SEM_Example_Lavaan_Air Pollution Impacts on Vulnerable Communities in California

Devan Addison-Turner and Nicholas Camacho

2022-11-13

R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see http://rmarkdown.rstudio.com.

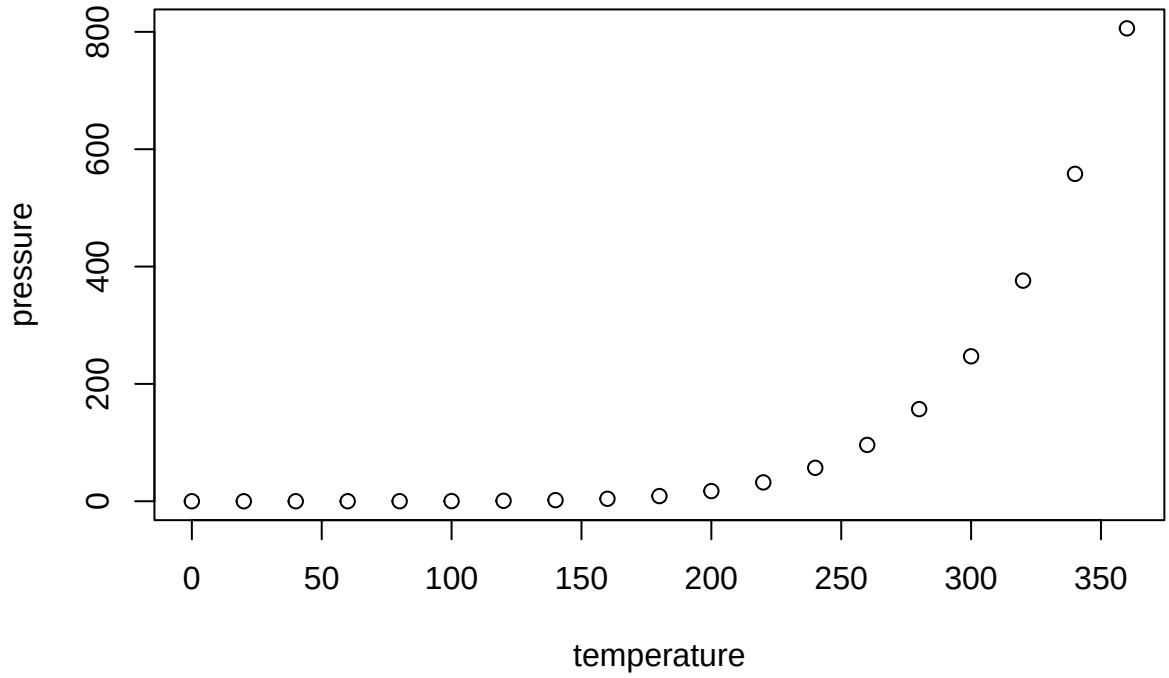
When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
summary(cars)
```

```
##      speed           dist       
##  Min.   : 4.0   Min.   :  2.00  
##  1st Qu.:12.0   1st Qu.: 26.00  
##  Median :15.0   Median : 36.00  
##  Mean   :15.4   Mean   : 42.98  
##  3rd Qu.:19.0   3rd Qu.: 56.00  
##  Max.   :25.0   Max.   :120.00
```

Including Plots

You can also embed plots, for example:



Note that the `echo = FALSE` parameter was added to the code chunk to prevent printing of the R code that generated the plot.

```
#Structural Equations Modeling (SEM) Example
#BIO 202 - Ecological Statistics
#Stanford University
#Fall 2022
#Instructors: Dr. Tad Fukami and Dr. Jesse Miller

#Project Background:
##Research Questions -
#1.Does Pollution impact Asthma?
#2.Does Pollution and Poverty impact Asthma?

#Location: The entire state of California

#Why is this study important? This study identifies metrics and techniques
#that are useful to quantify the risk and impacts of Pollution on
#vulnerable populations in California. Moreover, this study allows
#us to evaluate the causation of risks and impacts to help
#better inform decision-making to guide
#policy and investment in under-served and underrepresented communities.

#Data sample size (n = 8,035 observations)
#8,035 datapoints represent geogrpahic boundaries within California (area).
#Each census tract has approx. 4,200 people on average.
```

```
#Pro-Tips: Make sure to select the PDF format when creating a
#new RMarkdown file and install tinytex package to Knit file to PDF.
#Install tinytex package to Knit RMarkdown File to PDF.
#install.packages('tinytex')
#install.packages("tinytex", repos = "http://cran.us.r-project.org")
#tinytex::install_tinytex()
# to uninstall TinyTeX, run tinytex::uninstall_tinytex()
#Reference: https://yihui.org/tinytex/

#SEM Primary References:
#1. Grace, James B. 2006. Structural Equation Modeling
#and Natural Systems, Cambridge University Press
#2. Lefcheck, J. 2019. Structural Equation Modeling in R for Ecology and Evolution.


#Why Do SEM?
#SEM tests the direct and indirect effects on pre-assumed causal relationships:
#Enlisted by ecologists because
#1. Theory oriented
#2. Can analyze complex causal networks
#3. Permits causal inferences
#4. Model flexibility
#5. More intuitive interpretation

#Several conditions are deemed necessary, but not sufficient to establish
#causation.
#1. There is an empirical association between the variables - they
#are significantly correlated.
#2. A common cause of the two variables has been
#ruled out, and the two variables have a theoretical connection.
#3. One variable precedes the other, and if the preceding variable changes, the
#outcome variable also changes (and not vice versa).
#4. Even when some of the conditions of causation are not fully met,
#causal inference may still be justifiable

#Why did SEM work for us and our data?
#1. We have pre-assumed causal relationships between variables.
#2. Answering our research questions requires causal analysis.

#Why was SEM an appropriate choice for us?

#1. We want to understand the causal relationship between
#specific variables and we have an a priori hypothesis about this
#relationship. Furthermore, the data we have is suitable for causal analysis and SEM.

#2. Specifically, we are interested in the relationship between
#asthma, pollution, and poverty.

#3. Asthma was measured as the age-adjusted
#rate of emergency department visits for asthma.

#4. Pollution was measured as the average of percentiles from the Pollution Burden
#indicators, which is an aggregation of different pollution sources from
```

```
#buildings, transportation, and the natural environment.

#5. Poverty was measured as the percent of population living below two times
#the federal poverty level.

#6. All of our observed variables were measured as continuous data.

#Data structure is suitable for SEM

#Variables and Descriptions
#Variable: Asthma:
#Description: Age-adjusted rate of emergency department visits for asthma
#Data Source: California Office of Health Hazard Assessment Pollution
#Data Type Continuous

#Variable: Pollution
##Description: Average of percentiles from the Pollution Burden indicators
#(with a half weighting for the Environmental Effects indicators)
#Data Source: California Office of Health Hazard Assessment Pollution
#Data Type Continuous


#Variable: Poverty
#Description: Percent of population living below two times the federal
#(nationwide) poverty level of $18,755 annually
#Data Source: California Office of Health Hazard Assessment
#Data Type: Continuous

#Variable: EH_2022 = Extreme Heat
#Description: Number of extreme heat days in the year 2022
#Data Source: Cal-Adapt
#Data Type: Continuous

#Fundamentals of SEM
#The independent (x) variables are referred to as exogenous, while the
#dependent (y) variables are called endogenous.
#Endogenous variables can have influences on other endogenous variables.
#Commonly, causal order will flow from left to right across a model,
#although this will not always be the case.The ways in which variables in a
#model are connected are important.

#SEM is essentially a linear model framework that models regression equations
#with latent variables, and therefore it is possible to model the
#relationship between our variables of interest. To create our model in R,
#we chose to use the lavaan package, which stands for Latent variable analysis.
#This package is the most commonly used package for SEM, and in our experience,
#offers the most streamlined and intuitive process for creating your model and
#running goodness-of-fit tests. While we're not planning on getting into
#latent variables in this presentation, it is important to know that they are
#basically unobserved variables that we can construct from our observed data.
#We didn't use latent variables in our model, but it is possible to include
#them using lavaan.
```

```
#lavaan (LAtent VAriable ANalysis)
#Used for multivariate statistical modeling
#(i.e path analysis, confirmatory factor analysis, structural equation modeling)
#Streamlined formulas for SEM compared to other packages
#Creates a model and tests Goodness-of-fit

#2 types of variables:
#1. Observed variable (exists within the dataset)
#2. Latent Variable (Unobserved data created from existing variables

#The SEM process is composed of the following five steps and decisions:
#1. Construct a path diagram that shows the measurement and
#structural model of interest.
#2.Identify the level of measurement for each item and
#check distributional assumptions.
#3. Ensure that the fitting function you chose is based on
#measurement types (e.g., maximum likelihood for
#continuous measures, weighted least square for ordinal measures).
#4. Move through the model testing process in a logical fashion.
#5. Fit the model using the appropriate fitting function and carefully
#assess model fit by using a set of indexes.

#Checklist
#Part 1. Hypothesis
#1.1 Develop a hypothesis to test. For this study we used
#Research Questions
#1.Does Pollution impact Asthma?
#2.Does Pollution and Poverty impact Asthma?
#1.2 Establish paths of causal correlation between variables and
#the meaning for using these variables
#1.3 Select at least 3 variables that you believe have
#causal correlation to establish paths for influencing outcome
#1.4 Load in data
#1.5 Identify your observed or latent variables (unobserved or lagging)
#1.6 Data review to check for absent data, No NA's, no missing data points

#Part 2. Null/Baseline Model and User Model Evaluation
#2.1 There may be other ways to analyze categorical or ordinary data but we
#focused on using continuous
#2.2 This is how we applied SEM for our dataset composed of continuous data
#2.3 Model specification and create measurement models
#2.4 Choose the function to apply to SEM analysis such as linear regression
#2.5 Read in your data into a new model SEM_M1
#2.6 Make a new dataframe for SEM_M1
#2.7 Layering SEM_M2 and SEM_M3 over df1 similar to a filter
#2.8 What is the reason to include 1 in the models
#(Why would we want to change the intercept (mean)?)
#2.9 Paths of causal correlation are being tested
#2.10 Make sure there is no covariance testing inside measurement models
#(repetitive and throws off the model analysis) ()

#Part 3. Path Analysis
#3.1 Verify that your apriori assumption was correct when you first
```

```r
#constructed a path diagram prior to modeling.
#3.2 Look at the direct and/or indirect effects of causal correlation.
#3.3 Graph global p-value and R-Squared to assess quality of hypothesis
#and model tests

#install.package(lavaan) --- Latent Variable Analysis
#(however we did not use latent variables)
#only used observed variables for this study
#(main package to perform SEM analysis)
#install.package(sem) -- may interfere with lavaan package
#install.packages("semPaths") -- This function creates a path diagram
#of a SEM model (or general linear model), which is then plotted using qgraph
#install.packages("semptools") -- Provides tools for
#structural equation modeling, many of which extend the 'lavaan' pack- age;
#for example, to pool results from multiple imputations,
#probe latent interactions, or test measurement invariance.

#set working directory for file reference locations inside from my computer
setwd("~/Documents/GitHub/devanaddisonturner.github.io/BIO 202_Ecological Statistics/SEM Models")
#loads in Lavaan package
library(lavaan)
```

```
## This is lavaan 0.6-12
## lavaan is FREE software! Please report any bugs.
```

```r
#imports SEM_Example_Data csv. file from my working directory as SEM Model 1
SEM_M1 <- read.csv("SEM_Example_Data_California.csv")
#Make a new dataframe for SEM_M1
df1<-data.frame(SEM_M1)
df2<-na.omit(df1)

#Hypothesis Testing
#3 variables
#2 models using linear regression
#Continuous data

#Data Preparation
#SEM in lavaan is sensitive to NAs and continuous data
#Before setting up model, omit rows with missing data

#Syntax structure reference using Lavaan
#https://lavaan.ugent.be/tutorial/syntax1.html


#How do we include Extreme Heat in both M2 and M3 for future researcn?

##Baseline Model
  SEM_M2 <-   '
  # regressions
    Poverty ~ Pollution
  # Poverty ~ Pollution*EH_2022
  # Pollution ~ Poverty
  # Asthma ~ Pollution
```

```
    Asthma ~ Poverty
'
# Poverty ~~ Pollution
#SEM_M2_LM <- lm(Asthma ~ Poverty, data = SEM_M1)
SEM_M2_LM1 <- lm(Asthma ~ Pollution, data = SEM_M1)
#Pollution Burden is correlated with the
#rate of Asthma in California
#The rate of Asthma in California increases with Pollution Burden
summary(SEM_M2_LM1)
```

```
##
## Call:
## lm(formula = Asthma ~ Pollution, data = SEM_M1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -55.12 -20.97  -6.58  13.26 193.13
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 30.94979    1.16723   26.52   <2e-16 ***
## Pollution    0.49292    0.02621   18.80   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 29.92 on 8022 degrees of freedom
##   (11 observations deleted due to missingness)
## Multiple R-squared:  0.04221,    Adjusted R-squared:  0.04209
## F-statistic: 353.6 on 1 and 8022 DF,  p-value: < 2.2e-16
```

