Displaying Health Data

Cases, Techniques, Solutions

Colloquium + Live-Webcast + Recording
Medical Sciences Building (MBS) 160
University of Victoria
November 28 – 30 , 1 – 3 pm PST

University of Victoria
INSTITUTE ON AGING & LIFELONG HEALTH
island health
British Columbia BSERVATORY Population & Public Health
IALSA
Health Sciences Center
UCF UNIVERSITY OF CENTRAL FLORIDA

Displaying Health Data

Cases, Techniques, Solutions

Health Data

DAY 1
2018-11-28
Wednesday

13:00 Transactional data of Island Health: How patients vote with their feet
Dr. Ken Moselle (Island Health) and Dr. Andriy Koval (BC Observatory, UCF)

14:00 Visualizing logistic regression with the "coloring book" technique: A study in ggplot2
Dr. Andriy Koval (BC Observatory for Population and Public Health, UCF)

Substance Use

DAY 2
2018-11-29
Thursday

13:00 Nuances of information sharing and data display in a mobile application for students with substance use disorder
Dr. Barbara (Basia) Andraka-Christou (University of Central Florida)

14:00 Optimizing public health surveillance through reproducible reporting: Response to opioid crisis on Vancouver Island
Shannon Tracey (University of Victoria) and Maritia Gully (Island Health)

Pipelines & Dashboards

DAY 3
2018-11-30
Friday

13:00 Building pipelines and dashboards for practitioners: Mobilizing knowledge with reproducible reporting
Dr. Will Beasley (University of Oklahoma Health Sciences Center)

14:00 Constructing workflows for reproducible analytics: Suppressing small counts for provincial chronic disease dashboard
Dr. Andriy Koval (BC Observatory, UCF) and Anthony Leamon (Island Health)

Visualizing Logistic Regression with a “coloring book” technique:

A study in ggplot2

Andriy
Koval

British Columbia
OBSERVATORY
Population & Public Health

CIHR IRSC
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada

UCF
UNIVERSITY OF CENTRAL FLORIDA

Previous work

- 5 voices, 10 stories: Groningen Harmonization Exercise. IALSA, University of Groningen, 2016-04-21 (slides)

- When Notebooks are not Enough: Constructing Workflows for Reproducible Analytics. Matrix Institute Colloquium Series, University of Victoria, 2018-10-31 (slides)

- *Power of Population Data Science* webinar series. Population Data BC, 2018-11-01 (slides, video)

Today

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon
0.1 Modeling form: univariate logistic regression with categorical predictors
0.2 Graphical form: faceted scatterplot in ggplot2
0.3 Coloring book: Mapping informed expectations from predictors onto color

B. Workflow Highlights

1.0 "Let no one ignorant of geometry enter": (my) scripts were written to be read by humans
1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects
1.2 Layers of Isolation: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)
1.3 Two essential means of production: knitr::stitch() vs rmarkdown::render()

Code along by forking github.com/andkov/ipdln-2018-hackathon

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon by Statistics Canada in Banff

International Population Data Linkage Conference 2018
The LIDIC Hackathon: LInked Data Innovation Challenge

Information for Participants

Date and Time: September 11, 2018 afternoon

Sponsors: We are grateful for sponsorship of this workshop by Statistics Canada and IBM.

Description: Participants will engage in a team-based analysis of a complex, linked, synthesized dataset provided by Statistics Canada. This synthesized data base links socioeconomic and mortality data representing the Canadian population. The data based was derived from existing linked data available at Statistics Canada.

Objectives:

- To encourage innovative thinking about complex linked databases
- To stimulate interdisciplinary and inter-jurisdictional data collaborations
- To facilitate an environment for creative thinking about data
- To promote networking amongst participants

Number of records: 4,346,649
Number of variables: 34

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon by Statistics Canada in Banff

```
ls_model$predicted_values %>% glimpse(50) # predicted values
```

```
Observations: 3,883
Variables: 9
$ PR           <fct> Alberta, Alberta, Alberta...
$ age_group    <fct> 65, 60, 30, 80, 55, 40, 6...
$ female       <fct> FALSE, FALSE, TRUE, FALSE...
$ educ3        <fct> high school, more than hi...
$ marital      <fct> mar_cohab, mar_cohab, mar...
$ poor_health  <fct> FALSE, FALSE, FALSE, TRUE...
$ FOL          <fct> English only, English onl...
$ dv_hat       <dbl> 1.8628432, 2.3139500, 6.1...
$ dv_hat_p     <dbl> 0.8656280, 0.9100258, 0.9...
```

Originally:

Number of records: 4,346,649
Number of variables: 34

Data recreated from model parameters

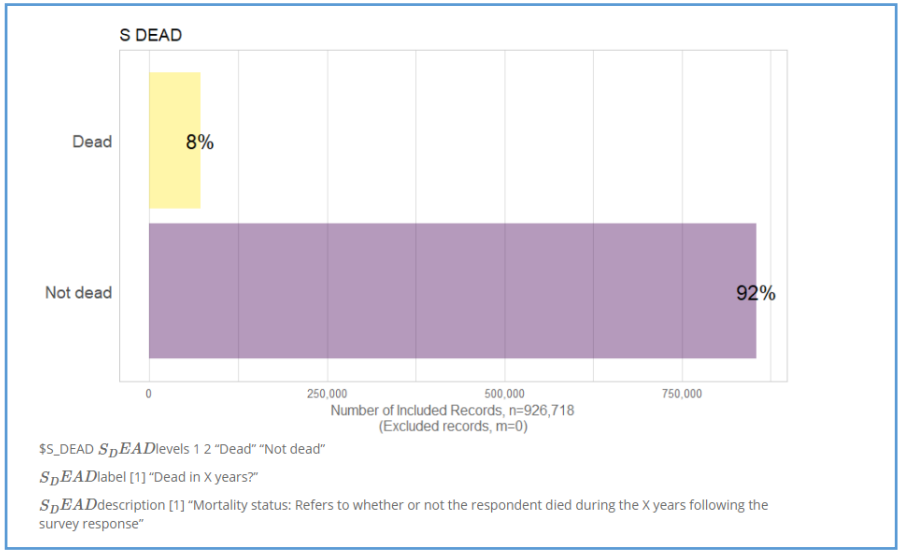
You can use this data to recreate the graphs from this talk with the script *./reports/graphing-phase-only/graphing-phase-only.R*

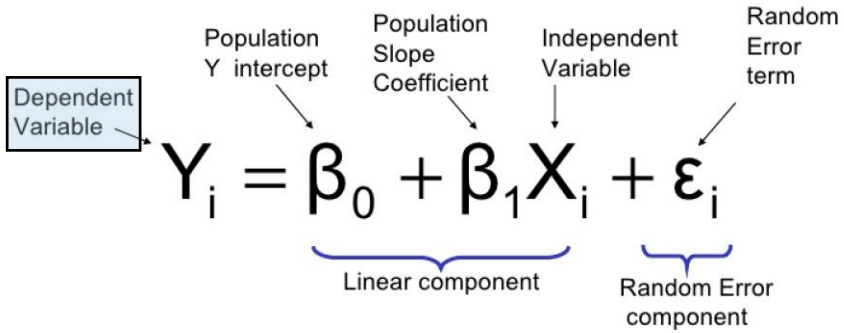
A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Dead in X years





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

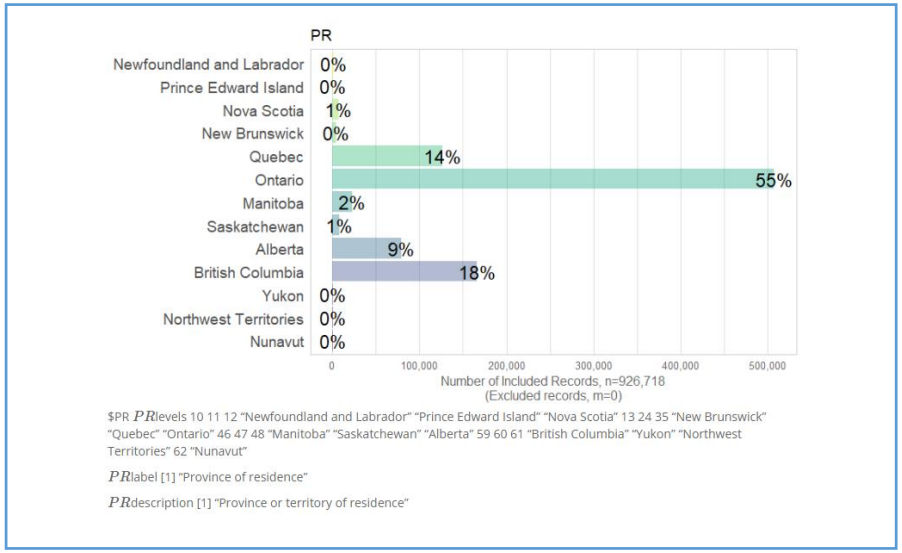
Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

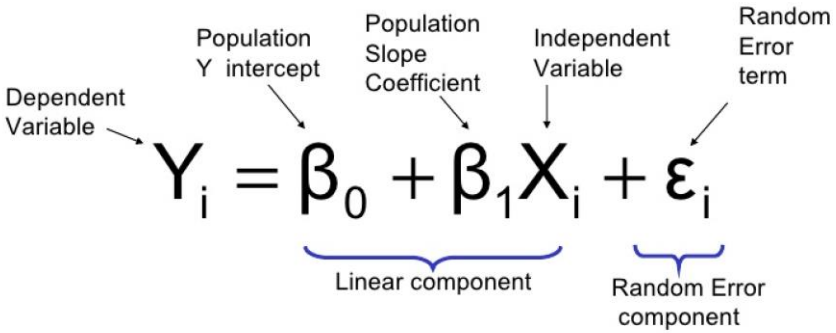
A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Province of residence





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

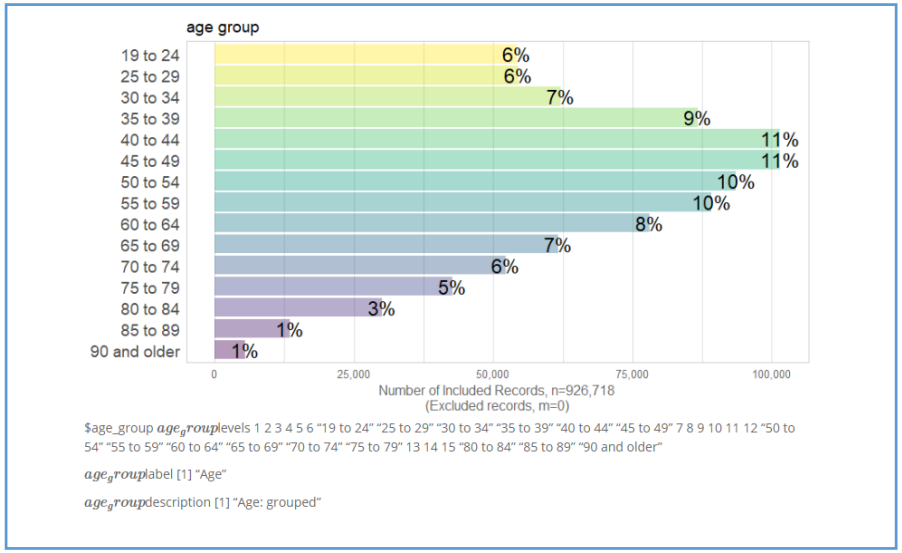
Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

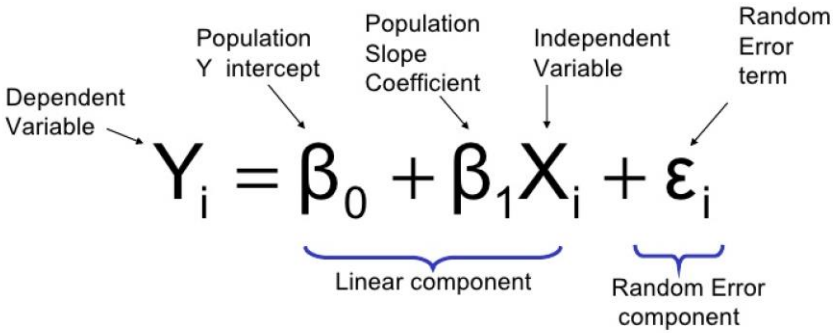
A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

5-year age category





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

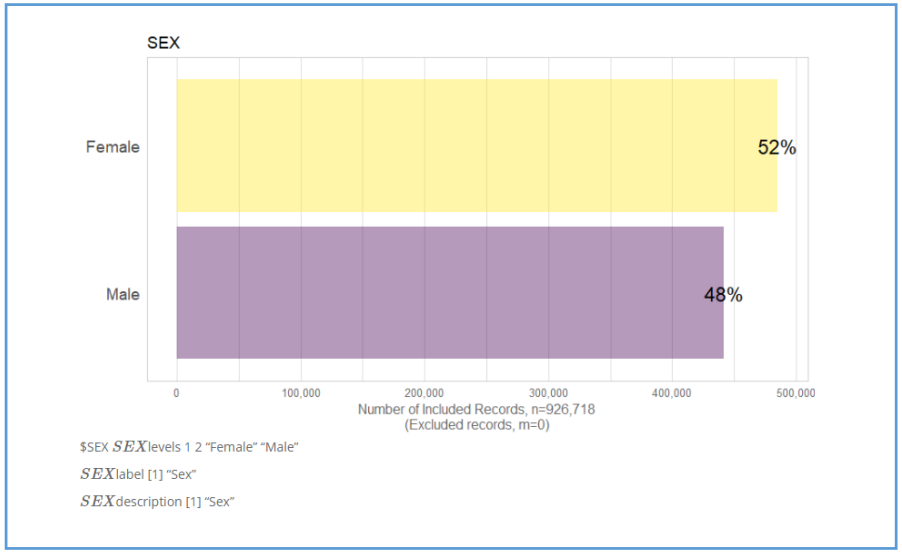
Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

