icddr,b

Ordinary least square regression model for continuous outcomes

Dr. Rina Das

Nutrition and Clinical Services Division, icddr,b

Learning objectives

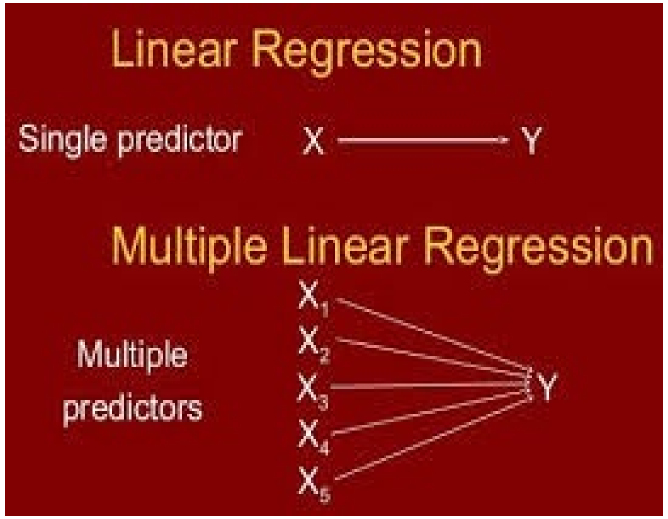
By the end of the session we will be able to:

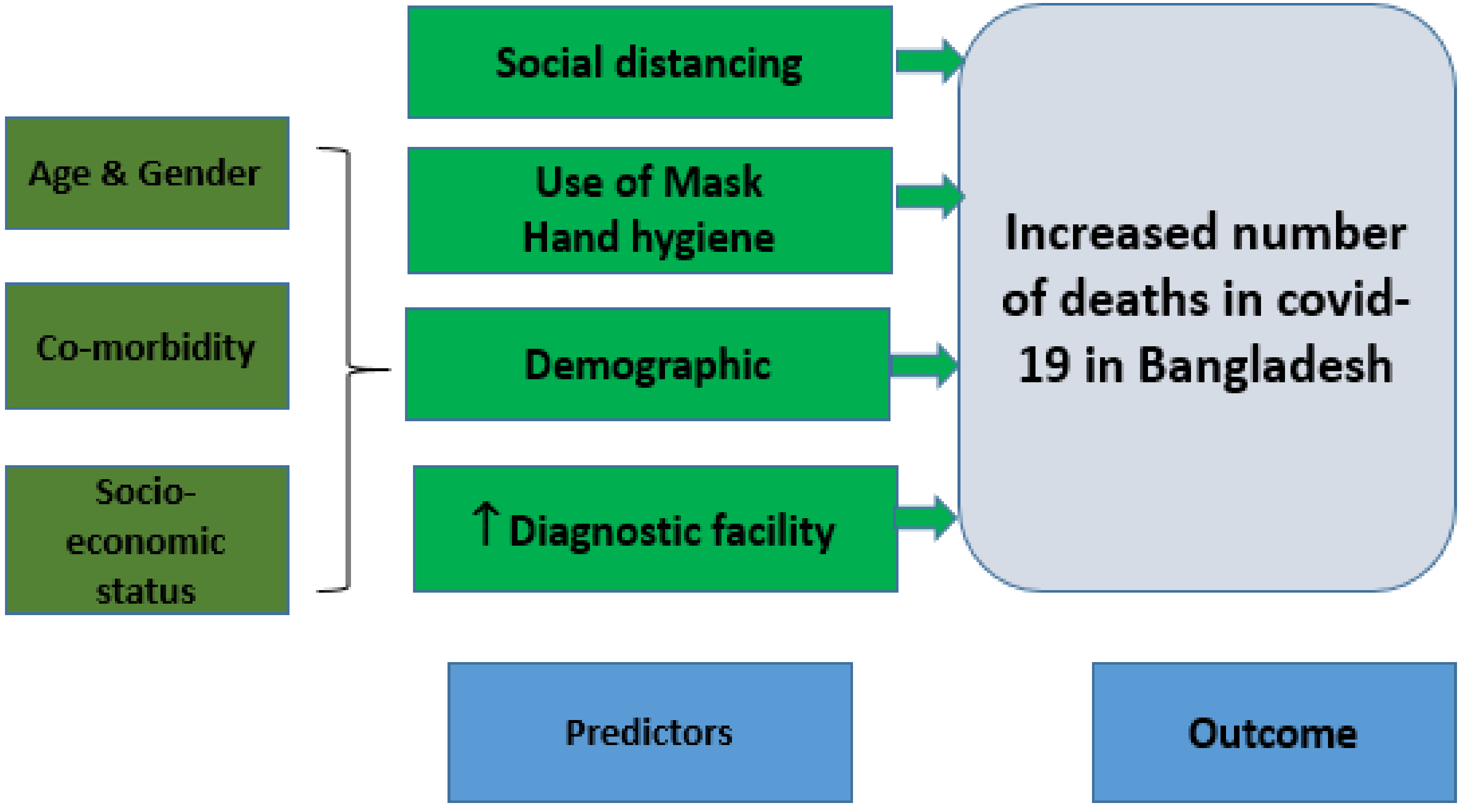
1. Build statistical model for investigating associations

- Identify important exposure variables for inclusion in a model
- Perform multiple linear regression analysis for epidemiological investigations

2. Interpret the findings and present them in a clear and concise manner

Why Regression?





Regression

- Regression analysis attempts to explain the variation in a dependent variable using the variation in independent variable(s)

- We use regression to estimate the unknown effect of changing one variable over another (Stock and Watson, 2003, ch. 4)

Purpose of regression

Estimation

- Estimate association between outcome and exposure adjusted for other covariates

Prediction

- Use an estimated model to predict the outcome of the given covariates in a new dataset

Difference ??

- **Correlation**

  Measures direction and strength of association

- **Linear regression**

  To quantify the magnitude of association when it is assumed that the linear relationship exists between dependent and independent variable

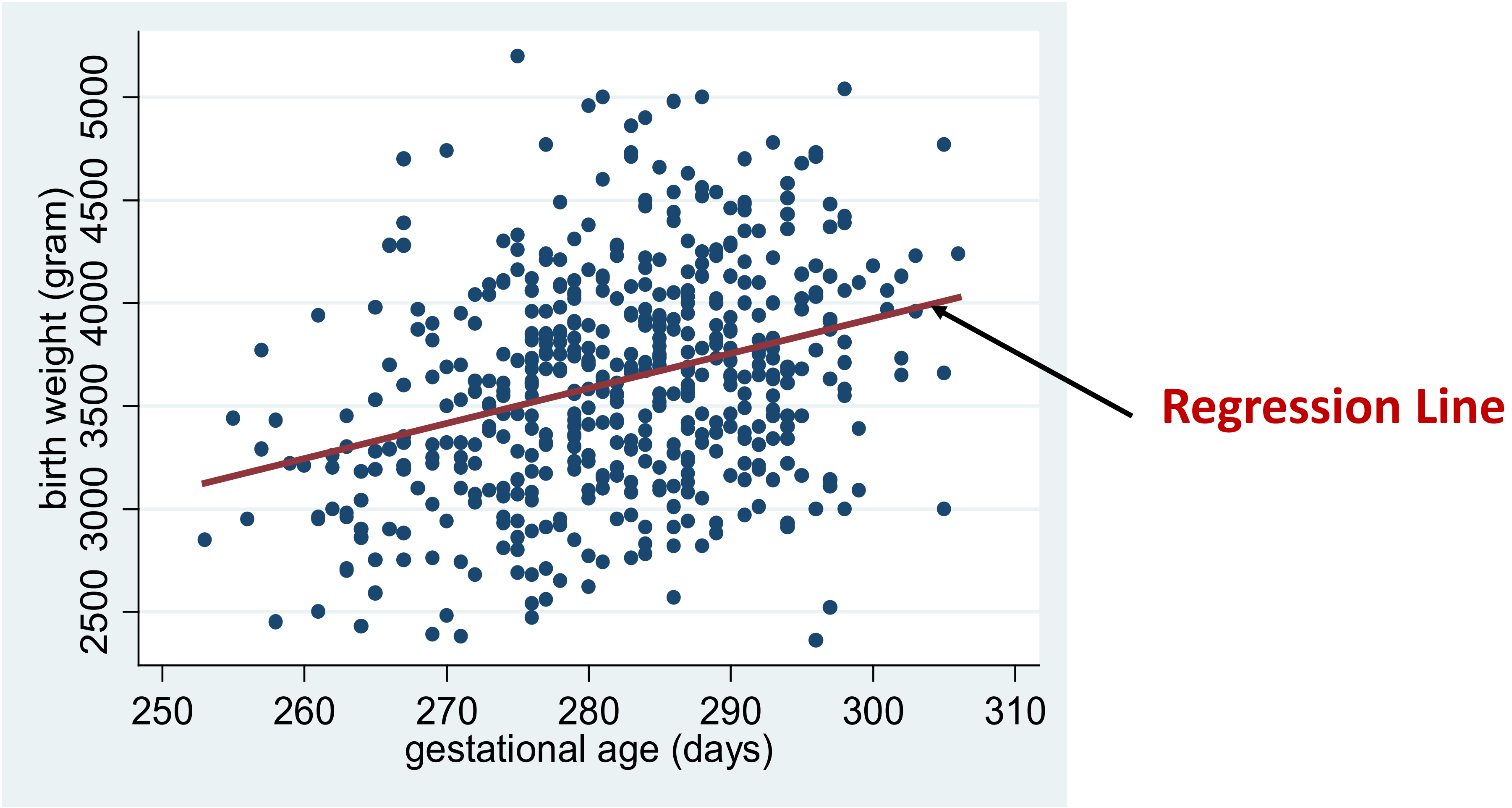
Variables

Independent variable:
It is a variable that stands alone and isn't changed by the other variables we are trying to measure. These are more or less *controlled*.

Dependent variable:
Dependent variables are not controlled but instead are simply *measured*. Dependent Variables depend on what the independent variable is

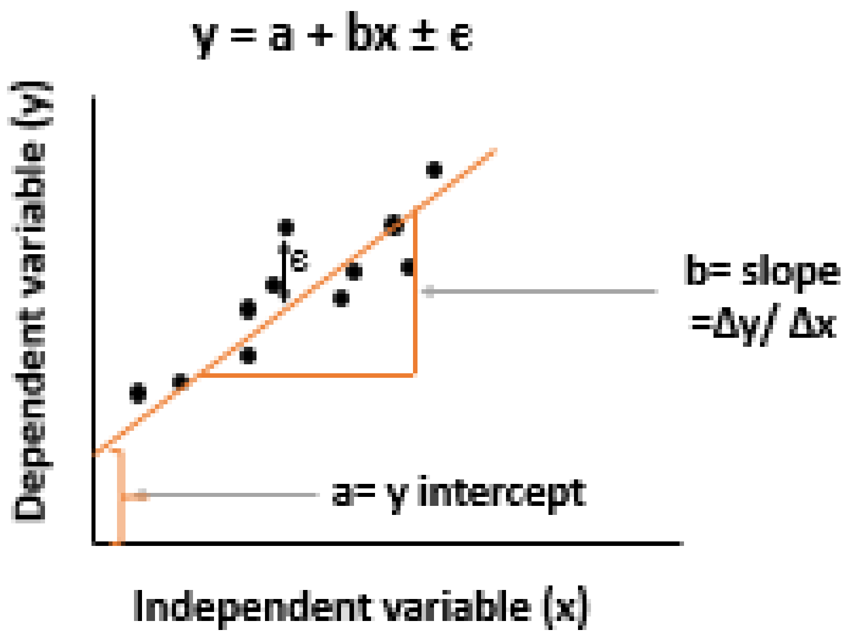
Independent variable(s) change the dependent variable

Regression



birth weight (gram)
2500
3000
3500
4000
4500
5000
250
260
270
280
290
300
310
gestational age (days)
Regression Line

