Draft-Structure

1. Introduction
2. Data & Prepare Work
3. Background
4. Analysis(SQL for R)
5. Prediction(R and Excel)
6. Conclusion
7. Future Work
8. References

Introduction



HIV is a hot topic in the world especially in China, children do not have enough knowledge about how to prevent HIV infection. As the graph shows above, China is the country has the most HIV infection population in this database group which includes Australia, Brazil, China, India, Japan, Korea, Russia. And recently, CCR5 gene modification is another “breaking through” news that make worldwide biologists indignation. In this paper, we will see what is HIV distribution in China (Mainland) using Rsudio with SQL.

Graph RStudio Code:

//SQL

* City\_Set <- sqldf("select HIV1.CC,HIV1.SEX from HIV\_SQL\_W HIV1 where (HIV1.CC =‘AUS 'or HIV1.CC ='BRA‘ or HIV1.CC ='CHN‘ or HIV1.CC ='IND‘ or HIV1.CC =‘JPN‘ or HIV1.CC ='KOR‘ or HIV1.CC ='RUS')")
* > p\_Citytest <-ggplot(data = City\_Set, aes(x=CC))+geom\_bar(aes(fill=SEX))
* > print(p\_Citytest)
* > ggsave("World\_HIV.pdf")

//Rsudio build in

* > BigCounty\_plot <- subset(HIV\_09DEC18\_16\_19\_30\_87,CC %in% c("RUS","CAN","CHN","USA","BRA","AUS","IND","JPN","KOR"))
* > p\_BigCounty\_HIV <- ggplot(data=BigCounty\_plot,aes(x=CC))+geom\_bar(aes(fill=SEX))
* > print(p\_BigCounty\_HIV)

Data & Prepare Work

The Databases are from United States Census

<https://www.census.gov/data-tools/demo/hiv/hiv.html?s_appName=hiv&s_refdate=&s_age=&menu=dual_grid&s_ccode=>

and

<https://www.census.gov/data-tools/demo/hiv/hiv.html?s_appName=hiv&s_refdate=&s_age=&menu=dual_grid&s_ccode=CHN>

By doing this, simply changed the Database Column name to abbreviation. It will make getting data much easier because Rstudio consider blank space as invalid character. Also, defined data frame that can use in SQL later.

* names(HIV\_09DEC18\_16\_19\_30\_87)<- c("S","CC","GA","RD","PS","SC","SEX","A","SID","C","DT","CO","SN","AU","Y","T","PI","VT","NC","ND","PR","IR","ST","TT","SA")
* names(HIV\_07DEC18\_12\_21\_21\_55)<- c("S","CC","GA","RD","PS","SC","SEX","A","SID","C","DT","CO","SN","AU","Y","T","PI","VT","NC","ND","PR","IR","ST","TT","SA")
* > HIV\_SQL <- data.frame(HIV\_07DEC18\_12\_21\_21\_55)
* > HIV\_SQL\_W <- data.frame(HIV\_09DEC18\_16\_19\_30\_87)

Background

HIV- Human Immunodeficiency Virus could cause AIDS. It can destroy the immune system, and make it easier for people to get sick. There are three majority ways that people can get HIV.

From Planned Parenthood:”

1. Having vaginal or anal sex,
2. Sharing needles or syringes for shooting drugs, getting stuck with a needle that has HIV-infected blood on it
3. Getting HIV-infected blood, semen, or vaginal fluids into open cuts or sores on the body.”

Analysis

Gender, Population and Sub-Group

Gender:



From the graph, Bisexual is the most popular group that infect HIV. Most of them are homosexual so Homosexual is much easier than other people to get HIV infection.

Rstudio Code:

//SQL

* Gender\_set <- sqldf("select HIV.SEX from HIV\_SQL HIV")
* > p\_Gender\_set <- ggplot(data=Gender\_set,aes(x=SEX))+geom\_bar()
* > print(p\_Gender\_set)

//GGplot

* p\_Gender <- ggplot(data=HIV\_07DEC18\_12\_21\_21\_55,aes(x=SEX))+geom\_bar()

Sub-Group



(Blood donors, Drug users, Homosexuals, Homosexuals & bisexuals, Homosexuals/bisexuals, IVDU, IVDU & drug users, Pregnant women, STI pts.)

