Honors College Thesis

An Analysis of Factors Associated with HIV and Hepatitis C Infection

By Ali Charlson
Major: Statistics
Oklahoma State University

Table of Contents

Acknowledgements

I would like to thank Dr. Lan Zhu for being my thesis director for this project. Her guidance and readiness to answer questions helped me explore a new topic within statistics that I had not encountered at this in-depth of a level, and she helped me learn a lot about categorical data analysis that will be helpful in my future with statistics.

I would also like to thank Dr. Carla Goad for being my second reader and also my advisor throughout my time as an undergraduate. Her support initially encouraged me to switch my major to statistics and I have enjoyed working with her on this project as well.

Lastly, I would like to thank Ebonie Hill-Williamson for being my honors advisor over all four and a half years of my undergraduate education. She was always ready to answer my (many) questions about honors courses, switching my major, honors projects, or anything else that I needed help with. Being a member of the Honors College at Oklahoma State University has been a great addition to my time as an undergraduate and I am grateful for all of the support I have received over the years.

Introduction

Human Immunodeficiency Virus (HIV) and Hepatitis C Virus (HCV) are both potentially life threatening diseases when left untreated. According to the CDC, approximately one in four people diagnosed with HIV are also diagnosed with HCV, and it has been shown that chronic HCV can sometimes advance faster in people already diagnosed with HIV (Center for Disease Control and Prevention 2015). The two diseases are spread in a similar manner, and it is possible that the factors associated with having HIV, HCV, or both are also similar. Being able to identify the factors associated with either disease could be helpful information for future treatments and preventive measures for patients diagnosed with HIV, HCV, or both.

HIV by itself is a virus that, when left untreated, can turn into Acquired Immunodeficiency Syndrome (AIDS) at its most advanced stage, which severely incapacitates the body's ability to fight infections. HIV is transmitted by contact of infected bodily fluids through either the sharing of injection equipment (needles and syringes) or unprotected sexual acts. In early stages of HIV, a person will experience flu-like symptoms and as the disease continues to multiply; the person's immune system can continue to weaken leading to higher rates of other infections. HCV is a similar disease primarily affecting the liver, and is also transmitted through sharing injection equipment or unprotected sexual acts. HCV symptoms include jaundice, joint pain, and fever, though many cases of HCV are asymptomatic. Both HIV and HCV are diseases that can require lifelong treatment, and medical intervention is always necessary for a patient to manage either disease.

Coinfection of HIV and HCV is a common problem as both diseases are transmitted in the same ways. In fact, fifty to ninety percent of people with HIV who inject drugs are also diagnosed with HCV, suggesting it is possible that injecting drugs and sharing injection

equipment is contributing to higher rates of coinfection of the two diseases (Center for Disease Control and Prevention 2015). This project aims to examine factors that could be associated with people diagnosed with HIV, HCV, or both diseases using data from a sample region in China with high HIV and HCV prevalence rates in order to discover relevant information on how infections of HIV and HCV might be determined and monitored. Overall, the project is an observational analysis with two main goals. The first goal is to identify potential factors associated with coinfection rates of HIV and HCV across China, and analyze in what proportion of population those factors are appearing. The second goal is to examine possible predictors of infection with HIV and HCV. In doing so, meaningful comparisons can be made between predictors associated with either disease. Identifying demographic and behavioral factors found in areas with high HIV and HCV prevalence rates could be useful for future research for both diseases.

Data Information

The data set used for this project was collected from an area in China from May 2004 to September 2012. There are 4,443 (n = 4,443) observations with 18 possible variables per observation. Some observations had missing variables that were not recorded. A variable list of the different recorded variables is presented below. All data is categorical, with the nominal variables being sex, education level, duration of use, marital status, nation, occupation, shared syringe, rehabilitation, sexuality, drug injection, manner of drug use, HCV, HIV, and both HIV and HCV. The ordinal variables are age and initial age of use. Throughout the analysis, HCV, HIV, and both HCV and HIV are used as dependent variables. All other variables are independent.

