that the values have a large range. The minimum being 0 pounds per year spent to 156 pounds spent per year. Due to the large maximum value the mean is shifted to a higher value of 19.3. We can see however that the majority of the participants spend little to no money on gambling as the median value is 6. Looking at the histogram, we can see that this is that case.



Multivariate Plots
For the Multivariate plots, I will only make comments on plots that i notice trends on.

When looking at all of the plots relating a variable to sex, the only visually noticeable trend would be between the amount gambled per year: Men spent more per year than women. There is also a slight trend between the amount of money earned per week and the amount gambled per year. There appears to be a slight upward linear trend, indicating that the more earned per week, the more likely the participant was to spend money on gambling. This however is a speculation, and would need further analysis to confirm. The last trend i noticed was there is a slight linear trend when comparing verbal scores and parental status. The participants with a higher family status tended to score higher on the verbal test. Further analysis is needed to determine the true relationship.

Plots are located after problems 6-10

2 Problems 6-10

Problem 6:

Suppose that Y is a random variable with pdf, $f(y) = \lambda e^{-\lambda y}$ Determing the mean and variance of Y. The mean is just the expectaion value.

$$E[y] = \int y\lambda e^{-\lambda y} dy \tag{1}$$

Solving via integration by parts, the mean is therefore:

$$E[y] = \frac{-e^{-\lambda y}(\lambda y + 1)}{\lambda} \tag{2}$$

To find the Variance:

$$\begin{gathered} Var[y] = E[y^2] - E[y]^2 \\ E[y^2] = \int y^2 \lambda e^{-\lambda y} dy = \frac{e^{-\lambda y}(\lambda^2 y^2 + 2\lambda y + 2)}{\lambda^2} \end{gathered} \tag{3}$$

Therefore the variance:

$$Var[y] = \frac{e^{-\lambda y}(\lambda^2 y^2 + 2\lambda y + 2)}{\lambda^2} + \frac{-e^{-2\lambda y}(\lambda y + 1)^2}{\lambda^2} \tag{4}$$

Problem 7:

Determing the mean and variance of 4Y +2.

Using the following property:

$$\begin{gathered} E[aY + b] = aE[Y] + b \\ E[4Y + 2] = 4e[y] + 2 \end{gathered} \tag{5}$$

Therefore the mean is:

$$E[4Y + 2] = \frac{-4e^{-\lambda y}(\lambda y + 1)}{\lambda} + 2 \tag{6}$$

Using a similar argument for the variance:

$$\begin{gathered} Var[aY + b] = a^2 Var[Y] = 16Var[Y] \\ Var[4Y + 2] = \frac{16e^{-\lambda y}(\lambda^2 y^2 + 2\lambda y + 2)}{\lambda^2} + \frac{-16e^{-2\lambda y}(\lambda y + 1)^2}{\lambda^2} \end{gathered} \tag{7}$$

Problem 8:

Suppose that random variables X and Y have a joint PDF given by $f(x,y) = 6(1-y); 0 <= x <= y <= 1$. Determine $f_x, f_y, E(X), E(Y), E(XY), and cov(X,Y)$.

$$f_x = \int_x^1 (6-6y)dy$$
$$f_x = 3x^2 - 6x + 3$$

$$f_y = \int_0^y (6-6y)dx$$
$$f_y = -6y^2 + 6y$$

$$E[x] = \int_0^1 x f_x dx = \int_0^1 (3x^3 - 6x^2 + 3x)dx$$
$$E[x] = \frac{1}{4}$$

$$E[y] = \int_0^1 x f_y dy = \int_0^1 y(-6y^2 + 6y)dy \tag{8}$$
$$E[y] = \frac{1}{2}$$

$$E[xy] = \int_0^1 \int_0^1 xyf(x,y)dxdy = \int_0^1 \int_0^1 xy(6-6y)dxdy$$
$$E[xy] = 1$$

$$Cov[xy] = E[xy] - E[x]E[y]$$
$$Cov[xy] = \frac{7}{8}$$

Problem 9:

Suppose that $y = (Y_1, Y_2)^T$ is a bivariate normal random vectore with mean and variance;

$$E[y] = \begin{pmatrix} 1 & 2 \end{pmatrix}^T$$
$$Var[y] = \begin{pmatrix} 1 & 0.25 \\ 0.25 & 2 \end{pmatrix} \tag{9}$$

