Introduction to Multiple Linear Regression (MLR)

Nicky Wakim

2024-02-05

Learning Objectives

1. Understand equations and visualizations that helps us build multiple linear regression model.
2. Fit MLR model (in R) and understand the difference between fitted regression plane and regression lines.
3. Identify the population multiple linear regression model and define statistics language for key notation.
4. Based off of previous SLR work, understand how the population MLR is estimated.
5. Interpret MLR (population) coefficient estimates with additional variable in model

Reminder of what we learned in the context of SLR

- SLR helped us establish the foundation for a lot of regression
  - But we do not usually use SLR in analysis

What did we learn in SLR??

Model Fitting

- Ordinary least squares (OLS)
- `lm()` function in R

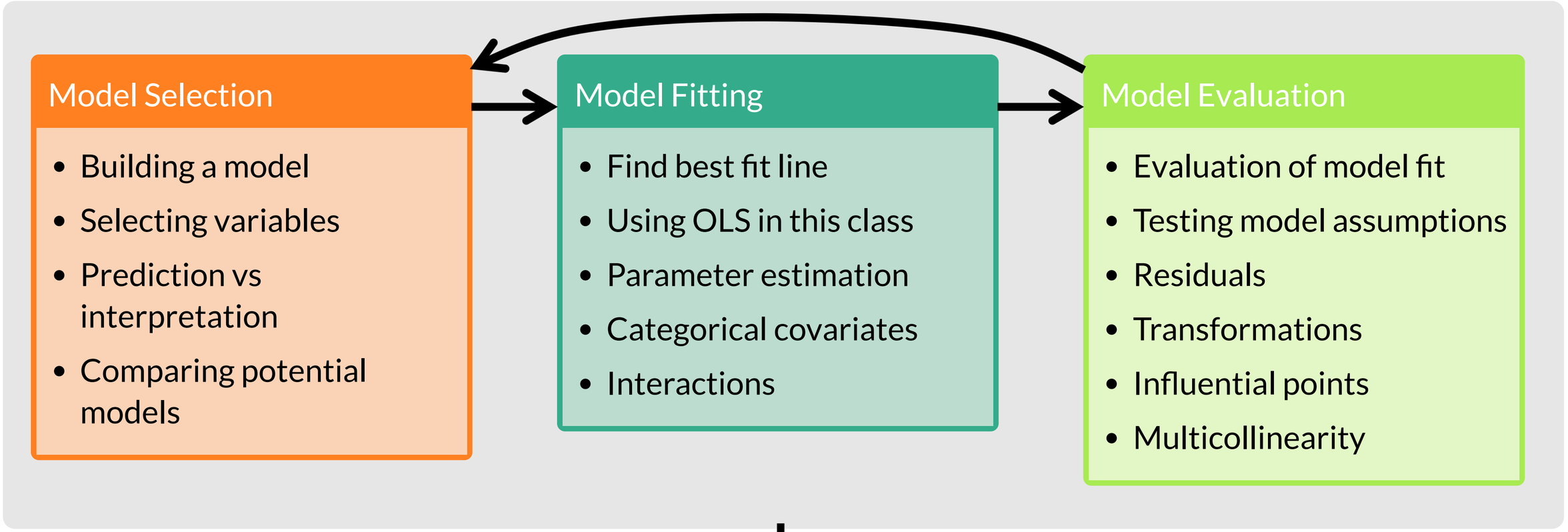
Model Use

- Inference for variance of residuals
- Hypothesis testing for coefficients
- Interpreting population coefficient estimates
- Calculated the expected mean for specific X values
- Interpreted coefficient of determination

Model Evaluation/Diagnostics

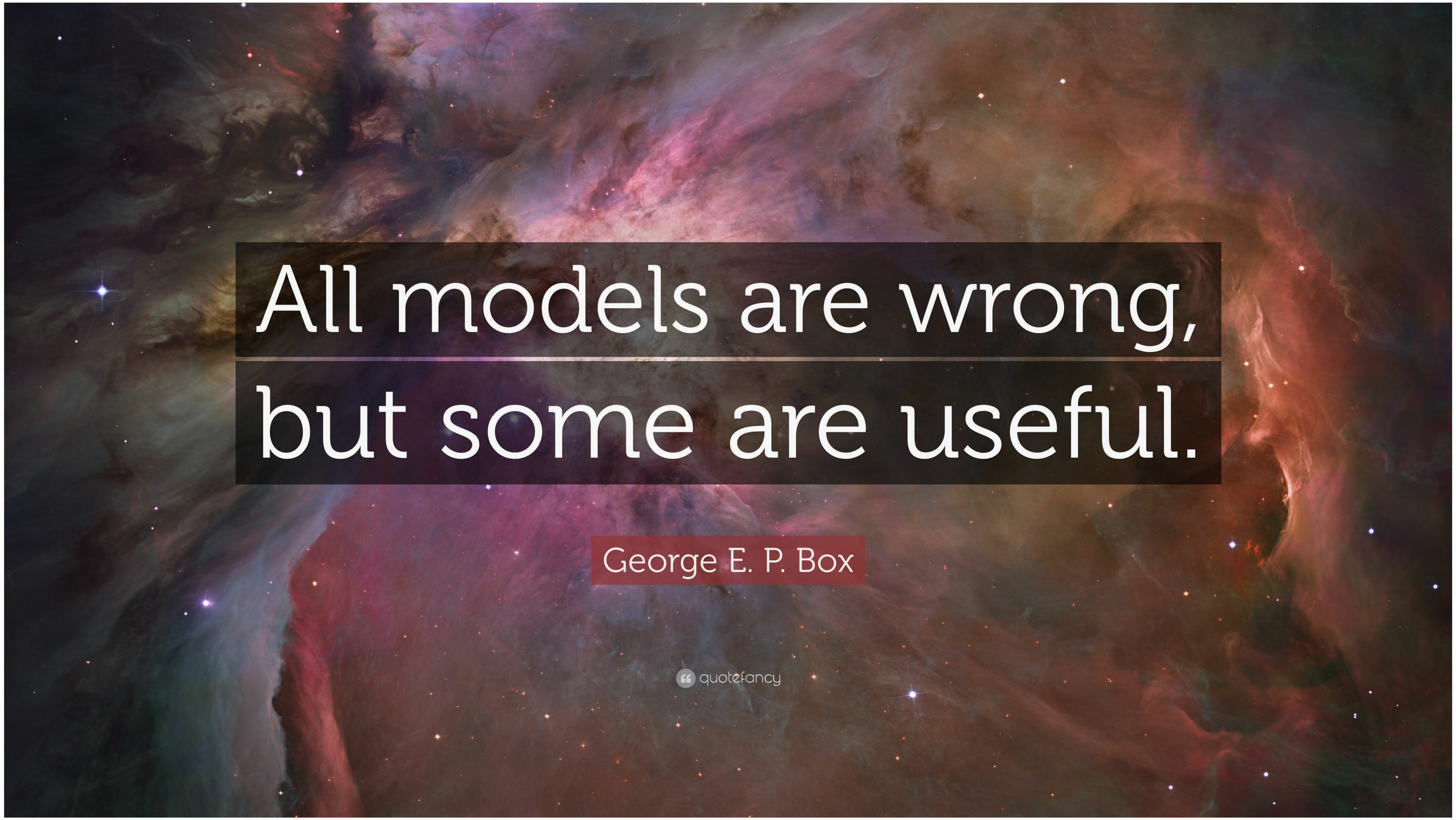
- LINE Assumptions
- Influential points
- Data Transformations