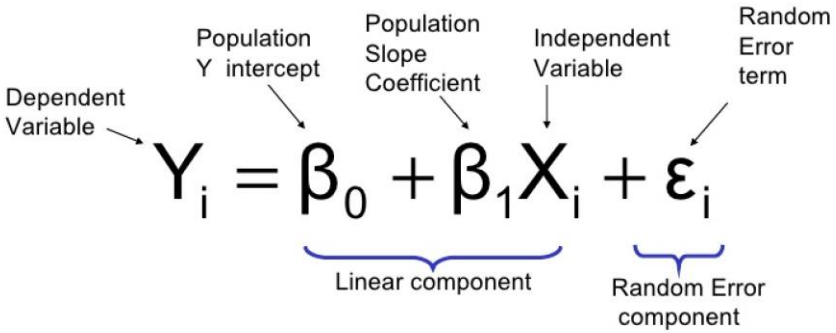
A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Sex





Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

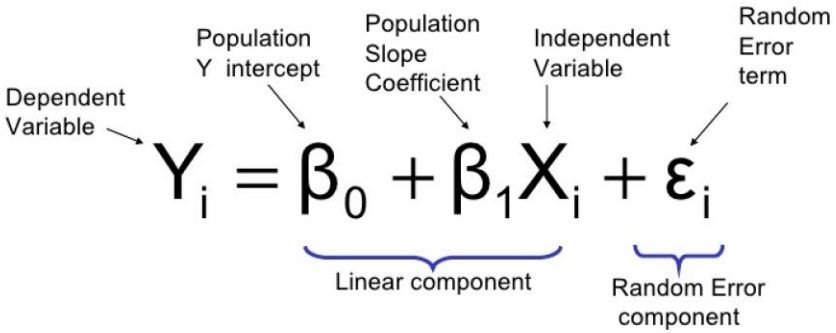
A. Graphing Technique

0.1 Modeling form

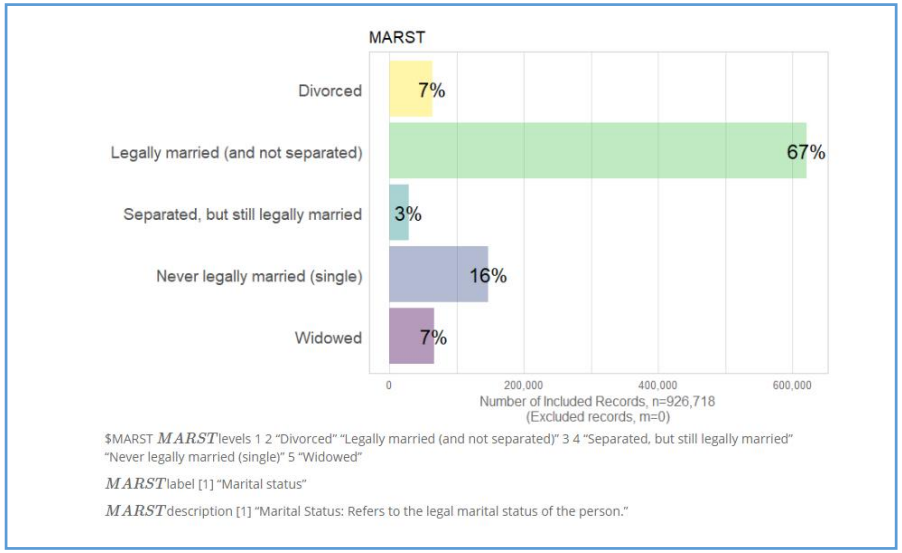
dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Marital Status

```
# because `still legaly married` is more legal than human
,marital = car::recode(
  MARST, "
  'Divorced'                              = 'sep_divorced'
  ;'Legally married (and not separated)'   = 'mar_cohab'
  ;'Separated, but still legally married'  = 'sep_divorced'
  ;'Never legally married (single)'        = 'single'
  ;'Widowed'                               = 'widowed'
  ")
,marital = factor(marital, levels = c(
  "sep_divorced","widowed","single","mar_cohab"))
```



Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5



$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

A. Graphing Technique

0.1 Modeling form

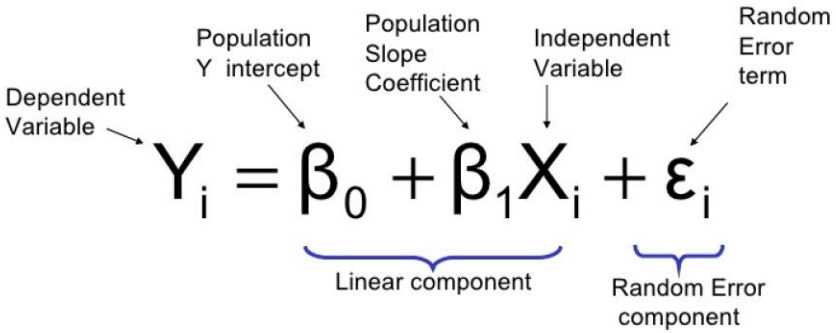
dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Highest Degree

```
# because even only 5 may be too granular for our purposes
,educ3 = car::recode(
  HCDD, "
   'None'                                                                                                 = 'less than high school'
  ;'High school graduation certificate or equivalency certificate'                                        = 'high school'
  ;'Other trades certificate or diploma'                                                                  = 'high school'
  ;'Registered apprenticeship certificate'                                                                = 'more than high school'
  ;'College, CEGEP or other non-university certificate or diploma from a program of 3 months to less than 1 year' = 'more than high school'
  ;'College, CEGEP or other non-university certificate or diploma from a program of 1 year to 2 years'   = 'more than high school'
  ;'College, CEGEP or other non-university certificate or diploma from a program of more than 2 years'   = 'more than high school'
  ;'University certificate or diploma below bachelor level'                                              = 'more than high school'
  ;'Bachelors degree'                                                                                     = 'more than high school'
  ;'University certificate or diploma above bachelor level'                                              = 'more than high school'
  ;'Degree in medicine, dentistry, veterinary medicine or optometry'                                      = 'more than high school'
  ;'Masters degree'                                                                                       = 'more than high school'
  ;'Earned doctorate degree'                                                                              = 'more than high school'
  ")
,educ3 = factor(educ3, levels = c(
  "less than high school"
  , "high school"
  , "more than high school"
)
)
```

```
# # because we want/need to inspect newly created variables
ds1 %>% group_by(educ3) %>% summarize(n = n())
```

```
# A tibble: 3 x 2
  educ3                      n
  <fct>                  <int>
1 less than high school  902326
2 high school           1403807
3 more than high school 2040516
```



Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

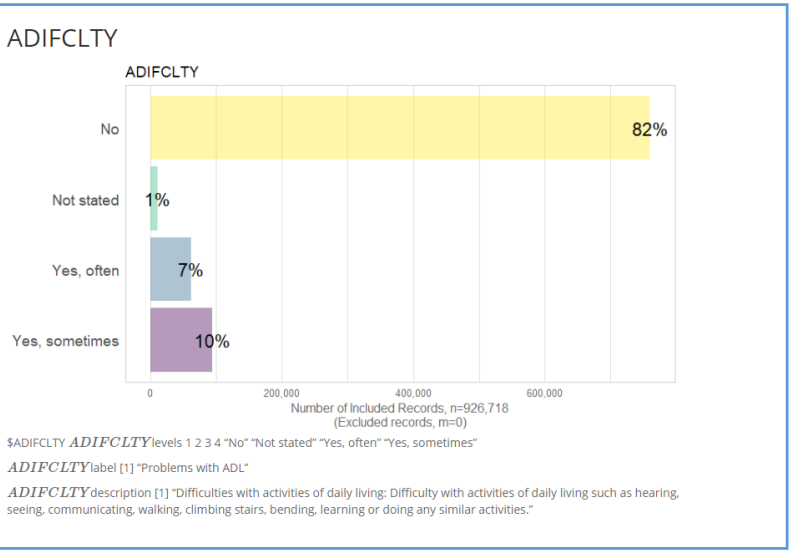
A. Graphing Technique

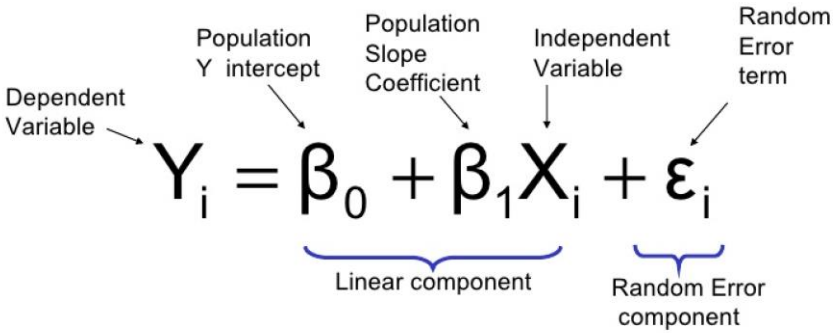
0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Activities of Daily Living

```
# ADIFCLTY              "Problems with ADL" (physical & cognitive)
# DISABFL               "Problems with ADL" (physical & social)
# because this is what counts practically
,poor_health = ifelse(ADIFCLTY %in% c("Yes, often","Yes, sometimes")
                      &
                        DISABFL %in% c("Yes, often","Yes, sometimes"),
                      TRUE, FALSE
)
,poor_health = factor(poor_health, levels = c("TRUE","FALSE"))
```





$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

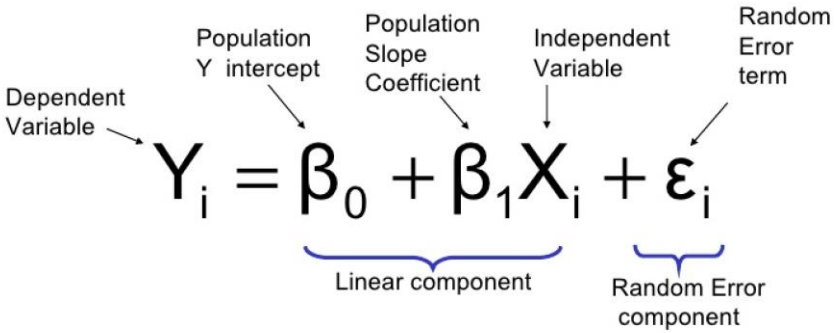
A. Graphing Technique

0.1 Modeling form

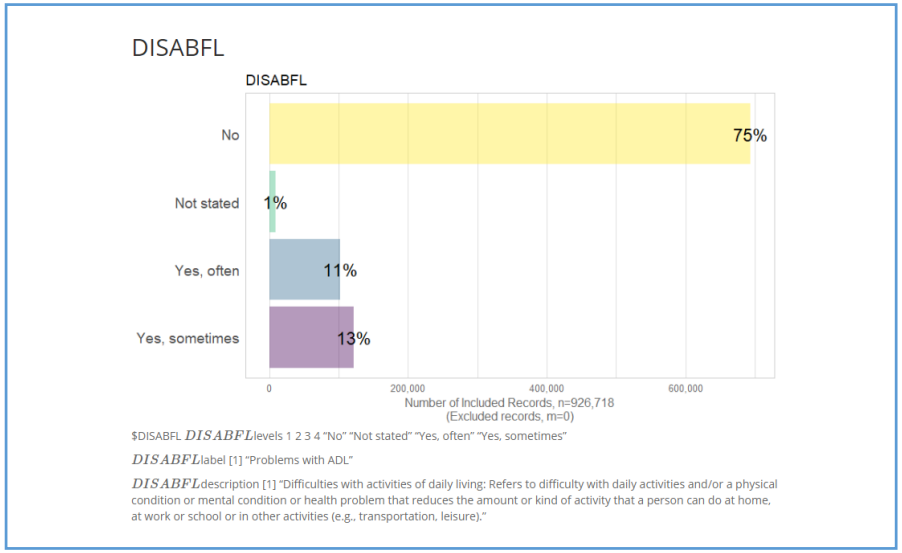
dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

Activities of Daily Living

```
# ADIFCLTY                "Problems with ADL" (physical & cognitive)
# DISABFL                 "Problems with ADL" (physical & social)
# because this is what counts practically
,poor_health = ifelse(ADIFCLTY %in% c("Yes, often","Yes, sometimes")
                      &
                        DISABFL %in% c("Yes, often","Yes, sometimes"),
                      TRUE, FALSE
)
,poor_health = factor(poor_health, levels = c("TRUE","FALSE"))
```



Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5



$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

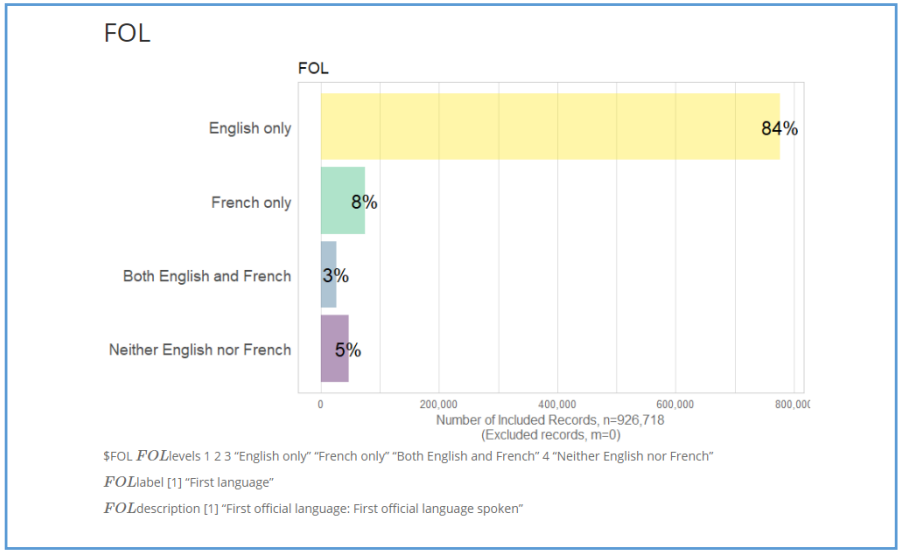
Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

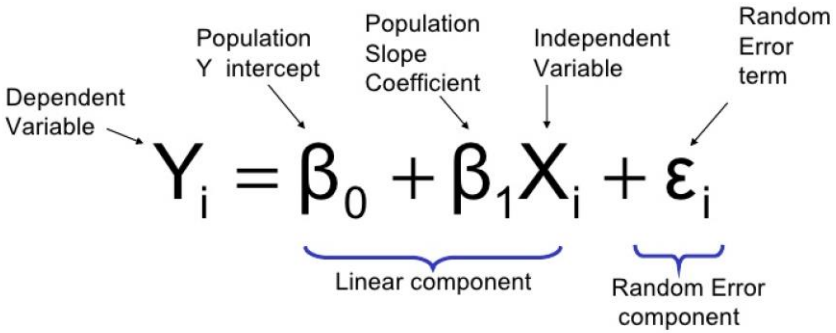
A. Graphing Technique

0.1 Modeling form

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

First Official Language





Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

A. Graphing Technique

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

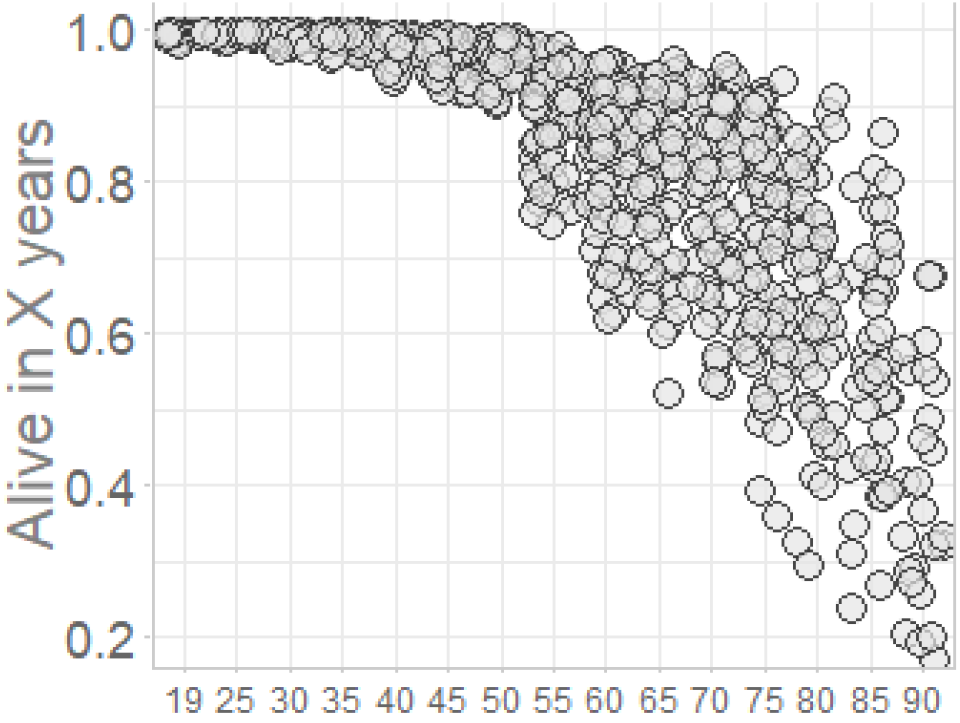
0.2 Graphical form

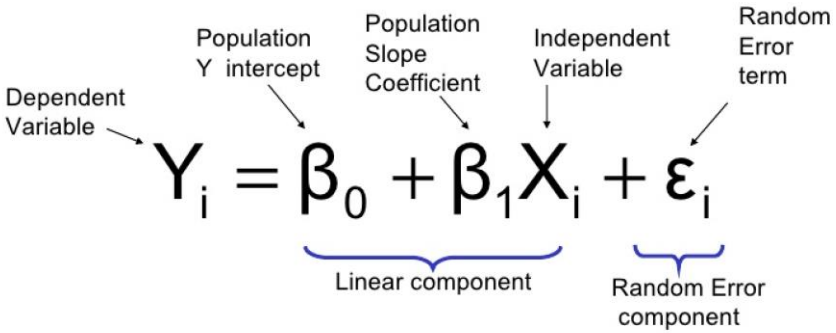
LEGEND

point = person
Y-axis = probability R is dead in X years
X-axis = age group (floor of 5-year category)

The higher the dot = the higher the chance to be alive in X years

Visualizing probability instead of log-odds because it is more intuitive





Source: https://towardsdatascience.com/how-are-logistic-regression-ordinary-least-squares-regression-related-1deab32d79f5

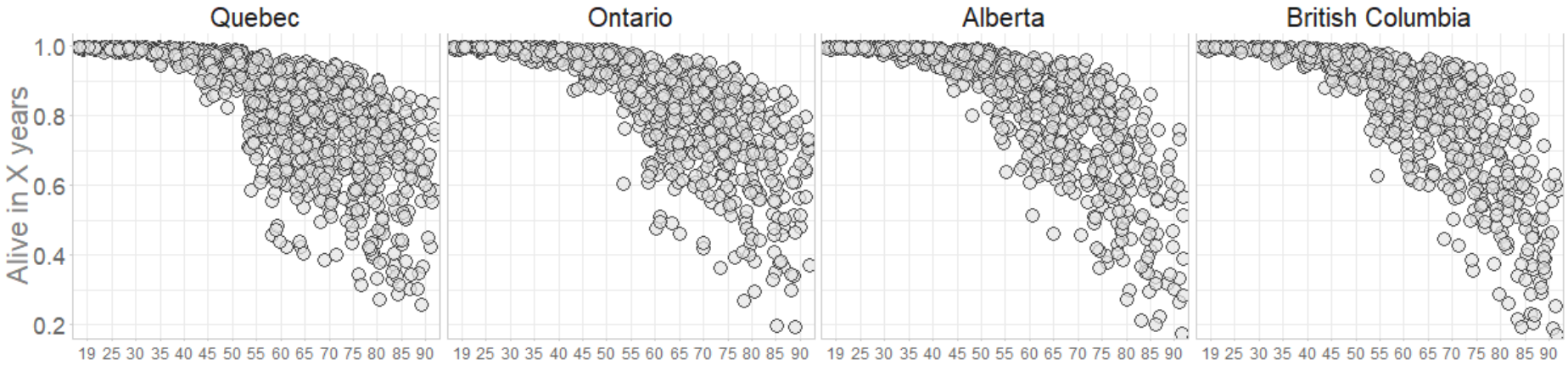
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1X_1 + b_2X_2 + \ldots + b_pX_p$$

Source: http://sphweb.bumc.bu.edu/otlt/MPH-Modules/BS/BS704_Multivariable/BS704_Multivariable8.html

A. Graphing Technique

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL

0.2 Graphical form



LEGEND

Facet = Province of residence

A. Graphing Technique

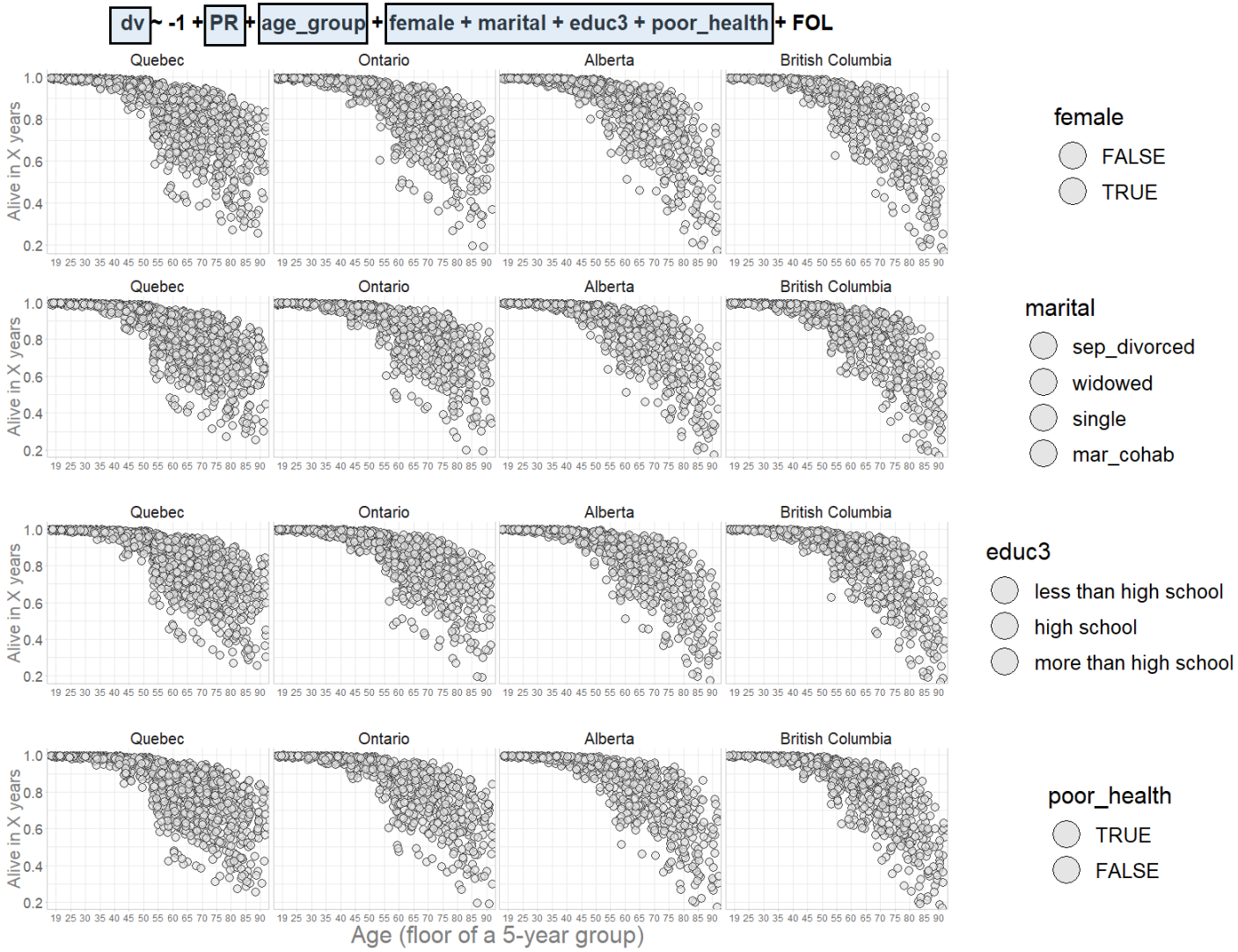
0.2 Graphical form

LEGEND

Rows = duplicate of each other (for now).