Regression Model

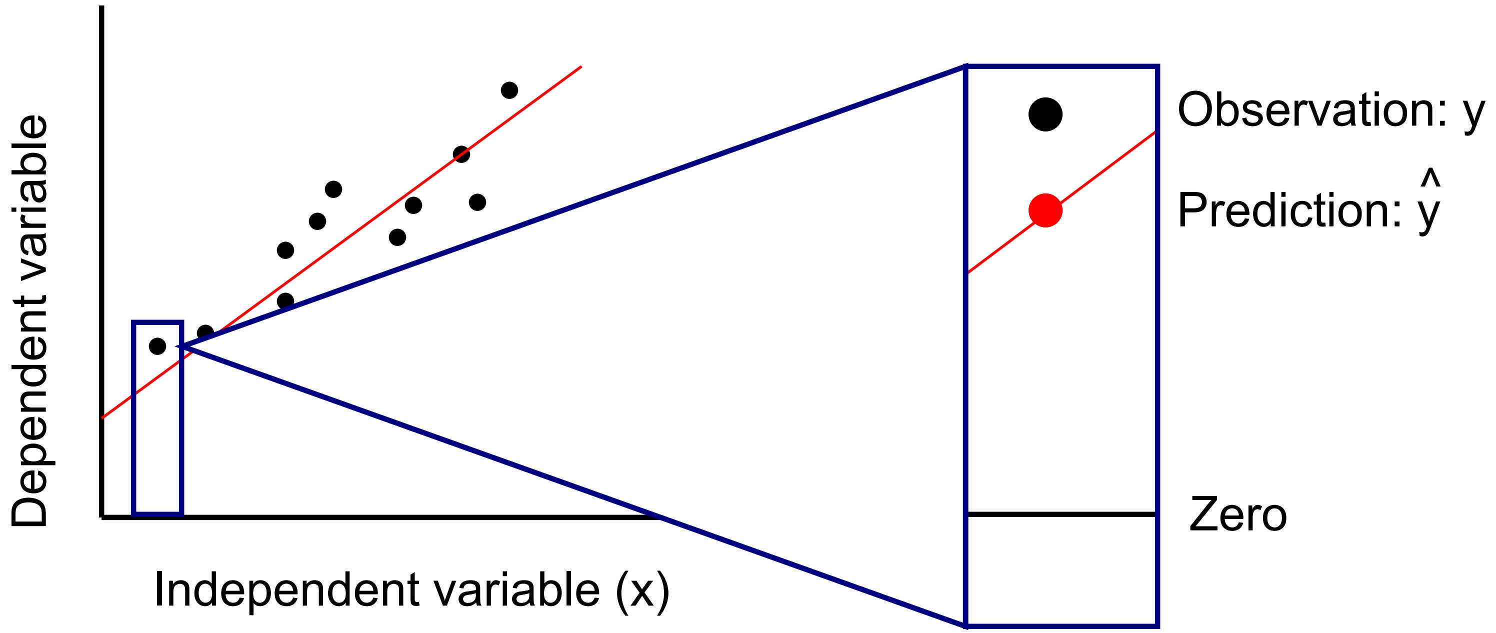


Simple regression fits a straight line to the data

Slope (b): Slope (b) represents the rate of change in dependent variable (y) for unit change in independent variable (x)

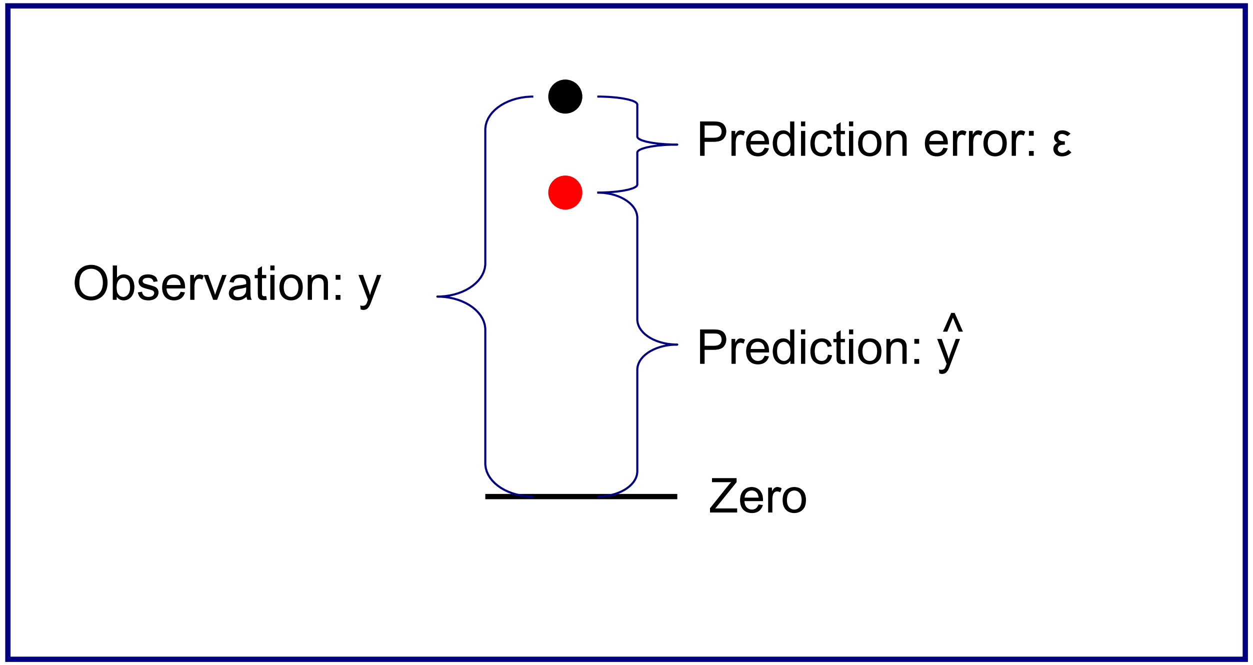
Intercept (a): Expected mean value of dependent variable (y) when there is no effect of independent variable on dependent variable (x=0)

$y = a + bx \pm \epsilon$



The function will make a prediction for each observed data point

The observation is denoted by y and the prediction is denoted by ŷ

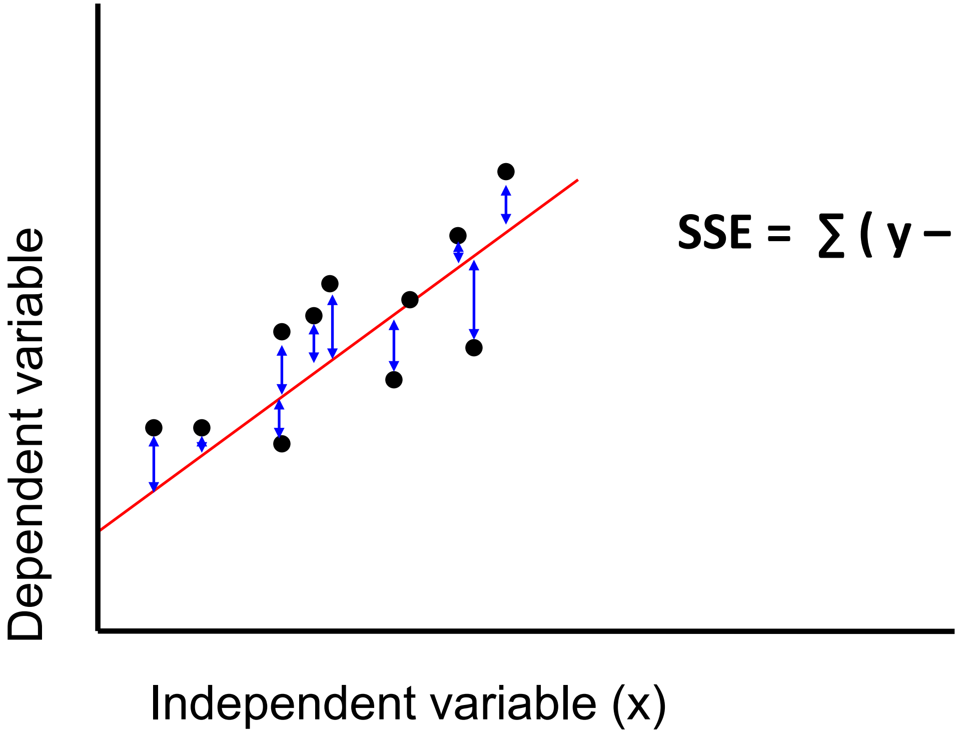


For each observation, the variation can be described as:

$$y = \hat{y} + \varepsilon$$

Actual = Explained + Error

Least Square Regression Line

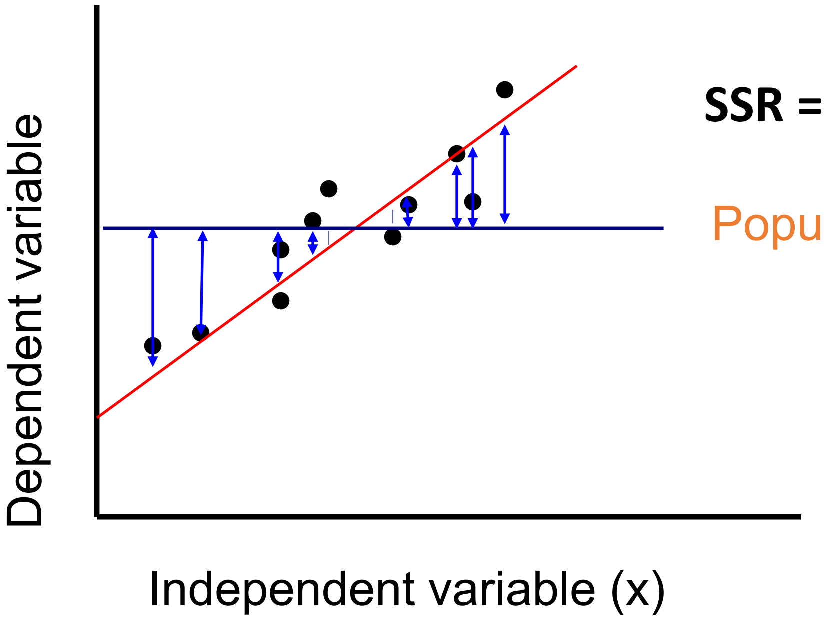


$$SSE = \sum (y - \hat{y})^2$$

A least squares regression selects the line with the lowest total sum of squared prediction errors

This value is called the **Sum of Squares of Error**, or SSE

Sum of Square Regression



$$SSR = \sum (\hat{y} - \bar{y})^2$$

Population mean: $\bar{y}$

The **Sum of Squares Regression** (SSR) is the sum of the squared differences between the prediction for each observation and the population mean

Coefficient of Determination, R^2

The proportion of total variation (SST) that is explained by the regression (SSR) is known as the Coefficient of Determination, and is often referred to as R^2

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$

measure of explained variation:

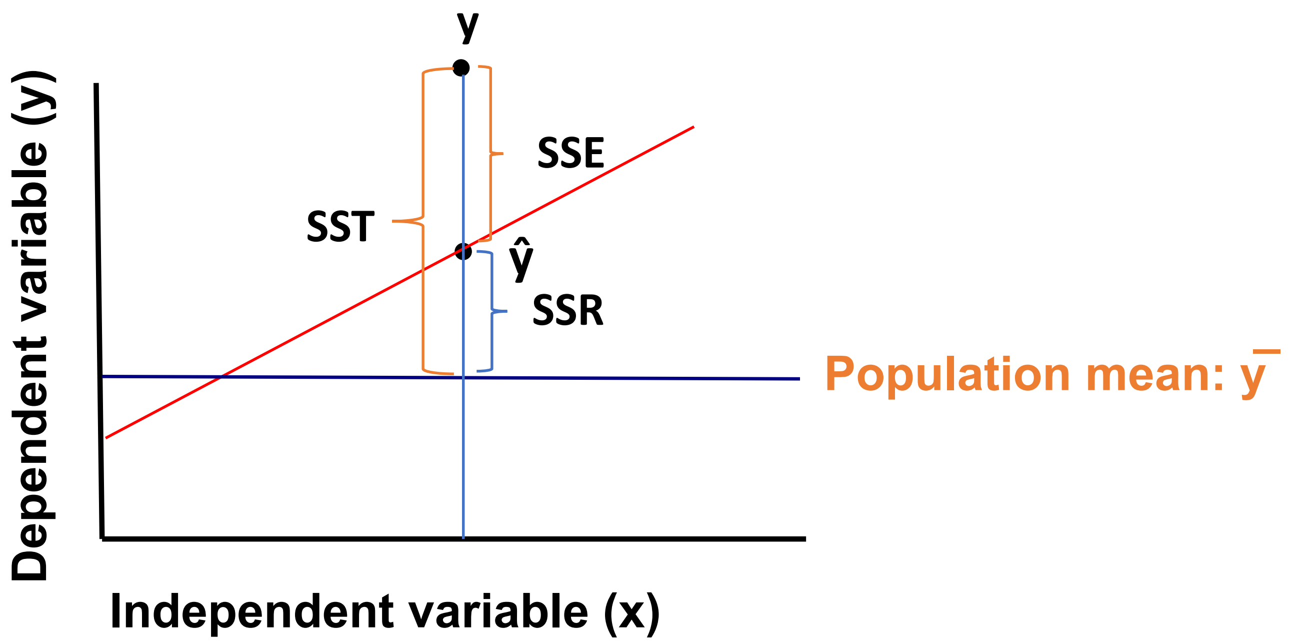
$SSR = \sum (\hat{y} - \bar{y})^2$

measure of unexplained variation:

$SSE = \sum (y - \hat{y})^2$

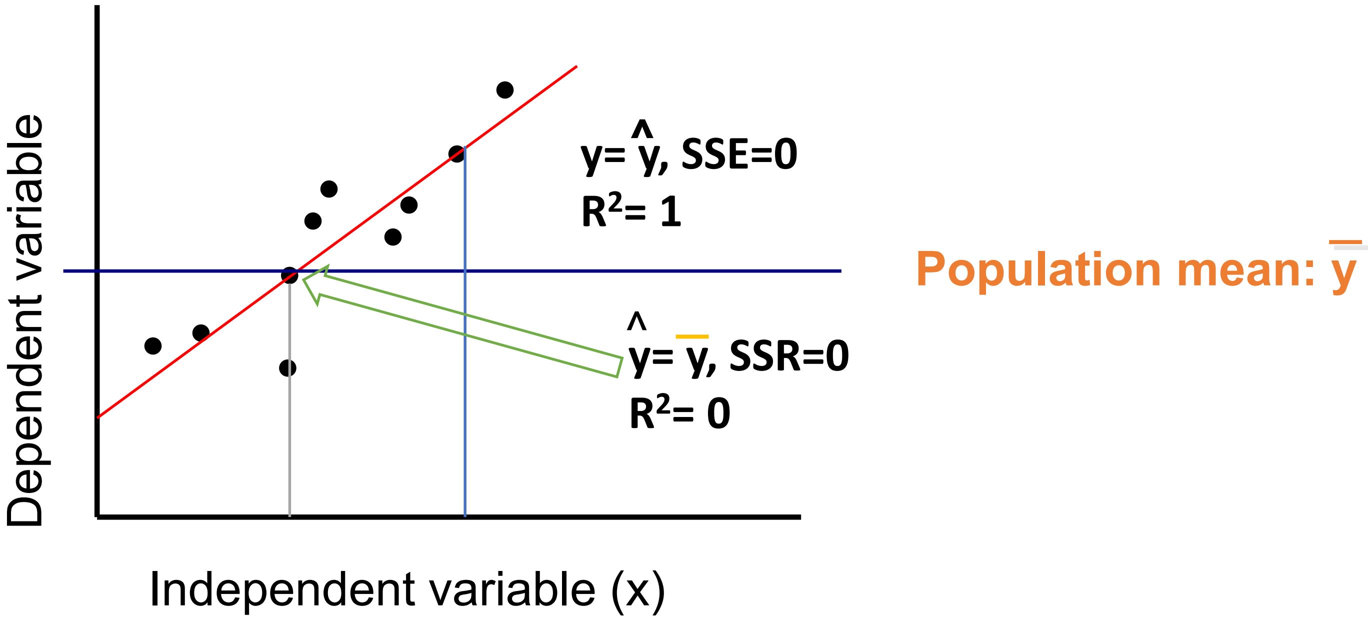
measure of total variation in y:

$SST = SSR + SSE$



Range of R^2

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR + SSE}$$



The value of R^2 can range between 0 and 1, and the higher its value the more accurate the regression model is. It is often expressed in percentage

Example

x	y	ŷ	ȳ	y-ŷ	(y-ŷ)²	ŷ-ȳ	(ŷ-ȳ)²
1	4	3	7	4-3=1	1	-4	16
2	6	4	7	2	4	-3	9
3	6	5	7	1	1	-2	4
4	7	6	7	1	1	-1	1
5	8	7	7	1	1	0	0
6	7	8	7	-1	1	1	1
7	8	9	7	-1	1	2	4
8	10	10	7	0	0	3	9
Total=					Σ(y-ŷ)²=10		Σ(ŷ-ȳ)²=44

$$R^2 = \frac{SSR}{SST} = \frac{SSR}{SSR+SSE} = \frac{44}{44+10} = 0.8148 = 81.48\%$$

Multiple Regression

- Multiple regression forms a 'linear combination' of <u>multiple variables</u> to best predict an outcome

- Then assess the contribution of each predictor variable to the equation, separately and/or combinedly

When to use
Multiple Linear Regression

1 (One) Dependent variable

[Continuous]

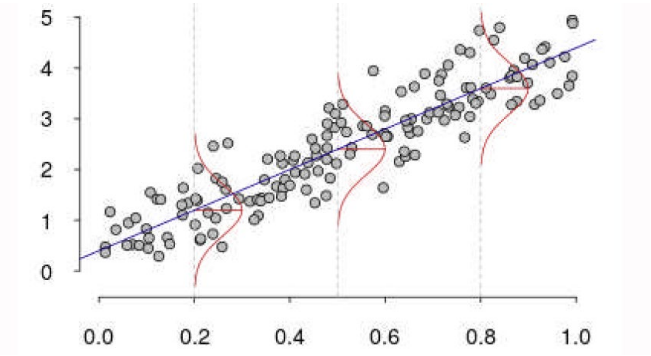
vs.

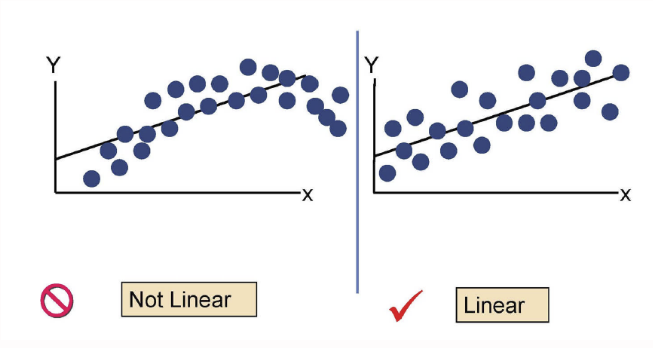
Several Independent variables

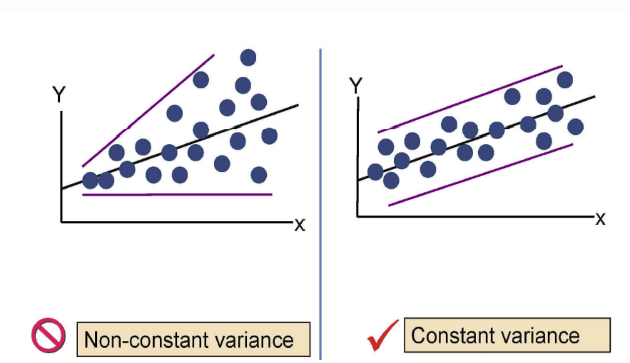
[Continuous and/or Categorical]

Assumption of linear regression

- Data normally distributed: outcome variable is distributed normally at each value of predictor variable
- Have linear relationship between predictor and outcome variables
- The variance of outcome variable at every value of predictor variable is the same
- The residuals are independent







Overview: statistical modelling in linear regression

Conceptual framework

1. Literature Review
2. Identification of the
-Exposure variables
- Confounders
-Effect modifiers

Preliminary analysis

1. Data exploration and data cleaning
2. Classification of the variables
3. Data reduction (categorization)

Univariate and bivariate analysis

1. Tables and visual plots
2. Cross tabulations
3. Unadjusted measure of effect (crude)
3. Correlations
4. Simple regression

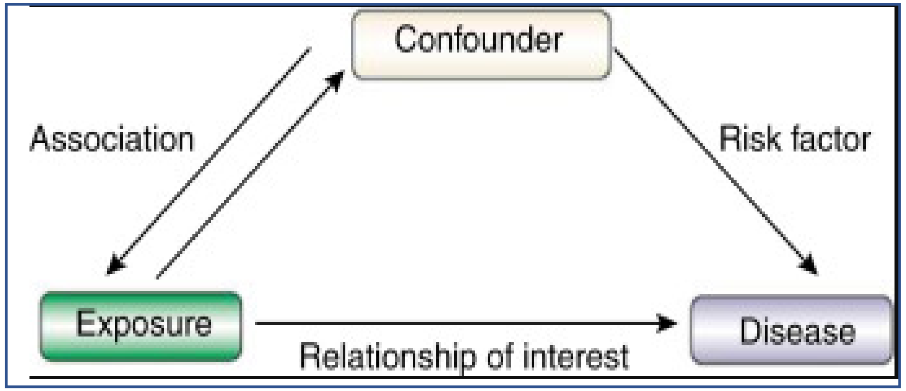
Multivariable analysis

1. Multiple linear regression
2. Adjusted for confounders
3. Investigate interaction
4. Post estimation diagnostics

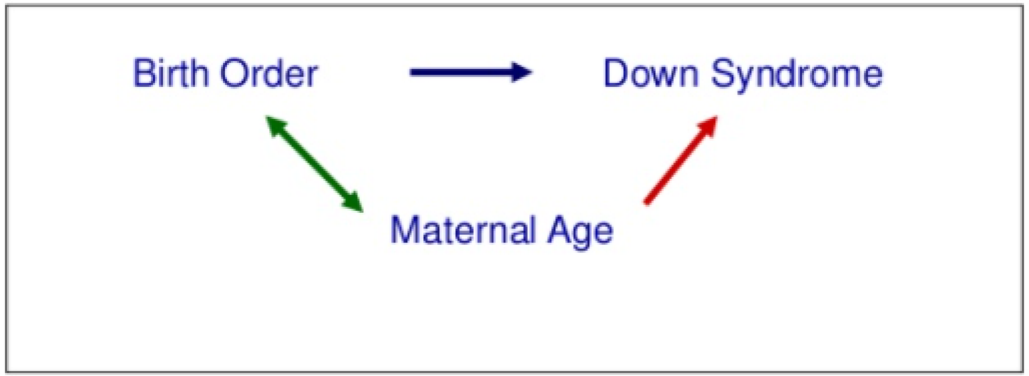
Confounder

Confounding happens when another variable is:

- Associated with the outcome
- Associated with the explanatory variable of interest
- Not on the causal pathway



Simultaneous adjustment for all exposures (if multiple exposures)



Maternal age is correlated with birth order and a risk factor even if birth order is low

There is no statistical test for confounding

https://www.sciencedirect.com/science/article/pii/S0085253815529748: Confounding: What it is and how to deal with it

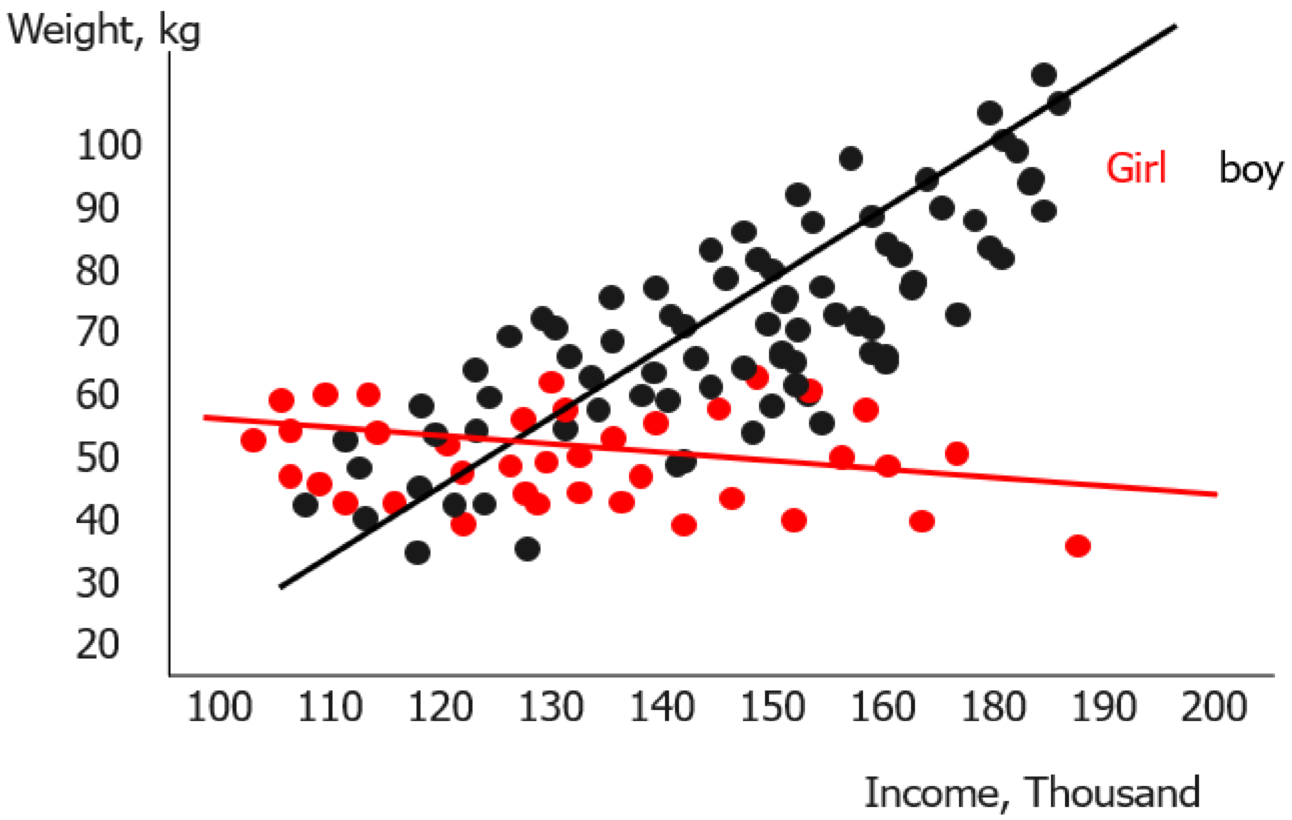
Control the confounder

- Adding a confounding variable as a predictor in the linear model allow for controlling the confounder

- Adjusting for confounding in this way we assume that:
  - ✓ The confounder has been measured perfectly
  - ✓ The association between confounder and outcome is perfectly linear

- Confounding is a property of the real-world system we are attempting to model
- Requires a thorough knowledge of the subject area

Interaction / effect modification



- Mediating factor– should not be adjusted/controlled
- Multivariable methods itself can be used to assess effect modification
- Need to report the effect modifier

Example of Multiple Linear Regression

- **Data Source:** BDHS data, 2018
- **Objective of the analysis:** To identify the factors associated with **height-for-age z score** among under 5 Bangladeshi children
- **Dependent variable:** Height-for-age z score (HAZ)

- **Independent/exposure variables:** ?????