Let's map that to our regression analysis process



Model Use (Inference)

- Inference for coefficients
- Hypothesis testing for coefficients
- Inference for expected Y given X


All models are wrong,
but some are useful.
George E. P. Box
quotefancy

Learning Objectives

1. Understand equations and visualizations that helps us build multiple linear regression model.
2. Fit MLR model (in R) and understand the difference between fitted regression plane and regression lines.
3. Identify the population multiple linear regression model and define statistics language for key notation.
4. Based off of previous SLR work, understand how the population MLR is estimated.
5. Interpret MLR (population) coefficient estimates with additional variable in model

Simple Linear Regression vs. Multiple Linear Regression

Simple Linear Regression

We use **one predictor** to try to explain the variance of the outcome

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Multiple Linear Regression

We use **multiple predictors** to try to explain the variance of the outcome

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- Has $k + 1$ total coefficients (including intercept) for k predictors/covariates
- Sometimes referred to as ***multivariable*** linear regression, but *never multivariate*

- The models have similar "LINE" assumptions and follow the same general diagnostic procedure

Going back to our life expectancy example

- Let's say many other variables were measured for each country, including food supply
  - **Food Supply** (kilocalories per person per day, kc PPD): the average kilocalories consumed by a person each day.
- In SLR, we only had one predictor and one outcome in the model:
  - **Life expectancy** = the average number of years a newborn child would live if current mortality patterns were to stay the same.
  - **Adult literacy rate** is the percentage of people ages 15 and above who can, with understanding, read and write a short, simple statement on their everyday life.

- Do we think adult female literacy rate is going to explain a lot of the variance of life expectancy between countries?

Loading the (new-*ish*) data

```
# Load the data - update code if the csv file is not in the same location on your c
# If you need to download the file, please go to ur shared folder under Data > Slid
gapm <- read_excel("data/Gapminder_vars_2011.xlsx",
                   na = "NA")  # important!!!!

gapm_sub <- gapm %>%
  drop_na(LifeExpectancyYrs, FemaleLiteracyRate, FoodSupplykcPPD)

glimpse(gapm_sub)
```

```
Rows: 72
Columns: 18
$ country                <chr> "Afghanistan", "Albania", "Angola",…
$ CO2emissions           <dbl> 0.4120, 1.7900, 1.2500, 5.3600, 4.6…
$ ElectricityUsePP       <dbl> NA, 2210, 207, NA, 2900, 1810, 258,…
$ FoodSupplykcPPD        <dbl> 2110, 3130, 2410, 2370, 3160, 2790,…
$ IncomePP               <dbl> 1660, 10200, 5910, 18600, 19600, 70…
$ LifeExpectancyYrs      <dbl> 56.7, 76.7, 60.9, 76.9, 76.0, 73.8,…
$ FemaleLiteracyRate     <dbl> 13.0, 95.7, 58.6, 99.4, 97.9, 99.5,…
$ population             <dbl> 2.97e+07, 2.93e+06, 2.42e+07, 9.57e…
```

Can we improve our model by adding food supply as a covariate?