Variable	Coding
Sex	0 for male, 1 for female
Age	1 for < 25 years, 2 for 25-35 years, 3 for 35-45 years, 4 for > 45 years
Age.initial – age of initial use	1 for < 20 years old, 2 for 20-30 years old, 3 for > 30 years old
Duration – duration of use	1 for <1 year, 2 for 1-5 years, 3 for 5-10 years, 4 for 10-15 years, 5 for > 15 years
Marital.status	0 for unmarried, 1 for married, 2 for divorced
Education	0 for illiterate, 1 for primary, 2 for junior, 3 for senior, 4 for college

Nation	0 for others, 1 for Han
Occupation	0 for unemployed, 1 for peasant, 2 for services, 3 for staff
Shared.s – shared syringe	0 for no, 1 for yes
Shared.3 – shared syringe for three months	0 for no, 1 for yes
Rehabilitation	0 for no, 1 for yes
Sexuality – participation in sexual acts	0 for no, 1 for yes
Inject – use of injection methods	0 for no, 1 for yes
Manner – manner of which drugs are taken	0 for mixed, 1 for by mouth, 2 for injected
HIV	0 for not diagnosed, 1 for diagnosed
HCV	0 for not diagnosed, 1 for diagnosed
HIV.HCV	0 for neither HIV or HCV, 1 for either HIV or HCV, 2 for HIV + HCV

Methodology

To satisfy the first goal of the project, it is necessary to verify that all of the variables being used in the analysis are factors significantly associated with HIV, HCV, or both. This is done using RxC contingency tables. These tables first provide descriptive analysis of the categorical data and display how it is divided across each level of the dependent variable. Each table is used to observe the proportions of the data divided into the different variable categories. The Pearson's Chi-square Test is then performed on each of the tables to identify which of the variables are significantly associated with HIV or HCV or both. The hypothesis being tested by the Chi-square test is:

H_0: the chosen variable and HIV.HCV are independent

H_a: the chosen variable and HIV.HCV are not independent

For use of the Chi-square test to be valid, two assumptions must be checked. The first assumption is that the data in each cell for each variable is only contributing to that cell. For example, for the variable gender the RxC contingency table is:

	Neither disease	**One disease**	**Both HIV and HCV**	**Totals**
Male	1,564	781	268	2,613
Female	337	115	38	490
Totals	1,901	896	306	3,103

It is obvious that none of the data in the male cells can also be contributing to the data in the female cells, and vice versa. This logic holds for the tables of all other variables, so this assumption is checked, and the Chi-square Test can still be used. The second assumption is that each table should have at least twenty subjects. As the data set has a total of n = 4,443 observations, all of the tables are able to have at least twenty subjects, so the second assumption is checked as well. The formula for the degrees of freedom for the Chi-square Test depends on

the number of rows and columns of the table being tested, so each test for each variable will have a different number of degrees of freedom, but that will not affect the accuracy of the Chi-square test for each individual variable. Contingency tables and their corresponding degrees of freedom and p-values are listed in the results section and Appendix A.

The second goal of the project is to examine the different predictors for contracting just HIV or just HCV. Examining models for the different diseases allows meaningful comparisons to be made between what factors are possible predictors of having HIV or HCV. Individually, HIV and HCV are binary response variables, so logistic regression is used to create the two separate models for HIV and HCV. HIV is the response variable for the first model and forward stepwise regression is used for variable selection in order to find the most accurate fit possible for this model. This is done by creating a model with no predictors:

HIV ~ 1

a model with all predictors:

HIV ~ age + age.of.initial.use + duration + gender + marital.status + education + nation + occupation + shared.syringe + rehabilitation + sexuality + inject.drug + drug.manner

and then searching through all possible models within this range to find the most accurate one by comparing relative AIC values. The null hypothesis being tested here is:

$\boldsymbol{H_0}$**:** all $\beta_i's = 0$

$\boldsymbol{H_a}$**:** at least one $\beta_i \neq 0$

The same process is then repeated for the second model, with HCV as the response variable. It is also tested using the hypothesis:

$\boldsymbol{H_0}$**:** all $\beta_i's = 0$

$\boldsymbol{H_a}$**:** at least one $\beta_i \neq 0$

This logistic regression has four different assumptions that must be checked. The first is that the response variable is binary, which has been previously demonstrated to be true in the Data Information section. The second assumption is that the response variables are coded correctly, with a value of '1' meaning that the event in question occurs and a value of '0' meaning that the event does not occur. This is also demonstrated in the Data Information section. The model must also be correctly fitted, which is checked by the use of forward stepwise regression. Lastly, the sample size must be adequately large enough, as it is recommended that for each predictor in the model, there are at least ten observations for each variable. The sample size used for both models is over 3,000, so this is satisfied.

Results: Chi-Square Test for Independence

An example of two of the contingency tables analyzed are displayed below, along with the resulting p-values and degrees of freedom from the Chi-square Test. The first table displays the nominal variable 'injection' and the second table displays the ordinal variable 'age.' Nominal and ordinal variables were the only types used in this analysis. The contingency tables for all other variables are displayed in Appendix A.