BDHS data, 2018

	caseid	v003	v005	v008	v011	v012	v021	v024	v025	v106	v113	v116	v130	v1:
1	1 17 4	4	664194	1415	1194	18	1	barisal	rural	higher	tube ...	pit l...	islam	
2	1 48 2	2	664194	1415	989	35	1	barisal	rural	primary	tube ...	pit l...	islam	
3	1 52 4	4	664194	1415	1152	21	1	barisal	rural	secon...	not a...	not a...	islam	
4	1 52 4	4	664194	1415	1152	21	1	barisal	rural	secon...	not a...	not a...	islam	
5	1 61 2	2	664194	1415	1035	31	1	barisal	rural	primary	tube ...	hangi...	islam	
6	1 69 2	2	664194	1415	1143	22	1	barisal	rural	secon...	tube ...	pit l...	islam	
7	1 74 2	2	664194	1415	1138	23	1	barisal	rural	higher	tube ...	pit l...	islam	
8	1104 1	1	664194	1415	934	40	1	barisal	rural	higher	tube ...	venti...	buddhism	
9	1117 2	2	664194	1415	1065	29	1	barisal	rural	secon...	tube ...	pit l...	islam	
10	1131 2	2	664194	1415	1162	21	1	barisal	rural	primary	tube ...	pit l...	islam	
11	2 7 2	2	677993	1415	971	37	2	barisal	rural	primary	tube ...	venti...	islam	
12	2 25 2	2	677993	1415	1126	24	2	barisal	rural	secon...	tube ...	pit l...	islam	
13	2 32 2	2	677993	1415	1043	31	2	barisal	rural	higher	tube ...	venti...	islam	
14	2 43 3	3	677993	1415	1114	25	2	barisal	rural	secon...	not a...	not a...	islam	
15	2 43 3	3	677993	1415	1114	25	2	barisal	rural	secon...	not a...	not a...	islam	
16	2 54 2	2	677993	1415	1085	27	2	barisal	rural	secon...	tube ...	pit l...	islam	
17	2 61 4	4	677993	1415	1076	28	2	barisal	rural	secon...	tube ...	pit l...	islam	
18	2 65 2	2	677993	1415	1063	29	2	barisal	rural	primary	tube ...	pit l...	islam	
19	2 72 2	2	677993	1415	1122	24	2	barisal	rural	secon...	tube ...	pit l...	islam	
20	2 76 2	2	677993	1415	1120	24	2	barisal	rural	primary	tube ...	pit l...	islam	
21	2 76 2	2	677993	1415	1120	24	2	barisal	rural	primary	tube ...	pit l...	islam	
22	2 79 2	2	677993	1415	1199	18	2	barisal	rural	primary	tube ...	pit l...	islam	
23	2 87 2	2	677993	1415	1034	31	2	barisal	rural	primary	tube ...	hangi...	islam	
24	2 94 2	2	677993	1415	1127	24	2	barisal	rural	primary	tube ...	pit l...	islam	
25	2 97 2	2	677993	1415	1106	25	2	barisal	rural	primary	tube ...	pit l...	islam	
26	3 7 4	4	726617	1416	938	39	3	barisal	rural	no ed...	not a...	not a...	islam	

Variables

Filter variables here

Name	Label
edu	Education
religion	Mother's religion
PartnerEdu	Partner's Education
deci_healt	Woman's own heal...
deci_purch	Making major hou...
deci_visit	Visits to her family...
all_dec	All three decisions
none_dec	None of the three ...
toilet	type of toilet facility
contraceptive	Use contraceptive ...
Beating	Domestic violence

Variables Snapshots

Properties

Variables	
Name	caseid
Label	case identification
Type	str15
Format	%15s
Value label	
Notes	
Data	
Filename	DataBDHS_2018.dta
Label	
Notes	
Variables	79
Observations	8,759
Size	1.68M
Memory	64M
Sorted by	

Statistical modelling

Conceptual framework

1. Literature Review
2. Identification of the
-Exposure variables
- Confounders
-Effect modifiers

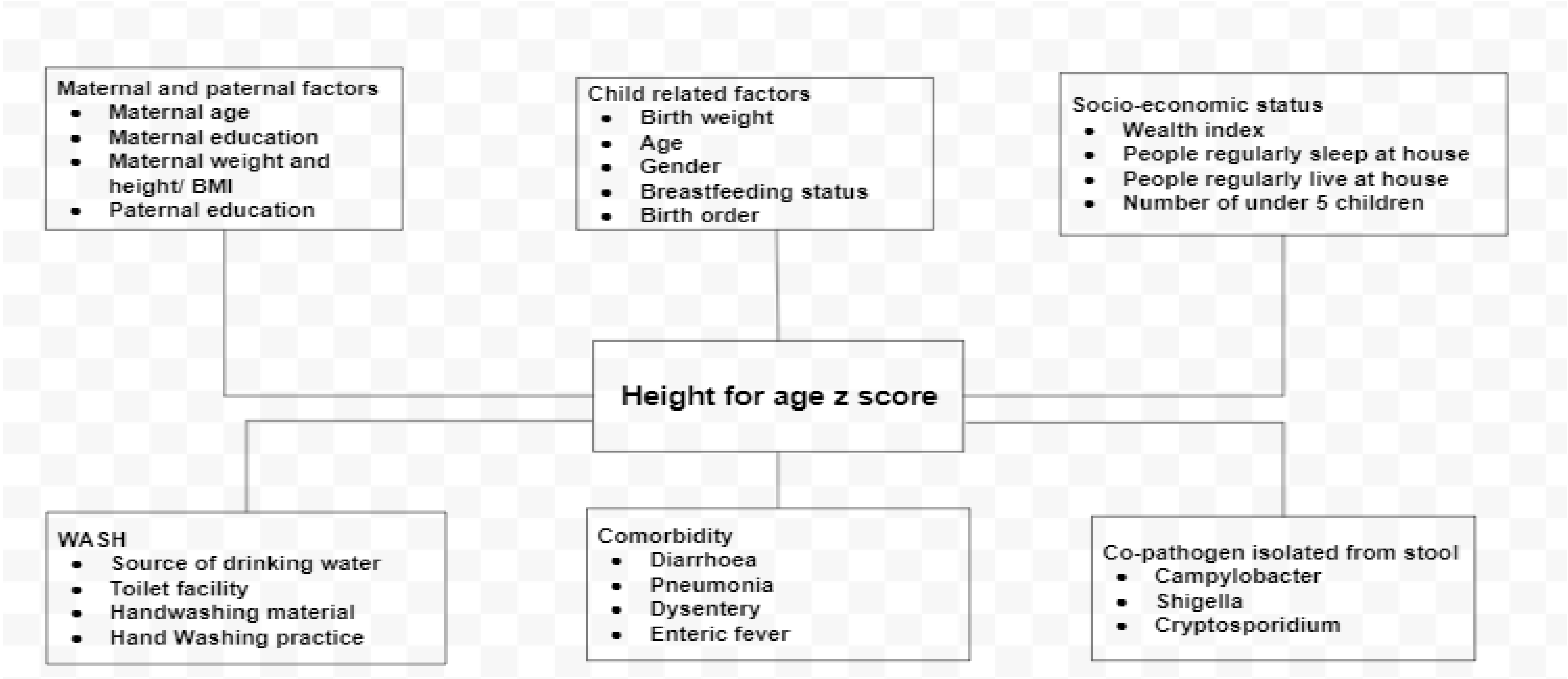
Preliminary analysis

1. Data exploration and data cleaning
2. Classification of the variables
3. Data reduction (categorization)

Univariate and bivariate analysis

1. Tables and visual plots
2. Cross tabulations
3. Correlations
4. Simple regression

Conceptual framework



Use of linear regression analysis

Example research question

Average maternal BMI in the urban area is better than rural area in Bangladesh, 2021

```
. tabstat bmi, by(v025)

Summary for variables: bmi
     by categories of: v025 (type of place of residence)
```

v025	mean
urban	23.64351
rural	22.13953
Total	22.66033

Use of linear regression analysis

Example research question

Average maternal BMI in the urban area is better than rural area in Bangladesh, 2021

```
. tabstat bmi, by(v025)

Summary for variables: bmi
     by categories of: v025 (type of place of residence)

    v025 |      mean
---------+----------
   urban |  23.64351
   rural |  22.13953
---------+----------
   Total |  22.66033
--------------------
```

Deference??

Use of linear regression analysis

Example research question

Average maternal BMI in the urban area is better than rural area in Bangladesh, 2021

```
. tabstat bmi, by(v025)

Summary for variables: bmi
     by categories of: v025 (type of place of residence)

    v025 |      mean
---------+----------
   urban |  23.64351
   rural |  22.13953
---------+----------
   Total |  22.66033
--------------------
```

Simply, we can do the ***t-test*** between two groups

Stata/SE 14.1 - F:\Dropbox (Personal)\WorkShop\Data_BDHS\BHS_RinaApa.dta

```
. ttest bmi, by( v025 )

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   urban |   2,986    23.64351    .0803364    4.389929    23.48599    23.80103
   rural |   5,637    22.13953    .0501222    3.763176    22.04127    22.23779
---------+--------------------------------------------------------------------
combined |   8,623    22.66033    .0436652    4.054759    22.57474    22.74593
---------+--------------------------------------------------------------------
    diff |            1.503988    .0903398                1.326901    1.681076
------------------------------------------------------------------------------
    diff = mean(urban) - mean(rural)                              t =  16.6481
Ho: diff = 0                                     degrees of freedom =     8621

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Variables: caseid, v012, v024, v025, v190, hw70, hw71, hw72, hw73, work, Mdelivery, Antevisit, anc2, Diarrhea, fever

F:\Dropbox (Personal)\WorkShop\Data_BDHS

Stata/SE 14.1 - F:\Dropbox (Personal)\WorkShop\Data_BDHS\BHS_RinaApa.dta

```
. ttest bmi, by( v025 )

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
   urban |   2,986    23.64351    .0803364    4.389929    23.48599    23.80103
   rural |   5,637    22.13953    .0501222    3.763176    22.04127    22.23779
---------+--------------------------------------------------------------------
combined |   8,623    22.66033    .0436652    4.054759    22.57474    22.74593
---------+--------------------------------------------------------------------
    diff |            1.503988    .0903398                1.326901    1.681076
------------------------------------------------------------------------------
    diff = mean(urban) - mean(rural)                              t =  16.6481
Ho: diff = 0                                     degrees of freedom =     8621

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 1.0000         Pr(|T| > |t|) = 0.0000          Pr(T > t) = 0.0000
```

Command

F:\Dropbox (Personal)\WorkShop\Data_BDHS

Stata/SE 14.1 - F:\Dropbox (Personal)\WorkShop\Data_BDHS\BHS_RinaApa.dta

File Edit Data Graphics Statistics User Window Help

```
       _cons     25.1475   .1554567   161.77   0.000     24.84277    25.45223

. reg bmi ib2.v025

      Source |       SS           df       MS      Number of obs   =     8,623
-------------+----------------------------------   F(1, 8621)      =    277.16
       Model |  4415.38245         1  4415.38245   Prob > F        =    0.0000
    Residual |  137339.519     8,621  15.9308107   R-squared       =    0.0311
-------------+----------------------------------   Adj R-squared   =    0.0310
       Total |  141754.901     8,622  16.4410695   Root MSE        =    3.9913

------------------------------------------------------------------------------
         bmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        v025 |
      urban  |   1.503988   .0903398    16.65   0.000     1.326901    1.681076
       _cons |   22.13953   .0531612   416.46   0.000     22.03532    22.24373
------------------------------------------------------------------------------
```

Command

Variables: caseid, v012, v024, v025, v190, hw70, hw71, hw72, hw73, work, Mdelivery, Antevisit, anc2, Diarrhea, fever

Review: 1 use ..., 2 do ... 1, 3 ttes... 1, 4 use ... 4, 5 clear, 6 use ..., 7 kee..., 8 do ... 1, 9 kee..., 1 sav..., 1 ttes..., 1 ttes..., 1 tab..., 1 ttes..., 1 reg ..., 1 reg ...

F:\Dropbox (Personal)\WorkShop\Data_BDHS

Stata/SE 14.1 - F:\Dropbox (Personal)\WorkShop\Data_BDHS\BHS_RinaApa.dta

File Edit Data Graphics Statistics User Window Help

Since t-test and regression give the same results, then why will go to regression analysis instead of t-test

```
       _cons |   25.1475   .1554567   161.77   0.000     24.84277    25.45223

. reg
                                                            ,623
                                                            7.16
       Model |  4415.38245         1  4415.38245   Prob > F        =    0.0000
    Residual |  137339.519     8,621  15.9308107   R-squared       =    0.0311
-------------+----------------------------------   Adj R-squared   =    0.0310
       Total |  141754.901     8,622  16.4410695   Root MSE        =    3.9913

------------------------------------------------------------------------------
         bmi |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        v025 |
      urban  |   1.503988   .0903398    16.65   0.000     1.326901    1.681076
       _cons |   22.13953   .0531612   416.46   0.000     22.03532    22.24373
------------------------------------------------------------------------------
```

Command

F:\Dropbox (Personal)\WorkShop\Data_BDHS

Stata/SE 14.1 - F:\Dropbox (Personal)\WorkShop\Data_BDHS\BHS_RinaApa.dta

Since t-test and regression give the same results, then why will go to regression analysis instead of t-test

We can adjust other covariates

We can adjust cluster and sampling weight if needed

We can adjust time for longitudinal data

