Lesson 13: Purposeful model selection

Nicky Wakim

2024-03-04

Learning Objectives

1. Understand the overall steps for purposeful selection as a model building strategy
2. Apply purposeful selection to a dataset using R
3. Use different approaches to assess the linear scale of continuous variables in logistic regression

Regression analysis process

Model Selection

- Building a model
- Selecting variables
- Prediction vs interpretation
- Comparing potential models

model fit stat.

Model Fitting

- Find best fit line
- Using OLS in this class
- Parameter estimation
- Categorical covariates
- Interactions

Model Evaluation

- Evaluation of model fit
- Testing model assumptions
- Residuals
- Transformations
- Influential points
- Multicollinearity

Model Use (Inference)

- Inference for coefficients
- Hypothesis testing for coefficients
- Inference for expected Y given X
- Prediction of new Y given X

Learning Objectives

1. Understand the overall steps for purposeful selection as a model building strategy

2. Apply purposeful selection to a dataset using R

3. Use different approaches to assess the linear scale of continuous variables in logistic regression

"Successful modeling of a complex data set is part science, part statistical methods, and part experience and common sense."

Hosmer, Lemeshow, and Sturdivant Textbook, pg. 101

Overall Process

0. Exploratory data analysis
1. Check unadjusted associations in simple linear regression – Lab 3
2. Enter all covariates in model that meet some threshold
    - One textbook suggest $p < 0.2$ or $p < 0.25$: great for modest sized datasets
    - PLEASE keep in mind sample size in your study
    - Can also use magnitude of association rather than, or along with, p-value
3. Remove those that no longer reach some threshold
    - Compare magnitude of associations to unadjusted version (univariable)
4. Check scaling of continuous and coding of categorical covariates
5. Check for interactions
6. Assess model fit
    - Model assumptions, diagnostics, overall fit

Process with snappier step names

Pre-step: Exploratory data analysis (EDA)

Step 1: Simple linear regressions / analysis

Step 2: Preliminary variable selection

Step 3: Assess change in coefficients

Step 4: Assess scale for continuous variables

Step 5: Check for interactions

Step 6: Assess model fit

Learning Objectives

1. Understand the overall steps for purposeful selection as a model building strategy

2. Apply purposeful selection to a dataset using R

3. Use different approaches to assess the linear scale of continuous variables in logistic regression

Pre-step: Exploratory data analysis

- Things we have been doing over the quarter in class and in our project
- I will not discuss some of the methods mentioned in our lab and data management class
  - I am only going to introduce additional exploratory functions

A few things we can do:

- Check the data
- Study your variables
- Missing data?
- Explore simple relationships and assumptions

Pre-step: Exploratory data analysis: Check the data

- Get to know the potential values for the data
  - Categories
  - Units
- Then make sure the summary of values makes sense
  - If minimum or maximum look outside appropriate range
  - For example: a negative value for a measurement that is inherently positive (like population or income)

Search GAPMINDER Donate Resources About Log in

Home > Download the data > Documentation

Documentation

Gapminder combines data from multiple sources into unique coherent time-series that can't be found elsewhere.

Most of our data are not good enough for detailed numeric analysis. They are only good enough to revolutionize people's worldview. But we only fill in gaps whenever we believe we know roughly what the numbers would have been, had they existed. The uncertainties are often large. But we comfort ourselves by knowing the errors in peoples worldview are even larger. Our data is constantly improved by feedback in our data forum from users finding mistakes.

We fill in all gaps: Our data is more consistent over time and space than most other sources, because we dare to fill all the gaps in the sources. We dare this because our purpose is to show people the big picture, and they won't understand it if its full of holes.

We use current geographic boundaries: We show the world history as if country borders had always been the same as today. Read more here.

Below are links to documentation describing how we have combined the sources in each case. For the sake of transparency, whenever allowed to share the underlying data, we make our complete calculations available for download, often in Excel files. In most of these files the details are not documented, as we haven't had time to describe every little step in our data process. But our data is constantly being improved by people who help find problems. If you have questions, we will try to answer them in our data-forum.

Each documentation page has a version number and links to the previous versions. Whenever we update the data, or make other significant changes in the documentation, we make a new version.

Data combined by Gapminder
Average age at 1st marriage (girls)
Babies per woman (total fertility rate)
Child Mortality Rate, under age five
GDP per capita in constant PPP dollars
Gini
HIV/AIDS
Income Mountains
Infant Mortality Rate, under age one
Legal slavery
Life Expectancy at Birth
Maternal mortality
Population
World Health Chart, data sources

This list only includes data that we have somehow modified or calculated ourselves. The complete list of data we use is here »

https://www.gapminder.org/data/documentation/

Pre-step: Exploratory data analysis: Check the data

- Look at a summary for the raw data
- Typical use:

```r
library(skimr)
skim(gapm)
```

- Some `skim()` help

Pre-step: Exploratory data analysis: Check the data

- Look at a summary for the raw data
- Typical use:

```
library(skimr)
skim(gapm)
```

- Some `skim()` help
- Note that `skim(gapm)` looks different because I had to create factors
- I am breaking down the `skim()` function into the categorical and continuous variables only because I want to show them on the slides

```
skim(gapm_sub1) %>% yank("factor")
```

Variable type: factor

skim_variable	n_missing	complete_rate	ordered	n_unique	top_counts
four_regions	0	1.00	FALSE	4	Asi: 57, Afr: 54, Eur: 49, Ame: 35
income_levels1	1	0.99	~~FALSE~~ TRUE	4	Hig: 56, Upp: 55, Low: 52, Low: 31
income_levels2	1	0.99	FALSE	2	Hig: 111, Low: 83

A : 2

Pre-step: Exploratory data analysis: Check the data

```
skim(gapm_sub1) %>% yank("numeric")
```

Summary( )

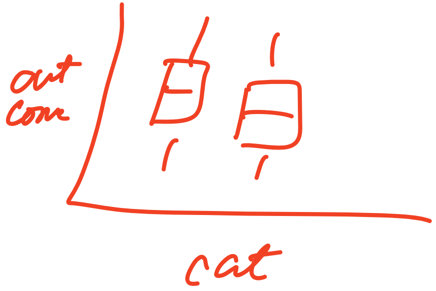
Variable type: numeric

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
CO2emissions	4	0.98	4.55	6.10	0.03	0.64	2.41	6.22	41.20	
ElectricityUsePP	58	0.70	4220.92	5964.07	31.10	699.00	2410.00	5600.00	52400.00	
FoodSupplykcPPD	27	0.86	2825.06	443.59	1910.00	2490.00	2775.00	3172.50	3740.00	
IncomePP	2	0.99	16704.45	19098.61	614.00	3370.00	10100.00	22700.00	129000.00	
LifeExpectancyYrs	8	0.96	70.66	8.44	47.50	64.30	72.70	76.90	82.90	
FemaleLiteracyRate	115	0.41	81.65	21.95	13.00	70.97	91.60	98.03	99.80	
WaterSourcePrct	1	0.99	84.84	18.64	18.30	74.90	93.50	99.07	100.00	
Latitude	0	1.00	19.11	23.93	-42.00	4.00	17.33	40.00	65.00	
Longitude	0	1.00	21.98	66.52	-175.00	-5.75	21.00	49.27	179.14	
population_mill	0	1.00	35.95	136.87	0.00	1.73	7.57	24.50	1370.00	

