Week 5: Linear Regression

Video 1: Introduction to Linear Regression

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Hea

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

Key Concepts

- Advantages of regression models
- Linear regression
- Log odds scale
- Logistic regression

Advantages of Regression Models

- Generally more efficient than stratification-based methods when data are sparse
- Modelling of a continuous outcomes
- Specify continuous and categorical exposures, confounders and modifiers
- Specify interactions to model effect modification
- Model nonlinear relationships between exposure and outcome and other covariates

Regression Models

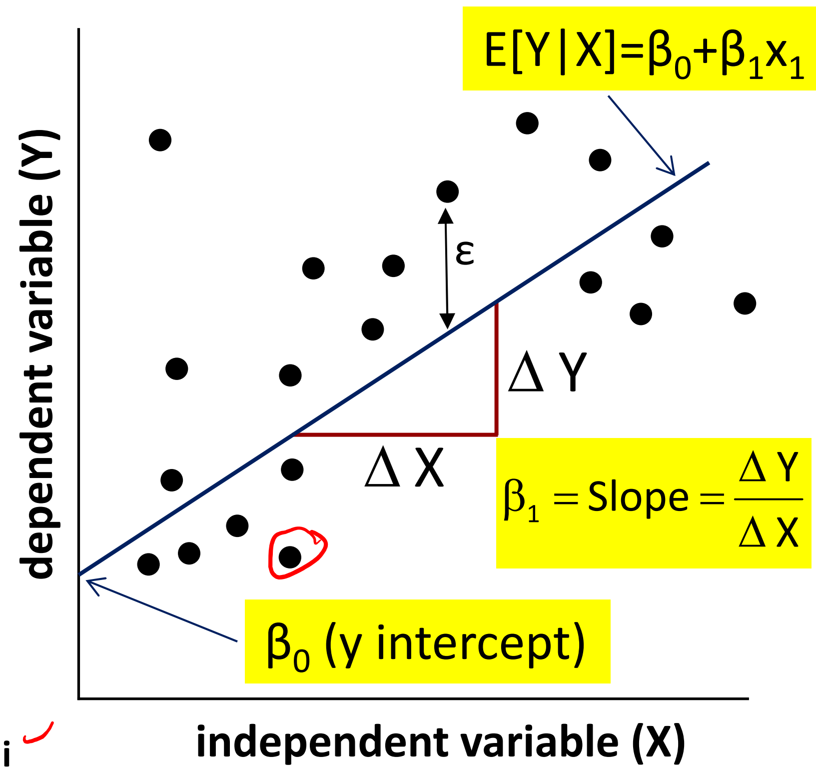
y	=	$\beta_0+\beta_1 x_1+\ldots\beta_i x_i+\varepsilon$
Dependent		Independent
Predicted		Predictor variables
Response variable		Explanatory variables
Outcome variable		Covariables

Simple Linear Regression

- Predict a continuous dependent (outcome) variable y from a continuous independent (exposure) variable x

- Simple linear regression fits a straight line to the data using the least squares method.

- Regression line: $\mathbf{E[Y|X]=\beta_0+\beta_1 x_1}$
  - Often presented as y = mx + b where
    - b=y-intercept
    - m=slope=$\Delta y/\Delta x$ (rise/run)



- Individual predicted value: $\mathbf{Y_i=\beta_0+\beta_1 x_{1i}+\varepsilon_i}$
  - β_0 =y-intercept (where the line crosses the Y-axis)
  - β_1 =slope=$\Delta y/\Delta x$ =average change in y when x changes by one unit
  - X_1 is a known constant
  - ε, the error, is an observation's deviation from the conditional mean, N(0, σ2)

