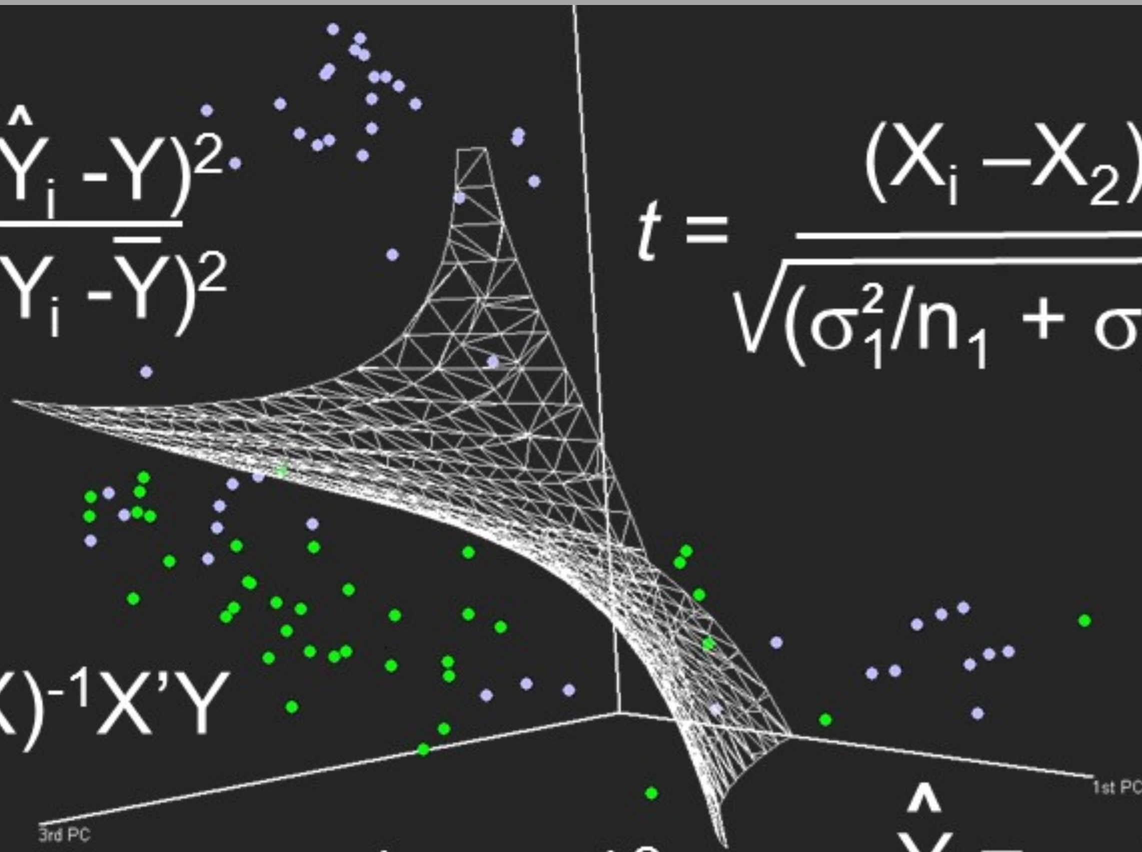
Advanced Methods of Analysis

Lean Six Sigma Black Belt

$$R^2 = \frac{\Sigma(\hat{Y}_i - Y)^2}{\Sigma(Y_i - \bar{Y})^2}$$

$$t = \frac{(X_i - X_2)}{\sqrt{(\sigma_1^2/n_1 + \sigma_2^2/n_2)}}$$



$$\hat{\beta} = (X'X)^{-1}X'Y$$

$$\hat{Y} = \alpha + \beta\bar{X}$$

$$\chi^2 = \Sigma\left[\frac{(f_o - f_e)^2}{f_e}\right]$$

Woody Bendle

Advanced methods of analysis & Innovation

Today's Topics

Regression

- Types of Regression / Data
  - Cross Section
  - Time Series
  - Univariate
  - Multivariate

Cluster Analysis

- K-Means

Innovation

- Introduction to the i^3 Innovation Process

Advanced methods of analysis

Regression

Definition:

Regression analysis is a statistical process for estimating the relationships among variables *(data)*

Advanced methods of analysis

Regression

Definition:

Regression analysis is a statistical process for estimating the relationships among variables *(data)*

- How does one thing affect another? *Univariate*

Advanced methods of analysis
Regression

Definition:

Regression analysis is a statistical process for estimating the relationships among variables *(data)*

- How does one thing affect another? *Univariate*
- How do several things together affect something else? *Multivariate*

Advanced methods of analysis

Regression

Definition:

Regression analysis is a statistical process for estimating the relationships among variables *(data)*

- How does one thing affect another? *Univariate*
- How do several things together affect something else? *Multivariate*
- What is the relationship over time? *Time Series*

Advanced methods of analysis
Regression

Theoretical

Hypothesis:
“I think ____ affects ____ in some way”

I think the amount of rain affects how much the grass grows

Vs.

A Theoretical

I have a bunch of data, let’s dump it into the machine and see what pops out…

Advanced methods of analysis

Regression

***Theoretical* approach**

Pros:

You’re approaching a problem with a belief construct

Scientific Method

Cons:

You’re approaching a problem with a limited belief construct

Advanced methods of analysis
Regression

A Theoretical **approach**

Pros:

Dispassionate

Avoids *"Curse of Knowledge"*

Can find relationships that challenge convention

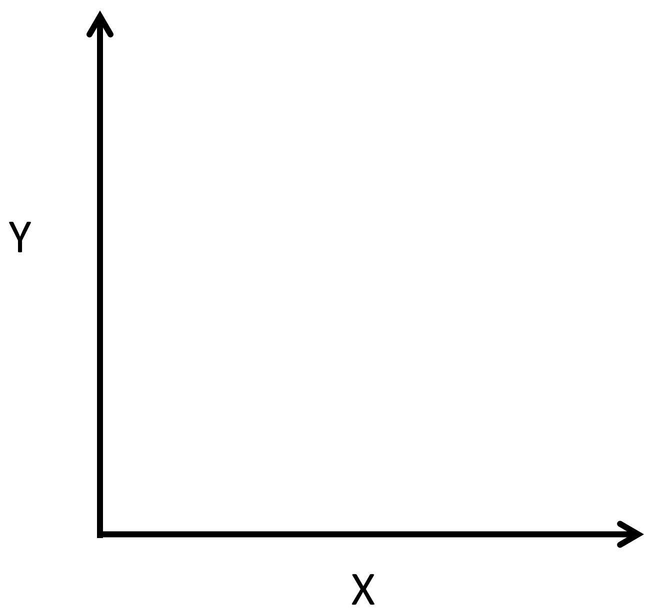
Cons:

You're approaching a problem with a limited belief construct

Results can lead to bad / wrong conclusions

Advanced methods of analysis
Univariate Regression

- Suspected relationship between two things
- Often shown using X-Y Graph



Y "Dependent Variable"

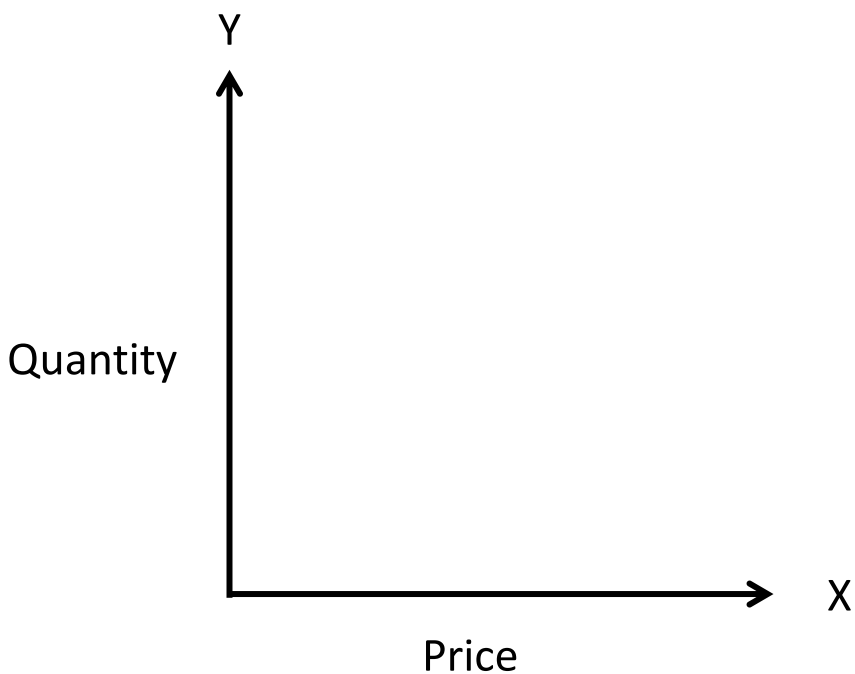
X "Independent Variable"