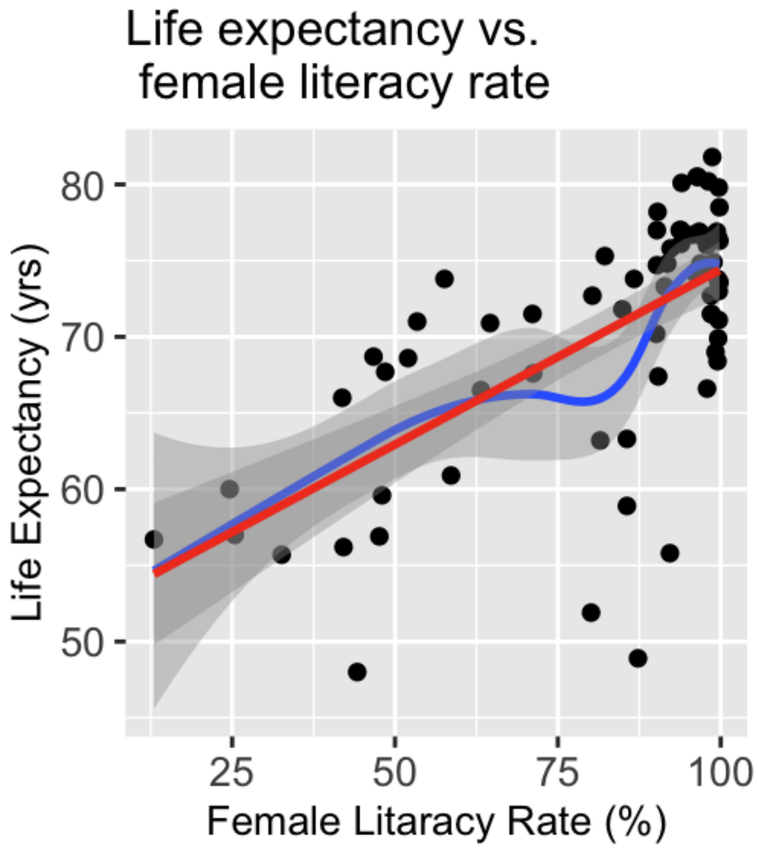
Simple linear regression population model

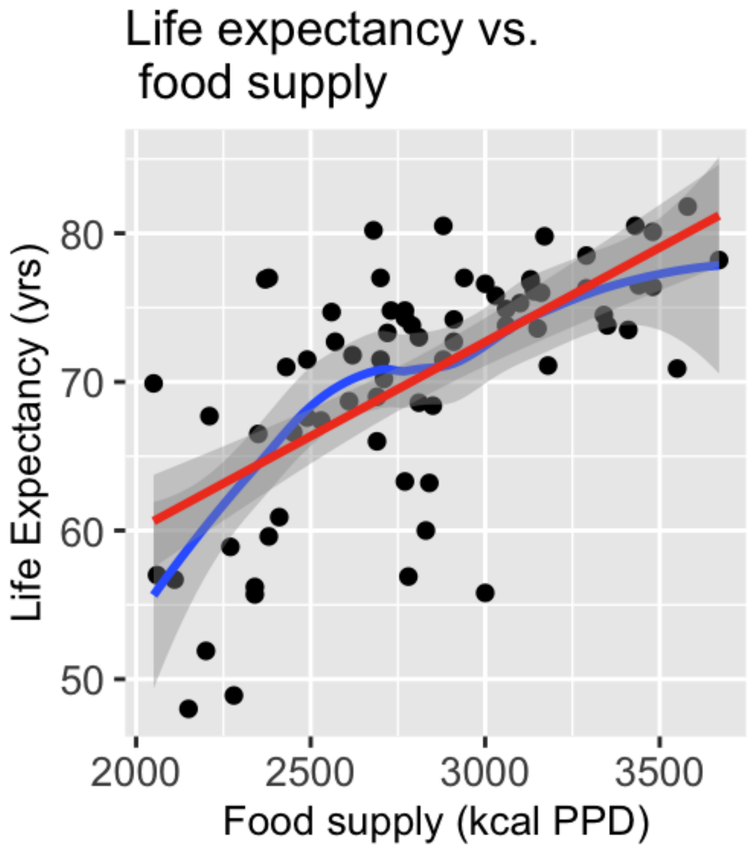
$$\text{Life expectancy} = \beta_0 + \beta_1 \text{Female literacy rate} + \epsilon$$
$$\text{LE} = \beta_0 + \beta_1 \text{FLR} + \epsilon$$

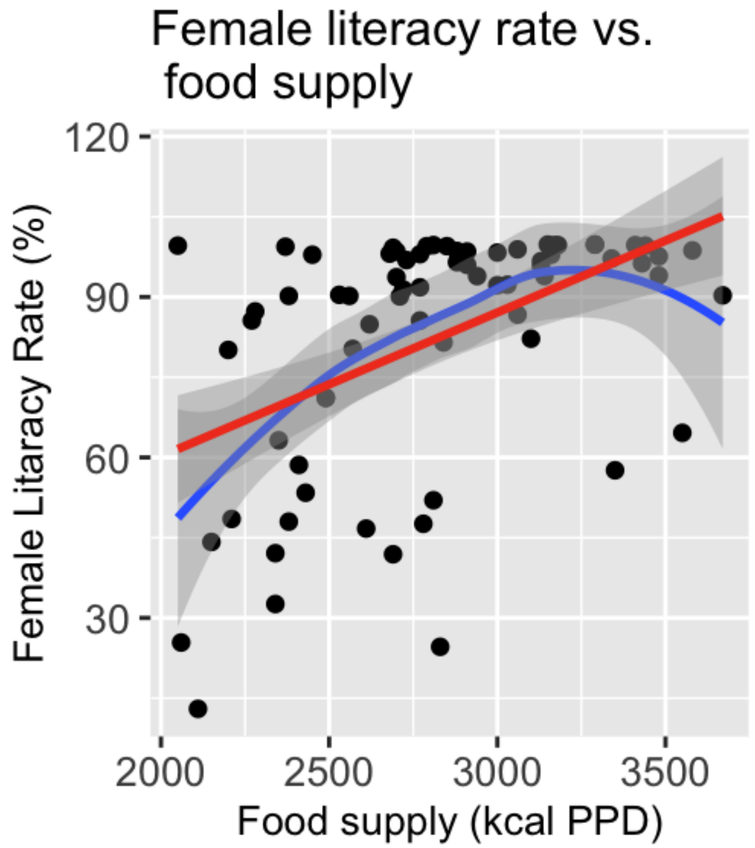
Multiple linear regression population model (with added Food Supply)

$$\text{Life expectancy} = \beta_0 + \beta_1 \text{Female literacy rate} + \beta_2 \text{Food supply} + \epsilon$$
$$\text{LE} = \beta_0 + \beta_1 \text{FLR} + \beta_2 \text{FS} + \epsilon$$