IVDU = Intravenous Drug Use

STI pts. = Sexually transmitted infection

First group combine Homosexual, Homosexual & bisexual, Homosexual/bisexual and STI pts. because they both belongs to SEX infection. This is the majority group of HIV infection.

Second group combine Drug users, IVDU, IVDU & drug users. Basically drug users shared needles if one of the group member got HIV. The virus will through the needle to infect others.

These are most two common ways to get HIV infection from our data. The result is exactly the same order as background mentioned.

Rstudio Code:

//SQL

* Subgroup\_set <- sqldf("select HIV.SEX, HIV.PS from HIV\_SQL HIV where (HIV.PS = 'Blood donors' or HIV.PS ='Drug users' or HIV.PS ='Homosexuals' or HIV.PS ='Homosexuals & bisexuals' or HIV.PS='IVDU' or HIV.PS='IVDU & drug users'or HIV.PS='Pregnant women'or HIV.PS='STI pts.')")
* p\_Subgroup\_set <- ggplot(data=Subgroup\_set,aes(x=PS))+geom\_bar(aes(fill=SEX))

//Rstudio build in

* Subgroup\_plot <- subset(HIV\_07DEC18\_12\_21\_21\_55, PS %in% c("Blood donors","Drug users","Homosexuals","Homosexuals & bisexuals","IVDU","IVDU & drug users","Pregnant women","STI pts."))

Population





First graph: Beijing, Chengdu, Guangzhou, Hangzhou, Ningbo, Qingdao, Shanghai, Shenyang, Shenzhen, Suzhou, Tianjin, Wuhan, Wuxi

Second graph: Dehong Prefecutre, Guangxi Autonomous Region, Kunming, Urban Area, Urumqi, Yuxi

It’s really easy to understand why Big City such as Beijing has more HIV infection because it has large population base. However, Small City also have HIV infection rising trend based on their poor medical environment and lack self-protection awareness.

Rstudio Code:

//SQL

* BigCity\_set <- sqldf("select HIV.SEX,HIV.GA from HIV\_SQL HIV where (HIV.GA = 'Beijing' or HIV.GA='Chengdu' or HIV.GA='Guangzhou' or HIV.GA='Hangzhou' or HIV.GA='Ningbo' or HIV.GA='Qingdao' or HIV.GA='Shanghai' or HIV.GA='Shenyang' or HIV.GA='Shenzhen' or HIV.GA='Suzhou' or HIV.GA='Tianjin' or HIV.GA='Wuhan' or HIV.GA='Wuxi’)”)
* p\_BigCity\_set <- ggplot(data=BigCity\_set,aes(x=GA))+geom\_bar(aes(fill=SEX))

//Rstudio

* BigCity\_plot <- subset(HIV\_07DEC18\_12\_21\_21\_55, GA %in% c("Beijing","Chengdu","Dalia","Guangzhou","Hangzhou","Ningbo","Qingdao","Shanghai","Shenyang","Shenzhen","Suzhou","Tianjin","Wuhan","Wuxi"))
* p\_BigCity\_HIV <- ggplot(data=BigCity\_plot,aes(x=GA))+geom\_bar(aes(fill=SEX))

//SQL

* SmallCity\_set <-sqldf("select HIV.SEX,HIV.GA from HIV\_SQL HIV where (HIV.GA ='Dehong Prefecture'or HIV.GA ='Guangxi Autonomous Region'or HIV.GA ='Kunming'or HIV.GA ='Urban area'or HIV.GA ='Urumqi'or HIV.GA ='Yunan Province'or HIV.GA ='Yuxi’)”)
* p\_SmallCity\_set <- ggplot(data=SmallCity\_set,aes(x=GA))+geom\_bar(aes(fill=SEX))+geom\_text(aes(x=GA,label=..count..),stat="count")