$Y = f(X)$

Advanced methods of analysis
Univariate Regression

Common Univariate Examples

Demand Function in Economics



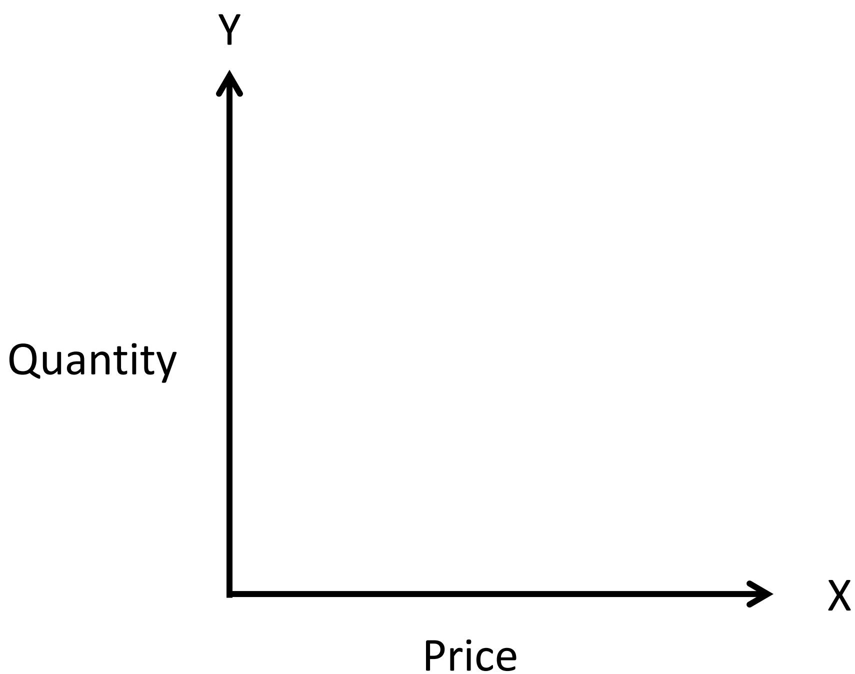
$$Y = f(X)$$

$$\text{Quantity} = f(\text{Price})$$

Advanced methods of analysis
Univariate Regression

Common Univariate Examples

Demand Function in Economics



$Y = f(X)$

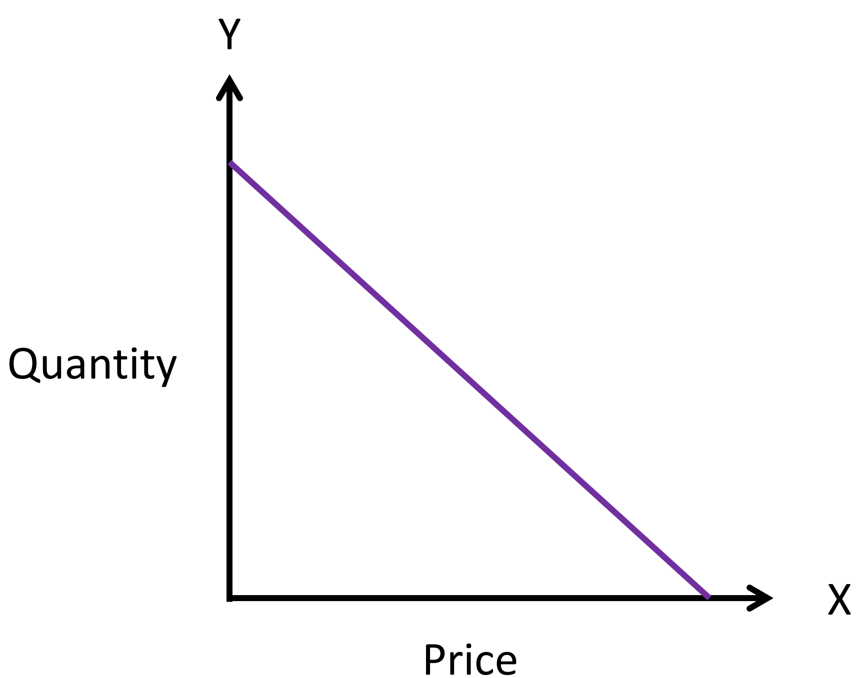
$$\underset{(\,-\,)}{\text{Quantity} = f(\text{Price})}$$

Negative or Inverse Relationship

Advanced methods of analysis
Univariate Regression

Common Univariate Examples

Demand Function in Economics



$Y = f(X)$

$\text{Quantity} = f(\text{Price})$
(-)

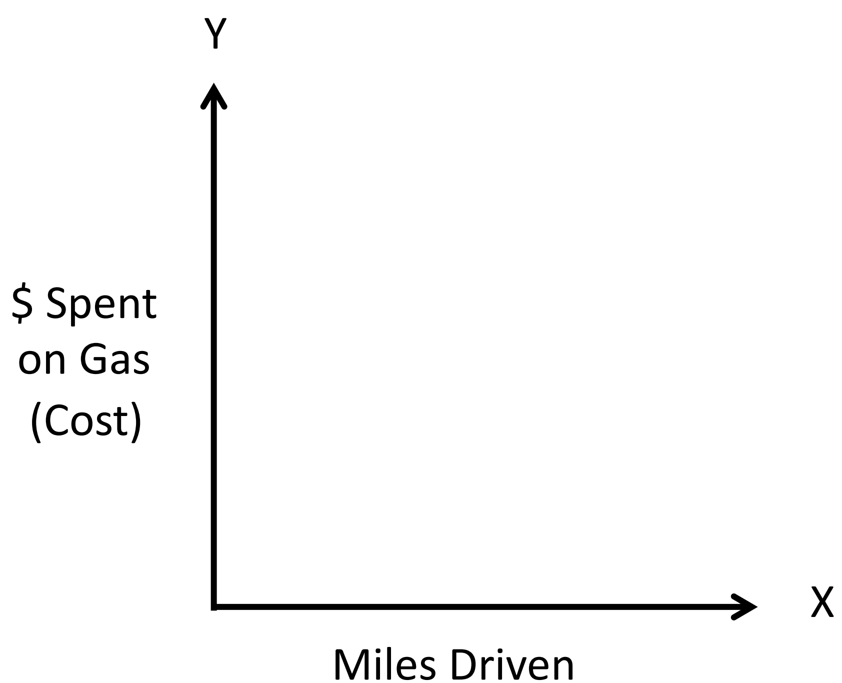
Negative or Inverse Relationship

Advanced methods of analysis

Univariate Regression

Common Univariate Examples

Cost Function



$Y = f(X)$

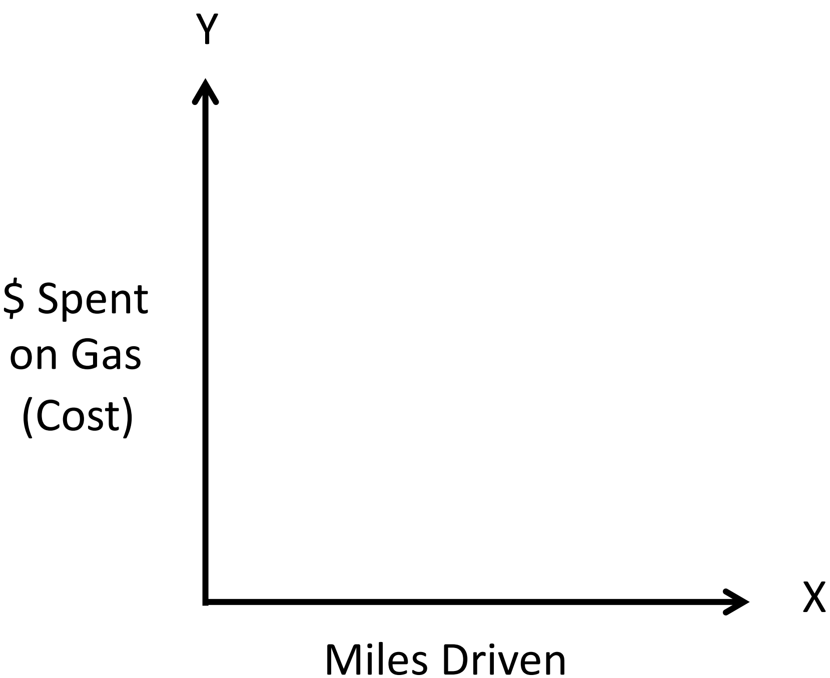
\$ Spent on Gas = f(Miles Driven)

Advanced methods of analysis

Univariate Regression

Common Univariate Examples

Cost Function



$Y = f(X)$

\$ Spent on Gas = f(Miles Driven)

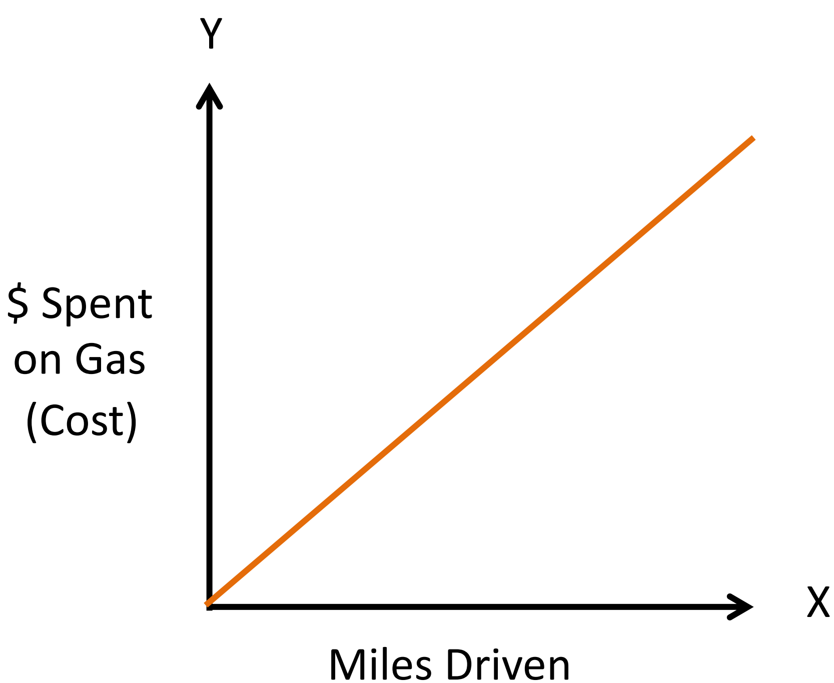
(+)