```
       _cons |   25.1475   .1554567   161.77   0.000    24.84277    25.45223

. reg

       Model |  4415.38245         1  4415.38245   Prob > F        =    0.0000
    Residual |  137339.519     8,621  15.9308107   R-squared       =    0.0311

        v025
       urban |  1.503988   .0903398    16.65   0.000    1.326901    1.681076
       _cons |  22.13953   .0531612   416.46   0.000    22.03532    22.24373
```

Variables: caseid, v012, v024, v025, v190, hw70, hw71, hw72, hw73, work, Mdelivery, Antevisit, anc2, Diarrhea, fever

F:\Dropbox (Personal)\WorkShop\Data_BDHS

Since t-test and regression give the same results, then why will go to regression analysis instead of t-test

We can adjust other covariates
We can adjust cluster and sampling weight if needed
We can adjust time for longitudinal data

This will not possible in t-test analysis, then we can use regression analysis

Data Exploration

```
. sum hw70

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+---------------------------------------------------------
        hw70 |      7,849   -1.382771    1.320636      -5.86       5.88
```

```
. tab b4

     sex of |
      child |      Freq.     Percent        Cum.
------------+-----------------------------------
       male |      4,567       52.14       52.14
     female |      4,192       47.86      100.00
------------+-----------------------------------
      Total |      8,759      100.00
```

```
. sum hw70, detail

          height/age standard deviation (new who)
-------------------------------------------------------------
      Percentiles      Smallest
 1%        -4.48          -5.86
 5%        -3.41          -5.86
10%        -2.94          -5.85       Obs               7,849
25%        -2.23          -5.85       Sum of Wgt.       7,849

50%        -1.45                      Mean          -1.382771
                        Largest       Std. Dev.      1.320636
75%         -.59           5.59
90%          .21           5.71       Variance       1.744079
95%          .73           5.87       Skewness       .4471885
99%          2.2           5.88       Kurtosis       4.754577
```

Grouping

✓Meaningful cut-off points
✓Standard grouping

```
. sum hw70
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hw70	7,849	-1.382771	1.320636	-5.86	5.88

```
. sum hw72
```

Variable	Obs	Mean	Std. Dev.	Min	Max
hw72	7,831	-.5411365	1.148866	-4.99	4.99

Stunting: HAZ < – 2,
(< 5 years of age)

```
. tab stunting
```

stunting	Freq.	Percent	Cum.
non-stunted	5,388	68.65	68.65
stunted	2,461	31.35	100.00
Total	7,849	100.00	

```
. tab wasting
```

wasting	Freq.	Percent	Cum.
non-wasted	7,164	91.48	91.48
wasted	667	8.52	100.00
Total	7,831	100.00	

Categorization

type of toilet facility	Freq.	Percent	Cum.
flush to piped sewer system	180	2.06	2.06
flush to septic tank	1,424	16.26	18.31
flush to pit latrine	276	3.15	21.46
flush to somewhere else	301	3.44	24.90
flush, don't know where	198	2.26	27.16
ventilated improved pit latrine (vip)	1,040	11.87	39.03
pit latrine with slab	1,855	21.18	60.21
pit latrine without slab/open pit	2,319	26.48	86.69
no facility/bush/field	75	0.86	87.54
hanging toilet/latrine	75	0.86	88.40
not a dejure resident	1,016	11.60	100.00
Total	8,759	100.00	

```
. tab toilet
```

type of toilet facility	Freq.	Percent	Cum.
improved	4,775	57.81	57.81
unimproved	3,485	42.19	100.00
Total	8,260	100.00	

highest educational level	Freq.	Percent	Cum.
no education	642	7.33	7.33
primary	2,548	29.09	36.42
secondary	4,115	46.98	83.40
higher	1,454	16.60	100.00
Total	8,759	100.00	

Education	Freq.	Percent	Cum.
below secondary	3,190	36.42	36.42
Secondary and above	5,569	63.58	100.00
Total	8,759	100.00	

Univariate analysis

- **Categorical variables** – Frequency tables
- **Quantitative variables** – Descriptive statistics
  - ✓ Central tendency (mean, median)
  - ✓ Dispersion (min, max, range, IQR, SD, coefficient of variation)
  - ✓ Shape of distribution (skewness, kurtosis)
- **Visualize data using plots**
  - ✓ Histograms (normally distributed/ not), bar charts, pie charts

Univariate analysis

Categorical variable

```
. tabulate b4
```

sex of child	Freq.	Percent	Cum.
male	4,567	52.14	52.14
female	4,192	47.86	100.00
Total	8,759	100.00	

Continuous variable

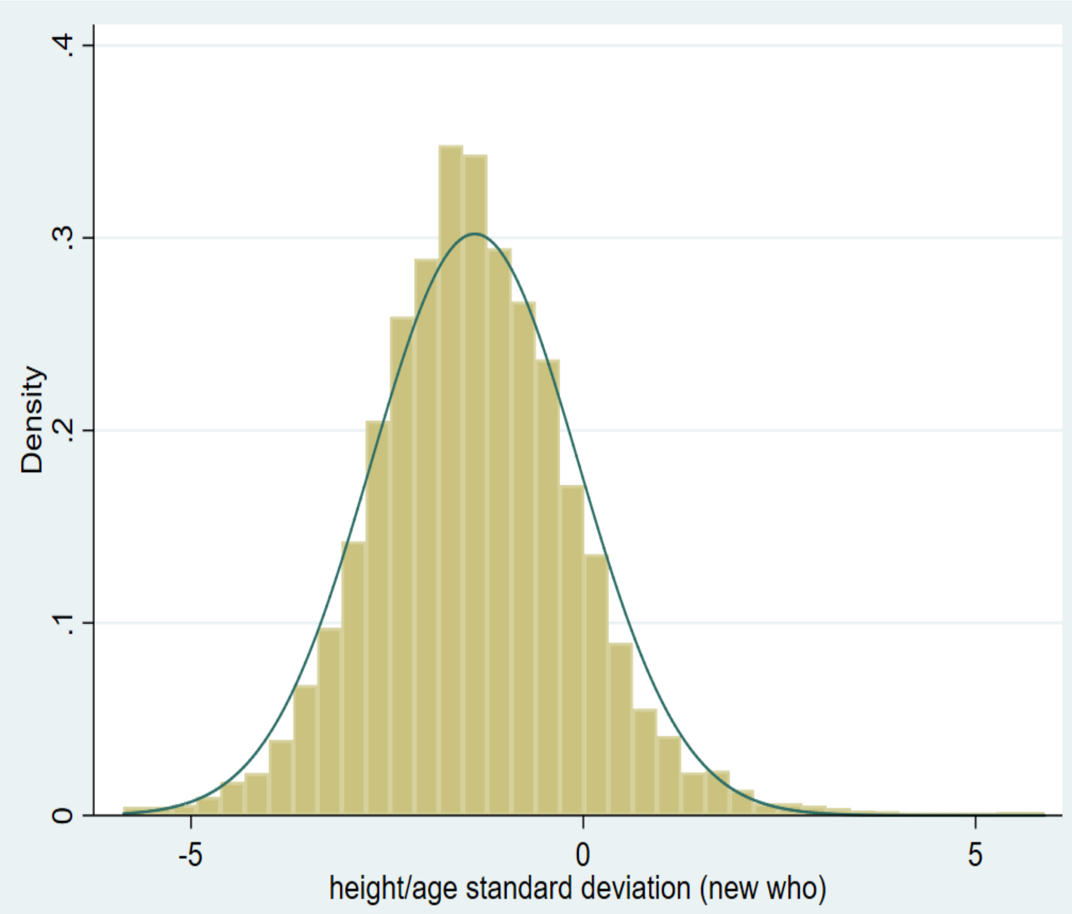
```
. sum hw70, detail
```

height/age standard deviation (new who)

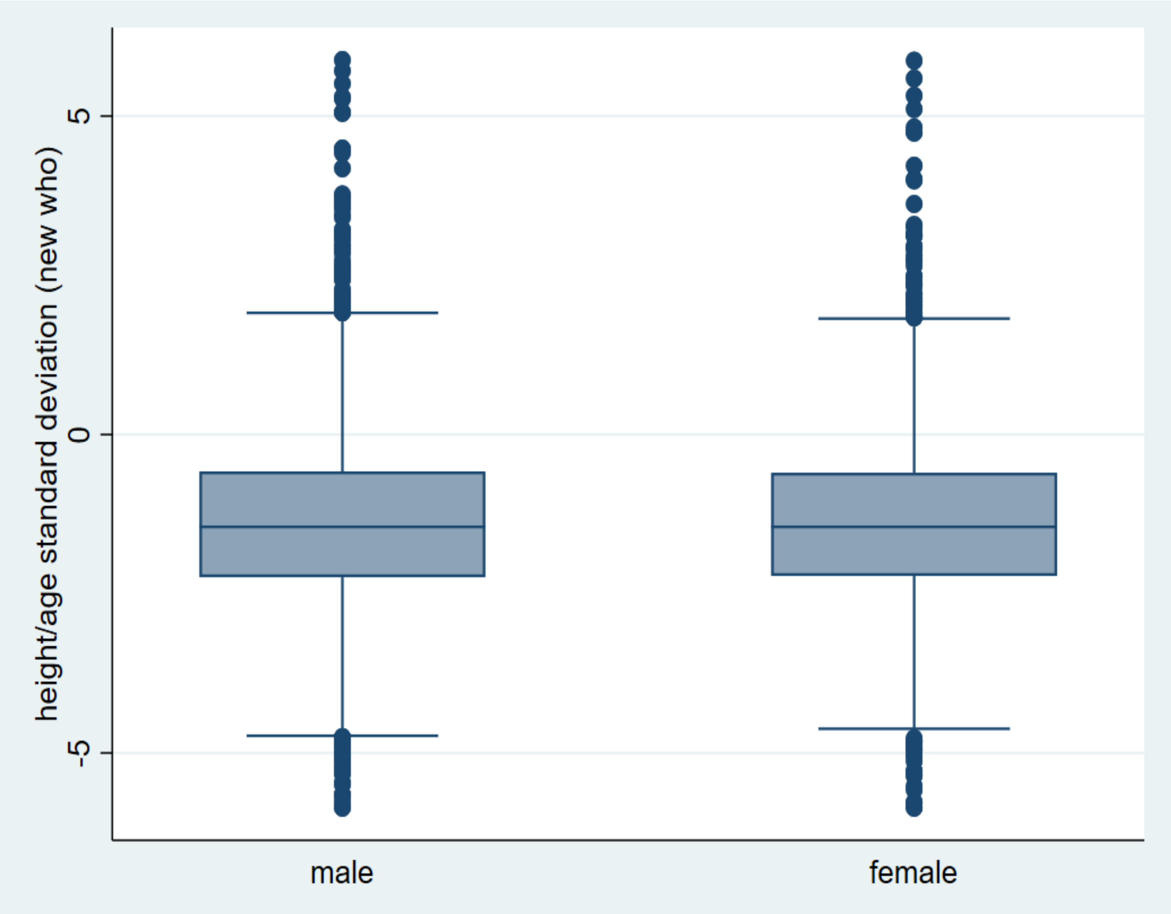
	Percentiles	Smallest		
1%	-4.48	-5.86		
5%	-3.41	-5.86		
10%	-2.94	-5.85	Obs	7,849
25%	-2.23	-5.85	Sum of Wgt.	7,849
50%	-1.45		Mean	-1.382771
		Largest	Std. Dev.	1.320636
75%	-.59	5.59		
90%	.21	5.71	Variance	1.744079
95%	.73	5.87	Skewness	.4471885
99%	2.2	5.88	Kurtosis	4.754577

Data visualization

histogram hw70, normal



graph box hw70, over(b4)



Bivariate analysis

Cross tabulation

sex of child	stunting non-stunt	stunted	Total
male	2,805 52.06	1,287 52.30	4,092 52.13
female	2,583 47.94	1,174 47.70	3,757 47.87
Total	5,388 100.00	2,461 100.00	7,849 100.00

Source	SS	df	MS
Model	.009383704	1	.009383704
Residual	1958.66612	7,847	.249606998
Total	1958.6755	7,848	.249576389

Number of obs	=	7,849
F(1, 7847)	=	0.04
Prob > F	=	0.8463
R-squared	=	0.0000
Adj R-squared	=	-0.0001
Root MSE	=	.49961

