DA 460 - Assignment 0

Marjorie Blanco

08 July, 2018

# Part 1 R

## The Data: Dr. Arbuthnot’s Baptism Records

# View data  
arbuthnot

# View dimension  
dim(arbuthnot)

## [1] 82 3

# View names of these columns (or variables)  
names(arbuthnot)

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

## Some Exploration

# View single column of a data  
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

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

## Problem 1

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

There are 82 records for girls baptized in the arbuthnot data set. The total count of girls baptized is 453841.

# extract just the counts of girls baptized  
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

#total number of girls baptized  
sum(arbuthnot$girls)

## [1] 453841

# records of girls baptized  
length(arbuthnot$girls)

## [1] 82

dim(arbuthnot)[1]

## [1] 82

## Problem 2

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

There is an apparent trend in the number of girls baptized over the years. There is steady increase of girls baptized starting in 1660. The trend can be described as polynomial.

# summary stats  
summary(arbuthnot$girls)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2722 4457 5718 5535 7150 7779

summary(arbuthnot$boys)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2890 4759 6073 5907 7576 8426

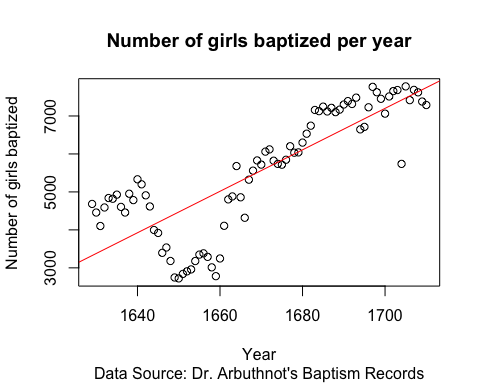
# observation with max girls births  
arbuthnot[arbuthnot$girls==max(arbuthnot$girls), ]

## year boys girls  
## 77 1705 8366 7779

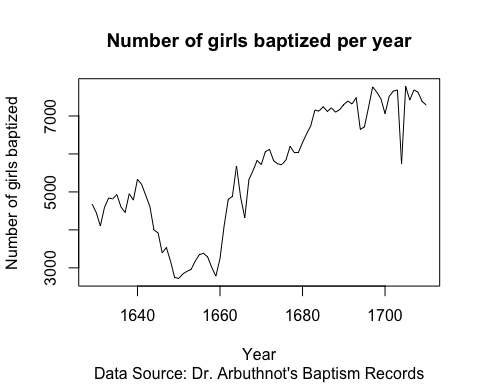
# observation with min girls births  
arbuthnot[arbuthnot$girls==min(arbuthnot$girls), ]

## year boys girls  
## 22 1650 2890 2722

# create a simple plot of the number of girls baptized per year  
plot(x = arbuthnot$year, y = arbuthnot$girls,   
 main="Number of girls baptized per year",  
 xlab="Year", ylab="Number of girls baptized", sub="Data Source: Dr. Arbuthnot's Baptism Records")   
abline(lm(arbuthnot$girls ~ arbuthnot$year), col='red')

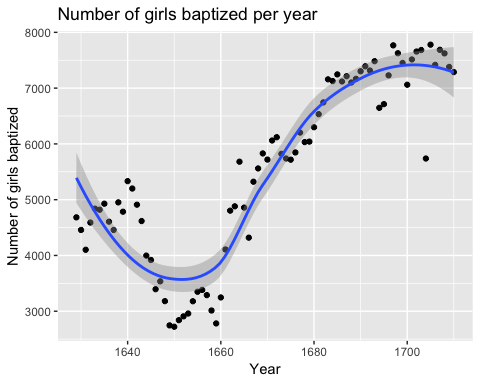


# create a simple line plot of the number of girls baptized per year  
plot(x = arbuthnot$year, y = arbuthnot$girls, type="l",   
 main="Number of girls baptized per year",  
 xlab="Year", ylab="Number of girls baptized", sub="Data Source: Dr. Arbuthnot's Baptism Records")



arbplot = data.frame(arbuthnot$year, arbuthnot$girls)  
ggplot(arbplot, aes(arbuthnot$year,arbuthnot$girls)) +   
 geom\_point() +   
 geom\_smooth() +  
 ggtitle("Number of girls baptized per year") +  
 xlab("Year") + ylab("Number of girls baptized")

