Lab 0

Priya Shaji

February 1, 2019

Analyzing Arbuthnot's dataset using basic R commands

Load the dataset

```
source("more/arbuthnot.R")
```

View the dataset

```
arbuthnot
```

```
##    year boys girls
## 1  1629 5218  4683
## 2  1630 4858  4457
## 3  1631 4422  4102
## 4  1632 4994  4590
## 5  1633 5158  4839
## 6  1634 5035  4820
## 7  1635 5106  4928
## 8  1636 4917  4605
## 9  1637 4703  4457
## 10 1638 5359  4952
## 11 1639 5366  4784
## 12 1640 5518  5332
## 13 1641 5470  5200
## 14 1642 5460  4910
## 15 1643 4793  4617
## 16 1644 4107  3997
## 17 1645 4047  3919
## 18 1646 3768  3395
## 19 1647 3796  3536
## 20 1648 3363  3181
## 21 1649 3079  2746
## 22 1650 2890  2722
## 23 1651 3231  2840
## 24 1652 3220  2908
## 25 1653 3196  2959
## 26 1654 3441  3179
## 27 1655 3655  3349
## 28 1656 3668  3382
## 29 1657 3396  3289
## 30 1658 3157  3013
## 31 1659 3209  2781
```

```
## 32 1660 3724  3247
## 33 1661 4748  4107
## 34 1662 5216  4803
## 35 1663 5411  4881
## 36 1664 6041  5681
## 37 1665 5114  4858
## 38 1666 4678  4319
## 39 1667 5616  5322
## 40 1668 6073  5560
## 41 1669 6506  5829
## 42 1670 6278  5719
## 43 1671 6449  6061
## 44 1672 6443  6120
## 45 1673 6073  5822
## 46 1674 6113  5738
## 47 1675 6058  5717
## 48 1676 6552  5847
## 49 1677 6423  6203
## 50 1678 6568  6033
## 51 1679 6247  6041
## 52 1680 6548  6299
## 53 1681 6822  6533
## 54 1682 6909  6744
## 55 1683 7577  7158
## 56 1684 7575  7127
## 57 1685 7484  7246
## 58 1686 7575  7119
## 59 1687 7737  7214
## 60 1688 7487  7101
## 61 1689 7604  7167
## 62 1690 7909  7302
## 63 1691 7662  7392
## 64 1692 7602  7316
## 65 1693 7676  7483
## 66 1694 6985  6647
## 67 1695 7263  6713
## 68 1696 7632  7229
## 69 1697 8062  7767
## 70 1698 8426  7626
## 71 1699 7911  7452
## 72 1700 7578  7061
## 73 1701 8102  7514
## 74 1702 8031  7656
## 75 1703 7765  7683
## 76 1704 6113  5738
## 77 1705 8366  7779
## 78 1706 7952  7417
## 79 1707 8379  7687
## 80 1708 8239  7623
```

```
## 81 1709 7840  7380
## 82 1710 7640  7288
```

Dimensions of the dataset

```
dim(arbuthnot)

## [1] 82  3
```

Names of columns of the Arbuthnot Dataset

```
names(arbuthnot)

## [1] "year"  "boys"  "girls"
```

Number of boys baptized each year

```
arbuthnot$boys

##  [1] 5218 4858 4422 4994 5158 5035 5106 4917 4703 5359 5366 5518 5470 5460
## [15] 4793 4107 4047 3768 3796 3363 3079 2890 3231 3220 3196 3441 3655 3668
## [29] 3396 3157 3209 3724 4748 5216 5411 6041 5114 4678 5616 6073 6506 6278
## [43] 6449 6443 6073 6113 6058 6552 6423 6568 6247 6548 6822 6909 7577 7575
## [57] 7484 7575 7737 7487 7604 7909 7662 7602 7676 6985 7263 7632 8062 8426
## [71] 7911 7578 8102 8031 7765 6113 8366 7952 8379 8239 7840 7640
```

EXERCISE 1

What command would you use to extract just the counts of girls baptized?

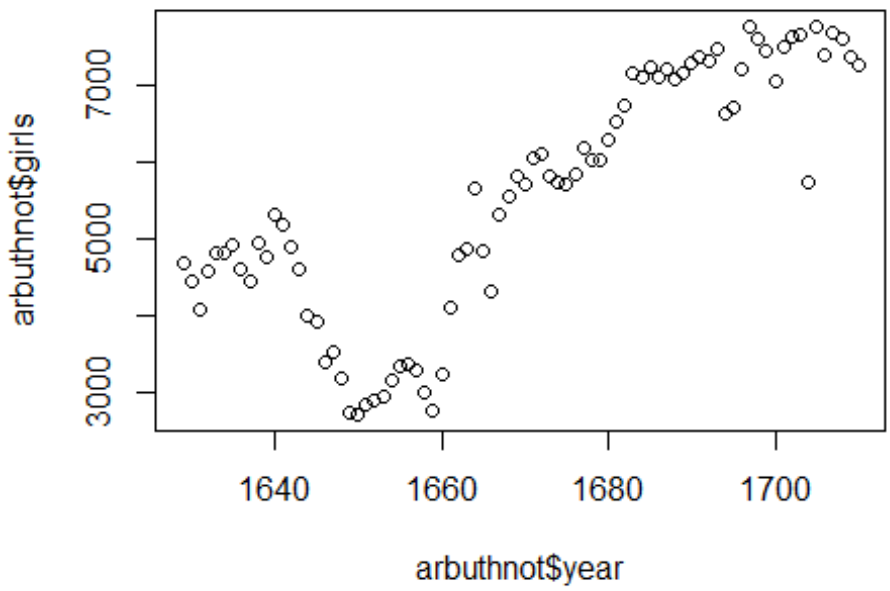
Answer 1:

```
arbuthnot$girls

##  [1] 4683 4457 4102 4590 4839 4820 4928 4605 4457 4952 4784 5332 5200 4910
## [15] 4617 3997 3919 3395 3536 3181 2746 2722 2840 2908 2959 3179 3349 3382
## [29] 3289 3013 2781 3247 4107 4803 4881 5681 4858 4319 5322 5560 5829 5719
## [43] 6061 6120 5822 5738 5717 5847 6203 6033 6041 6299 6533 6744 7158 7127
## [57] 7246 7119 7214 7101 7167 7302 7392 7316 7483 6647 6713 7229 7767 7626
## [71] 7452 7061 7514 7656 7683 5738 7779 7417 7687 7623 7380 7288
```

