

When Notebooks are not Enough: Constructing Workflows for Reproducible Analytics

Andriy Koval

Matrix Institute Colloquium Series

University of Victoria

2018-10-31

github.com/andkov/ipdIn-2018-hackathon

When notebooks are not enough

Last time at the Matrix Institute (2018-10-17)

- (Data) Science is about creating **software**!
- **Tradeoff** “Exploration vs Engineering”
- **Limitations** of Notebooks (by Neil Ernst)
 - Parameter configuration
 - Hidden state
 - Longevity and version control
 - Testing and modularity
 - Notebook carpentry

Today: Do reproducible projects overcome these limitations?

A. Graphing Technique

- 0.0 **Data & Context** : Mortality factors of Canadian immigrants at [IPDLN-2018 hackathon](#) by Statistics Canada in Banff
- 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 **Autonomous phases**: data cleaning, statistical modelling, graph production
- 1.3 **Layers of Isolation**: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)
- 1.4 Two essential **means of production**: [knitr::stitch\(\)](#) vs [rmarkdown::render\(\)](#)

C. Conclusions

- 2.0 **Different than Notebooks**: sacrifices simplicity for agility via layers of isolation
- 2.1 **R (+ .Rmd) = .html (+ .pdf)** : moving away from *data playing* towards *data science*
- 2.2 **Reproducible projects**: moving away from notebooks towards software
- 2.3 **Looking back** to Neil Ernst talk:
 - Parameters and configuration
 - Hidden state
 - Longevity and version control
 - Testing and modularity
 - Notebook carpentry

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



A. Graphing Technique

0.1 Modeling form

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

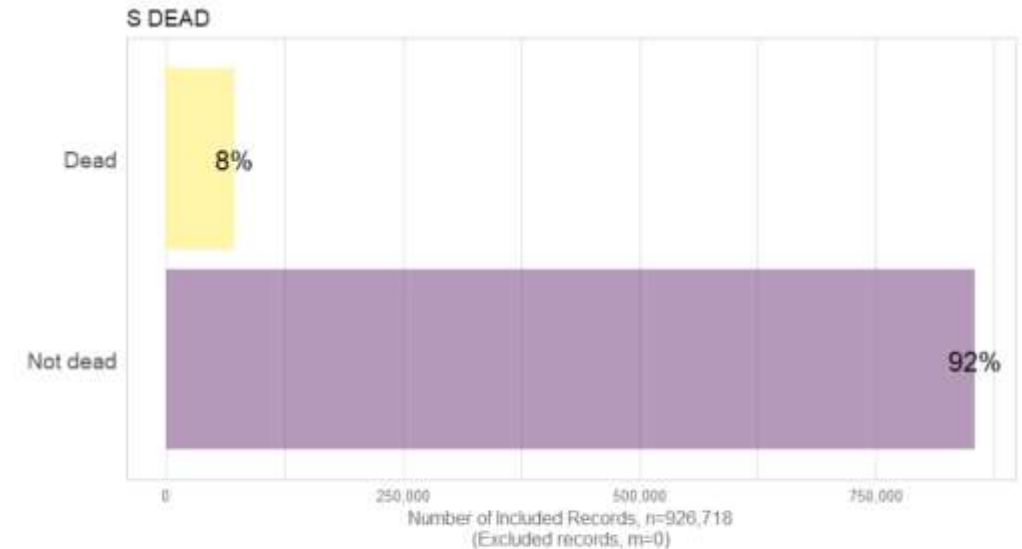
Dead in X years

Diagram illustrating the components of a linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Labels and components:

- Dependent Variable:** Y_i
- Population Y intercept:** β_0
- Population Slope Coefficient:** β_1
- Independent Variable:** X_i
- Random Error term:** ϵ_i
- Linear component:** $\beta_0 + \beta_1 X_i$
- Random Error component:** ϵ_i



`$S_DEAD $pEADlevels 1 2 "Dead" "Not dead"`

`$pEADlabel [1] "Dead in X years?"`

`$pEADdescription [1] "Mortality status: Refers to whether or not the respondent died during the X years following the survey response"`

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

A. Graphing Technique

0.1 Modeling form

$dv \sim -1 + \text{PR} + \text{age_group} + \text{female} + \text{marital} + \text{educ3} + \text{poor_health} + \text{FOL}$

Province of residence

Dependent Variable

Population Y intercept

Population Slope Coefficient

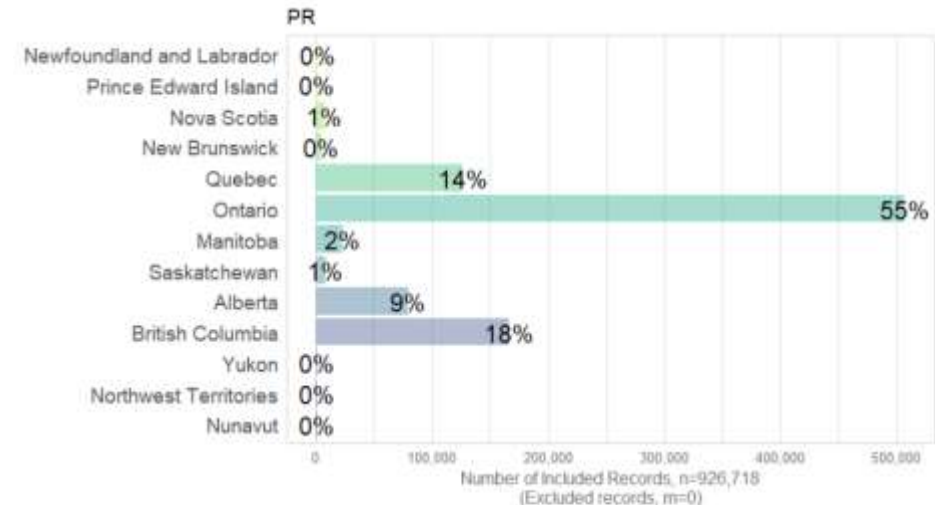
Independent Variable

Random Error term

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Linear component

Random Error component



\$PR\$levels 10 11 12 "Newfoundland and Labrador" "Prince Edward Island" "Nova Scotia" 13 24 35 "New Brunswick"
"Quebec" "Ontario" 46 47 48 "Manitoba" "Saskatchewan" "Alberta" 59 60 61 "British Columbia" "Yukon" "Northwest
Territories" 62 "Nunavut"
\$PRlabel\$ [1] "Province of residence"
\$PRdescription\$ [1] "Province or territory of residence"

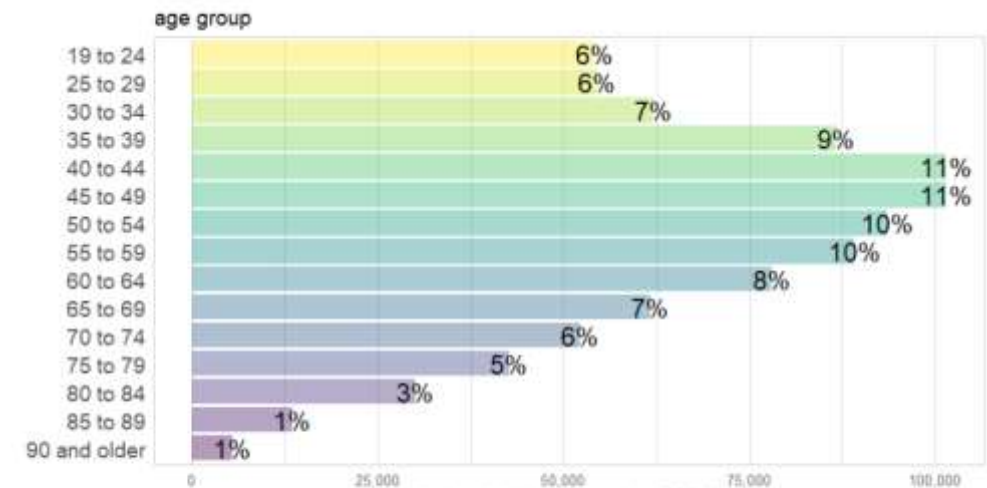
$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

A. Graphing Technique

0.1 Modeling form

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

5-year age category



\$age_group age_grouplevels 1 2 3 4 5 6 "19 to 24" "25 to 29" "30 to 34" "35 to 39" "40 to 44" "45 to 49" 7 8 9 10 11 12 "50 to 54" "55 to 59" "60 to 64" "65 to 69" "70 to 74" "75 to 79" 13 14 15 "80 to 84" "85 to 89" "90 and older"
age_grouplabel [1] "Age"
age_groupdescription [1] "Age: grouped"

Dependent Variable → $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Population Y intercept → β_0

Population Slope Coefficient → β_1

Independent Variable → X_i

Random Error term → ϵ_i

Linear component: $\beta_0 + \beta_1 X_i$

Random Error component: ϵ_i

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

A. Graphing Technique

0.1 Modeling form

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

Sex

Dependent Variable → $Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$

Population Y intercept → β_0

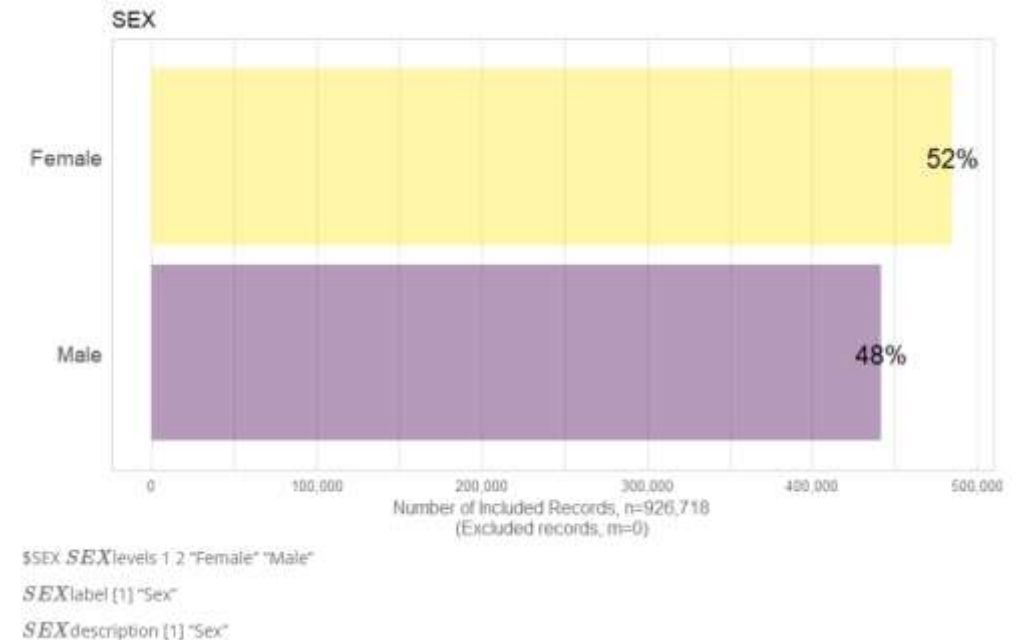
Population Slope Coefficient → β_1

Independent Variable → X_i

Random Error term → ϵ_i

Linear component: $\beta_0 + \beta_1 X_i$

Random Error component: ϵ_i



$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

A. Graphing Technique

0.1 Modeling form

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

Marital Status

```
# because 'still legally 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"))
```

Dependent Variable →

Population Y intercept →

Population Slope Coefficient →

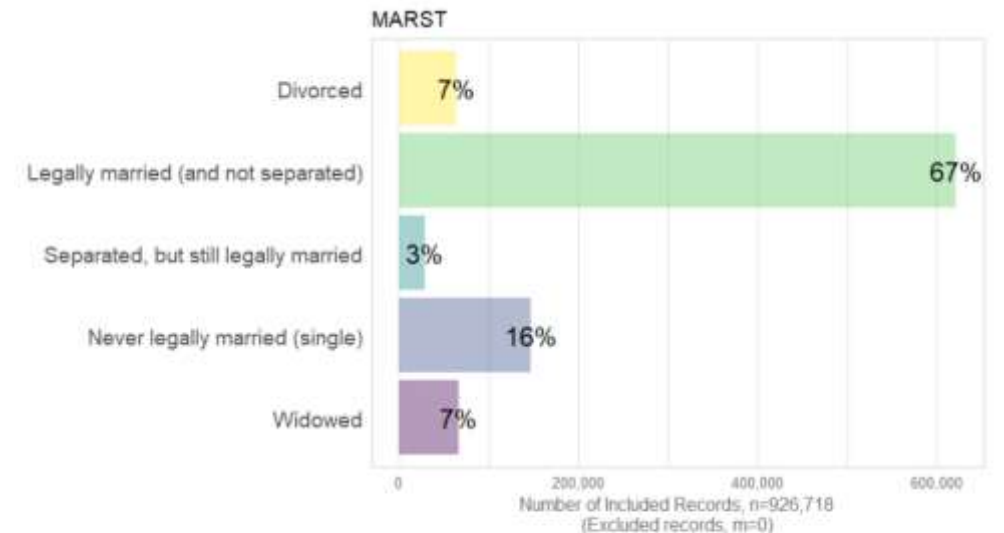
Independent Variable →

Random Error term →

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Linear component

Random Error component



\$MARST MARSTlevels 1 2 "Divorced" "Legally married (and not separated)" 3 4 "Separated, but still legally married" "Never legally married (single)" 5 "Widowed"

MARSTlabel [1] "Marital status"

MARSTdescription [1] "Marital Status; Refers to the legal marital status of the person;"

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

A. Graphing Technique

0.1 Modeling form

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

Highest Degree

```
# because more than 3 categories is too fragmented.
educ5 = car::recode(
  HCD0,
  none = "less than high school"
  "high school graduation certificate or equivalency certificate" = "high school"
  "Other trades certificate or diploma" = "high school"
  "Registered apprenticeship certificate" = "high school"
  "College, CEGEP or other non-university certificate or diploma from a program of 3 months to less than 1 year" = "college"
  "College, CEGEP or other non-university certificate or diploma from a program of 1 year to 2 years" = "college"
  "College, CEGEP or other non-university certificate or diploma from a program of more than 2 years" = "college"
  "University certificate or diploma below bachelor level" = "college"
  "Bachelors degree" = "graduate"
  "University certificate or diploma above bachelor level" = "graduate"
  "Degree in medicine, dentistry, veterinary medicine or optometry" = "graduate"
  "Masters degree" = "graduate"
  "Earned doctorate degree" = "Dr."
)
educ5 = factor(educ5, levels = c(
  "less than high school",
  "high school",
  "college",
  "graduate",
  "Dr."
))
```

```
ds1 %>% group_by(educ5) %>% summarize(n = n())
```

A tibble: 5 x 2

educ5	n
<fct>	<int>
1 less than high school	902326
2 high school	1587347
3 college	1555485
4 graduate	269945
5 Dr.	31546

