Building Statistical Models using Regression

Asad Khan

SHRS, UQ

16th September 2010

Overview

- Aspects of Modeling
- Data Exploration
- Linear Regression Models
- Model Building with Working Examples
- Regression Diagnostics

Aspects of Modelling

- To investigate whether an association exists between the variables
- To measure the strength (as well as direction) of association between the variables
- To study the form of the relationship

Choice of a model depends on the type of outcome:

- For *continuous* outcome variables, relationship could be linear or non-linear, examined by linear or non-linear regression models
- For *categorical outcome* variables, logistic regression is usually used to examine possible relationship

Data Exploration

To explore the distribution of the outcome variable, we can use a number of plots:

- Stem and leaf
- Box plot
- Histogram

We can also use normality tests to investigate distributions

Scatter plot is widely used to investigate linear relationship between two continuous variables

If the pattern is linear or approximately linear, we can

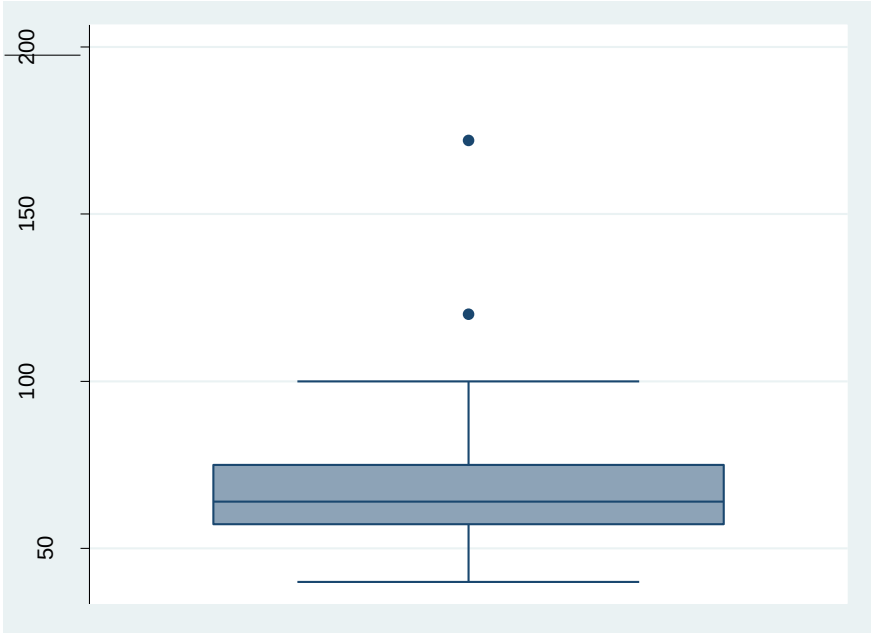
- compute Pearson's correlation coefficient to find strength and direction of association between the variables
- build a linear regression model to regress the effects of explanatory variables on the outcome variable

Let's first examine the distribution of the outcome variable (e.g. weight) through ***stem*** and ***box*** plots

Stata: `stem weight`

Stata: `graph box weight`

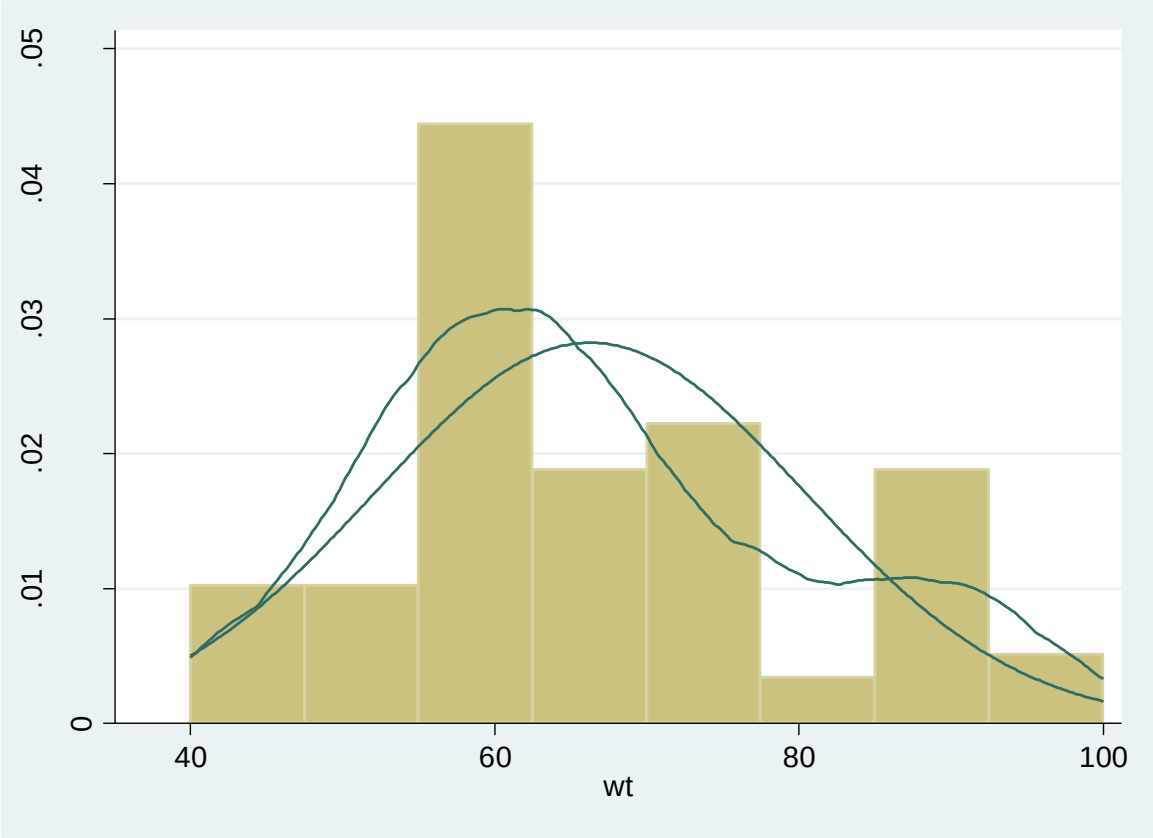
```
Stem-and-leaf plot for weight

  4*  | 05556789
  5*  | 01145555556777888899
  6*  | 000000122223355555589
  7*  | 00002222345568
  8*  | 25555799
  9*  | 00223
 10*  | 00
 11*  |
 12*  | 0
 13*  |
 14*  |
 15*  |
 16*  |
 17*  | 2
```