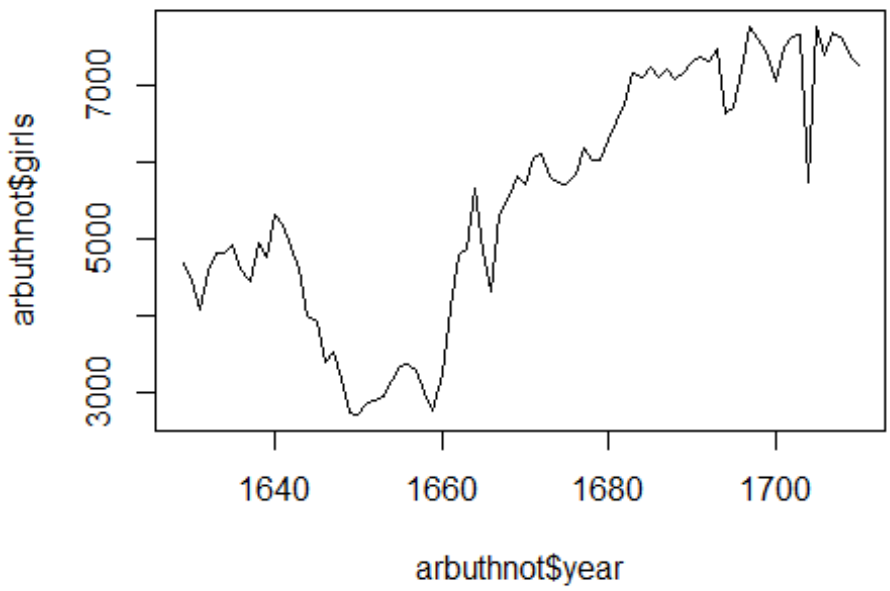
Create a simple plot of the number of girls baptized per year with the command

```
plot(x = arbuthnot$year, y = arbuthnot$girls)
```



The first argument in the plot function specifies the variable for the x-axis and the second for the y-axis. If we wanted to connect the data points with lines, we could add a third argument, the letter l for line.

```
plot(x = arbuthnot$year, y = arbuthnot$girls, type = "l")
```



EXERCISE 2

Is there an apparent trend in the number of girls baptized over the years? How would you describe it?

ANSWER 2: By analyzing the plot which shows the number of girls born each year, there is a gradual increase in the count of girls born from yea 1660 to 1700

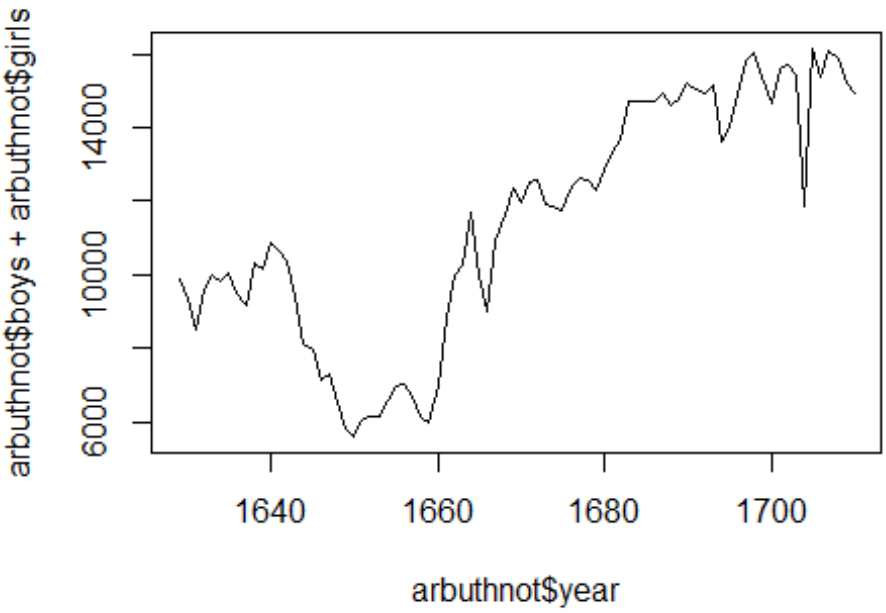
Add the vector for baptisms for boys and girls

```
arbuthnot$boys + arbuthnot$girls
```

```
##  [1]  9901  9315  8524  9584  9997  9855 10034  9522  9160 10311 10150
## [12] 10850 10670 10370  9410  8104  7966  7163  7332  6544  5825  5612
## [23]  6071  6128  6155  6620  7004  7050  6685  6170  5990  6971  8855
## [34] 10019 10292 11722  9972  8997 10938 11633 12335 11997 12510 12563
## [45] 11895 11851 11775 12399 12626 12601 12288 12847 13355 13653 14735
## [56] 14702 14730 14694 14951 14588 14771 15211 15054 14918 15159 13632
## [67] 13976 14861 15829 16052 15363 14639 15616 15687 15448 11851 16145
## [78] 15369 16066 15862 15220 14928
```