Dependent Variable → Y_i = β_0 + $\beta_1 X_i$ + ϵ_i

Population Y intercept → β_0

Population Slope Coefficient → β_1

Independent Variable → X_i

Random Error term → ϵ_i

Linear component

Random Error component

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

A. Graphing Technique

0.1 Modeling form

$dv \sim -1 + PR + age_group + female + marital + \boxed{educ3} + poor_health + FOL$

Highest Degree

```
# because even only 3 may be too granular for our purposes
educ3 = car::recode(
  hcbd,
  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, CSEP or other non-university certificate or diploma from a program of 3 months to less than 1 year" = "more than high school"
  "College, CSEP or other non-university certificate or diploma from a program of 1 year to 2 years" = "more than high school"
  "College, CSEP 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
```

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

Random Error term

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Linear component

Random Error component

$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

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"))
```

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

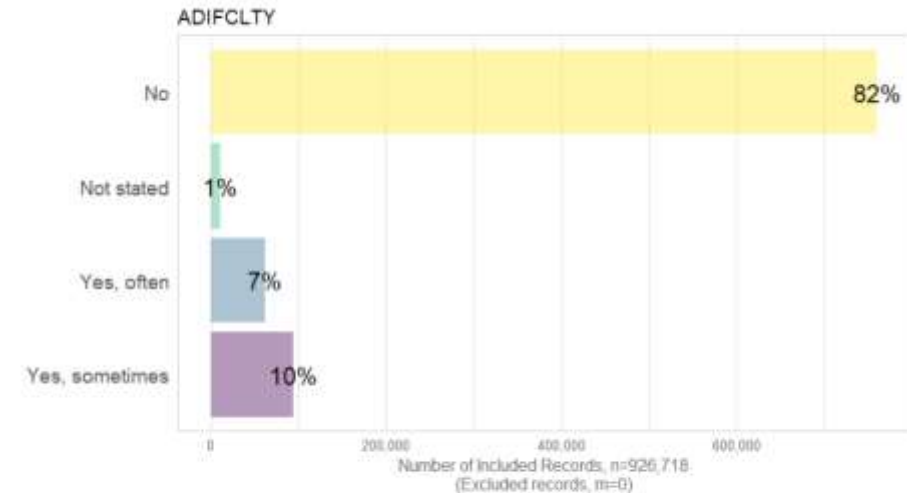
Random Error term

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Linear component

Random Error component

ADIFCLTY



\$ADIFCLTY ADIFCLTYlevels 1 2 3 4 "No" "Not stated" "Yes, often" "Yes, sometimes"

ADIFCLTYlabel [1] "Problems with ADL"

ADIFCLTYdescription [1] "Difficulties with activities of daily living: Difficulty with activities of daily living such as hearing, seeing, communicating, walking, climbing stairs, bending, learning or doing any similar activities."

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

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"))
```

Dependent Variable

Population Y intercept

Population Slope Coefficient

Independent Variable

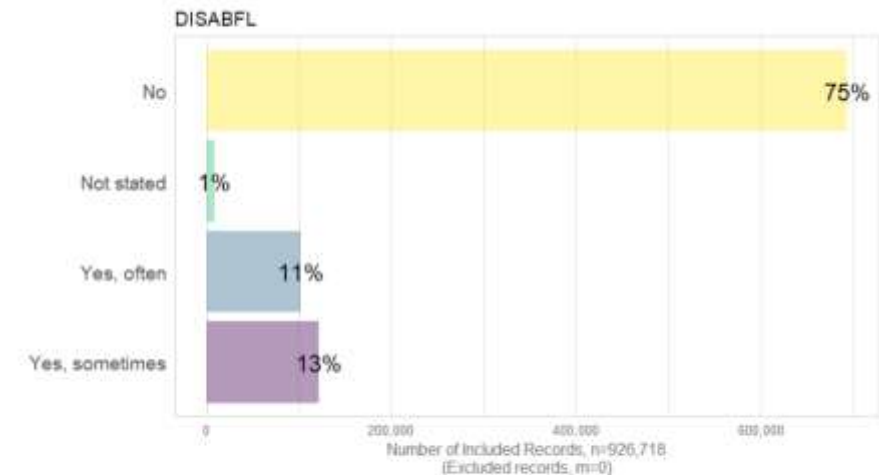
Random Error term

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Linear component

Random Error component

DISABFL



\$DISABFL DISABFLlevels 1 2 3 4 "No" "Not stated" "Yes, often" "Yes, sometimes"

DISABFLlabel [1] "Problems with ADL"

DISABFLdescription [1] "Difficulties with activities of daily living: Refers to difficulty with daily activities and/or a physical condition or mental condition or health problem that reduces the amount or kind of activity that a person can do at home, at work or school or in other activities (e.g., transportation, leisure)."

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

A. Graphing Technique

0.1 Modeling form

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

First Official Language

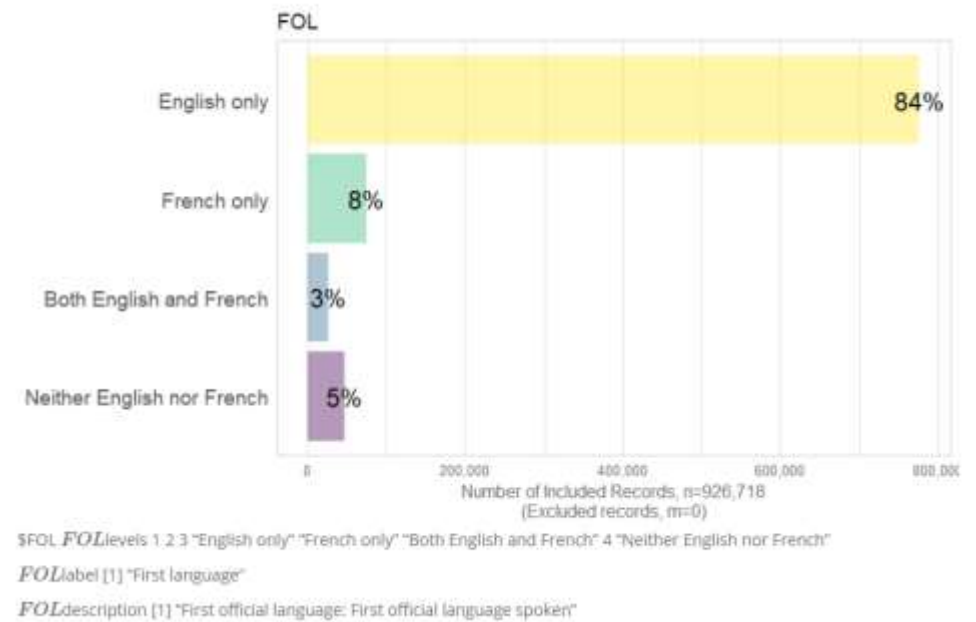
Diagram illustrating the components of a linear regression model:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Labels and components:

- Dependent Variable:** Y_i
- Population Y intercept:** β_0
- Population Slope Coefficient:** β_1
- Independent Variable:** X_i
- Random Error term:** ϵ_i
- Linear component:** $\beta_0 + \beta_1 X_i$
- Random Error component:** ϵ_i

FOL



$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

A. Graphing Technique

0.2 Graphical form

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

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

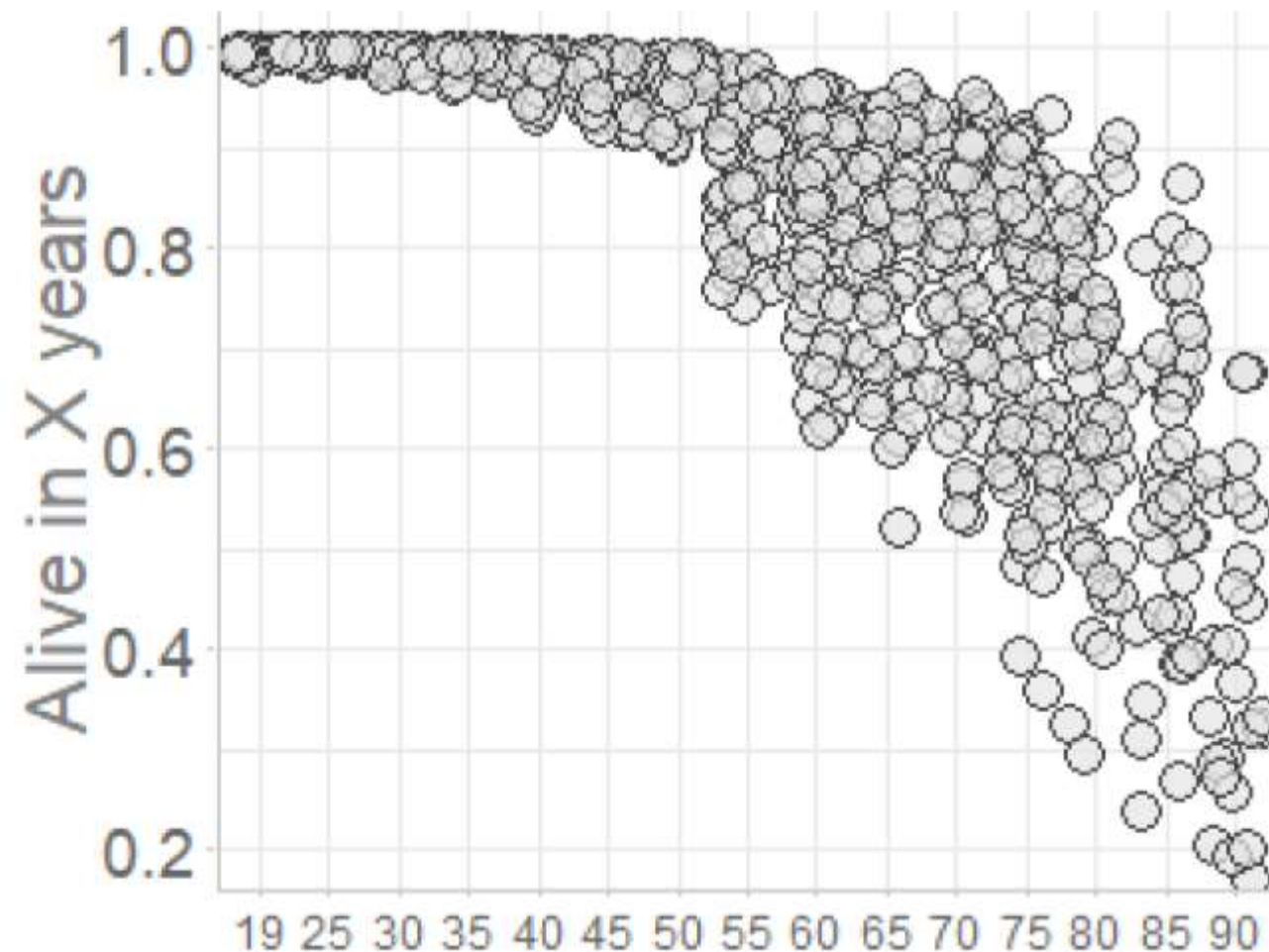
Visualizing probability instead of log-odds because it is more intuitive

Diagram illustrating the components of the linear regression equation:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i$$

Labels and components:

- Dependent Variable: Y_i
- Population Y intercept: β_0
- Population Slope Coefficient: β_1
- Independent Variable: X_i
- Random Error term: ϵ_i
- Linear component: $\beta_0 + \beta_1 X_i$
- Random Error component: ϵ_i

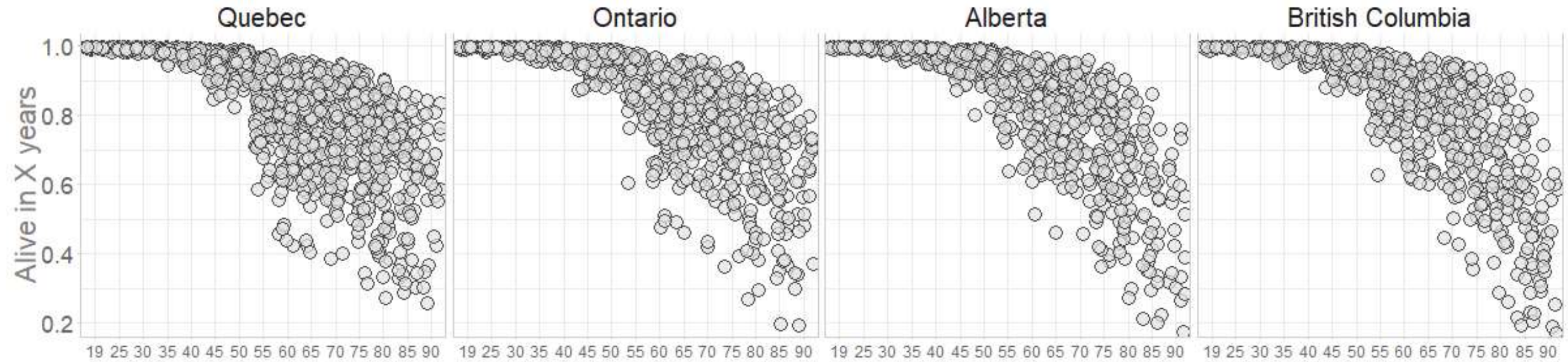


$$\ln\left(\frac{\hat{p}}{(1-\hat{p})}\right) = b_0 + b_1 X_1 + b_2 X_2 + \dots + b_p X_p$$

