Homework 08-10

Christina Lee

11/13/2021

1. Describe your substantive interest and the general questions(s) you would like to answer (eg, "Does more education cause people to become more liberal?"). Be sure to frame it in a such a way that you are proposing a hypothesis (or multiple hypotheses) that might be either confirmed or disproven by the results of your analysis.

Here, I have a data set that looked at social factors, economic factors, immunization factors and other health related factors such as alcohol consumption that might affect overall life expectancy. The data was collected from 2000 to 2015 for all countries (182 total). Since our data set should not be temporal data, I have decided to focus on 2014 in particular (data for this year is more complete). I have also picked out specific variables that I wish to examine, which I have created an entire new data set with just data collected from 2014 for all countries and with just the variables that I picked out.

A few hypothesis that I am interested in answering are:
H1: "Does higher immunization coverage of a specific vaccine at birth results in a higher life expectancy? Or does immunization of all vaccines produce the same effect on life expectancy?"
H2: "Does higher Human Development Index in terms of Income Composition of Resources results in higher life expectancy rate?"
H3: If there is, in fact, a specific vaccine that greatly influences life expectancy than the rest of the vaccine types, I would then ask "Does the combination of higher immunization coverage of a specific vaccine and Human Development Index results in a significantly higher life expectancy rate?"
H4: "Does all chosen predicting variables really affect life expectancy at all?"
H5: "Among all predicting variable, which has the most impact on life expectancy variation?"

2. Describe the data set you have found, including its source, its contents, and why it was collected originally.

The data set was shared and found on Kaggle. In addition, the data set for health factors is originally created by The Global Health Observatory (GHO) data repository under World Health Organization (WHO), which they keep track of the health status and multiple related factors for all country population. It is a public data set that can be used by anyone for health analysis and other purposes. The economic factors are collected from United Nation website.

3. What is your dependent variable? Why are you interested in explaining it? What do you hypothesize are the major factors that influence or cause it?

The dependent variable in this study is life expectancy. The reasons why I am interested in explaining life expectancy in this study is because, although there are studies that looked at the social, economic, mortality factors and other health related factors such as alcohol consumption, etc. on life expectancy, however, immunization coverage and human development index was rarely – or never, taken into consideration in the past. Therefore, it is interesting to see how they each might affect life expectancy when compared individually, when both are compared together with life expectancy, or when compared with other independent variables to examine mixed effects on life expectancy.

4. What are your independent variables, and why have you chosen these? Prior to running your regression, what effects do you expect them to have on the dependent variable? Which of these variables do you

think affect other of the independent variables, and how might that affect your final results?

The independent variables in this study are 1) Status (developed/developing), 2) Measles (number of reported cases per 1000 population), 3) Schooling (total years), 4) Income Composition of Resources (range 0-1), 5) Alcohol (liters of pure alcohol consumption per capita per year), 6) Population, 7) Hepatitis B (immunization coverage among 1-year-olds (%)), 8) Polio (immunization coverage among 1-year-olds (%)), 9) Diphtheria (Diphtheria, tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)).

As mentioned in Q3, I am interested in examining how important immunization coverage (Hepatitis B, Polio and Diphtheria immunizations), and Human Development Index in terms of Income Composition of Resources may affect life expectancy, which I am also interested in seeing how these variables may interact with social and economic factors –it is why I included variables such as schooling, status and population.

As for how might variables affect other independent variables. I personally think higher immunization coverage, Human Development Index in terms of Income Composition of Resources, and more schooling, will result in higher life expectancy rate. And developed countries will have higher life expectancy.

I think alcohol consumption has more to do with country culture (doesn't interact much with other independent variables). However, higher alcohol consumption may, perhaps, result in lower schooling and higher Human Development Index in terms of income composition of resources (more money to purchase and drink alcohol) and a lower life expectancy rate.

Since I've created this data set from scratch by manually entering each variable and data, this data set is already organized and ready to be loaded.

Load the data set:

```
LifeExpectancy_2014 <- read.csv("Life_Expectancy_2014_DataSet.csv", sep = ",",
                                header = T, stringsAsFactors = F )

# save the data set as a data frame
LifeExpectancy_2014DF <- data.frame(LifeExpectancy_2014)
```

5. Explain and show in detail how you rename and recode the variables you are examining, and what units each are measured in.