Positive Relationship

Advanced methods of analysis
Univariate Regression

Common Univariate Examples

Cost Function



$Y = f(X)$

$ Spent on Gas = f(Miles Driven)
(+)

Positive Relationship

Advanced methods of analysis

Univariate Regression

Common Univariate Examples

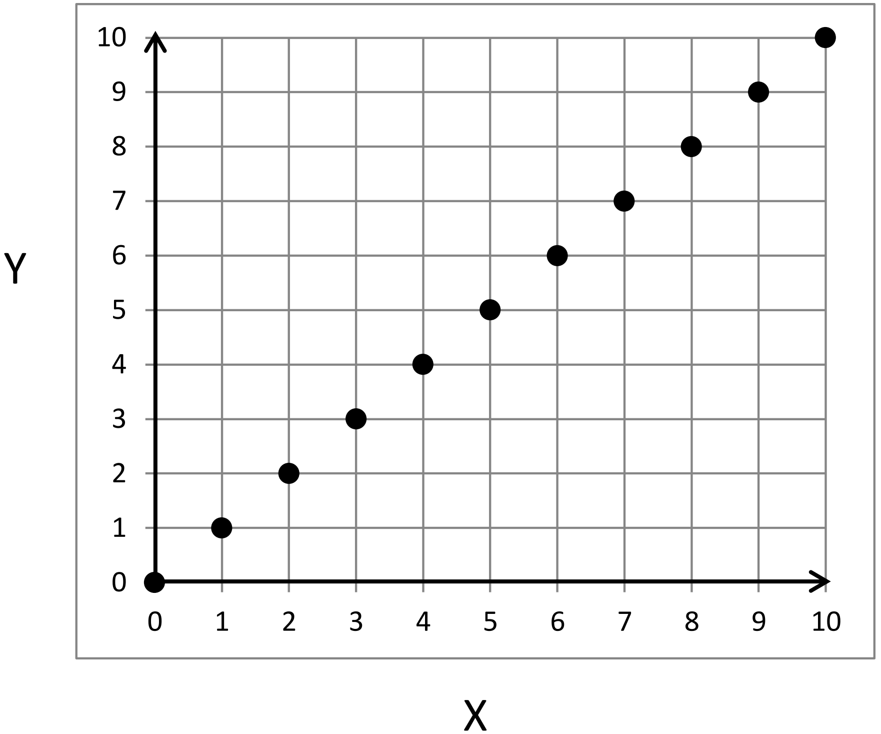
What are some other examples of these types of univariate relationships?

-
-
-
-
-
-
-

Advanced methods of analysis
Univariate Regression

Viewing / Visualizing Data

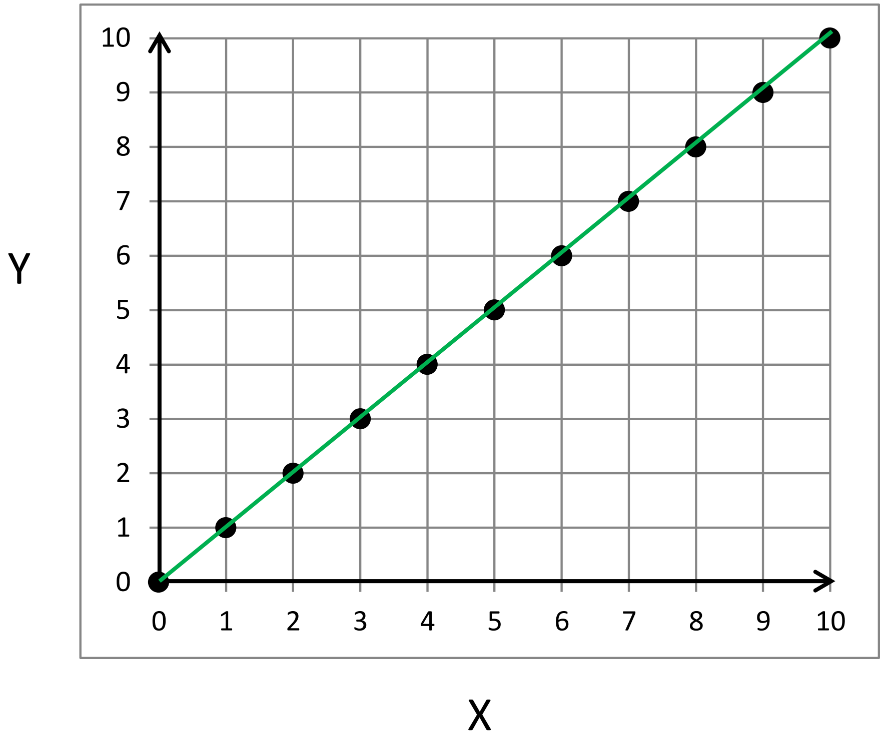
Y	**X**
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Advanced methods of analysis
Univariate Regression

Viewing / Visualizing Data

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10

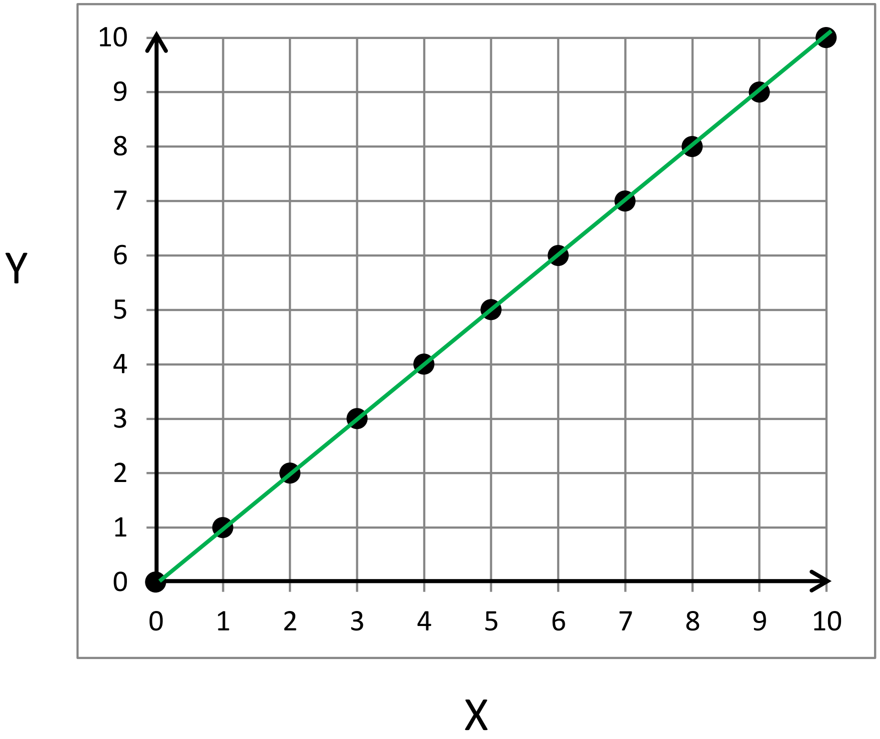


Advanced methods of analysis

Univariate Regression

Equation for a line

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



$$Y = mX + b$$

m = slope of the line
= "rise over run"

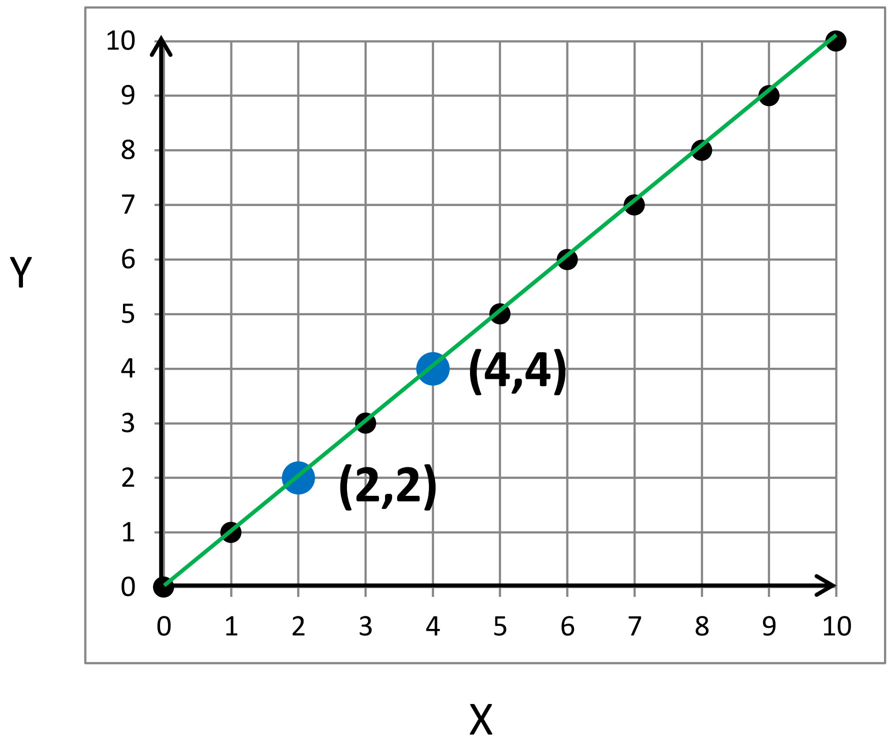
b = Y intercept

Advanced methods of analysis

Univariate Regression

Equation for a line $Y = mX + b$

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Solving for “m” if we know two (X,Y) points