//Rstudio

* SmallCity\_plot <- subset(HIV\_07DEC18\_12\_21\_21\_55, GA %in% c("Dehong Prefecture","Guangxi Autonomous Region","Kunming","Urban area","Urumqi","Yunnan Province","Yuxi"))
* p\_SmallCity\_HIV <- ggplot(data=SmallCity\_plot,aes(x=GA))+geom\_bar(aes(fill=SEX))

Prediction

The HIV population in China is growing since 1986, the Database doesn’t include the whole data for 2017 so the trending line will decrease a little that could cause error to our prediction. From the graph, we could say the number of people who infect HIV will still increasing in the future, so I used two predict function: Linear and Holt-Winters in Rsudio, Liner and Exponential in Excel Slover.



(From 1986 to 2017 )

Rsudio Code:

//SQL

* Year\_Set <- sqldf("select HIV.Y, HIV.SEX from HIV\_SQL HIV")
* p\_Yeartest <- ggplot(data=Year\_Set, aes(x=Y))+geom\_bar(aes(fill=SEX))

//GGplot

* p\_Year\_HIV <- ggplot(data=HIV\_07DEC18\_12\_21\_21\_55,aes(x=Y))+geom\_bar(aes(fill=factor(..count..)))+geom\_text(aes(x=Y,label=..count..),stat="bin")

(I set x value to Year and y value to Count in Rstudio that will make easier to do prediction

* x <- c(1986,1987,1988,1989,1990,1991,1992,1993,1994,1995,1996,1997,1998,1999,2000,2001,2002,2003,2004,2005,2006,2007,2008,2009,2010,2011,2012,2013,2014,2015,2016,2017)
* y <- c(25,35,35,33,49,51,54,26,247,23,133,195,369,306,653,907,604,342,852,551,1619,653,615,1007,781,944,1366,1890,1660,1507,1277,197)

)