Plot of the total number of baptisms per year with the command

```
plot(arbuthnot$year, arbuthnot$boys + arbuthnot$girls, type = "l")
```



The proportion of newborns that are boys

```
arbuthnot$boys / (arbuthnot$boys + arbuthnot$girls)
```

```
##  [1] 0.5270175 0.5215244 0.5187705 0.5210768 0.5159548 0.5109082 0.5088698
##  [8] 0.5163831 0.5134279 0.5197362 0.5286700 0.5085714 0.5126523 0.5265188
## [15] 0.5093518 0.5067868 0.5080341 0.5260366 0.5177305 0.5139059 0.5285837
## [22] 0.5149679 0.5322023 0.5254569 0.5192526 0.5197885 0.5218447 0.5202837
## [29] 0.5080030 0.5116694 0.5357262 0.5342132 0.5361942 0.5206108 0.5257482
## [36] 0.5153557 0.5128359 0.5199511 0.5134394 0.5220493 0.5274422 0.5232975
## [43] 0.5155076 0.5128552 0.5105507 0.5158214 0.5144798 0.5284297 0.5087122
## [50] 0.5212285 0.5083822 0.5096910 0.5108199 0.5060426 0.5142178 0.5152360
## [57] 0.5080788 0.5155165 0.5174905 0.5132301 0.5147925 0.5199527 0.5089677
## [64] 0.5095857 0.5063659 0.5123973 0.5196766 0.5135590 0.5093183 0.5249190
## [71] 0.5149385 0.5176583 0.5188268 0.5119526 0.5026541 0.5158214 0.5181790
## [78] 0.5174052 0.5215362 0.5194175 0.5151117 0.5117899
```

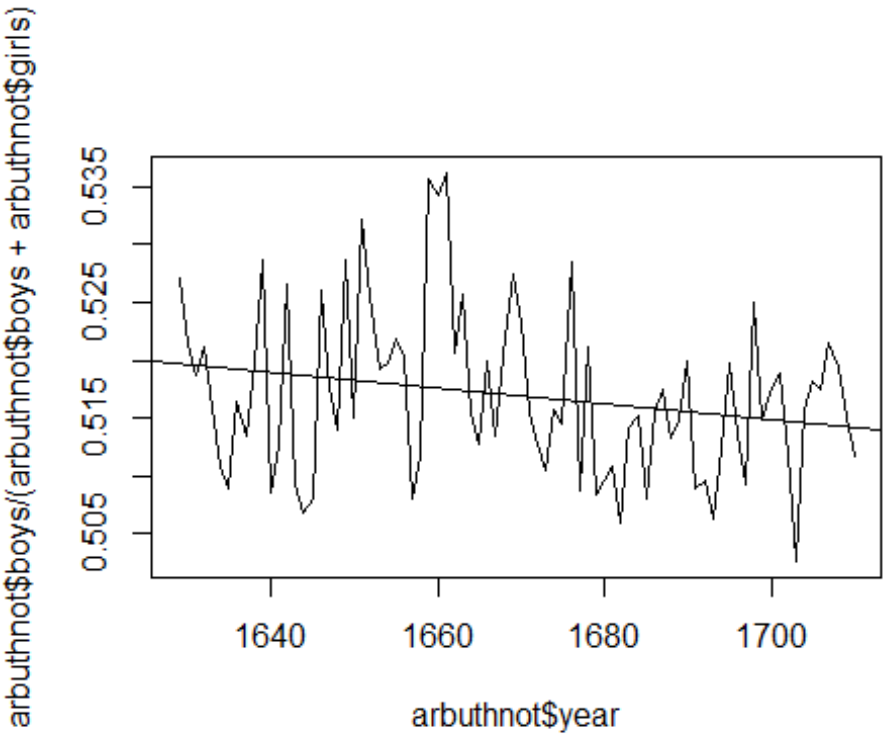
EXERCISE 3

Answer 3:

plot of the proportion of boys over time create a regression line to analyze the proportion over time