Poll Everywhere Question 1

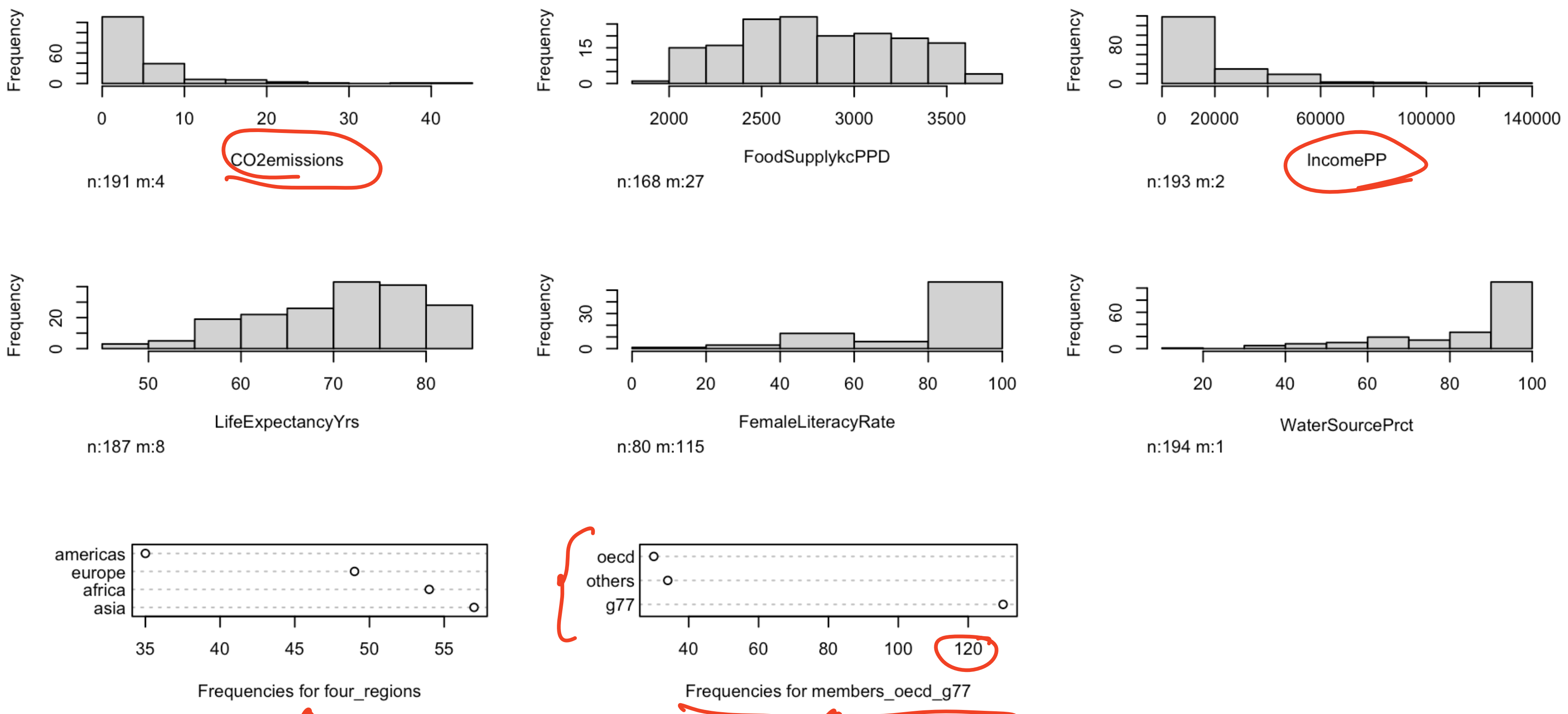
Pre-step: Exploratory data analysis: Study your variables

- Started this a little bit in previous slide (`skim()`), but you may want to look at things like:
  - Sample size
  - Counts of missing data
  - Means and standard deviations
  - IQRs
  - Medians
  - Minimums and maximums
- Can also look at visuals
  - Continuous variables: histograms (in `skimr() a little)
  - Categorical variables: ~~frequency plots~~ boxplots