BREAK

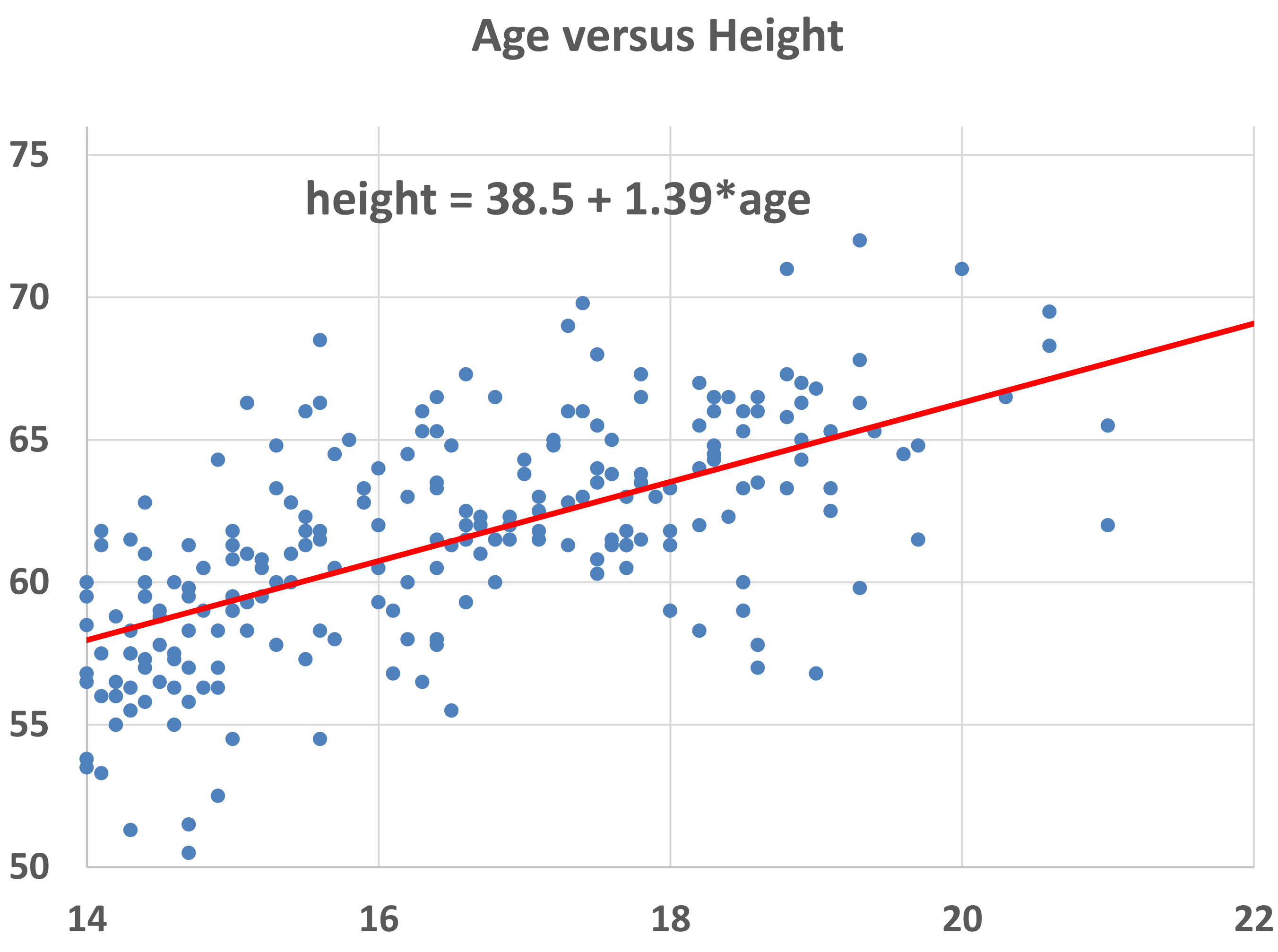
Week 5: Linear Regression

Video 2: Linear Regression Example