Variable: Injection

	Neither disease	**One disease**	**Both HIV and HCV**	**Totals**
No	1,348	342	73	1,763
Yes	553	554	233	1,340
Totals	1,901	896	306	3,103

p-value: <2.2e-16, degrees of freedom: 2

Variable: Age

	Neither disease	**One disease**	**Both HIV and HCV**	**Totals**
< 25	834	413	153	1,400
25 - 35	797	372	129	1,298
35 – 45	231	98	23	352
> 45	39	13	1	53
Totals	1,901	896	306	3,103

p-value: 0.04725, degrees of freedom: 6

The Chi-square Test was evaluated using a significance level of $\alpha = .05$. Every variable from the data set was found to have a statistically significant association with the ordinal variable HIV.HCV. Every p-value was less than .05, with most of the p-values also being less than .01. Thus, for every variable we reject the null hypothesis that the variable is independent from HIV.HCV. These significant associations mean that any of the variables recorded could be related to someone having HIV, HCV, or both HIV and HCV. Since each variable is at least significantly associated with HIV, HCV, or both HIV and HCV, the regression analysis will examine all of the variables as potential predictors.

Results: Logistic Regression - HIV

The forward stepwise regression found the best fitting model for predicting HIV to be:

$$\textit{HIV} \sim -1.659 - 0.584*\textit{education2} - 1.65*\textit{education3} - 1.70*\textit{education4} - 2.59*\textit{education5} + 1.28*\textit{inject.drug} + 0.219*\textit{marital.status} + 0.039*\textit{duration2} + 0.430*\textit{duration3} + 0.406*\textit{duration4} + 0.419*\textit{duration5}$$

The final model was calculated with n = 3,495 and had an AIC value of 3,229.9, which was the lowest of all possible models evaluated. AIC is used to compare the relative quality of a set of statistical models, and a relatively lower AIC value means that the model provides the best fit possible for the data. After analyzing the best fitted model, the logistic regression resulted in inject.drug, marital.status, and all levels of education being found as significant predictors for HIV at $\alpha = .05$. This means that each of these predictors have a statistically significant effect on the final outcome of the dependent variable HIV. We can reject the null hypothesis that all of the $\beta_i's$ are equal to zero and conclude that these variables are significant predictors of HIV. Duration of use was not found to be a statistically significant predictor at any of its levels, as all p-values were greater than .05 for the coefficients.

Interpreting the exact meaning of the $\beta_i's$ requires a closer analysis. Logistic regression is done on a log-odds scale and the applications of each coefficient are not immediately clear. For example, the exact interpretation of the coefficient for the variable 'gender' shows that for a one unit increase in the value of gender (essentially being female as opposed to male), there is a predicted one unit increase in the log-odds of HIV. Log-odds are calculated by the formula:

$$\boldsymbol{log}\left(\frac{\boldsymbol{p}}{(\mathbf{1}-\boldsymbol{p})}\right)$$

where p is the probability of having a diagnosis of the disease in question

While this formula is necessary for logistic regression to run properly and identify which variables are contributing to a significant change in HIV, the log-odds scale is not as useful when trying to explain what the magnitude of that change is. For actual interpretation, the odds ratio of each significant predictor can be examined for better information.

HIV Odds Ratios for Significant Predictors

Variable	Odds Ratio	P-Value
Education2	0.55766160	1.60 e-08
Education3	0.19175682	< 2 e-16
Education4	0.18318705	4.04 e-15
Education5	0.07472614	7.28 e-07
Inject.Drug	3.586770	< 2 e-16
Marital.Status1	1.33108839	0.0197

These ratios allow a conclusion to be drawn from each significant predictor. The odds ratio for the variable ‘education’ is less than one, so we can say that for a one unit increase in education (going from 0- illiterate to 1- primary school), the odds of having HIV for an individual who went to primary school are roughly 0.558 times lower compared to the odds of having HIV for an individual who is in the illiterate category, holding all else constant. Each increase in education level is resulting in the odds of having HIV for an individual at the higher education level being lower than the odds of having HIV at the baseline of illiterate education level. For the other two variables, the odds ratio is above one so the interpretation is slightly different. For example, the odds ratio for inject.drug means that a one unit change in inject.drug (going from 0 - not injecting to 1 - injecting) results in the odds of having HIV being about 3.58 times higher for someone who participates in drug injection behavior when compared to the odds of having HIV for someone who does not inject drugs, holding all else constant. Similar conclusions can be made for the other variables. Marital.status (going from unmarried - 0 to married - 1) resulted in the odds of having HIV being 1.33 times higher for a married person than an unmarried person.