A. Graphing Technique

0.2 Graphical form

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



LEGEND

Facet = Province of residence

A. Graphing Technique

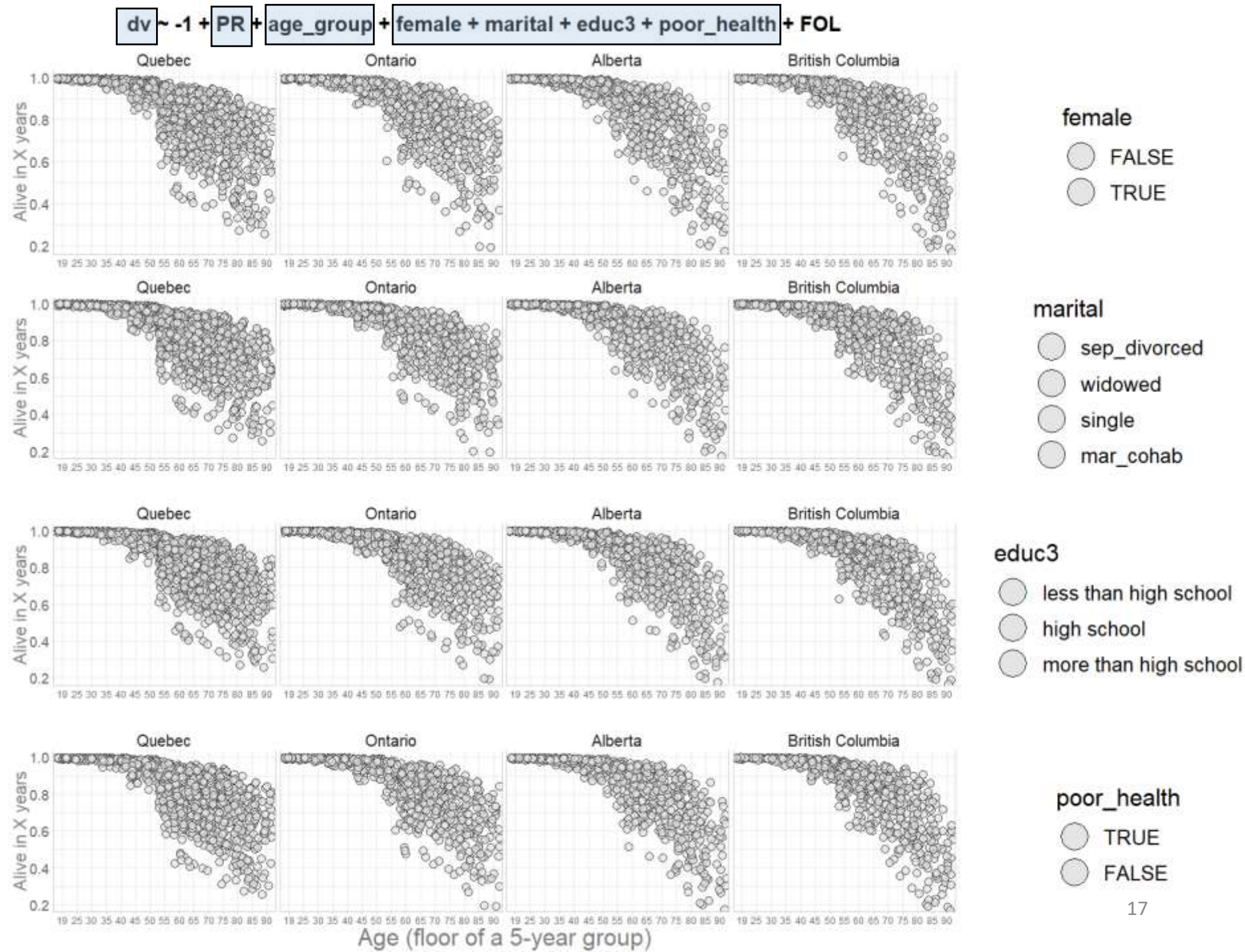
0.2 Graphical form

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

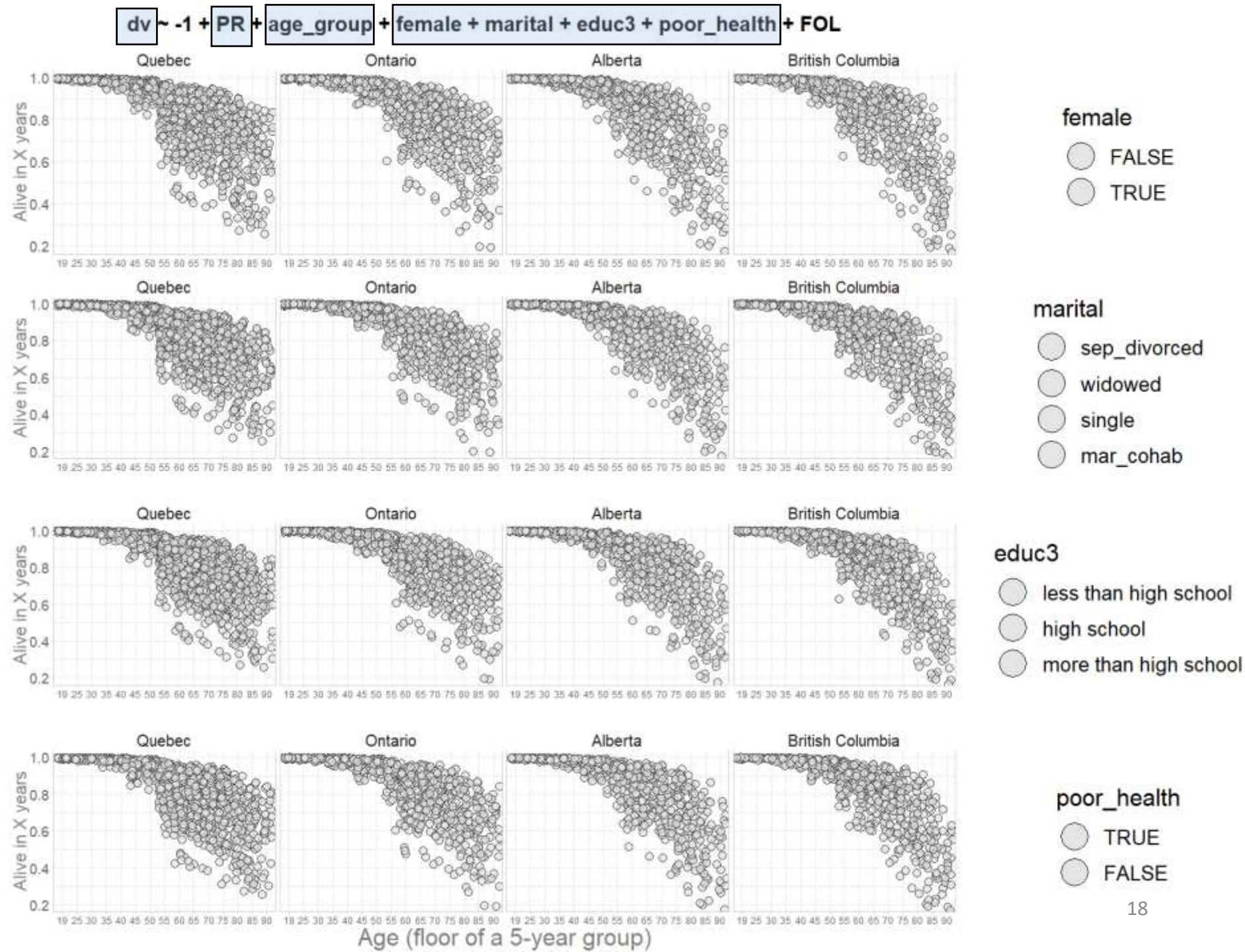
What should the “reference group” be for each predictor?

What do we expect based on existing research?

Informed expectation

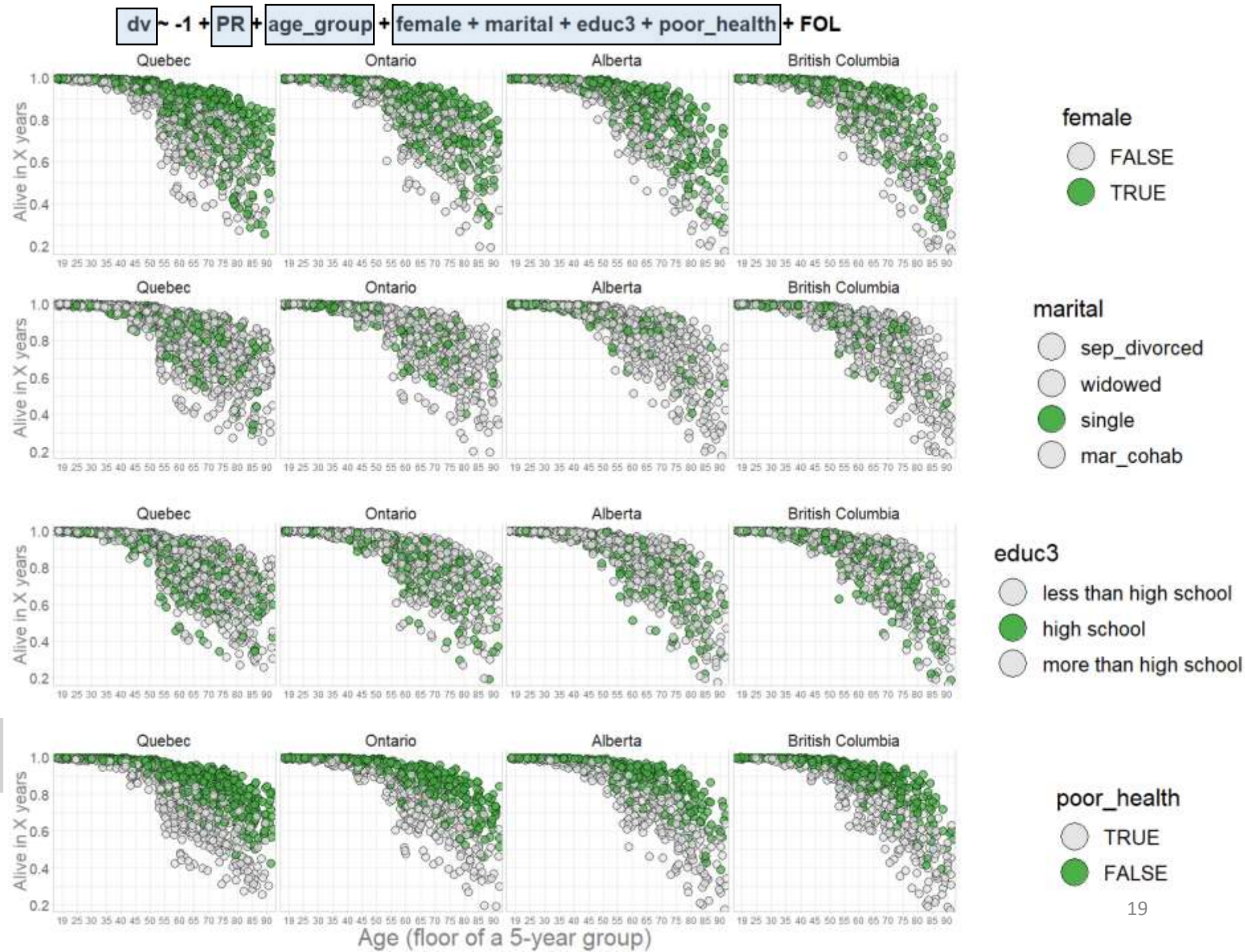
Reference group

?



A. Graphing Technique

0.3 Coloring book



A. Graphing Technique

0.3 Coloring book

QUESTION

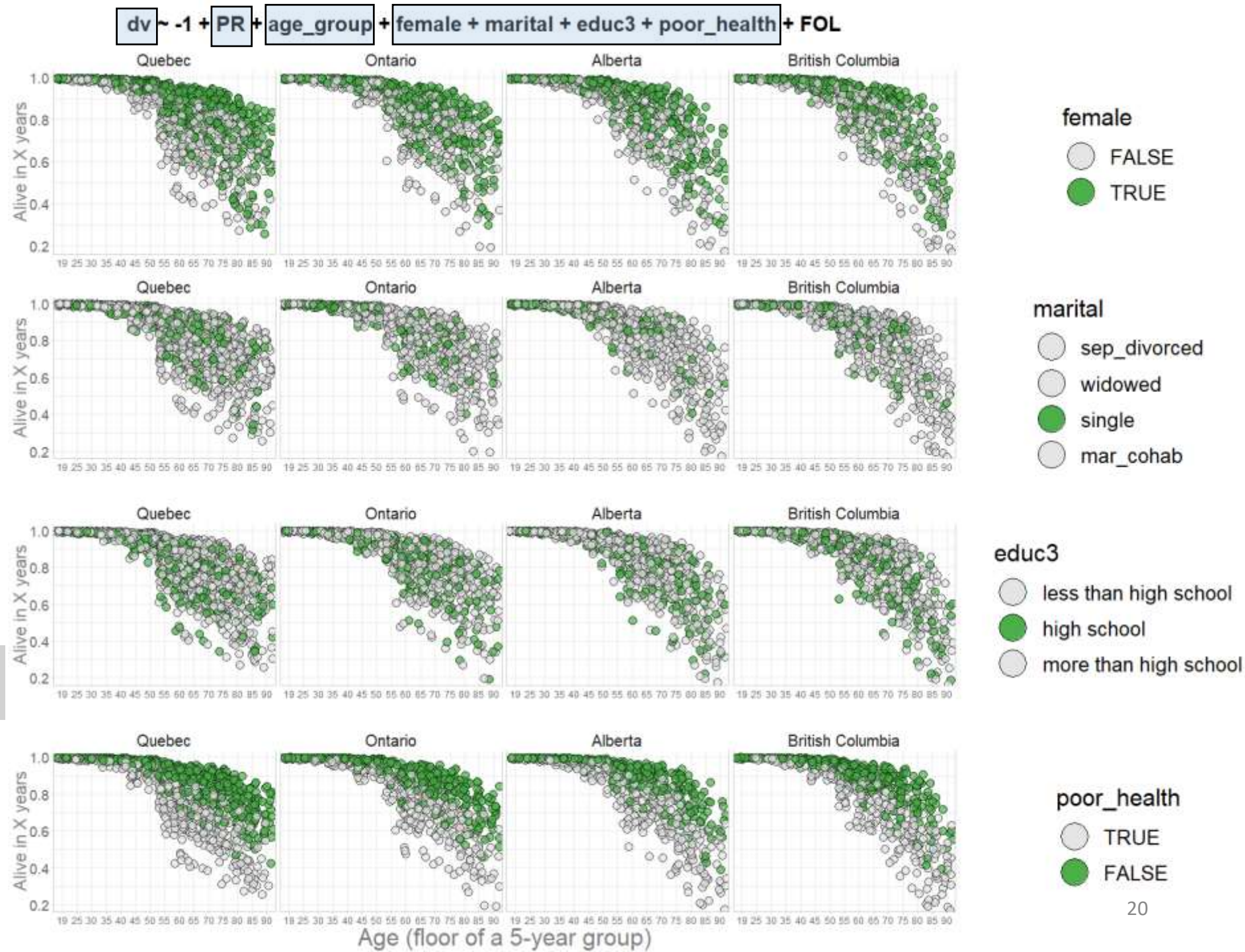
Compared to reference group, what levels of predictors are expected to **increase** the mortality risk?