Variables (units):
Country -> country
Year -> year
Status -> Developed or Developing status
Life Expectancy -> life expectancy in age
Measles -> number of reported cases per 1000 population
Schooling -> number of years of schooling
Income Composition of Resources -> Human Development Index in terms of income composition resources (index ranging from 0 to 1)
Alcohol -> recorded per capita (15+) consumption (in liters of pure alcohol) per year
Population -> total population of country
Hepatitis B -> Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
Polio -> Polio (Pol3) immunization coverage among 1-year-olds (%)
Diphtheria -> Diphtheria, tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)

For easier analysis I've decided to recode the only categorical variable, which is "Status" to numeric binaries of 1 = developed,0 = developing.

Next, there are a few missing values for some of the independent variables which are labeled as NA ("not available"). Hence, I have decided to replace the NA value for each independent variable with calculated median of that independent variable (EX: Schooling). For example, for all NA values under "Schooling", I will calculate the median value for Schooling by going through the collected data from all countries in 2014 and replace all NAs in Schooling with that number, which is 13.

```
LifeExpectancy_2014DF$Status <- as.factor(LifeExpectancy_2014DF$Status) # convert Status into factors
LifeExpectancy_2014DF$Status <- as.numeric(LifeExpectancy_2014DF$Status) # convert Status into numeric
LifeExpectancy_2014DF$Status[LifeExpectancy_2014DF$Status == 2] <- 0 # make binary 0 to developing
LifeExpectancy_2014DF$Status[LifeExpectancy_2014DF$Status == 1] <- 1 # make binary 1 to developed


LifeExpectancy_2014DF$Schooling[(is.na(LifeExpectancy_2014DF$Schooling))] <- 13 # NA to median

LifeExpectancy_2014DF$Income.Composition.of.Resources[
  (is.na(LifeExpectancy_2014DF$Income.Composition.of.Resources))]<- 0.7305 # NA to median

LifeExpectancy_2014DF$Alcohol[(is.na(LifeExpectancy_2014DF$Alcohol))] <- 0.26 # NA to median

LifeExpectancy_2014DF$Hepatitis.B[(is.na(LifeExpectancy_2014DF$Hepatitis.B))] <- 93 # NA to median
LifeExpectancy_2014DF$Hepatitis.B <-
  as.integer(LifeExpectancy_2014DF$Hepatitis.B) # convert Hep B into integer


#LifeExpectancy_2014DF
```

6. Before running a multiple regression, run a few bivariate regressions of Y on some of your X variables. What do you infer? Which of these do you think might change with the addition of multiple variables?

Based on the hypothesis that I am interested in testing, I will run my bivariate regressions as follows to examine their individual affects on life expectancy to better compare the changes and shifts in coefficients when I run the full multiple regression:
1st bivariate regression: regress income composition of resources on life expectancy
2nd bivariate regression: regress Hep B on life expectancy
3rd bivariate regression: regress Polio on life expectancy
4th bivariate regression: regress Diphtheria on life expectancy
5th bivariate regression: regress Schooling on life expectancy
6th bivariate regression: regress Alcohol on life expectancy

Run 1st bivariate regression (regress Income Composition of Resources on Life Expectancy):

```
bv1 <- lm(Life.Expectancy ~ Income.Composition.of.Resources, data =LifeExpectancy_2014DF)

summary(bv1)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ Income.Composition.of.Resources,
##     data = LifeExpectancy_2014DF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
```

```
## -18.6175  -1.9371  -0.0592   2.2480  10.2685
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      36.462      1.395   26.14   <2e-16 ***
## Income.Composition.of.Resources  49.905      1.945   25.66   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.953 on 180 degrees of freedom
## Multiple R-squared:  0.7853, Adjusted R-squared:  0.7841
## F-statistic: 658.3 on 1 and 180 DF,  p-value: < 2.2e-16
```

Run 2nd bivariate regression (regress Hep B on Life Expectancy):

```
bv2 <- lm(Life.Expectancy ~ Hepatitis.B, data =LifeExpectancy_2014DF)

summary(bv2)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ Hepatitis.B, data = LifeExpectancy_2014DF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.276  -4.992   1.190   4.543  19.257
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 61.02019    2.26339  26.960  < 2e-16 ***
## Hepatitis.B  0.12477    0.02612   4.777 3.67e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.036 on 180 degrees of freedom
## Multiple R-squared:  0.1125, Adjusted R-squared:  0.1076
## F-statistic: 22.82 on 1 and 180 DF,  p-value: 3.675e-06
```