Results: Logistic Regression - HCV

For the dependent variable HCV, the forward stepwise regression found the best fitting model to be:

$$HCV \sim 0.185 + 0.251*rehabilitation - 0.010*education2 + 0.032*education3 - 0.015*education4 - 0.132*education5$$

The final model was calculated with n = 3,246 and had an AIC value 3,922.8 which was relatively lower than the AIC for the other models. The logistic regression found rehabilitation and the fifth level of education (education5) to be the only statistically significant predictors of the log odds of HCV at $\alpha = .05$. This means we can reject the null hypothesis that all of the $\beta_i's$ are the same, and that the $\beta_i's$ for rehabilitation and education5 are statistically significantly different than zero. The odds ratios for the significant predictors are:

HCV Odds Ratios for Significant Predictors

Variable	Odds Ratio	P-Value
Rehabilitation	1.2849910	< 2 e-16
Education5	0.8766618	.00768

The odds ratio for rehabilitation means that when there is a one unit change in rehabilitation (going from 0- no rehabilitation to 1- rehabilitation), the odds of having HCV are roughly 3.54 times higher for an individual who has gone through rehabilitation than for someone who has not gone through rehabilitation. The only significant change in the education variable was the difference between an individual who was in the illiterate category and an individual who had completed college. The odds of having HCV for someone who had completed college were 0.88 times lower than the odds of having HCV for someone who had not completed any schooling.

Graphics

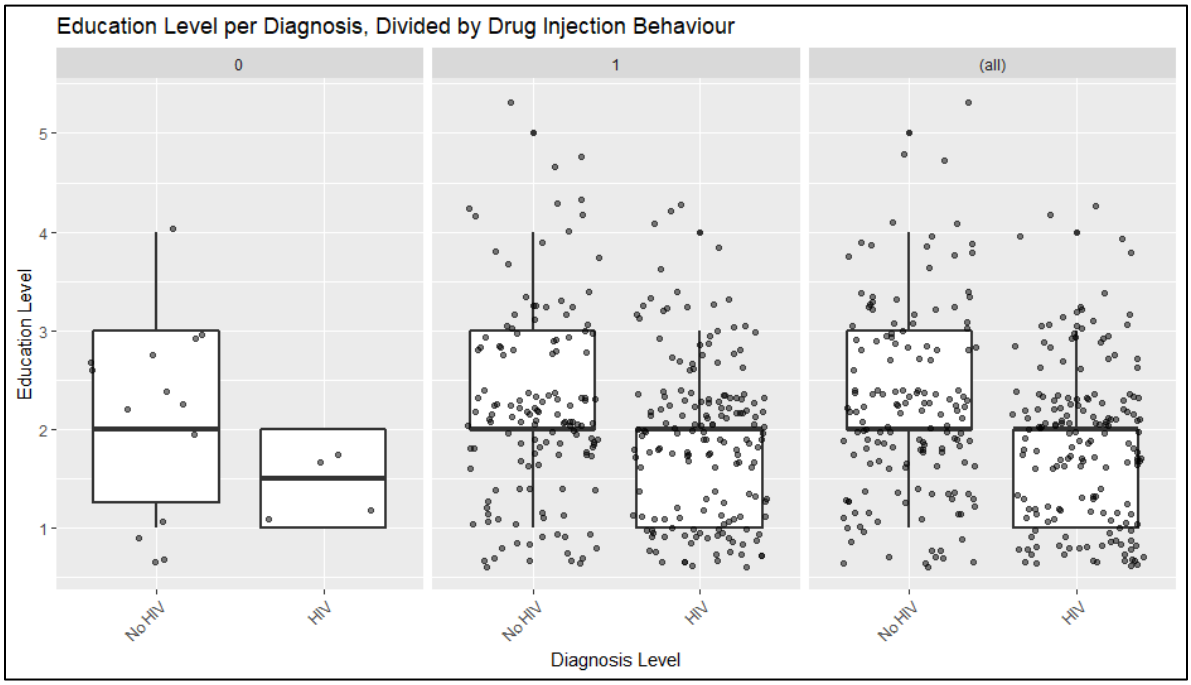

Fig. 1: *Data spread across the significant predictors from the HIV model, 0 = did not participate in drug injection, 1 = did participate in drug injection*