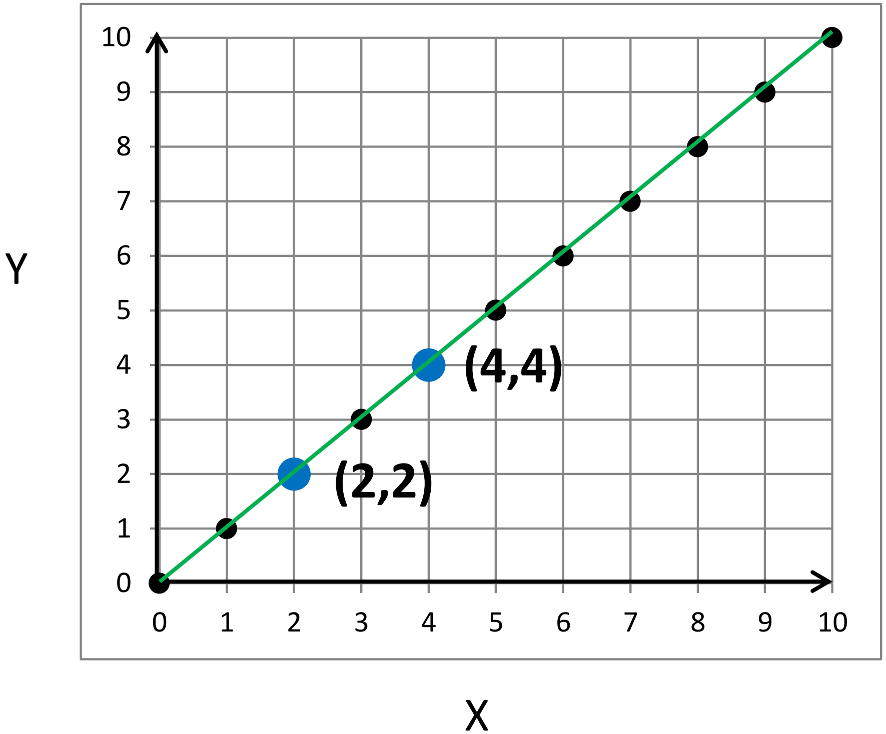
$$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

Advanced methods of analysis

Univariate Regression

Equation for a line $Y = mX + b$

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Solving for "m" if we know two (X,Y) points

$$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

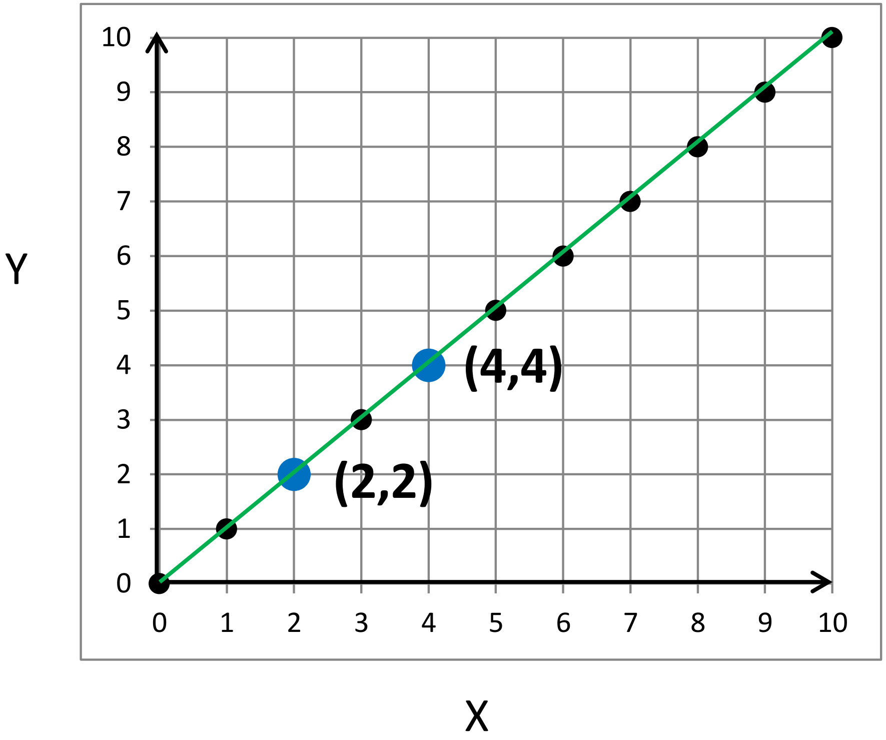
$$m = \frac{4 - 2}{4 - 2}$$

Advanced methods of analysis

Univariate Regression

Equation for a line $\quad Y = mX + b$

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Solving for "m" if we know two (X,Y) points

$$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

$$m = \frac{4 - 2}{4 - 2}$$

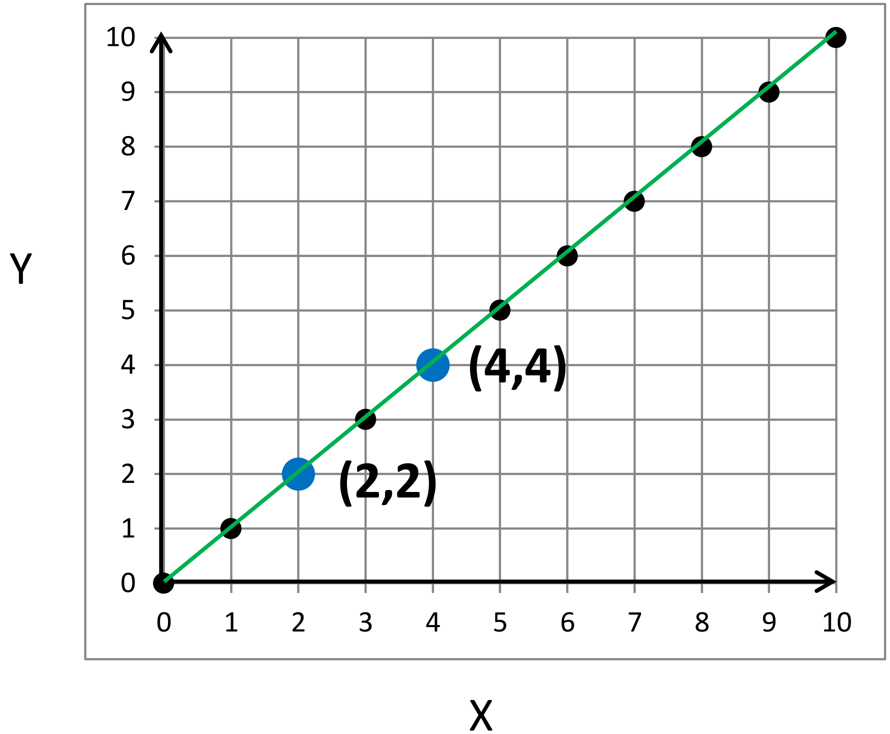
$$m = \frac{2}{2}$$

Advanced methods of analysis

Univariate Regression

Equation for a line $\quad Y = mX + b$

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Solving for "m" if we know two (X,Y) points

$$m = \frac{Y_2 - Y_1}{X_2 - X_1}$$

$$m = \frac{4 - 2}{4 - 2}$$

$$m = \frac{2}{2}$$

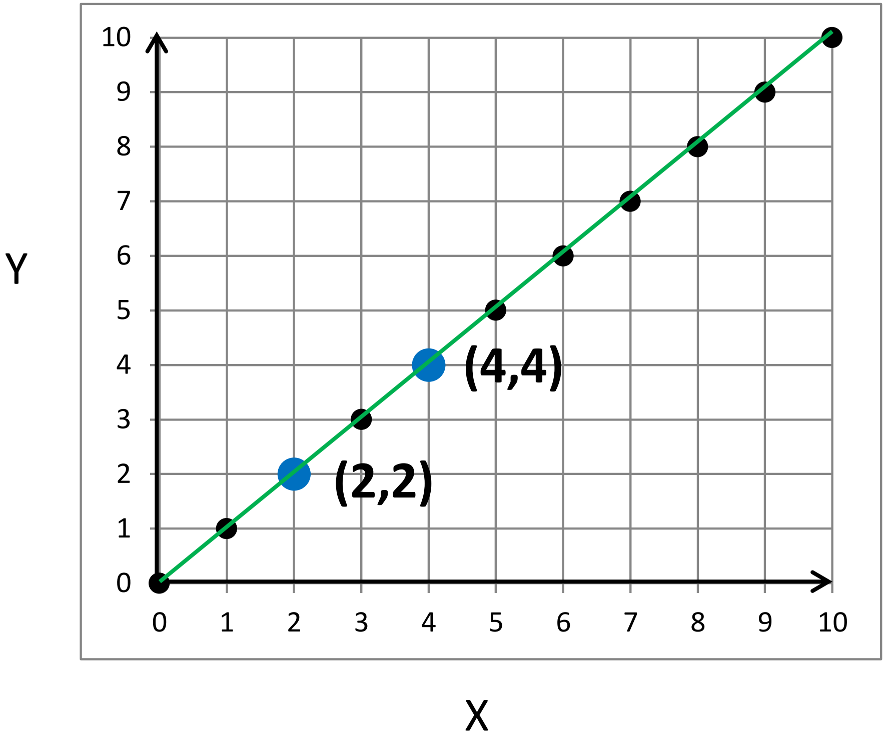
$$m = 1$$

Advanced methods of analysis

Univariate Regression

Equation for a line $Y = mX + b$

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Now solving for "b"