Informed expectation

Moderately increased risk

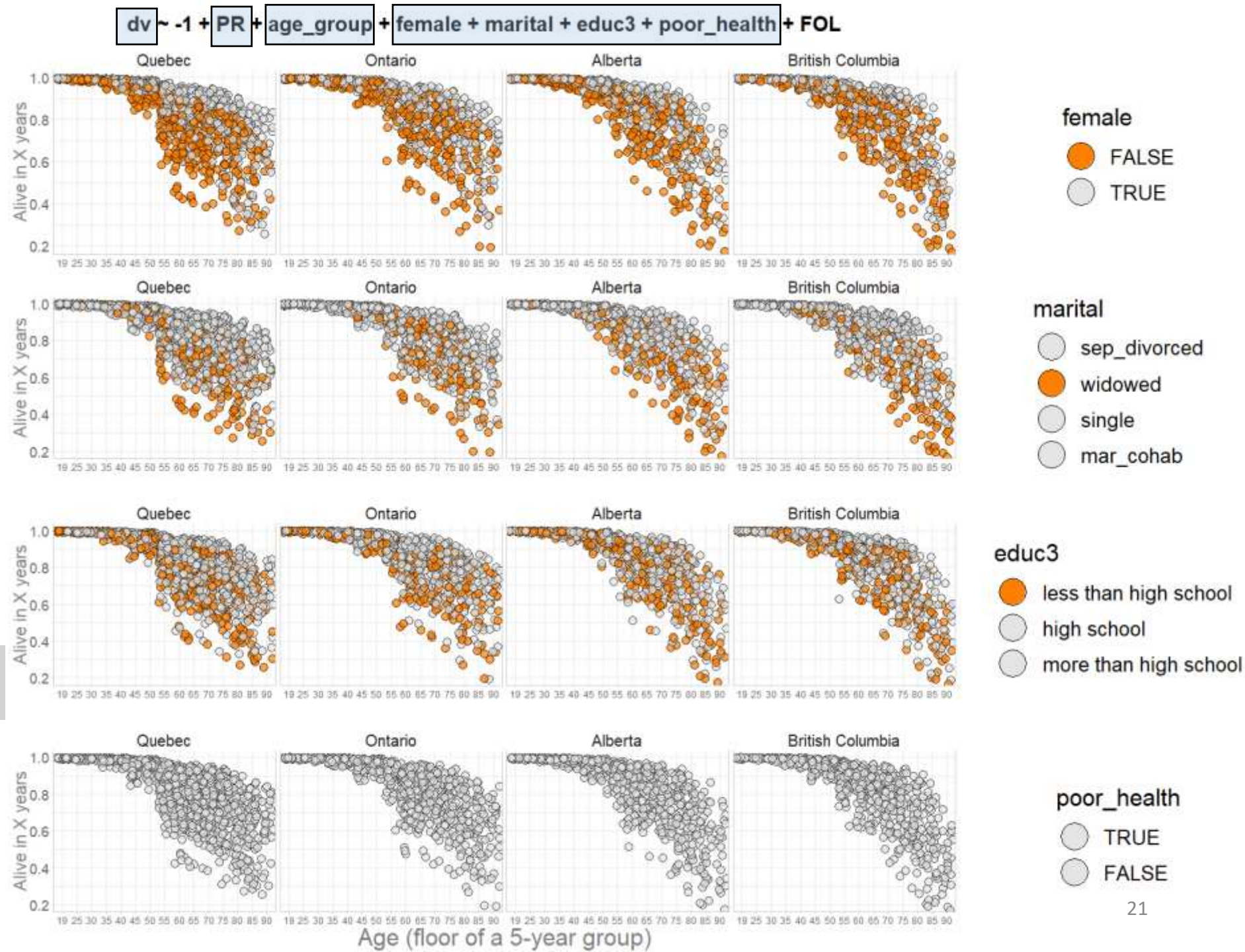
?

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?

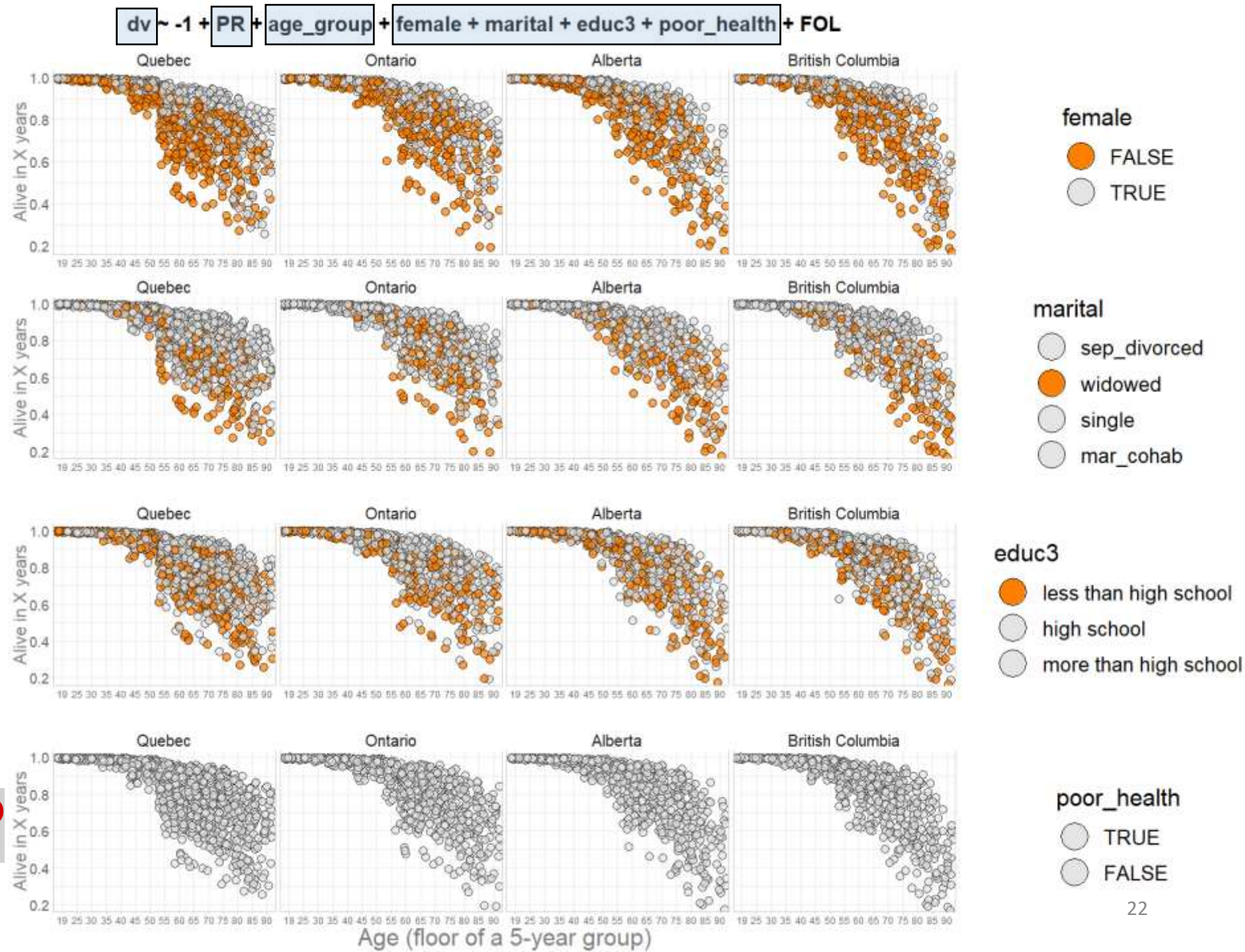
Informed expectation

Moderately increased risk

Reference group

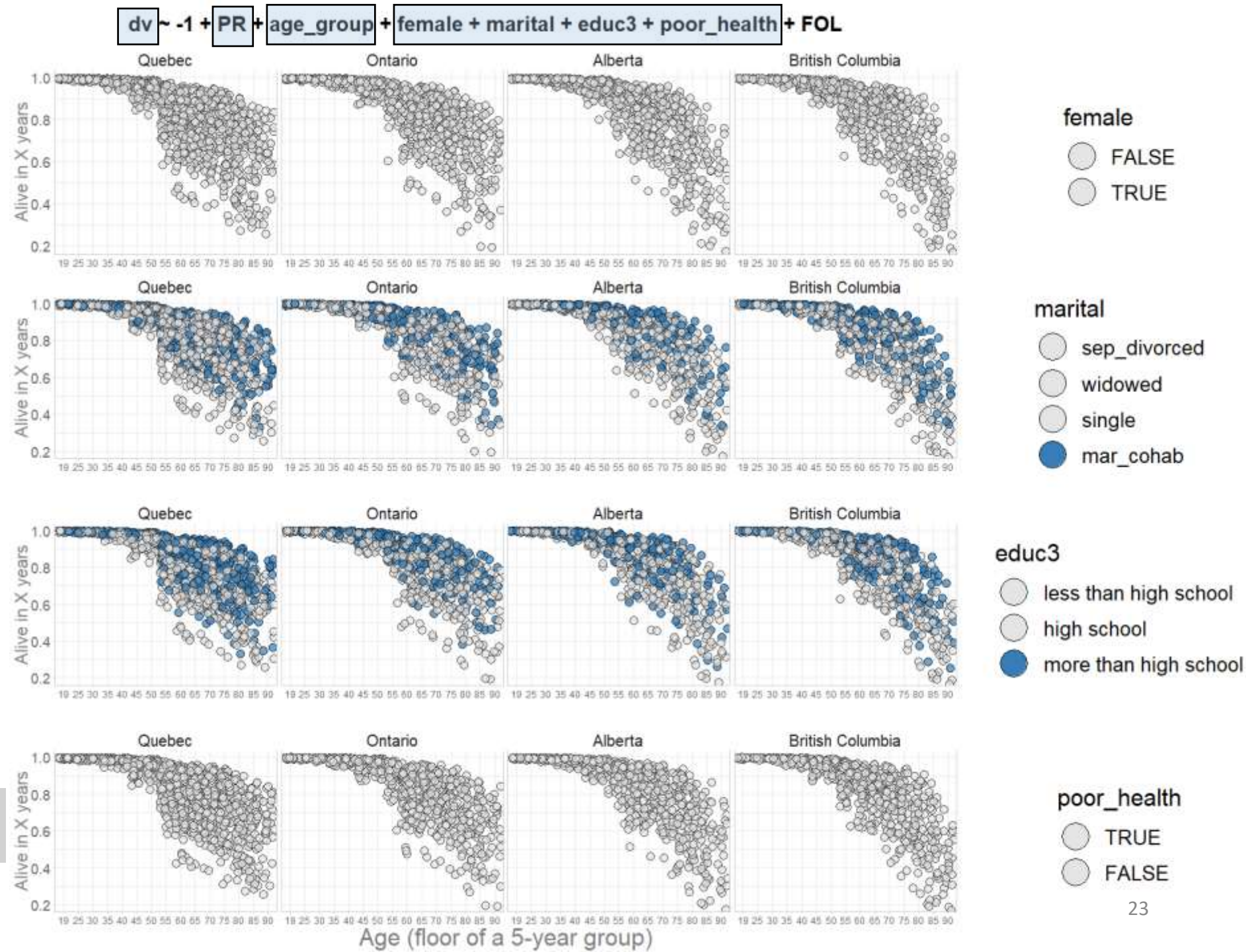
Moderately decreased risk

?



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

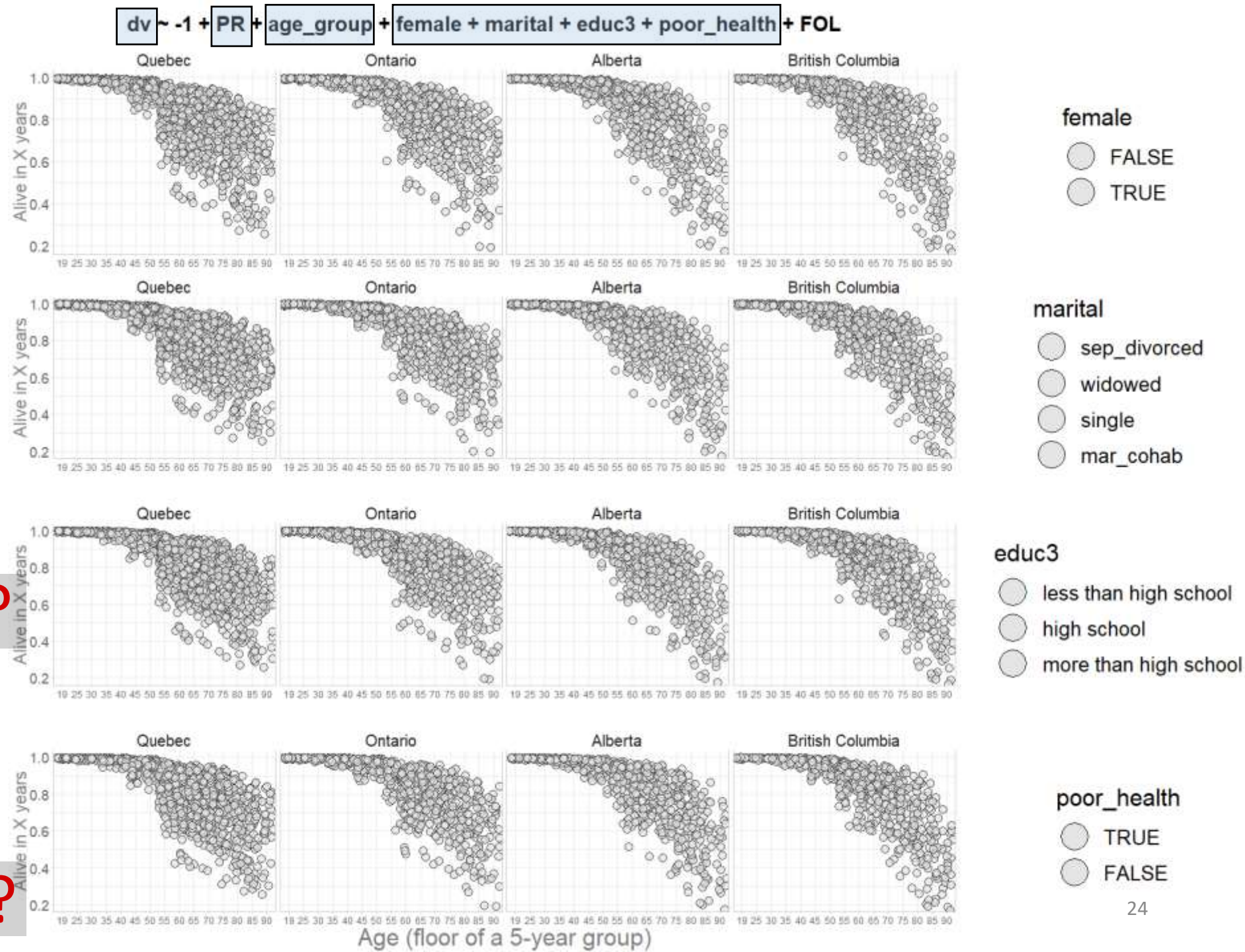
Substantially increased risk ?

Moderately increased risk

Reference group

Moderately decreased risk

Substantially decreased risk ?