Run 3rd bivariate regression (regress Polio on Life Expectancy):

```
bv3 <- lm(Life.Expectancy ~ Polio, data =LifeExpectancy_2014DF)

summary(bv3)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ Polio, data = LifeExpectancy_2014DF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -22.9628  -5.5099   0.6316   4.5307  24.8316
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 56.03937    2.48733  22.530  < 2e-16 ***
## Polio        0.18101    0.02843   6.366 1.56e-09 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.707 on 180 degrees of freedom
## Multiple R-squared:  0.1838, Adjusted R-squared:  0.1793
## F-statistic: 40.53 on 1 and 180 DF,  p-value: 1.561e-09
```

Run 4th bivariate regression (regress Diphtheria on Life Expectancy):

```
bv4 <- lm(Life.Expectancy ~ Diphtheria, data =LifeExpectancy_2014DF)

summary(bv4)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ Diphtheria, data = LifeExpectancy_2014DF)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.210  -5.117   1.098   4.686  21.412
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 59.84519    2.21863  26.974  < 2e-16 ***
## Diphtheria   0.13814    0.02547   5.424 1.86e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.908 on 180 degrees of freedom
## Multiple R-squared:  0.1405, Adjusted R-squared:  0.1357
## F-statistic: 29.42 on 1 and 180 DF,  p-value: 1.858e-07
```

Run 5th bivariate regression (regress Schooling on Life Expectancy):

```
bv5 <- lm(Life.Expectancy ~ Schooling, data =LifeExpectancy_2014DF)

summary(bv5)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ Schooling, data = LifeExpectancy_2014DF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.4380  -2.8027   0.4028   3.5487  11.2745
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.4926     1.7127   24.23   <2e-16 ***
## Schooling     2.3266     0.1297   17.94   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.11 on 180 degrees of freedom
## Multiple R-squared:  0.6412, Adjusted R-squared:  0.6392
## F-statistic: 321.7 on 1 and 180 DF,  p-value: < 2.2e-16
```

Run 6th bivariate regression (regress alcohol on life expectancy):

```
bv6 <- lm(Life.Expectancy ~ Alcohol, data =LifeExpectancy_2014DF)

summary(bv6)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ Alcohol, data = LifeExpectancy_2014DF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -25.2067  -4.4134   0.5669   5.5814  15.4669
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  68.0225     0.6867   99.05  < 2e-16 ***
## Alcohol       1.0665     0.1315    8.11 7.57e-14 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.3 on 180 degrees of freedom
## Multiple R-squared:  0.2676, Adjusted R-squared:  0.2636
## F-statistic: 65.78 on 1 and 180 DF,  p-value: 7.573e-14
```

1st bivariate:

I would infer that, income composition of resources has a positive effect on life expectancy and the effect is big (p value is extremely small) — by increasing 1 unit of income composition of resources would lead to an increase of 49.905 years of age towards life expectancy in average. However, you don't know if there is a chained causation between income composition of resources, life expectancy or another variable (X2). Such that, by adding X2 into the model the affect X1 (income compos.) may disappear or diminish, but we will have to test it out to see whether that is true. Moreover, the R squared is fairly close to 1 (0.7853); many of the variation in life expectancy is explained by income composition of resources.

2nd bivariate:

I would infer that, the percentage of Hep B immunization among 1 year olds do have a positive effect on life expectancy, and the effect is fairly significant (p value is very small). Such that, by increasing 1 unit of Hep B leads to an increase of 0.12477 years of age to life expectancy. However, the shift is tiny for each increase in Hep B immunization percentage among 1 year olds. In addition, the R squared is not that close to 1 (0.1125); not much variation in life expectancy is explained by Hep B.

It would be interesting to see the changes in the effect of Hep B and the shift in life expectancy if we run a multiple regression with all the important immunizations (combination of Hep B, Polio and Diphtheria) on life expectancy. However, for this to be tested correctly, it is beneficial to make sure that these independent variables not too "similar" or too correlated with each other (not measuring the same effect, because if so we can just combine, average and summarize them as "the total effect of immunization", which there is no need to spend time on examining their individual effects on life expectancy variation individually on a multiple regression analysis).