```
library(effects)
```

```
## Loading required package: carData
```

```
## lattice theme set by effectsTheme()
## See ?effectsTheme for details.
```

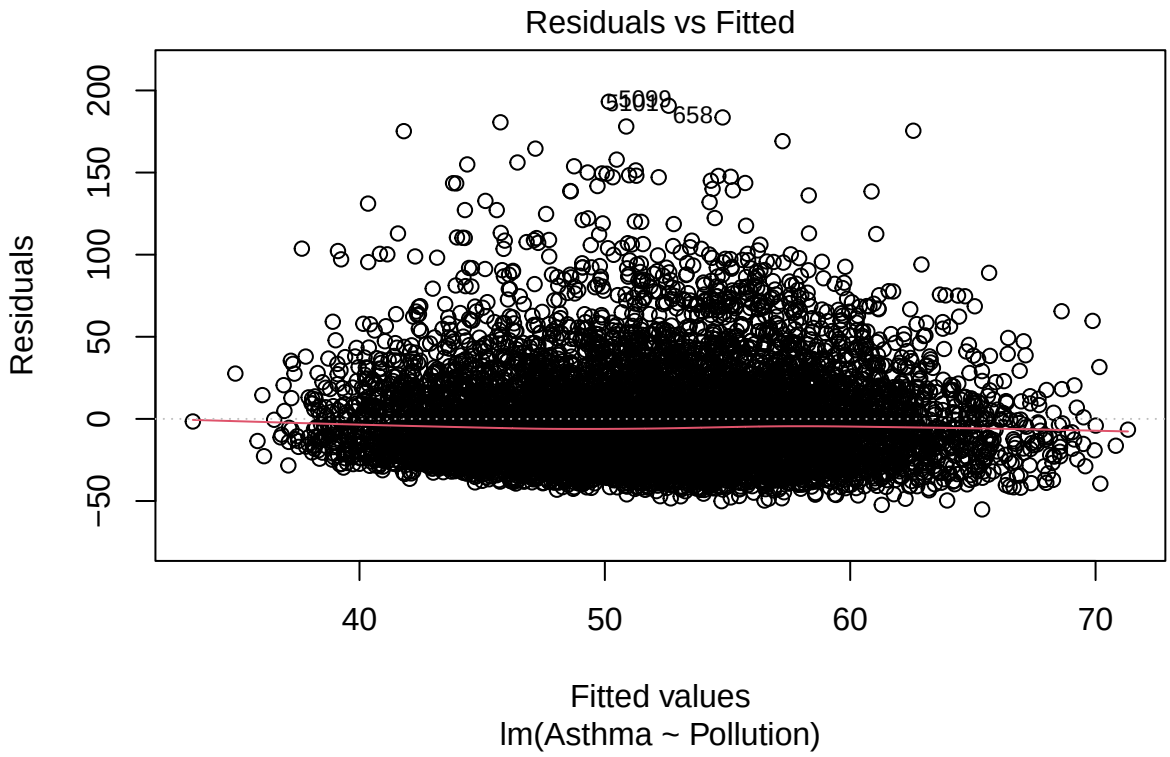
```
library(ggplot2)
library(dplyr)
```

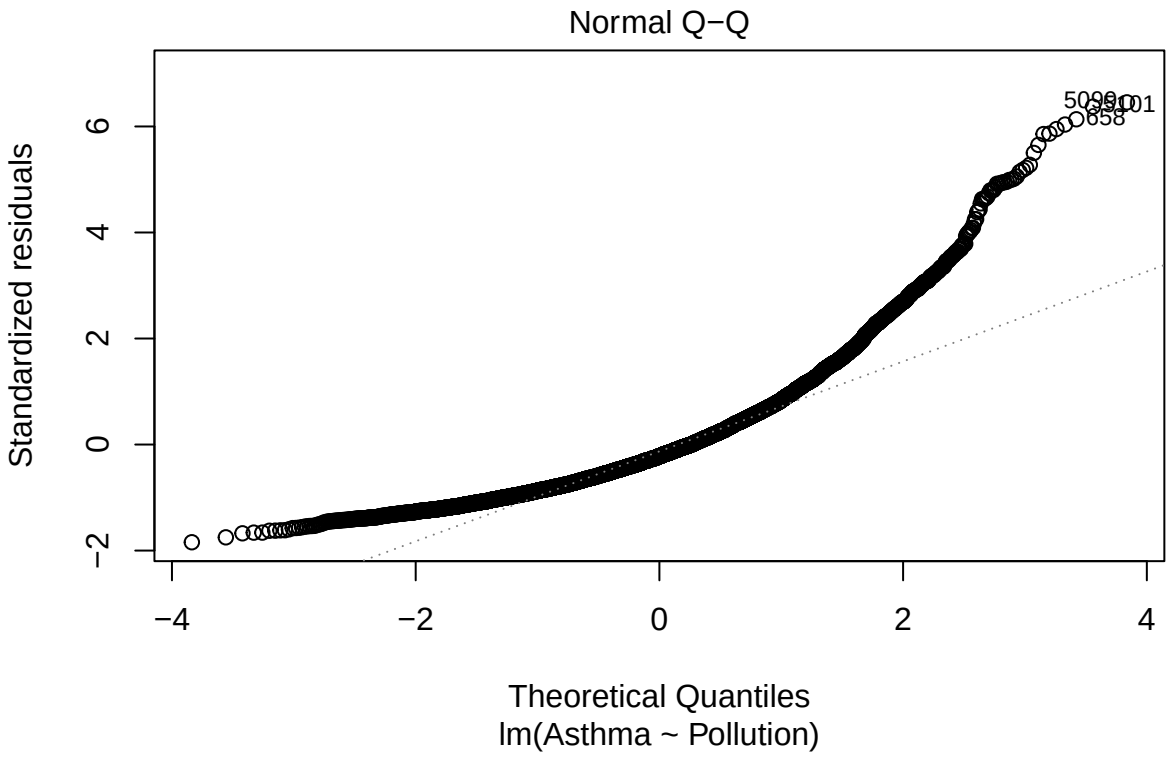
```
##
## Attaching package: 'dplyr'
```

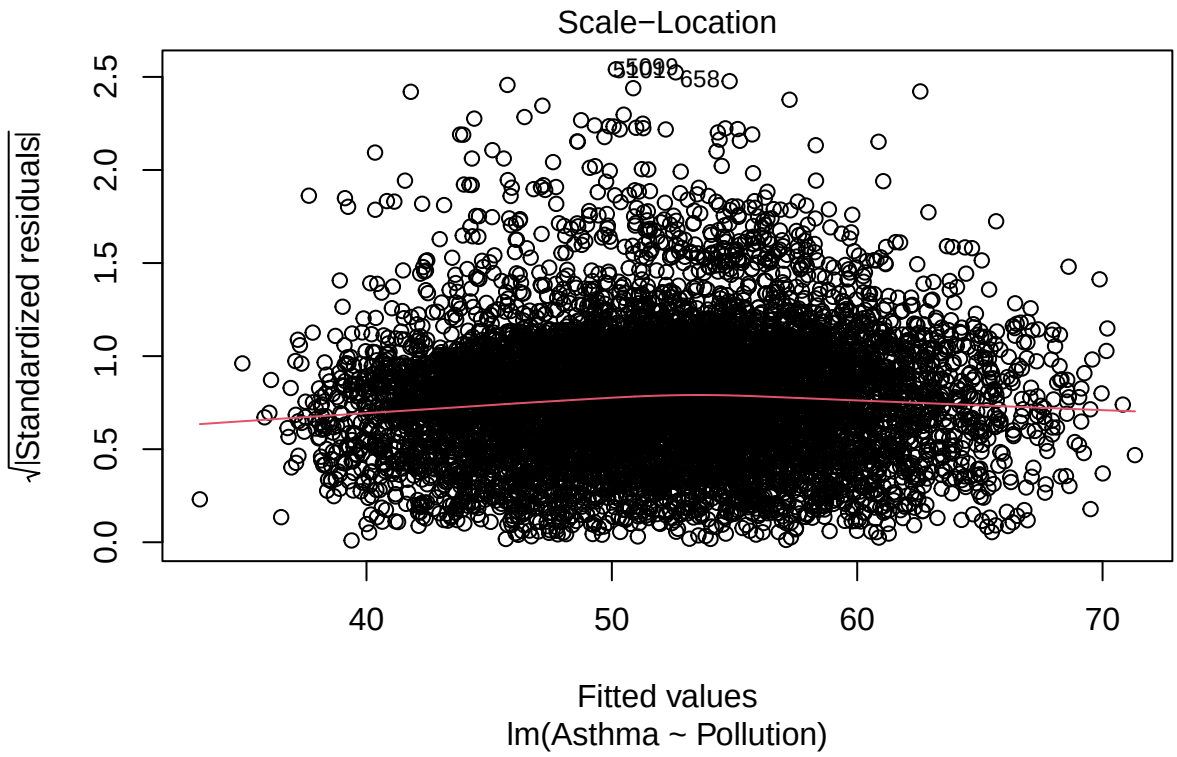
```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

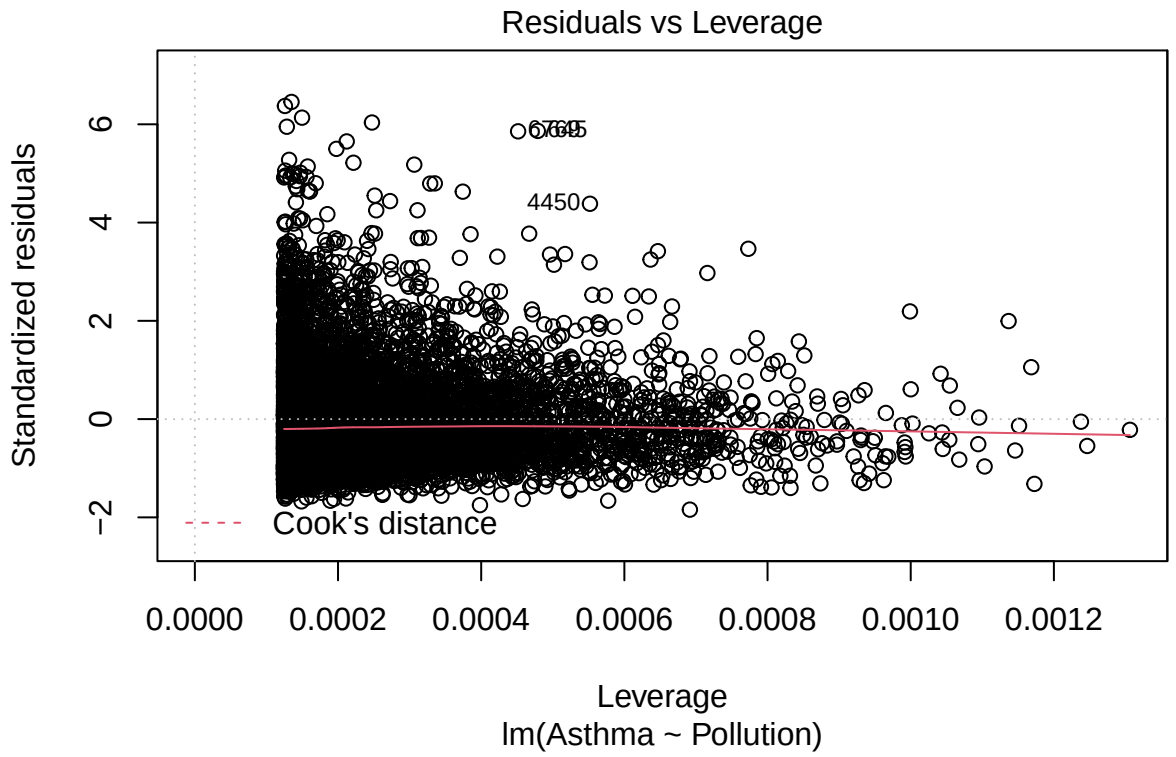
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
plot(lm(Asthma ~ Pollution, data = SEM_M1))
```

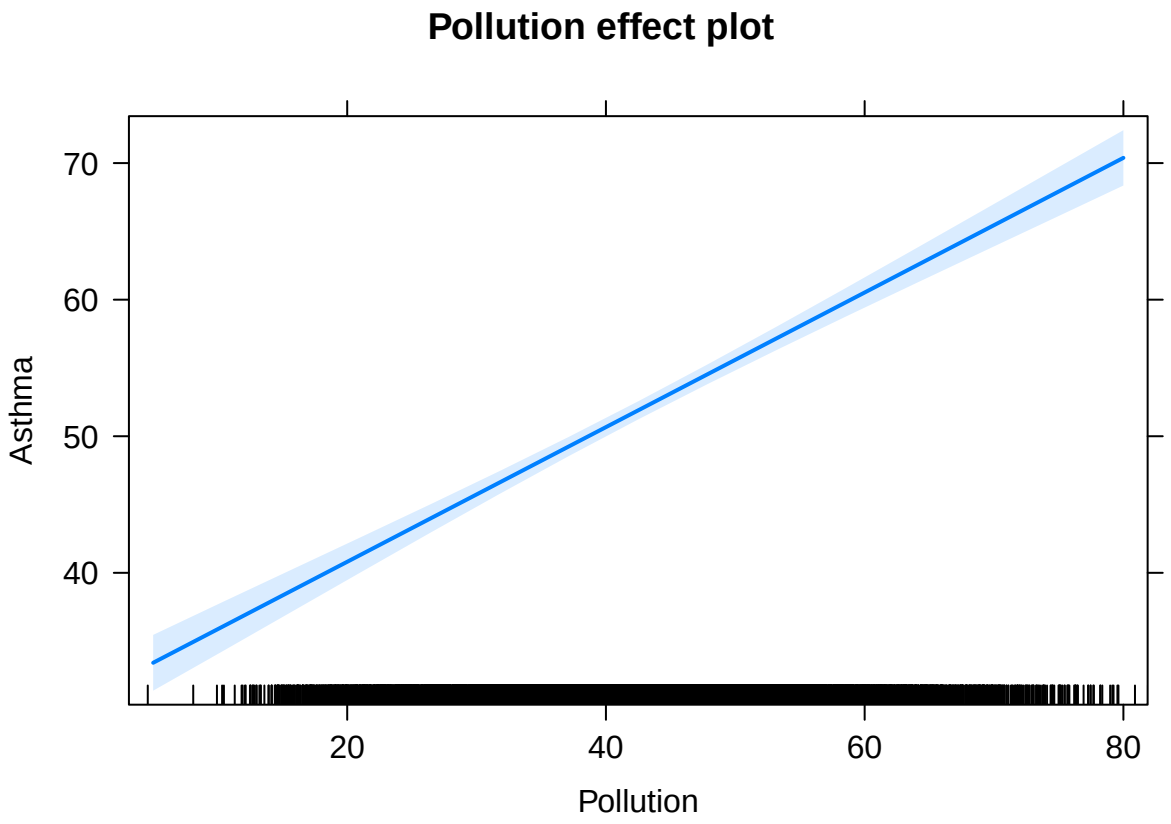



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Asthma ~ Pollution)
−4
−2
0
2
4
−2
0
2
4
6
5099
658


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
40
50
60
70
Fitted values
lm(Asthma ~ Pollution)




```
plot(allEffects(SEM_M2_LM1))
```




```
#call in sem function to be performed into new model that will be fitted
#from SEM analysis from SEM_M2
fit1b <- sem(SEM_M2, data=df2)   #Baseline Model
#fit the model then view output
summary(fit1b)
```

```
## lavaan 0.6-12 ended normally after 1 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         4
##
##   Number of observations                          7959
##
## Model Test User Model:
##
##   Test statistic                                 5.113
##   Degrees of freedom                                 1
##   P-value (Chi-square)                           0.024
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
```

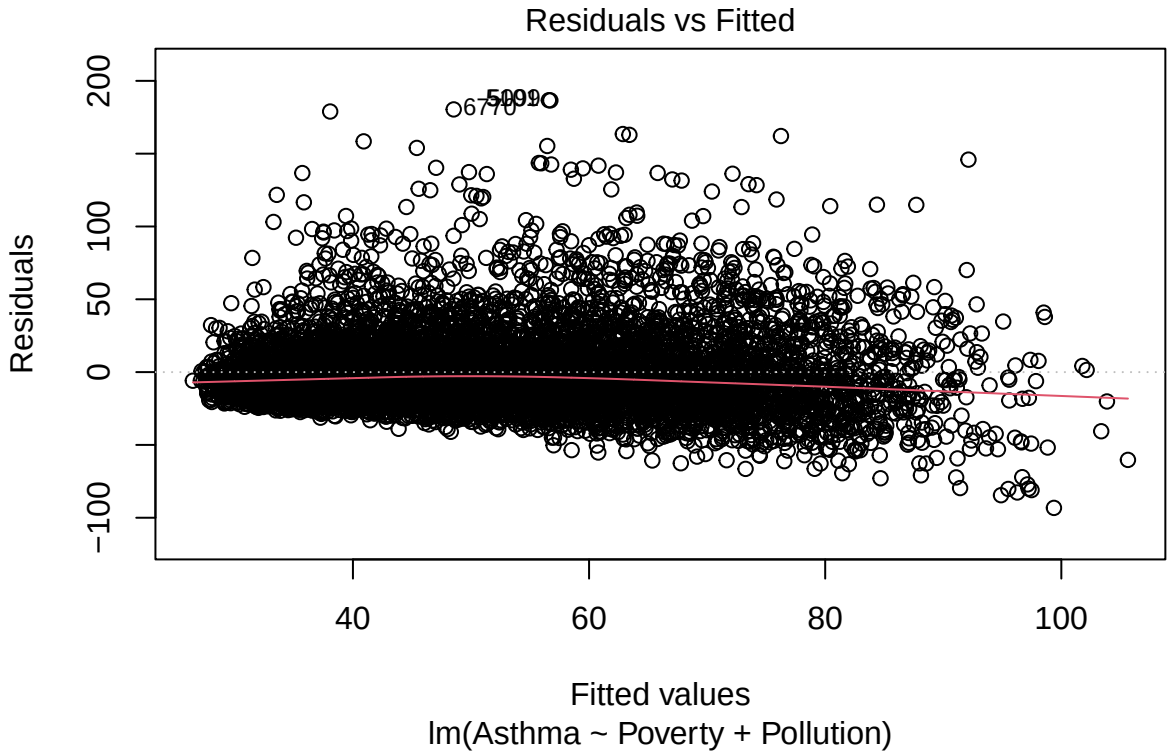
```
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Poverty ~
##     Pollution         0.538    0.015   36.006    0.000
##   Asthma ~
##     Poverty           0.834    0.016   51.258    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .Poverty         287.074    4.551   63.083    0.000
##    .Asthma          702.587   11.137   63.083    0.000
```

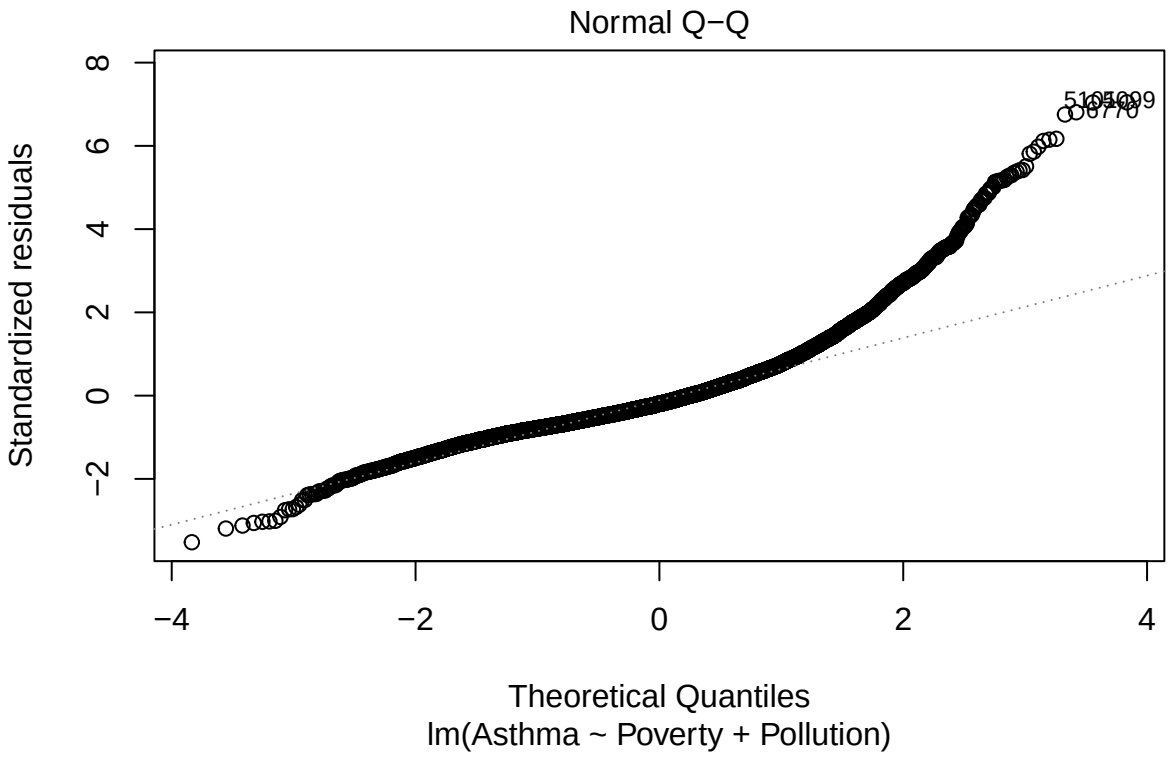
```
#Syntax Structure
#Reference: https://benwhalley.github.io/just-enough-r/path-models.html
#  m ~ x
# y ~ x + m