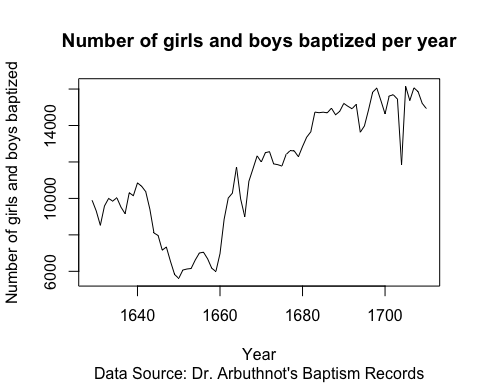
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



# mathematical expressions  
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  
arbuthnot$total = arbuthnot$boys + arbuthnot$girls  
plot(x = arbuthnot$year, y = arbuthnot$total, type="l",   
 main="Number of girls and boys baptized per year",  
 xlab="Year", ylab="Number of girls and boys baptized", sub="Data Source: Dr. Arbuthnot's Baptism Records")



# compute the ratio of the number of boys to the number of girls baptized in 1629 (expression)  
arbuthnot$boys / arbuthnot$girls

## [1] 1.114243 1.089971 1.078011 1.088017 1.065923 1.044606 1.036120  
## [8] 1.067752 1.055194 1.082189 1.121656 1.034884 1.051923 1.112016  
## [15] 1.038120 1.027521 1.032661 1.109867 1.073529 1.057215 1.121267  
## [22] 1.061719 1.137676 1.107290 1.080095 1.082416 1.091371 1.084565  
## [29] 1.032533 1.047793 1.153901 1.146905 1.156075 1.085988 1.108584  
## [36] 1.063369 1.052697 1.083121 1.055242 1.092266 1.116143 1.097744  
## [43] 1.064016 1.052778 1.043112 1.065354 1.059647 1.120575 1.035467  
## [50] 1.088679 1.034100 1.039530 1.044237 1.024466 1.058536 1.062860  
## [57] 1.032846 1.064054 1.072498 1.054359 1.060974 1.083128 1.036526  
## [64] 1.039092 1.025792 1.050850 1.081931 1.055748 1.037981 1.104904  
## [71] 1.061594 1.073219 1.078254 1.048981 1.010673 1.065354 1.075460  
## [78] 1.072132 1.090022 1.080808 1.062331 1.048299

# computed for all years simultaneously  
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

# check if boys outnumber girls in each year with  
arbuthnot$boys > arbuthnot$girls

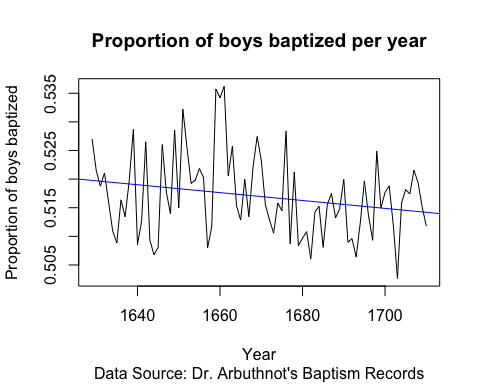
## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE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

## Problem 3

Now, make a plot of the proportion of boys over time. What do you see?

The proportion of boys that are baptised over time is betweeen 50.27% and 53.62%. This is a time-series plot.

plot(x = arbuthnot$year, y = arbuthnot$boys / (arbuthnot$boys + arbuthnot$girls), type="l",   
 main="Proportion of boys baptized per year",  
 xlab="Year", ylab="Proportion of boys baptized", sub="Data Source: Dr. Arbuthnot's Baptism Records")  
abline(lm((arbuthnot$boys/(arbuthnot$boys + arbuthnot$girls)) ~ arbuthnot$year), col='blue')



# Part 2 R

## The Data: Present Baptism Records

# View data  
present

# View dimension  
dim(present)

## [1] 63 3

# View names of these columns (or variables)  
names(present)

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

## Some Exploration

# View single column of a data  
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

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

## Problem 1

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

There are 63 records for girls baptized in the present data set. The total count of girls baptized is 1.130166510^{8}.

# extract just the counts of girls baptized  
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

#total number of girls baptized  
sum(present$girls)

## [1] 113016646

# records of girls baptized  
length(present$girls)

## [1] 63

dim(present)[1]

## [1] 63

## Problem 2

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

There is an apparent trend in the number of girls baptized over the years. There is steady increase of girls baptized from 1949 to 1969 and from 1970 to 2002. The trend can be described as polynomial.