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Hea

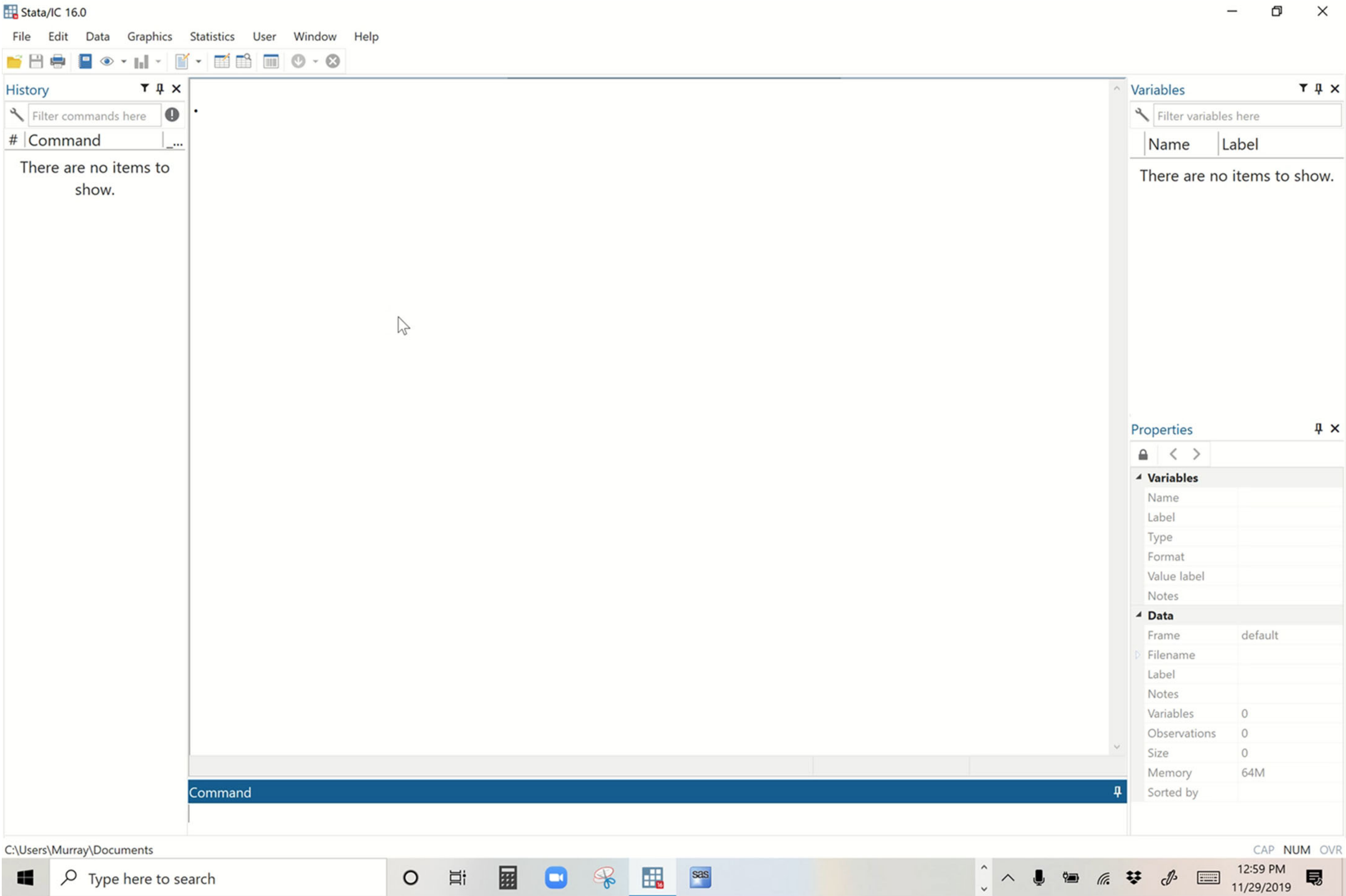
HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