3rd bivariate:

I would infer that, the percentage of Polio immunization among 1 year olds have a positive effect on life

expectancy, and the effect is fairly significant (p value is very small). Such that, by increasing 1 unit of Polio leads to an increase of 0.18101 years of age to life expectancy. However, as with Hep B, the shift is tiny for each increase in Polio immunization percentage among 1 year olds. In addition, the R squared is not close to 1 (0.1838), which not much variation in life expectancy can be explained by Polio immunization percentage among 1-year-olds.

4th bivariate:

I would infer that, the percentage of Diphtheria immunization among 1 year olds have a positive effect on life expectancy, and the effect is fairly significant (p value is very small). Such that, by increasing 1 unit of Polio leads to an increase of 0.13814 years of age to life expectancy. However, as with Hep B, the shift is tiny for each unit increase in Diphtheria immunization percentage among 1 year olds. In addition, the R squared is not close to 1 (0.1405), which not much variation in life expectancy can be explained by Diphtheria immunization percentage among 1-year-olds.

5th bivariate:

I would infer that, the average year of schooling have a positive effect on life expectancy, and the effect is significant (p value is very small). Such that, by increasing 1 unit of Schooling leads to an increase of 2.3266 years of age to life expectancy. This shift is pretty big for each unit increase in years of schooling on life expectancy. In addition, the R squared is fairly close to 1 (0.6412); a fair amount of variation in life expectancy can be explained by Schooling.

6th bivariate:

I would infer that, the total alcohol consumption (in liters of pure alcohol) per year (2014 in our case) per capita have a positive effect on life expectancy, and the effect is significant (p value is very small). Such that, by increasing 1 unit of Alcohol leads to an increase of 1.0665 years of age to life expectancy. This result is pretty counter-intuitive, and one may ask why by drinking more alcohol can lead to longer life expectancy? It may be that the richer the country/citizens, the more alcohol they can afford to drink, but also the richer the country/citizens the longer the life expectancy. Therefore, it is not that drinking more alcohol can increase life expectancy, but that income composition of resources (we will use this index since GDP is not part of our model), causes both. Thus, to examine whether this theory is true, we can regress income composition of resources and alcohol on life expectancy to see if the coefficient for alcohol is still positive.

In addition, the R squared is not that close to 1 (0.2676); not much of the variation in life expectancy can be explained by alcohol.

7. Run your full multiple regression using lm() and present your results using the output from the stargazer R package. Interpret the coefficients. What do they tell you substantively? Which variables seem to have the biggest substantive impact? Which ones could you actually change with some intervention, and how big a difference do you think that could make?

```
library(stargazer) # call stargazer
```

```
##
## Please cite as:
##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.
##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```
multi_regression <- lm(Life.Expectancy ~ Income.Composition.of.Resources +
                         Hepatitis.B + Polio + Diphtheria + Schooling +
                         Status + Alcohol, data =LifeExpectancy_2014DF)
```

```
stargazer(multi_regression, no.space=TRUE, header = FALSE, type ='latex')
```

Table 1:

	Dependent variable:
	Life.Expectancy
Income.Composition.of.Resources	51.748***
	(5.058)
Hepatitis.B	0.024
	(0.023)
Polio	0.040*
	(0.021)
Diphtheria	−0.008
	(0.025)
Schooling	−0.258
	(0.252)
Status	0.405
	(0.962)
Alcohol	−0.024
	(0.090)
Constant	33.790***
	(1.860)
Observations	182
R^2	0.800
Adjusted R^2	0.792
Residual Std. Error	3.880 (df = 174)
F Statistic	99.413*** (df = 7; 174)
Note:	*p<0.1; **p<0.05; ***p<0.01

```
#summary (multi_regression)
```

Looking at the results for the multiple regression, it tells me that:

1. Income composition of resources has a positive effect on life expectancy and the effect is significant (p value is very small). By increasing 1 unit of income composition of resources index will lead to an increase of 51.748 in age towards life expectancy. A 1 unit increase in the index causes a big shift/increase in life expectancy.