A. Graphing Technique

0.3 Coloring book

QUESTION

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

Informed expectation

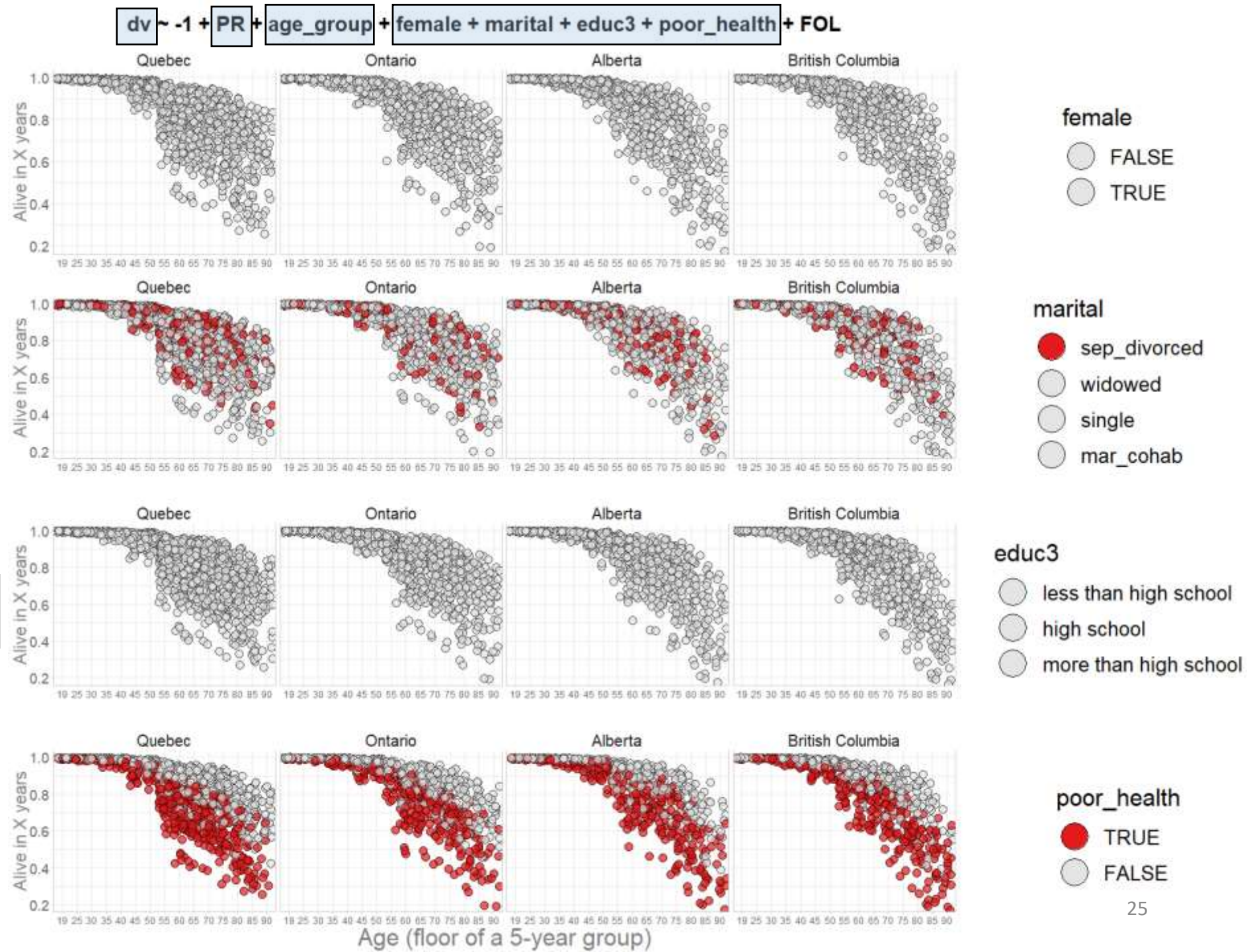
Substantially increased risk

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

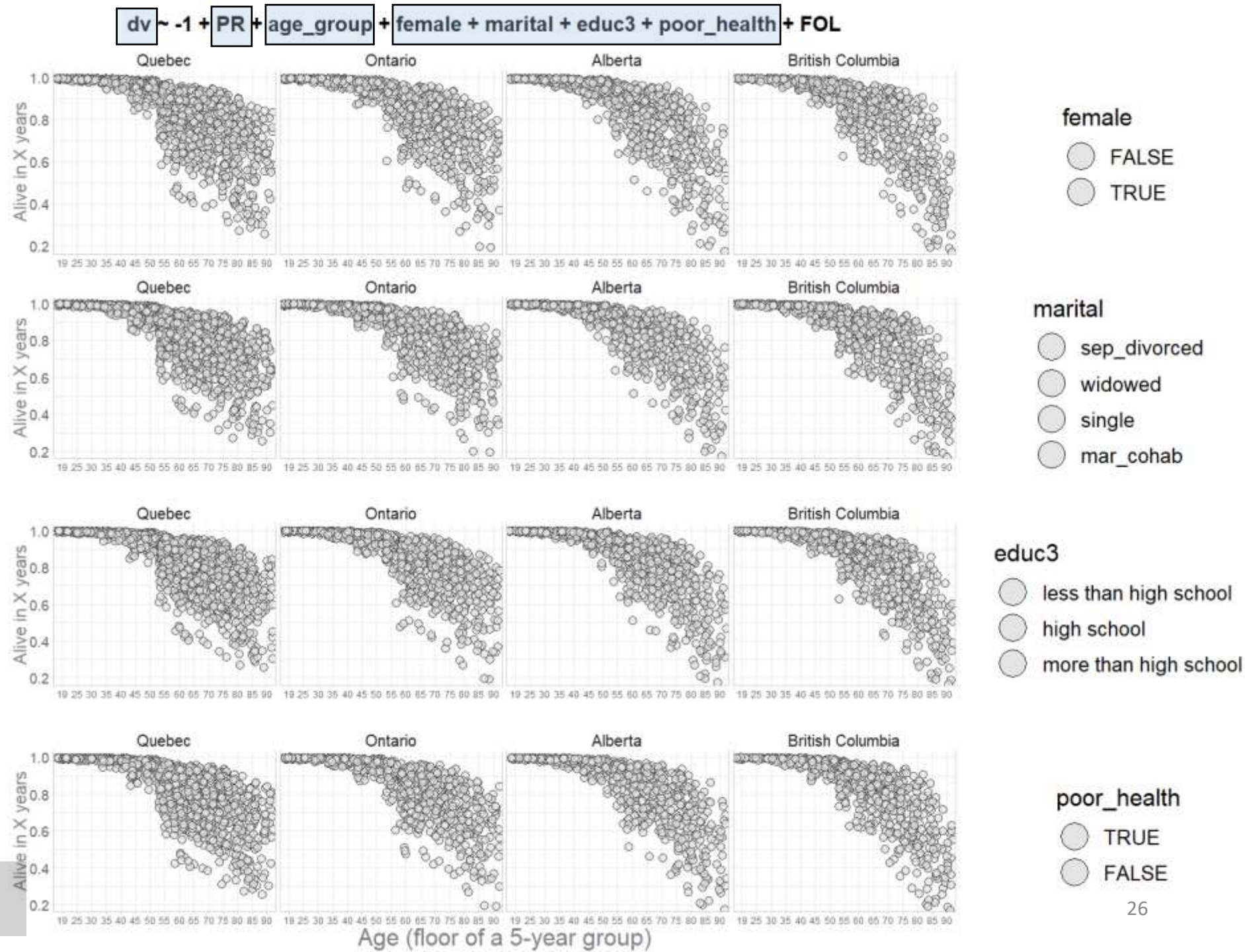
Substantially increased risk

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

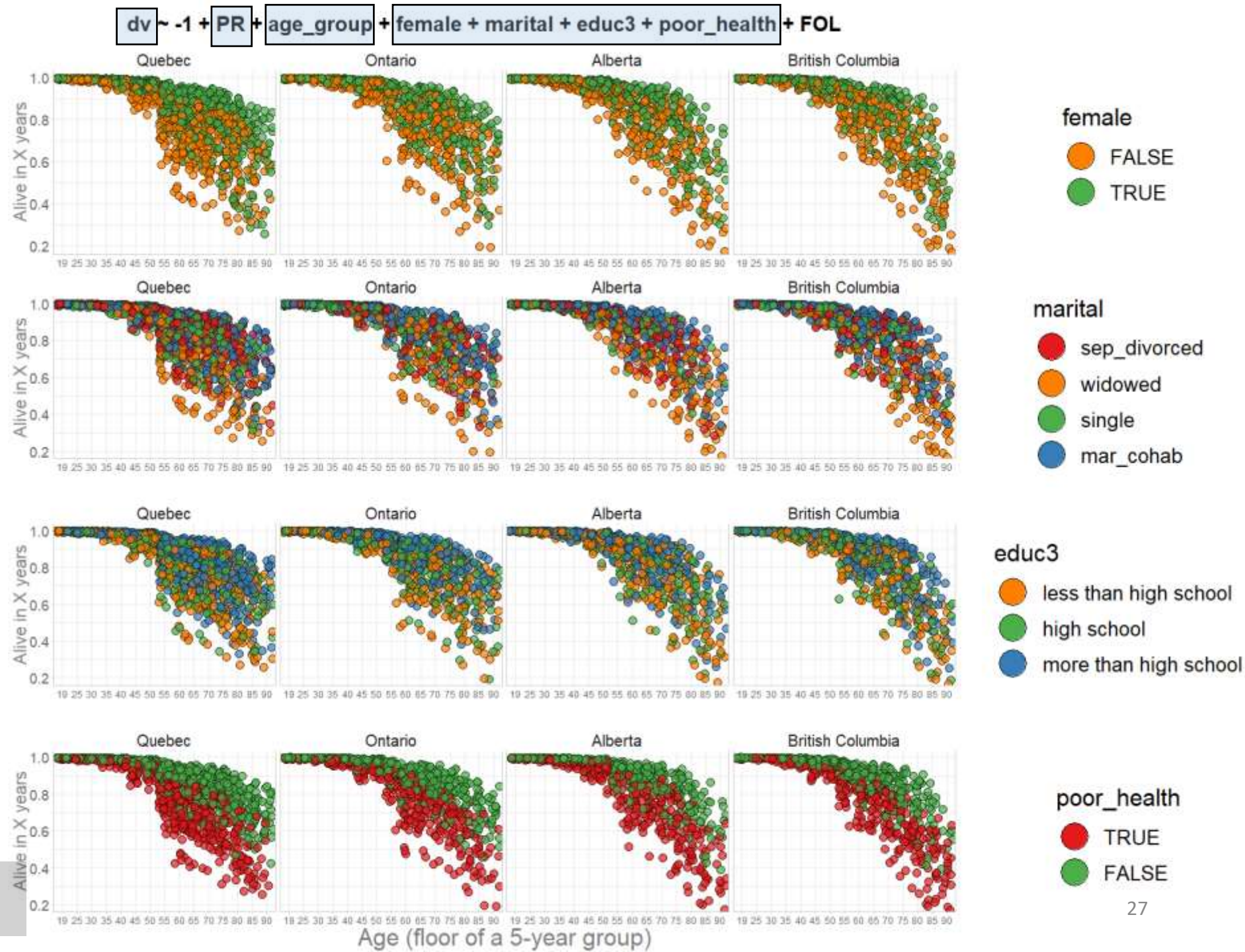
Substantially increased risk

Moderately increased risk

Reference group

Moderately decreased risk

Substantially decreased risk



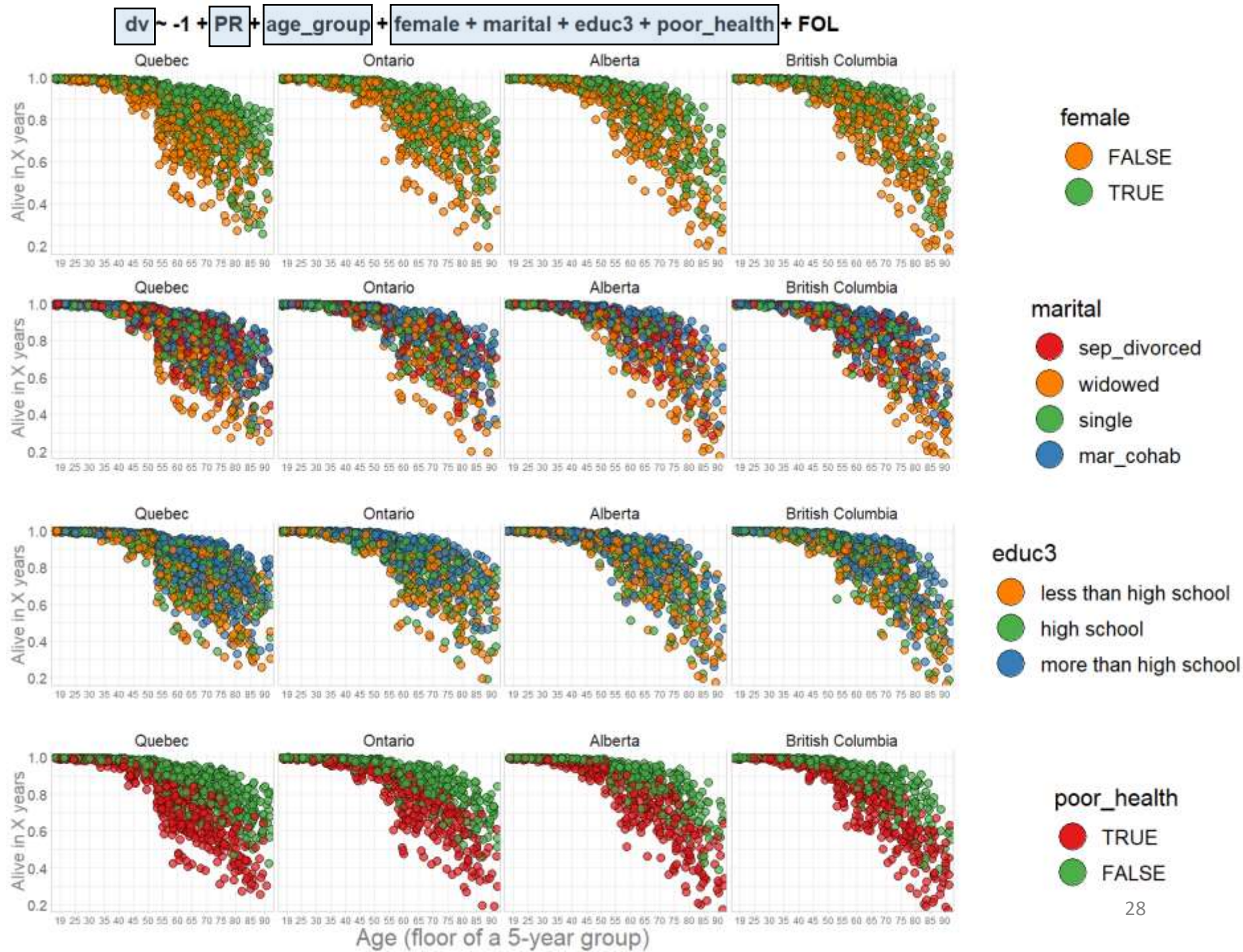
A. Graphing Technique

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

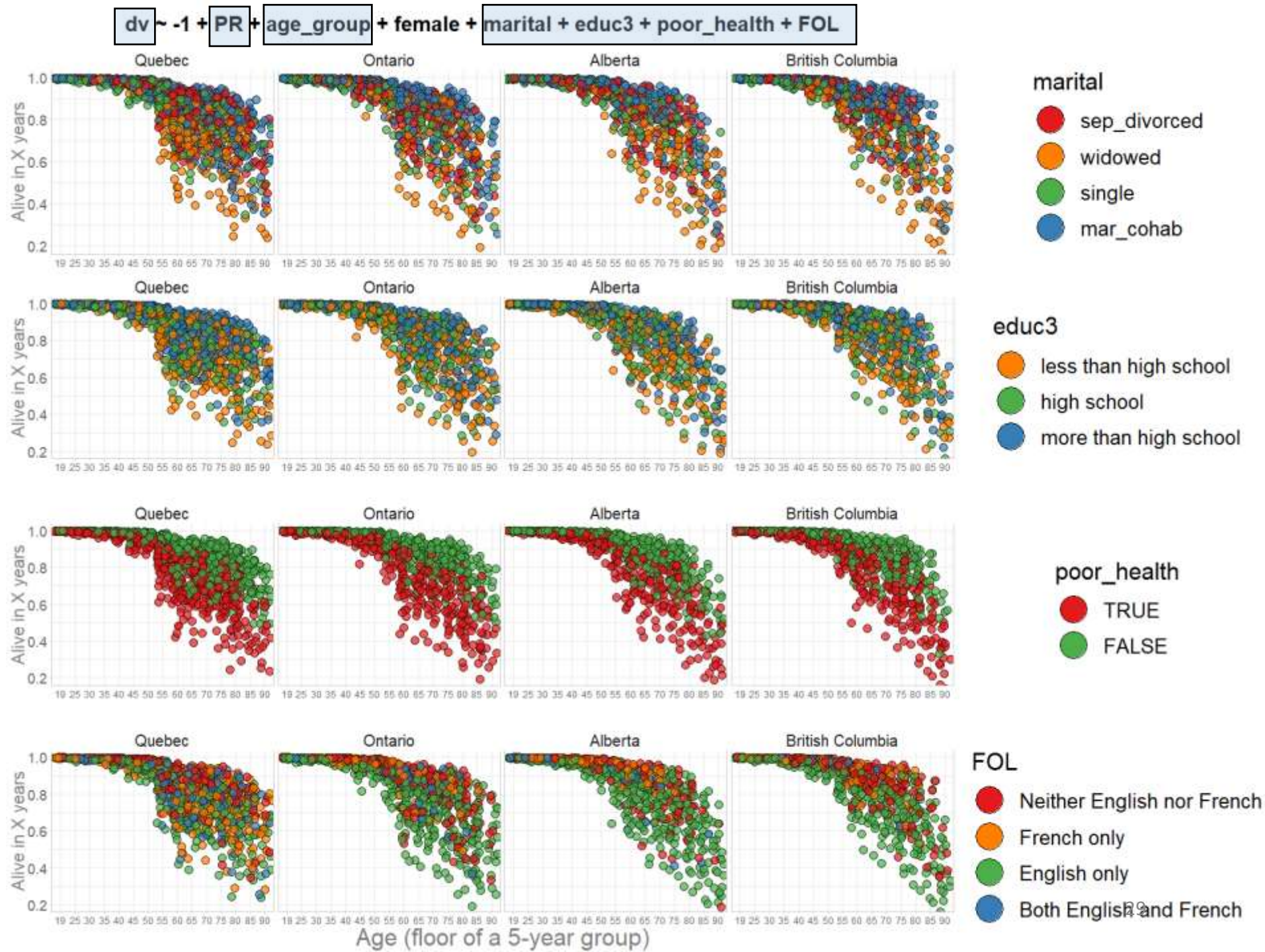
A. Graphing Technique

0.3 Coloring book

NOTICE

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

We can adjust what is being displayed



So how would you organize this production?

I cannot describe the workflow in the remaining time

But I can help you learn through reproduction

Here are some principles to keep in mind as you study the project

B. Workflow Highlights

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

How to reproduce

- 0. Clone this repository (either via git or from the browser)
- i. Launch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run `./reports/graphing-phase-only/graphing-phase-only.R` to load the model solution and start producing graphs

Background

- Information for Participants
- Data Codebook

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("../data-unshared/derived/0-metador.rds")
ds0      <- readRDS("../data-unshared/derived/1-greeted.rds")
```

Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats `eda1` but for subsample of first-generation immigrants

Result of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yielded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this [slide deck](#) and presented during the closing plenary of IDPDL-2018 Conference in Banff.

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.

Expect to read scripts

Main README should provide a map

<https://github.com/andkov/ipdln-2018-hackathon/README.md>

B. Workflow Highlights

1.1 [RAnalysisSkeleton](#) by Will Beasley: basic starting point for reproducible projects

Keep recognizable structure over projects

The screenshot shows the GitHub repository 'wibeasley / RAnalysisSkeleton'. The repository has 185 commits, 1 branch, 0 releases, 3 contributors, and is licensed under GPL-2.0. The repository description is 'Files and settings commonly used in analysis projects with R'. The repository contains a directory structure with folders like 'analysis', 'data-public', 'data-unshared', 'documentation', 'manipulation', 'reports', 'stitched-output', 'utility', and files like '.gitattributes', '.gitignore', 'LICENSE', 'NEWS', 'RAnalysisSkeleton.Rproj', 'README.md', and 'config.yml'. The README.md file is open, showing the title 'R Analysis Skeleton' and a description: 'This project contains the files and settings commonly used in analysis projects with R. A developer can start an analysis repository more quickly by copying these files.'

The screenshot shows the GitHub repository 'andkov / ipdln-2018-hackathon'. The repository has 115 commits, 1 branch, 0 releases, 1 contributor, and is licensed under GPL-2.0. The repository description is 'Repository to accompany a hackathon at IPDLN conference at Banff, Sep 2018'. The repository contains a directory structure with folders like 'data-public', 'data-unshared', 'libs', 'manipulation', 'reports', 'sandbox', 'scripts', 'utility', and files like '.gitignore', 'LICENSE', 'NEWS', 'README.md', and 'ipdln-2018-hackathon.Rproj'. The README.md file is open, showing the title 'ipdln-2018-hackathon' and a description: 'Demonstrating coloring-book technique of graph production in ggplot2 during data linkage hackathong at IPDLN-2018 conference at Banff, Sep 2018.'

Notice structural similarities

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

How to reproduce

- 0. Clone this repository (either via git or from the browser)
- i. Launch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run `[./reports/graphing-phase-only/graphing-phase-only.R]` to load the model solution and start producing graphs

Background

- [Information for Participants](#)
- [Data Codebook](#)

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0      <- readRDS("./data-unshared/derived/1-greeter.rds")
```

Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats `eda1` but for subsample of first-generation immigrants

Result of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yielded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this [slide deck](#) and presented during the closing plenary of IDPDL-2018 Conference in Banff.

Branch: master ▾ [ipdlN-2018-hackathon](#) / README.md

 andkov Update README.md

Try to keep tasks separate:

- Data cleaning
- Statistical modeling
- Graph production

Tasks are narratives to be told

Here are some examples

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

How to reproduce

- 0. Clone this repository (either via git or from the browser)
- i. Launch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run `[./reports/graphing-phase-only/graphing-phase-only.R]` to load the model solution and start producing graphs

Background

- [Information for Participants](#)
- [Data Codebook](#)

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0 <- readRDS("./data-unshared/derived/1-greeter.rds")
```

Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats `eda1` but for subsample of first-generation immigrants

Result of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yielded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this [slide deck](#) and presented during the closing plenary of IDPDL-2018 Conference in Banff.

Screenshots of linked dynamic document

```
# declare where you will store the product of this script
path_save <- "./data-unshared/derived/ls_guide.rds"
```

```
POBDER <- list(
  "levels" = c(