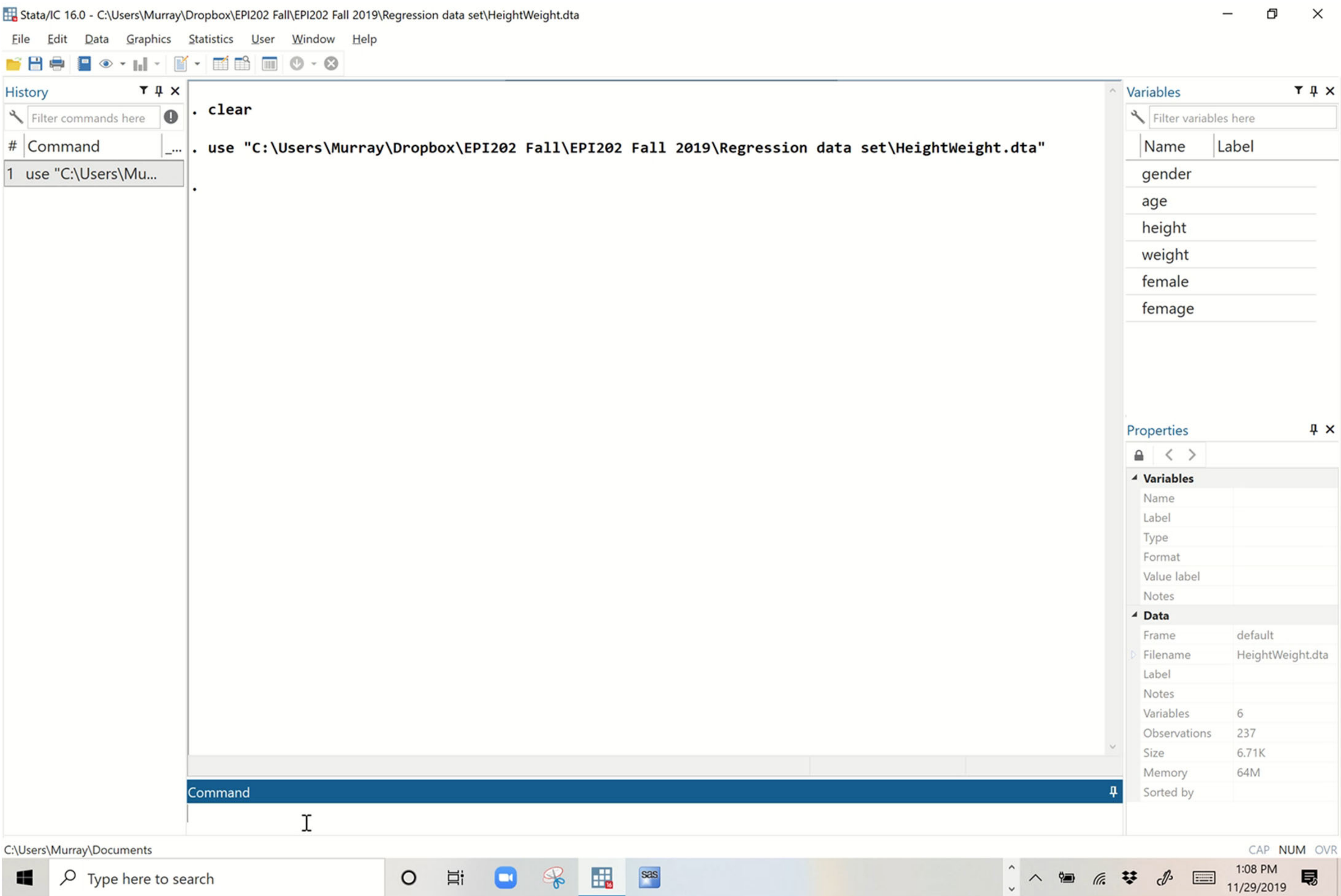
Linear Regression



Linear Regression Example

- Download the HeightWeight dataset from the Regression I module on Canvas
  - The dataset is available in multiple formats
    - CSV
    - Excel
    - R
    - SAS
    - Stata


Stata/IC 16.0
File
Edit
Data
Graphics
Statistics
User
Window
Help
History
Filter commands here
Command
There are no items to show.
Variables
Filter variables here
Name
Label
There are no items to show.
Properties
Variables
Name
Label
Type
Format
Value label
Notes
Data
Frame
default
Filename
Label
Notes
Variables
0
Observations
0
Size
0
Memory
64M
Sorted by
Command
C:\Users\Murray\Documents
CAP
NUM
OVR
Type here to search
12:59 PM
11/29/2019