Excluding students with weight >100, we can draw ***histogram*** along with ***kernel density*** and ***normal plot*** to examine the distribution of weight (<=100kg)

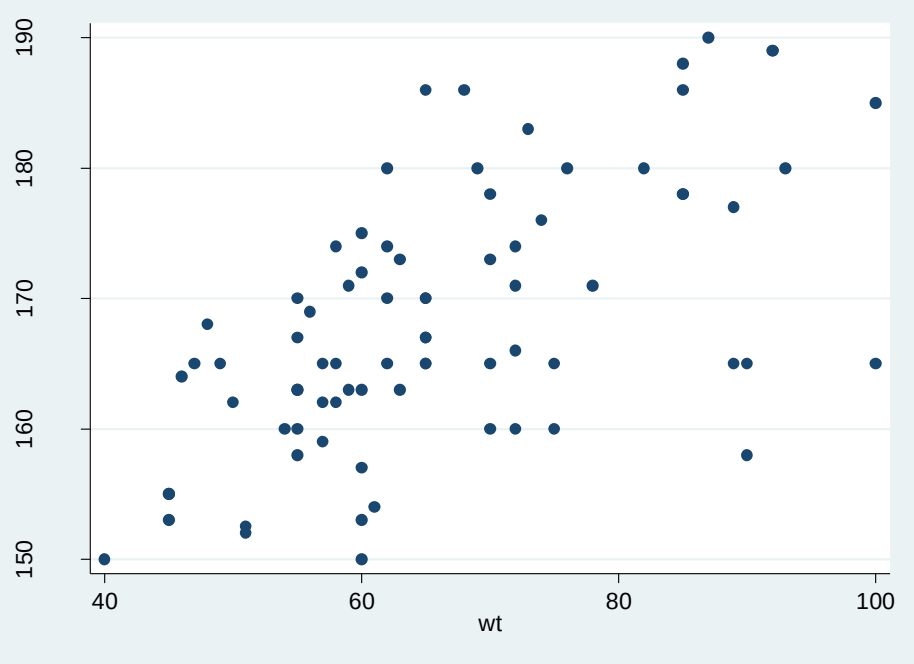
Stata: `gen wt=weight if weight<=100`

Stata: `hist wt, normal kdensity`



To draw a scatter plot of **wt** and **height**, type:

Stata: `scatter wt height`

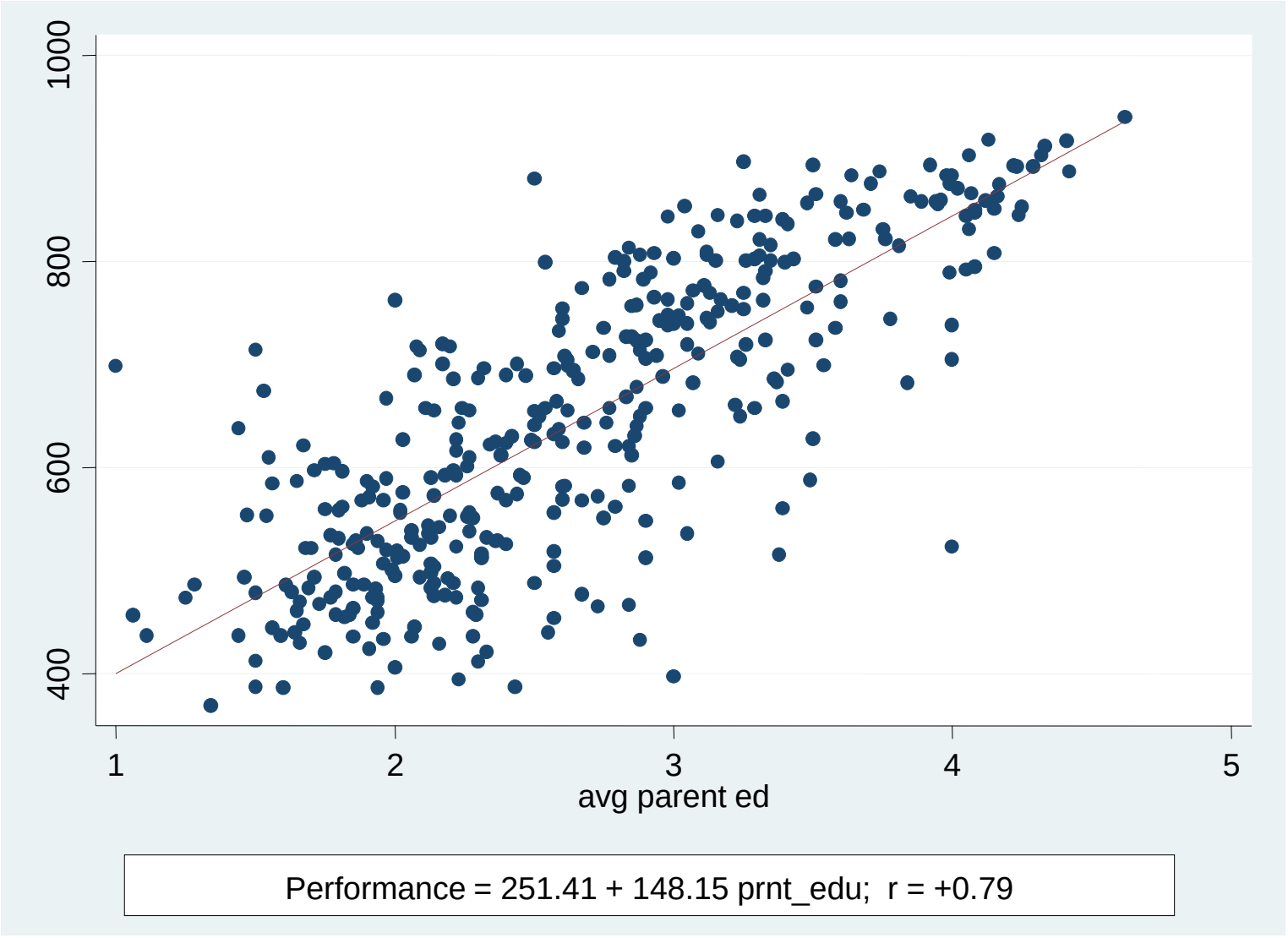


Assuming linear relationship between the variables, let's compute the Pearson's correlation coefficient