Notice that FOL is not displayed

The book is ready for coloring



A. Graphing Technique

0.3 Coloring book

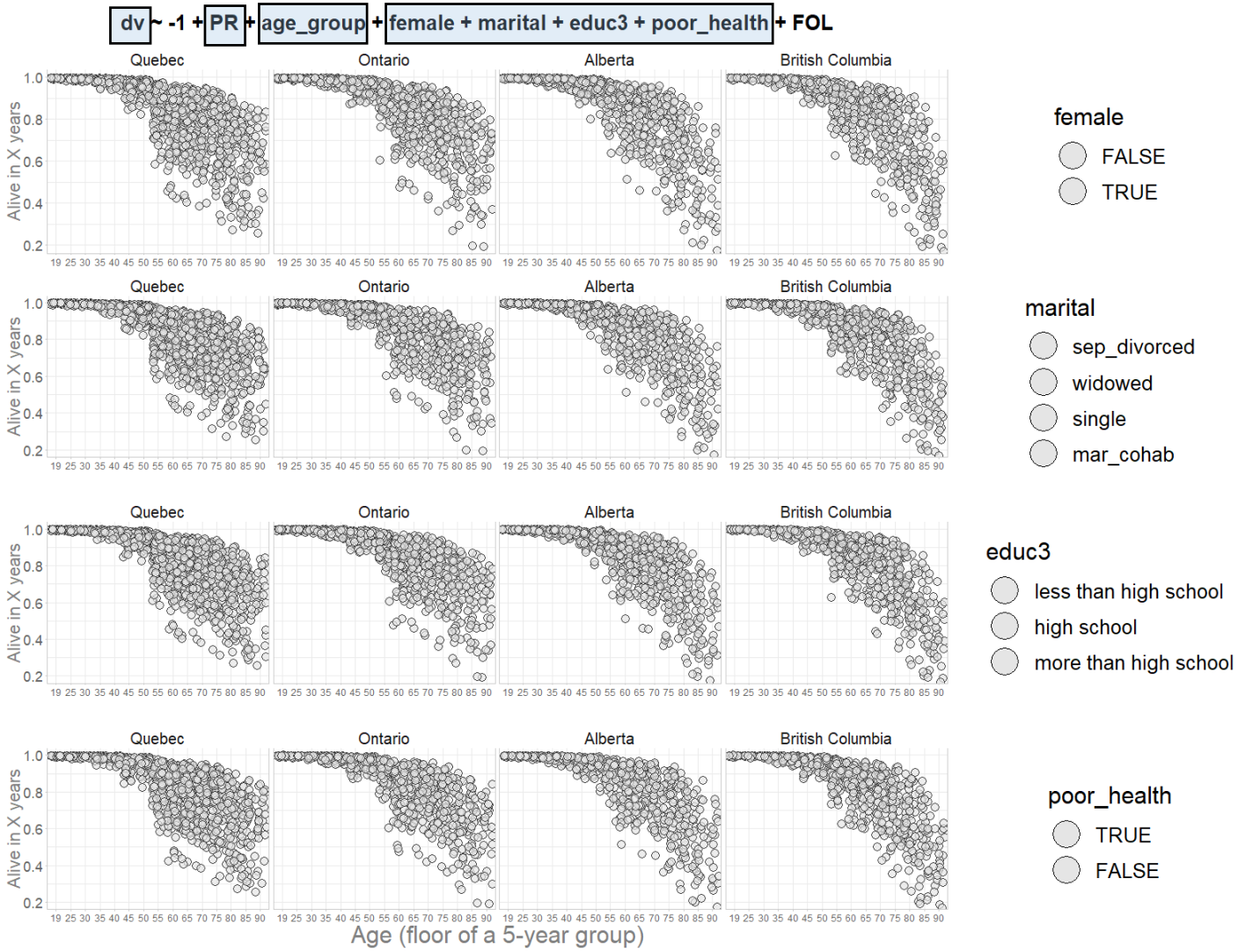
QUESTION

What should the "reference group" be for each predictor?

What do we expect based on existing research?

Informed expectation

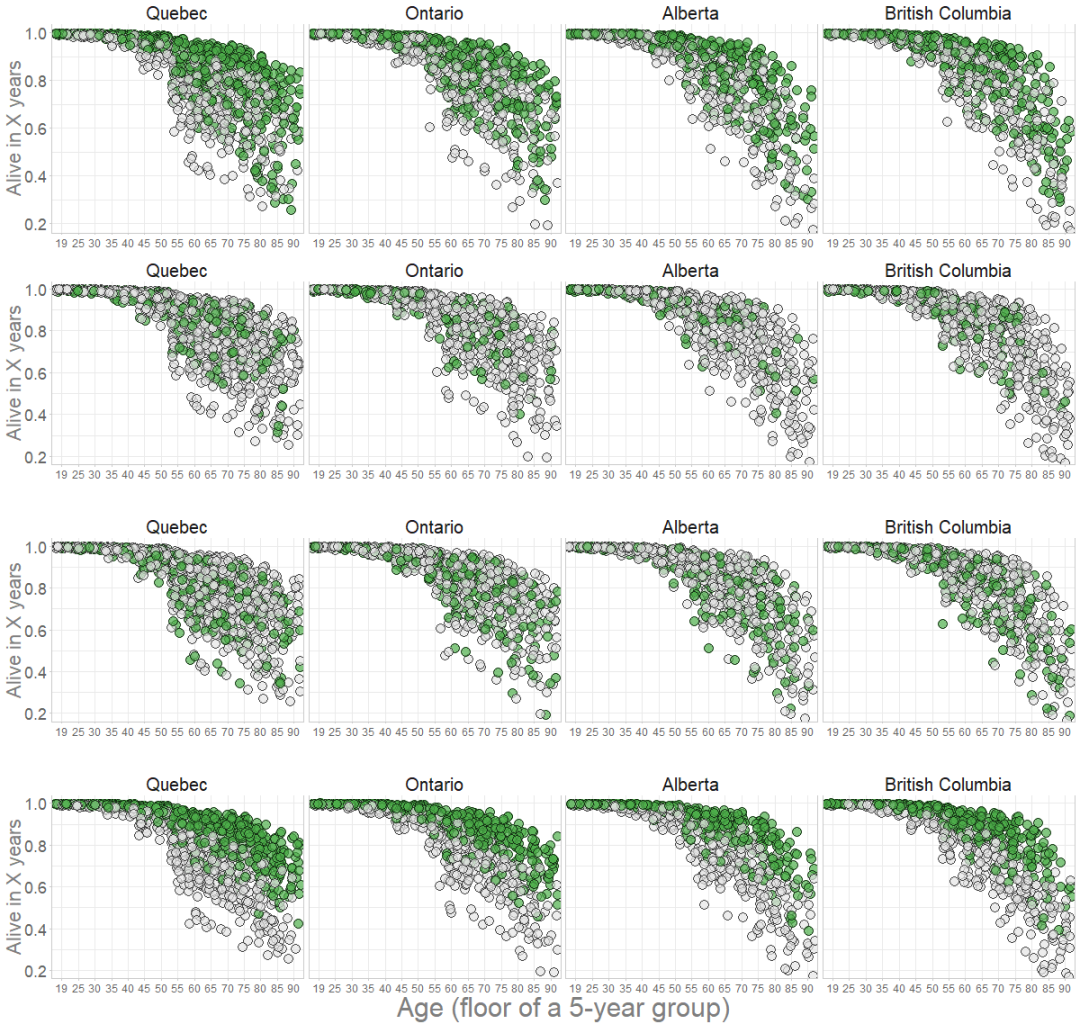
Reference group ?



A. Graphing Technique

0.3 Coloring book

dv ~ -1 + PR + age_group + female + marital + educ3 + poor_health + FOL



female
- FALSE
- TRUE

marital
- sep_divorced
- widowed
- single
- mar_cohab

educ3
- less than high school
- high school
- more than high school

poor_health
- TRUE
- FALSE

Informed expectation

Reference group

A. Graphing Technique

0.3 Coloring book

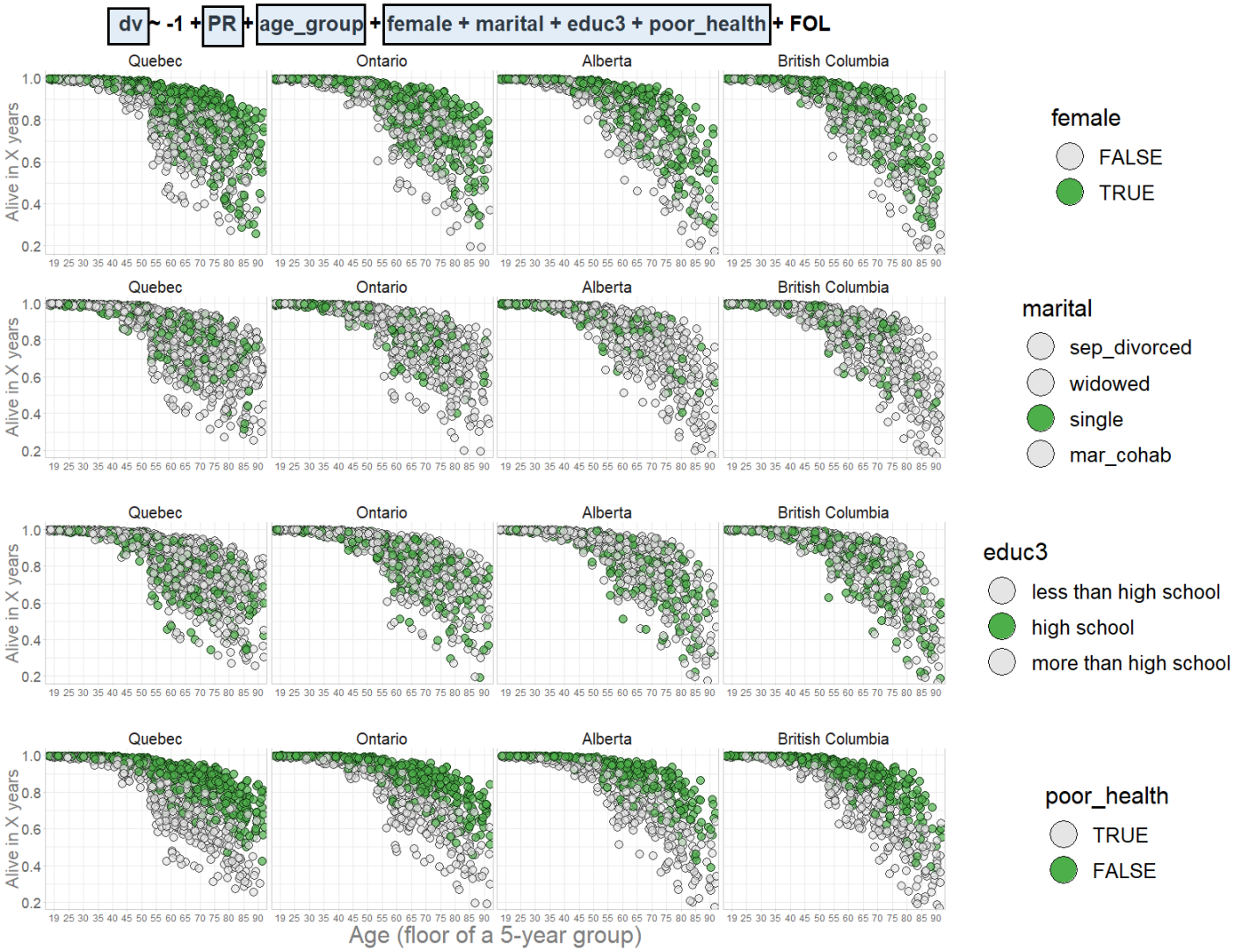
QUESTION

Compared to reference group, what levels of predictors are expected to **increase** the mortality risk?

Informed expectation

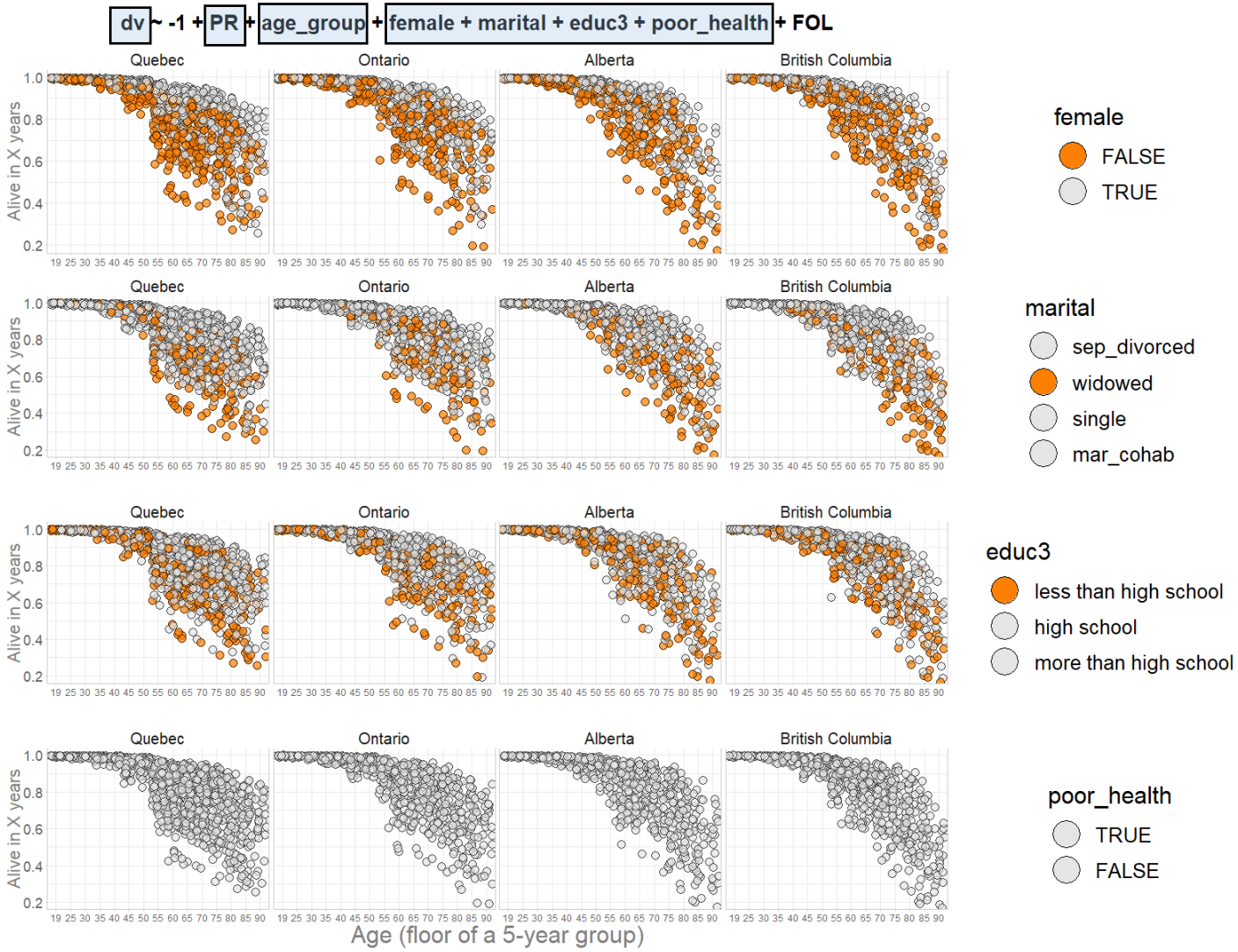
Moderately increased risk ?