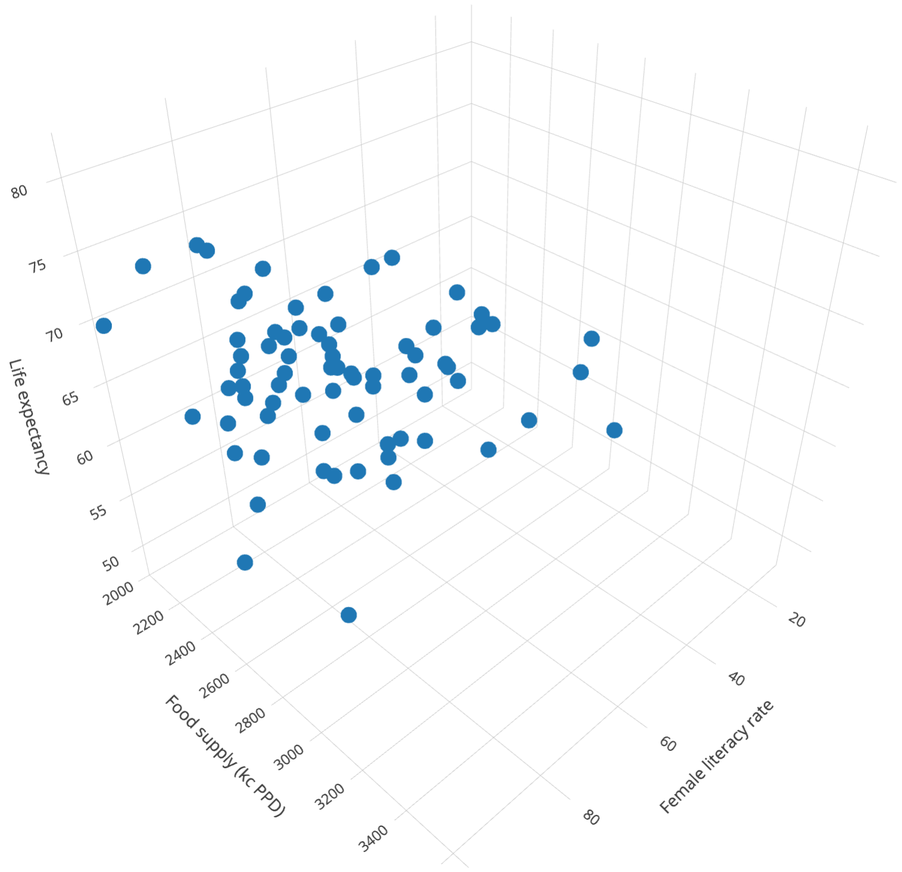
Visualize relationship between life expectancy, female literacy rate, and food supply







Visualize relationship in 3-D



Poll Everywhere Question 1

Learning Objectives

1. Understand equations and visualizations that helps us build multiple linear regression model.
2. Fit MLR model (in R) and understand the difference between fitted regression plane and regression lines.
3. Identify the population multiple linear regression model and define statistics language for key notation.
4. Based off of previous SLR work, understand how the population MLR is estimated.
5. Interpret MLR (population) coefficient estimates with additional variable in model

How do we fit a multiple linear regression model in R?

New population model for example:

$$\text{Life expectancy} = \beta_0 + \beta_1 \text{Female literacy rate} + \beta_2 \text{Food supply} + \epsilon$$

```
# Fit regression model:
mr1 <- lm(LifeExpectancyYrs ~ FemaleLiteracyRate + FoodSupplykcPPD,
          data = gapm_sub)
tidy(mr1, conf.int=T) %>% gt() %>% tab_options(table.font.size = 35) %>% fmt_number
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	33.595	4.472	7.512	0.000	24.674	42.517
FemaleLiteracyRate	0.157	0.032	4.873	0.000	0.093	0.221
FoodSupplykcPPD	0.008	0.002	4.726	0.000	0.005	0.012

Fitted multiple regression model:

$$\widehat{\text{Life expectancy}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{Female literacy rate} + \widehat{\beta}_2 \text{Food supply}$$

$$\widehat{\text{Life expectancy}} = 33.595 + 0.157\ \text{Female literacy rate} + 0.008\ \text{Food supply}$$

Don’t forget `summary()` to extract information!

```
1  summary(mr1)
```

```
Call:
lm(formula = LifeExpectancyYrs ~ FemaleLiteracyRate + FoodSupplykcPPD,
    data = gapm_sub)

Residuals:
    Min      1Q  Median      3Q     Max
-17.715  -2.328   1.052   3.022   9.083