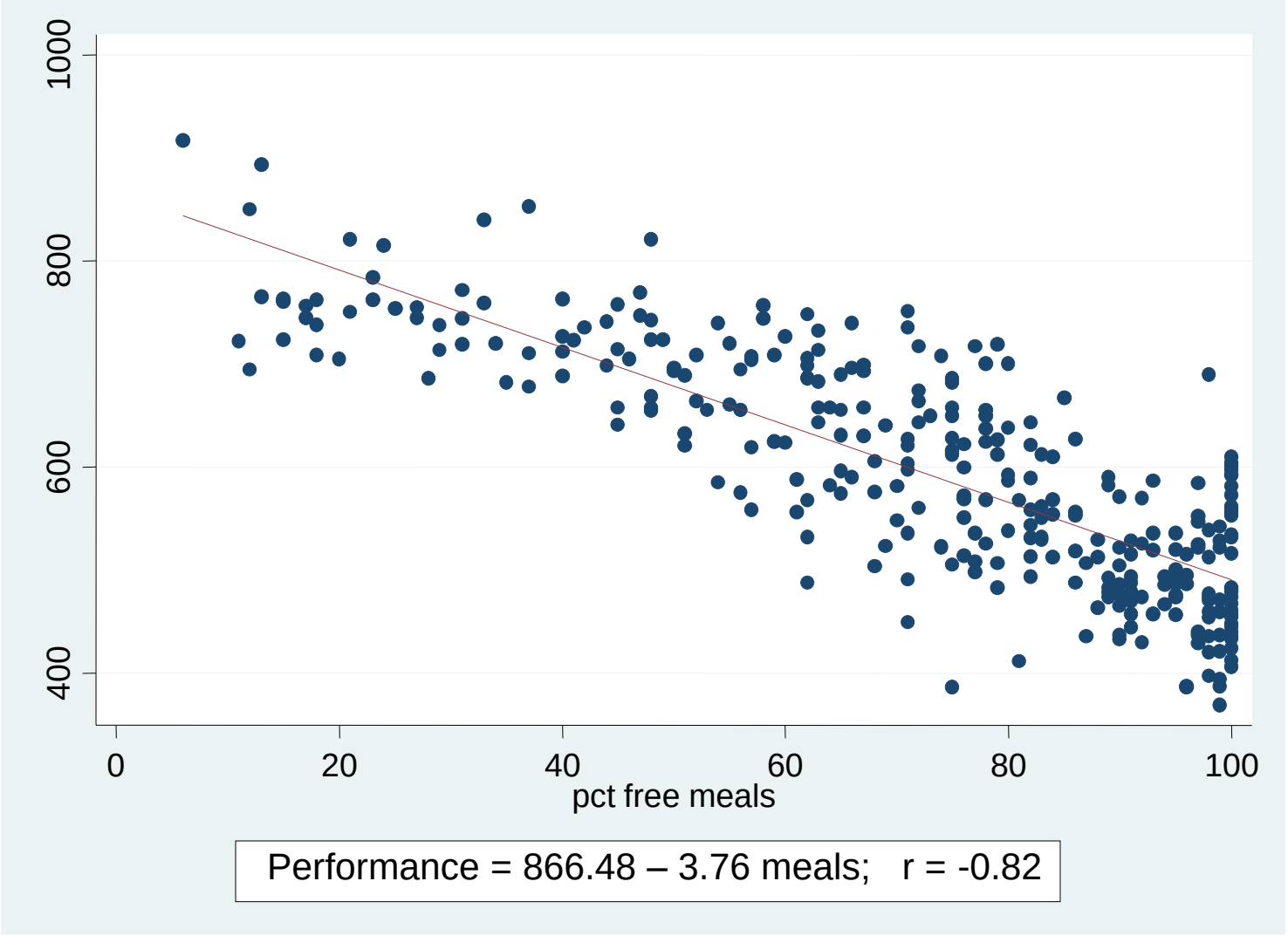
Stata: `corr wt height`

r=0.6043

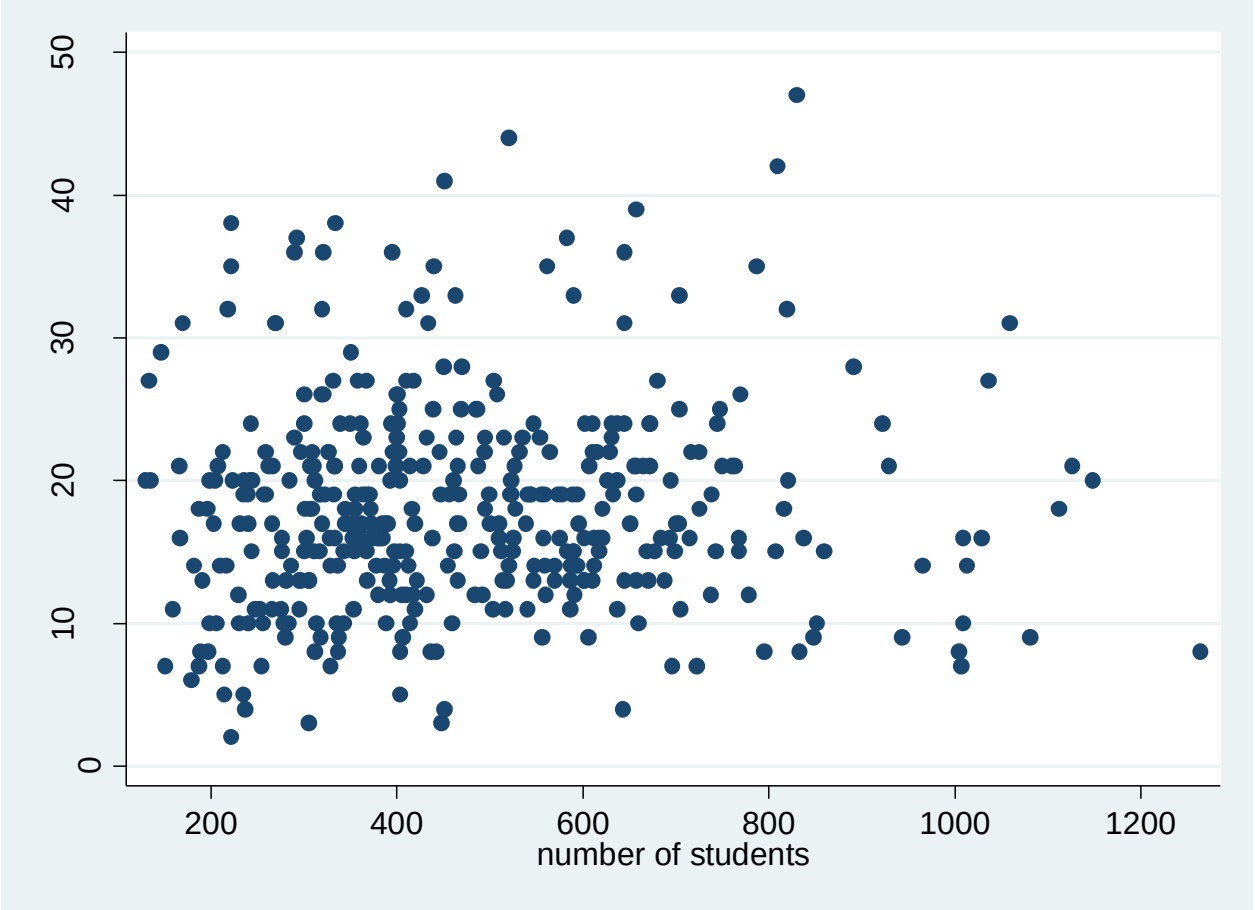
Example of positive correlation and slope



Example of negative correlation and slope



Any relationship?



r = 0?