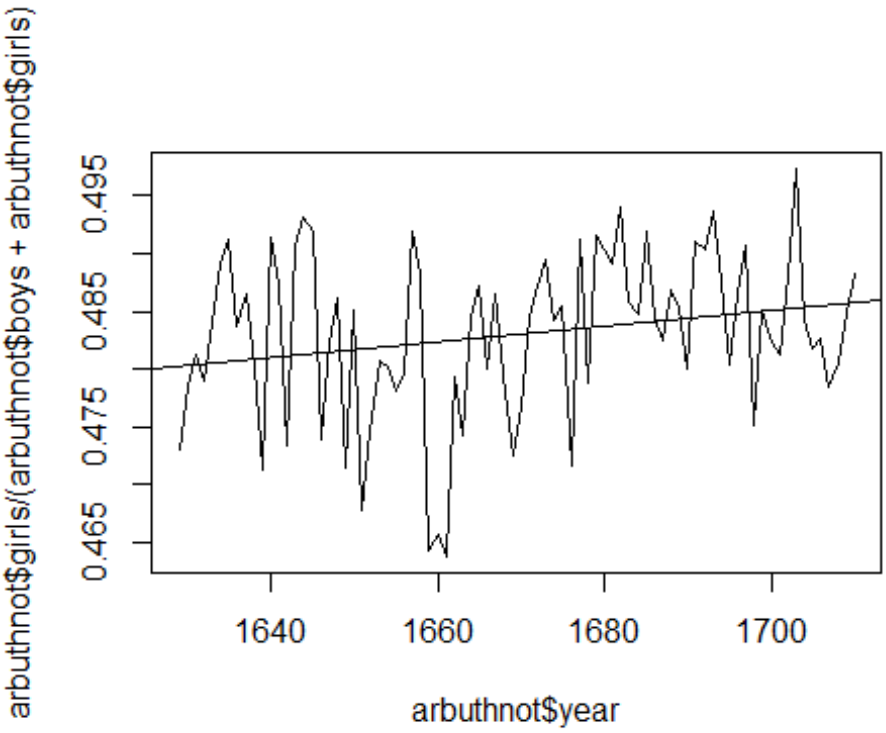
```
plot(arbuthnot$boys / (arbuthnot$boys + arbuthnot$girls)~arbuthnot$year, type
= "l",data=arbuthnot)
```

```
abline(lm(arbuthnot$boys / (arbuthnot$boys + arbuthnot$girls) ~
arbuthnot$year,data=arbuthnot))
```



plot of the proportion of girls over time create a regression line to analyze the proportion over time

```
plot(arbuthnot$girls / (arbuthnot$boys + arbuthnot$girls)~arbuthnot$year,
type = "l")
abline(lm(arbuthnot$girls / (arbuthnot$boys + arbuthnot$girls) ~
arbuthnot$year))
```



We infer that, the regression line of proportion of boys over time decreases over the years and is above 0.5 and the regression line of proportion of girls over time increases over years and is below 0.5. Therefore, the number of boys increases over the years as compared to number of girls.

Do boys outnumber girls in each year

```
arbuthnot$boys > arbuthnot$girls
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [71] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE

## yes
```

On Your Own

Analyzing present day birth records dataset in United States using basic R commands

1) Load the dataset

```
source("more/present.R")
```

2) View the dataset

```
present
```

```
##    year    boys   girls
## 1  1940 1211684 1148715
## 2  1941 1289734 1223693
## 3  1942 1444365 1364631
## 4  1943 1508959 1427901
## 5  1944 1435301 1359499
## 6  1945 1404587 1330869
## 7  1946 1691220 1597452
## 8  1947 1899876 1800064
## 9  1948 1813852 1721216
## 10 1949 1826352 1733177
## 11 1950 1823555 1730594
## 12 1951 1923020 1827830
## 13 1952 1971262 1875724
## 14 1953 2001798 1900322
## 15 1954 2059068 1958294
## 16 1955 2073719 1973576
## 17 1956 2133588 2029502
## 18 1957 2179960 2074824
## 19 1958 2152546 2051266
## 20 1959 2173638 2071158
## 21 1960 2179708 2078142
## 22 1961 2186274 2082052
## 23 1962 2132466 2034896
## 24 1963 2101632 1996388
## 25 1964 2060162 1967328
## 26 1965 1927054 1833304
## 27 1966 1845862 1760412
## 28 1967 1803388 1717571
## 29 1968 1796326 1705238
## 30 1969 1846572 1753634
## 31 1970 1915378 1816008
## 32 1971 1822910 1733060
## 33 1972 1669927 1588484
## 34 1973 1608326 1528639
## 35 1974 1622114 1537844
## 36 1975 1613135 1531063
## 37 1976 1624436 1543352
## 38 1977 1705916 1620716
## 39 1978 1709394 1623885
## 40 1979 1791267 1703131
## 41 1980 1852616 1759642
## 42 1981 1860272 1768966
## 43 1982 1885676 1794861
## 44 1983 1865553 1773380
## 45 1984 1879490 1789651
## 46 1985 1927983 1832578
## 47 1986 1924868 1831679
```

```
## 48 1987 1951153 1858241
## 49 1988 2002424 1907086
## 50 1989 2069490 1971468
## 51 1990 2129495 2028717
## 52 1991 2101518 2009389
## 53 1992 2082097 1982917
## 54 1993 2048861 1951379
## 55 1994 2022589 1930178
## 56 1995 1996355 1903234
## 57 1996 1990480 1901014
## 58 1997 1985596 1895298
## 59 1998 2016205 1925348
## 60 1999 2026854 1932563
## 61 2000 2076969 1981845
## 62 2001 2057922 1968011
## 63 2002 2057979 1963747
```

3)Dimensions of the dataset

```
dim(present)

## [1] 63  3
```

4)Names of columns of the present Dataset

```
names(present)

## [1] "year"  "boys"  "girls"
```

Number of boys born each year

```
present$boys

##  [1] 1211684 1289734 1444365 1508959 1435301 1404587 1691220 1899876
##  [9] 1813852 1826352 1823555 1923020 1971262 2001798 2059068 2073719
## [17] 2133588 2179960 2152546 2173638 2179708 2186274 2132466 2101632
## [25] 2060162 1927054 1845862 1803388 1796326 1846572 1915378 1822910
## [33] 1669927 1608326 1622114 1613135 1624436 1705916 1709394 1791267
## [41] 1852616 1860272 1885676 1865553 1879490 1927983 1924868 1951153
## [49] 2002424 2069490 2129495 2101518 2082097 2048861 2022589 1996355
## [57] 1990480 1985596 2016205 2026854 2076969 2057922 2057979
```

EXERCISE 1

What command would you use to extract just the counts of girls born?