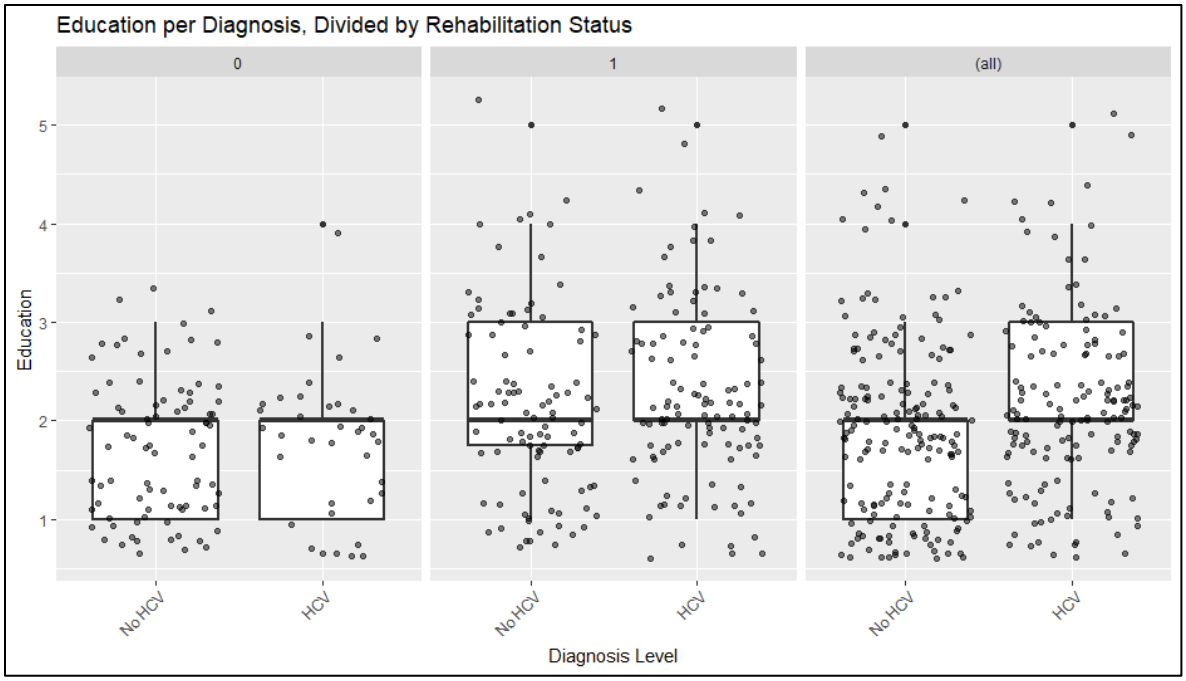

Fig. 2: *Data spread across significant rehabilitation predictor from the HCV model, 0 = did not have rehabilitation, 1 = did have rehabilitation*

Discussion

The results of each Chi-square test for independence on the contingency tables are fairly straight forward, showing that all variables examined could potentially be associated with whether or not an individual has just HIV, just HCV, or both. None of the factors analyzed were found to be independent from the HIV.HCV variable, which could be useful information for healthcare professionals when examining patients. Knowing that all of these factors are associated with the presence of HIV and HCV could help doctors monitor, diagnose, and treat patients who may be at risk for HIV, HCV, or both. Additionally, patients who are already diagnosed with just HIV or just HCV could take preventive measures to control for changes in factors associated with both diagnoses, to lower the possibility of contracting a second disease as well. These associated factors can help create a more complete understanding of who is being diagnosed with HIV, HCV, or both diseases.

The regression analysis results for HIV can be used to make broader predictions about who could be diagnosed with HIV. Injecting drugs is a very significant predictor of contracting HIV, so preventing and monitoring this activity in patients could be helpful. Additionally, marital status and education level changes were also found to be useful in predicting HIV, so public health professionals should be aware of these in at risk areas as well. The more significant information that is known about people who are diagnosed with HIV, the easier it will be for new diagnoses to be prevented.

The HCV regression analysis holds similar information as the HIV analysis; however the factors used in prediction are different. Rehabilitation was the main significant factor used to predict if a person might contract HCV or not. Interestingly, having gone through rehabilitation was associated more with having HCV over not having it. This is unexpected, as going through

rehabilitation is generally regarded as a positive choice that should be associated with positive outcomes. It is possible that someone who has already gone through drug rehabilitation has struggled more severely with drug use over their life and has been exposed to more situations where HCV could be transmitted, despite going through rehabilitation that should have been helpful. Regardless, the associations between rehabilitation attendance and HCV could be studied further, as the outcome is surprising. Additionally, completion of college was the only education level that seemed to lower the possibility of having HCV when compared to having no education level. Differences between each education level could also be looked at to examine if each increase in education level (elementary to middle school, middle to high school, etc.) are also having a significant impact on an individual having HCV.