2. Hep B immunization percentage among 1 year olds has a positive effect on life expectancy and the effect is not significant (p value is > 0.05). By increasing 1 unit of Hep B will lead to an increase of 0.024028 in age towards life expectancy. A 1 unit increase in the Hep B immunization percentage among 1 year olds causes a small shift/increase in life expectancy.

3. Polio immunization percentage among 1 year olds has a positive effect on life expectancy and the effect is close to being somewhat significant but still considered as insignificant (p value is 0.0558). By increasing 1 unit of Polio will lead to an increase of 0.039960 in age towards life expectancy. A 1 unit increase in the Polio immunization percentage among 1 year olds causes a small shift/increase in life expectancy.

4. Diphtheria immunization percentage among 1 year olds has a negative effect on life expectancy and the effect is not significant (p value is > 0.05). By increasing 1 unit of Diphtheria will lead to a decrease of

0.008325 in age towards life expectancy. A 1 unit increase in the Diphtheria immunization percentage among 1 year olds causes an extremely small shift/decrease in life expectancy.

5. Average year of schooling has a negative effect on life expectancy and the effect is not significant (p value is > 0.05). By increasing 1 unit of Schooling will lead to a decrease of 0.258160 in age towards life expectancy. A 1 unit increase in average years of schooling will cause a small shift/decrease in life expectancy.

6. The status (whether it is developed or developing) of a country has a positive effect on life expectancy and the effect is not significant (p value is > 0.05). By increasing 1 unit of Status (going from developing to developed) will lead to an increase of 0.405168 in age towards life expectancy. A 1 unit increase in Status will cause a small shift/increase in life expectancy.

7. The total alcohol consumption (in liters of pure alcohol) per year (2014 in our case) per capita have a negative effect on life expectancy and the effect is not significant (p value is > 0.05). By increasing 1 unit of Alcohol will lead to a decrease of 0.024032 in age towards life expectancy. A 1 unit increase in Alcohol will cause an extremely small shift/ decrease in life expectancy.

Furthermore, the R squared is fairly close to 1 (0.8); majority of the variation in life expectancy can be explained by the model. It seems like income composition of resources has the most impact on life expectancy since a 1 unit increase in the index will, apparently, cause an increase of 51.748087 in age towards life expectancy. It is also the only variable with a significant effect in the model.

Variables that I might be able to change with some interventions are, for example, immunization coverage of important vaccines (Hep B, Polio, Measles and Diphtheria). To do that, countries may impose stricter immunization mandates in 1-year olds and provide free vaccines as well in poorer countries to prevent illnesses that may contribute to shorter lifespan to a country's average life expectancy.

8. How have any of the coefficients changed from the bivariate regressions? What can you infer from that? How do you think your various independent variables interact and affect each other? Try to find an example where a variable appears signficant in the bivariate regression, but not in the full regression. Is this an example of a spurious or a chained causal pathway?

The coefficients that changed significantly from the bivariate regression and is worth mentioning are the coefficients for:

Diphtheria: The coefficient for Diphtheria in the bivariate regression model displayed a positive and significant effect on life expectancy, however, it ended up displaying a negative and insignificant effect on life expectancy in the full regression model. This is not the case with the two other immunization variables (Hep B and Polio) although both variable's effects on life expectancy became insignificant in the full regression model. My guess is that, Diphtheria may be displaying its strong positive effect through some other variable in the full regression model, which the only to find out is to remove one variable at a time, until Diphtheria's coefficient finally retrieves its significant & positive effect on life expectancy. To figure out why Hep B and Polio's positive effect on life expectancy became insignificant, we can also utilize the same method as well.

Schooling: The coefficient for Schooling in the bivariate regression model displayed a positive and significant effect on life expectancy, but displayed a negative and insignificant effect on life expectancy in the full regression model. My guess for this change in effect and significance, which is pretty counter-intuitive is that, its strong impact may be going through some other variables such as income composition of resources. To test this theory, we can remove income composition of resources from the model and see if the coefficient for Schooling becomes positive and significant.

Alcohol: The coefficient for Alcohol in the bivariate regression model displayed a positive and significant effect on life expectancy, but displayed a negative and insignificant effect on life expectancy in the full regression model, which makes much more sense. I am guessing that the reason for this change in Alcohol's coefficient is due to the omitted income composition of resources in the bivariate model. Hence, it may be that the richer the country as a whole/citizens, the more alcohol they can afford to drink, but also the richer the country/citizens the longer the life expectancy. So, it is not that drinking causes increased life expectancy,

but that income composition of resources causes both, which results a spurious relationship between the variables.