# summary stats  
summary(present$girls)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1148715 1711404 1831679 1793915 1965538 2082052

summary(present$boys)

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1211684 1799857 1924868 1885600 2058524 2186274

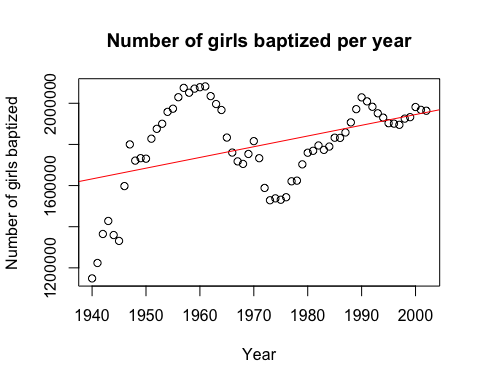
# observation with max girls births  
arbuthnot[present$girls==max(arbuthnot$girls), ]

## [1] year boys girls total  
## <0 rows> (or 0-length row.names)

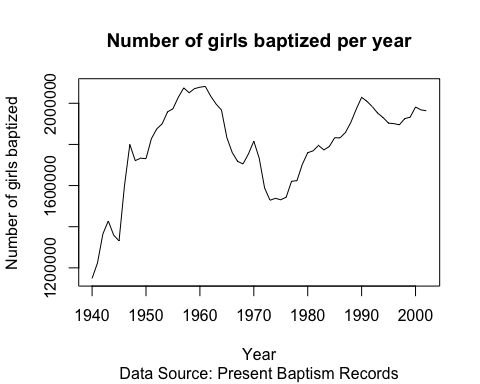
# observation with min girls births  
arbuthnot[present$girls==min(arbuthnot$girls), ]

## [1] year boys girls total  
## <0 rows> (or 0-length row.names)

# create a simple plot of the number of girls baptized per year  
plot(x = present$year, y = present$girls,   
 main="Number of girls baptized per year",  
 xlab="Year", ylab="Number of girls baptized")   
abline(lm(present$girls ~ present$year), col='red', sub="Data Source: Present Baptism Records")

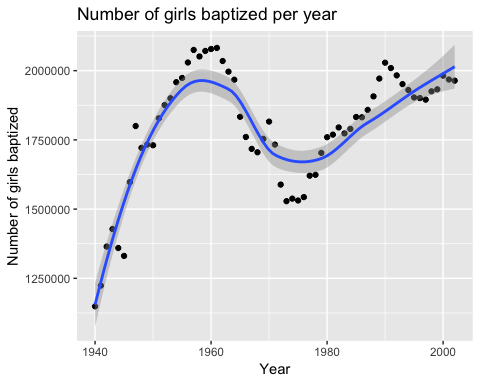


# create a simple line plot of the number of girls baptized per year  
plot(x = present$year, y = present$girls, type="l",   
 main="Number of girls baptized per year",  
 xlab="Year", ylab="Number of girls baptized", sub="Data Source: Present Baptism Records")



arbplot = data.frame(present$year, present$girls)  
ggplot(arbplot, aes(present$year, present$girls)) +   
 geom\_point() +   
 geom\_smooth() +  
 ggtitle("Number of girls baptized per year") +  
 xlab("Year") + ylab("Number of girls baptized")

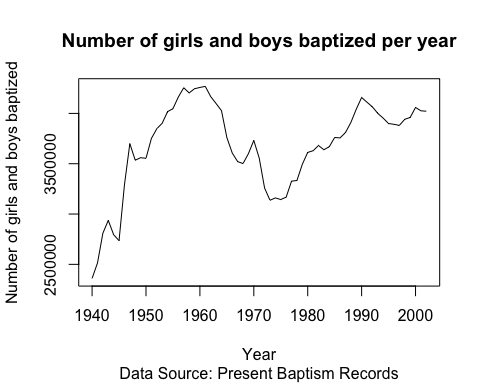
## `geom\_smooth()` using method = 'loess' and formula 'y ~ x'



# mathematical expressions  
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 baptisms per year  
present$total = present$boys + present$girls  
plot(x = present$year, y = present$total, type="l",   
 main="Number of girls and boys baptized per year",  
 xlab="Year", ylab="Number of girls and boys baptized", sub="Data Source: Present Baptism Records")



# compute the ratio of the number of boys to the number of girls baptized in 1629 (expression)  
present$boys / present$girls