What is the distribution, mean, and variance of $a^T y$, where $a = \begin{pmatrix} 1 & 3 \end{pmatrix}^T$

$$f(y) = a^T y = \begin{pmatrix} 1 & 3 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$
$$f(y) = Y_1 + 3Y_2 \tag{10}$$

For the mean:

$$E[a^T y] = a^T E[y]$$
$$E[a^T y] = \begin{pmatrix} 1 & 3 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{11}$$
$$E[a^T y] = 7$$

For the variance:

$$Var[a^t y] = aVar[y]a^T$$
$$Var[a^T y] = \begin{pmatrix} 1 & 3 \end{pmatrix} \begin{pmatrix} 1 & 0.25 \\ 0.25 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 3 \end{pmatrix} \tag{12}$$
$$Var[a^t y] = 20.5$$

Problem 10:

Using the same y as in problem 4, determone the distribution. mean, and variance of $z = Ay$, where

$$A = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \tag{13}$$

$$z = Ay = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix}$$
$$z = \begin{pmatrix} 2Y_1 + Y_2 \\ Y_1 + 2Y_2 \end{pmatrix} \tag{14}$$

For the mean:

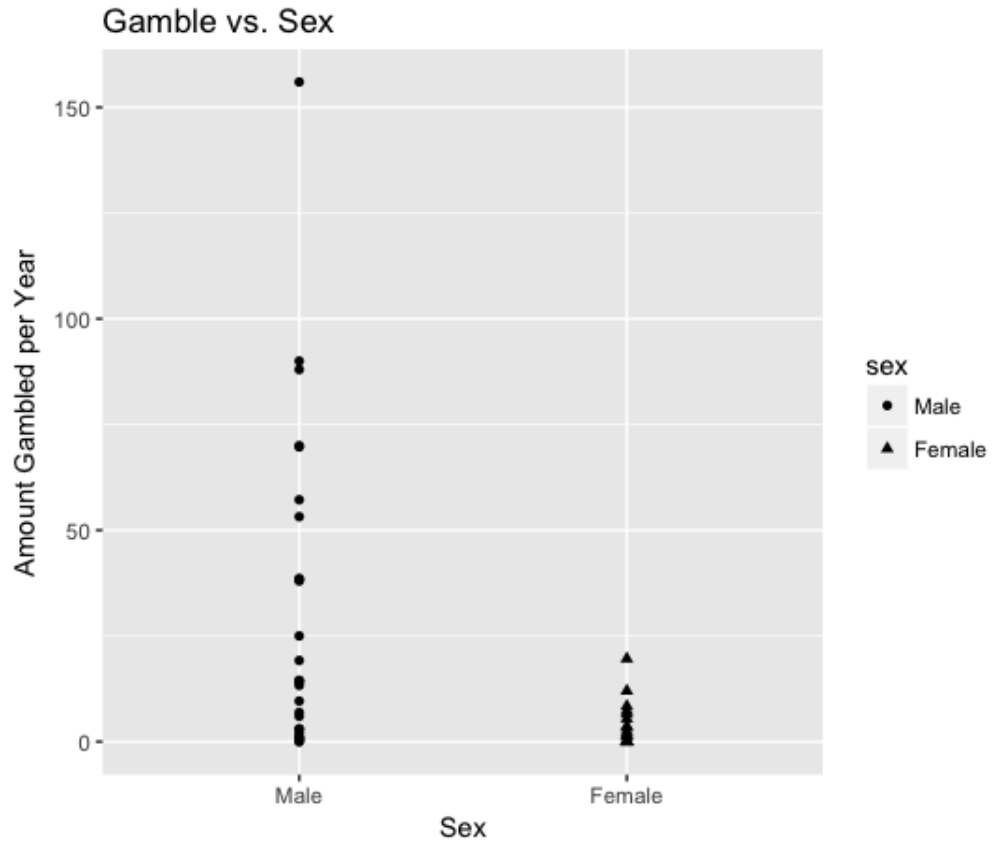
$$E[AY] = AE[Y]$$
$$E[AY] = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \end{pmatrix} \tag{15}$$
$$E[AY] = \begin{pmatrix} 4 \\ 3 \end{pmatrix}$$