One might assume that since HIV and HCV are contracted in similar ways, the most significant predictors for each model would be the same. It is a surprising result of this analysis to find that to not be true. The patients in this data set diagnosed with HIV over HCV have an entirely different set of predictors that relate to their diagnosis in a significant way. For example, as both HIV and HCV can be transmitted through injection drug use, it makes sense to assume that participating in injection drug use would be a significant predictor for both diseases. However, this analysis found that injection drug use was only predictive for someone being diagnosed with HIV rather than HCV, so perhaps there are additional biological or social factors that associate with injection drug use and diagnosis of HIV as well. Further study could be done examining these differences more in-depth.

Comparing these models is useful for healthcare practitioners as they monitor patients with HIV, HCV, or both. Changes in any of these variables could ultimately predict the contraction of an additional new disease, and trying to control or prevent these changes could

help many people from being diagnosed with a second difficult to manage disease. HIV and HCV are two very serious problems that require immediate and aggressive treatment. The more information that healthcare professionals have about who is at risk for HIV and HCV, the more opportunities there are for preventive measures and resources to be put into place. Having a list of potential characteristics and behaviors that are associated or predictive of HIV and HCV could allow for life saving treatments to be distributed in a more effective and useful manner.

Future Research

There is potential for future research relating to the coinfection rates of HIV and HCV. As the variable HIV.HCV has three levels (0 for neither disease, 1 for one disease, 2 for both diseases) ordinal logistic regression could be used to analyze the data with HIV.HCV as the response variable. This could present further information regarding what variables are possible significant predictors of contracting both diseases, in comparison to contracting just one disease or neither disease. There may be further interactions between the variables used in this project that did not show up when just using the binary HIV or HCV variables as responses. Analyzing HIV and HCV within these ordinal grouping levels would be a good way to potentially identify these interactions and assess further information about the diseases overall.

References

Analytics Vidhya Content Team. (November 1, 2015). *Simple Guide to Logistic Regression in R.* Retrieved from: https://www.analyticsvidhya.com/blog/2015/11/beginners-guide-on-logistic-regression-in-r/

Center for Disease Control and Prevention. *Hepatitis C FAQs for Health Professionals. Retrieved from:* https://www.cdc.gov/hepatitis/hcv/hcvfaq.htm#section2

Center for Disease Control and Prevention. *About HIV/AIDS*. Retrieved from: https://www.cdc.gov/hiv/basics/whatishiv.html

H. Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.

Lindquist, Martin. *Variable Selection.* Retrieved from: http://www.stat.columbia.edu/~martin/W2024/R10.pdf

Presnell, B. (May 28, 2000). *An Introduction to Categorical data Analysis Using R*. Retrieved from: http://web.stat.ufl.edu/~presnell/Courses/sta4504-2000sp/R/R-CDA.pdf

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Appendix A: Contingency Tables and Chi-Square Test Results

Variable: Gender

	Neither disease	One disease	Both HIV and HCV	Totals
Male	1,564	781	268	2,613
Female	337	115	38	490
Totals	1,901	896	306	3,103

p-value: 0.0009746, degrees of freedom: 2

Variable: Age

	Neither disease	One disease	Both HIV and HCV	Totals
< 25	834	413	153	1,400
25 - 35	797	372	129	1,298
35 – 45	231	98	23	352
> 45	39	13	1	53
Totals	1,901	896	306	3,103

p-value: 0.04725, degrees of freedom: 6

Variable: Age of Initial Use

	Neither disease	One disease	Both HIV and HCV	Totals
< 20	318	221	87	626
20 - 30	930	485	170	1,585
>30	653	190	49	892
Totals	1,901	896	306	3,103

p-value: < 2.2e-16, degrees of freedom: 4

Variable: Duration of Use (years)