#User Model
SEM_M3 <-   '
  # regressions
  # Pollution ~ Poverty
    Poverty ~ Pollution
  # Poverty ~ Pollution*EH_2022
    Asthma ~ Poverty + Pollution
  # Asthma ~ Poverty + (Pollution*EH_2022)
  # Asthma ~ Pollution + Poverty
  # EH_2022 ~ Asthma + Pollution
  # Asthma ~ EH_2022
  # EH_2022 ~ Poverty
'

#Example of Syntax Structure
#f3.syn <- '
#Love ~ Money
#Happiness ~ Money + Love
#Travel ~ Happiness + Money
#f3 <- sem(f3.syn, d, conditional.x=FALSE)
#semPaths_default(f3)

SEM_M2_LM2 <- lm(Asthma ~ Poverty + Pollution, data = SEM_M1)
summary(SEM_M2_LM2)
```

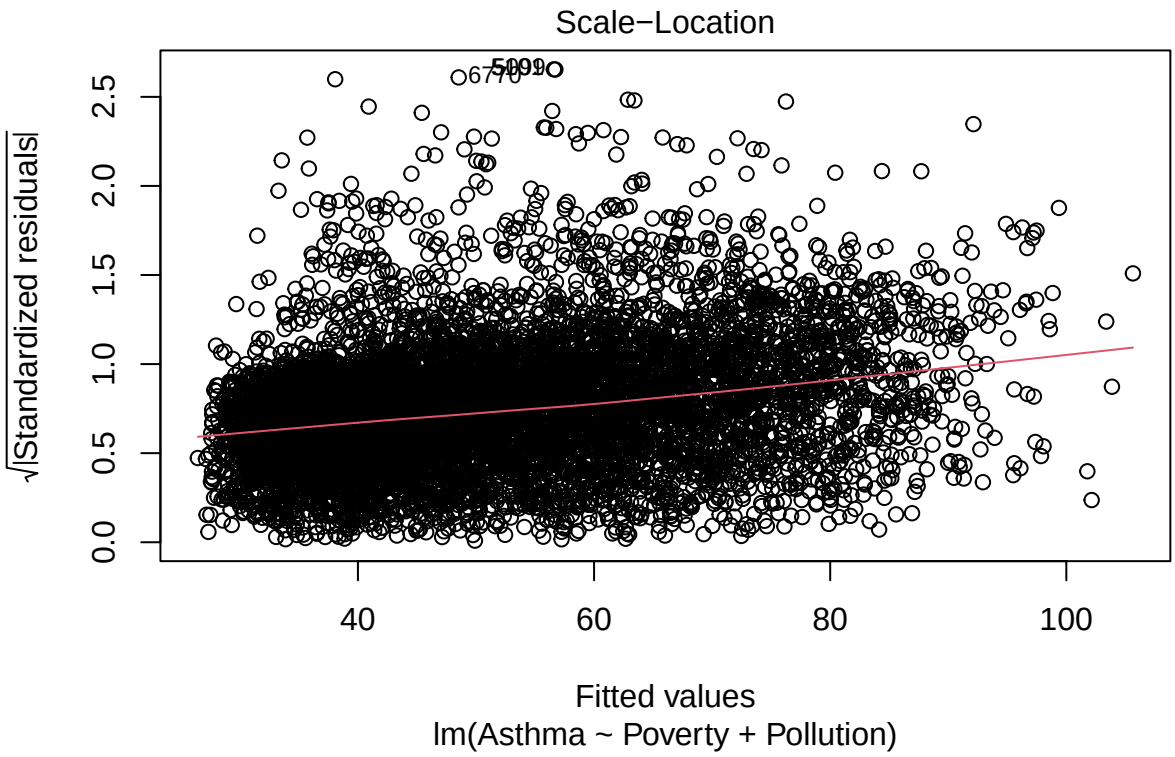
```
##
## Call:
## lm(formula = Asthma ~ Poverty + Pollution, data = SEM_M1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -93.245 -16.247  -5.176  10.478 186.695
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 23.96577    1.05053  22.813   <2e-16 ***
## Poverty      0.81874    0.01753  46.712   <2e-16 ***
```

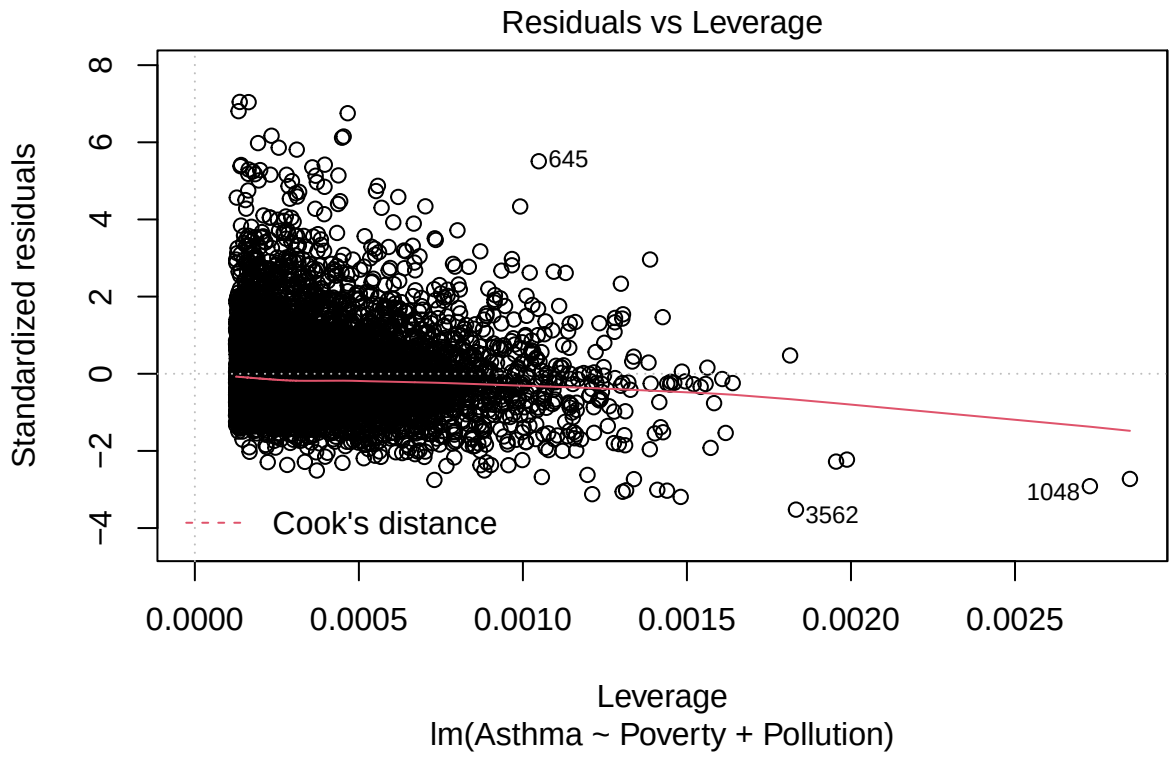
```
## Pollution    0.05691    0.02519   2.259   0.0239 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 26.5 on 7958 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.2487, Adjusted R-squared:  0.2485
## F-statistic:  1317 on 2 and 7958 DF,  p-value: < 2.2e-16
```

```
#Null/Baseline User Model SEM Tests Output
plot(lm(Asthma ~ Poverty + Pollution, data = SEM_M1))
```

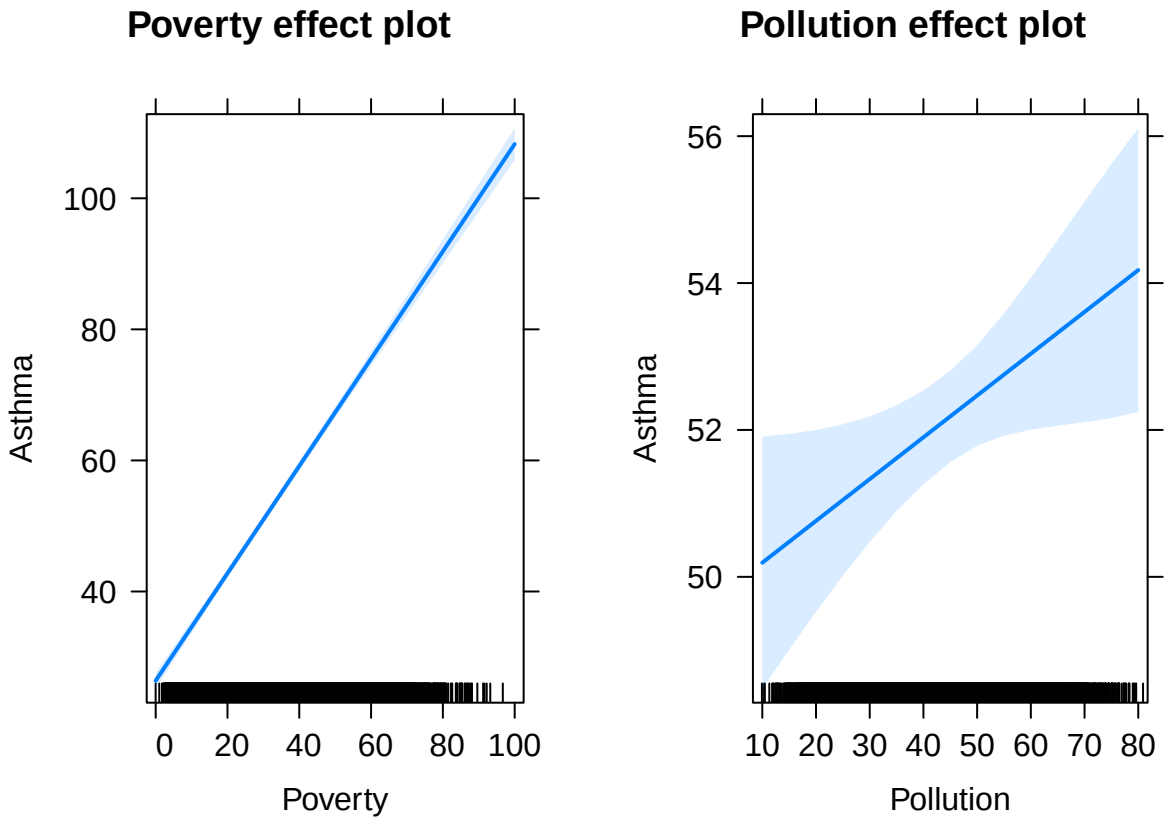



Normal Q−Q
Standardized residuals
8
6
4
2
0
−2
−4
−2
0
2
4
Theoretical Quantiles
lm(Asthma ~ Poverty + Pollution)

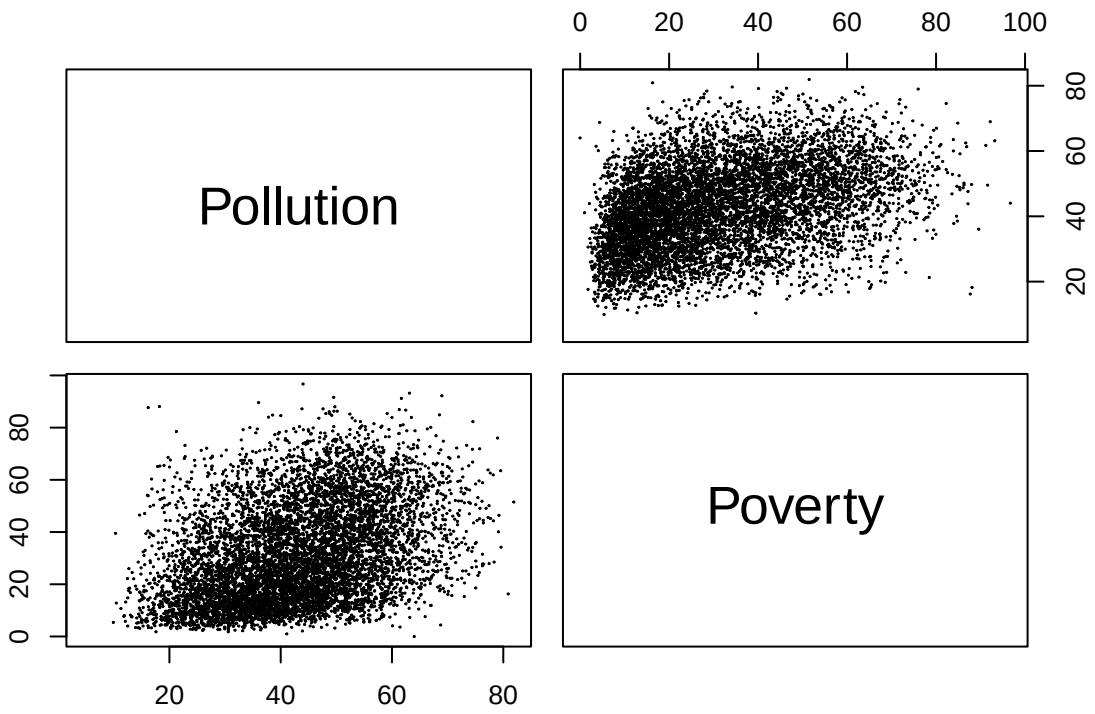

Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
2.0
2.5
40
60
80
100
Fitted values
lm(Asthma ~ Poverty + Pollution)



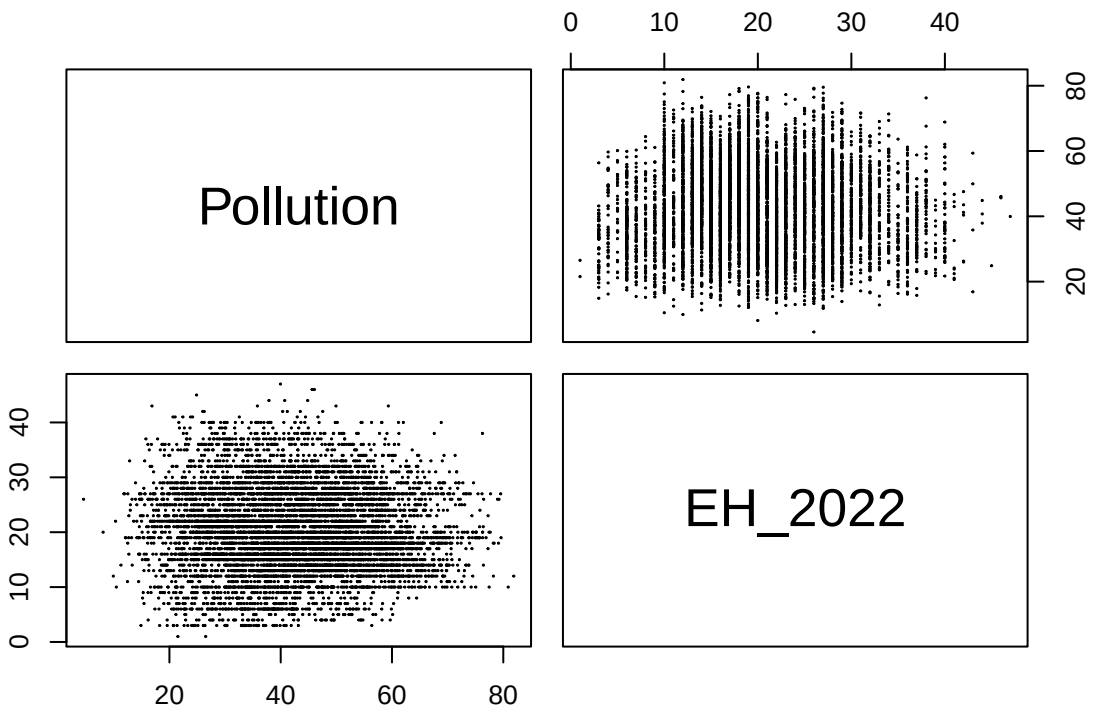

