Achieving Cold Fusion

Scott Bug, PhD;1,2\* Chien-Shiung Wu, PhD;2,3 Leonardo da Vinci; George Washington Carver, MS;1,2,3 and Marie Curie3

1 Impossibles Investigation Team, Creekshirebrook Academy of Thinks, Let Gade 27182, 1566 Copenhagen, Denmark

2 Department of Physics, University of North Science, 314 Newton Blvd, Springfield CT 06003

3 Statistical Consulting Unit, Creekshirebrook Academy of Thinks, 196 Normal Ave, Columbus, OH

\* Corresponding author

Let Gade 27182, 1566 Copenhagen, Denmark

965-555-5556

[scottbug@impossible.net](mailto:scottbug@impossible.net)

Funding:

# Abstract

A generalized linear model determining who is ill based on age, income, and whether or not someone lives on the east coast is ineffective at the 0.1 significance level.

# Introduction

It is common knowledge that sometimes people fall ill, and while it seems difficult to predict who will become ill, perhaps it is an attainable goal to determine who is already ill at any given moment.

## Background/rationale

Most people seem not to like being ill, (Wikipedia 2019) and science has proven that many diseases are communicable. Knowing who is ill may help to decrease transmission of communicable disease.

## Objectives

We hypothesize that characteristics of an individual will help to inform who is ill, and we hope to accomplish this using statistical techniques.

# Methods

##Study design

We will perform a generalized linear model describing who is ill based on certain demographic information. We will use R statistical software, version 3.5.0. (R Core Team 2018)

## Setting

Americentric boundaries on this study lead us to only be interested in people living in the US. Laziness causes us to disregard the period during which these data were collected.

## Participants

1. This is a cross-sectional study of adults living in the United States. Subjects must have a non-negative income.

## Variables

The outcome variable is binary, indicating whether or not the person is ill. The explanatory variables will be gender, income in dollars, age, and city of residence. We will test collinearity of some of these explanatory variables before developingn our final model.

## Data sources/ measurement

The study data comprises 149,999 adults between 25 and 65 years of age all living in the United States, taken from a data set of 150,000 such individuals. (Lepelaars 2018)

## Bias

This data set only contains individuals of age 25 to 65, so all other individuals will be excluded; any resulting model will likely not be useful for people outside this age range.

## Study size

One individual will be excluded because their listed income is under $0.

## Quantitative variables

All quantitative variables will be treated at face value; none were categorized nor transformed.

## Statistical methods

There are no missing data, fortunately. We will utilize a generalized linear model.

## Descriptive data

Here are descriptive statistics of the raw data set:

## Parsed with column specification:  
## cols(  
## Number = col\_double(),  
## City = col\_character(),  
## Gender = col\_character(),  
## Age = col\_double(),  
## Income = col\_double(),  
## Illness = col\_character()  
## )

## Skim summary statistics  
## n obs: 150000   
## n variables: 6   
##   
## ── Variable type:character ───────────────────────────────────────────────────────────────────────────────────────  
## variable missing complete n min max empty n\_unique  
## City 0 150000 150000 6 15 0 8  
## Gender 0 150000 150000 4 6 0 2  
## Illness 0 150000 150000 2 3 0 2  
##   
## ── Variable type:numeric ─────────────────────────────────────────────────────────────────────────────────────────  
## variable missing complete n mean sd p0 p25 p50  
## Age 0 150000 150000 44.95 11.57 25 35 45   
## Income 0 150000 150000 91252.8 24989.5 -654 80867.75 93655   
## Number 0 150000 150000 75000.5 43301.41 1 37500.75 75000.5  
## p75 p100 hist  
## 55 65 ▇▇▇▇▇▇▇▆  
## 1e+05 177157 ▁▁▂▃▇▂▁▁  
## 112500.25 150000 ▇▇▇▇▇▇▇▇

## Limitations

The primarly limitation of this study is its use of a toy data set. These data are all fictional. Also, even if they were not, this model is overly simplistic. Nonetheless, we will proceed.

## References

Lepelaars, Carlos. 2018. “Toy Dataset.” <https://www.kaggle.com/carlolepelaars/toy-dataset>.

R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <https://www.R-project.org/>.

Wikipedia. 2019. “Main Page — Wikipedia, the Free Encyclopedia.” <http://en.wikipedia.org/w/index.php?title=Main%20Page&oldid=889268954>.