	Neither disease	One disease	Both HIV and HCV	Totals
< 1	65	38	9	112
1 – 5	993	313	112	1,418
5 - 10	430	217	80	727
10 - 15	280	230	75	585
> 15	133	98	30	261
Totals	1,901	896	306	3,103

p-value: < 2.2e-16, degrees of freedom: 8

Variable: Marital Status

	Neither disease	One disease	Both HIV and HCV	Totals
Unmarried	338	197	70	605
Married	1,498	640	219	2,357
Divorced	65	59	17	141
Totals	1,901	896	306	3,103

p-value: 2.057e-05, degrees of freedom: 4

Variable: Education

	Neither disease	One disease	Both HIV and HCV	Totals
Illiterate	351	179	87	617
Primary	859	366	153	1,378
Junior	460	255	49	764
Senior	163	76	17	256
College	68	20	0	88
Totals	1,901	896	306	3,103

p-value: 5.42 e-08, degrees of freedom: 8

Variable: Nation

	Neither disease	One disease	Both HIV and HCV	Totals
Others	1,585	677	285	2,547
Han	316	219	21	556
Totals	1,901	896	306	3,103

p-value: 2.382e-12, degrees of freedom: 2

Variable: Occupation

	Neither disease	One disease	Both HIV and HCV	Totals
Unemployed	221	161	27	409
Peasant	1,489	643	263	2,395
Services	85	45	9	139
Staff	106	47	7	160
Totals	1,901	896	306	3,103

p-value: 1.118e-06, degrees of freedom: 6

Variable: Shared Syringe

	Neither disease	One disease	Both HIV and HCV	Totals
No	1,788	780	199	2,767
Yes	113	116	107	336
Totals	1,901	896	306	3,103

p-value: >2.2e-16, degrees of freedom: 2

Variable: Rehabilitation

	Neither disease	One disease	Both HIV and HCV	Totals
No	1,186	372	116	1,674
Yes	715	524	190	1,429
Totals	1,901	896	306	3,103

p-value: < 2.2e-16, degrees of freedom: 2

Variable: Sexuality

	Neither disease	One disease	Both HIV and HCV	Totals
No	734	290	108	1,132
Yes	1,097	554	186	1,837
Totals	1,831	844	294	2,969

p-value: 0.01576, degrees of freedom: 2

Variable: Injection

	Neither disease	One disease	Both HIV and HCV	Totals
No	1,348	342	73	1,763
Yes	553	554	233	1,340
Totals	1,901	896	306	3,103

p-value: <2.2e-16, degrees of freedom: 2

Variable: Manner of Use (of drugs)

	Neither disease	One disease	Both HIV and HCV	Totals
Mixed	256	281	120	657
By Mouth	1,401	411	102	1,914
Injected	171	157	76	404
Totals	1,828	849	298	2,975