Coefficients:
```

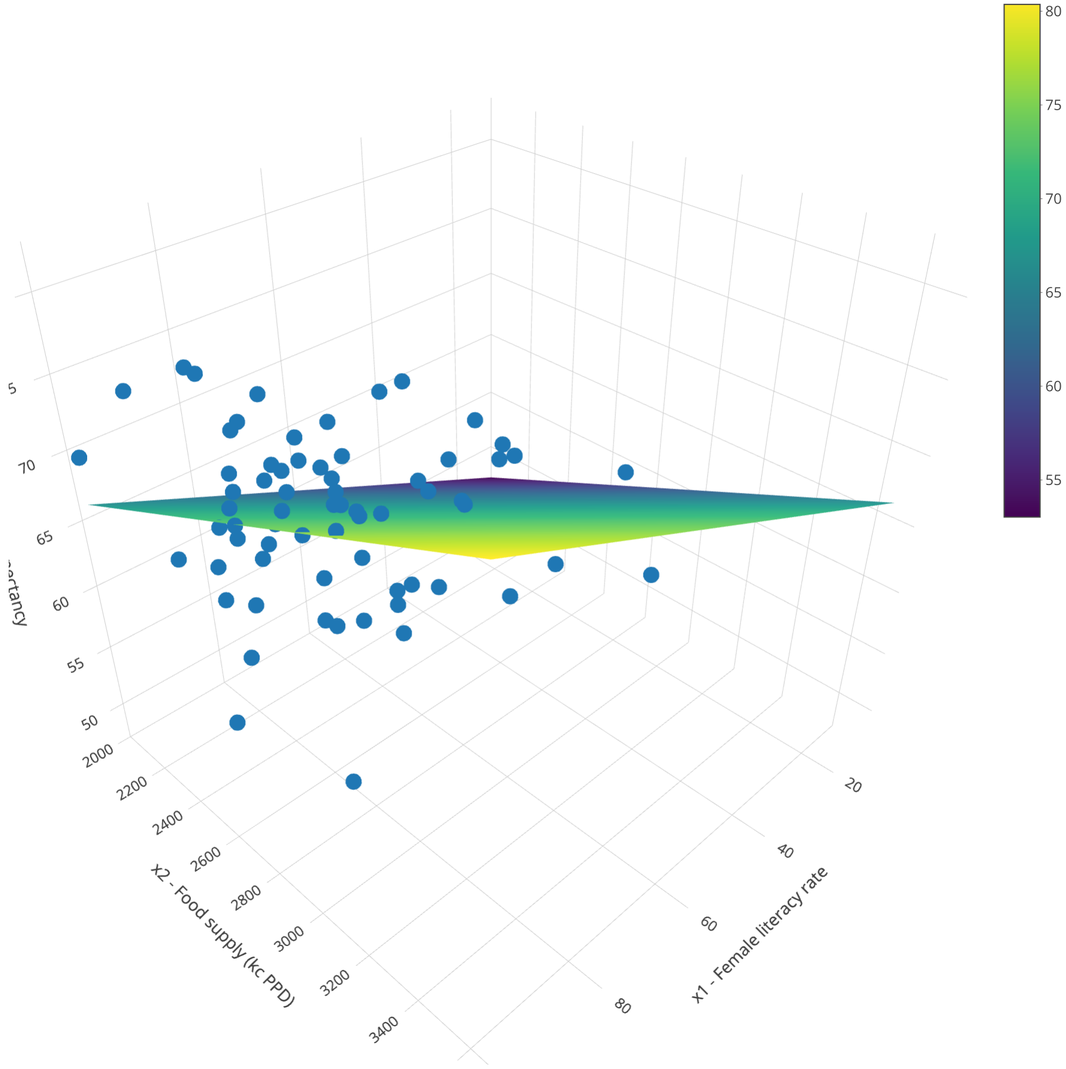
Visualize the fitted multiple regression model

- The fitted model equation

$$\widehat{Y} = \widehat{\beta}_0 + \widehat{\beta}_1 \cdot X_1 + \widehat{\beta}_2 \cdot X_2$$

  has three variables (Y, X_1, and X_2) and thus we need 3 dimensions to plot it

- Instead of a regression line, we get a **regression plane**
  - *See code in .qmd- file. I hid it from view in the html file.*



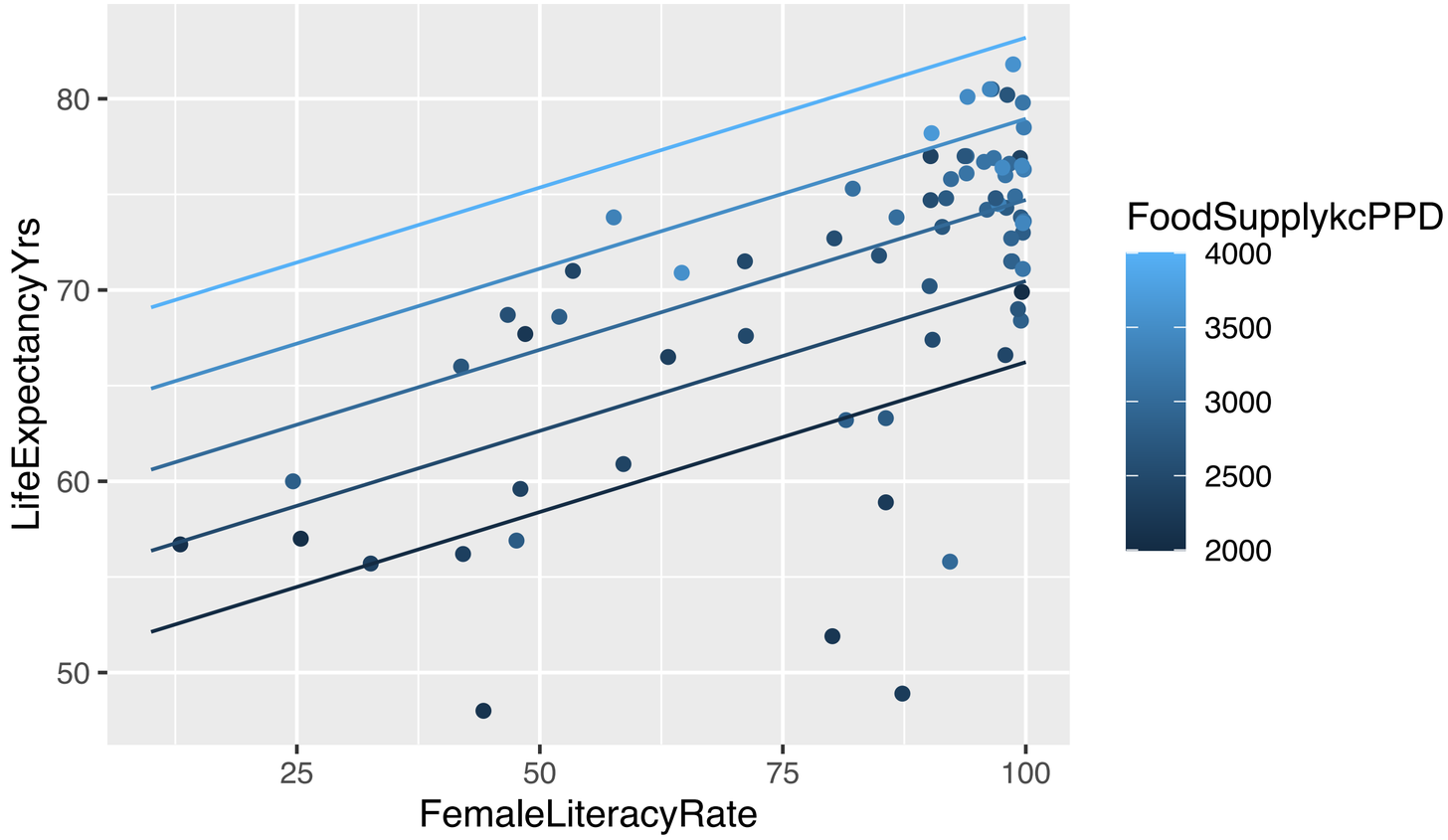
Regression lines for varying values of food supply

$$\widehat{\text{Life expectancy}} = \widehat{\beta}_0 + \widehat{\beta}_1 \text{Female literacy rate} + \widehat{\beta}_2 \text{Food supply}$$

$$\widehat{\text{Life expectancy}} = 33.595 + 0.157\ \text{Female literacy rate} + 0.008\ \text{Food supply}$$

- Note: when the food supply is held constant but the female literacy rate varies…
  - then the outcome values change along a **line**
- Different values of food supply give different lines
  - The intercepts change, but
  - the slopes stay the same (parallel lines)