9. How does what you see match, or not, your hypotheses from (4)? Why did/didn't it match what you expected?

Hypothesis or some thoughts from Q4 that did match with the full regression model is, as income composition of resources increases, life expectancy goes up significantly as well. Going from a still developing to a developed country also displayed a positive effect although the effect is not significant.

In addition, hypothesis that does not match with the model is, higher alcohol consumption leads to higher income composition of resources and lower life expectancy, since it is income composition of resources that is driving both higher alcohol consumption and increase in life expectancy. Another hypothesis that does not match with the model is, higher immunization coverage (Hep B, Polio, and Diphtheria) will increase life expectancy, since some variables did display a positive effect but they are insignificant and some variables also displayed a negative and insignificant effect on life expectancy in the full regression (opposite from the bivariate regression). My hypothesis for Schooling also does not match with the full regression model, since it display a negative and insignificant effect on life expectancy.

10. What do the R2 and adjusted R2 tell you about your model?

R squared is fairly close to 1 (0.800); this indicates that ~80% variation in life expectancy can be explained by the model or we have reduced the error by ~80%.

As for adjusted R squared (0.7919), it got smaller but it is not much smaller, reflecting the fact that most of the variables included in the model do have explanatory power.

11. How would you use one of the variable selection methods to choose a model with fewer variables? Select one of the methods (either one of the stepwise or criterion-based methods) and show which variables it would lead you to keep. Do you agree with its results?

I would use stepwise method, where I start by adding variables with low p-value and drop those whose p value rise to high. This method leaves me with income composition of resources and Diphtheria. I wouldn't say that the variables that I am left with does not make sense, but it does omit important variables such as Schooling, Alcohol and Status. In addition, perhaps Diphtheria is enough to represent the effect of immunization coverage which is why I am not too concerned if Hep B and Polio is omitted from the final model.

```
multi_regression_Step <- lm(Life.Expectancy ~ Income.Composition.of.Resources +
                          Diphtheria, data =LifeExpectancy_2014DF)

summary(multi_regression_Step)  # stepwise model
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ Income.Composition.of.Resources +
##     Diphtheria, data = LifeExpectancy_2014DF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -17.1407  -1.9104  -0.0579   2.4408  10.0440
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                     34.81129    1.51821  22.929   <2e-16 ***
## Income.Composition.of.Resources 48.19123    2.02988  23.741   <2e-16 ***
## Diphtheria                       0.03395    0.01328   2.555   0.0114 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.894 on 179 degrees of freedom
## Multiple R-squared:  0.7928, Adjusted R-squared:  0.7905
## F-statistic: 342.5 on 2 and 179 DF,  p-value: < 2.2e-16
```

12. What are your overall conclusions? What are the weaknesses of your results, and how could you improve them with better or different data?

What I can conclude from this model is that, only income composition of resources displayed a strong positive and significant effect (p value is <2e-16 and is very small) on life expectancy, with Polio being "almost" somewhat significant (p value is 0.0558). The rest of the variable all showed an insignificant effect on life expectancy whether it is positive or negative in the final model. I am guessing that the many of the variables (not all) are effecting life expectancy through income composition of resources, or other variables. To test this theory, we can try to remove income composition of resources from the final model and see if any of the variables will increase their significance.

In addition, R squared appears to be 0.8 (very close to 1). This may indicate that ~80% variation in life expectancy can be explained by the model and we have reduced error by 80%. Adjusted R squared did not shift too much (0.7919) from R squared as well, which may suggest that the variables in the model do have a large explanatory power.

In a greater scope, this means income composition of resources, out of all other factors, is a vital aspect to focus on if countries want to improve the overall life expectancy of their country since it does have a positive and significant effect on life expectancy.

The weakness of my results is that, most of the variables displayed an insignificant effect in the full regression model. This may indicate that there is an issue with variable selection to the multiple regression model. Hence, in order in improve variable selection methods such as backward elimination, forward elimination or Criterion-based approaches may be utilized, though they each have their own drawbacks and weaknesses as well.