Stata/IC 16.0 - C:\Users\Murray\Dropbox\EPI202 Fall\EPI202 Fall 2019\Regression data set\HeightWeight.dta
File Edit Data Graphics Statistics User Window Help
History
Filter commands here
Command
1 use "C:\Users\Mu...
. clear
. use "C:\Users\Murray\Dropbox\EPI202 Fall\EPI202 Fall 2019\Regression data set\HeightWeight.dta"
.
Variables
Filter variables here
Name Label
gender
age
height
weight
female
femage
Properties
Variables
Name
Label
Type
Format
Value label
Notes
Data
Frame default
Filename HeightWeight.dta
Label
Notes
Variables 6
Observations 237
Size 6.71K
Memory 64M
Sorted by
Command
C:\Users\Murray\Documents
CAP NUM OVR
Type here to search
1:08 PM
11/29/2019

Unnamed.pdf* - PDF Annotator

File Edit Tool View Extras Window Help

height	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	1.457297	.1107962	13.15	0.000	1.239011	1.675583
_cons	37.42901	1.828191	20.47	0.000	33.82719	41.03083

Modified

BREAK

Week 5: Linear Regression

Video 3: Linear Regression Adjusted for Covariates

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Hea

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

height	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	1.457297	.1107962	13.15	0.000	1.239011	1.675583
_cons	37.42901	1.828191	20.47	0.000	33.82719	41.03083

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$E(\text{height} \mid \text{Age}) = \beta_0 + \beta_1(\text{age})$$

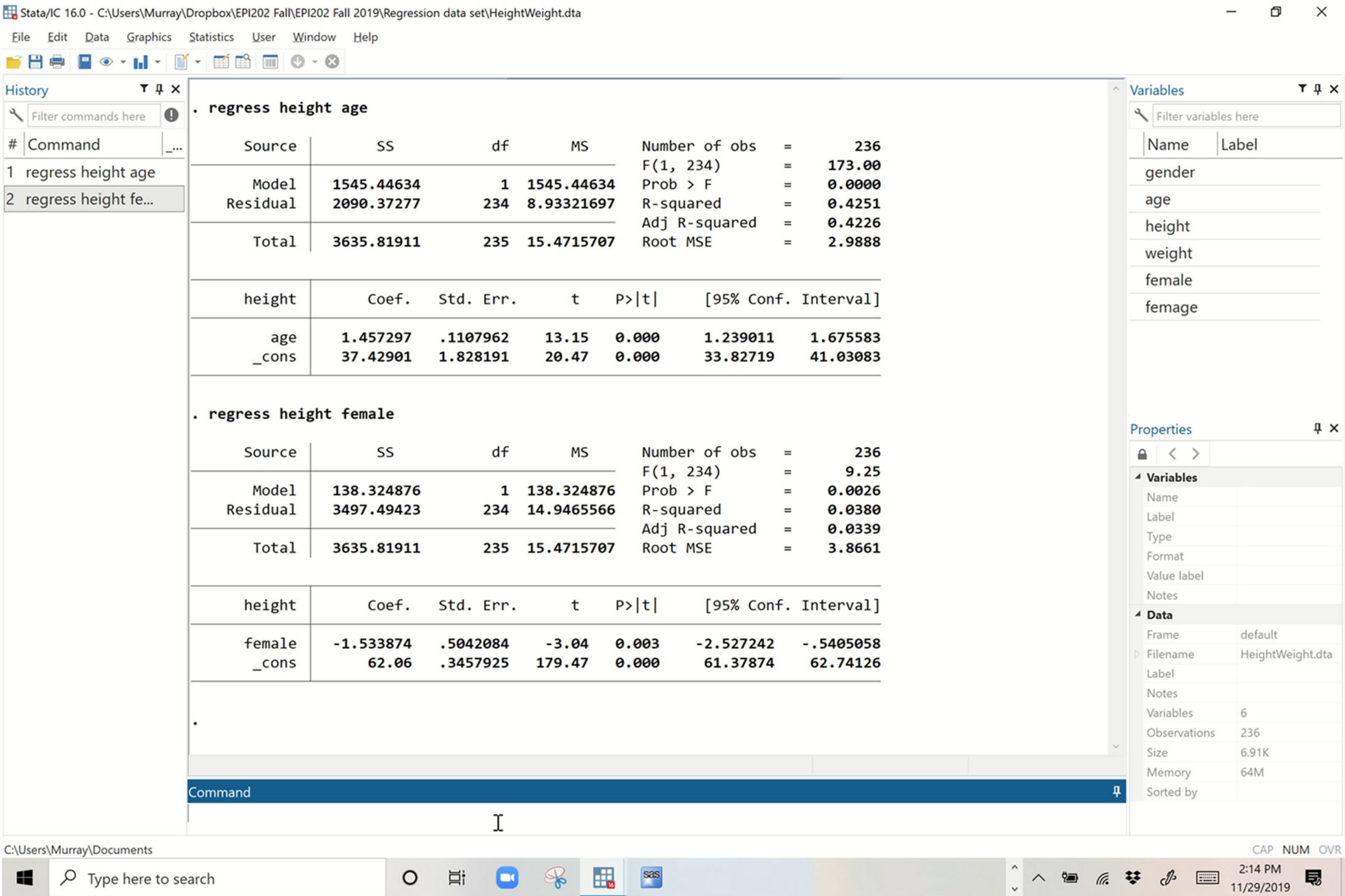
$$E(\text{height} \mid \text{Age}) = 37.43 + 1.46\,(\text{age})$$