```
1  (mr1_2d = ggPredict(mr1, interactive = T))
```



How do we calculate the regression line for 3000 kc PPD food supply?

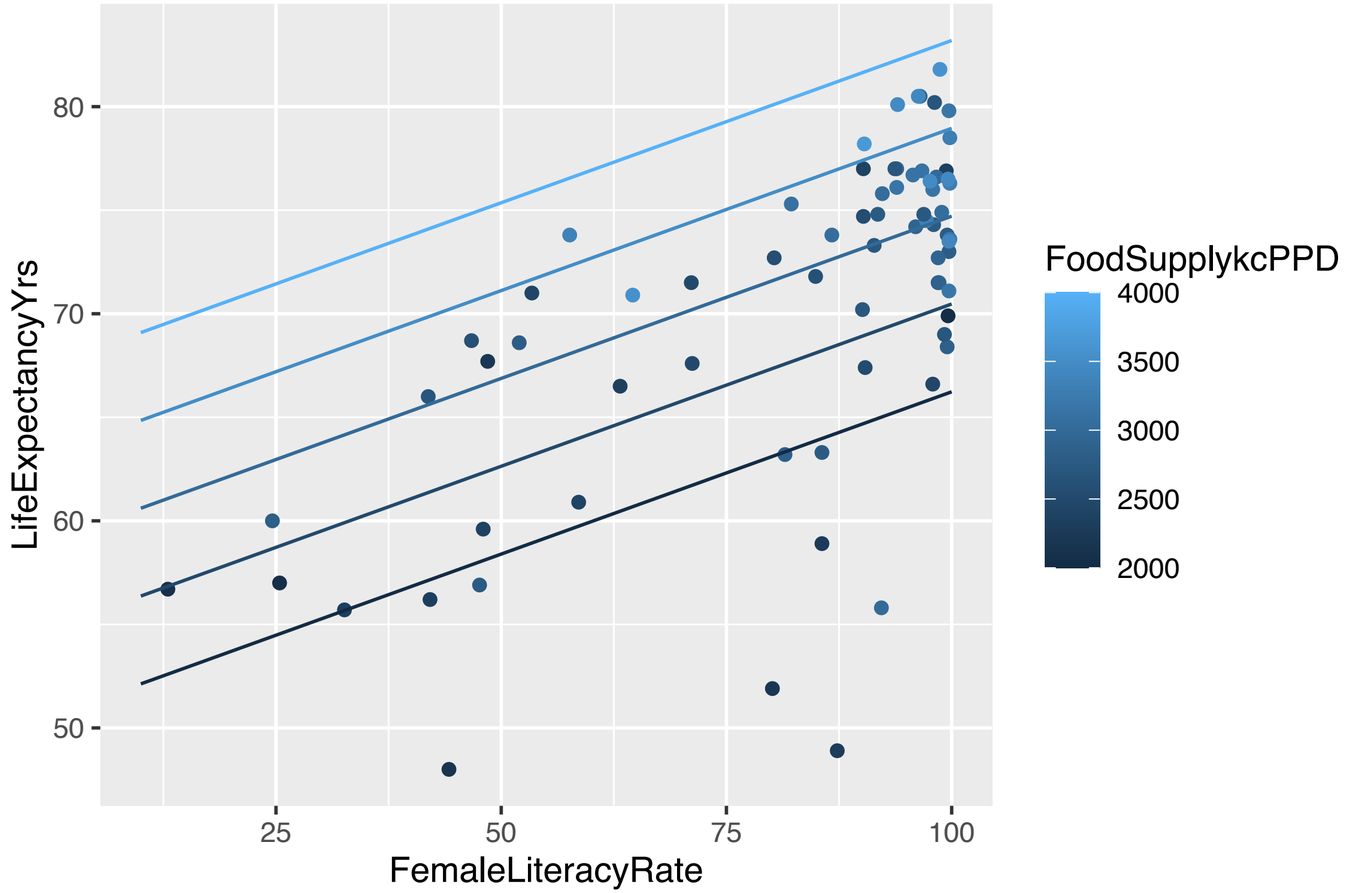
$$\widehat{\text{LE}} = 33.595 + 0.157\,\text{FLR} + 0.008\,\text{FS}$$

$$\widehat{\text{LE}} = 33.595 + 0.157\,\text{FLR} + 0.008 \cdot 3000$$

$$\widehat{\text{LE}} = 33.595 + 0.157\,\text{FLR} + 24$$

$$\widehat{\text{LE}} = 57.6 + 0.157\,\text{FLR}$$