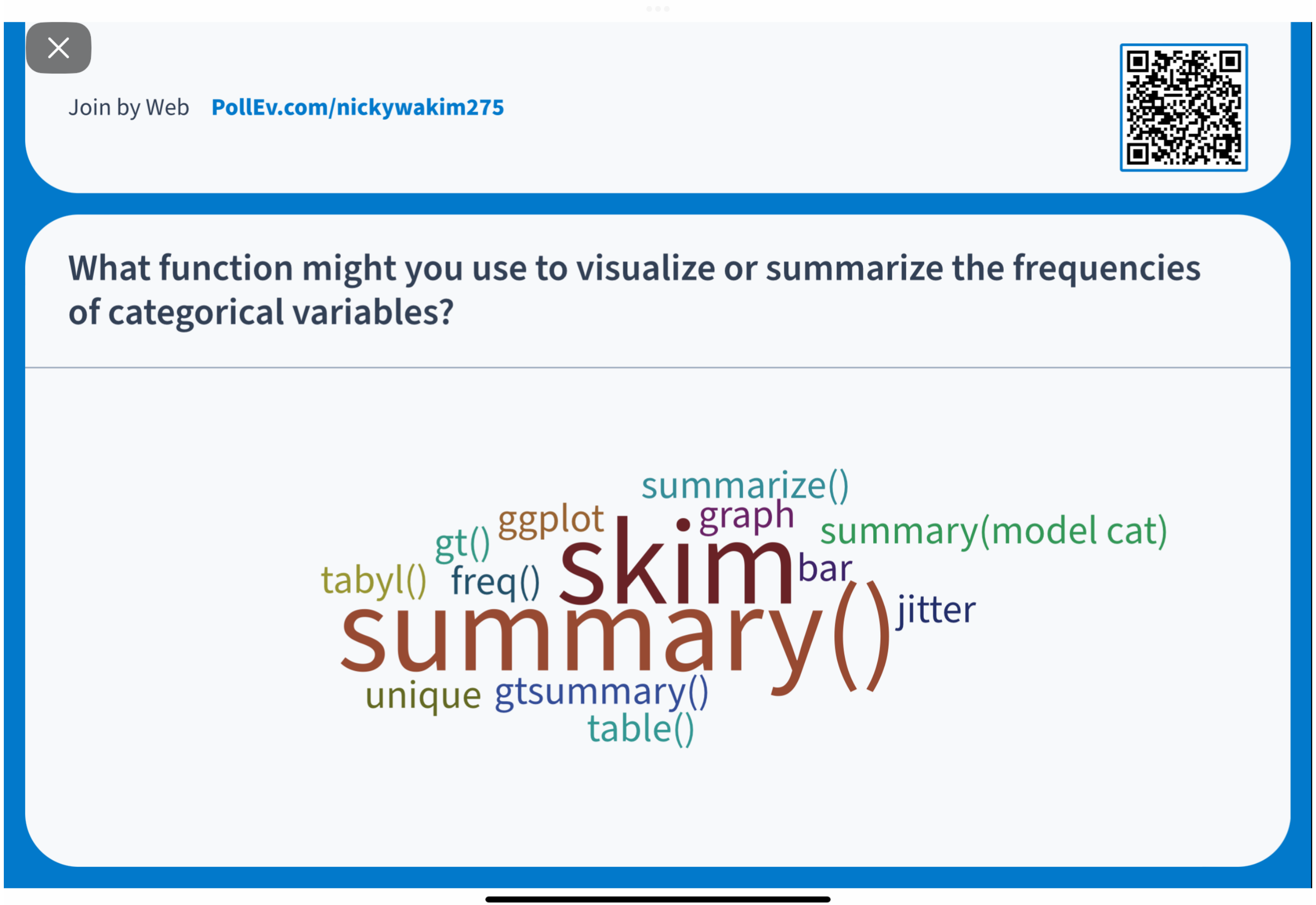


Pre-step: Exploratory data analysis: Study your variables

```
library(Hmisc)
hist.data.frame(gapm %>% select(-Longitude, -Latitude, -eight_regions, -six_regions, -geo, -`World bank, 4 income groups
```



Poll Everywhere Question 2



Pre-step: Exploratory data analysis: Missing data

- Why are there missing data?
- Which variables and observations should be excluded because of missing data?
- Will I impute missing data?

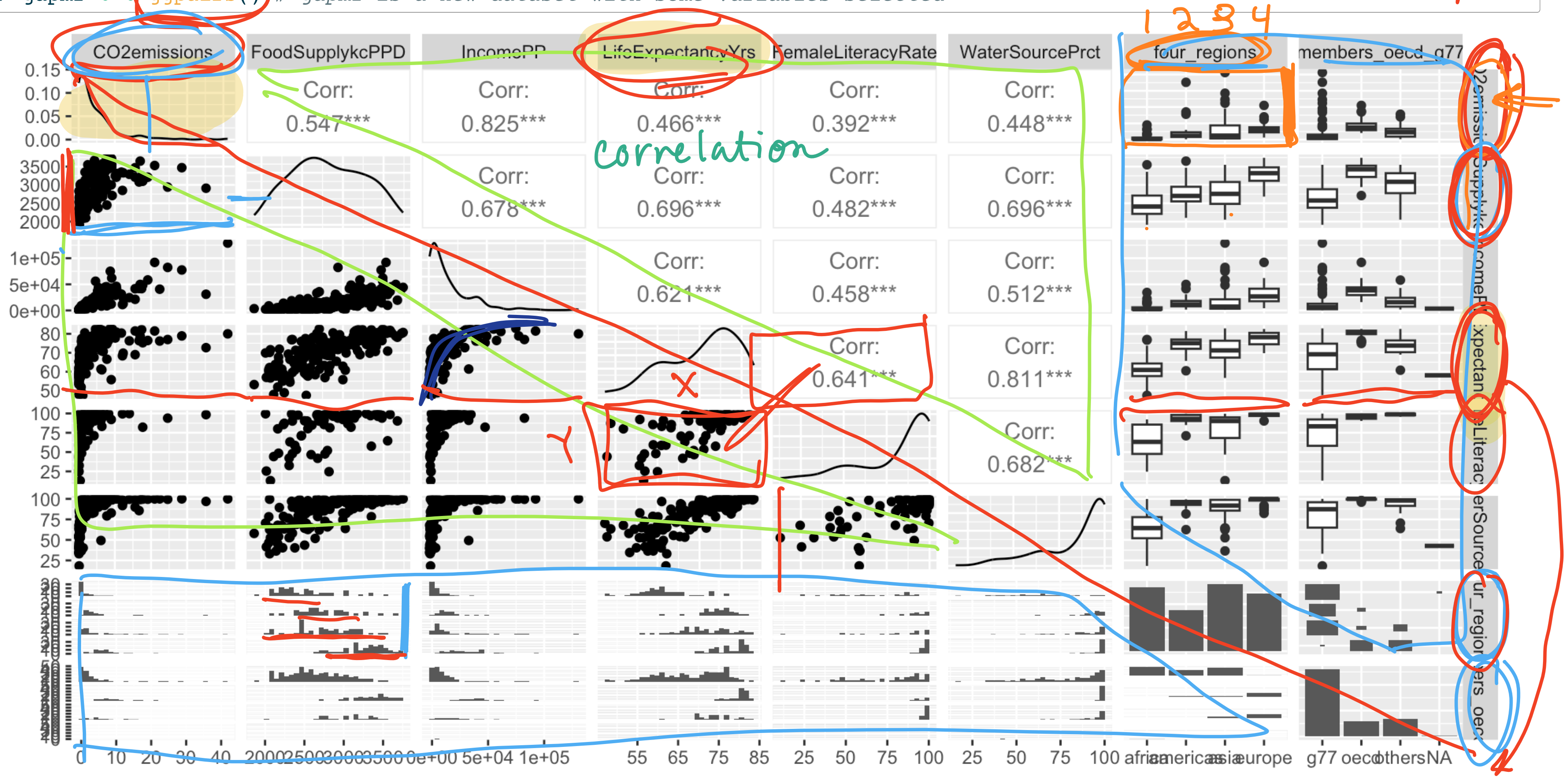
- Unfortunately, we don't have time to discuss missing data more thoroughly
- I will try to cover this topic more thoroughly in BSTA 513

- For the Gapminder dataset, we chose to use complete cases

Pre-step / Step 1 : Explore simple relationships and assumptions

do NOT put in Report!

```
1  gapm2 %>% ggpairs() # gapm2 is a new dataset with some variables selected
```



Poll Everywhere Question 3

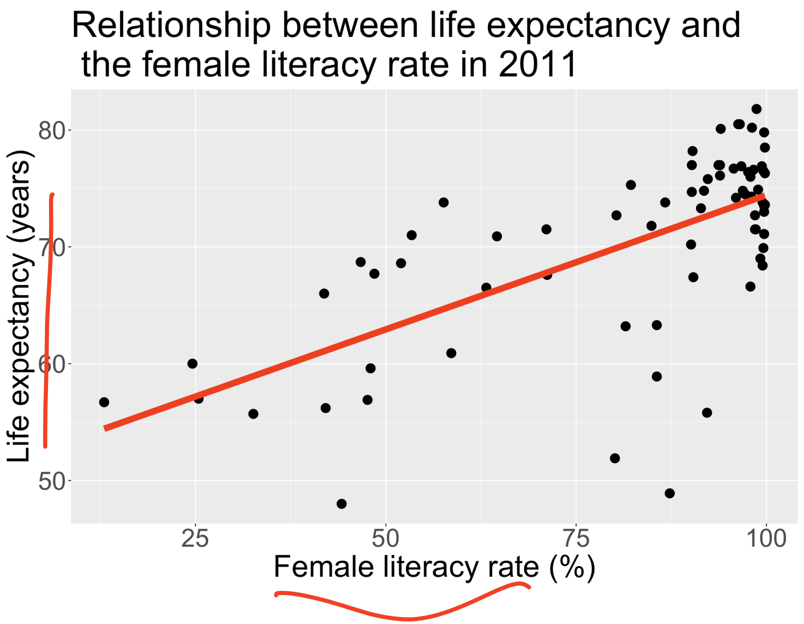
Step 1: Simple linear regressions / analysis