```
plot(allEffects(SEM_M2_LM2))
```



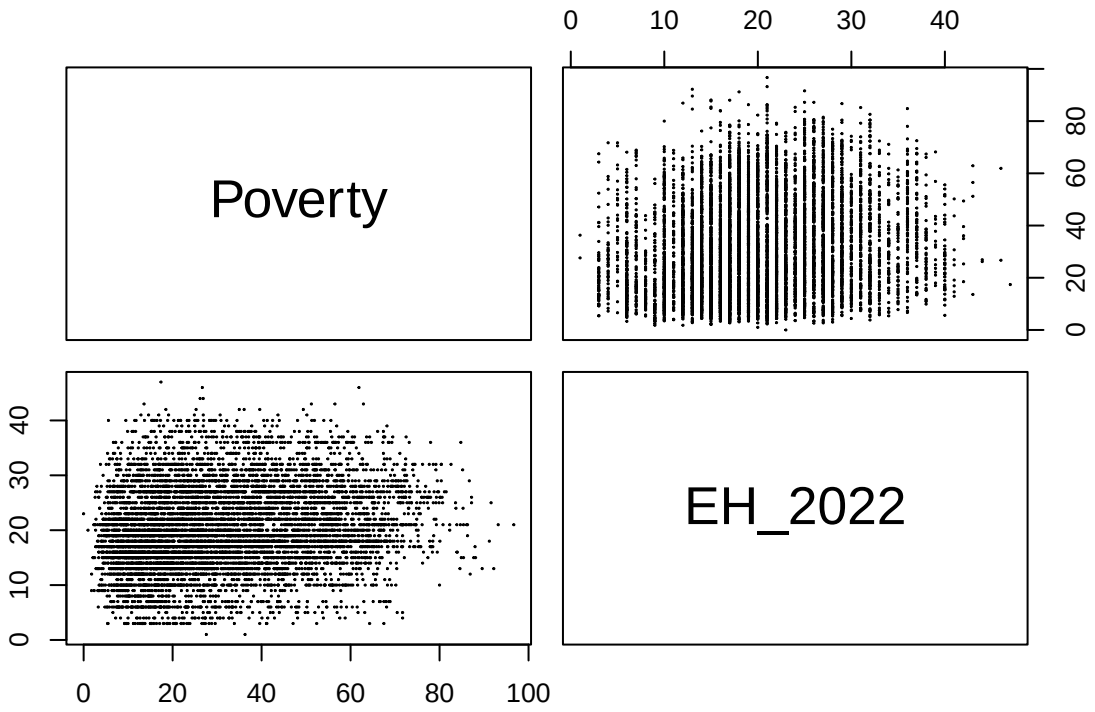

```
#Poverty and Pollution Burden are both correlated with the
#rate of Asthma in California
#Poverty has the strongest correlation with the rate of Asthma in California
#The rate of Asthma in California increases with Poverty (nationwide)
#and Pollution Burden
SEM_M1 %>% select(Pollution,  Poverty)  %>% pairs(cex = 0.1)
```




```
SEM_M1 %>% select(Pollution,  EH_2022)  %>% pairs(cex = 0.1)
```




```
SEM_M1 %>% select(Poverty,  EH_2022)  %>% pairs(cex = 0.1)
```

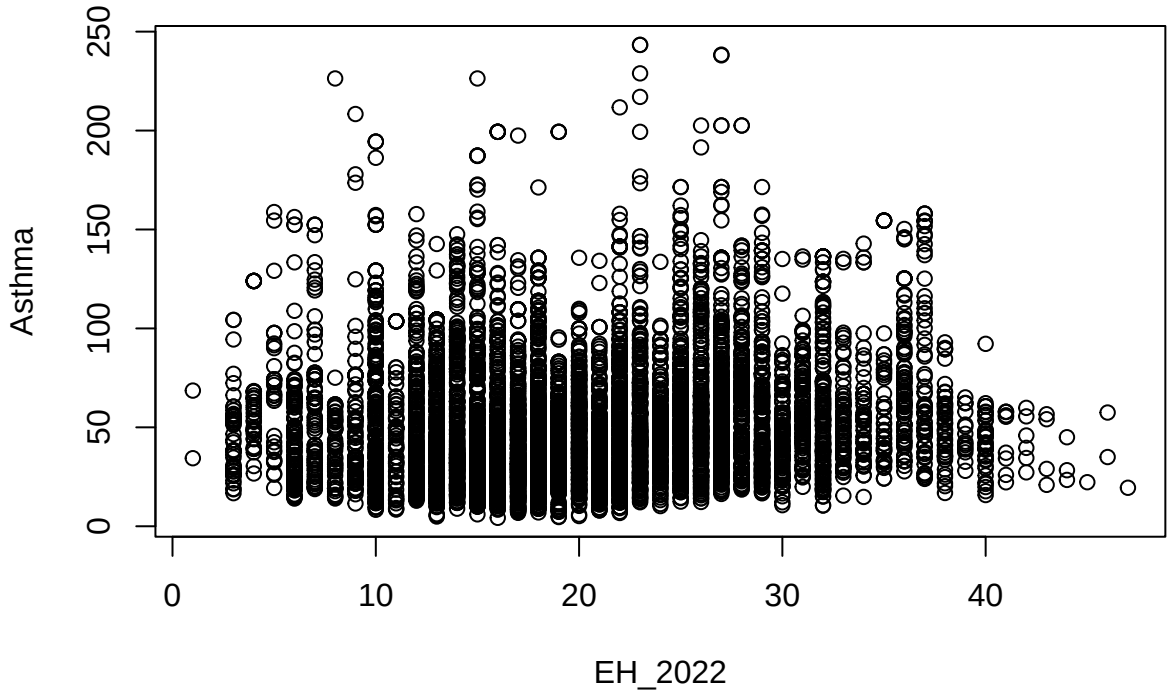



```
#There is a linear relationship between Pollution Burden and Poverty
#in California
#Pollution Burden and Poverty are correlated with each other
#Poverty is the dominant variable
#Pollution Burden increases with Poverty(low-income communities in California)
SEM_M2_LM3 <- lm(Asthma ~ EH_2022, data = SEM_M1)
summary(SEM_M2_LM3)
```

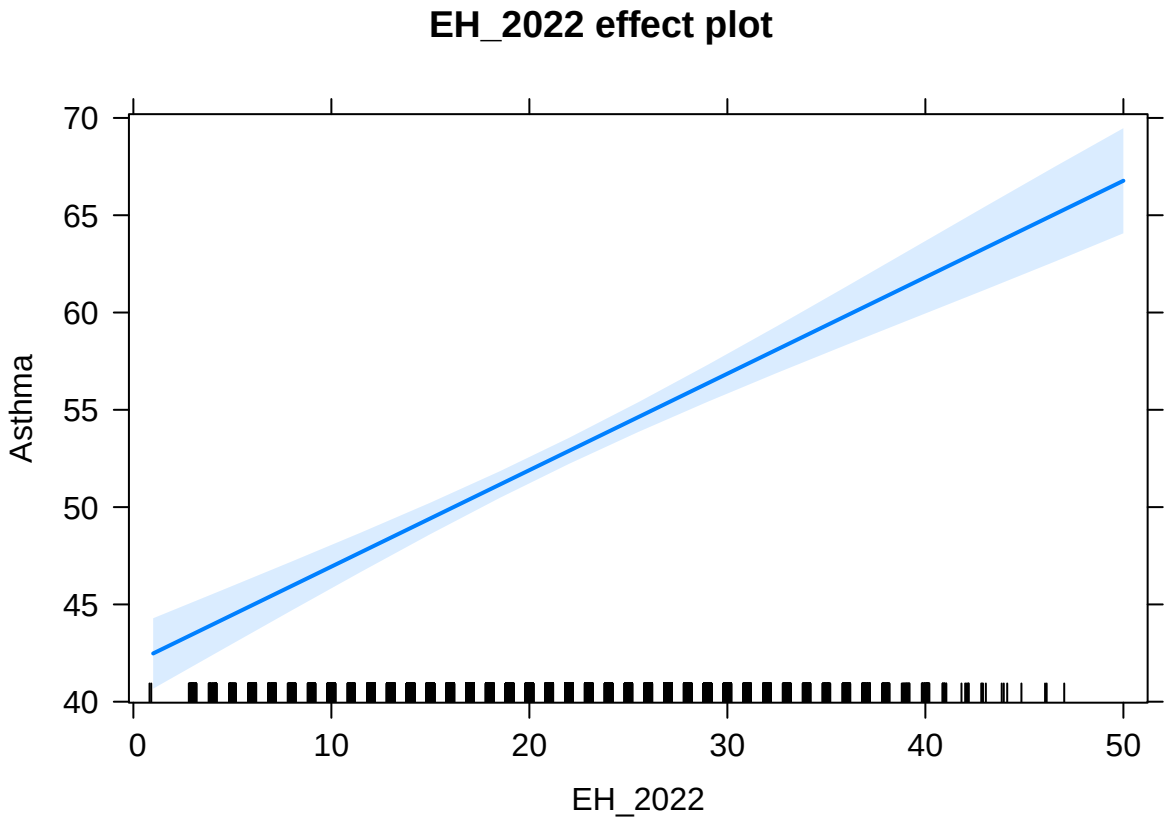
```
##
## Call:
## lm(formula = Asthma ~ EH_2022, data = SEM_M1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -47.368 -21.549  -6.657  13.899 189.904
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  41.9838     0.9651   43.50   <2e-16 ***
## EH_2022       0.4958     0.0448   11.07   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 30.35 on 8020 degrees of freedom
##   (13 observations deleted due to missingness)
## Multiple R-squared:  0.01504,    Adjusted R-squared:  0.01492
```

```
## F-statistic: 122.5 on 1 and 8020 DF,  p-value: < 2.2e-16
```

```
#User Model SEM Tests Output
plot(Asthma ~ EH_2022, data = SEM_M1)
```



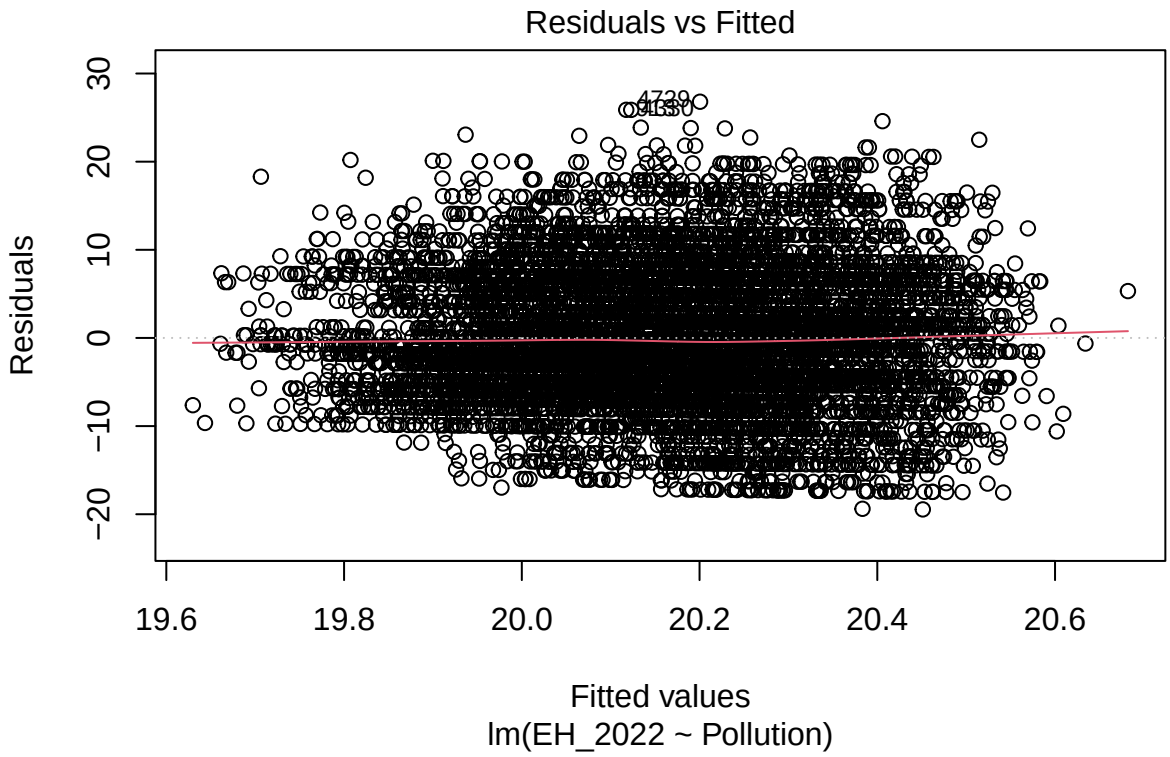