Linear Regression Model

Let Y denote a outcome variable and X denote an explanatory variable, then the simplest mathematical expression for how Y depends on X can be expressed as a linear function:

$$Y = (\alpha + \beta X) + \varepsilon$$

{ Outcome = Systematic variation + Residual variation }

where α is the y-intercept

β is the slope and

ε is the residual of the model and assumed to be **independently and normally** distributed with mean 0 and standard deviation σ [i.e. $\varepsilon \sim N(0,\sigma^2)$]

This formula says that Y is a linear function of the explanatory variable X with a slope of β and a Y-intercept α.

For a given set of data in which Y and X appear to be linearly related, we can estimate the unknown regression coefficients α and β by minimizing the error of prediction.

For the sample data, estimate the model by estimating what's called a predictive equation:

$$\hat{Y} = a + b\,X$$

where *a* and *b* are estimates of α and β, respectively

Fitted multiple regression model can be expressed as:

$$\hat{Y} = a + b_1 X_1 + b_2 X_2 + ... + b_k X_k$$

where X_1, X_2,..., X_k are explanatory variables (IVs)

Consider a linear model: *Y= 3 + 2 X*

If X=0, Y= 3 + 2 (0) = 3

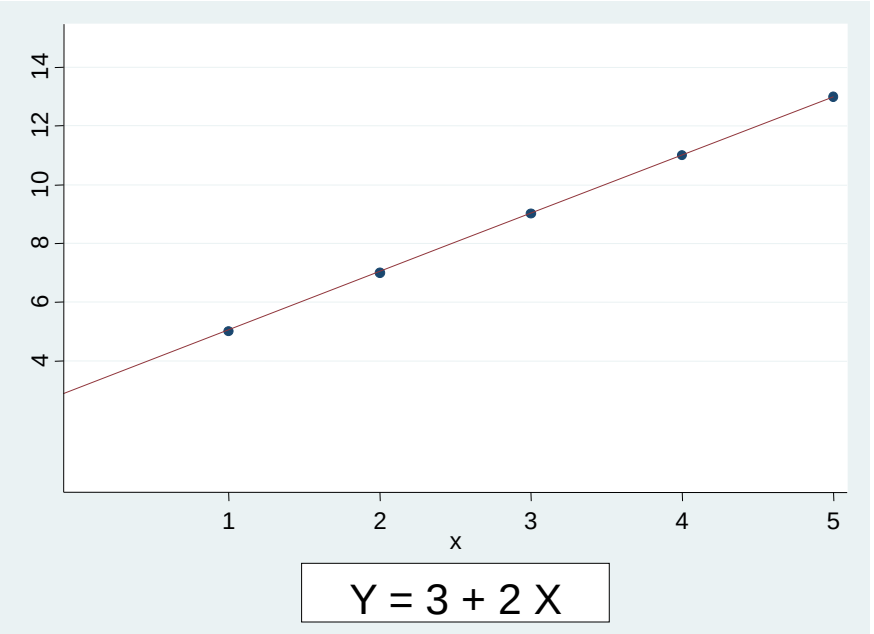
If X=1, Y= 3 + 2 (1) = 5

If X=2, Y= 3 + 2 (2) = 7

If X=3, Y= 3 + 2 (3) = 9

If X=4, Y= 3 + 2 (4) = 11

…



►1 unit increase in X results in 2 units increase in Y

►Positive slope implies upward slopping line

Consider another linear model: *Y= 3 - 2 X*

If X=0, Y= 3 - 2 (0) = 3

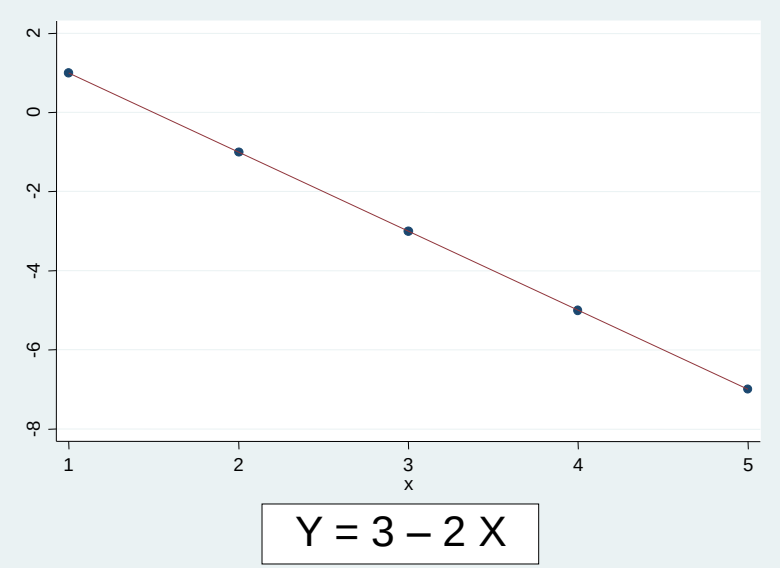
If X=1, Y= 3 - 2 (1) = 1

If X=2, Y= 3 - 2 (2) = -1

If X=3, Y= 3 - 2 (3) = -3

If X=4, Y= 3 - 2 (4) = -5

…



►1 unit increase in X results in 2 units decrease in Y

►Negative slope implies downward slopping line

Model Building using Working Examples