    "1" = "Born in province of residence"
    , "2" = "Born in another province"
    , "3" = "Born outside Canada "
  )
  , "label" = "Place of birth"
  , "description" = "Place of birth: Indicates whether the respondent was born in the same province that they live in"
)
PR <- list(
  "levels" = c(
    "10" = "Newfoundland and Labrador"
    , "11" = "Prince Edward Island"
    , "12" = "Nova Scotia"
    , "13" = "New Brunswick"
    , "14" = "Quebec"
    , "15" = "Ontario"
    , "16" = "Manitoba"
    , "17" = "Saskatchewan"
    , "18" = "Alberta"
    , "19" = "British Columbia"
    , "20" = "Yukon"
    , "21" = "Northwest Territories"
    , "22" = "Nunavut"
  )
  , "label" = "Province of residence"
  , "description" = "Province or territory of residence"
)
```

```
# create vector with names
block_names <- c("demographic", "identity", "economic", "immigration", "health")
item_names <- c("demographic", "identity", "economic", "immigration", "health")
# create a list object to hold all available metadata
ls_guide <- list()
ls_guide[["block"]] <- mget(block_names, envir = globalenv())
ls_guide[["item"]] <- mget(item_names, envir = globalenv())
```

```
# show components of this list object
ls_guide %>% lapply(names)
```

```
## $block
## [1] "demographic" "identity" "economic" "immigration" "health"
##
## $item
## [1] "SEX" "age_group"
## [3] "MARST" "EPCNT_PP_R"
## [5] "KID_group" "PR"
## [7] "FOL" "OLN"
## [9] "DIVISIN" "ABDERR"
## [11] "ABIDENT" "HCDO"
## [13] "COWD" "NOCBRD"
## [15] "TRMODE" "LOINCA"
## [17] "LOINCB" "d_licoratio_da_bef"
## [19] "RUINDFG" "RPAIR"
## [21] "POBDER" "OPBILN"
## [23] "IMMOER" "AGE_IMM_REVISED_group"
## [25] "YRIM_group" "CITSH"
## [27] "GENSTPOB" "ADIFCLTY"
## [29] "DISABFL" "DISABIL"
## [31] "S_DEAD" "COD1"
## [33] "COD1_CODES" "COD2"
## [35] "COD2_CODES"
```

B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

How to reproduce

- 0. Clone this repository (either via git or from the browser)
- i. Launch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run `[./reports/graphing-phase-only/graphing-phase-only.R]` to load the model solution and start producing graphs

Background

- [Information for Participants](#)
- [Data Codebook](#)

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0 <- readRDS("./data-unshared/derived/1-greeter.rds")
```

Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats `eda1` but for subsample of first-generation immigrants

Result of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yielded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this [slide deck](#) and presented during the closing plenary of IDPDL-2018 Conference in Banff.

Screenshots of linked dynamic document

```
# link to the source of the location mapping
path_input_micro <- "./data-unshared/raw/ipdln_synth_final.csv"
path_input_meta <- "./data-unshared/derived/ls_guide.rds"

# test whether the file exists / the link is good
testit::assert("File does not exist", base::file.exists(path_input_micro))
testit::assert("File does not exist", base::file.exists(path_input_meta))

# declare where you will store the product of this script
path_save <- "./data-unshared/derived/0-greeter.rds"
```

```
ds0 <- readr::read_csv(path_input_micro) %>% as.data.frame()
```

```
# basic inspection
ds0 %>% dplyr::glimpse(50)
```

```
## Observations: 4,346,649
## Variables: 34
## $ ABDERR_synth
## $ ABIDENT_synth
## $ ADIFCLTY_synth
## $ CITSM_synth
## $ COWD_synth
## $ DISABFL_synth
## $ DISABIL_synth
## $ DVISMIN_synth
## $ FOL_synth
## $ FPTIM_synth
## $ GENSTPOB_synth
## $ HCDD_synth
## $ IMMDER_synth
## $ LOINCA_synth
## $ LOINCB_synth
## $ MARST_synth
## $ NOCSBRD_synth
## $ OLN_synth
## $ POBDER_synth
## $ SEX_synth
## $ TRMODE_synth
## $ RPAIR_synth
## $ PR_synth
```

```
cat("Save results to ", path_save)
```

```
## Save results to ./data-unshared/derived/0-greeter.rds
```

```
saveRDS(ds1, path_save)
```

The R session information (including the OS info, R version and all packages used):

```
sessionInfo()
```

```
## R version 3.4.4 (2018-03-15)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows >= 8 x64 (build 9200)
... <int> >>, 40, 44, ...
```


B. Workflow Highlights

Screenshots of linked dynamic document

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

How to reproduce

- 0. Clone this repository (either via git or from the browser)
- i. Launch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run `[./reports/graphing-phase-only/graphing-phase-only.R]` to load the model solution and start producing graphs

Background

- [Information for Participants](#)
- [Data Codebook](#)

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0 <- readRDS("./data-unshared/derived/1-greeter.rds")
```

Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` repeats `eda1` but for subsample of first-generation immigrants

Result of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

This script yielded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this [slide deck](#) and presented during the closing plenary of IPDLN-2018 Conference in Banff.



B. Workflow Highlights

1.2 Autonomous phases: data cleaning, statistical modelling, graph production

How to reproduce

- 0. Clone this repository (either via git or from the browser)
- i. Launch RStudio project via .Rproj file
- ii. Execute `./manipulation/0-metador.R` to generate object with meta data
- iii. Examine `./reports/technique-demonstration/` to see how models were estimated.
- iv. Run `[./reports/graphing-phase-only/graphing-phase-only.R]` to load the model solution and start producing graphs

Background

- [Information for Participants](#)
- [Data Codebook](#)

Dynamic Documentation on Data Cleaning

- `./manipulation/0-metador.R` records the definition of available variables, their factor levels, labels, description, as well as additional meta data (e.g. colors, fonts, themes).
- `./manipulation/1-greeter.R` imports the raw data and perform general tweaks.

The product of these two scripts define the foundation of every subsequent analytic report.