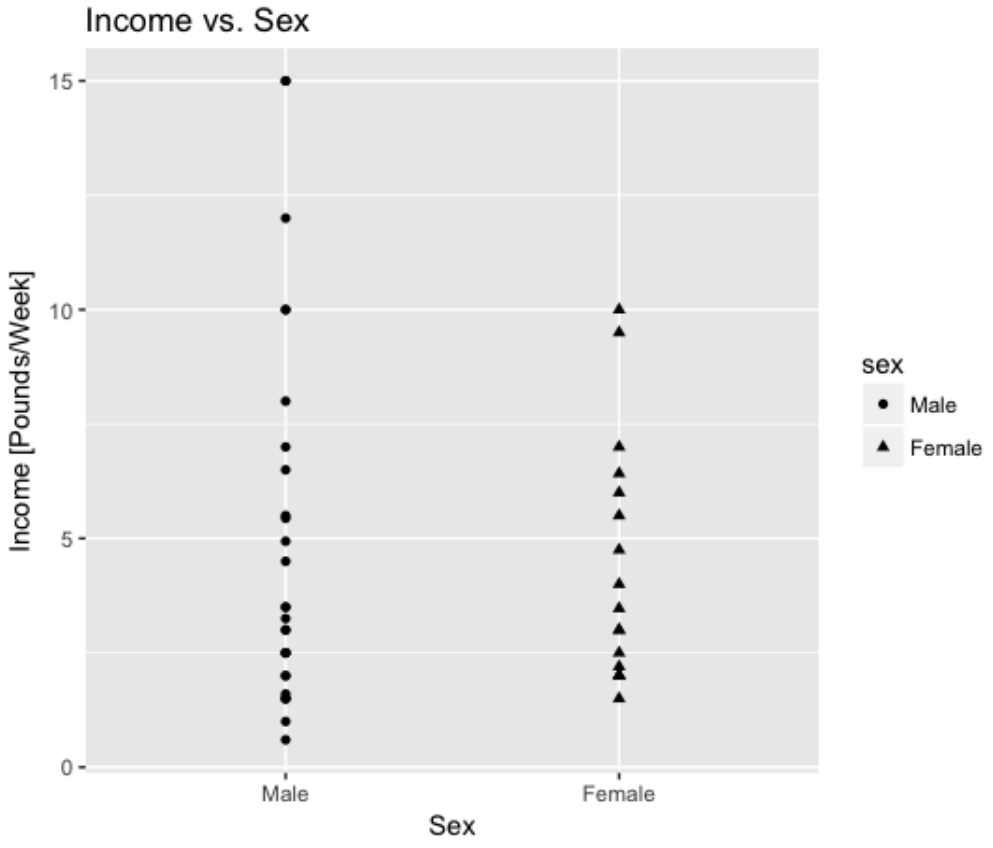
For the Variance:

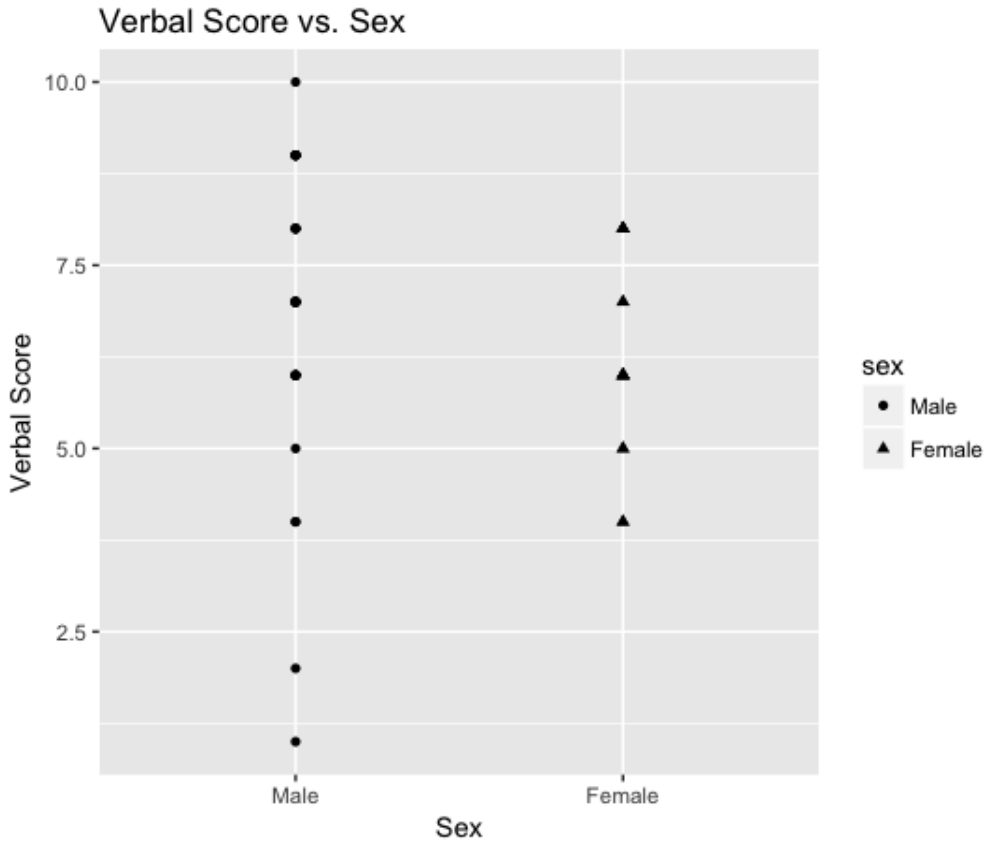
$$Var[AY] = AVar[Y]A^T$$

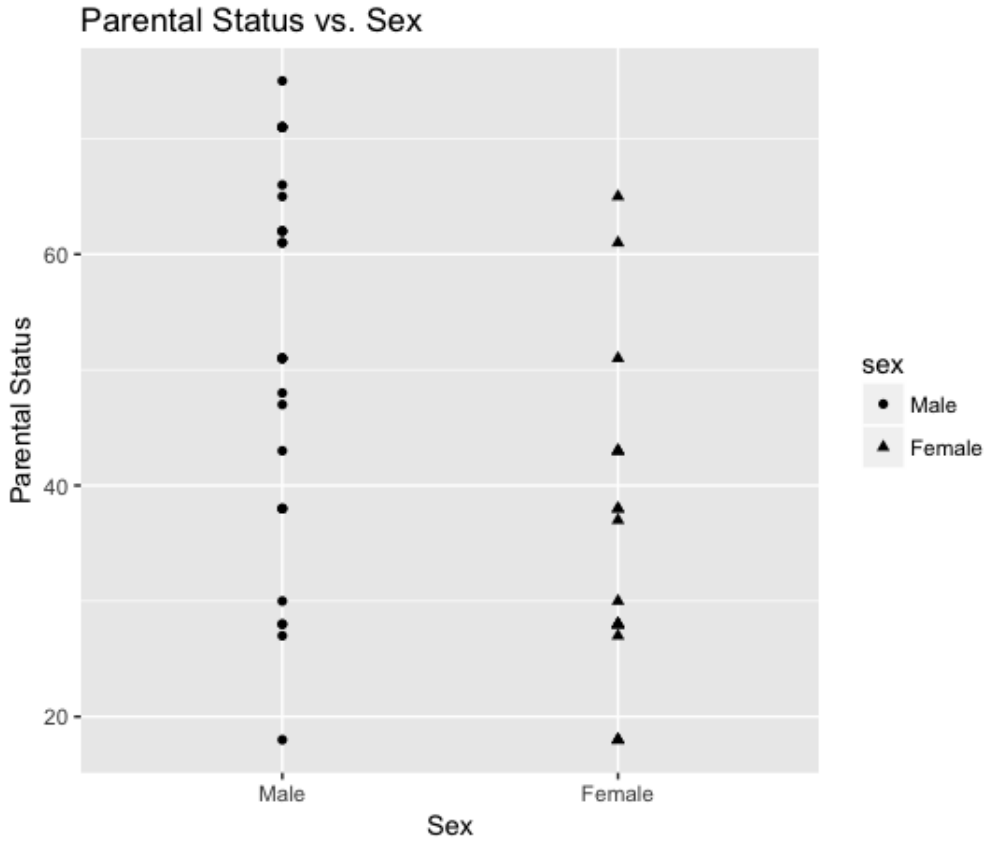
$$Var[AY] = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \begin{pmatrix} 1 & 0.25 \\ 0.25 & 2 \end{pmatrix} \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix} \tag{16}$$