```
plot(allEffects(SEM_M2_LM3))
```

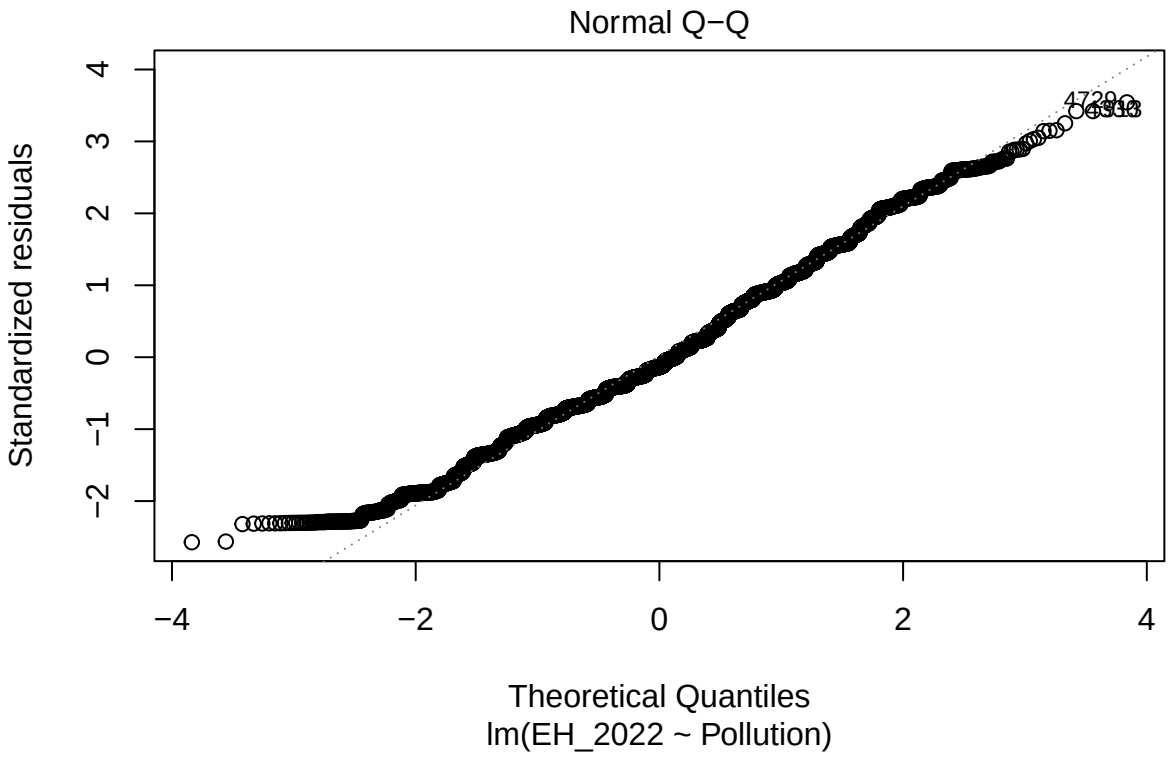


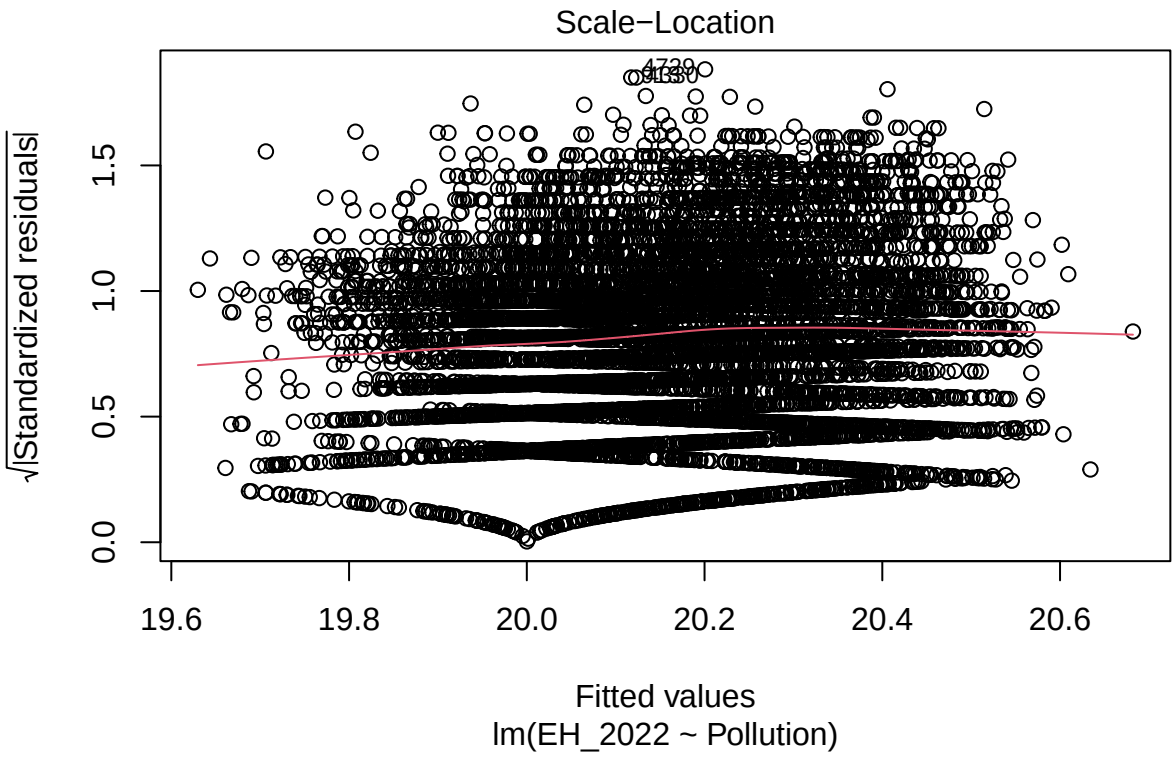

```
#Extreme Heat is correlated with the rate of Asthma in California
#The rate of Asthma increases with Extreme Heat in California
SEM_M2_LM4 <- lm(EH_2022 ~ Pollution, data = SEM_M1)
summary(SEM_M2_LM4)
```

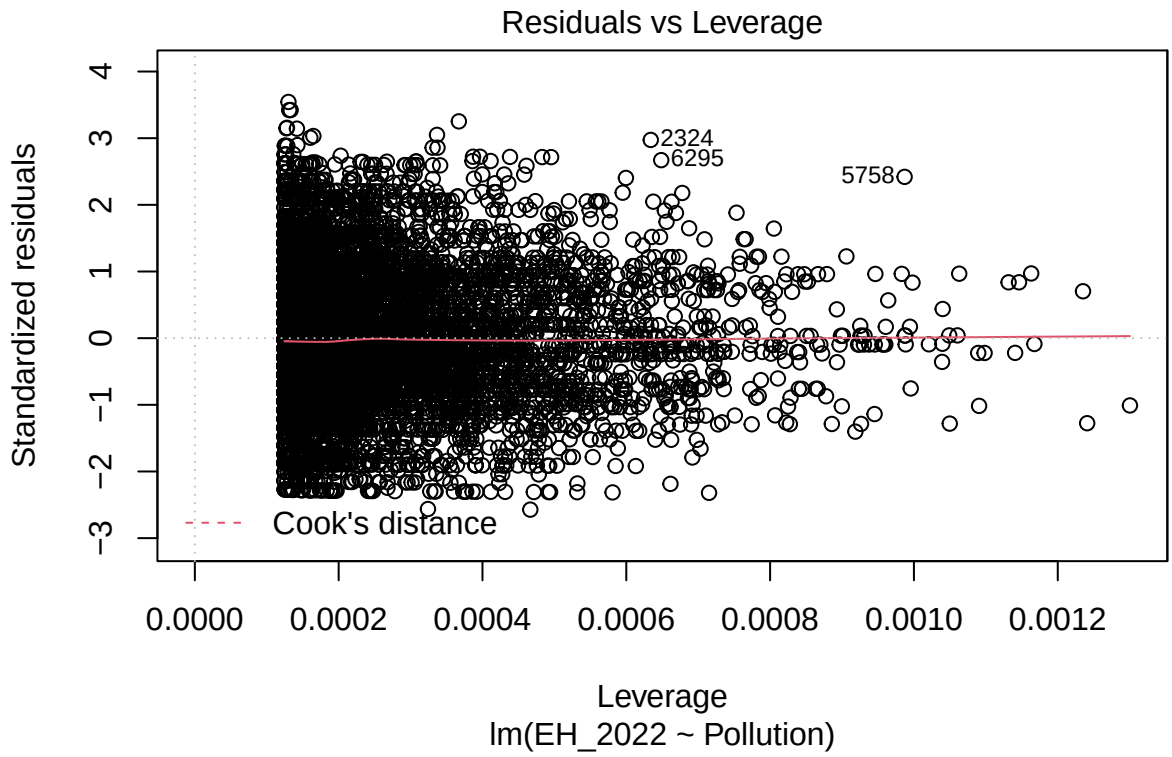
```
##
## Call:
## lm(formula = EH_2022 ~ Pollution, data = SEM_M1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.451  -5.151  -1.025   5.455  26.799
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 20.744401   0.294721  70.386   <2e-16 ***
## Pollution   -0.013609   0.006615  -2.057   0.0397 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.564 on 8031 degrees of freedom
##   (2 observations deleted due to missingness)
## Multiple R-squared:  0.0005268,  Adjusted R-squared:  0.0004023
## F-statistic: 4.233 on 1 and 8031 DF,  p-value: 0.03969
```

```
plot(lm(EH_2022 ~ Pollution, data = SEM_M1))
```

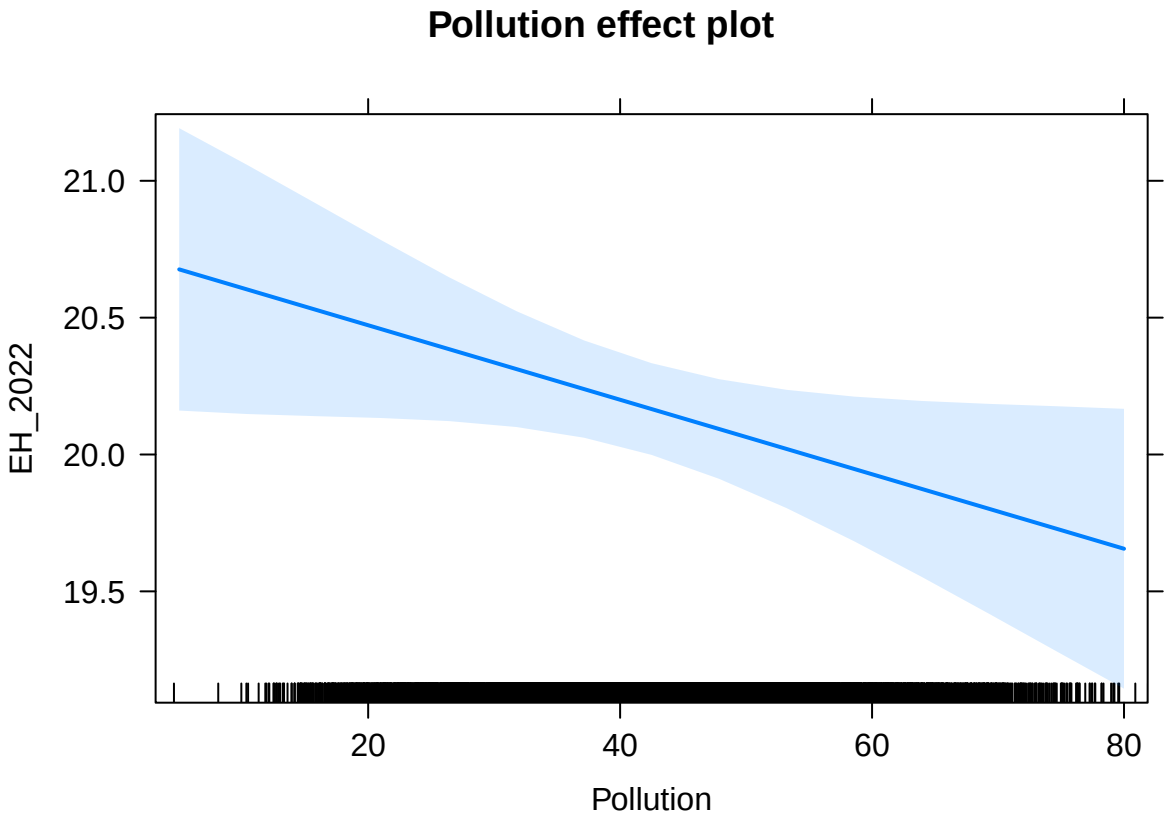



Normal Q−Q
Standardized residuals
−2
−1
0
1
2
3
4
−4
−2
0
2
4
Theoretical Quantiles
lm(EH_2022 ~ Pollution)


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
19.6
19.8
20.0
20.2
20.4
20.6
Fitted values
lm(EH_2022 ~ Pollution)



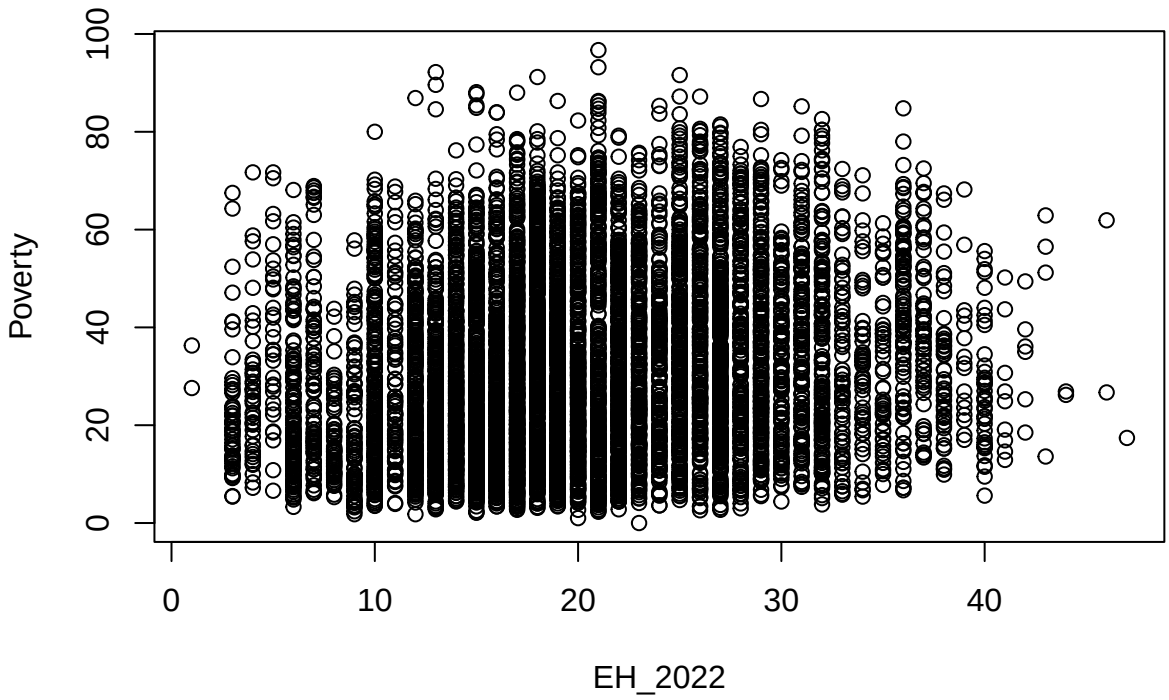