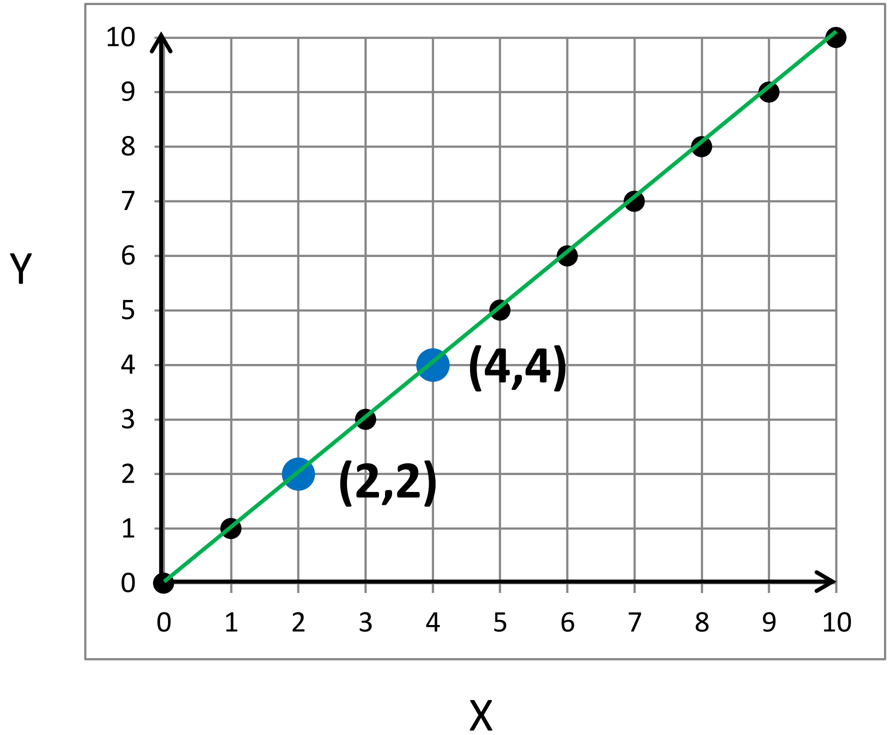
$Y = mX + b$

Advanced methods of analysis

Univariate Regression

Equation for a line $Y = mX + b$

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Now solving for “b”

$Y = mX + b$

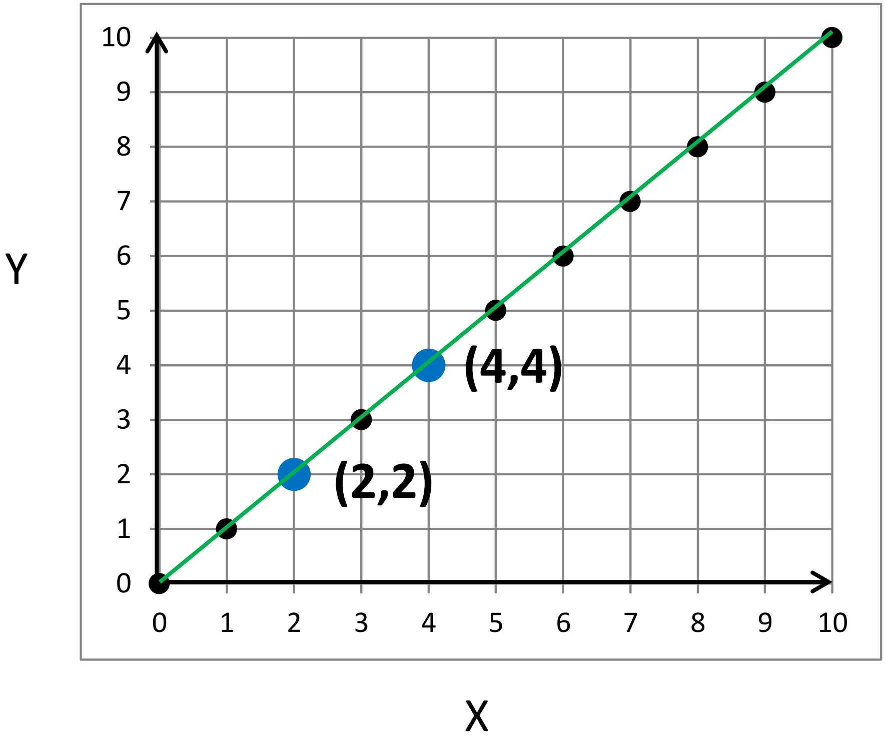
$Y = 1(X) + b$

Advanced methods of analysis

Univariate Regression

Equation for a line $Y = mX + b$

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Now solving for "b"

$Y = mX + b$

$Y = 1(X) + b$

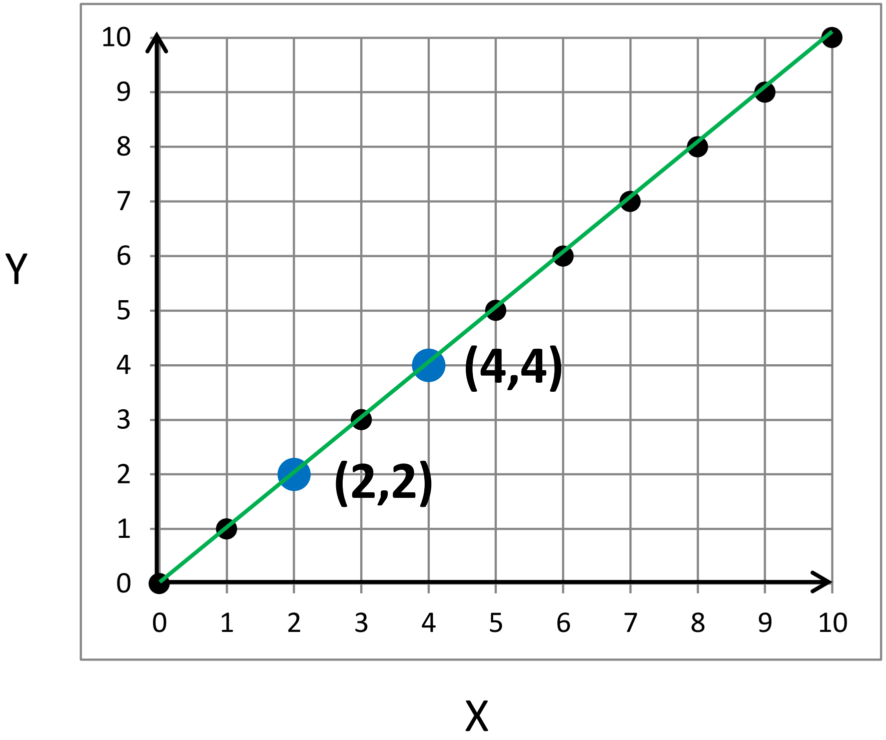
$2 = 1(2) + b$

Advanced methods of analysis

Univariate Regression

Equation for a line $Y = mX + b$

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Now solving for "b"

$Y = mX + b$

$Y = 1(X) + b$

$2 = 1(2) + b$

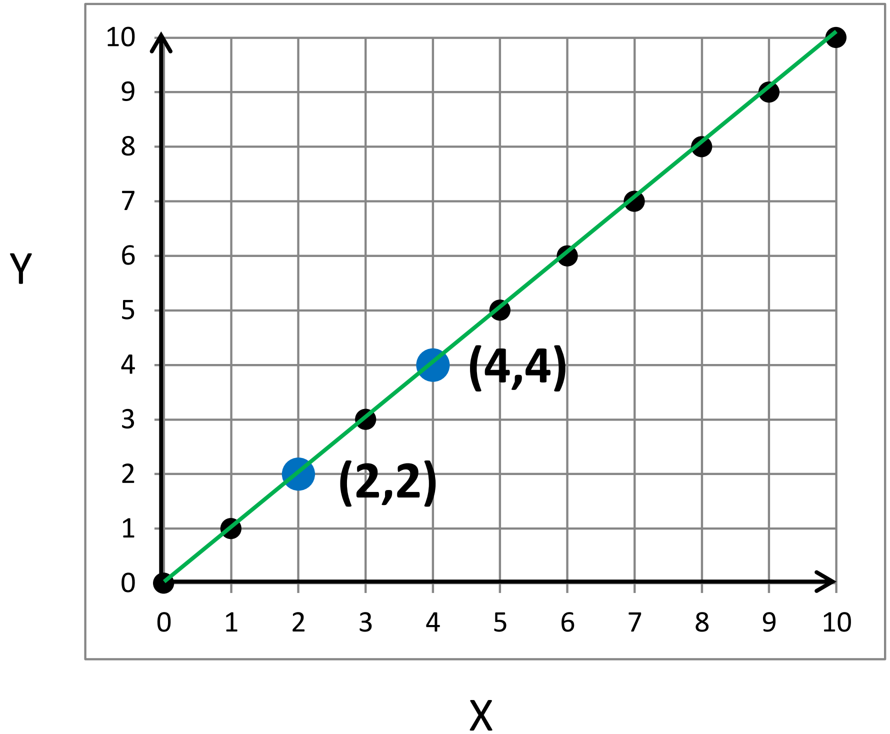
$2 = 2 + b$

Advanced methods of analysis

Univariate Regression

Equation for a line $\qquad Y = mX + b$

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Now solving for "b"