$$E(\text{height} \mid 15) = 37.43 + 1.46\,(15)$$

female Age height

female Age height

height	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
female	-1.533874	.5042084	-3.04	0.003	-2.527242	-.5405058
_cons	62.06	.3457925	179.47	0.000	61.37874	62.74126




```
. regress height age

      Source |       SS           df       MS      Number of obs   =       236
-------------+----------------------------------   F(1, 234)       =    173.00
       Model |  1545.44634         1  1545.44634   Prob > F        =    0.0000
    Residual |  2090.37277       234  8.93321697   R-squared       =    0.4251
-------------+----------------------------------   Adj R-squared   =    0.4226
       Total |  3635.81911       235  15.4715707   Root MSE        =    2.9888

------------------------------------------------------------------------------
      height |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.457297   .1107962    13.15   0.000     1.239011    1.675583
       _cons |   37.42901   1.828191    20.47   0.000     33.82719    41.03083
------------------------------------------------------------------------------

. regress height female

      Source |       SS           df       MS      Number of obs   =       236
-------------+----------------------------------   F(1, 234)       =      9.25
       Model |  138.324876         1  138.324876   Prob > F        =    0.0026
    Residual |  3497.49423       234  14.9465566   R-squared       =    0.0380
-------------+----------------------------------   Adj R-squared   =    0.0339
       Total |  3635.81911       235  15.4715707   Root MSE        =    3.8661

------------------------------------------------------------------------------
      height |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -1.533874   .5042084    -3.04   0.003    -2.527242   -.5405058
       _cons |      62.06   .3457925   179.47   0.000     61.37874    62.74126
------------------------------------------------------------------------------

.
```

height	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	1.465677	.1068379	13.72	0.000	1.255185	1.676169
female	-1.627297	.3758752	-4.33	0.000	-2.367845	-.8867485
_cons	38.0569	1.768544	21.52	0.000	34.57252	41.54128

BREAK

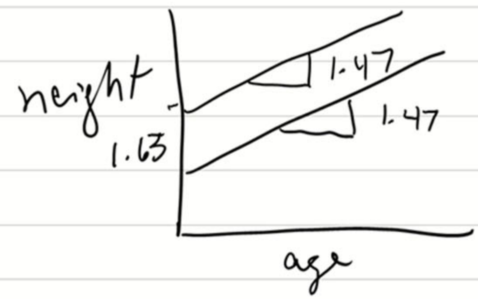
Week 5: Linear Regression

Video 4: Linear Regression with Interaction Terms to Account for Effect Measure Modification

EPI202 – Epidemiologic Methods II
Murray A. Mittleman, MD, DrPH
Department of Epidemiology, Harvard TH Chan School of Public Hea

HARVARD T.H. CHAN
SCHOOL OF PUBLIC HEALTH

$$\text{height} = 38.06 + 1.47(\text{age}) - 1.63(\text{female})$$



Stata/IC 16.0 - C:\Users\Murray\Dropbox\EPI202 Fall\EPI202 Fall 2019\Regression data set\HeightWeight.dta