Reference group



A. Graphing Technique

0.3 Coloring book



Informed expectation

Moderately increased risk

Reference group

A. Graphing Technique

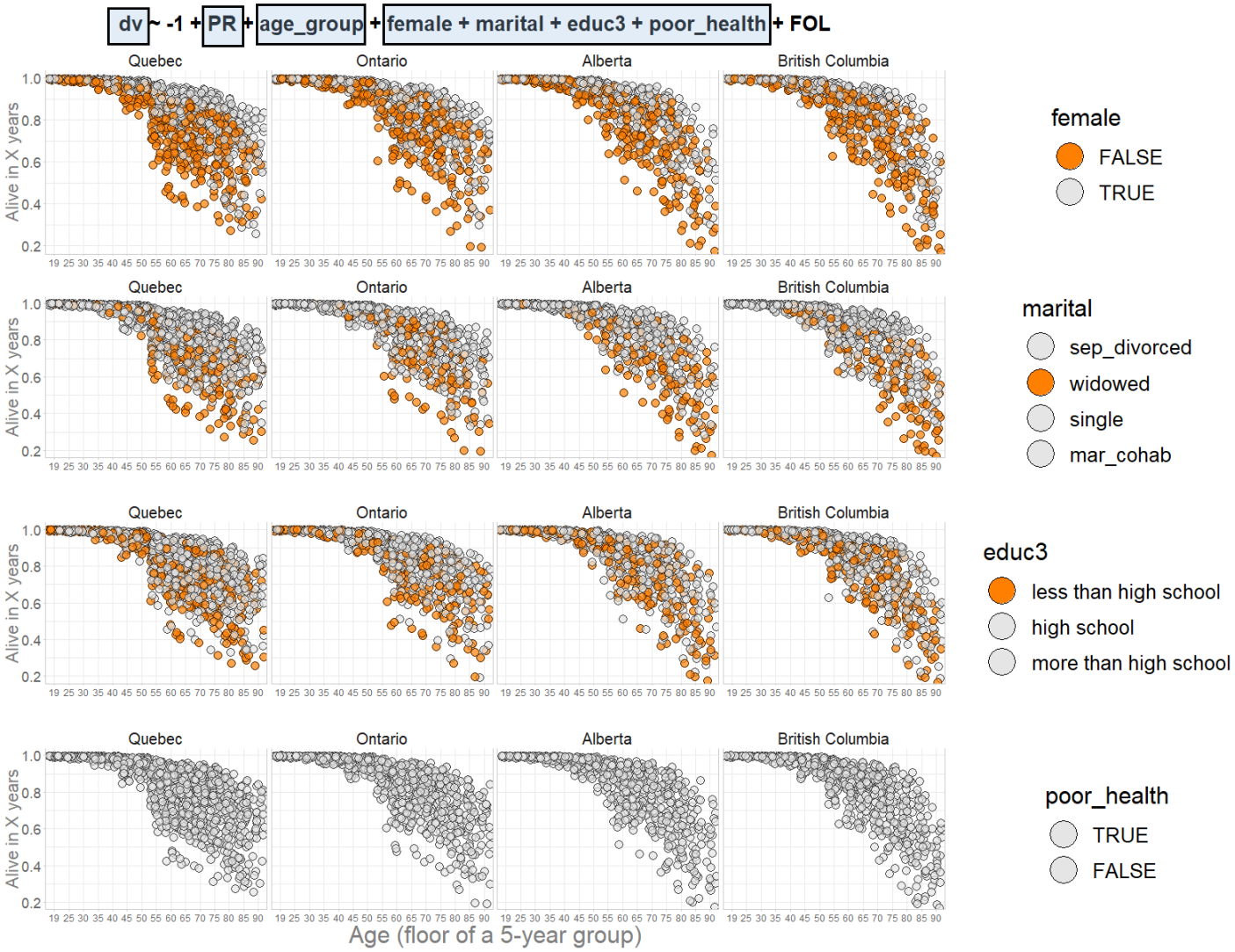
0.3 Coloring book

QUESTION

Compared to reference group, what levels of predictors are expected to **decrease** the mortality risk?

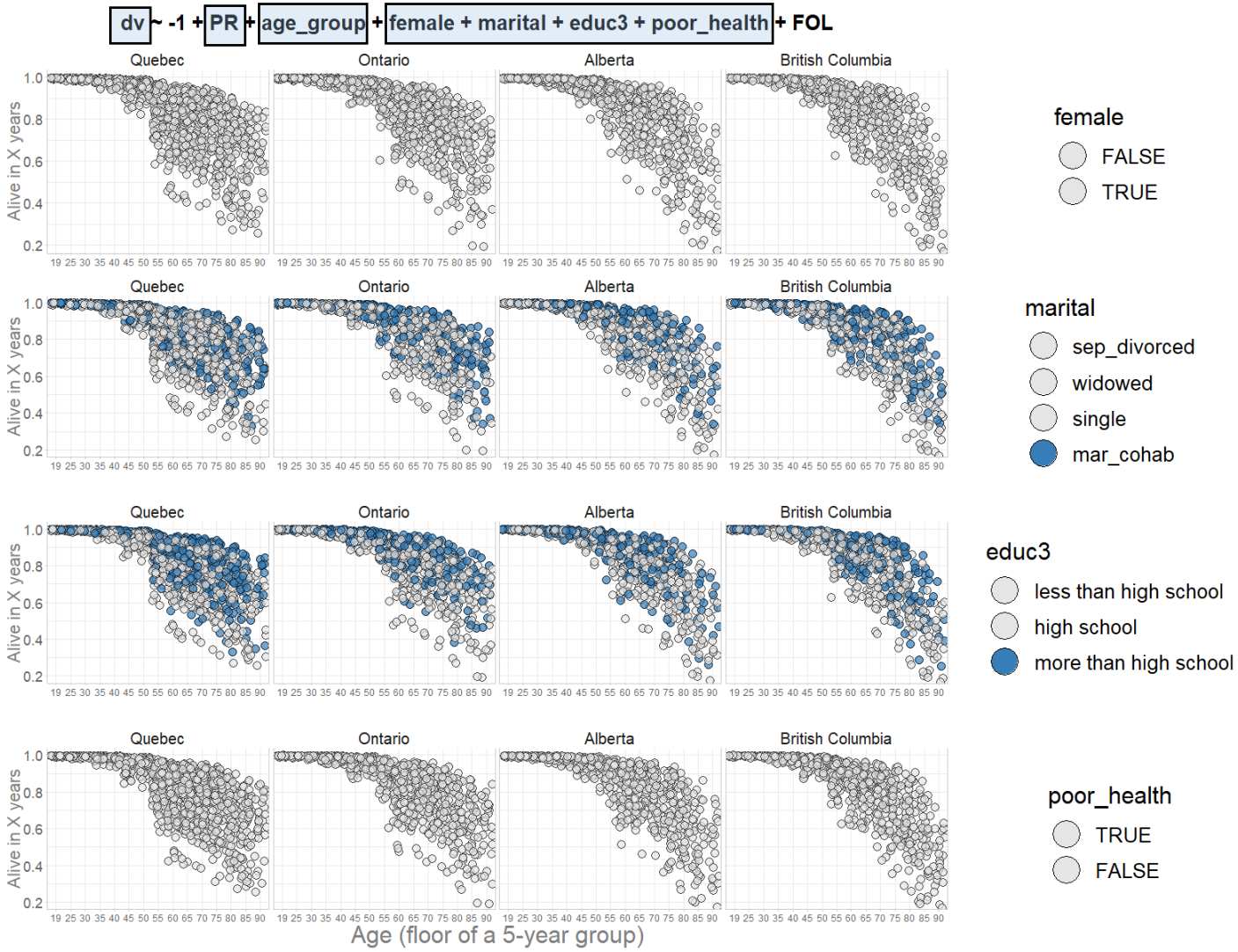
Informed expectation

- Moderately increased risk
- Reference group
- Moderately decreased risk ?



A. Graphing Technique

0.3 Coloring book



Informed expectation

- Moderately increased risk
- Reference group
- Moderately decreased risk

A. Graphing Technique

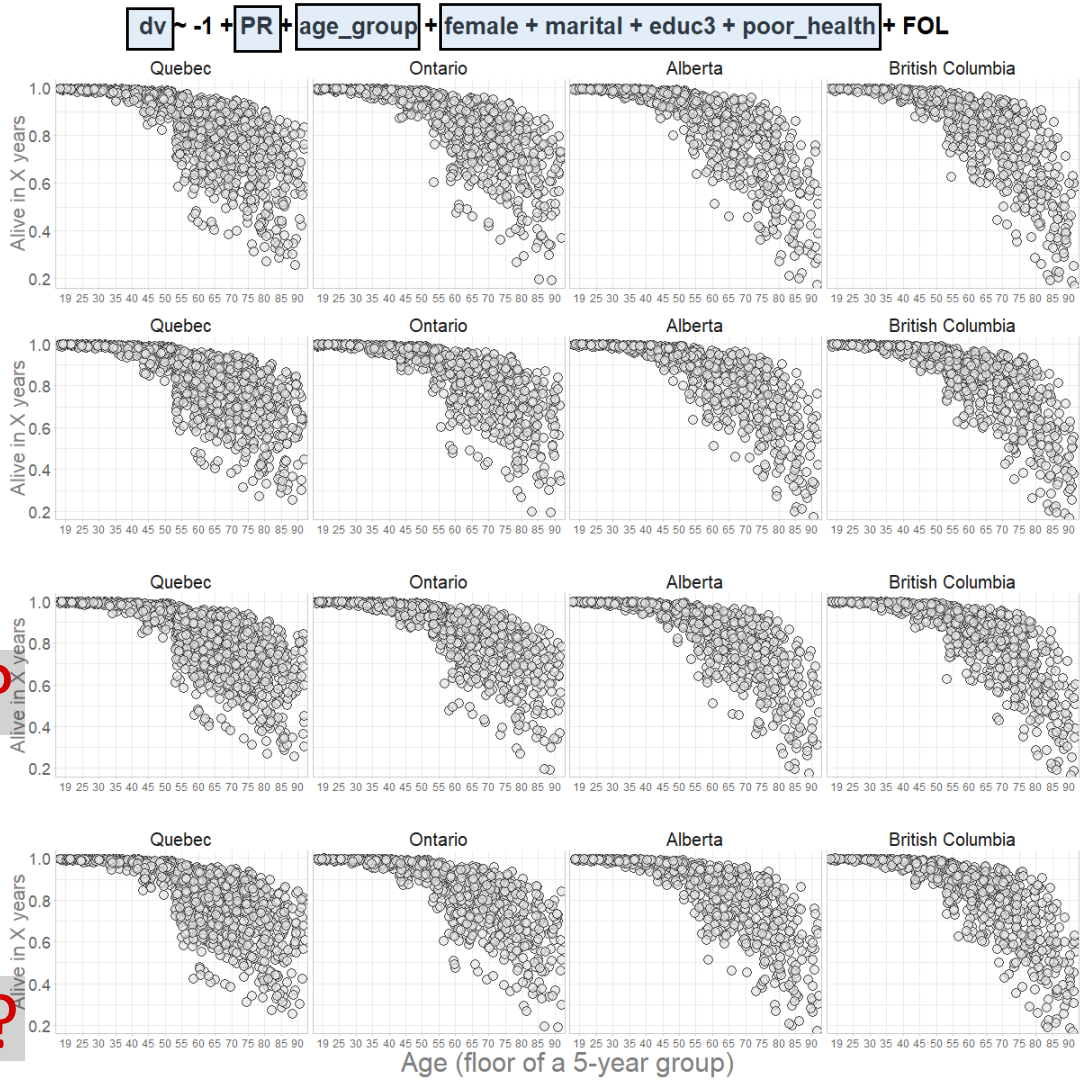
0.3 Coloring book

QUESTION

What levels of predictors are expected to affect mortality risk drastically?