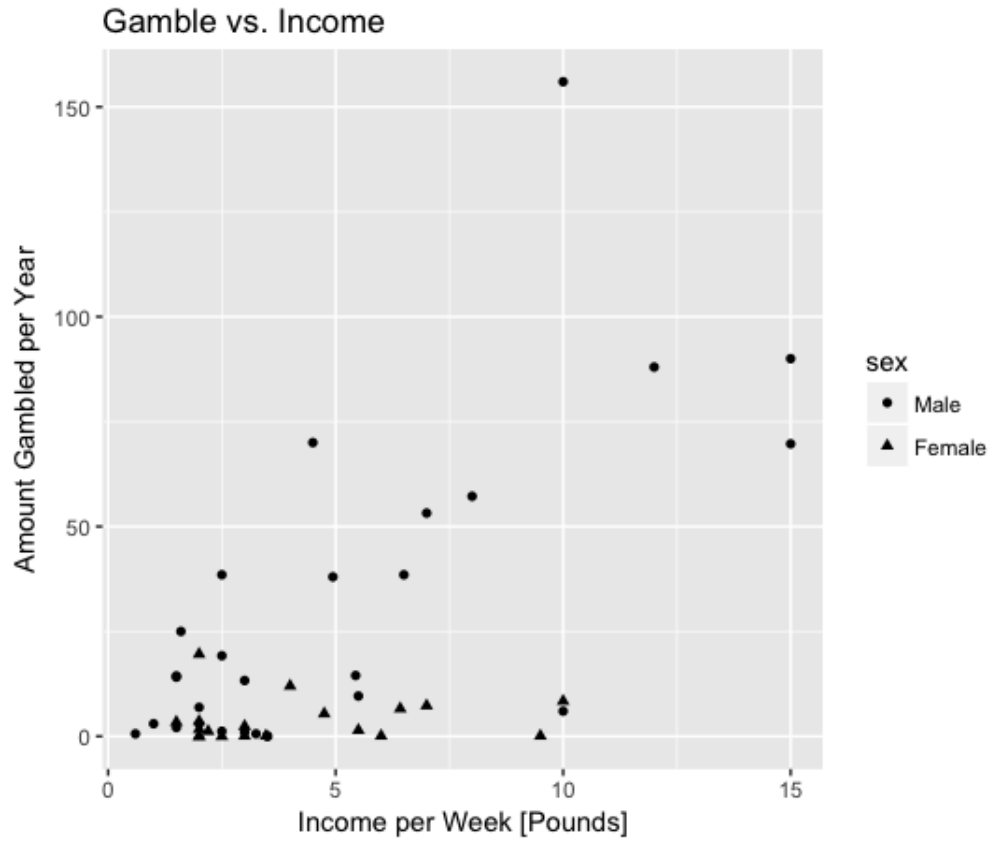
$$Var[AY] = \begin{pmatrix} 3.5 & 2.5 \\ -0.5 & 6 \end{pmatrix}$$