- For each covariate, we want to see how it relates to the outcome (without adjusting for other covariates)
- We can partially do this with **visualizations**
  - Helps us see the data we throw it into regression that makes assumptions (like our LINE assumptions)
  - `ggpairs()` can be a quick way to do it
  - `ggplot()` can make each plot
    - `+ geom_boxplot()` to make boxplots by groups for categorical covariates
    - `+ geom_jitter() + stat_summary()` to make non-overlaping points with group means for categorical covariates
    - `+ geom_point()` to make scatterplots for continuous covariates
- We need to run **simple linear regression**
  - We're calling regression with multi-level categories "simple" even though there are multiple coefficients

Step 1: Simple linear regressions / analysis

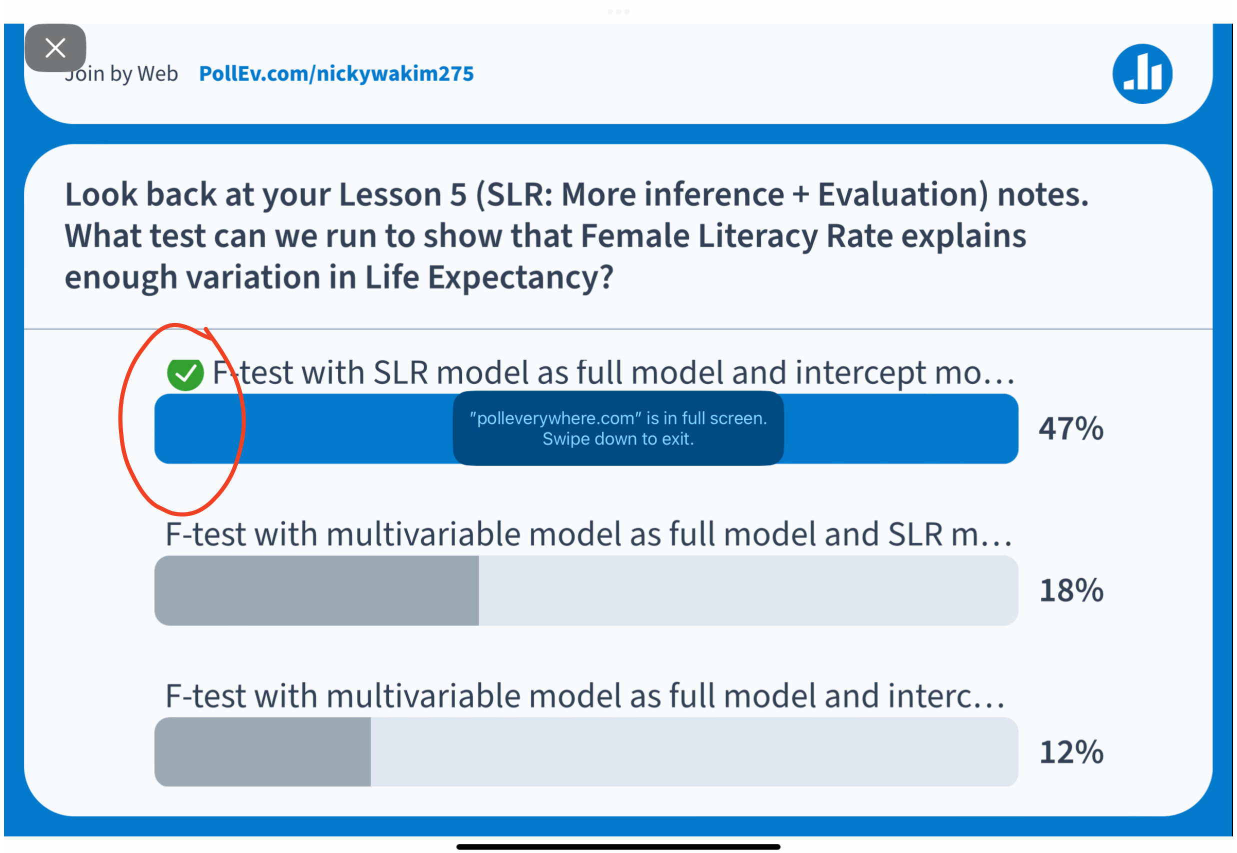
- Let's think back to our Gapminder dataset
- Always good to start with our main relationship: life expectancy vs. female literacy rate
  - *Throwback to Lesson 3 SLR when we first visualized and ran `lm()` for this relationship*

```
model_FLR = lm(LifeExpectancyYrs ~ FemaleLiteracyRate, data = gapm_sub)
```



term	estimate	std.error	statistic	p.value
(Intercept)	51.438	2.739	18.782	0.000
FemaleLiteracyRate	0.230	0.032	7.141	0.000

Poll Everywhere Question 4



$Y = \beta_0 + \underline{\beta_1 X} + \varepsilon$ ← full

Red:

$Y = \beta_0 + \varepsilon$

full:

$Y = \beta_0 + \beta_1 I(X=2) + \beta_2 I(X=3) + \varepsilon$

where $X = 1$ is Ref

Red: $Y = \beta_0 + \varepsilon$

Step 1: Simple linear regressions / analysis

- Let's do this with one other variable before I show you a streamlined version of SLR

```
model_WR = lm(LifeExpectancyYrs ~ four_regions, data = gapm_sub)
```

Step 1: Simple linear regressions / analysis

- If we do a good job visualizing the relationship between our outcome and each covariate, then we can proceed to a streamlined version of the F-test for each relationship
- First, I will select the variables that we are considering for model selection:

```
gapm2 = gapm_sub %>% select(LifeExpectancyYrs, CO2emissions, FoodSupplykcPPD,
                            IncomePP, FemaleLiteracyRate, WaterSourcePrct,
                            four_regions, members_oecd_g77)
```

- We need to make sure our dataset only contains the variables we are considering for the model:

```
gapm3 = gapm2 %>% select(-LifeExpectancyYrs)
```

Step 1: Simple linear regressions / analysis

- Now I can run the `lapply()` function, which allows me to run the same function multiple times over all the columns in `gapm3`
- For each covariate I am running: `lm(gapm2$LifeExpectancyYrs ~ x) %>% anova()` (F-test)
  - So I am fitting the simple linear regression and printing the ANOVA table with F-test (comparing model with & without the covariate)

```
lapply( gapm3, function(x) lm(gapm2$LifeExpectancyYrs ~ x) %>% anova() )
```

y y

```
$CO2emissions
Analysis of Variance Table

Response: gapm2$LifeExpectancyYrs
          Df Sum Sq Mean Sq F value   Pr(>F)
x          1  452.3  452.31  7.6536 0.007241 **
Residuals 70 4136.8   59.10
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

$FoodSupplykcPPD
Analysis of Variance Table

Response: gapm2$LifeExpectancyYrs
          Df Sum Sq Mean Sq F value    Pr(>F)
```

- We can scroll through the output to see the ANOVA table for each covariate

Step 1: Simple linear regressions / analysis

F-test: does X explain variation LE?