13. Calculations (using R):

    a. Derive the coefficients from your regression using the (X X)−1X Y formula. (If you run into problems using solve(), try using ginv() instead, which does the same thing but is a bit more robust.)

```
LifeExpectancy_matrix <- as.matrix (cbind(LifeExpectancy_2014DF$Status, LifeExpectancy_2014DF$Hepatitis
LifeExpectancy_2014DF$Income.Composition.of.Resources))

LifeExpectancy_matrix <- cbind(1,LifeExpectancy_matrix)
#head(LifeExpectancy_matrix)

solve (t(LifeExpectancy_matrix) %*% LifeExpectancy_matrix) %*%
  t(LifeExpectancy_matrix) %*% LifeExpectancy_2014DF$Life.Expectancy
```

```
##              [,1]
## [1,] 33.790405169
## [2,]  0.405168394
## [3,]  0.024027753
## [4,]  0.039960204
## [5,] -0.008325322
## [6,] -0.258160334
## [7,] -0.024031547
## [8,] 51.748086980
```

    b. For one of the coefficients, confirm its p value as shown in the regression output using the coefficient, its standard error, and pt() in R.

```
2* (1- pt(1.064,180)) # p value for Hep B
```

```
## [1] 0.288754
```

c. Calculate the R2 and adjusted R2 using R, and confirm that your results match the regression output.

```
ypred <- predict(multi_regression)

y <- LifeExpectancy_2014DF$Life.Expectancy

tss <- sum((y - mean(y))^2)
sse <- sum((y - ypred)^2)
r2 <- (tss-sse)/tss
r2
```

```
## [1] 0.7999744
```

```
n <- length(y)
k <- 7
dft <- n - 1
dfe <- n - k - 1
adj_r <- (tss/dft - sse/dfe) / (tss/dft)
adj_r
```

```
## [1] 0.7919274
```

d. Calculate the F statistic using R and confirm it against the regression output.

```
fstat <- (r2/k) / ((1-r2) / (n-k-1))
fstat  # F statistic
```

```
## [1] 99.41267
```

14. Add at least one quadratic term into your model and interpret the results. Is it significant? What is the effect of a 1-unit increase in that variable at its mean value?

```
multi_regressionQUAD <- lm(Life.Expectancy ~ Income.Composition.of.Resources +
                           Hepatitis.B + Polio + Diphtheria + Schooling +
                           Status + Alcohol + I(Income.Composition.of.Resources^2),
                           data =LifeExpectancy_2014DF)

summary (multi_regressionQUAD)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ Income.Composition.of.Resources +
##     Hepatitis.B + Polio + Diphtheria + Schooling + Status + Alcohol +
##     I(Income.Composition.of.Resources^2), data = LifeExpectancy_2014DF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.2661  -2.0785  -0.1935   2.6517  10.4284
##
## Coefficients:
##                                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)                       40.125018   6.551736   6.124 5.95e-09 ***
## Income.Composition.of.Resources   31.057432  21.133559   1.470   0.1435
## Hepatitis.B                        0.024027   0.022588   1.064   0.2889
```

```
## Polio                                  0.040730   0.020768   1.961   0.0515 .
## Diphtheria                            -0.008175   0.025439  -0.321   0.7483
## Schooling                             -0.254339   0.251885  -1.010   0.3140
## Status                                -0.239259   1.155244  -0.207   0.8362
## Alcohol                               -0.029630   0.089751  -0.330   0.7417
## I(Income.Composition.of.Resources^2) 15.891617  15.760106   1.008   0.3147
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.88 on 173 degrees of freedom
## Multiple R-squared:  0.8011, Adjusted R-squared:  0.7919
## F-statistic: 87.12 on 8 and 173 DF,  p-value: < 2.2e-16
```

I have decided to add a quadratic term on Income Composition of Resources. Since the coefficient on the x^2 term is positive, it means the curve is concave up and due to a p value that is > 0.05, there is no significant curved effect and the effect of income composition of resources is, at best, still linear. Therefore, a 1 unit change in income composition of resources will result in the same effect on life expectancy as seen in the original multiple regression in Q7.