```
plot(allEffects(SEM_M2_LM4))
```



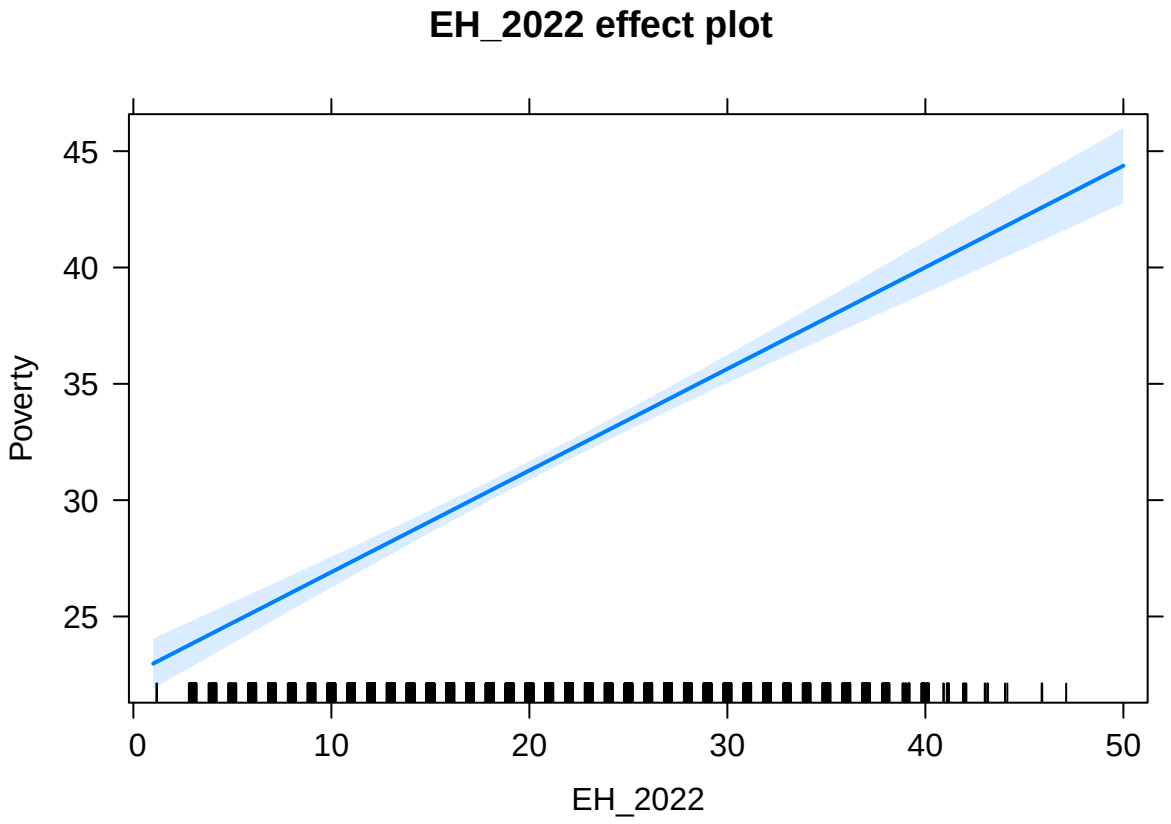

```
#Pollution and Extreme Heat are not correlated
SEM_M2_LM5 <- lm(Poverty ~ EH_2022, data = SEM_M1)
summary(SEM_M2_LM5)
```

```
##
## Call:
## lm(formula = Poverty ~ EH_2022, data = SEM_M1)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -34.409 -14.642  -3.331  12.647  64.990
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 22.53889    0.57527   39.18   <2e-16 ***
## EH_2022      0.43674    0.02673   16.34   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.97 on 7957 degrees of freedom
##   (76 observations deleted due to missingness)
## Multiple R-squared:  0.03246,    Adjusted R-squared:  0.03234
## F-statistic:   267 on 1 and 7957 DF,  p-value: < 2.2e-16
```

```
plot(Poverty ~ EH_2022, data = SEM_M1)
```



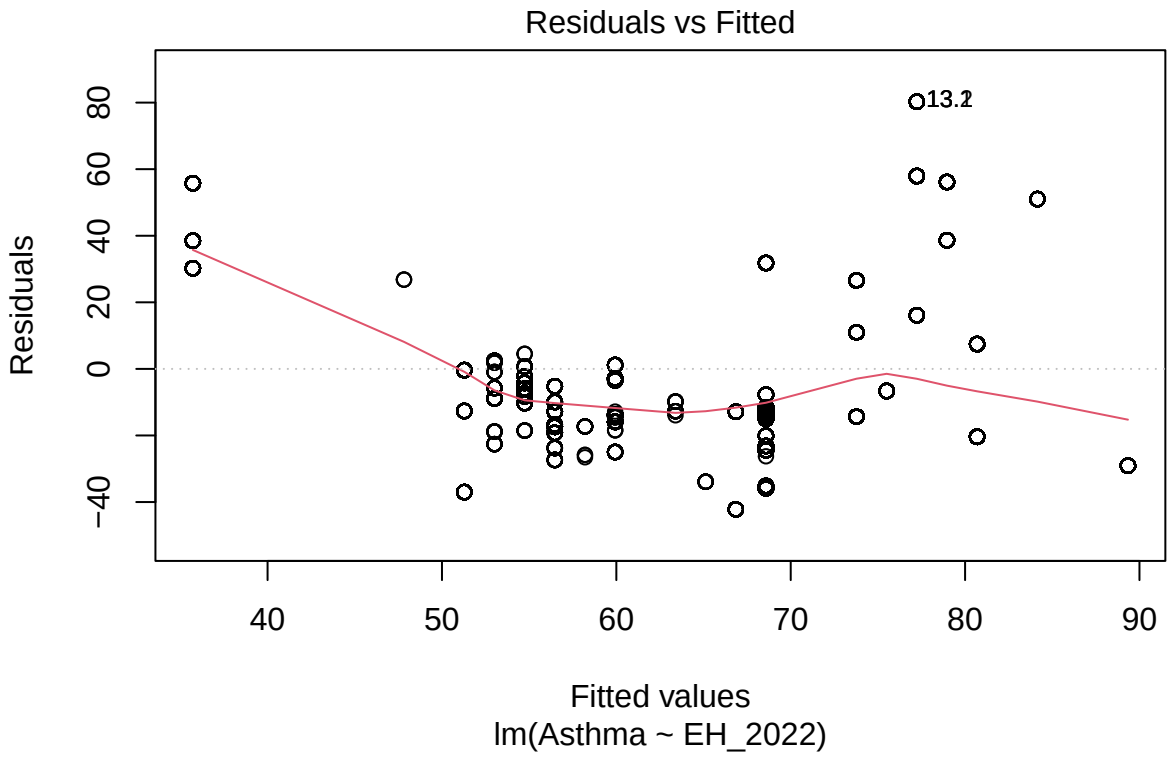

```
plot(allEffects(SEM_M2_LM5))
```

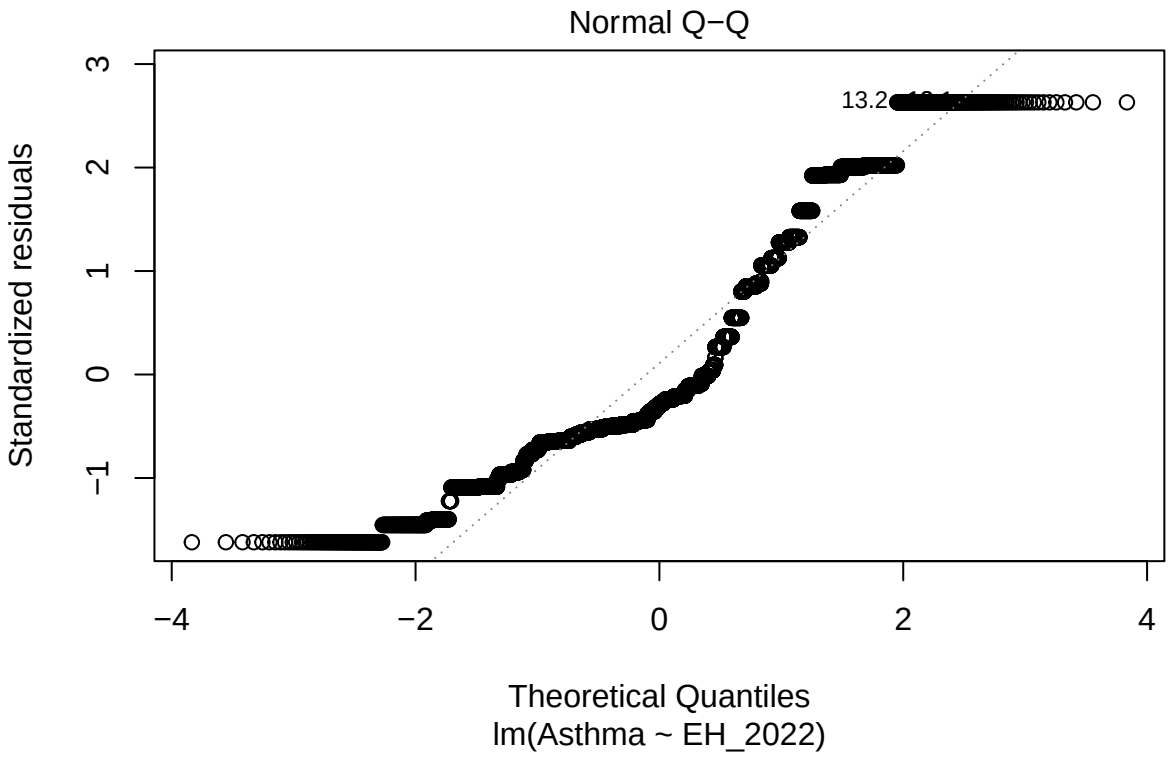


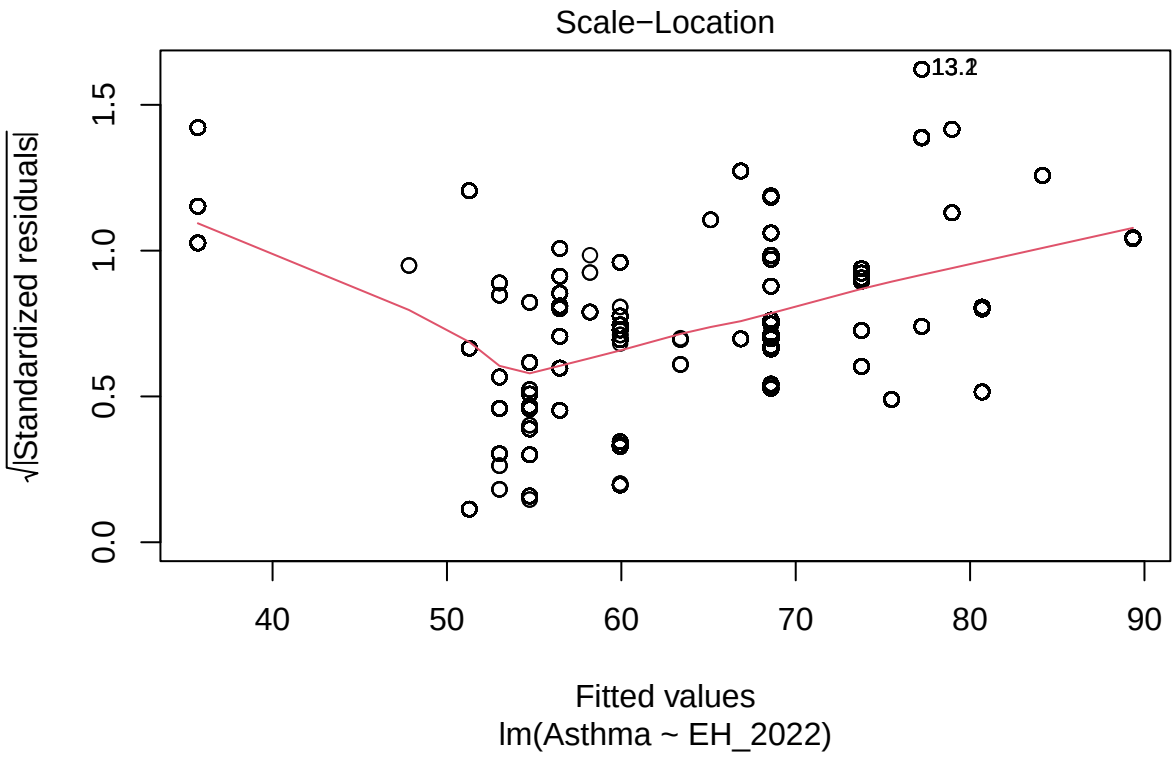

```
#Extreme Heat and Poverty are correlated
#Poverty increases with Extreme Heat
SEM_M2_LM6 <- lm(Asthma ~ EH_2022, Poverty, Pollution, data = SEM_M1)
summary(SEM_M2_LM6)
```

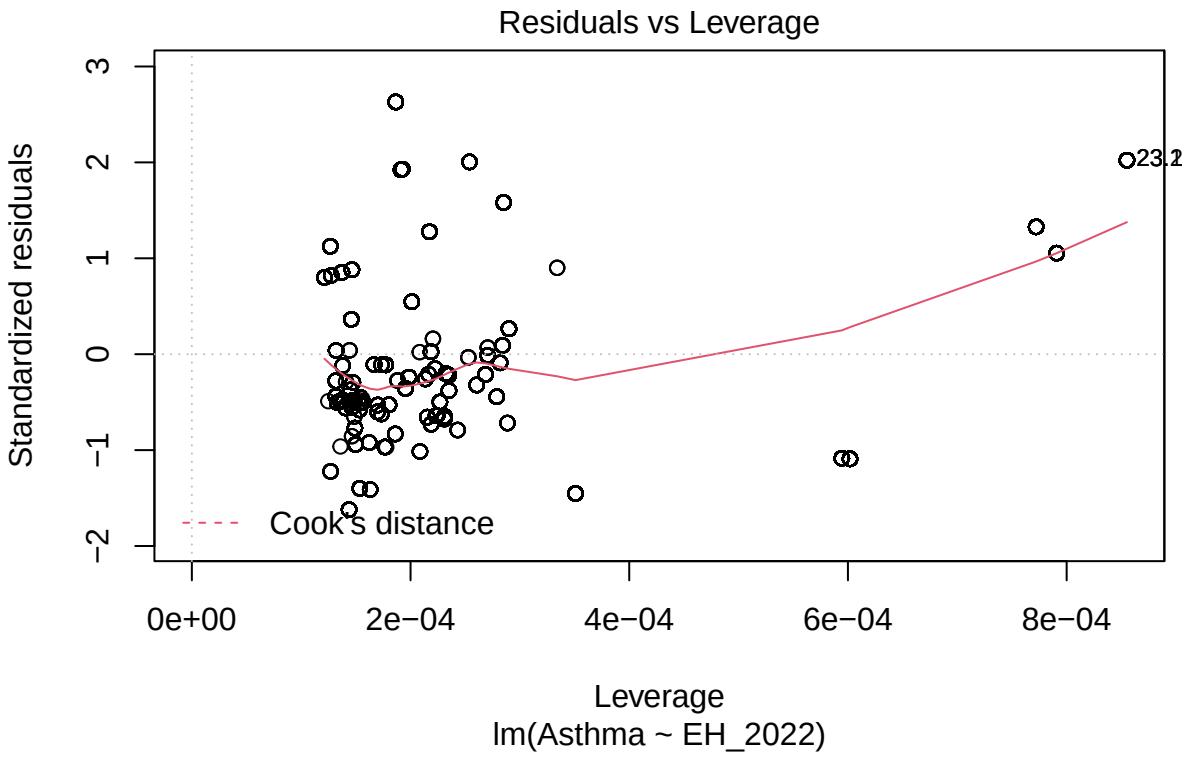
```
##
## Call:
## lm(formula = Asthma ~ EH_2022, data = SEM_M1, subset = Poverty,
##     weights = Pollution)
##
## Weighted Residuals:
##    Min     1Q Median     3Q    Max
## -337.7 -120.9  -61.3  166.9  548.1
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 27.06467    1.00615   26.90   <2e-16 ***
## EH_2022      1.72985    0.04253   40.67   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 208.4 on 7958 degrees of freedom
##   (74 observations deleted due to missingness)
## Multiple R-squared:  0.1721, Adjusted R-squared:  0.172
## F-statistic:  1654 on 1 and 7958 DF,  p-value: < 2.2e-16
```