- We can also filter the ANOVA table to just show the p-value for each F-test

```
1  sapply( gapm3, function(x) anova( lm(gapm2$LifeExpectancyYrs ~ x) )$`Pr(>F)` )
```

```
     CO2emissions FoodSupplykcPPD     IncomePP FemaleLiteracyRate
[1,]  0.007241207    1.187753e-09 3.557341e-06       6.894997e-10
[2,]           NA              NA           NA                 NA
     WaterSourcePrct four_regions members_oecd_g77
[1,]    1.148644e-17 1.857818e-13      7.55261e-05
[2,]              NA           NA               NA
```

- Row 1 is the p-value for the F-test
  - This will help us in Step 2

lower p-value = variance explained

Step 2: Preliminary variable selection

- Identify candidates for your first multivariable model by performing an F-test on each covariate's SLR
  - Using p-values from previous slide
  - If the p-value of the test is less than 0.25, then consider the variable a candidate

- Candidates for first multivariable model
  - All clinically important variables (regardless of p-value)
  - Variables with ~~univariate test~~ with p-value < 0.25

SLR F-test

- With more experience, you won't need to rely on these strict rules as much

Step 2: Preliminary variable selection

- From the previous p-values from the F-test on each covariate's SLR
  - Decision: we keep all the covariates since they all have a p-value < 0.25

```
sapply( gapm3, function(x) anova( lm(gapm2$LifeExpectancyYrs ~ x) )$`Pr(>F)` )
```

```
     CO2emissions FoodSupplykcPPD     IncomePP FemaleLiteracyRate
[1,]  0.007241207    1.187753e-09 3.557341e-06       6.894997e-10
[2,]           NA              NA           NA                 NA
     WaterSourcePrct four_regions members_oecd_g77
[1,]    1.148644e-17 1.857818e-13      7.55261e-05
[2,]              NA           NA               NA
```