Let's consider an example where we are interested in modeling the relationship between scores on various tests and different socio-demographic variables.

The data file (**hsb2.dta**) used for this exercise contains 200 observations from a sample of high school students with demographic information about the students, such as their gender (**`female`**), socio-economic status (**`ses`**) and ethnic background (**`race`**). It also contains a number of scores on standardized tests, including tests of reading (**`read`**), writing (**`write`**), mathematics (**`math`**) and social studies (**`socst`**).

To load ***hsb2.dta*** from the web, type the ***Stata*** command:

```
use http://www.ats.ucla.edu/stat/stata/notes/hsb2, clear
```

To learn more about contents of the dataset, type:

Stata: **`describe`**

To learn more about particular variables, type:

Stata: **`codebook female science`**

To obtain frequencies of **`female`**, type:

Stata: **`tab1 female`**

To obtain descriptive statistics of **`science`** score, type:

Stata: **`sum science, detail`**

To look at the first 5 cases of some selected variables, type:

Stata: **`list science math ses female in 1/5`**

Research question?

Suppose we want to investigate how scores in mathematics (**`math`**) and gender (**`female`**) are associated with the science scores (**`science`**)

For regression modeling, let's consider **`science`** as response variable, and **`math`** and **`female`** as explanatory variables

Distribution of `science` score

To look at the histogram of **`science`** along with normal and kernel density plots, type:

Stata: **`hist science, normal kdensity`**

To look at the boxplot of **`science`**, type:

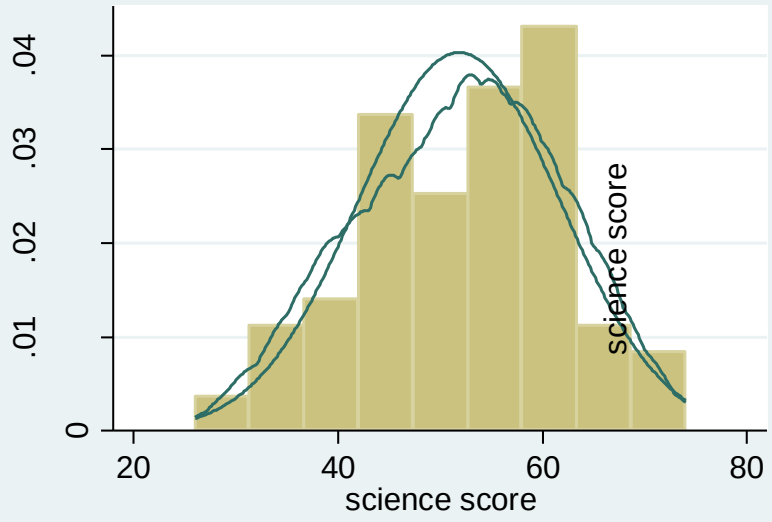
Stata: **`graph box science`**

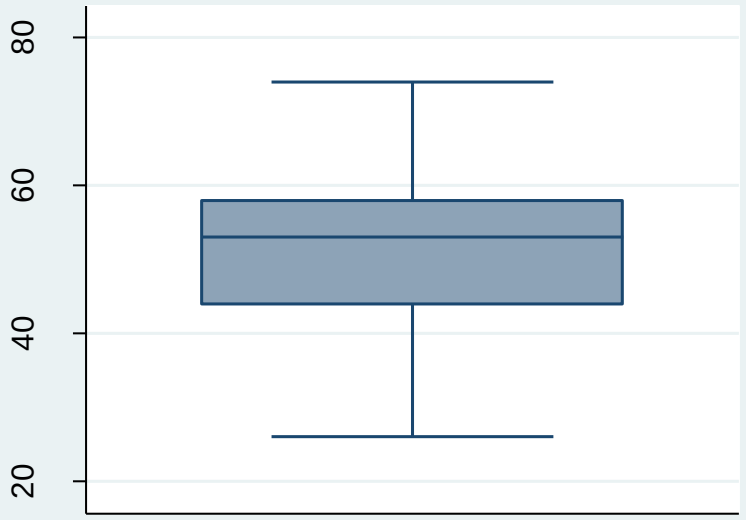
To check the normality of **`science`** using Q-Q plot, type:

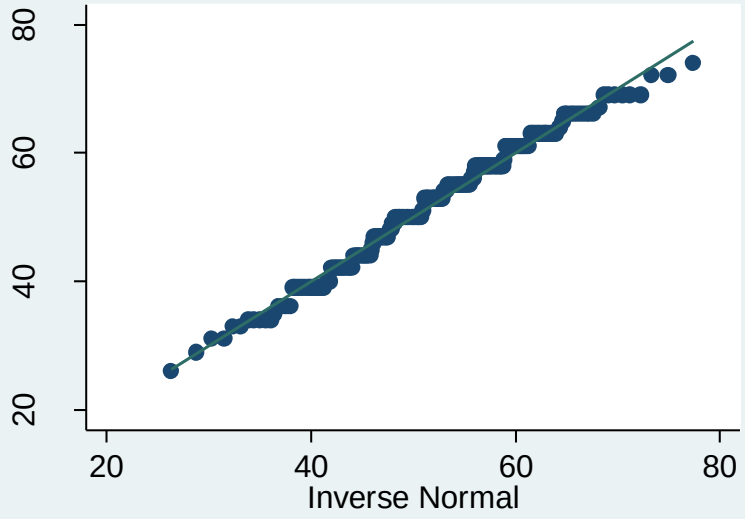
Stata: **`qnorm science`**

To check the normality of **`science`** using P-P plot, type:

Stata: **`pnorm science`**







Do these graphs suggest that science scores are normally distributed?

Exploring normality using statistical tests

To test the normality of **`science`** using a skewness/kurtosis test, type:

Stata: **`sktest science`**

To obtain Shapiro-Wilk W test for normality, type:

Stata: **`swilk science`**

To obtain Shapiro-Francia W' test for normality, type:

Stata: **`sfrancia science`**

Be aware that in these tests, the null hypothesis states that the variable (e.g. **`science`**) is normally distributed

Examining the relationships

To draw a scatter plot of **`science`** and **`math`**, type:

Stata: **`scatter science math`**

To draw a scatter plot of **`science`** and **`math`** along with a fitted line, type:

Stata: **`scatter science math ||lfit science math`**

To compute Pearson's correlation coefficient, type:

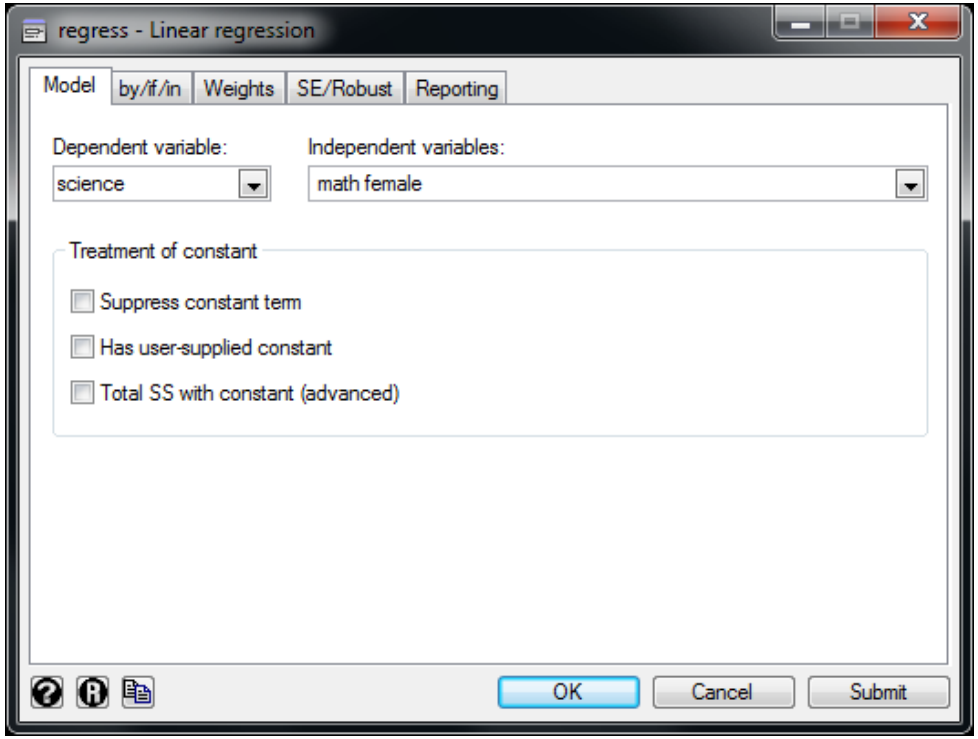
Stata: **`corr science math`**

To find how **`science`** regress on **`math`** and **`female`**, we can use **`regres`** command to run a linear regression

Stata: **`regress science math female`**

- If you want to use dialogue box to run your analysis, you can do so by using a Stata command '**db**'
- To obtain a dialogue box to run linear regression, type:

  Stata: **db regress**
- You then need to select dependent and independent variables from drop-down menu to run the model



Regression analysis output

Stata: `regress science math female`

Analysis of variance

Fit statistics

Source	SS	df	MS
Model	7993.54995	2	3996.77498
Residual	11513.95	197	58.4464469
Total	19507.5	199	98.0276382