Informed expectation

- Substantially increased risk ?
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk ?



female
- FALSE
- TRUE

marital
- sep_divorced
- widowed
- single
- mar_cohab

educ3
- less than high school
- high school
- more than high school

poor_health
- TRUE
- FALSE

A. Graphing Technique

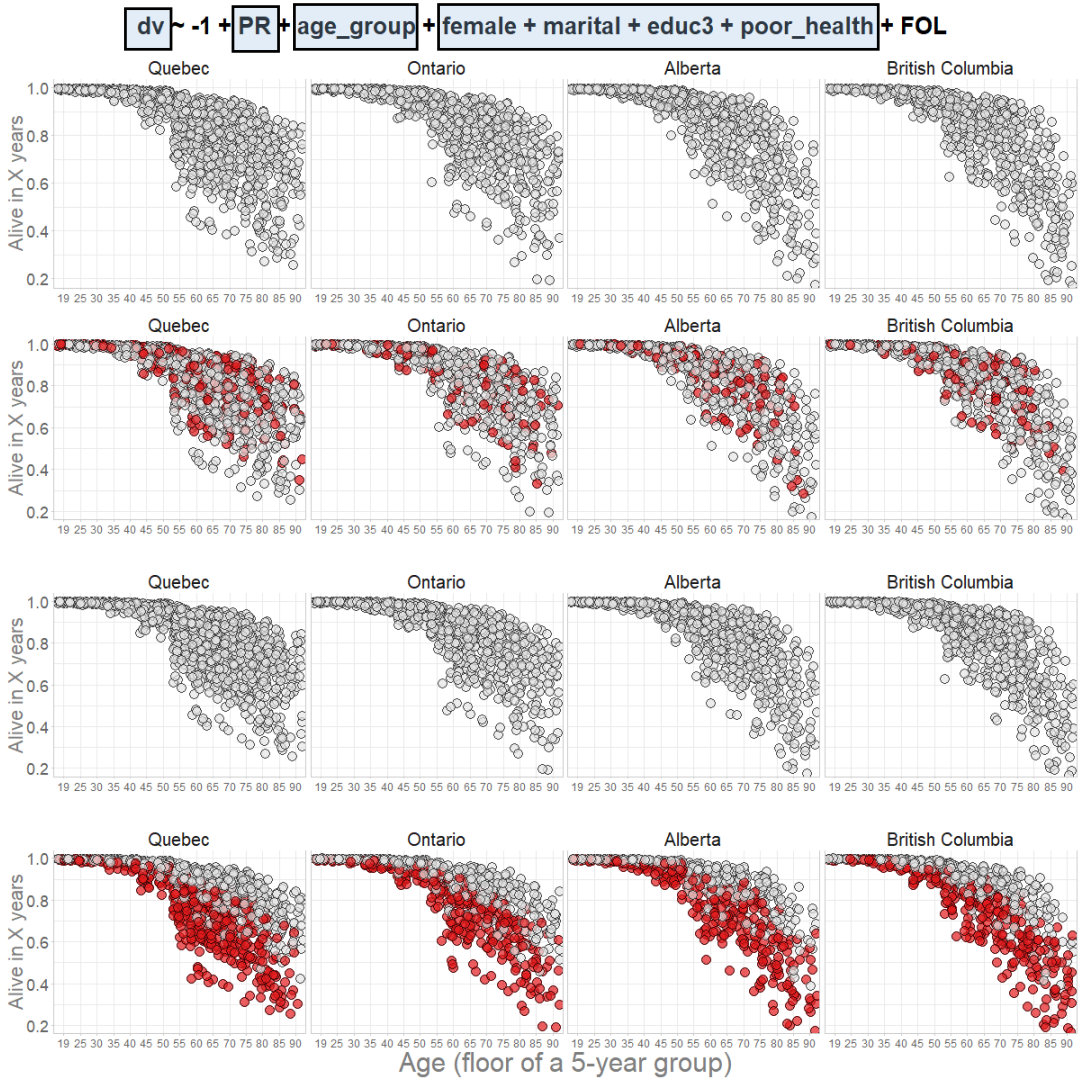
0.3 Coloring book

QUESTION

What levels of predictors are expected to affect mortality risk drastically?

Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



A. Graphing Technique

0.3 Coloring book

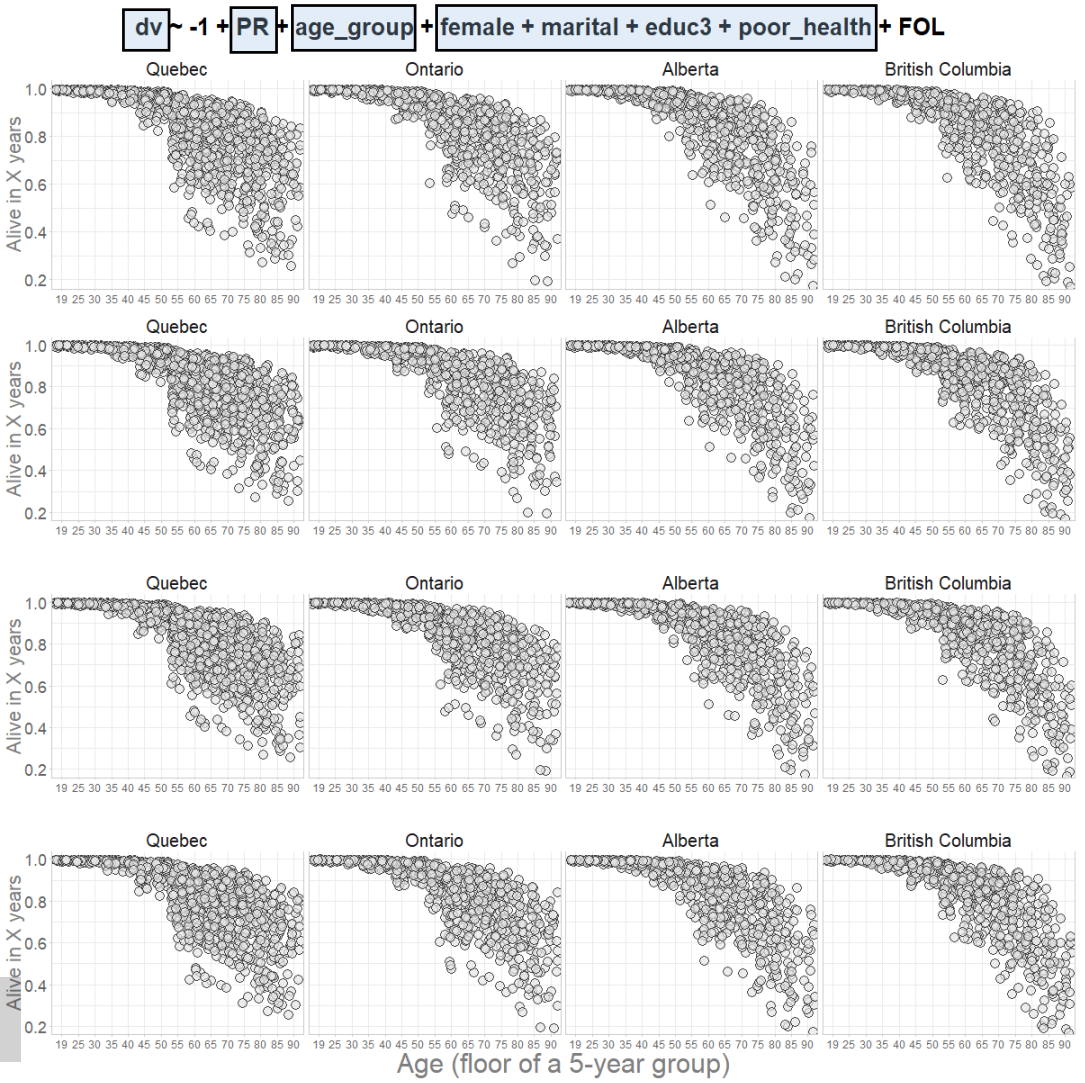
QUESTION

What levels of predictors are expected to affect mortality risk drastically?

No “very bad” and it’s ok.

Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



A. Graphing Technique

0.3 Coloring book

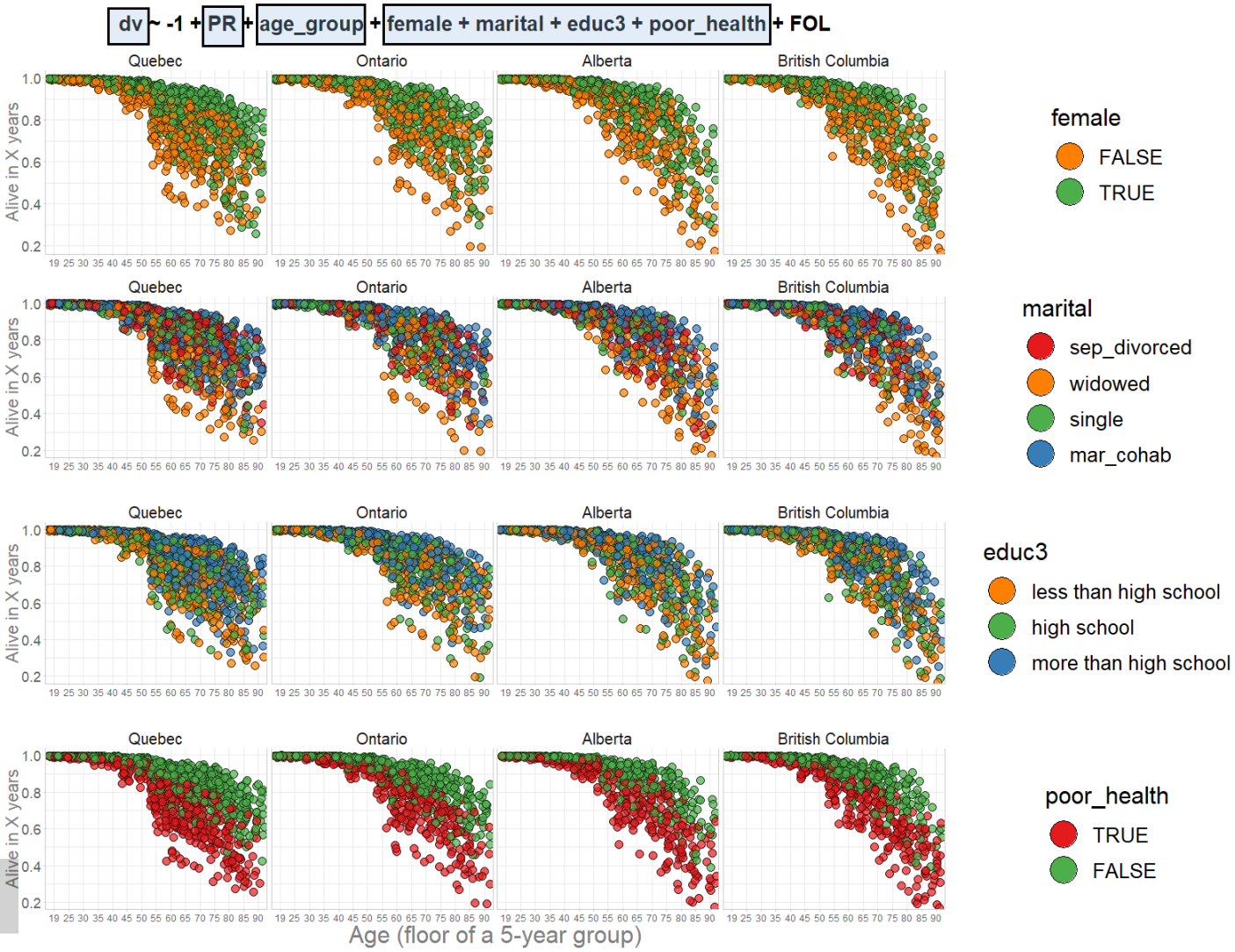
NOTICE

Plotting all colors at once may not be as informative as one would expect

May require too much tweaking to make useful

Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



A. Graphing Technique

0.3 Coloring book

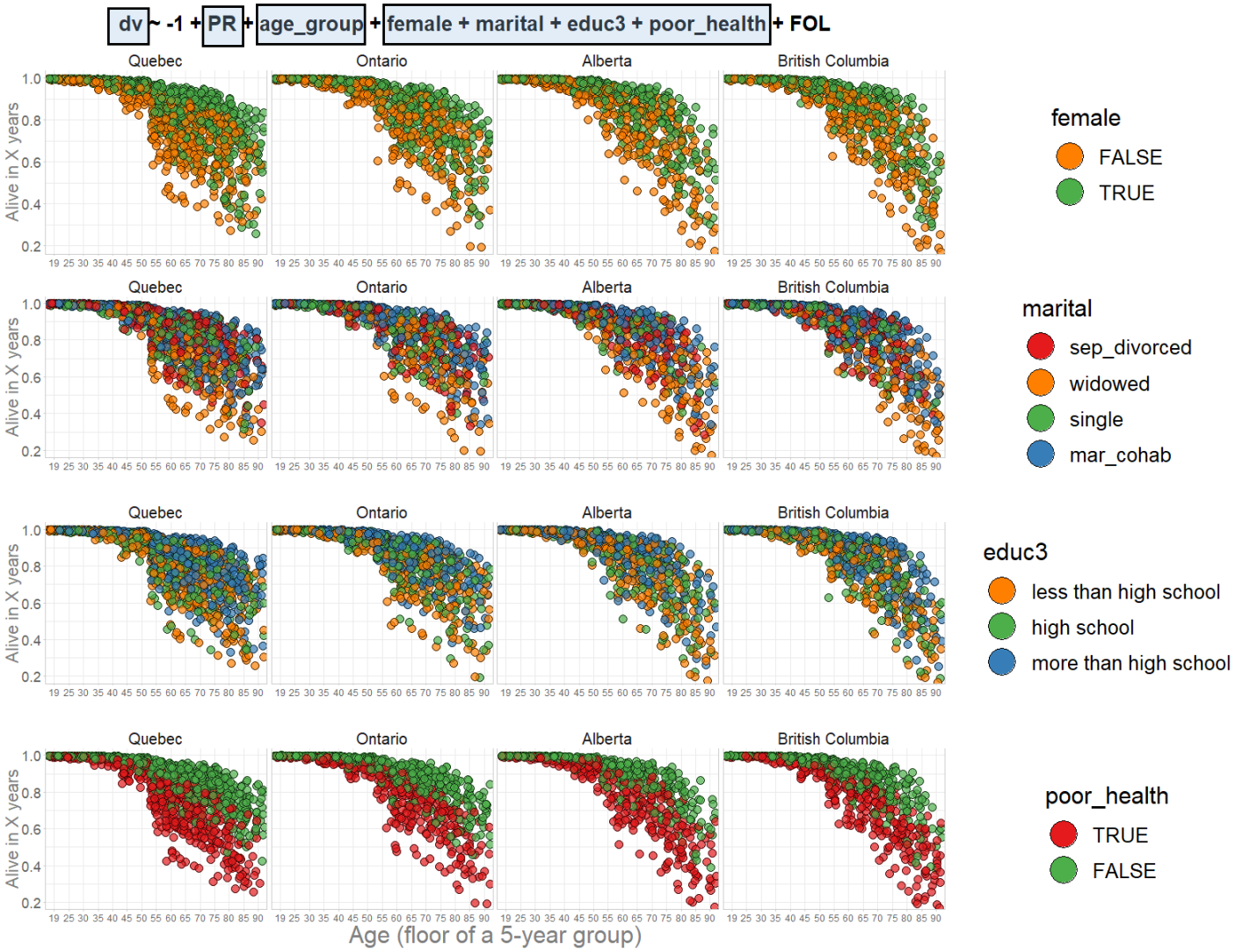
NOTICE

Note all predictors are worth visualizing, some are there for control.