all less than 0.25

Step 2: Preliminary variable selection

- Fit an **initial model** including any independent variable with p-value < 0.25 and clinically important variables

```
init_model = lm(LifeExpectancyYrs ~ FemaleLiteracyRate + CO2emissions + IncomePP +
                four_regions + WaterSourcePrct + FoodSupplykcPPD + members_oecd_g77,
                  data = gapm2)
tidy(init_model, conf.int = T) %>% gt() %>% tab_options(table.font.size = 30) %>%
  fmt_number(decimals = 4)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	37.5560	4.4083	8.5194	0.0000	28.7410	46.3710
FemaleLiteracyRate	0.0020	0.0352	0.0580	0.9539	−0.0684	0.0725
CO2emissions	−0.2860	0.1340	−2.1344	0.0368	−0.5539	−0.0181
IncomePP	0.0002	0.0001	2.4133	0.0188	0.0000	0.0003
four_regionsAmericas	9.8963	2.0031	4.9405	0.0000	5.8909	13.9017
four_regionsAsia	5.7849	1.5993	3.6172	0.0006	2.5870	8.9829
four_regionsEurope	7.1421	2.6994	2.6458	0.0104	1.7442	12.5399
WaterSourcePrct	0.1377	0.0658	2.0928	0.0405	0.0061	0.2693
FoodSupplykcPPD	0.0052	0.0021	2.4961	0.0153	0.0010	0.0093
members_oecd_g77oecd	−0.3317	2.5476	−0.1302	0.8968	−5.4259	4.7625
members_oecd_g77others	0.3341	2.2986	0.1453	0.8849	−4.2622	4.9304

Step 3: Assess change in coefficient

- This is where we start identifying covariates that we might remove
- I would start by using the p-value to guide me towards specific variables
  - Female literacy rate, but that's our main covariate
  - `members_oecd_g77` ✓
  - Maybe water source percent? ✓

g77: developing countries w/ common econ interests

term	estimate	std.error	statistic	p.value
(Intercept)	37.5560	4.4083	8.5194	0.0000
FemaleLiteracyRate	0.0020	0.0352	0.0580	0.9539
CO2emissions	−0.2860	0.1340	−2.1344	0.0368
IncomePP	0.0002	0.0001	2.4133	0.0188
four_regionsAmericas	9.8963	2.0031	4.9405	0.0000
four_regionsAsia	5.7849	1.5993	3.6172	0.0006
four_regionsEurope	7.1421	2.6994	2.6458	0.0104
WaterSourcePrct	0.1377	0.0658	2.0928	0.0405
FoodSupplykcPPD	0.0052	0.0021	2.4961	0.0153
members_oecd_g77oecd	−0.3317	2.5476	−0.1302	0.8968
members_oecd_g77others	0.3341	2.2986	0.1453	0.8849

- Some people will say you can use the p-value alone
  - I like to double check that those variables do not have a large effect on the other coefficients

main effect of explanatory var.

Step 3: Assess change in coefficient

- Very similar to the process we used when looking at confounders

initial model

- One variable at a time, we run the multivariable model with and without the variable
  - We look at the p-value of the F-test for the coefficients of said variable — oecd, g77, others: F-test p-value w/in MLR model
  - We look at the percent change for the coefficient ($\Delta\%$) of our explanatory variable

- General rule: We can remove a variable if…
  - p-value > 0.05 for the F-test of its own coefficients
  - AND change in coefficient ($\Delta\%$) of our explanatory variable is < 10%

→ 10% indicates confounder

Step 3: Assess change in coefficient

- Let's try this out on `members_oecd_g77`
  - ▶ Display the ANOVA table with F-statistic and p-value

term	df.residual	rss	df	sumsq	statistic	p.value
LifeExpectancyYrs ~ FemaleLiteracyRate + CO2emissions + IncomePP + four_regions + WaterSourcePrct + FoodSupplykcPPD + members_oecd_g77	61.000	999.201	NA	NA	NA	NA
LifeExpectancyYrs ~ FemaleLiteracyRate + CO2emissions + IncomePP + four_regions + WaterSourcePrct + FoodSupplykcPPD	63.000	1,000.988	−2.000	−1.787	0.055	0.947

- $\widehat{\beta}_{FLR,full} = 0.002, \widehat{\beta}_{FLR,red} = 0.0036$