Answer 1:

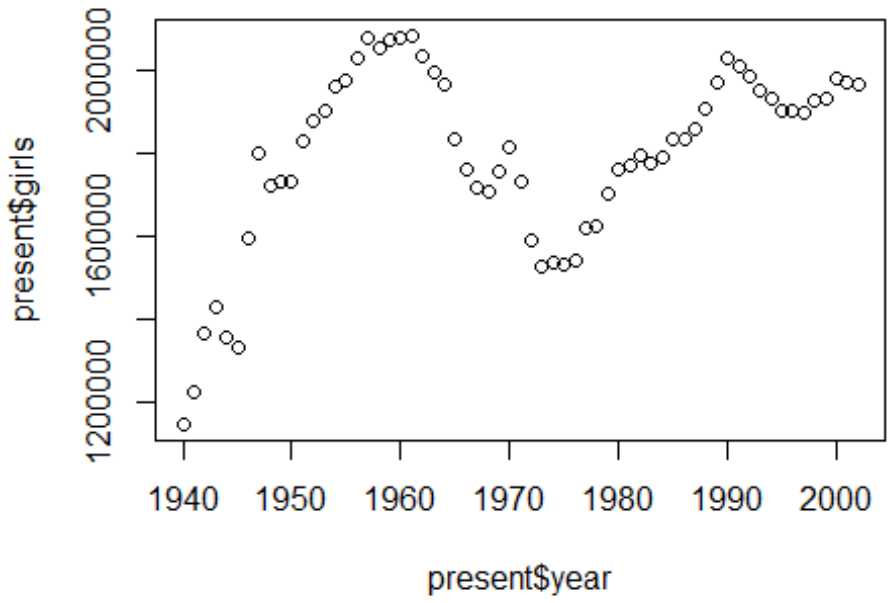
```
present$girls

##  [1] 1148715 1223693 1364631 1427901 1359499 1330869 1597452 1800064
##  [9] 1721216 1733177 1730594 1827830 1875724 1900322 1958294 1973576
## [17] 2029502 2074824 2051266 2071158 2078142 2082052 2034896 1996388
```

```
## [25] 1967328 1833304 1760412 1717571 1705238 1753634 1816008 1733060
## [33] 1588484 1528639 1537844 1531063 1543352 1620716 1623885 1703131
## [41] 1759642 1768966 1794861 1773380 1789651 1832578 1831679 1858241
## [49] 1907086 1971468 2028717 2009389 1982917 1951379 1930178 1903234
## [57] 1901014 1895298 1925348 1932563 1981845 1968011 1963747
```

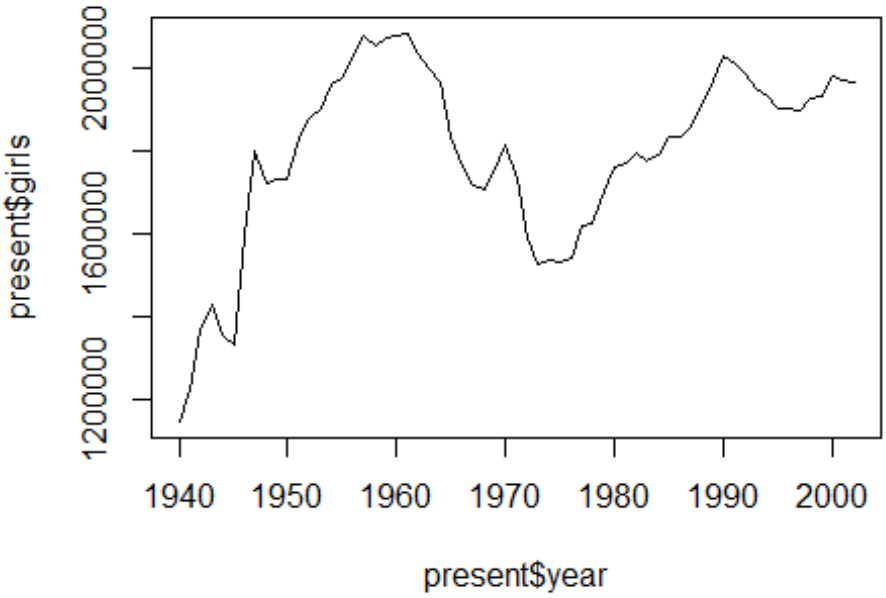
Create a simple plot of the number of girls born per year with the command

```
plot(x = present$year, y = present$girls)
```



The first argument in the plot function specifies the variable for the x-axis and the second for the y-axis. If we wanted to connect the data points with lines, we could add a third argument, the letter l for line.

```
plot(x = present$year, y = present$girls, type = "l")
```



EXERCISE 2

Is there an apparent trend in the number of girls baptized over the years? How would you describe it?