```
plot(lm(Asthma ~ EH_2022, Poverty, Pollution, data = SEM_M1))
```

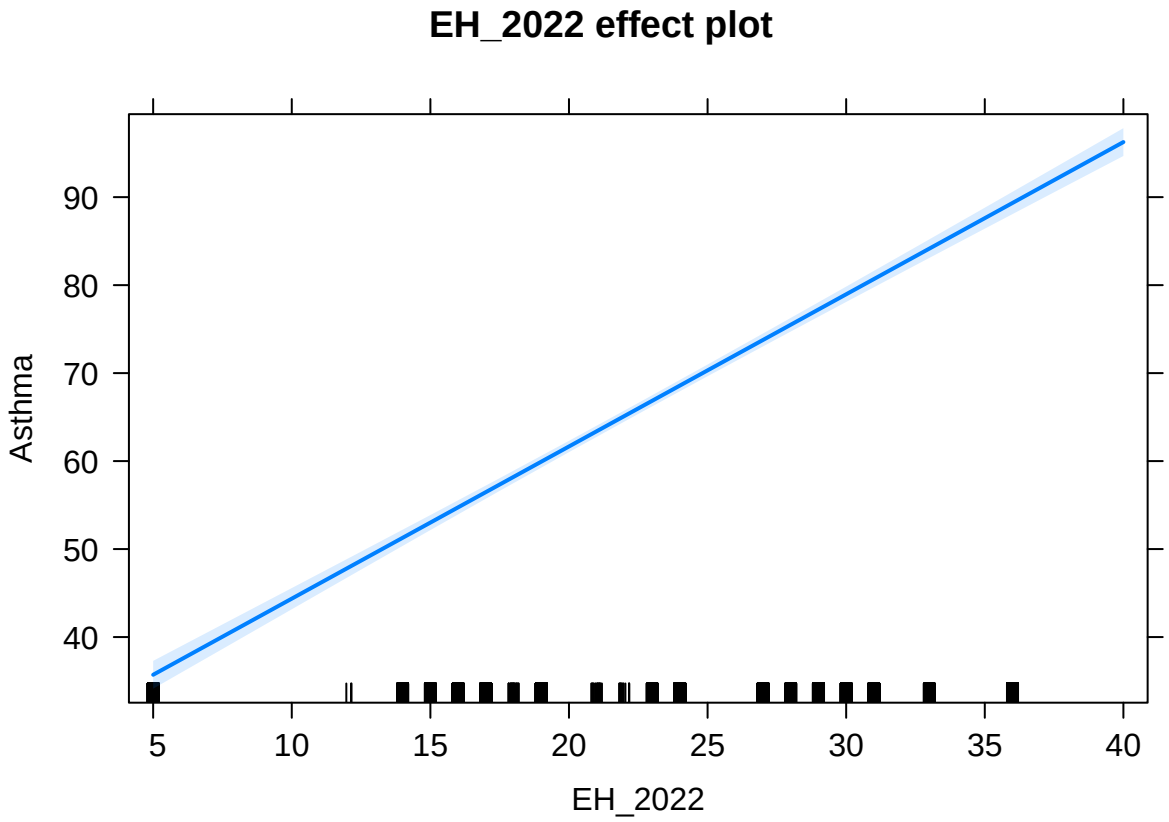



Normal Q−Q
Standardized residuals
Theoretical Quantiles
lm(Asthma ~ EH_2022)
13.2
−4
−2
0
2
4
−1
0
1
2
3


Scale−Location
√|Standardized residuals|
0.0
0.5
1.0
1.5
13.2
40
50
60
70
80
90
Fitted values
lm(Asthma ~ EH_2022)




```
plot(allEffects(SEM_M2_LM6))
```




```
#Pollution, Poverty, and Extreme Heat are correlated with rate of Asthma
#in California. The rate of Asthma increases with Pollution, Poverty,
#and extreme heat
fit1c <- sem(SEM_M3, data=df2)
summary(fit1c)
```

```
## lavaan 0.6-12 ended normally after 1 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         5
##
##   Number of observations                          7959
##
## Model Test User Model:
##
##   Test statistic                                 0.000
##   Degrees of freedom                                 0
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Regressions:
```

```
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Poverty ~
##     Pollution         0.538    0.015   36.006    0.000
##   Asthma ~
##     Poverty           0.819    0.018   46.701    0.000
##     Pollution         0.057    0.025    2.262    0.024
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .Poverty         287.074    4.551   63.083    0.000
##    .Asthma          702.136   11.130   63.083    0.000
```

```
#comparison of variance between two models
anova(fit1b, fit1c)
```

```
## Chi-Squared Difference Test
##
##       Df    AIC    BIC  Chisq Chisq diff Df diff Pr(>Chisq)
## fit1c  0 142393 142428 0.0000
## fit1b  1 142397 142425 5.1132     5.1132       1    0.02374 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
#ANOVA is a statistical test for estimating how a quantitative
#dependent variable changes according to the levels of one or more independent
#variables. ANOVA tests whether there is a difference in means of the
#groups at each level of the independent variable.
#The null hypothesis (HO) of the ANOVA is no difference in means,
#and the alternate hypothesis (Ha) is that the
#means are different from one another.
#Reference: https://www.scribbr.com/statistics/anova-in-r/

#1) The baseline is a null model, typically in which all of your
#observed variables are constrained to covary with no other variables
#The covariances are fixed to 0)--just individual variances are estimated.
#This is what is often taken as a 'reasonable' worst-possible fitting model,
#against which your fitted model is compared in order to
#calculate relative indexes of model fit (e.g., CFI/TLI).

#2) The chi-square statistic (labeled as the minimum function test statistic)
#is used to perform a test of perfect model fit, both for your specified and
#null/baseline models. It essentially is a measure of deviance between your
#model-implied variance/covariance matrix, and your
#observed variance/covariance matrix.

#3) Standards for what level of model fit is considered "acceptable"
#may differ from discipline to discipline, but at least according
#to Hu & Bentler (1999), you are within the realm of what is considered
#"acceptable". A CFI of .955 is often considered "good". Keep in mind, however,
#that both TLI and CFI are relative indexes of model fit--they compare
#the fit of your model to the fit of your (worst fitting) null model.
#Hu & Bentler (1999) suggested that you interpret/report both a
#relative and an absolute index of model fit. Absolute indexes of model fit
```

```
#compare the fit of your model to a perfect fitting model--RMSEA and SRMR
#are a couple of good candidates (the former is often calculated along with a
#confidence interval, which is nice).

#References:
#1. Brown, T. A. (2015). Confirmatory factor analysis for applied
#research (2nd Edition). New York, NY:

#2. Guilford Press., Hu, L., & Bentler, P. M. (1999).
#Cutoff criteria for fit indexes in covariance structure analysis:

#3. Conventional criteria versus new alternatives. Structural Equation Modeling,
#6, 1-55.

#4.  Kline, R. B. (2010). Principles and practice of structural equation modeling
#(3rd Edition). New York, NY: Guilford Press.

#5. Little, T. D. (2013). Longitudinal structural equation modeling. New York,
#NY: Guilford Press.
#https://stats.stackexchange.com/questions/140909/how-do-i-
#interpret-lavaan-output

#The model summary first lists the independent variables being tested
#('Pollution' and 'Poverty'). Next is the residual variance ('Residuals'),
#which is the variation in the dependent variable that isn't explained by
#the independent variables (predictor variables).

#The following columns provide all of the information needed to
#interpret the model:
#Df shows the degrees of freedom for each variable
#(number of levels in the variable minus 1).
#Sum sq is the sum of squares
#(a.k.a. the variation between the group means created by the levels of
#the independent variable and the overall mean).
#Mean sq shows the mean sum of squares (the sum of squares divided by
#the degrees of freedom).
#F value is the test statistic from the F-test
#(the mean square of the variable divided by the mean square of each parameter).
#Pr(>F) is the p-value of the F statistic,
#and shows how likely it is that the F-value calculated
#from the F-test would have occurred if the null hypothesis of
#no difference was true.

#Measurement of Fit for Linear Regression Models #https://www.theanalysisfactor.com/assessing-the-fit-o
#The RMSE is the square root of the variance of the residuals.
#It indicates the absolute fit of the model to the data-how close the
#observed data points are to the model's predicted values.
#Whereas R-squared is a relative measure of fit,
#RMSE is an absolute measure of fit. As the square root of a variance,
#RMSE can be interpreted as the standard deviation of the unexplained variance.
#It has the useful property of being in the same units as the response variable.
#Lower values of RMSE indicate better fit.
#RMSE is a good measure of how accurately the model predicts the response.
```

```r
#It's the most important criterion for fit if the main purpose of the model
#is prediction.


#https://www.scribbr.com/statistics/two-way-anova/
#fit statistics - Null/Baseline and User Model Fitting
#Issue of the accuracy of path coefficients is the evaluation of
#overall model fit.In historic path analysis, fit of the data to the
#overall model was not assessed. Maximum likelihood methods as well as other
#modern solution procedures provide for an additional assessment whether the
#data is  consistent with the specified model overall (typically through a
#chi square or related test of goodness of fit).Tests of overall model fit is
#a major advance associated with modern path analysis

summary(fit1b, fit.measures=TRUE)
```

```
## lavaan 0.6-12 ended normally after 1 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         4
##
##   Number of observations                          7959
##
## Model Test User Model:
##
##   Test statistic                                 5.113
##   Degrees of freedom                                 1
##   P-value (Chi-square)                           0.024
##
## Model Test Baseline Model:
##
##   Test statistic                              3476.632
##   Degrees of freedom                                 3
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    0.999
##   Tucker-Lewis Index (TLI)                       0.996
##
## Loglikelihood and Information Criteria:
##
##   Loglikelihood user model (H0)             -71194.304
##   Loglikelihood unrestricted model (H1)     -71191.747
##
##   Akaike (AIC)                              142396.608
##   Bayesian (BIC)                            142424.536
##   Sample-size adjusted Bayesian (BIC)       142411.825
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.023
```

```
##   90 Percent confidence interval - lower         0.007
##   90 Percent confidence interval - upper         0.044
##   P-value RMSEA <= 0.05                          0.986
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.008
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Poverty ~
##     Pollution         0.538    0.015   36.006    0.000
##   Asthma ~
##     Poverty           0.834    0.016   51.258    0.000
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .Poverty         287.074    4.551   63.083    0.000
##    .Asthma          702.587   11.137   63.083    0.000
```

```
summary(fit1c, fit.measures=TRUE)
```

```
## lavaan 0.6-12 ended normally after 1 iterations
##
##   Estimator                                         ML
##   Optimization method                           NLMINB
##   Number of model parameters                         5
##
##   Number of observations                          7959
##
## Model Test User Model:
##
##   Test statistic                                 0.000
##   Degrees of freedom                                 0
##
## Model Test Baseline Model:
##
##   Test statistic                              3476.632
##   Degrees of freedom                                 3
##   P-value                                        0.000
##
## User Model versus Baseline Model:
##
##   Comparative Fit Index (CFI)                    1.000
##   Tucker-Lewis Index (TLI)                       1.000
##
## Loglikelihood and Information Criteria:
##
```

```
##   Loglikelihood user model (H0)             -71191.747
##   Loglikelihood unrestricted model (H1)     -71191.747
##
##   Akaike (AIC)                              142393.495
##   Bayesian (BIC)                            142428.405
##   Sample-size adjusted Bayesian (BIC)       142412.516
##
## Root Mean Square Error of Approximation:
##
##   RMSEA                                          0.000
##   90 Percent confidence interval - lower         0.000
##   90 Percent confidence interval - upper         0.000
##   P-value RMSEA <= 0.05                             NA
##
## Standardized Root Mean Square Residual:
##
##   SRMR                                           0.000
##
## Parameter Estimates:
##
##   Standard errors                             Standard
##   Information                                 Expected
##   Information saturated (h1) model          Structured
##
## Regressions:
##                    Estimate  Std.Err  z-value  P(>|z|)
##   Poverty ~
##     Pollution         0.538    0.015   36.006    0.000
##   Asthma ~
##     Poverty           0.819    0.018   46.701    0.000
##     Pollution         0.057    0.025    2.262    0.024
##
## Variances:
##                    Estimate  Std.Err  z-value  P(>|z|)
##    .Poverty         287.074    4.551   63.083    0.000
##    .Asthma          702.136   11.130   63.083    0.000
```

```
inspect(fit1b)
```

```
## $lambda
##           Povrty Asthma Polltn
## Poverty        0      0      0
## Asthma         0      0      0
## Pollution      0      0      0
##
## $theta
##           Povrty Asthma Polltn
## Poverty   0
## Asthma    0      0
## Pollution 0      0      0
##
## $psi
##           Povrty Asthma Polltn
## Poverty   3
```

```
## Asthma    0      4
## Pollution 0      0      0
##
## $beta
##           Povrty Asthma Polltn
## Poverty        0      0      1
## Asthma         2      0      0
## Pollution      0      0      0
```

```
inspect(fit1c)
```

```
## $lambda
##           Povrty Asthma Polltn
## Poverty        0      0      0
## Asthma         0      0      0
## Pollution      0      0      0
##
## $theta
##           Povrty Asthma Polltn
## Poverty   0
## Asthma    0      0
## Pollution 0      0      0
##
## $psi
##           Povrty Asthma Polltn
## Poverty   4
## Asthma    0      5
## Pollution 0      0      0
##
## $beta
##           Povrty Asthma Polltn
## Poverty        0      0      1
## Asthma         2      0      3
## Pollution      0      0      0
```

```
#inspect() scans a data.frame object for errors that may affect the use of
#functions in model. By default, all variables are checked regarding
#the class (numeric or factor), missing values, and presence of
#possible outliers. The function will return a warning if the data
#looks like unbalanced, has missing values or possible outliers.
parameterEstimates(fit1b)
```

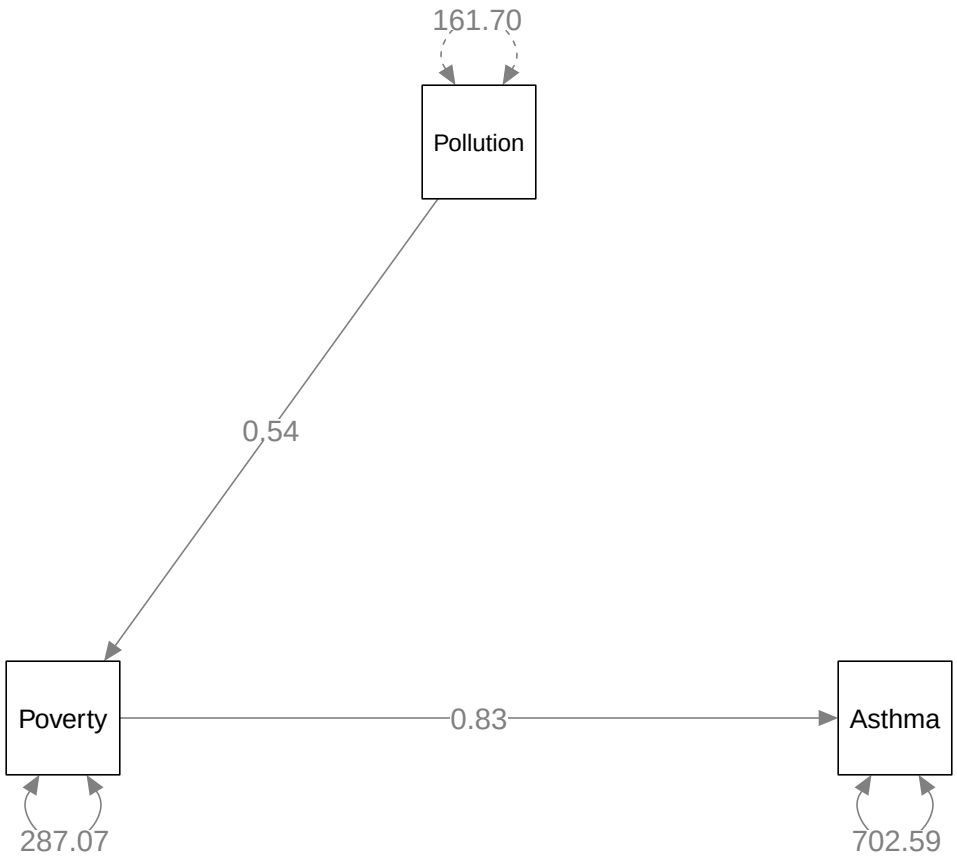
```
##         lhs op       rhs     est     se      z pvalue ci.lower ci.upper
## 1   Poverty  ~ Pollution   0.538  0.015 36.006      0    0.508    0.567
## 2    Asthma  ~   Poverty   0.834  0.016 51.258      0    0.802    0.865
## 3   Poverty ~~   Poverty 287.074  4.551 63.083      0  278.155  295.993
## 4    Asthma ~~    Asthma 702.587 11.137 63.083      0  680.758  724.416
## 5 Pollution ~~ Pollution 161.700  0.000     NA     NA  161.700  161.700
```