$Y = mX + b$

$Y = 1(X) + b$

$2 = 1(2) + b$

$2 = 2 + b$

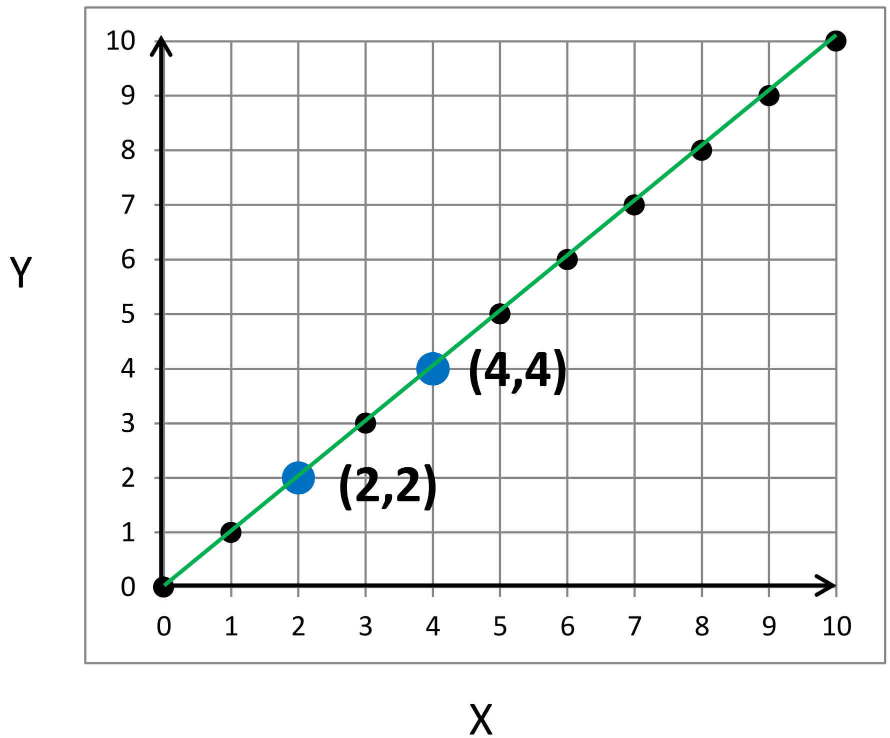
$(2 - 2) = (2 - 2) + b$

Advanced methods of analysis

Univariate Regression

Equation for a line $Y = mX + b$

Y	X
0	0
1	1
2	2
3	3
4	4
5	5
6	6
7	7
8	8
9	9
10	10



Now solving for "b"

$Y = mX + b$

$Y = 1(X) + b$

$2 = 1(2) + b$

$2 = 2 + b$

$(2 - 2) = (2 - 2) + b$

$0 = b$

Advanced methods of analysis

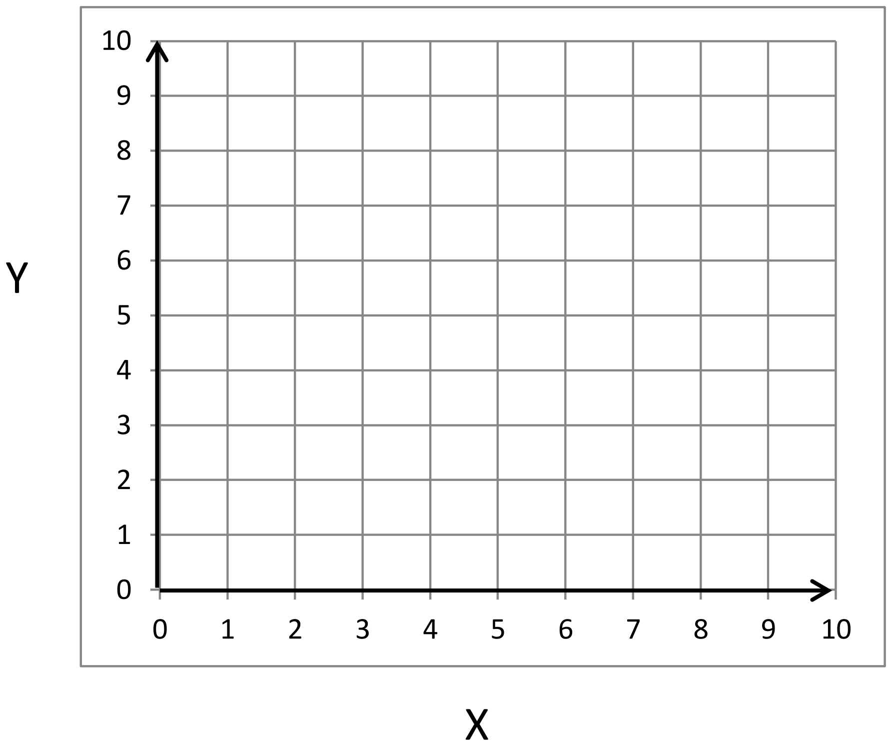
Univariate Regression

Exercise:

- Plot (X,Y) coordinates
- Using Line Equation, solve for **"m"** and **"b"**

Y	X
3	10
4	8
5	6
6	4
7	2

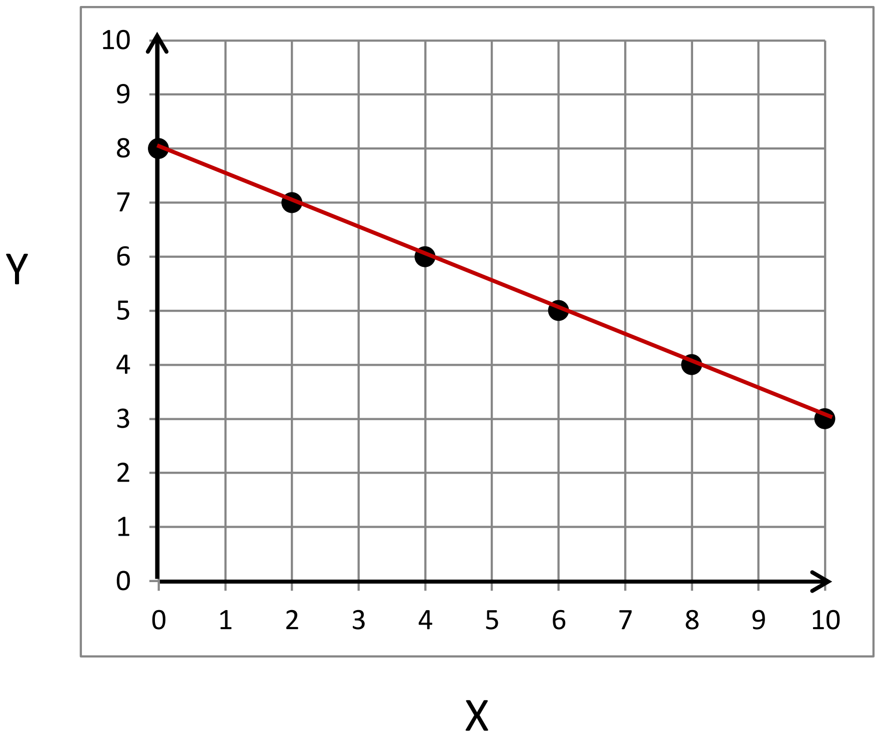
$$Y = mX + b$$



Advanced methods of analysis
Univariate Regression

Exercise:
- Plot (X,Y) coordinates
- Using Line Equation, solve for **"m"** and **"b"**

Y	X
3	10
4	8
5	6
6	4
7	2



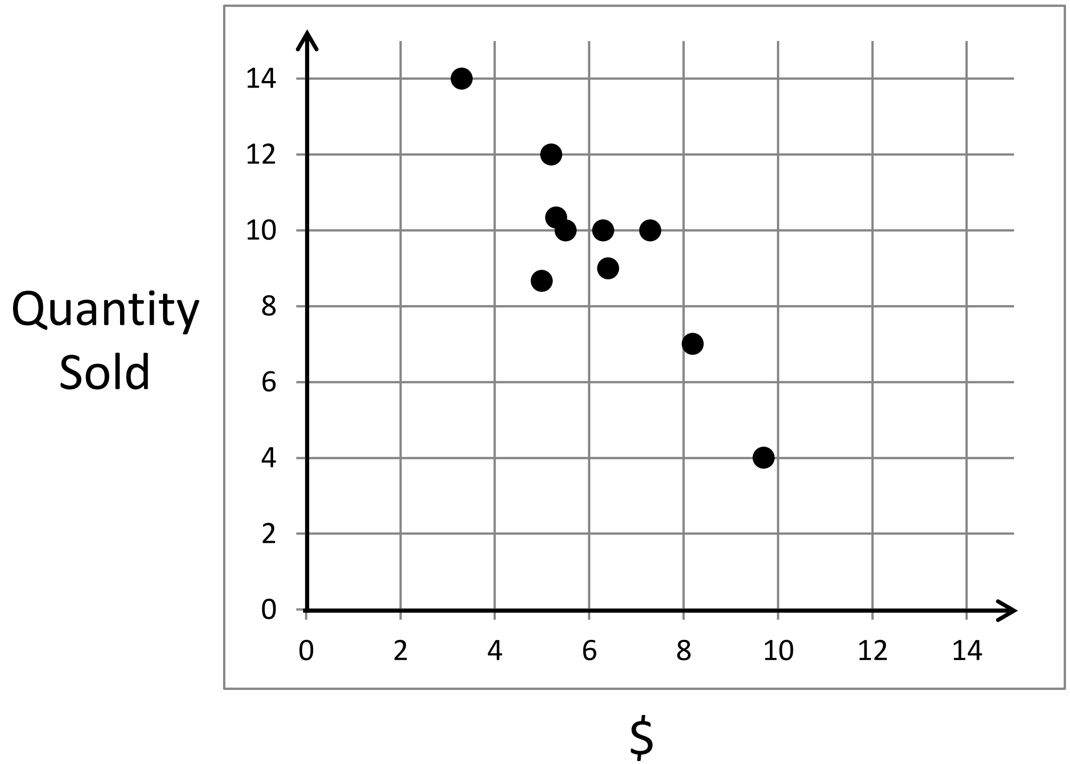
$$Y = mX + b$$

$$m = -0.5$$

$$b = 8$$

Real world data are rarely "perfect"