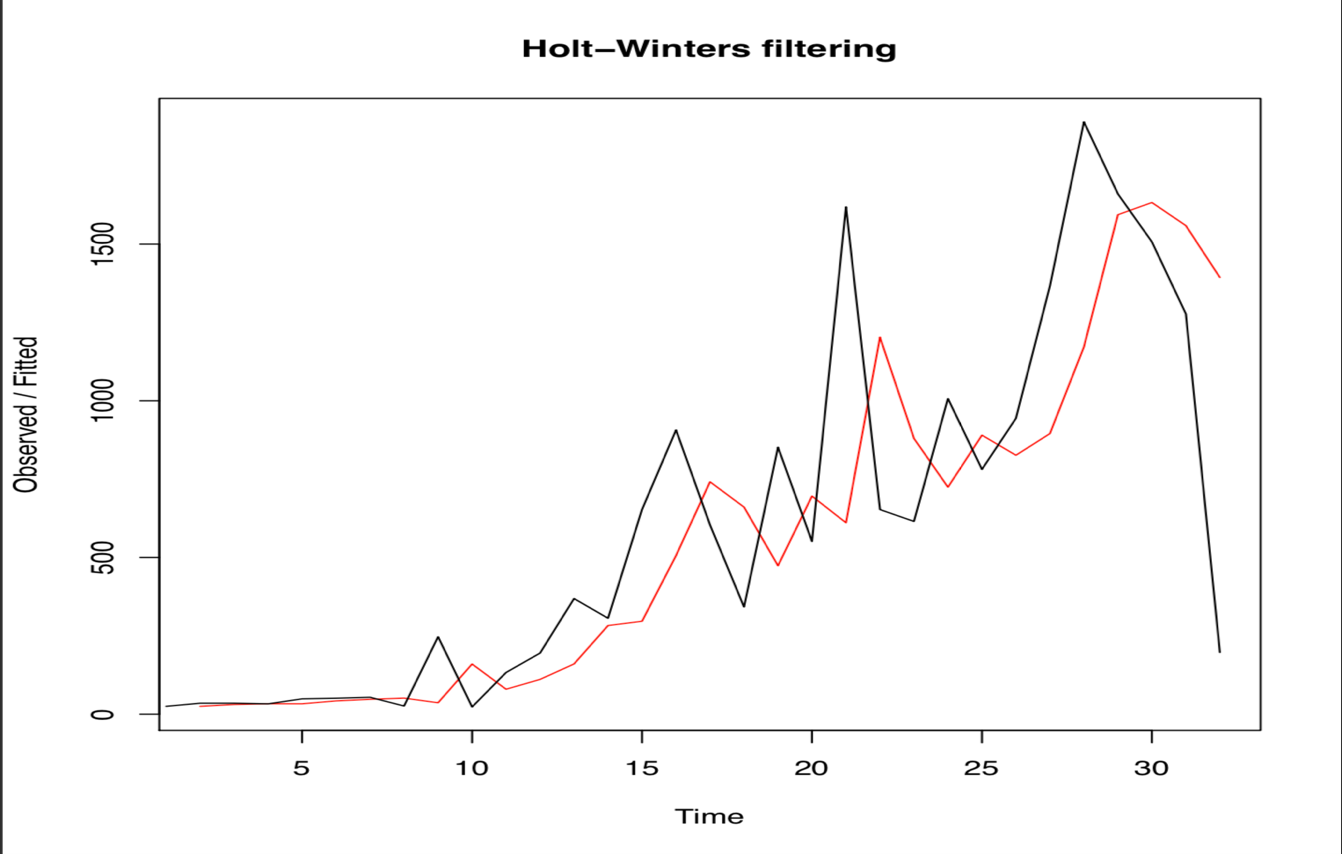


First, we check the correlation between x and y (using cor(x,y) in Rstudio), the value is 0.7964651 which is positive number, so we can consider using Linear Function. We got a=47.99707 and b = -95472.19300

Analysis of Linear Function:

* anova(fit)
* Analysis of Variance Table
* Response: x
* Df Sum Sq Mean Sq F value Pr(>F)
* y 1 6284544 6284544 52.047 4.977e-08 \*\*\*
* Residuals 30 3622414 120747
* ---
* Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Since Pr(>F) = 4.977e-08 it means the level under α=0.05 and it tells us there exists a liner relation between X and Y.