```
1  (mr1_2d = ggPredict(mr1, interactive = T))
```



Poll Everwhere Question 2

Learning Objectives

1. Understand equations and visualizations that helps us build multiple linear regression model.
2. Fit MLR model (in R) and understand the difference between fitted regression plane and regression lines.
3. Identify the population multiple linear regression model and define statistics language for key notation.
4. Based off of previous SLR work, understand how the population MLR is estimated.
5. Interpret MLR (population) coefficient estimates with additional variable in model

Population multiple regression model

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

or on the individual (observation) level:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_k x_{ik} + \epsilon_i, \quad \text{for } i = 1, 2, \ldots, n$$

Observable sample data

- Y is our dependent variable
  - Aka outcome or response variable
- $X_1, X_2, \ldots, X_k$ are our k independent variables
  - Aka predictors or covariates

Unobservable population parameters

- $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are **unknown** population **parameters**
  - From our sample, we find the population parameter estimates: $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$
- ϵ is the random error
  - And is still normally distributed
  - $\epsilon \sim N(0, \sigma^2)$ where σ^2 is the population parameter of the variance

Learning Objectives

1. Understand equations and visualizations that helps us build multiple linear regression model.
2. Fit MLR model (in R) and understand the difference between fitted regression plane and regression lines.
3. Identify the population multiple linear regression model and define statistics language for key notation.
4. Based off of previous SLR work, understand how the population MLR is estimated.
5. Interpret MLR (population) coefficient estimates with additional variable in model

How do we estimate the model parameters?

- We need to estimate the population model coefficients $\widehat{\beta}_0, \widehat{\beta}_1, \widehat{\beta}_2, \ldots, \widehat{\beta}_k$
- This can be done using the **ordinary least-squares method**
  - Find the $\widehat{\beta}$ values that **minimize** the sum of squares due to error (SSE)

$$SSE = \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

$$SSE = \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$$

$$SSE = \sum_{i=1}^{n} (Y_i - (\widehat{\beta}_0 + \widehat{\beta}_1 X_{i1} + \ldots + \widehat{\beta}_1 X_{ik}))^2$$

$$SSE = \sum_{i=1}^{n} (Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_{i1} - \ldots - \widehat{\beta}_1 X_{ik})^2$$

Technical side note (not needed in our class)

- The equations for calculating the $\widehat{\beta}$ values is best done using matrix notation (not required for our class)
- We will be using R to get the coefficients instead of the equation (already did this a few slides back!)
- How we have represented the population regression model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \epsilon$$

- How to represent population model with matrix notation:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$
$$\boldsymbol{Y}_{n\times 1} = \boldsymbol{X}_{n\times(k+1)}\boldsymbol{\beta}_{(k+1)\times 1} + \boldsymbol{\epsilon}_{n\times 1}$$

$$\boldsymbol{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}_{n\times 1} \qquad \boldsymbol{X} = \begin{bmatrix} 1 & X_{11} & X_{12} & \ldots & X_{1,k} \\ 1 & X_{21} & X_{22} & \ldots & X_{2,k} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & X_{n1} & X_{n2} & \ldots & X_{n,k} \end{bmatrix}_{n\times(k+1)}$$

$$\boldsymbol{\epsilon} = \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}_{n\times 1} \qquad \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}_{(k+1)\times 1}$$

LINE model assumptions

[L] **Linearity** of relationship between variables

The mean value of Y given any combination of $X_1, X_2, \ldots, X_k$ values, is a linear function of $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$:

$$\mu_{Y|X_1,\ldots,X_k} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k$$

[I] **Independence** of the Y values

Observations $(X_1, X_2, \ldots, X_k, Y)$ are independent from one another

[N] **Normality** of the Y's given X (residuals)

Y has a normal distribution for any any combination of $X_1, X_2, \ldots, X_k$ values

- Thus, the residuals are normally distributed

[E] **Equality** of variance of the residuals (homoscedasticity)

The variance of Y is the same for any any combination of $X_1, X_2, \ldots, X_k$ values

$$\sigma^2_{Y|X_1,X_2,\ldots,X_k} = Var(Y|X_1, X_2, \ldots, X_k) = \sigma^2$$

Summary of the LINE assumptions

- Equivalently, the residuals are independently and identically distributed (iid):
  - normal
  - with mean 0 and
  - constant variance σ^2

Variation: Explained vs. Unexplained

$$\sum_{i=1}^{n}(Y_i - \bar{Y})^2 = \sum_{i=1}^{n}(\widehat{Y}_i - \bar{Y})^2 + \sum_{i=1}^{n}(Y_i - \widehat{Y}_i)^2$$

$$SSY = SSR + SSE$$

- $Y_i - \bar{Y}$ = the deviation of Y_i around the mean $\bar{Y}$
  - (the **total** amount deviation unexplained at $X_{i1}, \ldots, X_{ik}$).
- $Y_i - \widehat{Y}_i$ = the deviation of the observation Y around the fitted regression line
  - (the amount deviation **unexplained** by the regression at $X_{i1}, \ldots, X_{ik}$).
- $\widehat{Y}_i - \bar{Y}$ = the deviation of the fitted value $\widehat{Y}_i$ around the mean $\bar{Y}$
  - (the amount deviation **explained** by the regression at $X_{i1}, \ldots, X_{ik}$).

Poll Everywhere Question 3

Building the ANOVA table

ANOVA table (k = # of predictors, n = # of observations)

Variation Source	df	SS	MS	test statistic	p-value
Regression	k	SSR	$MSR = \frac{SSR}{k}$	$F = \frac{MSR}{MSE}$	$F \sim F_{(k,n-k-1)}$
Error	$n-k-1$	SSE	$MSE = \frac{SSE}{n-k-1}$		
Total	$n-1$	SSY			