Obs	X $	Y # Sold
1	10	4
2	5	9
3	6	10
4	3	14
5	6	8
6	6	10
7	8	6
8	5	12
9	7	9
10	5	10

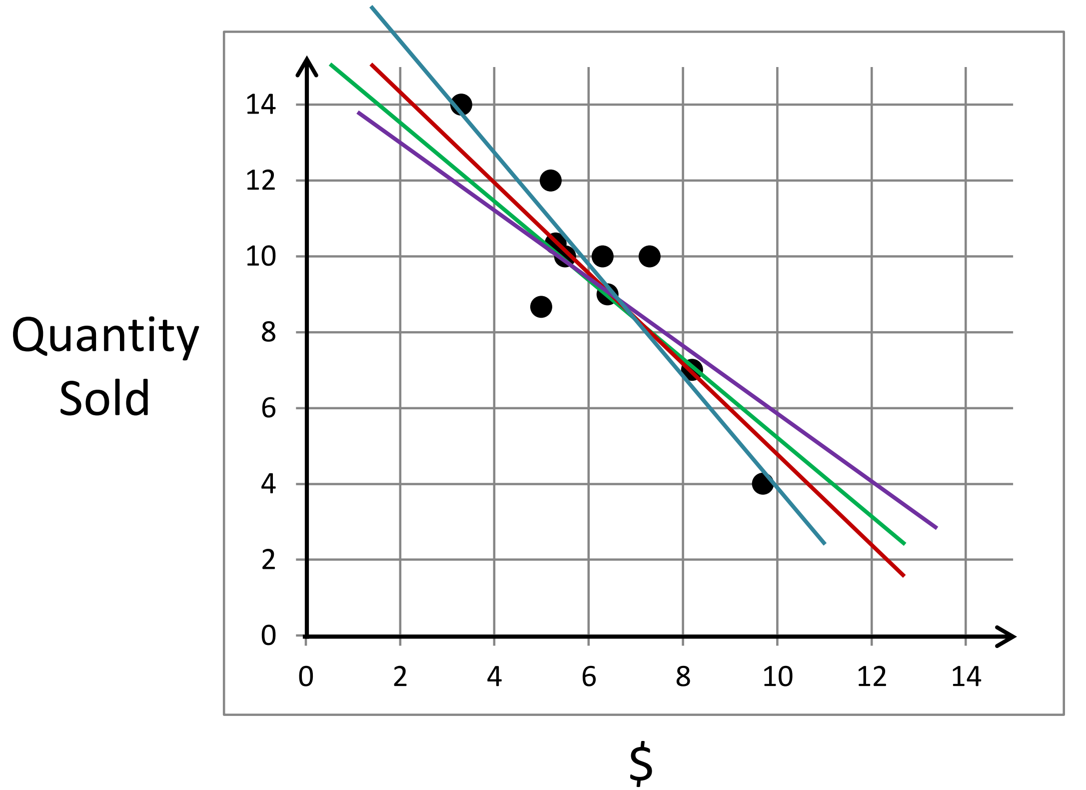


Advanced methods of analysis
Univariate Regression

Can draw a lot of lines through this scatter plot

How do you determine the “best” line?

Obs	X $	Y # Sold
1	10	4
2	5	9
3	6	10
4	3	14
5	6	8
6	6	10
7	8	6
8	5	12
9	7	9
10	5	10



Advanced methods of analysis

Univariate Regression

Equation for a line $Y = mX + b$

Regression Equation $Y = a + bX$

where… a = Y intercept

b = Slope

or… $Y = a + bX$

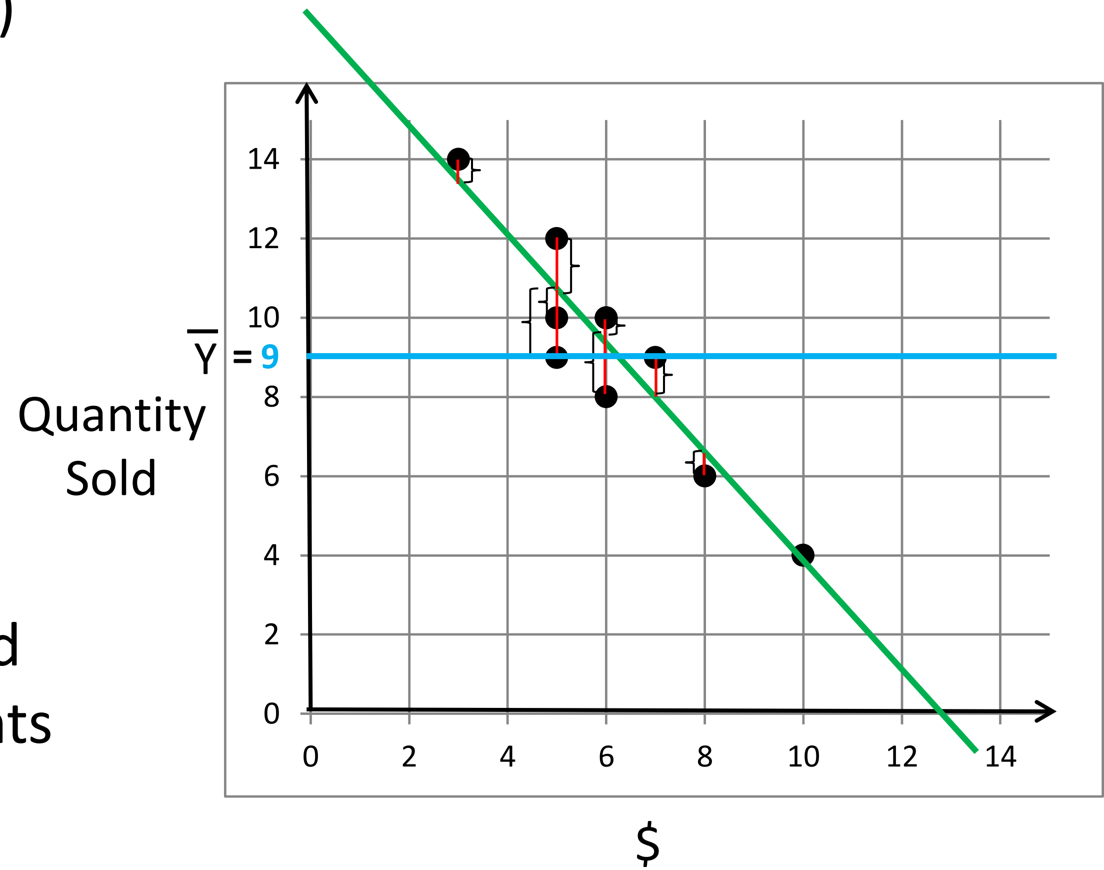
Advanced methods of analysis

Univariate Regression

Ordinary Least Squares (OLS)

$$\hat{Y} = a + bX$$

A good line is one that minimizes the sum of squared differences between the points and the line



Advanced methods of analysis

Univariate Regression

Ordinary Least Squares (OLS)

Some Equations…

$$\hat{Y} = a + bX$$

$$b = \frac{SS_{xy}}{SS_{xx}}$$

$$a = \bar{Y} - b\bar{X}$$

$\bar{Y}$ = mean of Y

$\bar{X}$ = mean of X

$$SS_{xy} = S\ X_i Y_i - \frac{(S\ X_i)\ (S\ Y_i)}{n}$$

$$SS_{xx} = S\ X_i^2 - \frac{(S\ X_i)^2}{n}$$

Advanced methods of analysis
Univariate Regression

OLS – Solving our example by hand

	Obs	X $	Y # Sold	X²	X x Y
	1	10	4	100	40
	2	5	9	25	45
	3	6	10	36	60
	4	3	14	9	42
	5	6	8	36	48
	6	6	10	36	60
	7	8	6	64	48
	8	5	12	25	60
	9	7	9	49	63
n =	10	5	10	25	50
	Sum	61	92	**Sum X^2** 405	516 **Sum X x Y**
	(Sum X)²	3,721			