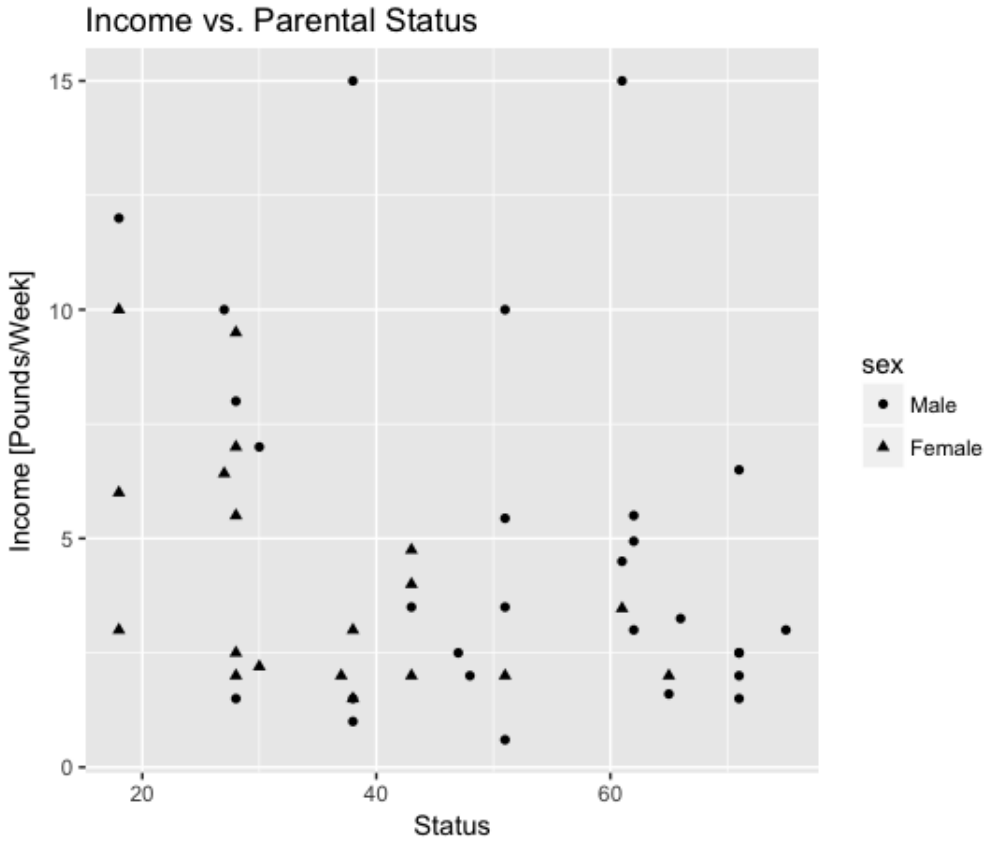

Gamble vs. Sex
Amount Gambled per Year
150
100
50
0
Male
Female
Sex
sex
Male
Female