```
parameterEstimates(fit1c)
```

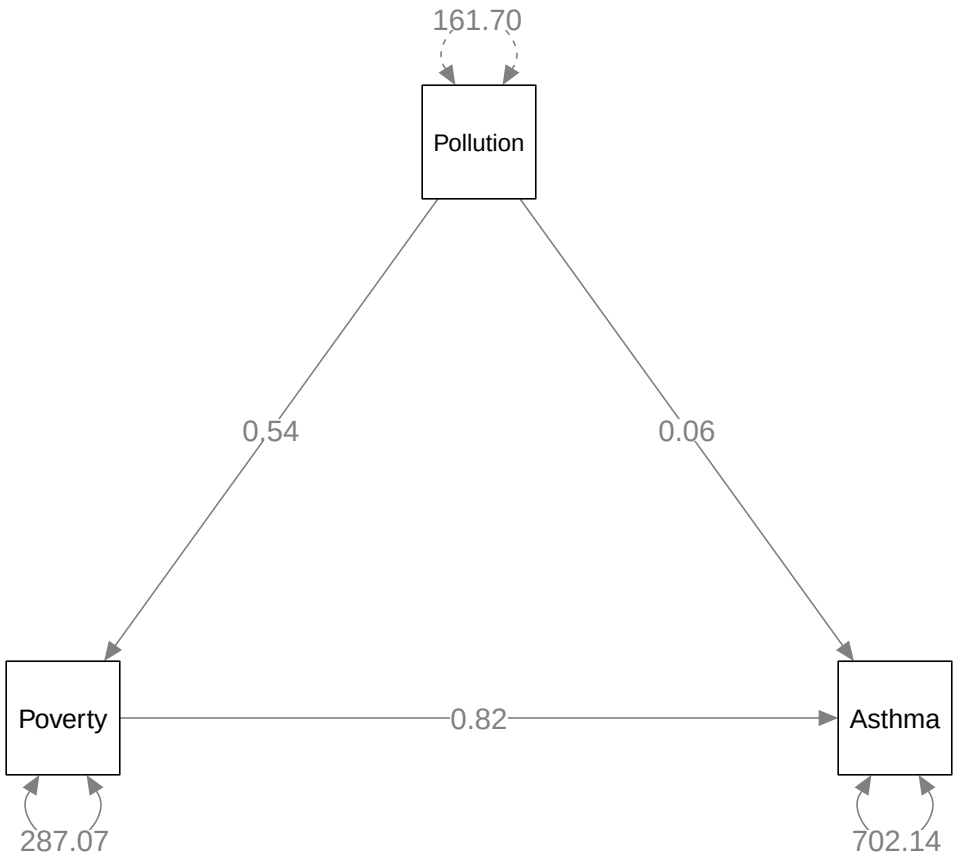
```
##         lhs op       rhs     est     se      z pvalue ci.lower ci.upper
```

```
## 1   Poverty  ~ Pollution   0.538  0.015 36.006  0.000    0.508    0.567
## 2    Asthma  ~   Poverty   0.819  0.018 46.701  0.000    0.784    0.853
## 3    Asthma  ~ Pollution   0.057  0.025  2.262  0.024    0.008    0.106
## 4   Poverty ~~   Poverty 287.074  4.551 63.083  0.000  278.155  295.993
## 5    Asthma ~~    Asthma 702.136 11.130 63.083  0.000  680.321  723.951
## 6 Pollution ~~ Pollution 161.700  0.000     NA     NA  161.700  161.700
```

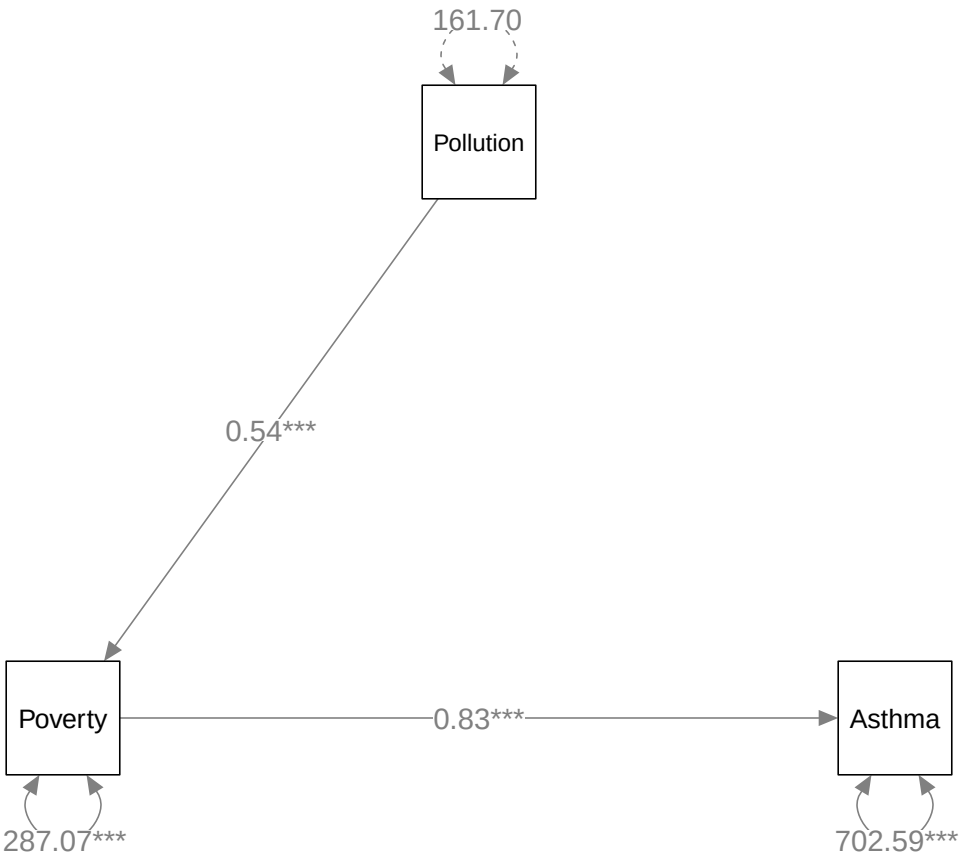
```
#Let us consider a simple path analysis model:
#Elements of graphical representation include:
#1. observed variables (in boxes)
#2. double-headed curved arrows (representing unresolved correlations)
#3. single-headed arrows (representing directed relationships residual variances

#Typically the symbols representing residual variances (error terms)
#are not enclosed by squares.

#There are several basic kinds of path models depending on their
#structure and interconnections.

#Path coefficients specify values for the parameters associated with
#pathways between variables.
#'semPlot' is a new package that can be used for unified visualizations of
#SEM models.
#Path diagrams and visual analysis of various SEM packages' output.
#Path diagrams including visualizations of the parameter estimates
#can be plotted with semPaths and visualizations of the implied and
#observed correlation structures
library(semPlot)
p_1b <- semPaths(fit1b, whatLabels = "est", #Baseline Model
         #  object, rsquare = TRUE
           sizeMan = 10,
           edge.label.cex = 1.15,
           style = "ram",
           nCharNodes = 0, nCharEdges = 0)
```



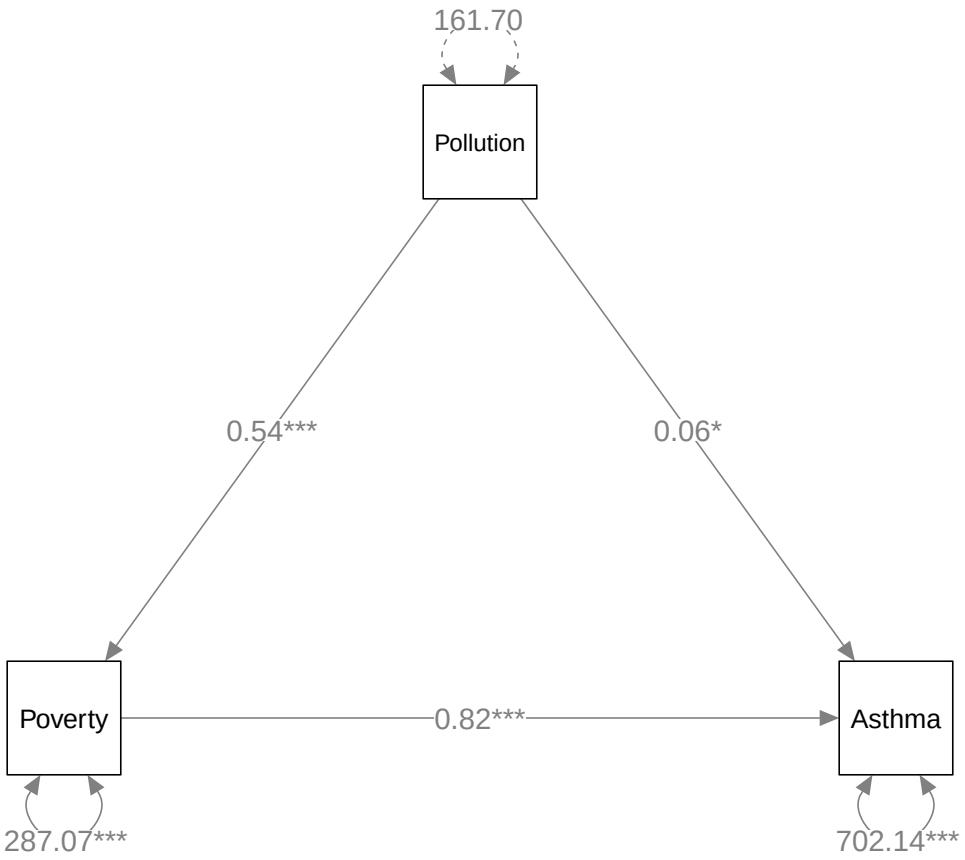

```
p_1c <- semPaths(fit1c, whatLabels = "est", #User Model
           sizeMan = 10,
           edge.label.cex = 1.15,
           style = "ram",
           nCharNodes = 0, nCharEdges = 0)
```




```
#load package ("semptools")
#plot significant paths based on p-value
#We know from the lavaan::lavaan() output that some paths are significant
#and some are not. In some disciplines, asterisks are conventionally added
#indicate this. However, semPlot::semPaths() does not do this.
#We can use mark_sig() to add asterisks based on the p-values
#of the free parameters.
#Reference
#https://cran.r-project.org/web/packages/semptools/vignettes/semptools.html
library(semptools)
p_1b_1 <- mark_sig(p_1b, fit1b)
plot(p_1b_1)
```




```
p_1c_1 <- mark_sig(p_1c, fit1c)
plot(p_1c_1)
```




```
#Advantages of SEM:
#1. Can test hypotheses based on multiple constructs that may be
#indirectly or directly related for both linear and nonlinear models.
#2. Compared to conventional multiple regression analyses,
#SEM has greater statistical power

#Disadvanatages of SEM:
#1. Simultaneous examination of multiple variables requires larger sample sizes
#for additional variables
#2. SEM cannot correct for weaknesses inherent in any type of study.
#3. An a priori specification is necessary for making sense of the
#statistical significance of relationships among variables

#Like any model, there are some advantages and disadvantages to SEM.

#1. The advantages of modeling using SEM compared to other techniques
#boils down to two points. Firstly, with SEM you can test hypothesis
#based on multiple constructs that may be indirectly or directly related for
#both linear and nonlinear models. Secondly, compared to conventional
#multiple regression analyses, SEM has greater statistical power. Regarding the
#disadvantages of SEM, they mostly concern data structure and the quality of
#your experimental design. The first disadvantage is that simultaneous
#examination of multiple variables requires larger sample sizes for additional variables.

#SEM Capabilities:
#SEM techniques can be applied to problems that range from strictly confirmatory
```

```
#to highly exploratory. Other methods, such as principal components analysis (PCA),
#do not have procedures for directly evaluating a-priori ideas about the
#precise nature of underlying factors that explain a set of correlations.

#SEM Challenges:
#SEM in lavaan is for the most part pretty intuitive and user friendly,
#however, we wanted to take some time to summarize potential challenges
#you might encounter. If you're going to encounter an error in your modeling,
#it's most likely going to be due to something with either your data structure
#or the syntax that you're using. Make sure you've taken the necessary steps
#to remove NAs and make your data types for your selected variables are uniform.
#1. Data structure
#2. Syntax structure
#3. Lavaan vs SEM package comparison
#4. Data interpretation from outputs

#Results:
#1. Null/Baseline Model has a p-value of 0.0000 and R-squared value of 0.000.
#2. User Model has a p-value of 0.024 and R-squared value of 0.237.
#User Model rejected the null hypothesis (p > 0.000)
#The global p-value needs to be above p > 0.05 to be a good fitting model.

#"Does the hypothesized model fit the data well?" This is a critical question
#in almost every application of structural equation modeling (SEM).
#The model chi-square statistic and several fit indices are commonly reported
#to address this question. Several model fit indices that are widely applied are
#considered, all of which are based on a fit function given a specific estimation method.

#The chi-square statistic (labeled as the minimum function test statistic)
#is used to perform a test of perfect model fit, both for your specified and
#null/baseline models. It essentially is a measure of deviance between your
#model-implied variance/covariance matrix, and your observed
#variance/covariance matrix. Our User Model rejected the null by having a (p < .001).
#The global p-value for our user model is 0.024.
#A perfect model or good quality model should have a p-value of (p> 0.05).
#Though, the global p-value of our baseline model was 0.000.
#Some statisticians (e.g., Klein, 2010) argue that the chi-square test of model
#fit is useful in evaluating the quality of a model, but most others
#discourage putting a lot stock in its interpretation, both for conceptual
#(i.e., the null of perfect fit is unreasonable) and practical (i.e, chi-square test
#is sensitive to sample size). RMSEA is an absolute fit index, in that it
#assesses how far a hypothesized model is from a perfect model.

#Fit indices that are considered include the root mean square error of
#approximation (RMSEA; Steiger, 1990; Steiger & Lind, 1980).
#Also, the comparative fit index (CFI; Bentler, 1990), and Tucker-Lewis index
#(TLI; Bentler & Bonett, 1980; Tucker & Lewis, 1973) compares the fit of a
#target model to the fit of an independent, or null, model. CFI and TLI larger
#than .95 indicate relatively good model-data fit in general. CFI and TLI are
#incremental fit indices that compare the fit of a hypothesized model (User Model)
#with that of a baseline model (i.e., a model with the worst fit).
#RMSEA value of < .05 indicates a "close fit," and
#that < .08 suggests a reasonable model-data fit. The RMSEA of our User Model
```

```
#came to be 0.023, while the RMSEA of the baseline was 0.000.
#The CFI and TFI of our User Model came to be 0.999 and 0.996 respectively,
#which is good because it should be > .90. The CFI and TFI for our
#baseline model were 1.000 and 1.000.
#The application of RMSEA, CFI, and TLI is heavily contingent on a set of cutoff criteria.

#Discussion Questions
#1.Based on the outputs of this SEM, do we have enough information to
#justify this particular path?
#2. One potential flaw could be your ecological assumption or
#missing variables that are not being considered?
#(There could be limitations based on the data collection from sources)
#3. What would best the approach to improve the quality of our SEM linear models?

#Future Research: Evaluate Extreme Heat Impacts (4th Variable)
#Research Questions -
#1.Does Pollution impact Asthma?
#2.Does Pollution and Poverty impact Asthma?
#3.Does Pollution, Poverty, and Extreme Heat impact asthma?

#Null/Baseline Model -
#1. Pollution Burden, Poverty, and Extreme Heat have a direct effect
#on the rate of Asthma in California.

#User Model -
#1. Pollution Burden, Poverty, and Extreme Heat have a direct effect on the
#rate of Asthma in California.
#2. Pollution and Extreme Heat have a direct effect on Poverty.
#3. Pollution Burden and Extreme Heat have a direct and
#indirect effect on the rate of Asthma in California mediated by Poverty.
```