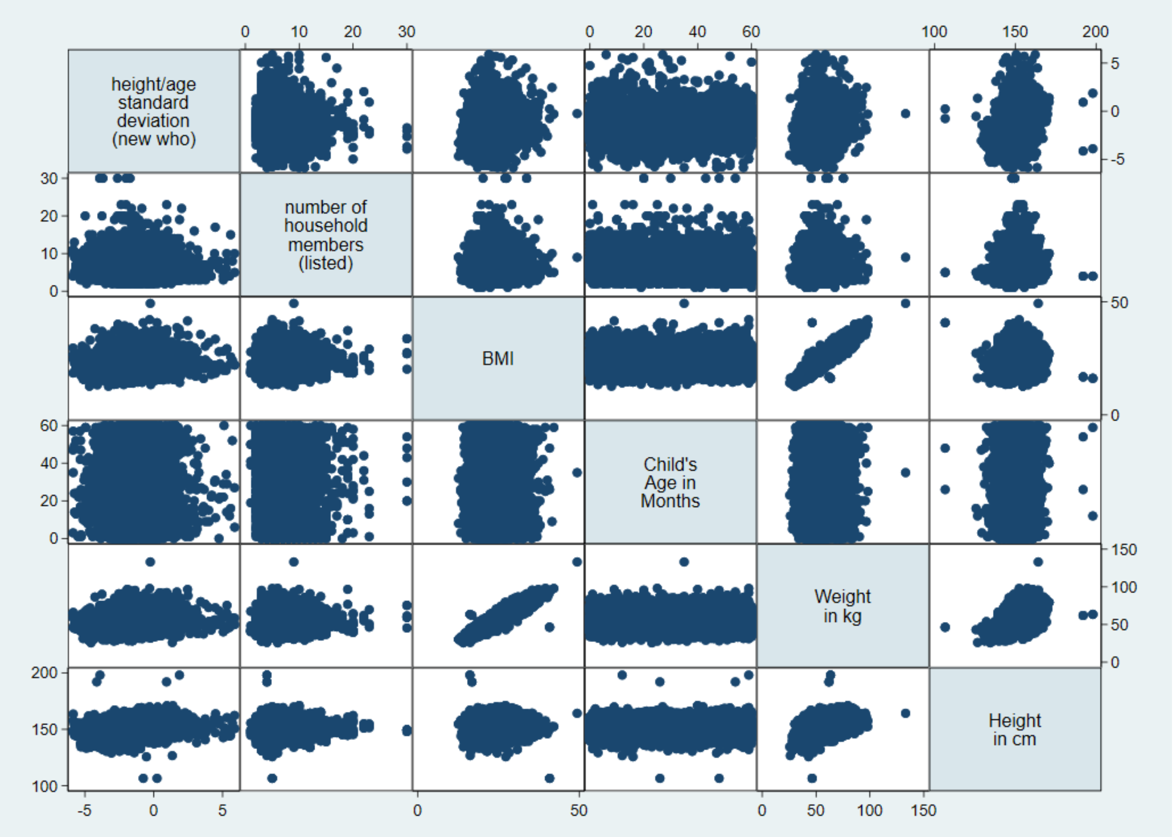
b4	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
stunting	-.0023568	.0121553	-0.19	0.846	-.0261845	.0214708
_cons	1.479399	.0068064	217.36	0.000	1.466056	1.492741

Bivariate analysis

Correlations and scatterplot matrix

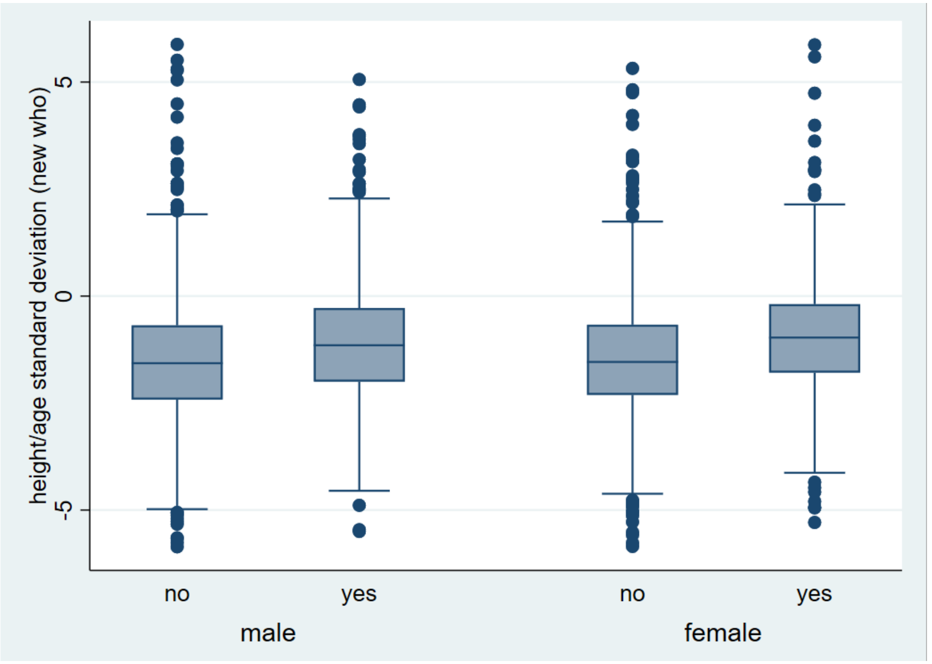
```
. pwcorr hw70 v136 bmi ChildAge w h, sig
```

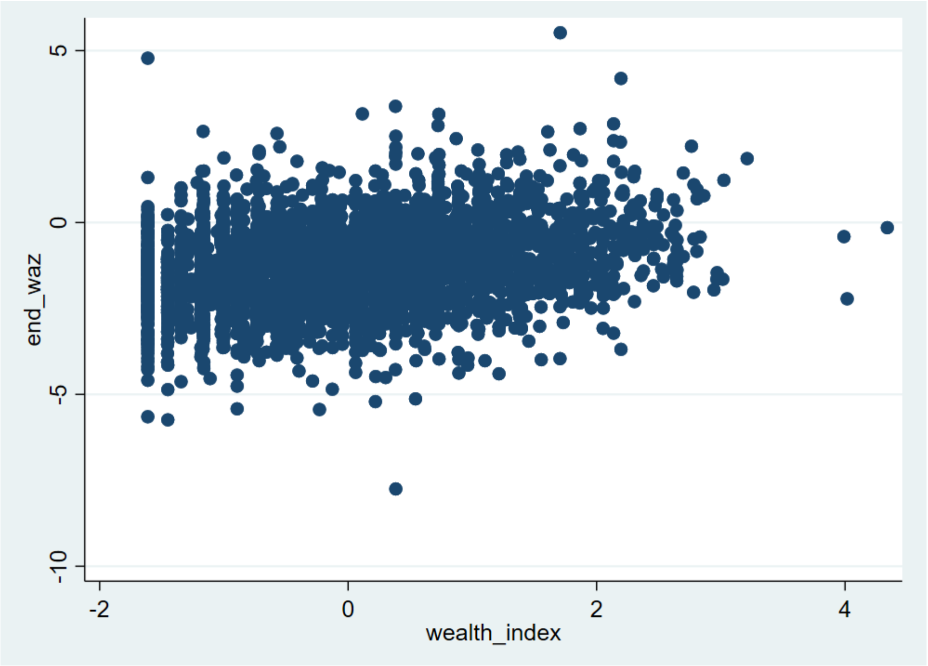
	hw70	v136	bmi	ChildAge	w	h
hw70	1.0000					
v136	-0.0050	1.0000				
	0.6588					
bmi	0.1640	0.0149	1.0000			
	0.0000	0.1657				
ChildAge	-0.0941	-0.0795	0.1146	1.0000		
	0.0000	0.0000	0.0000			
w	0.2548	0.0283	0.9240	0.1022	1.0000	
	0.0000	0.0085	0.0000	0.0000		
h	0.2825	0.0374	0.0442	-0.0020	0.4155	1.0000
	0.0000	0.0005	0.0000	0.8509	0.0000	



Visual plot

- **Box plot for a quantitative variable by groups of a categorical variable**
  - ✓ Weight for age z score for boys and girls in different age group
- **Scatterplots for two quantitative variables to show crude relationship**
  - ✓ Linear/ non linear, strong/ week/ no relationship
  - ✓ Weight for age z score and wealth index have non linear relationship





Simple linear regression

```
. reg hw70 i.b4

      Source |       SS           df       MS      Number of obs   =     7,849
-------------+----------------------------------   F(1, 7847)      =      0.34
       Model |  .594686997         1  .594686997   Prob > F        =    0.5593
    Residual |  13686.9412     7,847  1.74422597   R-squared       =    0.0000
-------------+----------------------------------   Adj R-squared   =   -0.0001
       Total |  13687.5359     7,848  1.74407949   Root MSE        =    1.3207

------------------------------------------------------------------------------
        hw70 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
          b4 |
     female  |  -.0174246   .0298415    -0.58   0.559    -.0759218    .0410726
       _cons |  -1.374431   .0206459   -66.57   0.000    -1.414902   -1.333959
------------------------------------------------------------------------------
```

Multivariable analysis

Multiple linear regression

1. Entry method
2. Forward selection
3. Backward elimination

Check for effect modifiers

1. Interaction
2. Estimate the effect of effect modifier

Model fit and diagnostics

1. Goodness of fit
2. Multicollinearity
3. Identify influential individuals

Hypothesis test

1. Likelihood ratio test
2. Wald test

Multiple Linear Regression

Entry method	A procedure for variable selection in which all variables in a block are entered in a single step
Forward Selection	A stepwise variable selection procedure in which variables are sequentially entered into the model
Backward Elimination	A variable selection procedure in which all variables are entered into the equation and then sequentially removed
Stepwise selection	Combination of Forward and Backward method

Variable selection

- If we are only interested in the **predictions** from the linear model, we can use all of the variables

- **To be considered:**
  - ✓ Literature review
  - ✓ Biological plausibility
  - ✓ Availability of data

Regression analysis

```
. reg hw70 ChildAge i.b4 i.Mdelivery i.Diarrhea bmi i.edu i.PartnerEdu i.toilet i.waterdrink i.v190
> i.v025
```

Source	SS	df	MS			
				Number of obs	=	4,481
				F(14, 4466)	=	32.51
Model	793.642104	14	56.6887217	Prob > F	=	0.0000
Residual	7787.29153	4,466	1.74368373	R-squared	=	0.0925
				Adj R-squared	=	0.0896
Total	8580.93364	4,480	1.91538697	Root MSE	=	1.3205

hw70	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ChildAge	-.0239668	.0018957	-12.64	0.000	-.0276833	-.0202503
b4						
female	.0658597	.0396016	1.66	0.096	-.0117792	.1434985
Mdelivery						
Non-caesarrean	-.1900547	.0463315	-4.10	0.000	-.2808873	-.099222
Diarrhea						
Yes	.0546046	.077705	0.70	0.482	-.0977357	.2069449
bmi	.0284534	.0054991	5.17	0.000	.0176724	.0392345
edu						
Secondary and above	.0804704	.0475508	1.69	0.091	-.0127527	.1736935

Interpretation of STATA output

Source	SS	df	MS
Model	793.642104	14	56.6887217
Residual	7787.29153	4,466	1.74368373
Total	8580.93364	4,480	1.91538697

Number of obs	=	4,481
F(14, 4466)	=	32.51
Prob > F	=	0.0000
R-squared	=	0.0925
Adj R-squared	=	0.0896
Root MSE	=	1.3205

hw70	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ChildAge	-.0240432	.001896	-12.68	0.000	-.0277602	-.0203261
b4						
female	.0690642	.0395836	1.74	0.081	-.0085393	.1466676
bmi	.0286517	.0055002	5.21	0.000	.0178686	.0394347
edu						
Secondary and above	.0748253	.0474852	1.58	0.115	-.0182692	.1679198
PartnerEdu						
Secondary and above	.2480062	.0468394	5.29	0.000	.1561778	.3398347
toilet						
unimproved	-.0154156	.0501998	-0.31	0.759	-.113832	.0830008
waterdrink						
Improved water	-.0000542	.0622986	-0.00	0.999	-.1221903	.122082
Diarrhea						
Yes	.0554164	.0777308	0.71	0.476	-.0969745	.2078073
Mdelivery						
Non-caesarrean	-.1919778	.0463377	-4.14	0.000	-.2828227	-.1011329
v190						
poorer	.0013483	.0600784	0.02	0.982	-.1164352	.1191317
middle	.0448888	.0660676	0.68	0.497	-.0846364	.174414
richer	.1664902	.0702273	2.37	0.018	.0288099	.3041704
richest	.3587349	.0795614	4.51	0.000	.2027551	.5147146

Post estimation

Assumptions in linear regression are based mostly on predictive values and residuals. In particular, we will consider the following assumptions:

Linearity- **Big deal if violated**

Homogeneity of variance- not as big deal if violated

Normality of residual- not as big deal if violated

Independence- **Huge deal if violated**

Multicoliniarity- **Big deal if violated**

Check for Multicoliniarty

```
. estat vif
```

Variable	VIF	1/VIF
ChildAge	1.02	0.984288
2.b4	1.00	0.995227
1.Mdelivery	1.21	0.825838
1.Diarrhea	1.01	0.994852
bmi	1.16	0.864292
1.edu	1.30	0.768287
1.PartnerEdu	1.41	0.710592
2.toilet	1.59	0.627674
1.waterdrink	1.34	0.745393
v190		
2	1.58	0.631035
3	1.72	0.580933
4	2.00	0.499700
5	2.46	0.407123
2.v025	1.17	0.854660
Mean VIF	1.43	