## [1] 1.054817 1.053969 1.058429 1.056767 1.055757 1.055391 1.058698  
## [8] 1.055449 1.053820 1.053760 1.053716 1.052078 1.050934 1.053399  
## [15] 1.051460 1.050742 1.051286 1.050672 1.049374 1.049480 1.048873  
## [22] 1.050057 1.047948 1.052717 1.047188 1.051137 1.048540 1.049964  
## [29] 1.053417 1.052997 1.054719 1.051845 1.051271 1.052129 1.054797  
## [36] 1.053605 1.052538 1.052569 1.052657 1.051749 1.052837 1.051615  
## [43] 1.050597 1.051976 1.050199 1.052061 1.050876 1.050000 1.049991  
## [50] 1.049720 1.049676 1.045849 1.050017 1.049955 1.047877 1.048928  
## [57] 1.047062 1.047643 1.047190 1.048791 1.047998 1.045686 1.047986

# computed for all years simultaneously  
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

# check if boys outnumber girls in each year with  
present$boys > present$girls

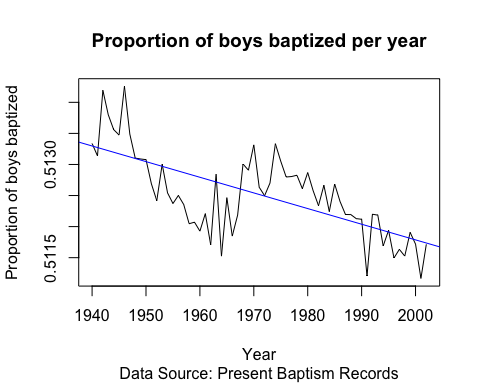
## [1] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [15] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [29] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE  
## [43] TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE TRUE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

## Problem 3

Now, make a plot of the proportion of boys over time. What do you see?

The proportion of boys that are baptised over time is betweeen 51.12% and 51.43%. This is a time-series plot.

plot(x = present$year, y = present$boys / (present$boys + present$girls), type="l",   
 main="Proportion of boys baptized per year",  
 xlab="Year", ylab="Proportion of boys baptized", sub="Data Source: Present Baptism Records")  
abline(lm((present$boys/(present$boys + present$girls)) ~ present$year), col='blue')



# Part 3

## Problem 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?

There are a total of 63 years included in this present data set. The dimensions of the data frame is 63, 4. There are 63 records and 4 variables. The variable names are: year, boys, girls, total. The year range for the present data set is from 1940 to 2002. The year range for the present data set is from 1940 to 2002.

dim(present)

## [1] 63 4

names(present)

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

## Problem 2

How do these counts compare to Arbuthnot’s? Are they on a similar scale?

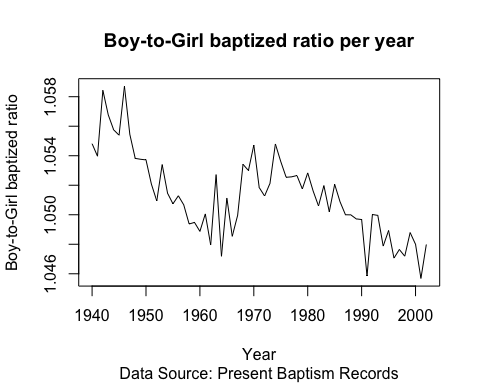
The counts are different for Arbuthnot and Present data set. The scale (year) are different for both data set.

## Problem 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.

There is a steady decrease of boy-to-girl ratio. Arbuthnot’s observation about boys being born in greater proportion than girls in the U.S. seems to hold for the mid 1940s, however even that proportion of (at its highest 1.06) isn’t particularly high and overall it decreases as the years pass.

plot(x = present$year, y = present$boys / present$girls, type="l",   
 main="Boy-to-Girl baptized ratio per year",  
 xlab="Year", ylab="Boy-to-Girl baptized ratio", sub="Data Source: Present Baptism Records")



## Problem 4

In what year did we see the most total number of births in the U.S.? You can refer to the help files or the R reference card <http://cran.r-project.org/doc/contrib/Short-> refcard.pdf to find helpful commands.

# calculate total baptisms per year  
present$total <- present$boys + present$girls  
present[which.max( present[,"total"]), "year"]

## [1] 1961

In 1961 the most total number of births in the U.S. The total was 4268.326 thousand.

year <- present[which.max(present[,"total"]), "year"]  
total <- present[which.max(present[,"total"]), "total"]

# Plot of the total number of baptisms per year  
plot(x = present$year, y = present$total, type="l",   
 main="Number of girls and boys baptized per year",  
 xlab="Year", ylab="Number of girls and boys baptized", sub="Data Source: Present Baptism Records")  
text(year, total, present[which.max(present[,"total"]), "total"], pos=4)  
points(year, total, pch = 19)