ANSWER 2: By analyzing the plot which shows the number of girls born each year, there is a gradual increase in the count of girls born from yea 1660 to 1700

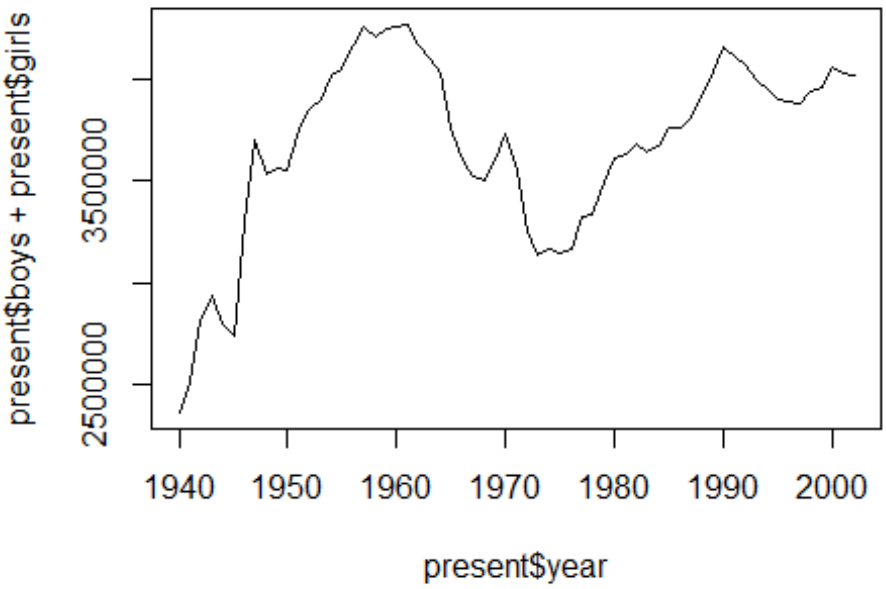
Add the vector for birth for boys and girls

```
present$boys + present$girls
```

```
##  [1] 2360399 2513427 2808996 2936860 2794800 2735456 3288672 3699940
##  [9] 3535068 3559529 3554149 3750850 3846986 3902120 4017362 4047295
## [17] 4163090 4254784 4203812 4244796 4257850 4268326 4167362 4098020
## [25] 4027490 3760358 3606274 3520959 3501564 3600206 3731386 3555970
## [33] 3258411 3136965 3159958 3144198 3167788 3326632 3333279 3494398
## [41] 3612258 3629238 3680537 3638933 3669141 3760561 3756547 3809394
## [49] 3909510 4040958 4158212 4110907 4065014 4000240 3952767 3899589
## [57] 3891494 3880894 3941553 3959417 4058814 4025933 4021726
```

Plot of the total number of births per year with the command

```
plot(present$year, present$boys + present$girls, type = "l")
```



The proportion of newborns that are boys

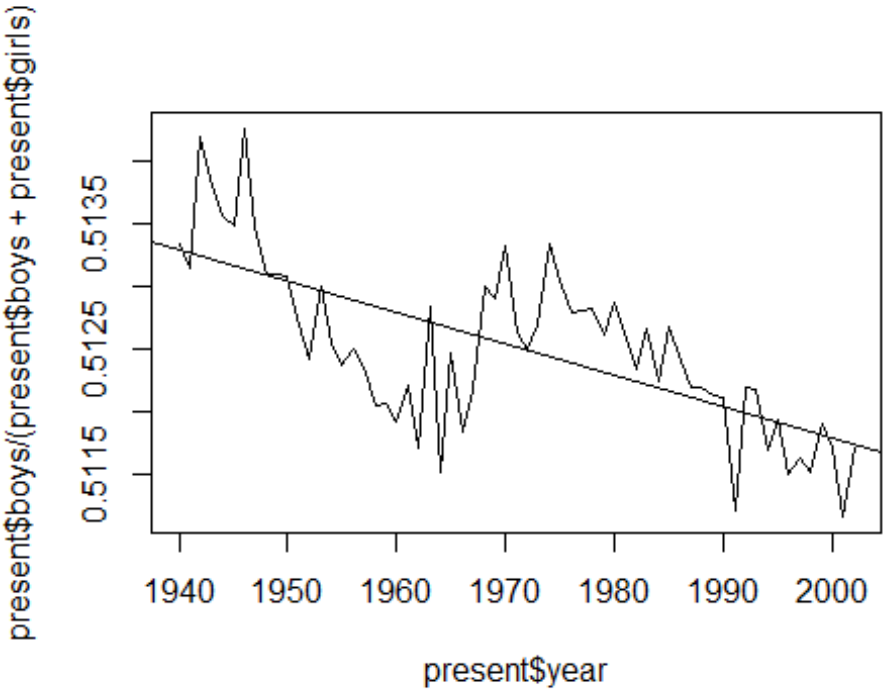
```
present$boys / (present$boys + present$girls)
```

```
##  [1] 0.5133386 0.5131376 0.5141926 0.5138001 0.5135613 0.5134745 0.5142562
##  [8] 0.5134883 0.5131024 0.5130881 0.5130778 0.5126891 0.5124173 0.5130027
## [15] 0.5125423 0.5123716 0.5125011 0.5123550 0.5120462 0.5120713 0.5119269
## [22] 0.5122088 0.5117064 0.5128408 0.5115250 0.5124656 0.5118474 0.5121866
## [29] 0.5130068 0.5129073 0.5133154 0.5126337 0.5124973 0.5127013 0.5133340
## [36] 0.5130513 0.5127982 0.5128057 0.5128266 0.5126110 0.5128692 0.5125792
## [43] 0.5123372 0.5126648 0.5122425 0.5126849 0.5124035 0.5121951 0.5121931
## [50] 0.5121286 0.5121179 0.5112054 0.5121992 0.5121845 0.5116894 0.5119398
## [57] 0.5114951 0.5116337 0.5115255 0.5119072 0.5117182 0.5111665 0.5117154
```