```
1 anova(mr1) %>% tidy() %>% gt() %>%
2    tab_options(table.font.size = 40) %>% fmt_number(decimals = 3)
```

term	df	sumsq	meansq	statistic	p.value
FemaleLiteracyRate	1.000	1,934.245	1,934.245	66.547	0.000
FoodSupplykcPPD	1.000	649.319	649.319	22.339	0.000
Residuals	69.000	2,005.556	29.066	NA	NA

Learning Objectives

1. Understand equations and visualizations that helps us build multiple linear regression model.
2. Fit MLR model (in R) and understand the difference between fitted regression plane and regression lines.
3. Identify the population multiple linear regression model and define statistics language for key notation.
4. Based off of previous SLR work, understand how the population MLR is estimated.

5. Interpret MLR (population) coefficient estimates with additional variable in model

Interpreting the estimated population coefficients

- For a population model:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

  - Where X_1 and X_2 are continuous variables
  - No need to specify Y because it required to be continuous in linear regression

General interpretation for $\widehat{\beta}_0$

The expected Y-variable is ($\widehat{\beta}_0$ units) when the X_1-variable is 0 X_1-units and X_2-variable is 0 X_1-units (95% CI: LB, UB).

General interpretation for $\widehat{\beta}_1$

For every increase of 1 X_1-unit in the X_1-variable, adjusting/controlling for X_2-variable, there is an expected increase/decrease of $|\widehat{\beta}_1|$ units in the Y-variable (95%: LB, UB).

General interpretation for $\widehat{\beta}_2$

For every increase of 1 X_2-unit in the X_2-variable, adjusting/controlling for X_1-variable, there is an expected increase/decrease of $|\widehat{\beta}_2|$ units in the Y-variable (95%: LB, UB).

Poll Everywhere Question 4

Getting these interpretations from our regression table

We fit the regression model in R and printed the regression table:

```
1 mr1 <- lm(LifeExpectancyYrs ~ FemaleLiteracyRate + FoodSupplykcPPD,
2           data = gapm_sub)
```

term	estimate	std.error	statistic	p.value	conf.low	conf.high
(Intercept)	33.595	4.472	7.512	0.000	24.674	42.517
FemaleLiteracyRate	0.157	0.032	4.873	0.000	0.093	0.221
FoodSupplykcPPD	0.008	0.002	4.726	0.000	0.005	0.012

Fitted multiple regression model: $\widehat{\text{LE}} = 33.595 + 0.157\ \text{FLR} + 0.008\ \text{FS}$

Interpretation for $\widehat{\beta}_0$

The expected life expectancy is 33.595 years when the female literacy rate is 0% and food supply is 0 0 kcal PPD (95% CI: 24.674, 41.517).

Interpretation for $\widehat{\beta}_1$

For every 1% increase in the female literacy rate, adjusting for food supply, there is an expected increase of 0.157 years in the life expectancy (95%: 0.093, 0.221).

Interpretation for $\widehat{\beta}_2$

For every 1 kcal PPD increase in the food supply, adjusting for female literacy rate, there is an expected increase of 0.008 years in life expectancy (95%: 0.005, 0.012).

Let's just examine the general interpretation vs. the example

General interpretation for $\widehat{\beta}_0$

The expected Y-variable is ($\widehat{\beta}_0$ units) when the X_1-variable is 0 X_1-units and X_2-variable is 0 X_1-units (95% CI: LB, UB).

General interpretation for $\widehat{\beta}_1$

For every increase of 1 X_1-unit in the X_1-variable, adjusting/controlling for X_2-variable, there is an expected increase/decrease of $|\widehat{\beta}_1|$ units in the Y-variable (95%: LB, UB).

General interpretation for $\widehat{\beta}_2$

For every increase of 1 X_2-unit in the X_2-variable, adjusting/controlling for X_1-variable, there is an expected increase/decrease of $|\widehat{\beta}_2|$ units in the Y-variable (95%: LB, UB).

Interpretation for $\widehat{\beta}_0$

The expected life expectancy is 33.595 years when the female literacy rate is 0% and food supply is 0 0 kcal PPD (95% CI: 24.674, 41.517).

Interpretation for $\widehat{\beta}_1$

For every 1% increase in the female literacy rate, adjusting for food supply, there is an expected increase of 0.157 years in the life expectancy (95%: 0.093, 0.221).

Interpretation for $\widehat{\beta}_2$

For every 1 kcal PPD increase in the food supply, adjusting for female literacy rate, there is an expected increase of 0.008 years in life expectancy (95%: 0.005, 0.012).