File Edit Data Graphics Statistics User Window Help

History

#	Command
1	regress height age
2	regress height fe...
3	regress height ag...

```
. regress height female

      Source |       SS           df       MS      Number of obs   =       236
-------------+----------------------------------   F(1, 234)       =      9.25
       Model |  138.324876         1  138.324876   Prob > F        =    0.0026
    Residual |  3497.49423       234  14.9465566   R-squared       =    0.0380
-------------+----------------------------------   Adj R-squared   =    0.0339
       Total |  3635.81911       235  15.4715707   Root MSE        =    3.8661

------------------------------------------------------------------------------
      height |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      female |  -1.533874   .5042084    -3.04   0.003    -2.527242   -.5405058
       _cons |      62.06   .3457925   179.47   0.000     61.37874    62.74126
------------------------------------------------------------------------------

. regress height age female

      Source |       SS           df       MS      Number of obs   =       236
-------------+----------------------------------   F(2, 233)       =    102.43
       Model |  1701.08304         2  850.541519   Prob > F        =    0.0000
    Residual |  1934.73607       233  8.30358829   R-squared       =    0.4679
-------------+----------------------------------   Adj R-squared   =    0.4633
       Total |  3635.81911       235  15.4715707   Root MSE        =    2.8816

------------------------------------------------------------------------------
      height |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         age |   1.465677   .1068379    13.72   0.000     1.255185    1.676169
      female |  -1.627297   .3758752    -4.33   0.000    -2.367845   -.8867485
       _cons |    38.0569   1.768544    21.52   0.000     34.57252    41.54128
------------------------------------------------------------------------------

.
```

Variables

Name	Label
gender	
age	
height	
weight	
female	
femage	

Properties

Variables	
Name	
Label	
Type	
Format	
Value label	
Notes	
Data	
Frame	default
Filename	HeightWeight.dta
Label	
Notes	
Variables	6
Observations	236
Size	6.91K
Memory	64M
Sorted by	