We can adjust what is being displayed

Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



A. Graphing Technique

0.3 Coloring book

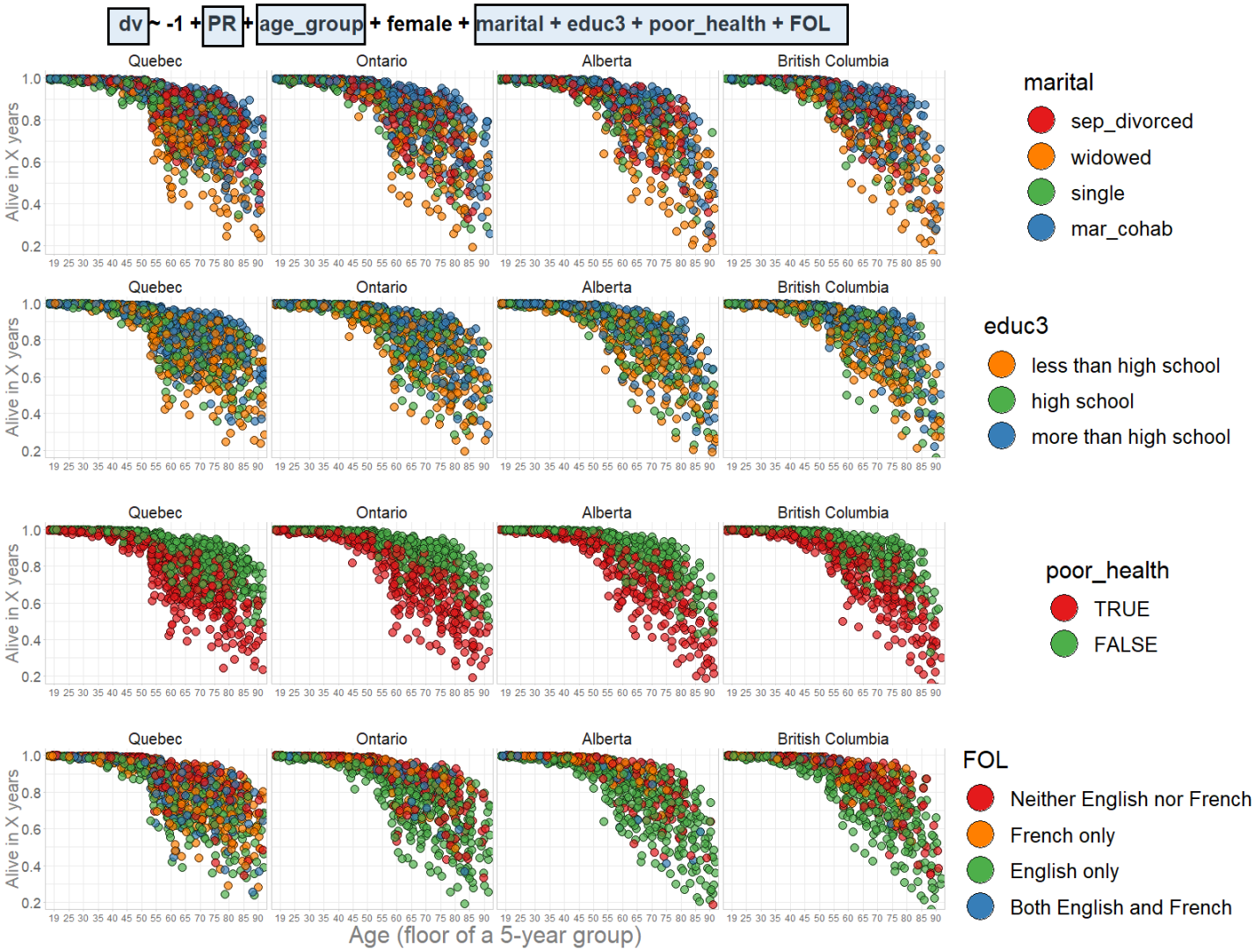
NOTICE

Note all predictors are worth visualizing, some are there for control.

We can adjust what is being displayed

Informed expectation

- Substantially increased risk
- Moderately increased risk
- Reference group
- Moderately decreased risk
- Substantially decreased risk



Let us try to reproduce these graphs!

Power of Population Data Science webinar series. Population Data BC, 2018-11-01 (slides, video)

Today

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon
0.1 Modeling form: univariate logistic regression with categorical predictors
0.2 Graphical form: faceted scatterplot in ggplot2
0.3 Coloring book: Mapping informed expectations from predictors onto color

B. Workflow Highlights

1.0 "Let no one ignorant of geometry enter": (my) scripts were written to be read by humans
1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects
1.2 Layers of Isolation: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)
1.3 Two essential means of production: knitr::stitch() vs rmarkdown::render()

Code along by forking github.com/andkov/ipdln-2018-hackathon

B. Workflow Highlights

1.0 “Let no one ignorant of geometry enter”: (my) scripts were written to be read by humans

Donald Knuth. "Literate Programming (1984)" in Literate Programming. CSLI, 1992, pg. 99.

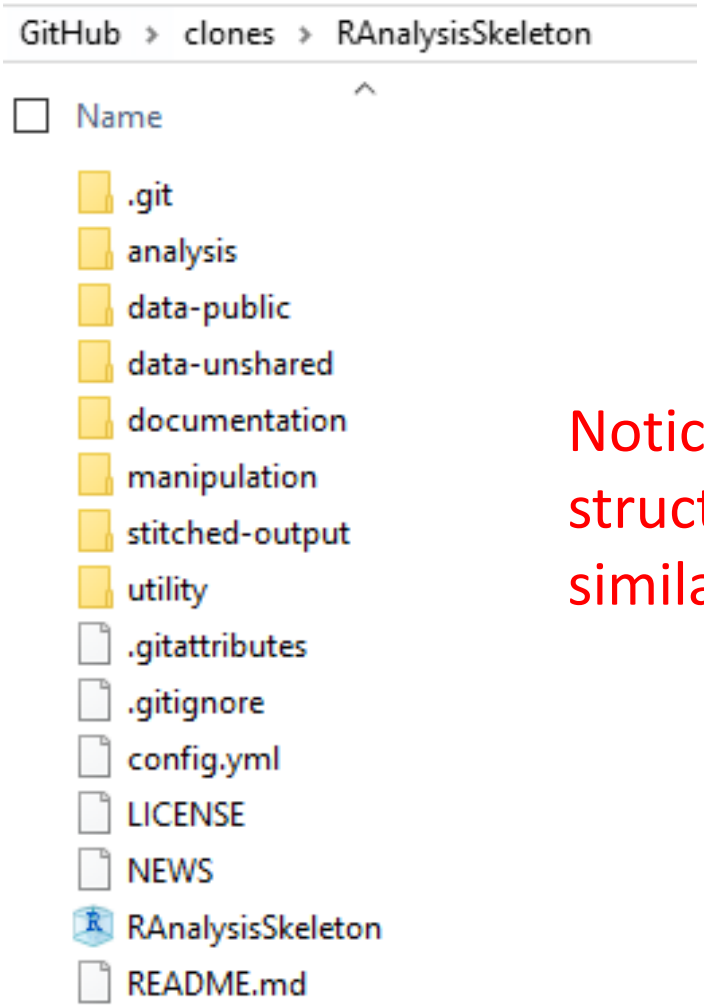
I believe that the time is ripe for significantly better documentation of programs, and that we can best achieve this by considering programs to be works of literature. Hence, my title: "Literate Programming."

Let us change our traditional attitude to the construction of programs: Instead of imagining that our main task is to instruct a computer what to do, let us concentrate rather on explaining to human beings what we want a computer to do.

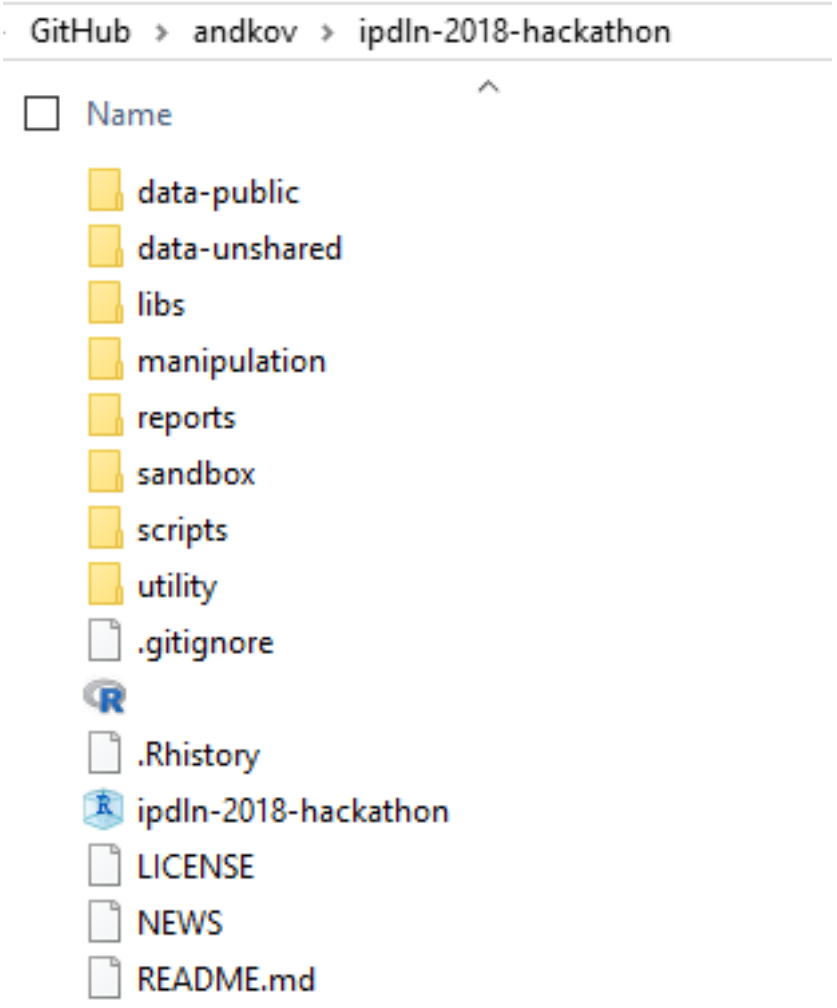
Source: http://www.literateprogramming.com/