```
ls_guide <- readRDS("./data-unshared/derived/0-metador.rds")
ds0 <- readRDS("./data-unshared/derived/1-greeted.rds")
```

Analytics during Hackathon

- `./reports/eda-1/eda-1` - prints frequency distributions of all variables.
- `./reports/eda-1/eda-1a-first-gen-immigrant` - repeats `eda1` but for subsample of first-generation immigrants






Result of these two EDAs informed development of the script to estimate and to graph models of immigrant mortality:

- `./reports/coloring-book-mortality/coloring-book-mortality.R` - implements analysis in the historic context of the IPDLN-2018-hackathon. Not a report, but a bare R script. Need to know the options before running. More for archeological purposes.

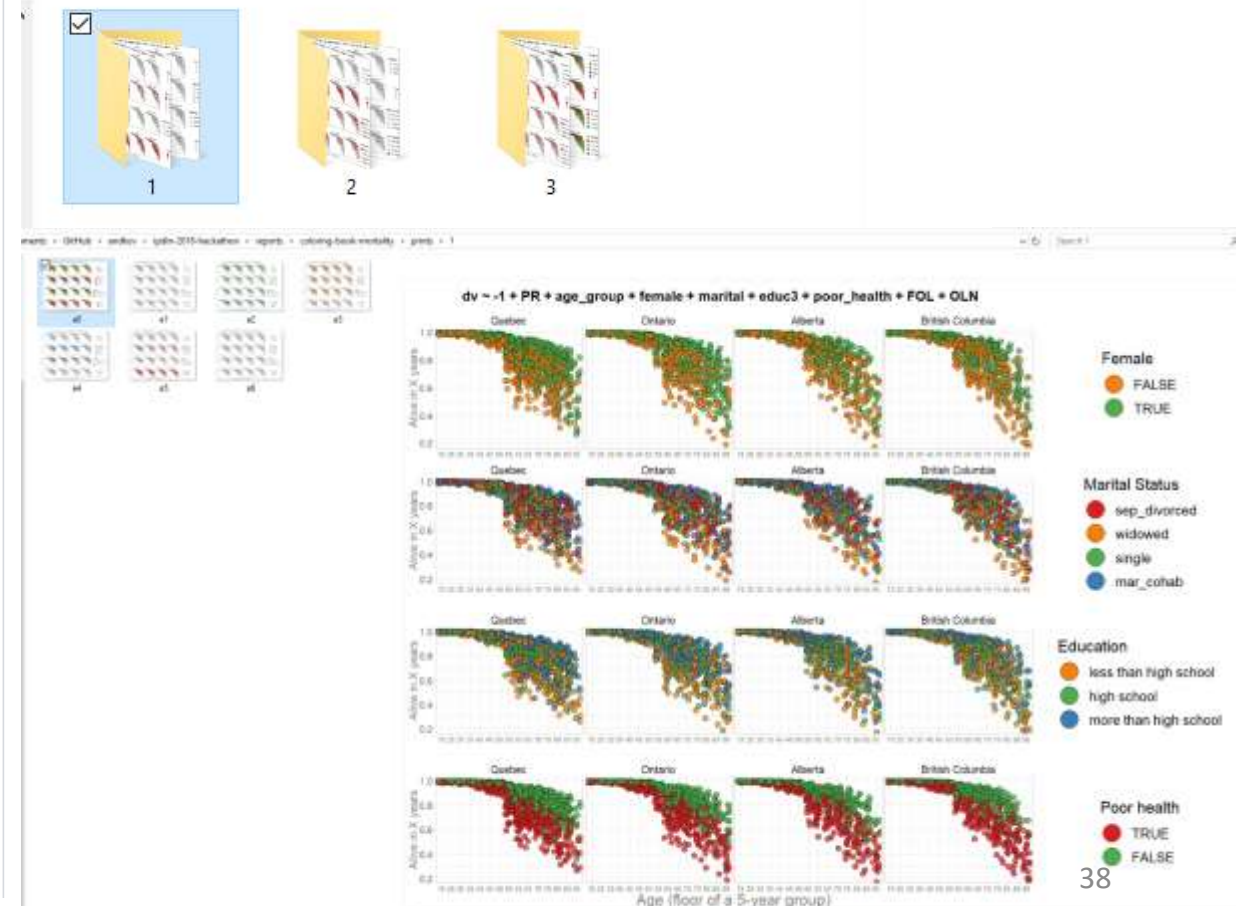
This script yielded a collection of printed graphs stored in `./reports/coloring-book-mortality/prints/`, visualizing three different collection of predictors from the same model. There were put together into this [slide deck](#) and presented during the closing plenary of IDPDL-2018 Conference in Banff.

Screenshots of project repository

ments > GitHub > andkov > ipdln-2018-hackathon > reports > coloring-book-mortality

<input type="checkbox"/> Name	Date
<input checked="" type="checkbox"/> prints	2018-09-13 08:02
 coloring-book-mortality	2018-09-12 15:23
 ipdln-2018-banff-hackathon-results-2018-09-14	2018-09-14 07:17
 results-part-1	2018-09-13 23:41
 results-part-2	2018-09-13 23:41
 results-presentation-script.md	2018-09-14 07:30

uments > GitHub > andkov > ipdln-2018-hackathon > reports > coloring-book-mortality > prints

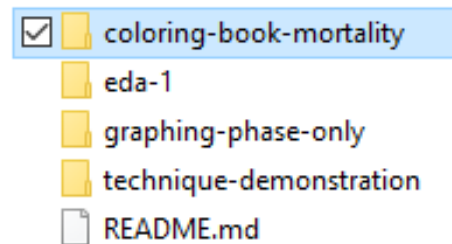
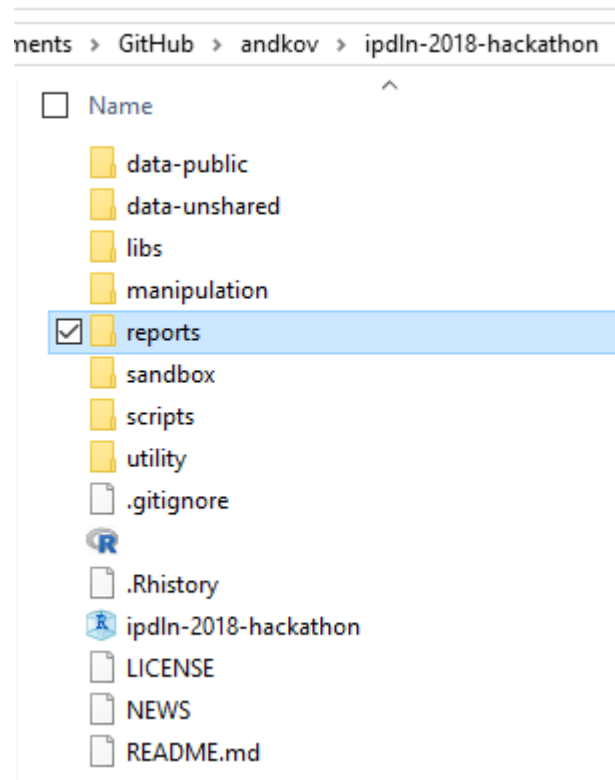


B. Workflow Highlights

1.3 **Layers of Isolation**: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)

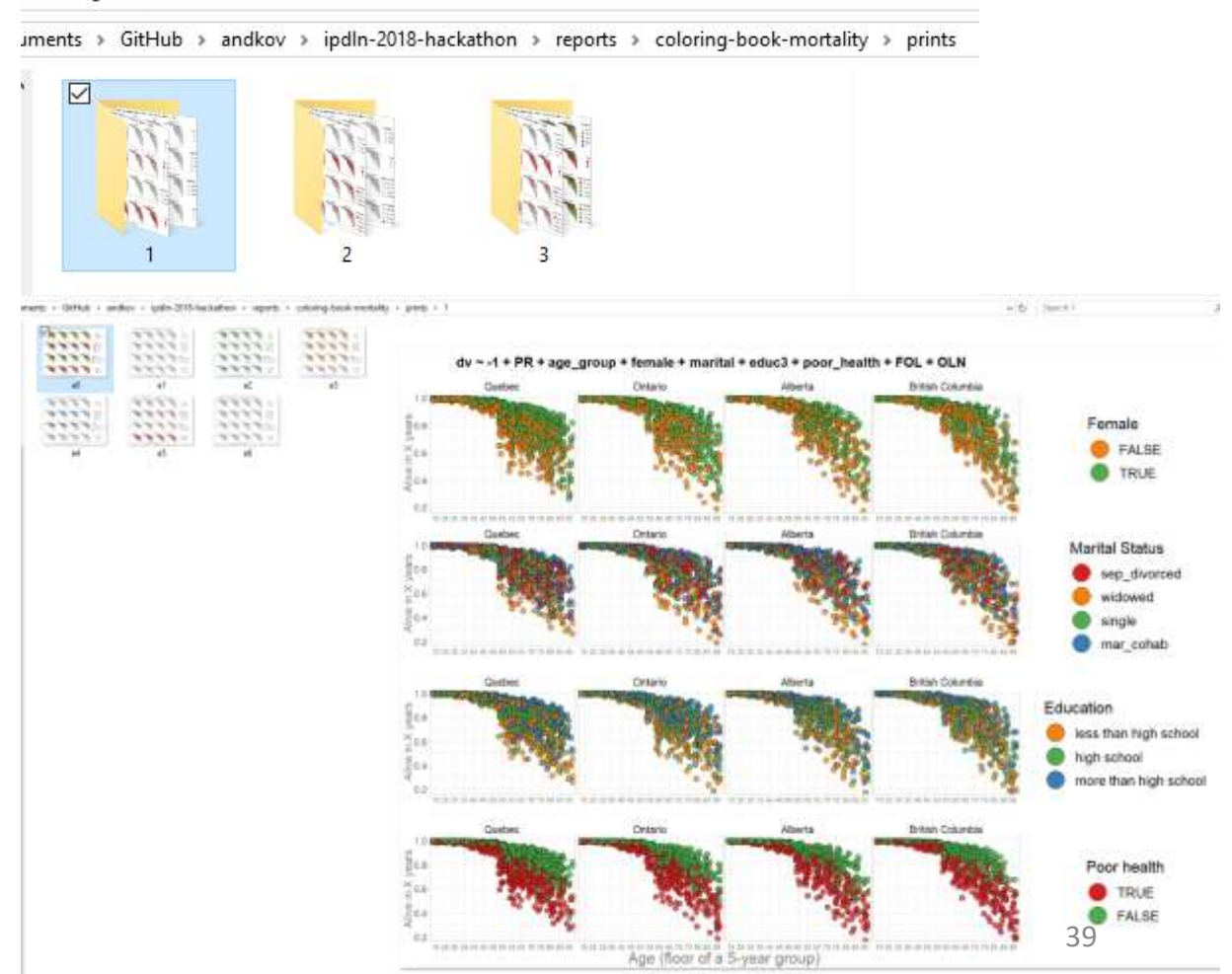
./reports/coloring-book-mortality/

Fails to separate modeling, graphing, and reporting



Screenshots of project repository

Name	Date
prints	2018-09-13 08:02
coloring-book-mortality	2018-09-12 15:23
ipdIn-2018-banff-hackathon-results-2018-09-14	2018-09-14 07:17
results-part-1	2018-09-13 23:41
results-part-2	2018-09-13 23:41
results-presentation-script.md	2018-09-14 07:30




B. Workflow Highlights

1.3 **Layers of Isolation**: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)

Technique demonstration


Branch: master ▾ ipdIn-2018-hackathon / README.md


 andkov Update README.md


- `./reports/technique-demonstration/` - a cleaned, simplified and heavily annotated .R + .Rmd version of [coloring-book-mortality.R](#) script. Optimized for learning the workflow with the original data. For full details consult its [stitched_output](#).
- `./reports/graphing-phase-only/` - focuses on the graphing phase of production. Fully reproducible: works with the results of the models estimated during [technical-demonstration](#), stored in `./data-public/dereived/technique-demonstration/`. For full details consult its [stitched_output](#)


ents > GitHub > andkov > ipdIn-2018-hackathon


☐ Name ^


 data-public


 data-unshared


 libs


 manipulation


☒  reports


 sandbox


 scripts


 utility


 .gitignore





 .Rhistory


 ipdIn-2018-hackathon


 LICENSE


 NEWS


 README.md









 coloring-book-mortality

 eda-1

☒  graphing-phase-only

 technique-demonstration

 README.md

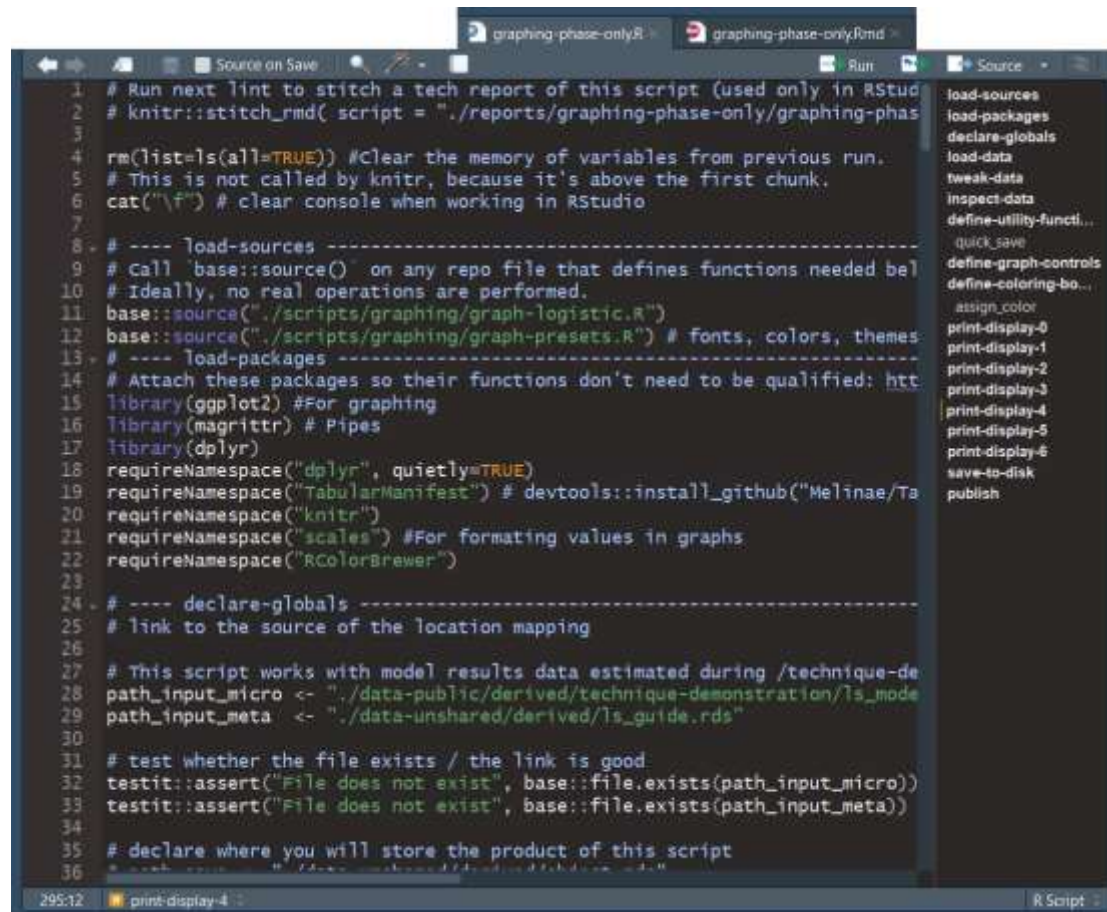
ents > GitHub > andkov > ipdIn-2018-hackathon > reports > graphing-phase-only				
<input type="checkbox"/> Name	Date modified	Type	Size	
 figure-png	2018-10-30 12:27	File folder		
 prints	2018-10-30 12:58	File folder		
 stitched_output	2018-10-30 13:48	File folder		
 graphing-phase-only.md	2018-10-30 13:40	MD File	24 KB	
<input checked="" type="checkbox"/>  graphing-phase-only	2018-10-30 13:43	R File	16 KB	
<input checked="" type="checkbox"/>  graphing-phase-only	2018-10-30 13:36	RMD File	5 KB	
 graphing-phase-only-1	2018-10-30 13:37	Chrome HTML Do...	2,805 KB	
 graphing-phase-only-2	2018-10-30 13:40	Chrome HTML Do...	2,771 KB	

B. Workflow Highlights

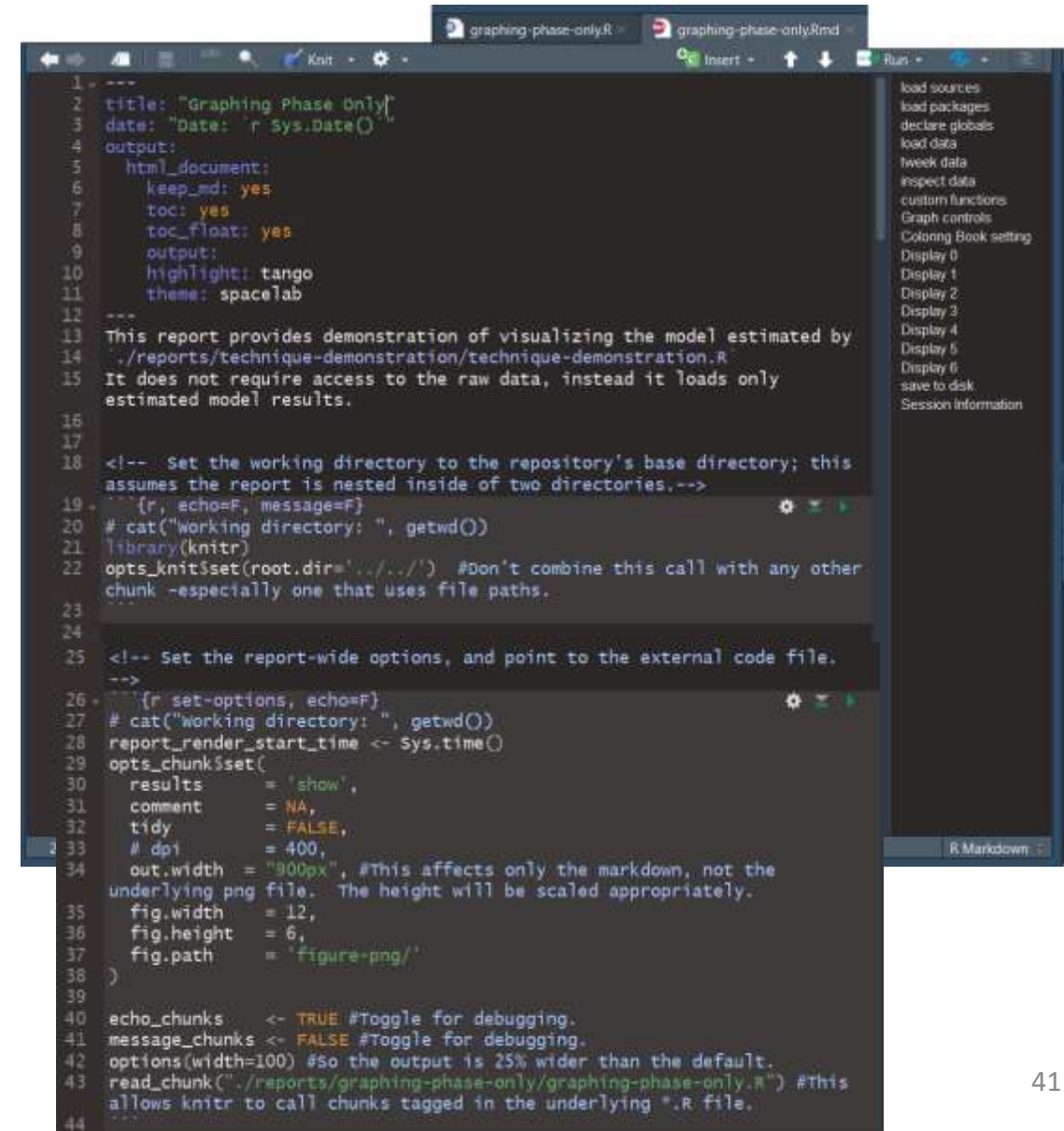
1.3 **Layers of Isolation**: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)

.R – stores analysis (what really happens)

.Rmd – stores presentation (how you tell about it)