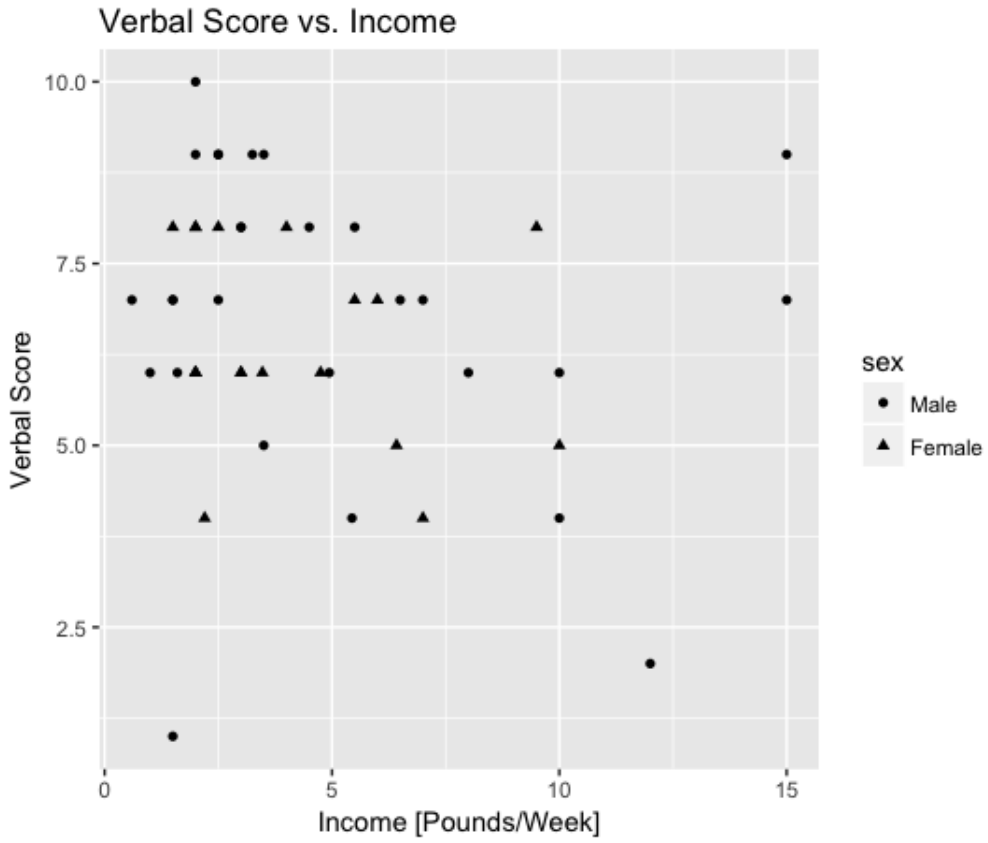

Income vs. Sex
Income [Pounds/Week]
15
10
5
0
Male
Female
Sex
sex
Male
Female

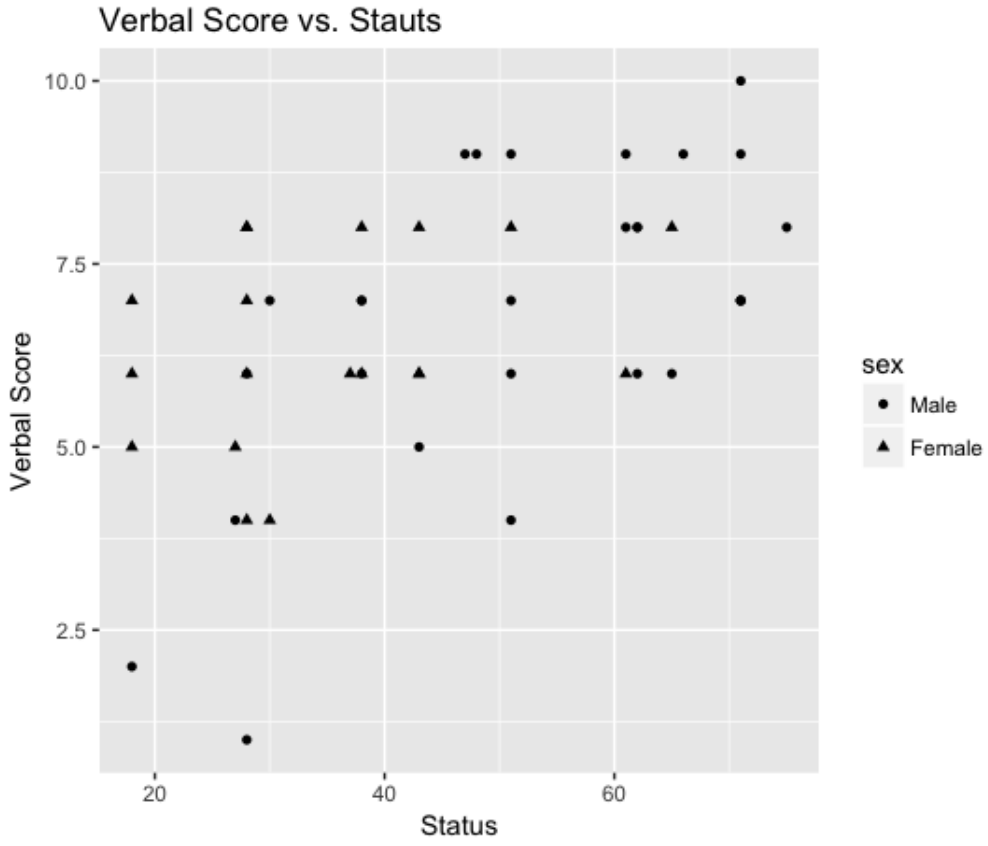

Verbal Score vs. Sex
10.0
7.5
5.0
2.5
Verbal Score
Male
Female
Sex
sex
Male
Female