p-value: <2.2e-16, degrees of freedom: 4

Appendix B: R Code

```
#loading in data set
library(readr)
thesis <- read_csv("~/Fall '17/thesis.xlsb.csv")
View(thesis)

#identifying all variables, as factor when necessary and creating three level
#variable

age <- thesis$age
age.initial <- thesis$age.of.initial.use
duration <- as.factor(thesis$Duration)
gender <- factor(thesis$gender)
marital.status <- factor(thesis$marital.status)
Education <- as.factor(thesis$education)
nation <- factor(thesis$nation)
occupation <- factor(thesis$occupation)
shared.s <- factor(thesis$shared.syringe)
shared.3 <- factor(thesis$Shared.3.months.)
rehab <- factor(thesis$rehabilitation)
sexuality <- factor(thesis$sexuality)
inject <- factor(thesis$inject.drug)
manner <- factor(thesis$drug.manner)
HCV <- factor(thesis$HCV)
HIV <- factor(thesis$HIV)
thesis$HIV.HCV <- thesis$HIV + thesis$HCV
HIV.HCV <- factor(thesis$HIV.HCV)

#chisquare test to calculate association of factors:
#r x c contingency tables
table1 <- table(gender, HIV.HCV)
dimnames(table1) <- list(c("Male","Female"),c("Neither","One", "Both"))
names(dimnames(table1)) <- c("Gender","HIV.HCV")
table1
chisq.test(table1)
#pvalue 0.0009746 - gender


table2 <- table(age, HIV.HCV)
dimnames(table2) <- list(c("< 25","25-35", "35-45", ">
45"),c("Neither","One", "Both"))
names(dimnames(table2)) <- c("Age","HIV.HCV")
table2
chisq.test(table2)
#pvalue 0.04725 - age

table3 <- table(age.initial, HIV.HCV)
dimnames(table3) <- list(c("< 20","20-30", "> 30"),c("Neither","One",
"Both"))
names(dimnames(table3)) <- c("Age of Initial Use","HIV.HCV")
table3
chisq.test(table3)
#pvalue < 2.2e-16 - age.initial

table4 <- table(duration, HIV.HCV)
dimnames(table4) <- list(c("< 1","1-5", "5-10", "10-15", ">
15"),c("Neither","One", "Both"))
names(dimnames(table4)) <- c("Duration (years)","HIV.HCV")
table4
chisq.test(table4)
#pvalue < 2.2e-16 - duration
```

```
table5 <- table(marital.status, HIV.HCV)
dimnames(table5) <- list(c("Unmarried","Married",
"Divorced"),c("Neither","One", "Both"))
names(dimnames(table5)) <- c("Marital Status","HIV.HCV")
table5
chisq.test(table5)
#pvalue 2.057e-05 - marital status

table6 <- table(education, HIV.HCV)
dimnames(table6) <- list(c("Illiterate","Primary", "Junior", "Senior",
"College"),c("Neither","One", "Both"))
names(dimnames(table6)) <- c("Education","HIV.HCV")
table6
chisq.test(table6)
#pvalue 5.42e-08 - education

table7 <- table(nation, HIV.HCV)
dimnames(table7) <- list(c("Others","Han"),c("Neither","One", "Both"))
names(dimnames(table7)) <- c("Nation","HIV.HCV")
table7
chisq.test(table7)
 #pvalue 2.382e-12 - nation

table8 <- table(occupation, HIV.HCV)
dimnames(table8) <- list(c("Unemployed", "Peasant", "Services",
"Staff"),c("Neither","One", "Both"))
names(dimnames(table8)) <- c("Employed","HIV.HCV")
table8
chisq.test(table8)
 #pvalue 1.118e-06 - employment status

table9 <- table(shared.s, HIV.HCV)
dimnames(table9) <- list(c("No","Yes"),c("Neither","One", "Both"))
names(dimnames(table9)) <- c("Shared Syringe","HIV.HCV")
table9
chisq.test(table9)
#pvalue < 2.2e-16 - shared syringe

table10 <- table(shared.3, HIV.HCV)
dimnames(table10) <- list(c("No", "Yes"),c("Neither","One", "Both"))
names(dimnames(table10)) <- c("Shared Syringe, 3 months","HIV.HCV")
table10
#remove this predictor as it only has 339 observations and is limiting the
data set analysis
chisq.test(table10)
#pvalue 0.0009885 - shared syringe, 3 months

table11 <- table(rehab, HIV.HCV)
dimnames(table11) <- list(c("No","Yes"),c("Neither","One", "Both"))
names(dimnames(table11)) <- c("Rehabilitation","HIV.HCV")
table11
chisq.test(table11)
#pvalue < 2.2e-16 - rehabilitation

table12 <- table(sexuality, HIV.HCV)
dimnames(table12) <- list(c("No","Yes"),c("Neither","One", "Both"))
names(dimnames(table12)) <- c("Sexuality","HIV.HCV")
table12
chisq.test(table12)
#pvalue .01576 - sexuality
```

```
table13 <- table(inject, HIV.HCV)
dimnames(table13) <- list(c("No", "Yes"),c("Neither","One", "Both"))
names(dimnames(table13)) <- c("Did they inject?","HIV.HCV")
table13
chisq.test(table13)
#pvalue < 2.2e-16 - injection

table14 <- table(manner, HIV.HCV)
dimnames(table14) <- list(c("Mixed", "By Mouth",
"Injected"),c("Neither","One", "Both"))
names(dimnames(table14)) <- c("Manner of Use","HIV.HCV")
table14
chisq.test(table14)
#pvalue 2.2e-16 - manner of use

#logistic regression for HIV using forward stepwise regression
data<-na.exclude(thesis)
data$HIV.HCV <- NULL
data$HCV <- NULL
data$HCVHIV <- NULL
data$Shared.3.months. <- NULL
min.model <- glm(HIV ~ 1, data=data, family = binomial())
summary(min.model)
model <- glm(HIV ~ ., data = data)
summary(model)
step(min.model, scope=list(lower=min.model, upper=model),
direction="forward", data = data)
final2 <- glm(HIV ~ Education + inject.drug + marital.status + duration, data
= thesis, family = binomial())
summary(final2)
#observations = 3495
#odds ratio
exp(coef(final2))

#logistic regression for HCV using forward stepwise regression
data<-na.exclude(thesis)
data$HIV.HCV <- NULL
data$HIV <- NULL
data$HCVHIV <- NULL
data$Shared.3.months. <- NULL
min.model <- glm(HCV ~ 1, data=data)
summary(min.model)
model <- glm(HCV ~ ., data = data)
summary(model)
step(min.model, scope=list(lower=min.model, upper=model),
direction="forward", data = data)
final3 <- glm(HCV ~ rehabilitation + Education, data = thesis)
summary(final3)
#observations = 3246
#odds ratio
exp(coef(final3))
```