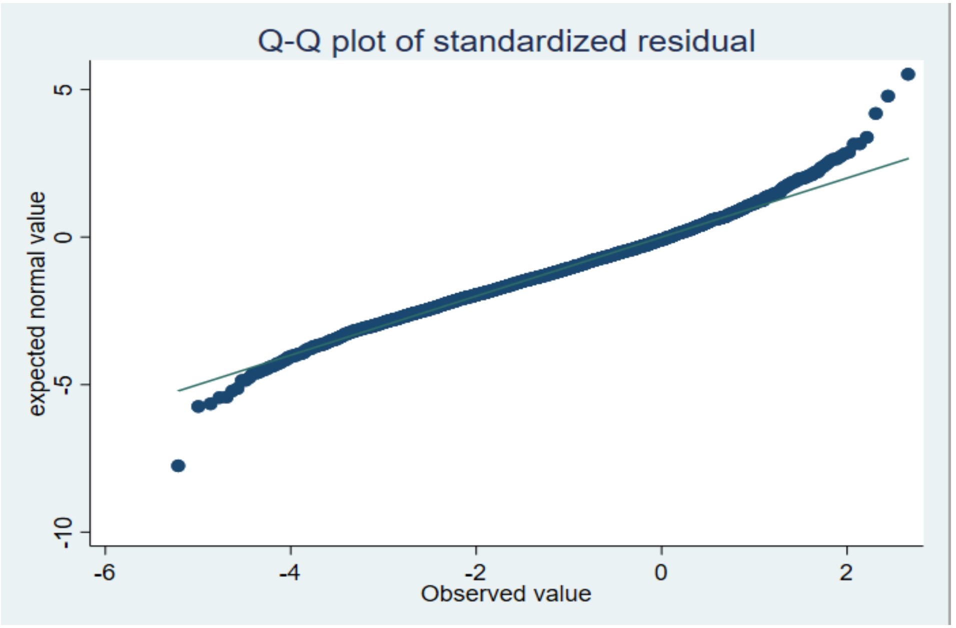
Variable	VIF	1/VIF
ChildAge	1.02	0.975687
2.b4	1.01	0.994776
1.Mdelivery	1.20	0.832963
v190		
2	1.57	0.637598
3	1.69	0.590228
4	1.96	0.510788
5	2.42	0.413228
bmi	144.93	0.006900
agem	1.10	0.911235
h	26.61	0.037575
w	177.47	0.005635
1.PartnerEdu	1.33	0.754624
2.toilet	1.59	0.628491
1.waterdrink	1.35	0.740207
2.v025	1.17	0.854790
Mean VIF	24.43	

- **Not associated with the outcome variable**

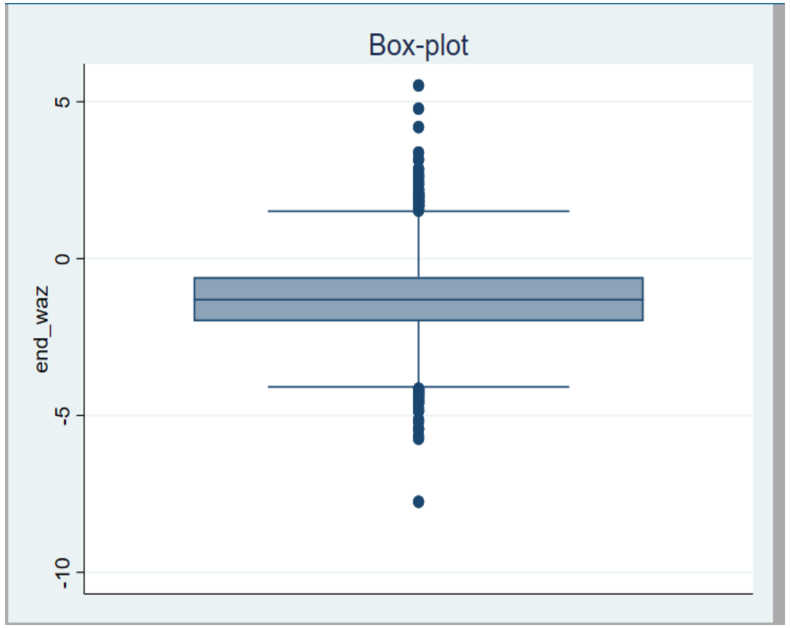
VIF: less than 5= no multicollinearity
VIF: between 5-10= moderate multicollinearity
VIF: more than 10: high multicollinearity

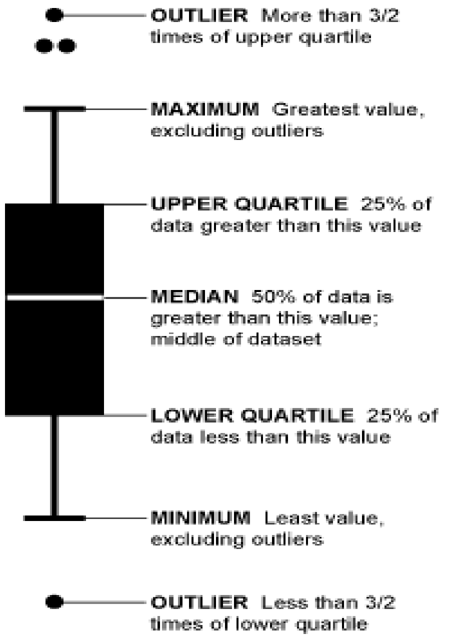
Normality of residual

Q-Q plot for standardized residuals

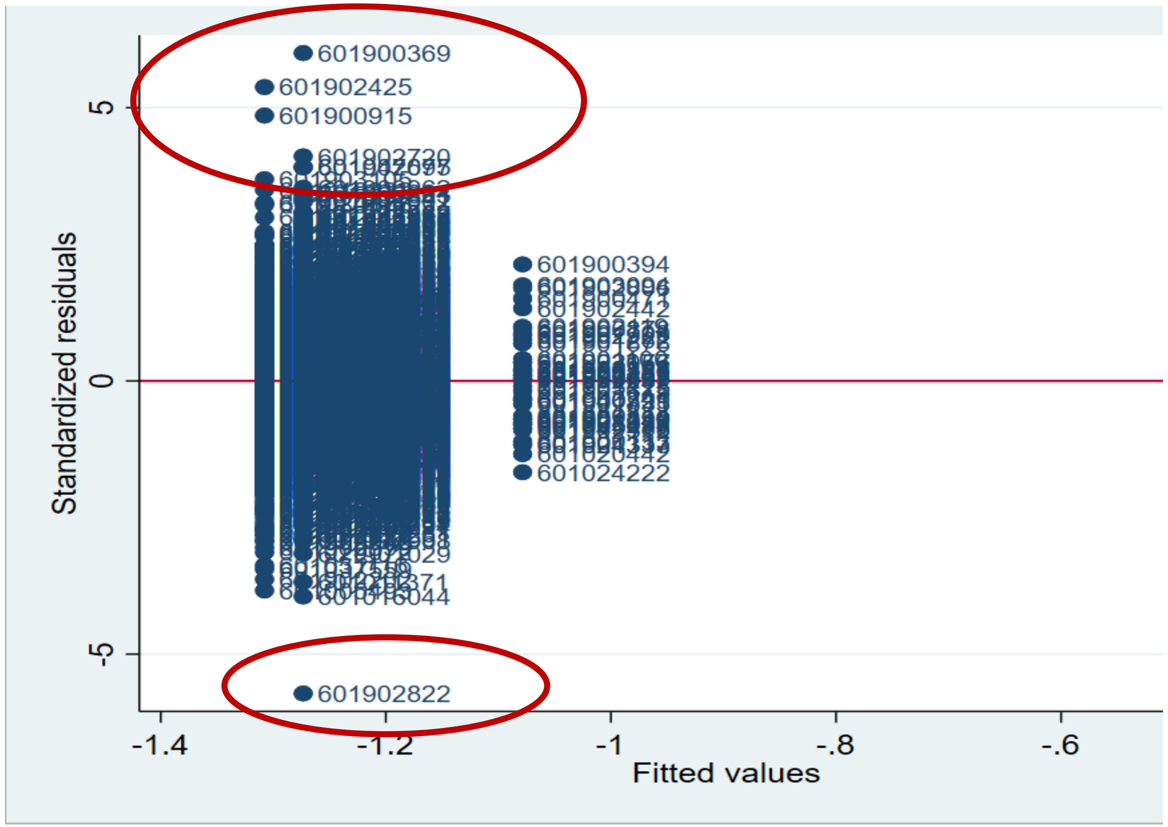


Box plot for residuals





Outlier: checking residuals



Rule of thumb: absolute (values) >2 or 3 suggests influential

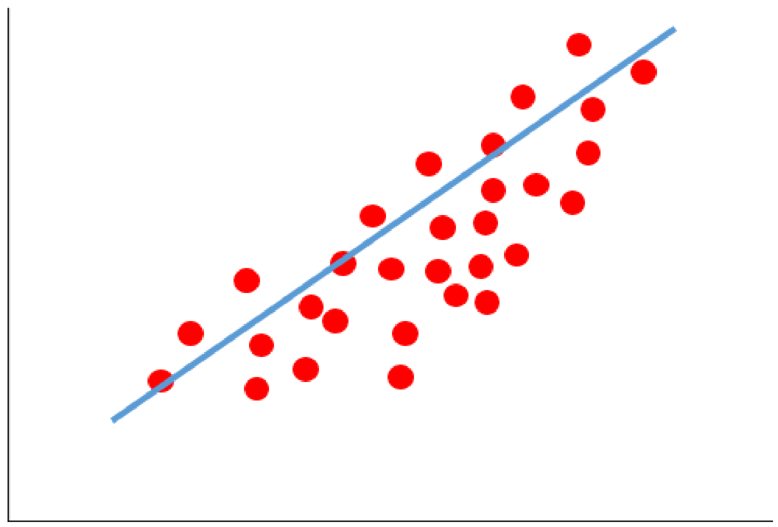
Points far from the rest need attention

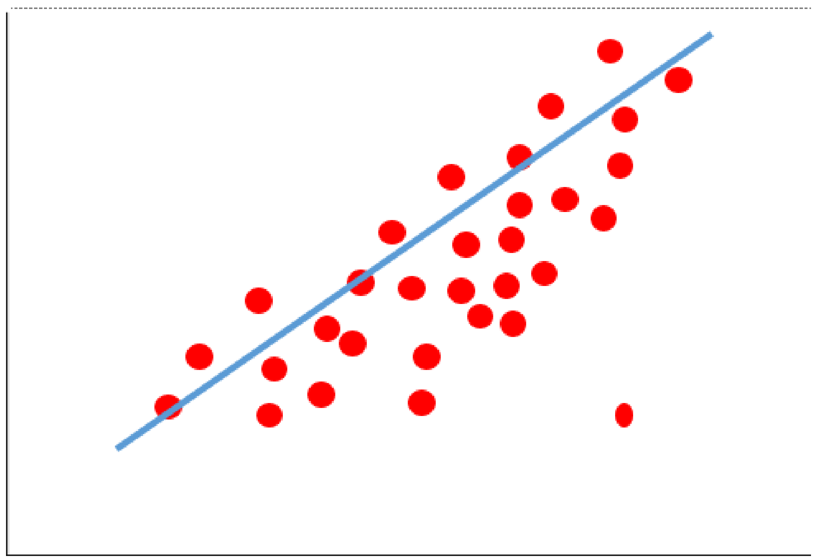
Next, need to check the leverage of observations with large residuals

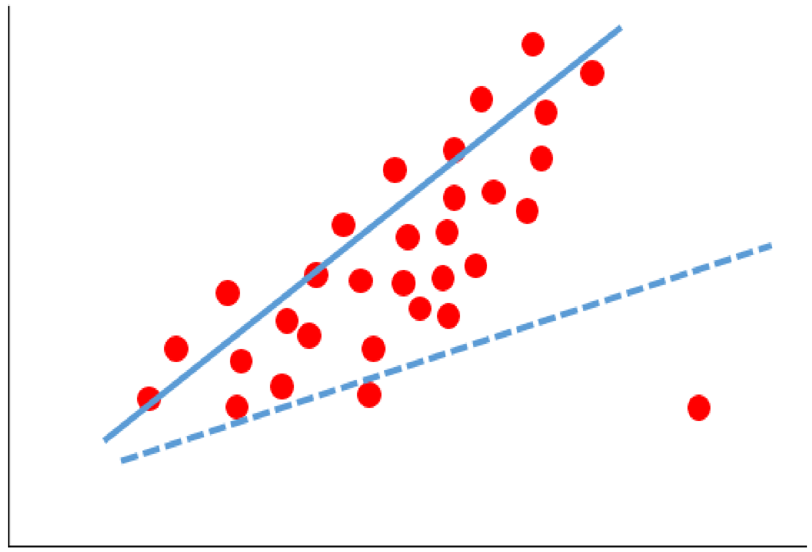
```
. tab outl
```

outl	Freq.	Percent	Cum.
0	4,432	98.91	98.91
1	49	1.09	100.00
Total	4,481	100.00	

Influential observation







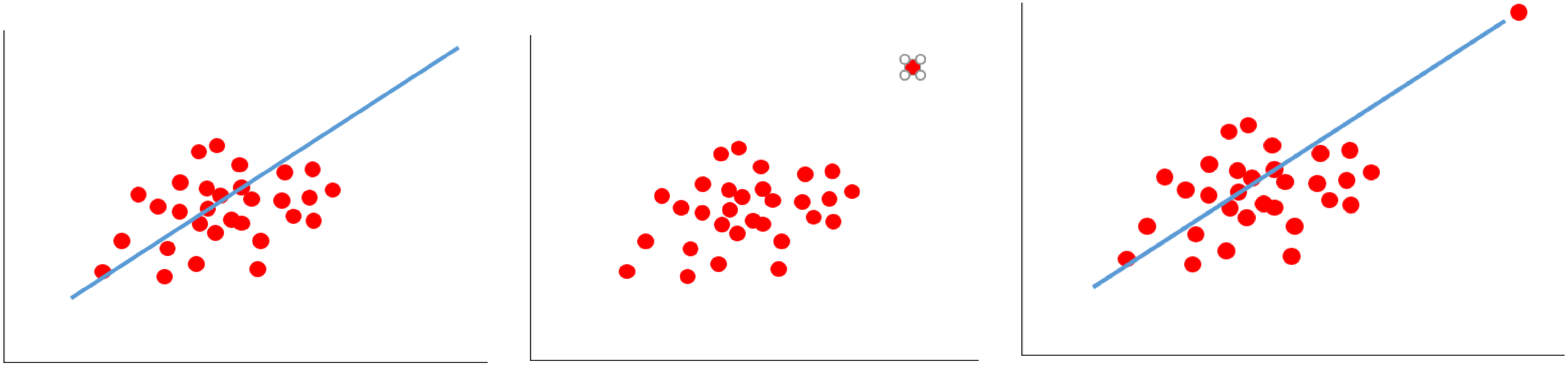
Influential observation: Cook's distance

```
. tab inf
```

inf	Freq.	Percent	Cum.
0	4,481	100.00	100.00
Total	4,481	100.00	

A general rule of thumb is that observations with a Cook's Distance of more than 1, is a possible outlier