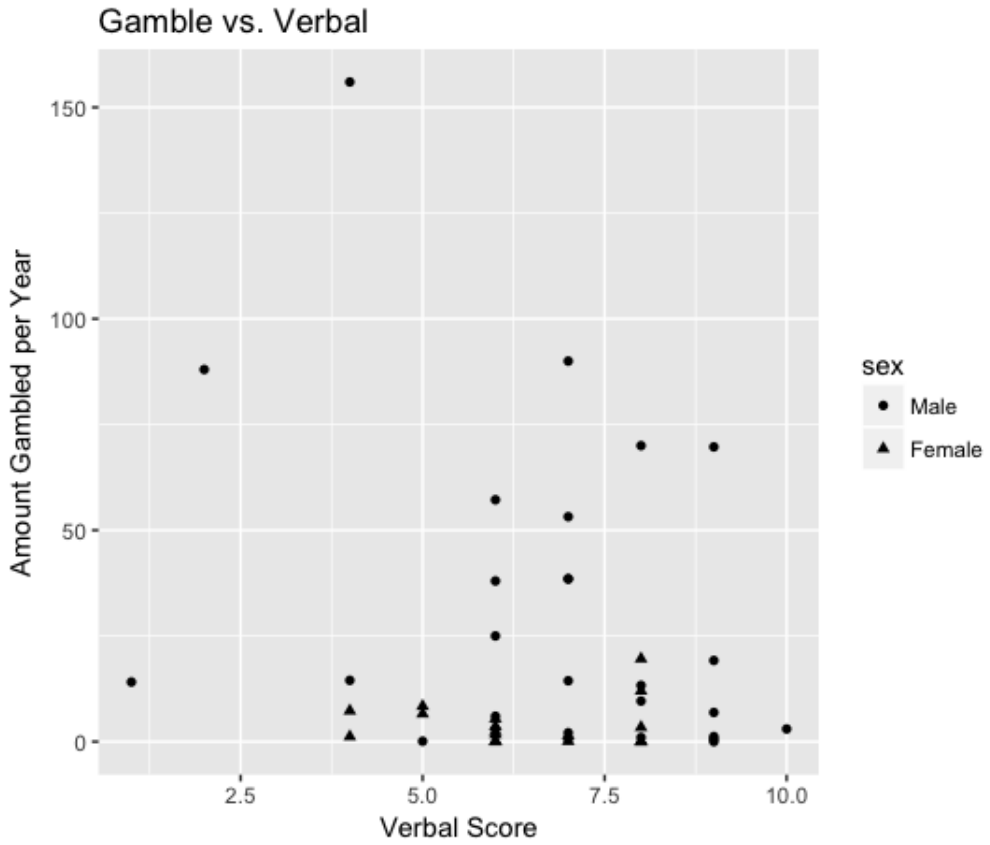

Parental Status vs. Sex
60
40
20
Parental Status
Male
Female
Sex
sex
Male
Female

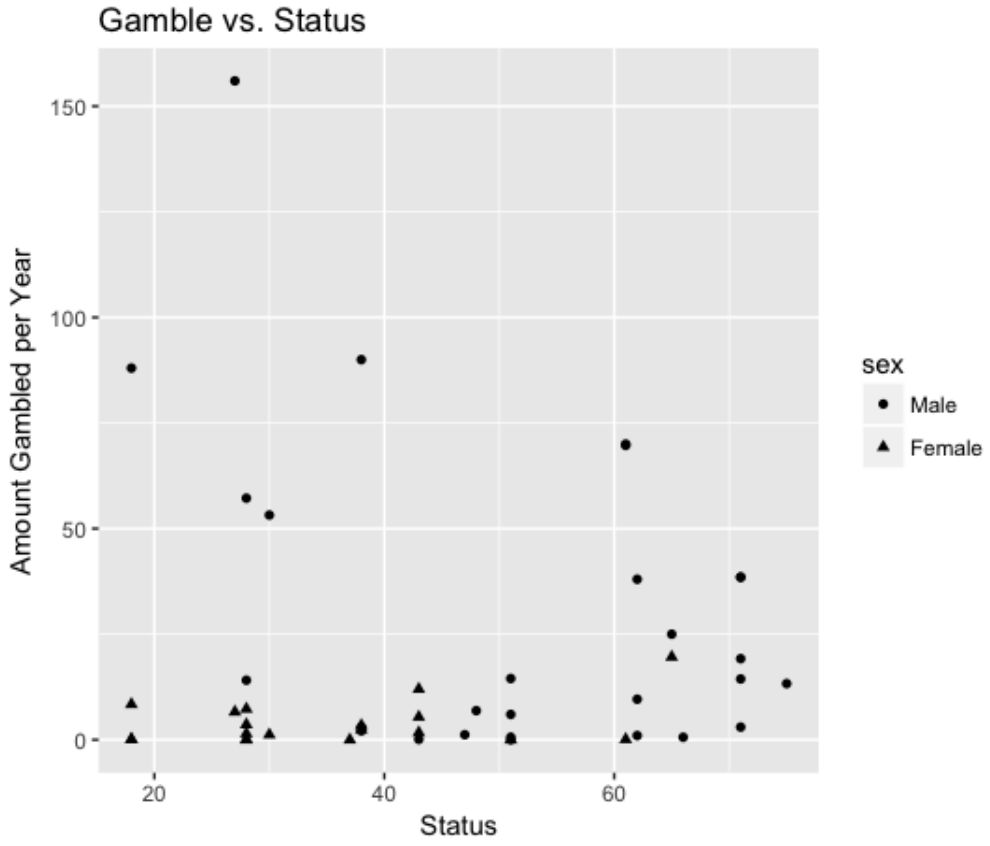

Gamble vs. Income
Amount Gambled per Year
150
100
50
0
0
5
10
15
Income per Week [Pounds]
sex
Male
Female


Income vs. Parental Status
Income [Pounds/Week]
15
10
5
0
20
40
60
Status
sex
Male
Female


Verbal Score vs. Income
Verbal Score
10.0
7.5
5.0
2.5
0
5
10
15
Income [Pounds/Week]
sex
Male
Female


Verbal Score vs. Stauts
Verbal Score
10.0
7.5
5.0
2.5
20
40
60
Status
sex
Male
Female


Gamble vs. Verbal
Amount Gambled per Year
150
100
50
0
2.5
5.0
7.5
10.0
Verbal Score
sex
Male
Female


Gamble vs. Status
Amount Gambled per Year
150
100
50
0
20
40
60
Status
sex
Male
Female