15. Add at least one interaction term to you model and interpret the results. Is it significant? What is the effect of a 1-unit increase in one of those interacted variables holding the other at its mean value?

```
multi_regressionINTER <- lm(Life.Expectancy ~ Income.Composition.of.Resources +
                              Hepatitis.B + Polio + Diphtheria + Schooling +
                              Alcohol + Status +
                              Income.Composition.of.Resources * Status, data=LifeExpectancy_2014DF)

summary (multi_regressionINTER)
```

```
##
## Call:
## lm(formula = Life.Expectancy ~ Income.Composition.of.Resources +
##     Hepatitis.B + Polio + Diphtheria + Schooling + Alcohol +
##     Status + Income.Composition.of.Resources * Status, data = LifeExpectancy_2014DF)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -16.3908  -1.9043  -0.0308   2.6221  10.7420
##
## Coefficients:
##                                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)                             33.943684   1.868939  18.162   <2e-16
## Income.Composition.of.Resources         51.475399   5.069992  10.153   <2e-16
## Hepatitis.B                              0.023100   0.022626   1.021    0.309
## Polio                                    0.041058   0.020803   1.974    0.050
## Diphtheria                              -0.007112   0.025491  -0.279    0.781
## Schooling                               -0.265900   0.252160  -1.054    0.293
## Alcohol                                 -0.021464   0.089680  -0.239    0.811
## Status                                 -13.217643  15.249537  -0.867    0.387
## Income.Composition.of.Resources:Status  15.372978  17.174350   0.895    0.372
##
## (Intercept)                            ***
## Income.Composition.of.Resources        ***
## Hepatitis.B
## Polio                                  .
## Diphtheria
```

```
## Schooling
## Alcohol
## Status
## Income.Composition.of.Resources:Status
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.883 on 173 degrees of freedom
## Multiple R-squared:  0.8009, Adjusted R-squared:  0.7917
## F-statistic: 86.99 on 8 and 173 DF,  p-value: < 2.2e-16
```

Here I want to examine what is the effect of income composition of resources on life expectancy in terms of the Status of a country (developed and developing) ? To test this, I will add an interaction term between income composition of resources and Status. The results are not significant since p value is > 0.05.

For developed countries (note: developed = 1), the effect of income composition of resources on life expectancy is :

51.475399 + 15.372978 x 1 = 66.85

For developing countries (note: developing = 0), the effect of income composition of resources on life expectancy is:

51.475399 + 15.372978 x 0 = 51.475399

Therefore, for developed countries, 1 unit increase of income composition of resources will lead to an increase of 66.85 years of age to life expectancy, where 1 unit increase of income composition of resources will lead to an increase of 51.475399 years of age to life expectancy for developing countries.

16. Test either the model in 14 or the model in 15 using the F test for nested models. That is, estimate the full model with the variable and quadratic term, or the variable and interaction, and then estimate the reduced model without either, and run the F test to establish whether those variables significantly improve your model.

```
complete <- lm(Life.Expectancy ~ Income.Composition.of.Resources +
                                 Hepatitis.B + Polio + Diphtheria + Schooling +
                                 Alcohol + Status +
                                 Income.Composition.of.Resources * Status, data=LifeExpectancy_2014DF)


reduced <- lm(Life.Expectancy ~ Hepatitis.B + Polio + Diphtheria + Schooling +
                                 Alcohol + Status, data=LifeExpectancy_2014DF)

anova(reduced, complete)
```

```
## Analysis of Variance Table
##
## Model 1: Life.Expectancy ~ Hepatitis.B + Polio + Diphtheria + Schooling +
##     Alcohol + Status
## Model 2: Life.Expectancy ~ Income.Composition.of.Resources + Hepatitis.B +
##     Polio + Diphtheria + Schooling + Alcohol + Status + Income.Composition.of.Resources *
##     Status
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    175 4196.0
## 2    173 2607.8  2    1588.2 52.678 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Here, I would like to test whether the pair of variables from Q15, income composition of resources and income composition of resources * Status, belong in our full model by running an F test to determine whether those variables significantly improve my model.

$$H_a = Complete\ model\ does\ better\ than\ reduced\ model$$

$$H_o = Complete\ model\ does\ no\ better\ than\ reduced\ model$$

Result: Since p value is much less than 0.05, we reject the null. This shows that the income composition of resources and income composition of resources * Status terms definitely both belong in the regression and that the complete model with the these variables is significantly better than the reduced model in explaining our Y, life expectancy.