Number of obs	= 200
F(2, 197)	= 68.38
Prob > F	= 0.0000
R-squared	= 0.4098
Adj R-squared	= 0.4038
Root MSE	= 7.645

science	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
math	.6631901	.0578724	11.46	0.000	.549061	.7773191
female	-2.168396	1.086043	-2.00	0.047	-4.310159	-.0266329
_cons	18.11813	3.167133	5.72	0.000	11.8723	24.36397

Coefficients estimates

Interpretations:

$F_{2,197}$=68.38, p<0.0001 suggests that the math scores and female significantly associated with the science scores

One measure of variability explained by the regression model can take the value of model SS as a proportion of the total variation in Y in the data, what is called ***coefficient of determination***:

$$R^2 = \frac{Model\ SS}{Total\ SS} = \frac{7993.54995}{19507.5} = 0.4098$$

► About 41% of total variability in science scores is explained by the linear regression model (i.e. by math scores and gender)

The estimated regression line is:

science=18.118+0.663*math-2.168*female

- for each unit increases in math score, the predicted science score would be increased by 0.663 units, holding the effect of gender constant
  - the effect of math scores on science scores is significantly different from zero (t_{197}=11.46, p<0.0001)

- predicted science score would be 2.168 points lower for female than male students, holding the effect of math constant
  - there is marginal evidence to suggest that the effect of gender on science scores is significantly different from zero (t_{197}=2.00, p=0.045)

Regression Diagnostics

Once we fit the regression model, next step would be to check the assumptions of the model to ensure whether the fitted model is adequate (post-estimation in Stata)

Multicolinearity?

To examine multicollinearity among the explanatory variables, type a post-estimation command:

Stata: `estat vif`

Homoscedasticity?

Constant variance, homoscedasticity, can be checked by plotting residuals against fitted values :

Stata: `rvfplot, yline(0)`

Stata: `hettest`

Normality of residuals?

We can calculate the residuals: $e = (Y - \hat{Y})$ using a post-estimation command:

Stata: `predict e, residual`

To check ***normality*** of residuals using a Q-Q plot:

Stata: `qnorm e`

To draw a kernel density plot with normal curve

Stata: `kdensity e, normal`

We can also examine residuals by using statistical tests

Outliers?

To identify possible outlier(s), we need to calculate studentized residuals (r):

Stata: `predict r, rstudent`

To examine studentized residuals that exceed +2 or -2,

Stata: `list id r if abs(r)>2`

To examine studentized residuals that exceed +3 or -3,

Stata: `list id r if abs(r)>3`

Once we decide on the outliers, we can rerun the model by excluding students with large studentized residuals (say, r>±2):

Stata: `regress science math female if abs(r)<=2`

Comparison of the two models

Variable	Original model (n=200)			Revised model (n=190)		
	Reg-coeff	SE	P-value	Reg-coeff	SE	P-value
Math	0.663	0.0579	0.0001	0.692	0.0510	0.0001
Female	-2.168	1.0860	0.045	-1.807	0.9620	0.062
Adj-R^2	0.4038			0.4987		

Observations??

Which model would you like to consider?

Before making a decision about the final model, we need to make sure that the associated assumptions are met

If not, we may wish to rerun the regression to examine the adequacy of the revised fitted model until all assumptions are met

Regression model with categorical IV (>2 cat)

[e.g. SES (1=low, 2=medium, 3=high)]

Stata: `xi: regress science math female i.ses`

```
i.ses             _Ises_1-3           (naturally coded; _Ises_1 omitted)

      Source |       SS       df       MS              Number of obs =     200
-------------+------------------------------           F(  4,   195) =   35.38
       Model |  8203.18844     4  2050.79711            Prob > F      =  0.0000
    Residual |  11304.3116   195  57.9708285            R-squared     =  0.4205
-------------+------------------------------           Adj R-squared =  0.4086
       Total |    19507.5   199  98.0276382            Root MSE      =  7.6139

------------------------------------------------------------------------------
     science |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
        math |   .6329366   .0598967    10.57   0.000     .5148081    .7510652
      female |  -1.895787    1.09374    -1.73   0.085    -4.052866    .2612914
     _Ises_2 |   1.745932    1.38343     1.26   0.208    -.9824743    4.474339
     _Ises_3 |   2.971343   1.564739     1.90   0.059    -.1146412    6.057328
       _cons |    17.87125   3.232592     5.53   0.000     11.49592    24.24658
------------------------------------------------------------------------------
```

Getting Help

Within Stata:

- To get help for a particular command (e.g. regression)

  ```
  help regression
  ```

- To obtain all references to a topic (e.g. logistic)

  ```
  search logistic
  ```

- To find relevant commands on a topic (e.g. anova)

  ```
  findit anova
  ```

Online Stata support : www.stata.com/support

AU/NZ distributor for ***Stata*** & ***StatTransfer***

www.survey-design.com.au

- Stata **GradPlan** arrangements for students

Thank You

Comments
Questions?