```
1 # Run next lint to stitch a tech report of this script (used only in RStudio)
2 # knitr::stitch_rmd( script = "../reports/graphing-phase-only/graphing-phase-only.Rmd")
3
4 rm(list=ls(all=TRUE)) #Clear the memory of variables from previous run.
5 # This is not called by knitr, because it's above the first chunk.
6 cat("\n") # clear console when working in RStudio
7
8 # ---- load-sources ----
9 # Call 'base::source()' on any repo file that defines functions needed below
10 # Ideally, no real operations are performed.
11 base::source("../scripts/graphing/graph-logistic.R")
12 base::source("../scripts/graphing/graph-presets.R") # fonts, colors, themes
13 # ---- load-packages ----
14 # Attach these packages so their functions don't need to be qualified: http
15 library(ggplot2) #For graphing
16 library(magrittr) # Pipes
17 library(dplyr)
18 requireNamespace("dplyr", quietly=TRUE)
19 requireNamespace("TabularManifest") # devtools::install_github("Melinae/Ta
20 requireNamespace("knitr")
21 requireNamespace("scales") #For formatting values in graphs
22 requireNamespace("RColorBrewer")
23
24 # ---- declare-globals ----
25 # link to the source of the location mapping
26
27 # This script works with model results data estimated during /technique-de
28 path_input_micro <- "../data-public/derived/technique-demonstration/ls_mode
29 path_input_meta <- "../data-unshared/derived/ls_guide.rds"
30
31 # test whether the file exists / the link is good
32 testit::assert("File does not exist", base::file.exists(path_input_micro))
33 testit::assert("File does not exist", base::file.exists(path_input_meta))
34
35 # declare where you will store the product of this script
36 # path_output <- "../reports/graphing-phase-only/graphing-phase-only.Rmd"
```



```
1 ---
2 title: "Graphing Phase Only"
3 date: "Date: r Sys.Date()"
4 output:
5   html_document:
6     keep_md: yes
7     toc: yes
8     toc_float: yes
9     output:
10       highlight: tango
11       theme: spacelab
12 ---
13 This report provides demonstration of visualizing the model estimated by
14 ../reports/technique-demonstration/technique-demonstration.R
15 It does not require access to the raw data, instead it loads only
16 estimated model results.
17
18 <!-- Set the working directory to the repository's base directory; this
19 assumes the report is nested inside of two directories.-->
20 {r, echo=F, message=F}
21 # cat("working directory: ", getwd())
22 library(knitr)
23 opts_knit$set(root.dir='../..') #Don't combine this call with any other
24 chunk -especially one that uses file paths.
25
26 <!-- Set the report-wide options, and point to the external code file.
27 -->
28 {r set-options, echo=F}
29 # cat("working directory: ", getwd())
30 report_render_start_time <- Sys.time()
31 opts_chunk$set(
32   results = 'show',
33   comment = NA,
34   tidy = FALSE,
35   # dpi = 400,
36   out.width = "900px", #This affects only the markdown, not the
37   underlying png file. The height will be scaled appropriately.
38   fig.width = 12,
39   fig.height = 6,
40   fig.path = 'figure-png/'
41 )
42
43 echo_chunks <- TRUE #Toggle for debugging.
44 message_chunks <- FALSE #Toggle for debugging.
45 options(width=100) #So the output is 25% wider than the default.
46 read_chunk("../reports/graphing-phase-only/graphing-phase-only.R") #This
47 allows knitr to call chunks tagged in the underlying *.R file.
48 ---
```


B. Workflow Highlights

1.3 **Layers of Isolation**: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)

.R – stores analysis (what really happens)

.Rmd – stores presentation (how you tell about it)

<https://raw.githubusercontent.com/andkov/ipdIn-2018-hackathon/master/reports/graphing-phase-only/graphing-phase-only-1.html>

load sources

load packages

declare globals

load data

tweek data

inspect data

custom functions

Graph controls

Coloring Book setting

Display 0

Display 1

Display 2

Display 3

Display 4

Display 5

Display 6

save to disk

Session Information

Graphing Phase Only

Date: 2018-10-30

This report provides demonstration of visualizing the model estimated by

`./reports/technique-demonstration/technique-demonstration.R` It does not require access to the raw data, instead it loads only estimated model results.

load sources

```
# Call `base::source()` on any repo file that defines functions needed below.  
# Ideally, no real operations are performed.  
base::source("../scripts/graphing/graph-logistic.R")  
base::source("../scripts/graphing/graph-presets.R") # fonts, colors, themes
```

load packages

```
# Attach these packages so their functions don't need to be qualified: http://r-pkgs.had.co.nz/namespace.html#  
search-path  
library(ggplot2) #For graphing  
library(magrittr) # Pipes  
library(dplyr)  
requireNamespace("dplyr", quietly=TRUE)  
requireNamespace("TabularManifest") # devtools::install_github("Melinae/TabularManifest")  
requireNamespace("knitr")  
requireNamespace("scales") #For formatting values in graphs  
requireNamespace("RColorBrewer")
```

declare globals

```
# Link to the source of the location mapping  
  
# This script works with model results data estimated during /technique-demonstration/  
path_input_micro <- "../data-public/derived/technique-demonstration/ls_model.rds"  
path_input_meta <- "../data-unshared/derived/ls_guide.rds"  
  
# test whether the file exists / the link is good
```


B. Workflow Highlights

1.3 **Layers of Isolation**: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)

Technique demonstration

- `./reports/technique-demonstration/` - a cleaned, simplified and heavily annotated .R + .Rmd version of [coloring-book-mortality.R](#) script. Optimized for learning the workflow with the original data. For full details consult its [stitched_output](#).
- `./reports/graphing-phase-only/` - focuses on the graphing phase of production. Fully reproducible: works with the results of the models estimated during [technical-demonstration](#), stored in `./data-public/dereived/technique-demonstration/` . For full details consult its [stitched_output](#)

Branch: master ▾ ipdIn-2018-hackathon / README.md

 andkov Update README.md

nents > GitHub > andkov > ipdIn-2018-hackathon

☐ Name

data-public

data-unshared

libs

manipulation


☒ reports

sandbox

scripts

utility

.gitignore



.Rhistory

ipdIn-2018-hackathon

LICENSE

NEWS

README.md

coloring-book-mortality






☒ eda-1

graphing-phase-only

technique-demonstration

README.md

nents > GitHub > andkov > ipdIn-2018-hackathon > reports > eda-1

<input type="checkbox"/> Name	Date modified	Type	Size
<div>figure-png</div>	2018-09-05 15:53	File folder	
<div> eda-1</div>	2018-09-11 13:17	Chrome HTML Do...	1,963 KB
<div>eda-1.md</div>	2018-09-11 13:17	MD File	40 KB
<div><input checked="" type="checkbox"/>  eda-1</div>	2018-10-30 17:51	R File	4 KB
<div><input checked="" type="checkbox"/>  eda-1</div>	2018-09-05 16:29	RMD File	4 KB
<div> eda-1a-first-gen-immigrant</div>	2018-10-30 17:52	Chrome HTML Do...	1,943 KB
<div>eda-1a-first-gen-immigrant.md</div>	2018-10-30 17:52	MD File	41 KB
<div><input checked="" type="checkbox"/>  eda-1a-first-gen-immigrant</div>	2018-10-30 17:49	RMD File	4 KB

B. Workflow Highlights

1.4 Two essential means of production: `knitr::stitch()` vs `rmarkdown::render()`

Technique demonstration

- `./reports/technique-demonstration/` - a cleaned, simplified and heavily annotated .R + .Rmd version of `coloring-book-mortality.R` script. Optimized for learning the workflow with the original data. For full details consult its `stitched_output`.
- `./reports/graphing-phase-only/` - focuses on the graphing phase of production. Fully reproducible: works with the results of the models estimated during `technical-demonstration`, stored in `./data-public/dereived/technique-demonstration/`. For full details consult its `stitched_output`

Branch: master ▾ ipdln-2018-hackathon / README.md

andkov Update README.md

ents > GitHub > andkov > ipdln-2018-hackathon

☐ Name

data-public

data-unshared

libs

manipulation

☒ reports

sandbox

scripts

utility

.gitignore

.Rhistory

ipdln-2018-hackathon

LICENSE

NEWS

README.md

coloring-book-mortality

eda-1

graphing-phase-only

☒ technique-demonstration

README.md

ents > GitHub > andkov > ipdln-2018-hackathon > reports > technique-demonstration				
<input type="checkbox"/> Name	Date modified	Type	Size	
figure-png	2018-10-30 13:30	File folder		
prints	2018-10-30 12:42	File folder		
stitched_output	2018-10-30 09:01	File folder		
technique-demonstration.md	2018-10-30 13:39	MD File	52 KB	
<input checked="" type="checkbox"/> technique-demonstration	2018-10-30 13:42	R File	28 KB	
<input checked="" type="checkbox"/> technique-demonstration	2018-10-30 12:45	RMD File	6 KB	
technique-demonstration-1	2018-10-30 13:34	Chrome HTML Do...	2,854 KB	
technique-demonstration-2	2018-10-30 13:39	Chrome HTML Do...	2,820 KB	

ents > GitHub > andkov > ipdln-2018-hackathon > reports > technique-demonstration > stitched_output				
<input type="checkbox"/> Name	Date modified	Type	Size	
<input checked="" type="checkbox"/> technique-demonstration	2018-10-30 13:43	Chrome HTML Do...	77 KB	
technique-demonstration.md	2018-10-30 13:43	MD File	55 KB	

A. Graphing Technique

- 0.0 **Data & Context** : Mortality factors of Canadian immigrants at [IPDLN-2018 hackathon](#) by Statistics Canada in Banff
- 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 **Autonomous phases**: data cleaning, statistical modelling, graph production
- 1.3 **Layers of Isolation**: analysis vs presentation using .R (+ .Rmd) => .html (+ .pdf)
- 1.4 Two essential **means of production**: [knitr::stitch\(\)](#) vs [rmarkdown::render\(\)](#)

C. Conclusions

- 2.0 **Different than Notebooks**: sacrifices simplicity for agility via layers of isolation
- 2.1 **R (+ .Rmd) = .html (+ .pdf)** : moving away from *data playing* towards *data science*
- 2.2 **Reproducible projects**: moving away from notebooks towards software
- 2.3 **Looking back** to Neil Ernst talk:
 - Parameters and configuration
 - Hidden state
 - Longevity and version control
 - Testing and modularity
 - Notebook carpentry