$$\Delta\% = 100\% \cdot \frac{\widehat{\beta}_{FLR,full} - \widehat{\beta}_{FLR,red}}{\widehat{\beta}_{FLR,full}} = 100\% \cdot \frac{0.002 - 0.0036}{0.002} = -74.41\%$$

- Based off the percent change, I would keep this in the model

can remove members vars BUT check Δ%

Step 3: Assess change in coefficient

- Let's try this out on water source percent (even though the p-value was < 0.05)

▶ Display the ANOVA table with F-statistic and p-value

term	df.residual	rss	df	sumsq	statistic	p.value
LifeExpectancyYrs ~ FemaleLiteracyRate + CO2emissions + IncomePP + four_regions + WaterSourcePrct + FoodSupplykcPPD + members_oecd_g77	61.000	999.201	NA	NA	NA	NA
LifeExpectancyYrs ~ FemaleLiteracyRate + CO2emissions + IncomePP + four_regions + members_oecd_g77 + FoodSupplykcPPD	62.000	1,070.944	−1.000	−71.744	4.380	0.041

- $\widehat{\beta}_{FLR,full} = 0.002, \widehat{\beta}_{FLR,red} = 0.034$

$$\Delta\% = 100\% \cdot \frac{\widehat{\beta}_{FLR,full} - \widehat{\beta}_{FLR,red}}{\widehat{\beta}_{FLR,full}} = 100\% \cdot \frac{0.002 - 0.034}{0.002} = -1561.06\%$$

- Based off the percent change (and p-value), I would keep this in the model

Poll Everywhere Question 5

Join by Web PollEv.com/nickywakim275

Which of the following is/are an important inclusion rule for variables when we are model building?

Including variables that are clinically significant 0% → decision

Including variables that are significant in the model 0% → guides us

Including variables that are confounders or effect modi… 0% → decision

✓ All of the above 74%

A and B only

Step 3: Assess change in coefficient

- At the end of this step, we have a **preliminary main effects model**
- Where the variables are excluded that met the following criteria:
  - P-value > 0.05 for the F-test of its own coefficients
  - Change in coefficient ($\Delta\%$) of our explanatory variable is < 10%
- In our example, the **preliminary main effects model** (end of Step 3) was the same as the **intiial model** (end of Step 2)

Remaining slides under construction

Learning Objectives

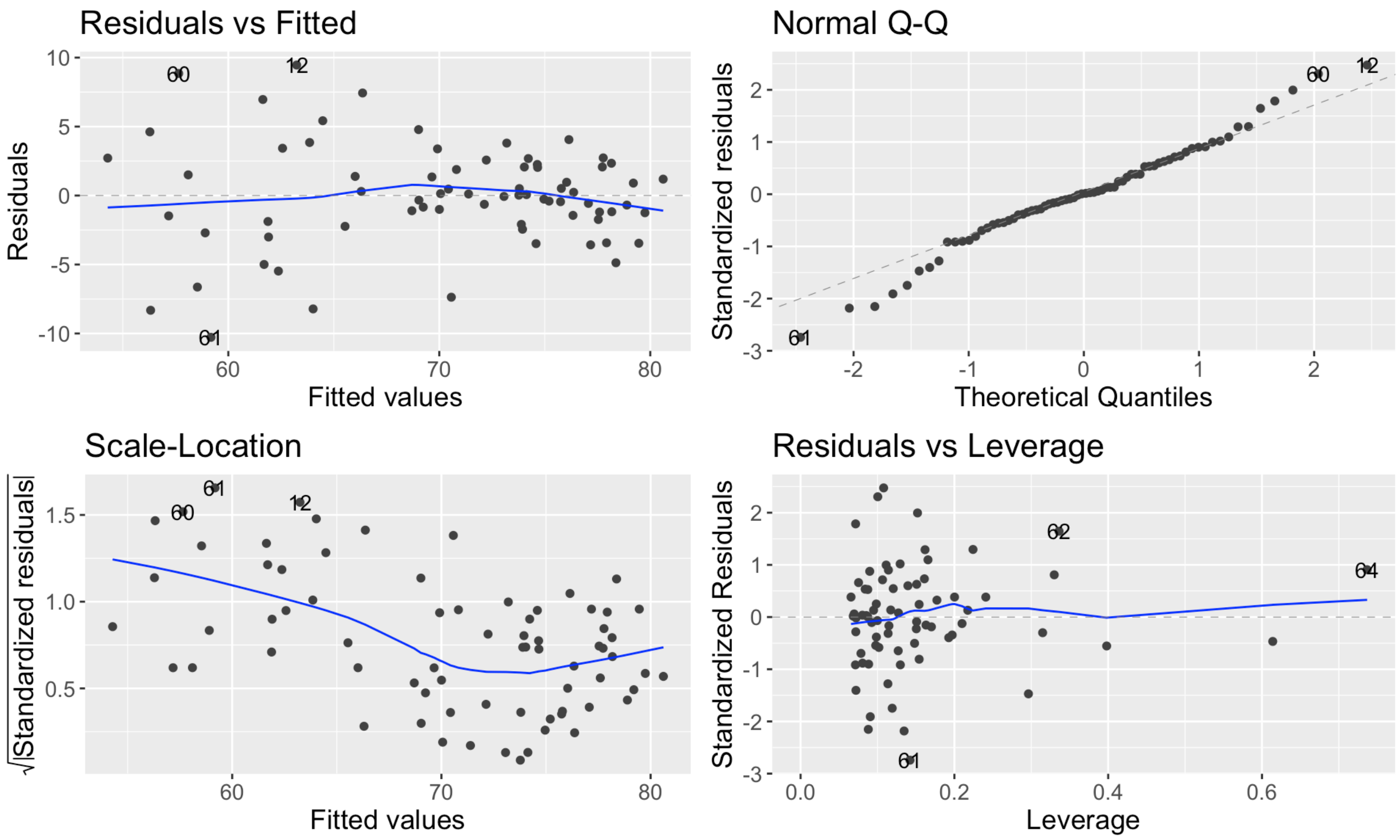
1. Understand the overall steps for purposeful selection as a model building strategy
2. Apply purposeful selection to a dataset using R
3. Use different approaches to assess the linear scale of continuous variables in logistic regression

Step 4: Assess scale for continuous variables

- We assume the linear regression model is linear for **each continuous variable**
- We need to assess linearity for continuous variables in the model
  - Do this through smoothed scatterplots that we introduced in Lesson 6 (SLR Diagnostics)
  - Residual plots (can be used in SLR) does not help us in MLR
  - Each term in MLR model needs to have linearity with outcome
- Three methods/approaches to address the violation of linearity assumption:
  - Approach 1: Quantile method/Indicator variables
  - Approach 2: Fractional Polynomials
  - Approach 3: Spline functions
- For our class, only implement **Approach 2 or 3**
- Model at the end of Step 4 is the **main effects model**

Step 4: Assess scale for continuous variables

- Residual plot does not help us with linearity in MLR



Step 4: Assess scale for continuous variables: Smoothed scatterplots

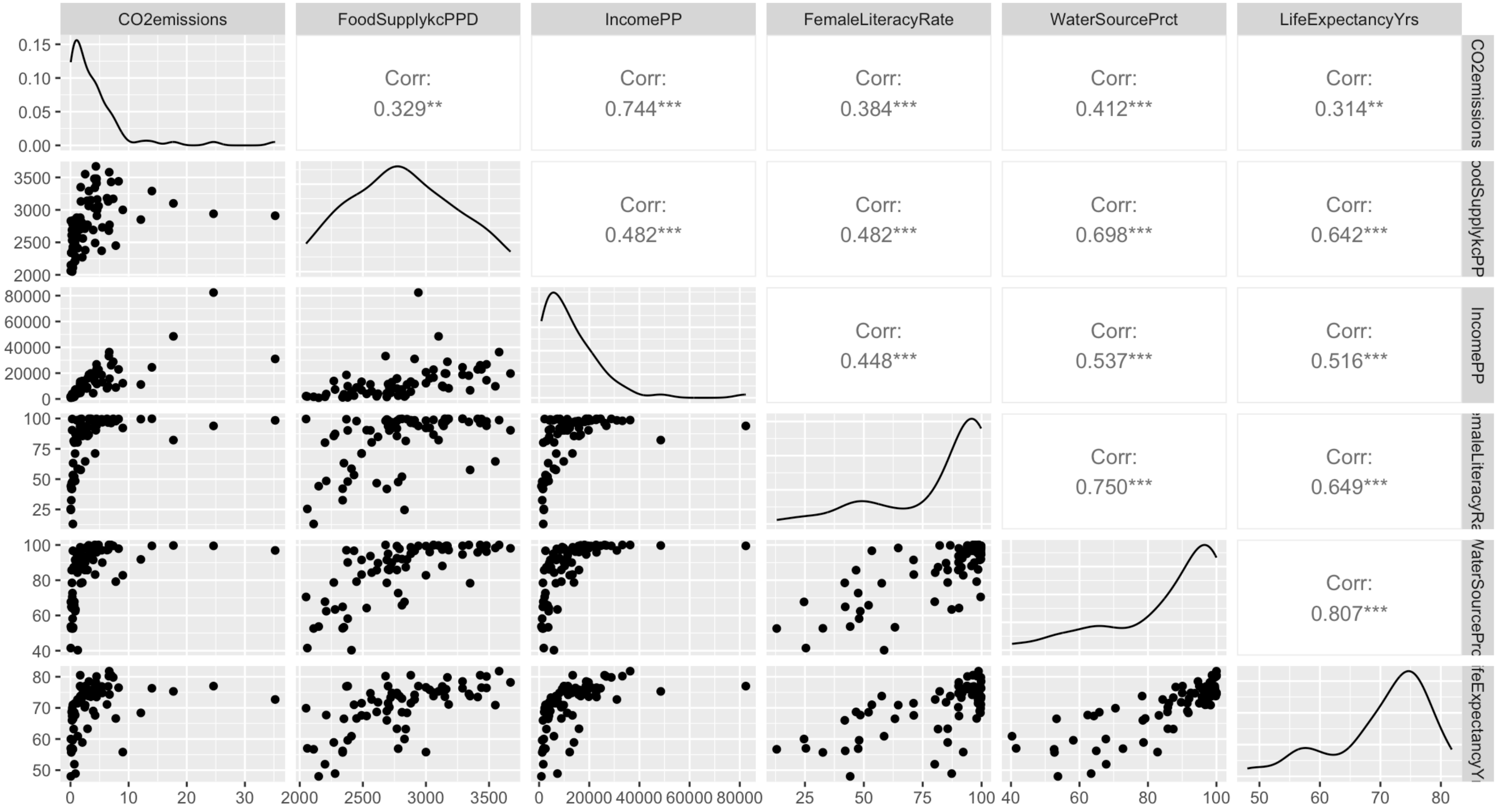
- **Only checking linearity**, not addressing linearity issues
- Can also identify extreme observations
  - Which can influence the assessment of linearity when using fractional polynomials or spline functions