High Leverage value



Observations with high leverage (extreme explanatory X values) cannot affect regression estimates

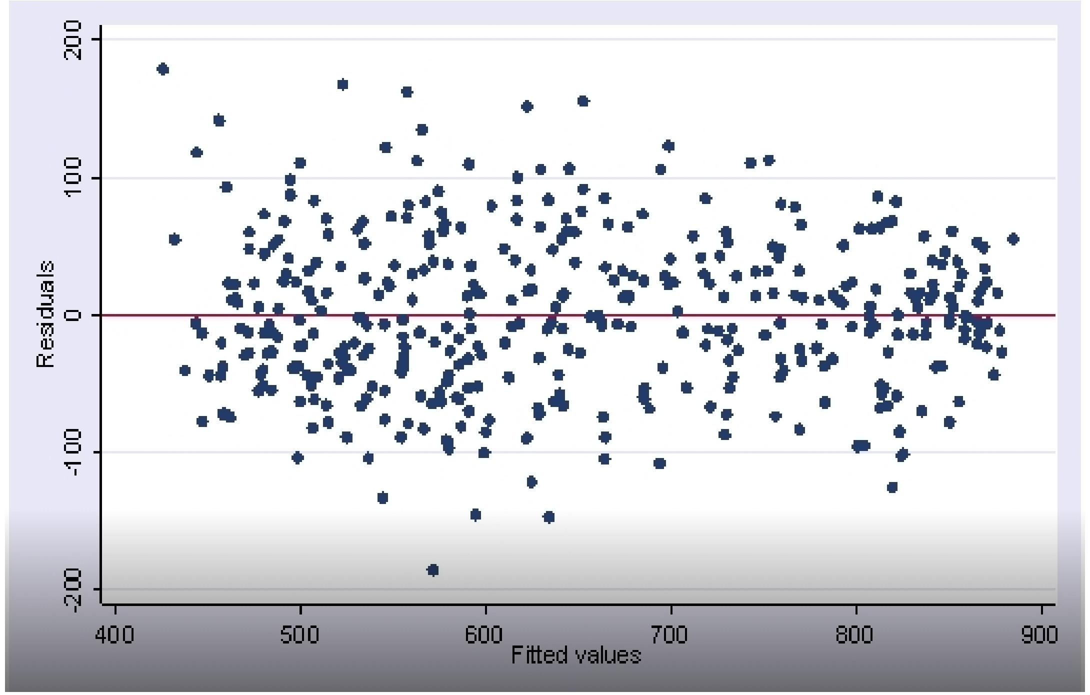
Leverage of observations

```
. tab cut

        cut |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |      4,626       99.44       99.44
          1 |         26        0.56      100.00
------------+-----------------------------------
      Total |      4,652      100.00

.
```

- Observations with high leverage (extreme explanatory X values) cannot affect regression estimates
- Observations with large residual values may not necessary to have high leverage, that means they may not have much impact on regression estimate
- Rule of thumb: values > 2 or 3 times the average leverage
- If some observations have high leverage, repeat analysis, without these subjects to see if the regression estimates change substantially, and report both results

Test of homogeneity

Non significant Chi square value indicates the absence of heteroskedasticity




```
. estat hettest

Breusch-Pagan / Cook-Weisberg test for heteroskedasticity
         Ho: Constant variance
         Variables: fitted values of hw70

         chi2(1)      =     1.66
         Prob > chi2  =   0.1973
```

Interactions

Girls who were exclusively breastfed, are predicted to have increased weight for age z score by 0.90 (0.90= 0.19 + 0.71)

```
. reg end_waz i.exclu_breastfed i.gender
```

Source	SS	df	MS			
Model	18.7390168	2	9.3695084	Number of obs	=	3,795
Residual	4871.82137	3,792	1.28476302	F(2, 3792)	=	7.29
				Prob > F	=	0.0007
Total	4890.56039	3,794	1.28902488	R-squared	=	0.0038
				Adj R-squared	=	0.0033
				Root MSE	=	1.1335

end_waz	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.exclu_breastfed	.4934698	.1305021	3.78	0.000	.2376087	.749331
gender						
Girl	-.0198342	.0372887	-0.53	0.595	-.092942	.0532736
_cons	-1.279513	.0242904	-52.68	0.000	-1.327136	-1.231889

```
. reg end_waz i.exclu_breastfed##i.gender
```

Source	SS	df	MS			
Model	28.1881643	3	9.39605476	Number of obs	=	3,795
Residual	4862.37223	3,791	1.2826094	F(3, 3791)	=	7.33
				Prob > F	=	0.0001
Total	4890.56039	3,794	1.28902488	R-squared	=	0.0058
				Adj R-squared	=	0.0050
				Root MSE	=	1.1325

end_waz	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
1.exclu_breastfed	.1949738	.170577	1.14	0.253	-.1394577	.5294053
gender						
Girl	-.0343688	.0376403	-0.91	0.361	-.1081659	.0394283
exclu_breastfed#gender						
1#Girl	.7181257	.2645766	2.71	0.007	.1993996	1.236852
_cons	-1.273418	.0243737	-52.25	0.000	-1.321205	-1.225631

Stratum specific estimate

Model for boys

Source	SS	df	MS
Model	345.891888	9	38.432432
Residual	2476.70201	2,194	1.12885233
Total	2822.5939	2,203	1.28125007

Number of obs	=	2,204
F(9, 2194)	=	34.05
Prob > F	=	0.0000
R-squared	=	0.1225
Adj R-squared	=	0.1189
Root MSE	=	1.0625

end_waz	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
agegroup						
12-23 Months	-.1139774	.0553802	-2.06	0.040	-.2225805	-.0053743
24-59 Months	-.3905528	.0559715	-6.98	0.000	-.5003154	-.2807902
edumoth						
Illiterate	-.1392287	.0717957	-1.94	0.053	-.2800232	.0015659
ppl_sleep_comb	-.0184868	.0089999	-2.05	0.040	-.0361361	-.0008376
wash_nurse						
Yes	.0954588	.0541061	1.76	0.078	-.0106457	.2015633
wealth_index	.3228925	.0266161	12.13	0.000	.2706971	.3750878
hand_washing						
Without soap	-.2192843	.0639184	-3.43	0.001	-.3446311	-.0939375
1.exclu_breastfed	.0057598	.1631773	0.04	0.972	-.3142384	.325758
STATUS						
case	-.1836669	.0473891	-3.88	0.000	-.276599	-.0907347
_cons	-.9054725	.0699354	-12.95	0.000	-1.042619	-.7683259

Model for girls

Source	SS	df	MS
Model	260.71973	9	28.9688589
Residual	1806.87774	1,581	1.14287017
Total	2067.59747	1,590	1.30037577

Number of obs	=	1,591
F(9, 1581)	=	25.35
Prob > F	=	0.0000
R-squared	=	0.1261
Adj R-squared	=	0.1211
Root MSE	=	1.0691

end_waz	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
agegroup						
12-23 Months	-.2557737	.0657349	-3.89	0.000	-.3847105	-.1268369
24-59 Months	-.519987	.0672311	-7.73	0.000	-.6518585	-.3881155
edumoth						
Illiterate	-.2979517	.0839957	-3.55	0.000	-.4627063	-.1331972
ppl_sleep_comb	-.0342381	.0108328	-3.16	0.002	-.0554863	-.0129899
wash_nurse						
Yes	.1239155	.0666349	1.86	0.063	-.0067865	.2546175
wealth_index	.2622395	.0312448	8.39	0.000	.200954	.3235251
hand_washing						
Without soap	-.1844716	.0761689	-2.42	0.016	-.3338742	-.0350689
1.exclu_breastfed	.623043	.1951464	3.19	0.001	.24027	1.005816
STATUS						
case	-.141141	.0564893	-2.50	0.013	-.2519428	-.0303391
_cons	-.7728289	.0855913	-9.03	0.000	-.9407133	-.6049445

Confounding vs Effect modifier

	Confounding	Effect modification
Study design	Literature review and collect data	Literature review and collect data
Analysis	If the stratum-specific measures are similar to each other, and at least 10% different than the crude, then the covariable is a confounder	If the stratum-specific measures are significantly different than each other, then the covariable is an effect modifier
Control of confounding or Describe interaction	**During study design** • Matching **During analysis** • Stratification • Multiple variable regression	As it is a biological phenomenon, effect modifier can not be controlled, it should be described • Stratification
Writing results	Report an adjusted measure of association that controls for the confounder	Report the stratum-specific measures of association

Hypothesis test

- **Null hypothesis (H0):** the complex model is not better than the simple model
- Likelihood Ratio Test (LRT) is generally the best
- Wald test makes an additional approximation

Model fit

Model 1

```
. reg hw70 ChildAge i.b4 i.Mdelivery i.Diarrhea bmi i.edu i.PartnerEdu i.toilet i.waterd
> rink i.v190
```

Source	SS	df	MS
Model	786.512357	13	60.5009505
Residual	7794.42128	4,467	1.74488947
Total	8580.93364	4,480	1.91538697

Number of obs	=	4,481
F(13, 4467)	=	34.67
Prob > F	=	0.0000
R-squared	=	0.0917
Adj R-squared	=	0.0890
Root MSE	=	1.3209

hw70	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ChildAge	-.0240432	.001896	-12.68	0.000	-.0277602	-.0203261
b4						
female	.0690642	.0395836	1.74	0.081	-.0085393	.1466676
Mdelivery						

Model 2

```
. reg hw70 ChildAge i.b4 i.Mdelivery i.Diarrhea bmi i.edu i.PartnerEdu i.toilet i.waterd
> rink i.v190 i.v025
```

Source	SS	df	MS
Model	793.642104	14	56.6887217
Residual	7787.29153	4,466	1.74368373
Total	8580.93364	4,480	1.91538697

Number of obs	=	4,481
F(14, 4466)	=	32.51
Prob > F	=	0.0000
R-squared	=	0.0925
Adj R-squared	=	0.0896
Root MSE	=	1.3205

hw70	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
ChildAge	-.0239668	.0018957	-12.64	0.000	-.0276833	-.0202503
b4						
female	.0658597	.0396016	1.66	0.096	-.0117792	.1434985
Mdelivery						
Non-caesarrean	-.1900547	.0463315	-4.10	0.000	-.2808873	-.099222

Likelihood ratio test

- It compares two regression models
- The models must have the same outcome and use the same method
- One model must be “nested” inside the other
- Same number of observations, no missing value

```
. lrtest B A

Likelihood-ratio test                                 LR chi2(1)  =      4.10
(Assumption: B nested in A)                           Prob > chi2 =    0.0429
```

A low p-value (<0.05) provides evidence that adding the extra variable, “residence:urban/rural” improves the model

Wald test in Stata

end_waz	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
agegroup						
12-23 Months	-.1656215	.0425845	-3.89	0.000	-.2491123	-.0821308
24-59 Months	-.4408054	.0432217	-10.20	0.000	-.5255454	-.3560653
gender						
Girl	-.0243261	.0352875	-0.69	0.491	-.0935104	.0448583
edumoth						
Illiterate	-.2402525	.0550158	-4.37	0.000	-.3481161	-.1323889
ppl_sleep_comb	-.016923	.0068721	-2.46	0.014	-.0303964	-.0034496
wash_nurse						
Yes	.1168969	.0421223	2.78	0.006	.0343123	.1994816
wiq						
Lower middle	.1733797	.0557806	3.11	0.002	.0640168	.2827427
Middle	.2854186	.0563743	5.06	0.000	.1748917	.3959456
Upper middle	.4290963	.0583157	7.36	0.000	.314763	.5434296
Richest	.7942495	.0619274	12.83	0.000	.6728351	.9156638
hand_washing						
Without soap	-.2134211	.0492194	-4.34	0.000	-.3099203	-.116922
1.exclu_breastfed	.2701748	.1258386	2.15	0.032	.0234566	.516893
1.pn35	-.1985539	.0791935	-2.51	0.012	-.3538202	-.0432877
STATUS						
case	-.1633458	.0365379	-4.47	0.000	-.2349817	-.0917099
_cons	-1.21805	.0628143	-19.39	0.000	-1.341203	-1.094897

Summary

- Multiple linear regression is used when we have a continuous outcome and predictor variables can be continuous or categorical
- We assume that the data are normally distributed and have linear relationship
- We can use any suitable method for Multiple linear regression which depends on the research question
- Report an adjusted measure of association that controls for the confounder
- Report the stratum-specific measures of association for the effect modifier
- We can use Likelihood Ratio Tests and Wald Tests to help us decide which variables are associated with the outcome

THANK YOU

icddr,b thanks its core donors for their on-going support

Government of the People's Republic of Bangladesh

Canada

SWEDEN

UKaid from the British people