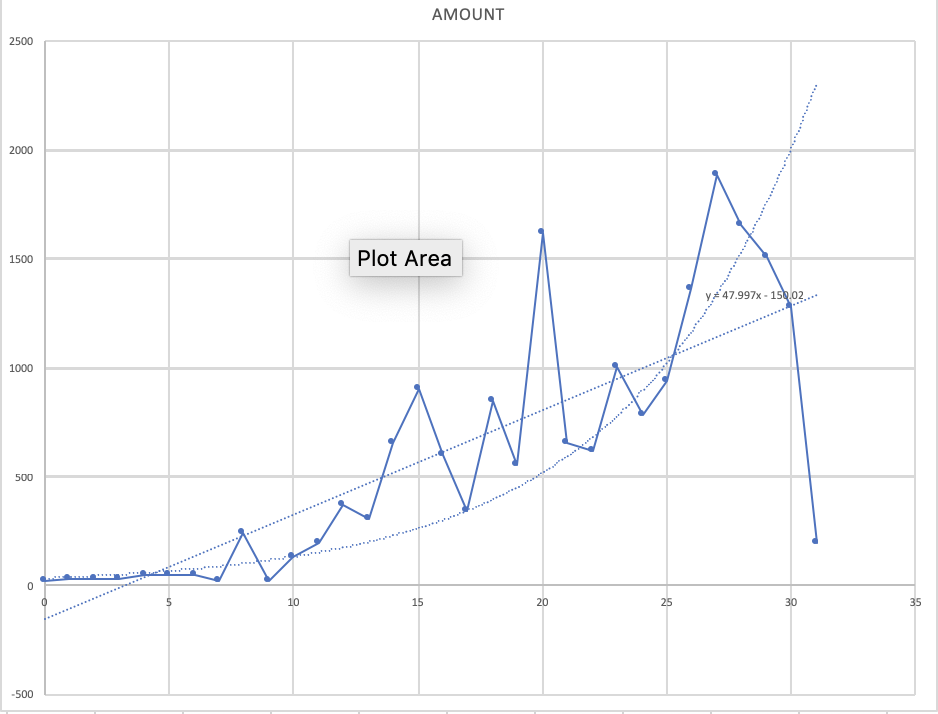


(Black line is original data and Red line is trending line)

Because the data shows Exponential smoothing. So I decided to use Holt-Winters to do prediction. After predict line fitted into data the SSE is 4453998 which mean Linear Function is better.

Future Predict based on Holt-Winters Model:

* Point Forecast Lo 80 Hi 80 Lo 95 Hi 95
* 33 690.7228 199.22649 1182.219 -60.95594 1442.402
* 34 690.7228 120.73183 1260.714 -181.00316 1562.449
* 35 690.7228 51.80905 1329.637 -286.41146 1667.857
* 36 690.7228 -10.37053 1391.816 -381.50692 1762.953
* 37 690.7228 -67.46777 1448.913 -468.82961 1850.275
* 38 690.7228 -120.55645 1502.002 -550.02174 1931.467
* 39 690.7228 -170.37829 1551.824 -626.21768 2007.663
* 40 690.7228 -217.47111 1598.917 -698.23992 2079.686
* 41 690.7228 -262.23956 1643.685 -766.70736 2148.153
* 42 690.7228 -304.99721 1686.443 -832.09954 2213.545



This is excel solver model to do trending line based on our Dataset

Linear and Exponential Function Compare:

|  |  |  |  |
| --- | --- | --- | --- |
| y=ax+b |  | SSE | 3622413.852 |
| a | 47.9970632 |  |  |
| b | -150.01705 |  |  |
|  |  |  |  |
|  |  |  |  |
| y=a\*e^b\*x |  | SSE | 4351813.453 |
| a1 | 180.334644 |  |  |
| b1 | 0.06768395 |  |  |

Compare with the SSE we can see Linear Function is better in our HIV case.

And the Liner predicion SSE in Rstuio is 3622414 which is pretty closed to 3622413.852 that is SSE in excel for Liner prediction.

From our models, we can see it’s really hard to predict HIV infection in CHINA, that’s because HIV infection depends on some effectors that not increasing or decreasing smoothly such as medication level, how many people had unprotected sex with HIV infection people or some breaking technology that could eliminate HIV virus in human body.

As I mentioned CCR5 ,a gene that most HIV virus will infect, there is only one person who successful modified CCR5 gene to cure HIV infection. So this technology is still not 100 percent sure can be used into cure HIV infection.

Conclusion

The HIV infection population may still increasing in the future. What we should do to prevent this by ourselves is getting more knowledges about HIV, increasing our self-protect awareness, don’t feel afraid if got HIV, get in touch with doctor as soon as possible.

For Chinese government, popularized HIV knowledges is really important. Let Chinese know HIV instead of saying HIV is dangerous. Also enhance urban area medical level to help people who don’t live in city could get a good medication.

Future Work

There are several topics that could improve this research. The first one is that analyze current HIV death rate under HAAT Therapy. The second is that Analyze after taking HIV Blocker how many people still get HIV

References

* <https://blog.csdn.net/tMb8Z9Vdm66wH68VX1/article/details/79544739>
* <https://blog.csdn.net/fengziyun/article/details/7757779>
* <https://blog.csdn.net/u012671171/article/details/43155351>
* <https://www.jianshu.com/p/098616aabed1>
* <https://www.rstudio.com/wp-content/uploads/2015/03/ggplot2-cheatsheet.pdf>
* <https://www.plannedparenthood.org/learn/stds-hiv-safer-sex/hiv-aids>