EXERCISE 3

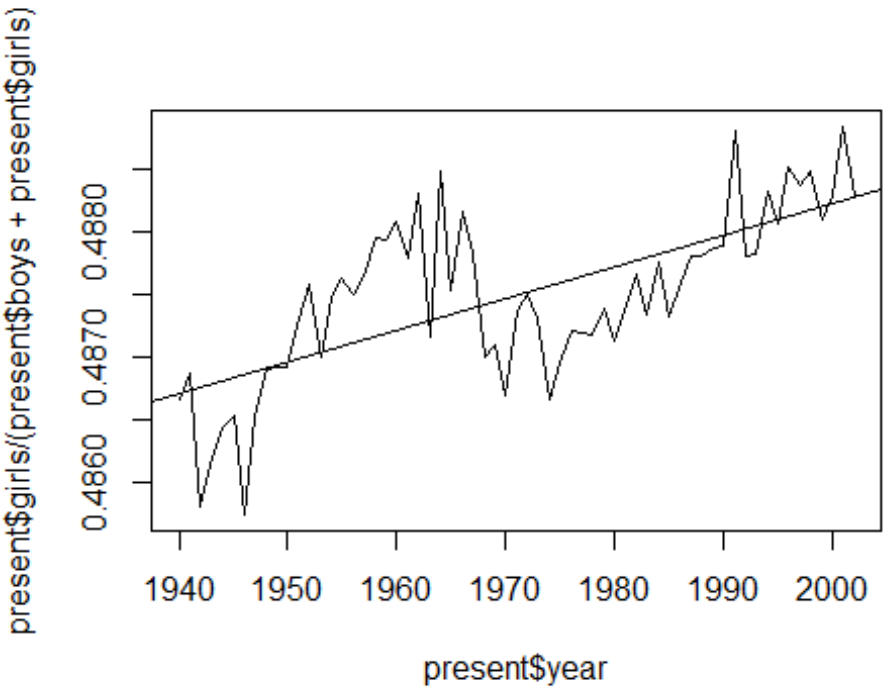
Answer 3: plot of the proportion of boys over time create a regression line to analyze the proportion over time

```
plot( present$boys / (present$boys + present$girls)~present$year, type = "l")
abline(lm(present$boys / (present$boys + present$girls) ~ present$year))
```



plot of the proportion of girls over time create a regression line to analyze the proportion over time

```
plot(present$girls / (present$boys + present$girls)~present$year, type = "l")
abline(lm(present$girls / (present$boys + present$girls) ~ present$year))
```



We infer that, the regression line of proportion of boys over time decreases over the years and is above 0.5 and the regression line of proportion of girls over time increases over years and is below 0.5. Therefore, the number of boys increases over the years as compared to number of girls.

Do boys outnumber girls in each year ?

```
present$boys > present$girls
```

```
##  [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE
## [57] TRUE TRUE TRUE TRUE TRUE TRUE TRUE

## yes
```

Questions

1) What years are included in this data set? What are the dimensions of the data frame and what are the variable or column names?

```
head(present)
```

```
##   year    boys   girls
## 1 1940 1211684 1148715
## 2 1941 1289734 1223693
## 3 1942 1444365 1364631
## 4 1943 1508959 1427901
```

```
## 5 1944 1435301 1359499
## 6 1945 1404587 1330869
```

```
tail(present)
```

```
##    year    boys   girls
## 58 1997 1985596 1895298
## 59 1998 2016205 1925348
## 60 1999 2026854 1932563
## 61 2000 2076969 1981845
## 62 2001 2057922 1968011
## 63 2002 2057979 1963747
```

```
dim(present)
```

```
## [1] 63  3
```

```
names(present)
```

```
## [1] "year"  "boys"  "girls"
```

Answer 1) The years included in this dataset are: 1940 to 2002

2) How do these counts compare to Arbuthnot's? Are they on a similar scale?

Answer2) Arbuthnot and present datasets are similar for the following cases:
Both have same no. of columns and same column names Both datasets are analyzing the birth of boys and girls over time.

Arbuthnot and present datasets differ in their counts: When we calculate the mean their is a difference in the counts of both the datasets.

arbuthnot's mean

```
mean(arbuthnot$boys + arbuthnot$girls)
```

```
## [1] 11441.74
```

present's mean

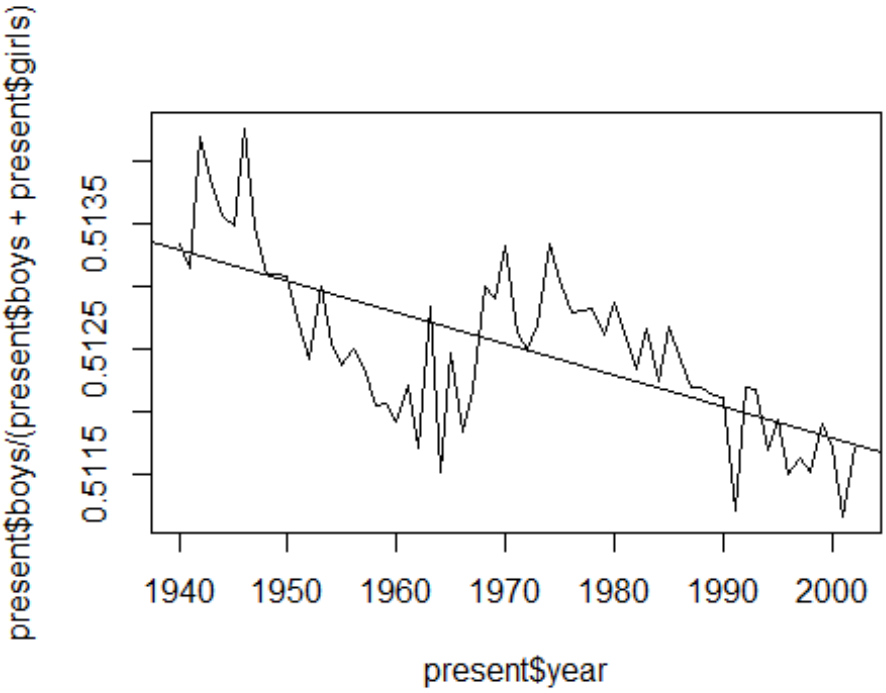
```
mean(present$boys + present$girls)
```

```
## [1] 3679515
```