Command

C:\Users\Murray\Documents

CAP NUM OVR

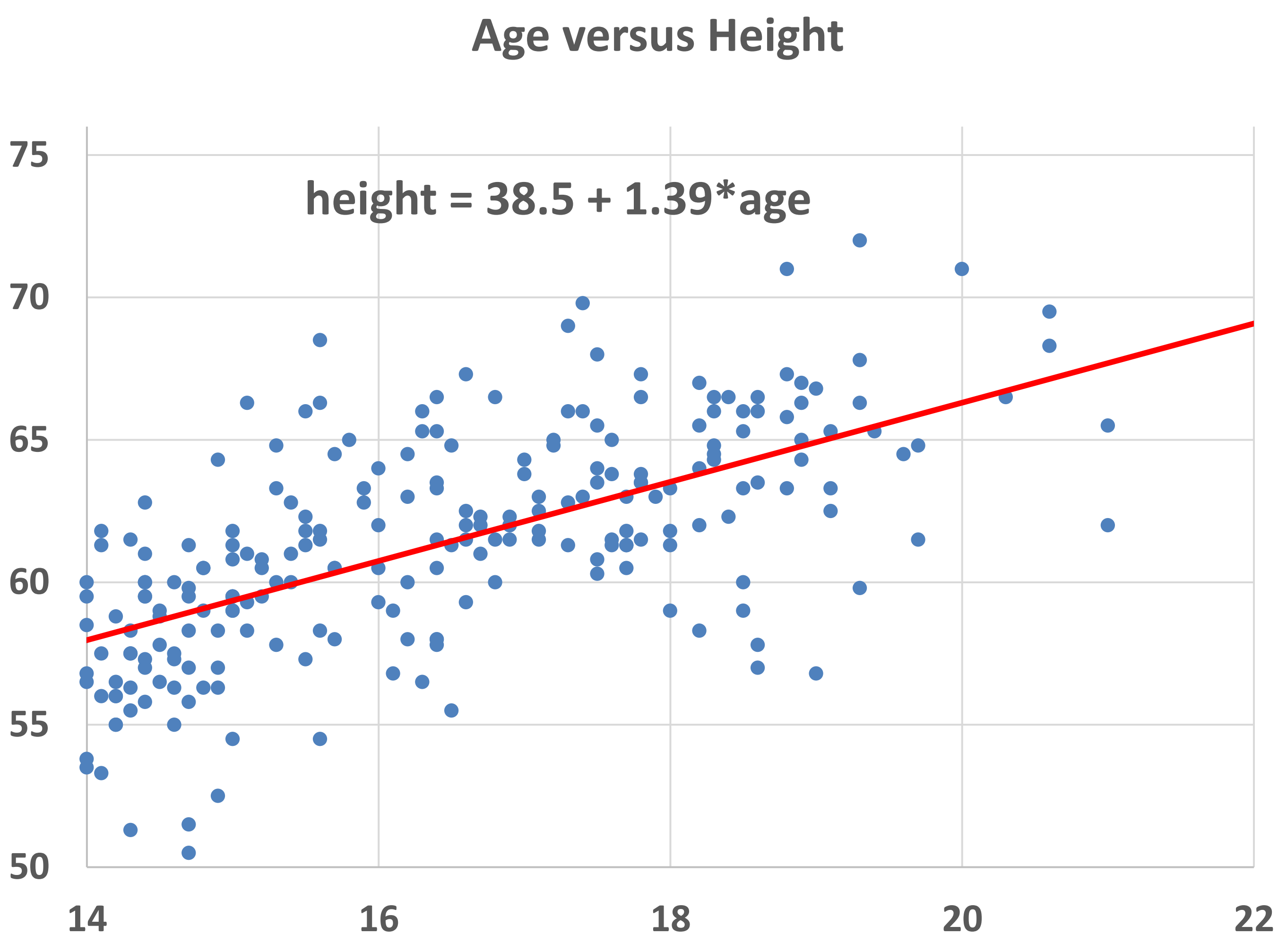
Type here to search

2:30 PM 11/29/2019

height	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
age	1.918139	.1449452	13.23	0.000	1.632562	2.203716
female	13.31463	3.393413	3.92	0.000	6.628783	20.00047
femage	-.9106005	.2056254	-4.43	0.000	-1.315732	-.5054687
_cons	30.64702	2.386661	12.84	0.000	25.94473	35.34932

$$height = \beta_0 + \beta_1(age) + \beta_2(female) + \beta_3(age * female)$$

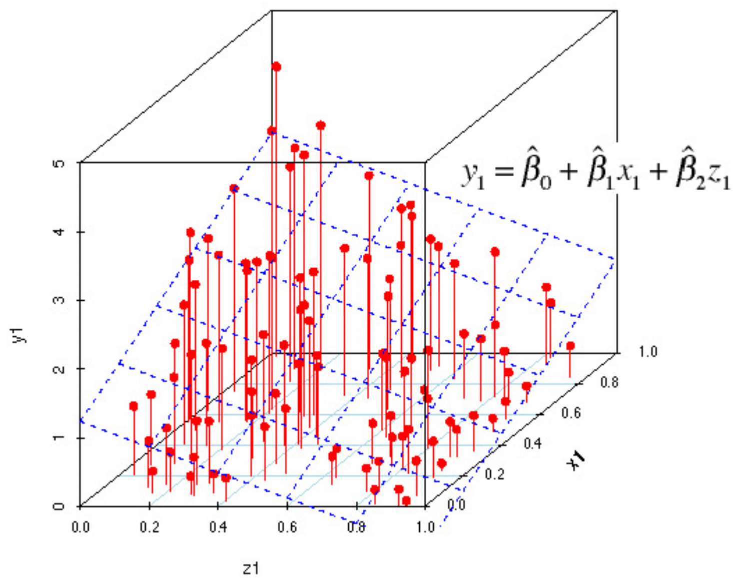
Linear Regression



Multiple Linear Regression

- Multiple linear regression models predict a continuous dependent (outcome) variable y from continuous and categorical independent variables x_i:

$$E[Y|X]=\beta_0+\beta_1 x_1 \ldots \beta_i x_i$$



- The regression line is the best-fit line through the points in the data
  - □ β_0, β_1, …, β_k are parameters
  - □ X_1, X_2, …,X_k are known constants
  - □ β_k = change in average outcome (difference in mean outcome) per unit change in Xi holding all other X's constant

BREAK