B. Workflow Highlights

1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects



Notice structural similarities



Keeps recognizable structure over projects

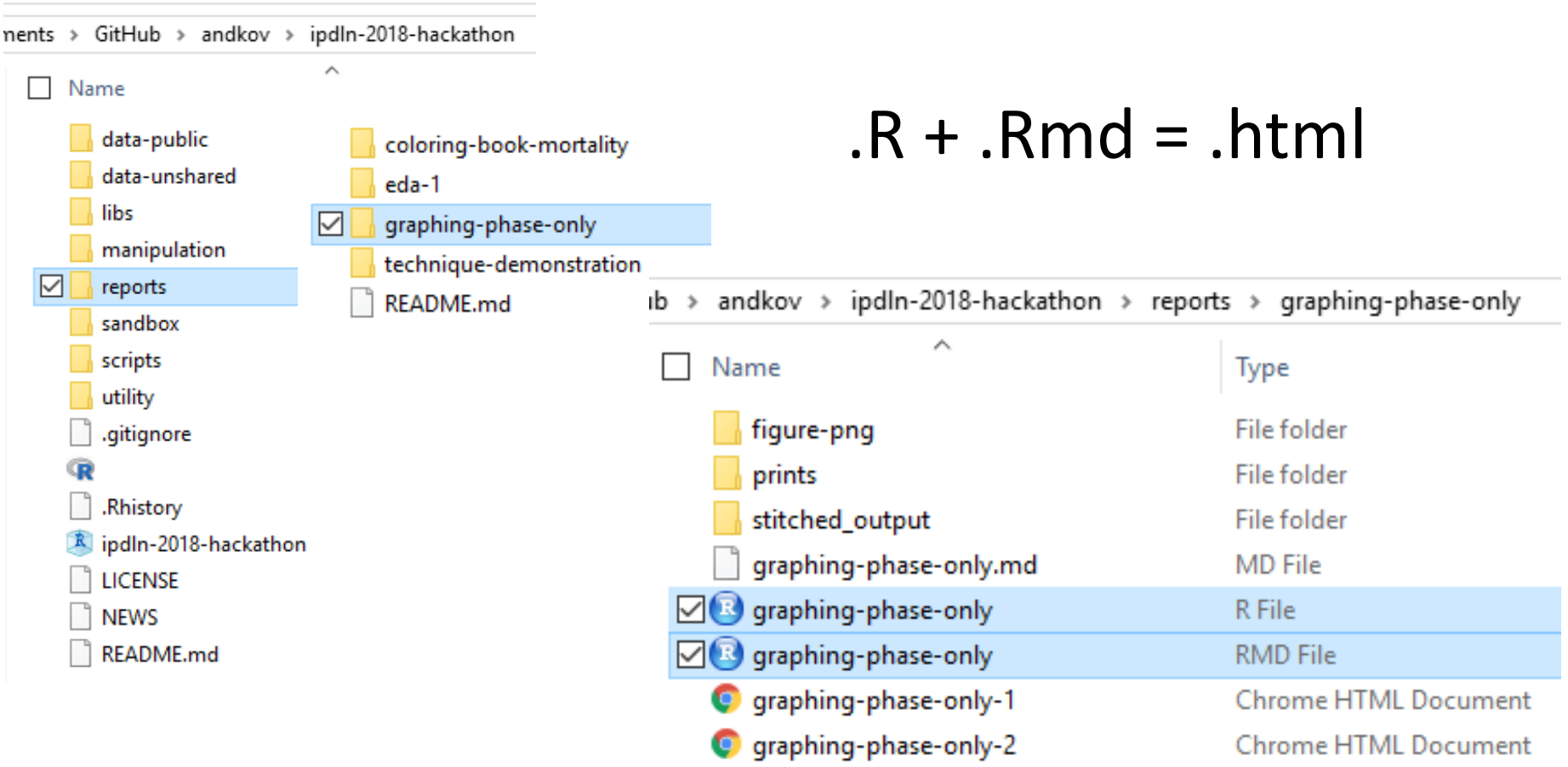
.R
stores analysis
(what really happens)

.Rmd
stores presentation
(how you tell about it)

.R + .Rmd = .html

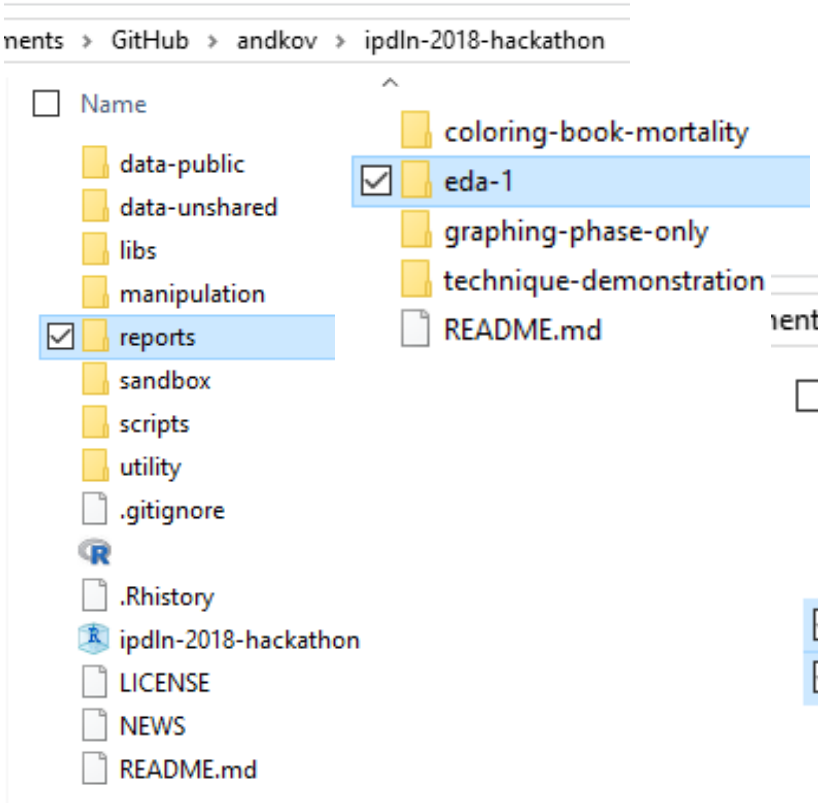
B. Workflow Highlights

1.2 Layers of Isolation: analysis vs presentation



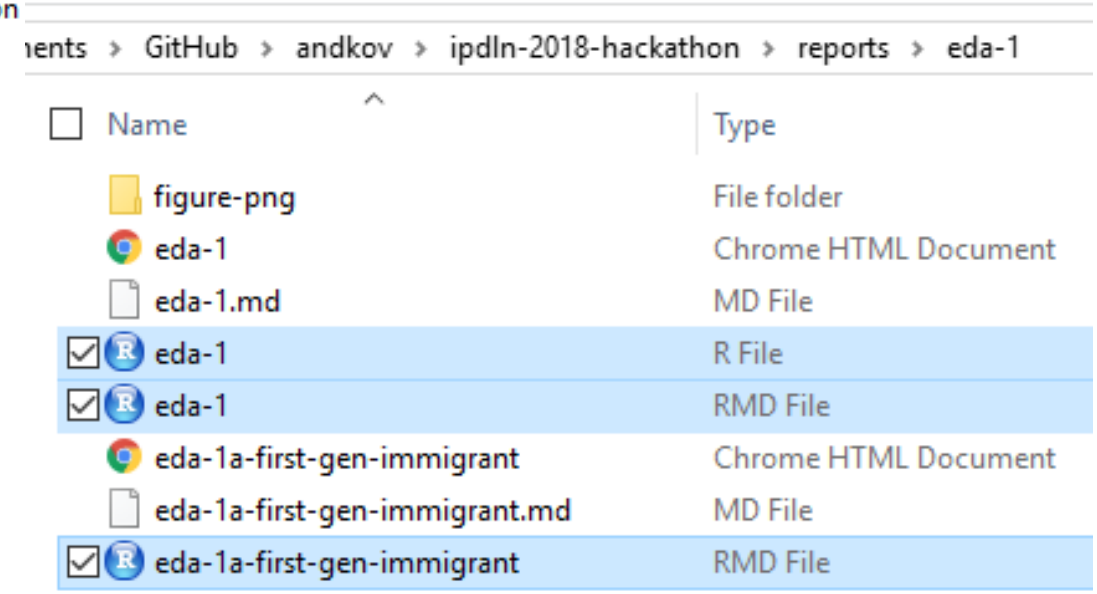
B. Workflow Highlights

1.2 Layers of Isolation: analysis vs presentation



$$.R + .Rmd_1 = .html_1$$

$$.R + .Rmd_2 = .html_2$$

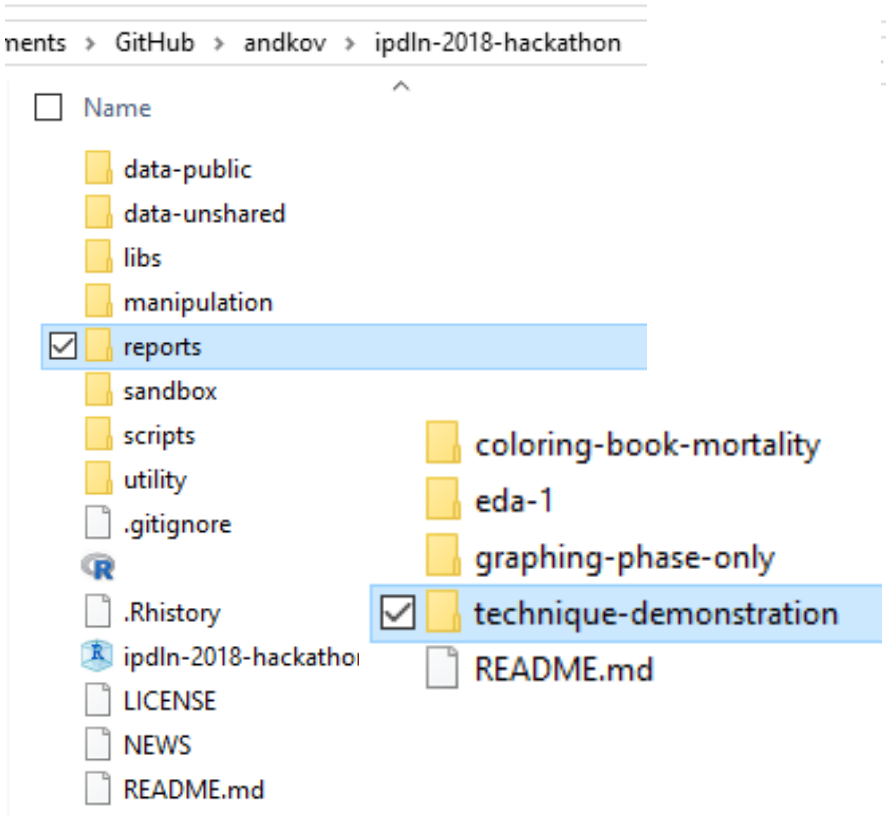


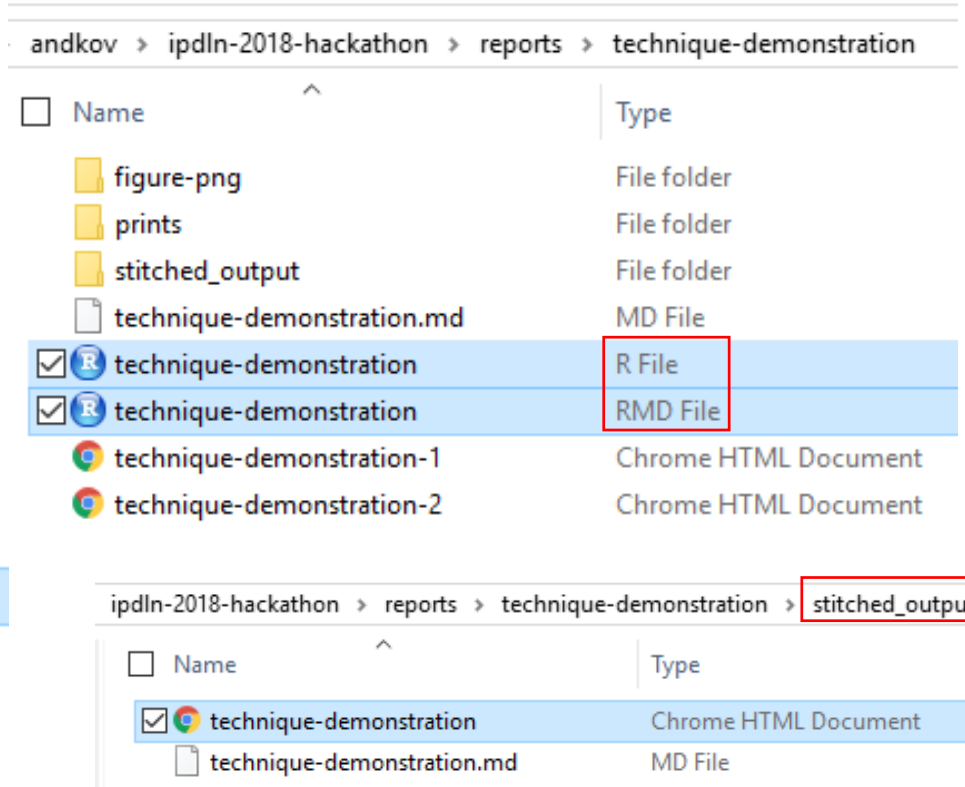
B. Workflow Highlights

1.3 Two essential means of production

rmarkdown::render(.R + .Rmd) = .html

knitr::stitch(.R) = .html





Conclusions

A. Graphing Technique

0.0 Data & Context : Mortality factors of Canadian immigrants at IPDLN-2018 hackathon
0.1 Modeling form: univariate logistic regression with categorical predictors
0.2 Graphical form: faceted scatterplot in ggplot2
0.3 Coloring book: Mapping informed expectations from predictors onto color

B. Workflow Highlights

1.0 "Let no one ignorant of geometry enter": (my) scripts were written to be read by humans
1.1 RAnalysisSkeleton by Will Beasley: basic starting point for reproducible projects
1.2 Layers of Isolation: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)
1.3 Two essential means of production: knitr::stitch() vs rmarkdown::render()

Code along by forking github.com/andkov/ipdln-2018-hackathon

Visualizing Logistic Regression with a “coloring book” technique:

A study in ggplot2

Please email questions to
aging@uvic.ca

Andriy
Koval

British Columbia
OBSERVATORY
Population & Public Health

CIHR IRSC
Canadian Institutes of Health Research / Instituts de recherche en santé du Canada

UCF
UNIVERSITY OF CENTRAL FLORIDA

Displaying Health Data

Cases, Techniques, Solutions

Colloquium + Live-Webcast + Recording
Medical Sciences Building (MBS) 160
University of Victoria
November 28 – 30 , 1 – 3 pm PST

Please email questions to
aging@uvic.ca

University of Victoria
INSTITUTE ON AGING & LIFELONG HEALTH
island health
British Columbia BSERVATORY Population & Public Health
IALSA
Health Sciences Center
UCF UNIVERSITY OF CENTRAL FLORIDA