- Plot the observed and smoothed values of outcome vs. continuous variable
- Helps us decide if the continuous variable can stay **as is** in the model
  - Problem: if not linear, then we need to represent the variable in a new way (Approaches 2-4)

Step 4: Assess scale for continuous variables: Smoothed scatterplots

- In Gapminder dataset, we have 5 continuous variables:
  - CO2 Emissions
  - Food Supply
  - Income
  - Female Literacy Rate
  - Water source percent
- Plot each of these agains the outcome, life expectancy

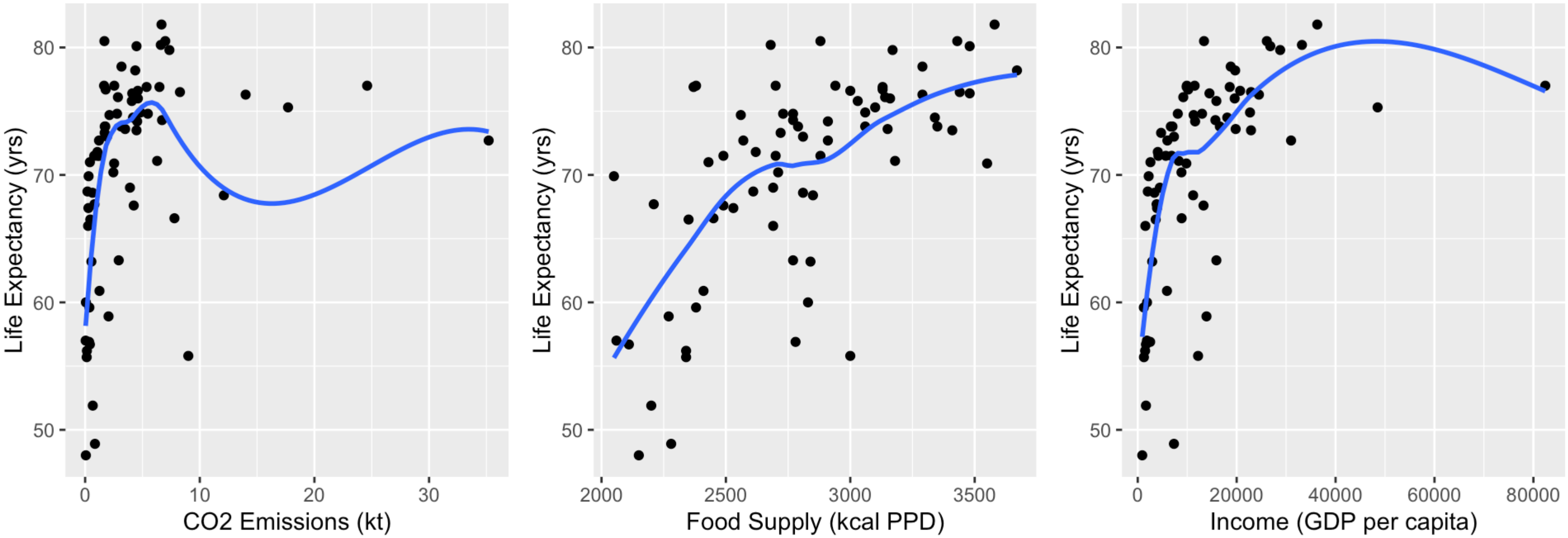
Step 4: Assess scale for continuous variables: Smoothed scatterplots

- We can quickly look at ggpairs() to identify variables



Step 4: Assess scale for continuous variables: Smoothed scatterplots

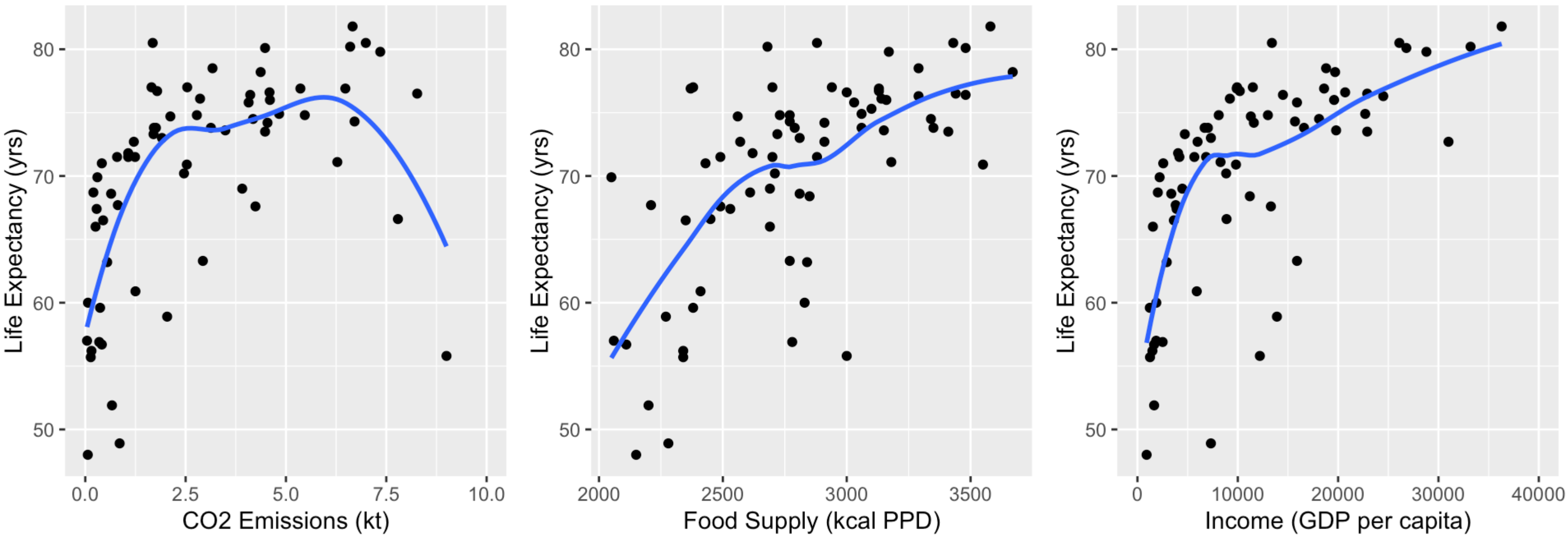
- Take a look at C02, Food Supply, and Income



- Food Supply looks admissible
- CO2 Emissions and Income do not look very linear, but I want to zoom into the area of the plots that have most of the data

Step 4: Assess scale for continuous variables: Smoothed scatterplots

- Zoom into areas on plots with more data



- Food Supply still looks admissible
- CO2 Emissions and Income not linear: will address this!!

Step 4: Assess scale for continuous variables

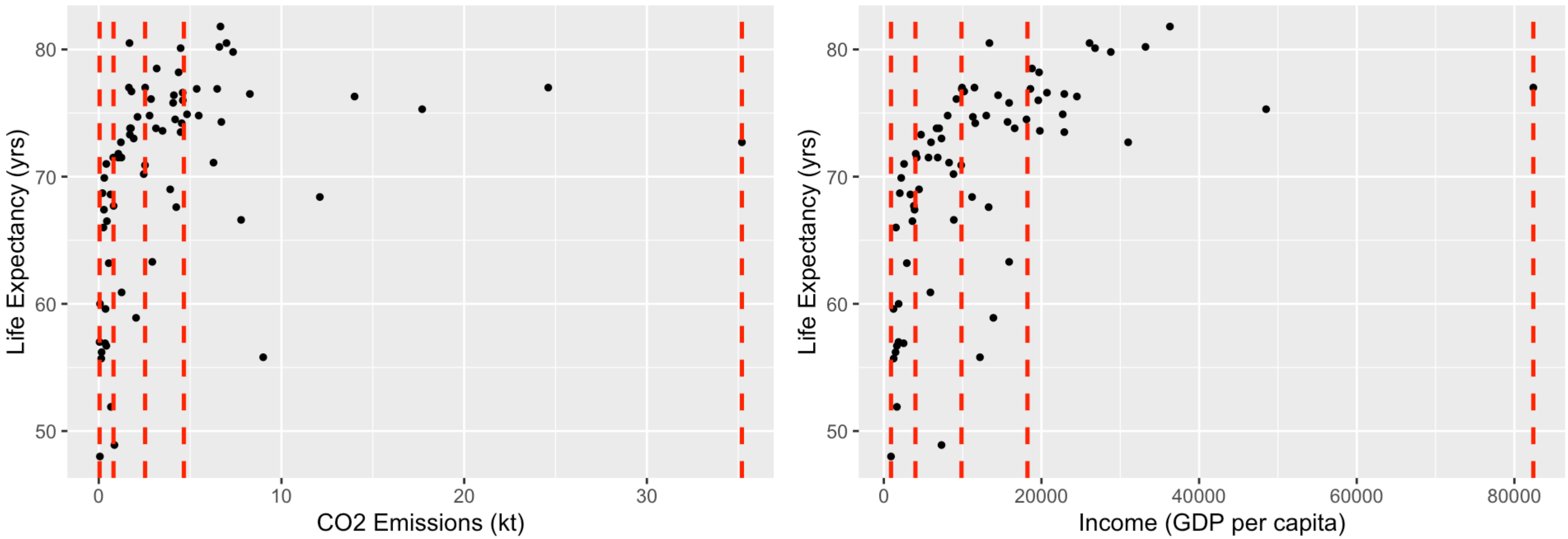
- Three methods/approaches to address the violation of linearity assumption:
  - Approach 1: Quantile method/Indicator variables
  - Approach 2: Fractional Polynomials
  - Approach 3: Spline functions

Step 4: Approach 1: Quantile method/Indicator variables

- Split a continuous variable into its quartiles
  - Create dummy variables corresponding to each quartile
  - Fit logistic regression with the dummy variables
  - Plot quartile midpoints vs. coefficient estimates for the respective dummy variables
- Disadvantages:
  - Takes some time to create new variables, especially with multiple continuous covariates
  - Start with quartiles, but might be more appropriate to use different splits
    - No set rules on this
- Advantage: graphical and visually helps

Step 4: Approach 1: Quantile method/Indicator variables

- Take a look at the quartiles within the scatterplot



Step 4: Approach 2: Fractional Polynomials

Step 4: Approach 3: Spline functions

Learning Objectives

1. Understand the overall steps for purposeful selection as a model building strategy

2. Apply purposeful selection to a dataset using R

3. Use different approaches to assess the linear scale of continuous variables in logistic regression

Step 5: Check for interactions

- Create a list of interaction terms from variables in the "main effects model" that has clinical plausibility
- Add the interaction variables, one at a time, to the main effects model, and assess the significance using a likelihood ratio test or Wald test
  - May keep interaction terms with p-value < 0.05
- Keep the main effects untouched, only simplify the interaction terms – locked!
- Use methods from Step 2 (comparing model with all interactions to a smaller model with interactions) to determine which interactions to keep
- The model by the end of Step 6 is called the preliminary final model

Step 6: Assess model fit

- Assess the adequacy of the model and check its fit
- Methods will be discussed later class
- If the model is adequate and fits well, then it is the Final model

Next time

- More details on steps 4-6 on Monday before quiz!