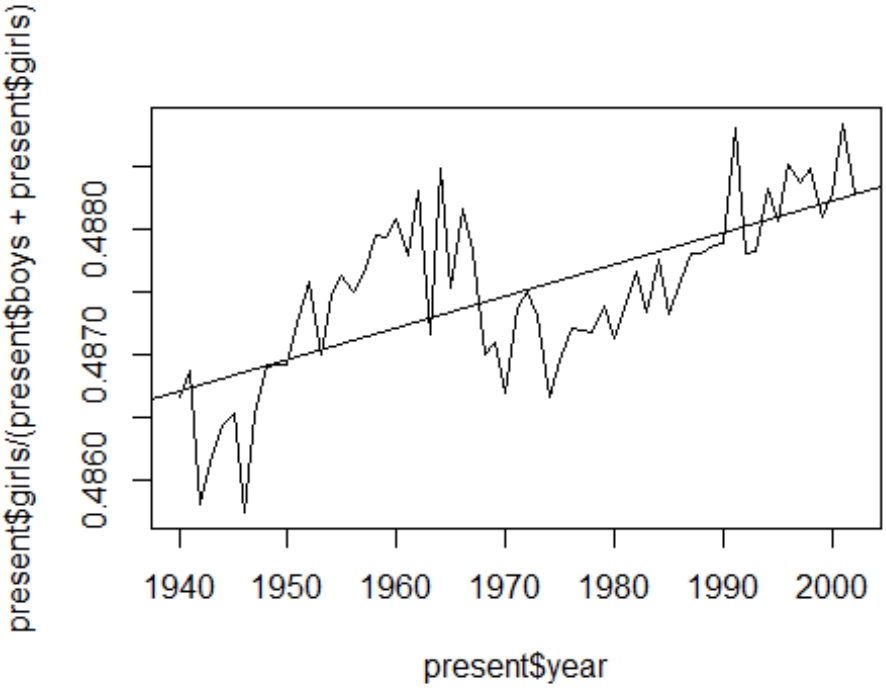
3) Make a plot that displays the boy-to-girl ratio for every year in the data set. What do you see? Does Arbuthnot's observation about boys being born in greater proportion than girls hold up in the U.S.? Include the plot in your response.

Answer 3)

```
plot(present$year, present$boys / (present$boys + present$girls), type = "l")
abline(lm(present$boys / (present$boys + present$girls) ~ present$year))
```




```
plot(present$year, present$girls / (present$boys + present$girls), type =
"l")
abline(lm(present$girls / (present$boys + present$girls) ~ present$year))
```



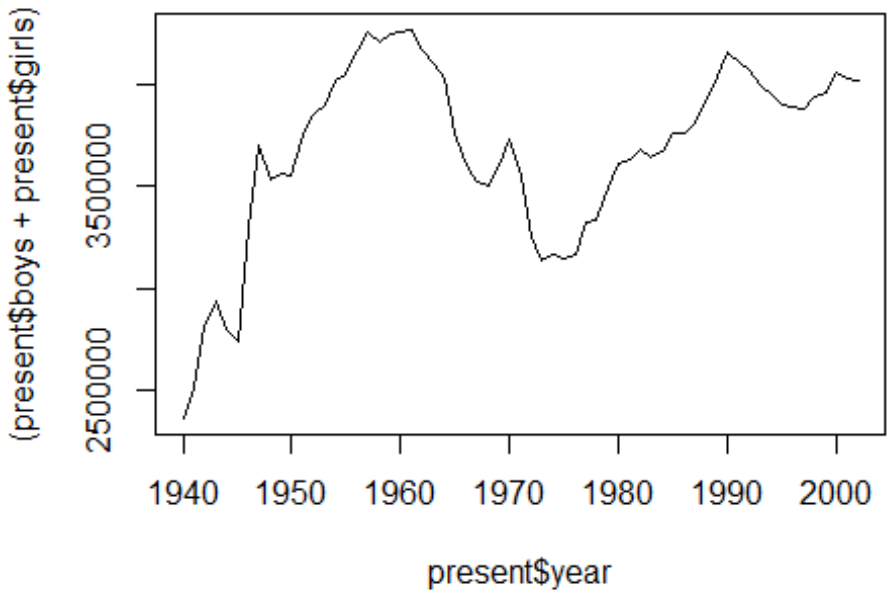
We infer that, the regression line of proportion of boys over time decreases over the years and is above 0.5 and the regression line of proportion of girls over time increases over years and is below 0.5. Therefore, the number of boys increases over the years as compared to number of girls.

Therefore Arbuthnot's observation about boys being born in greater proportion than girls hold up in the U.S.

4) In what year did we see the most total number of births in the U.S.?

Answer 4) Analyzing using plot

```
plot(present$year, (present$boys + present$girls), type = "l")
```



Analyzing using calculation, to be more precise

```
present$year[(present$boys + present$girls) == max(present$boys +
present$girls)]

## [1] 1961
```

Therefore, the year 1961 , it's the most total number of births in the U.S