Mean X	Mean Y
6	9

$$\hat{Y} = a + bX$$

$$= \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xy} = X_i Y_i - \frac{(X_i)(Y_i)}{n}$$

$$SS_{xy} = 516 - \frac{61 \times 92}{10} = \mathbf{-45.2}$$

Advanced methods of analysis

Univariate Regression

OLS – Solving our example by hand

	Obs	X $	Y # Sold	X^2	X x Y
	1	10	4	100	40
	2	5	9	25	45
	3	6	10	36	60
	4	3	14	9	42
	5	6	8	36	48
	6	6	10	36	60
	7	8	6	64	48
	8	5	12	25	60
	9	7	9	49	63
n =	10	5	10	25	50
	Sum	61	92	Sum X^2 405	516 Sum X x Y
	$(\text{Sum X})^2$	3,721			

Mean X	Mean Y
6	9

$$\hat{Y} = a + bX$$

$$= \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xy} = X_iY_i - \frac{(X_i)(Y_i)}{n}$$

$$SS_{xy} = 516 - \frac{61 \text{ x } 92}{10} = \mathbf{-45.2}$$

$$SS_{xx} = X_i^2 - \frac{(X_i)^2}{n}$$

$$SS_{xx} = 405 - \frac{3{,}721}{10} = \mathbf{32.9}$$

$$b = \frac{\mathbf{-45.2}}{\mathbf{32.9}} = \mathbf{-1.37}$$

Advanced methods of analysis

Univariate Regression

OLS – Solving our example by hand

	Obs	X $	Y # Sold	X^2	X x Y
	1	10	4	100	40
	2	5	9	25	45
	3	6	10	36	60
	4	3	14	9	42
	5	6	8	36	48
	6	6	10	36	60
	7	8	6	64	48
	8	5	12	25	60
	9	7	9	49	63
n =	10	5	10	25	50
	Sum	61	92	**Sum X^2** 405	516 **Sum X x Y**
	(Sum X)^2	3,721			

Mean X	Mean Y
6	9

$$\hat{Y} = a + bX$$

$$= \frac{SS_{xy}}{SS_{xx}}$$

$$SS_{xy} = \quad X_iY_i - \frac{(\quad X_i)(\quad Y_i)}{n}$$

$$SS_{xy} = 516 - \frac{61 \times 92}{10} = \mathbf{-45.2}$$

$$SS_{xx} = \quad X_i^2 - \frac{(\quad X_i)^2}{n}$$

$$SS_{xx} = 405 - \frac{3{,}721}{10} = \mathbf{32.9}$$

$$b = \frac{\mathbf{-45.2}}{\mathbf{32.9}} = \mathbf{-1.37}$$

$$= \bar{Y} - \quad \bar{X}$$

$$a = 9 - (-1.37 \times 6) = \mathbf{17.6}$$

$$\hat{Y} = 17.6 - 1.37 * X$$

Advanced methods of analysis

Univariate Regression

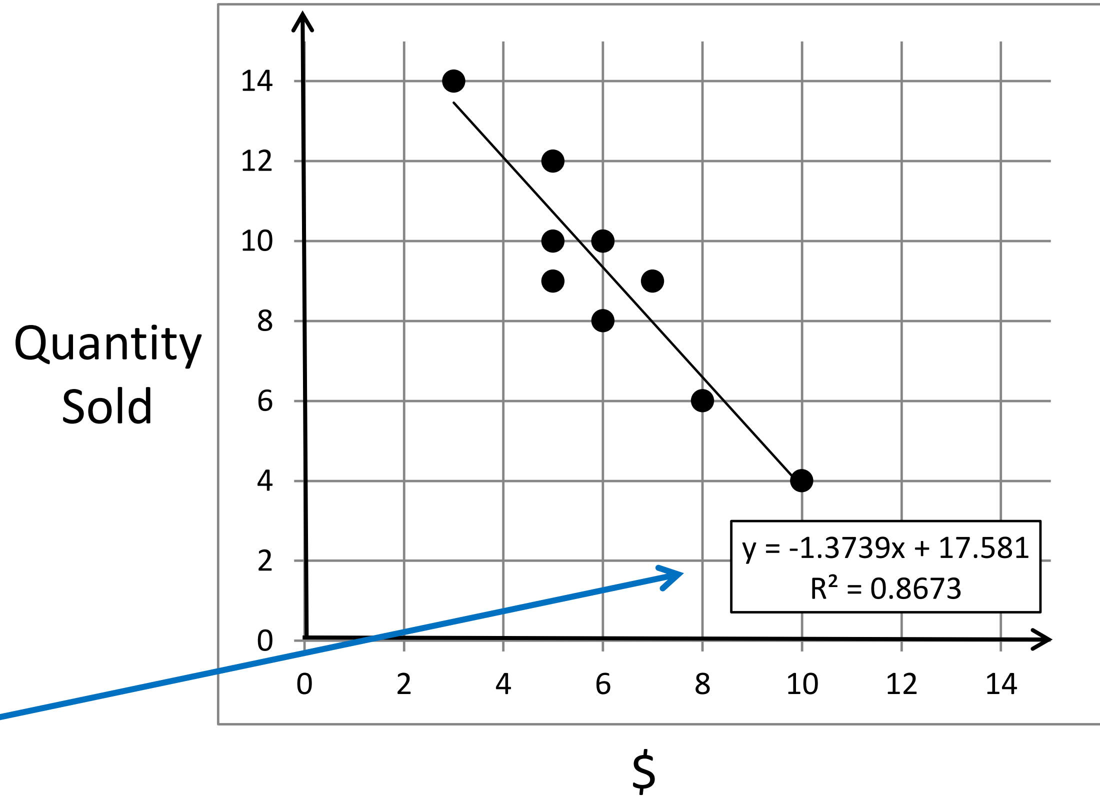
Ordinary Least Squares (OLS)

Using Excel

- XY Scatter Plot
- Add Trend Line

$$\hat{Y} = a + bX$$

$$\hat{Y} = 17.6 - 1.37 * X$$



Advanced methods of analysis

Univariate Regression

Ordinary Least Squares (OLS)

Using Excel

Data Analysis (Regression)

EXCEL SUMMARY OUTPUT

$$\hat{Y} = 17.6 - 1.37 * X$$

Regression Statistics	
Multiple R	0.93
R Square	0.87
Adjusted R Square	0.85
Standard Error	1.09
Observations	10.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	62.10	62.10	52.29	0.00
Residual	8	9.50	1.19		
Total	9	71.60			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	17.58	1.21	14.54	0.00	14.79	20.37	14.79	20.37
X Variable 1	-1.37	0.19	-7.23	0.00	-1.81	-0.94	-1.81	-0.94

Advanced methods of analysis

Univariate Regression

Ordinary Least Squares (OLS)

Assessing the Model

R Square

Also called the coefficient of determination. Reflects the percent of the variance (dispersion) in the data that is explained by the model

Higher is better

EXCEL SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.93
R Square	0.87
Adjusted R Square	0.85
Standard Error	1.09
Observations	10.00

$$\boxed{R^2 = \frac{S(\hat{Y}_i - \bar{Y})^2}{S(Y_i - \bar{Y})^2}}$$

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	62.10	62.10	52.29	0.00
Residual	8	9.50	1.19		
Total	9	71.60			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	17.58	1.21	14.54	0.00	14.79	20.37	14.79	20.37
X Variable 1	-1.37	0.19	-7.23	0.00	-1.81	-0.94	-1.81	-0.94

Advanced methods of analysis

Univariate Regression

Ordinary Least Squares (OLS)

Assessing the Model

Adjusted R Square

Adjusts the R^2 value to control for the number of variables included in the model

Adding variables will increase R^2

EXCEL SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.93
R Square	0.87
Adjusted R Square	0.85
Standard Error	1.09
Observations	10.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	62.10	62.10	52.29	0.00
Residual	8	9.50	1.19		
Total	9	71.60			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	17.58	1.21	14.54	0.00	14.79	20.37	14.79	20.37
X Variable 1	-1.37	0.19	-7.23	0.00	-1.81	-0.94	-1.81	-0.94

Advanced methods of analysis

Univariate Regression

Ordinary Least Squares (OLS)

Assessing the Model

EXCEL SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.93
R Square	0.87
Adjusted R Square	0.85
Standard Error	1.09
Observations	10.00

Model Standard Error

$$S_e = \sqrt{\frac{S(Y_i - \hat{Y})^2}{(n-2)}}$$

Is the standard deviation of the model's residual errors

Is a measure of the amount of "noise" in the data

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	62.10	62.10	52.29	0.00
Residual	8	9.50	1.19		
Total	9	71.60			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	17.58	1.21	14.54	0.00	14.79	20.37	14.79	20.37
X Variable 1	-1.37	0.19	-7.23	0.00	-1.81	-0.94	-1.81	-0.94

Advanced methods of analysis

Univariate Regression

Ordinary Least Squares (OLS)

Assessing the Model

EXCEL SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.93
R Square	0.87
Adjusted R Square	0.85
Standard Error	1.09
Observations	10.00

Coefficient Standard Error

$$S_b = \frac{S_e}{\sqrt{S(X_i - \bar{X})^2}}$$

Is essentially a measure of the signal-to-noise

The larger the coefficient standard error, the worse the signal-to-noise, meaning, less precise

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	62.10	62.10	52.29	0.00
Residual	8	9.50	1.19		
Total	9	71.60			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	17.58	1.21	14.54	0.00	14.79	20.37	14.79	20.37
X Variable 1	-1.37	0.19	-7.23	0.00	-1.81	-0.94	-1.81	-0.94

Advanced methods of analysis

Univariate Regression

Ordinary Least Squares (OLS)

Assessing the Model

t Statistic

Is the Coefficient divided by the standard error

Test of the hypothesis that the “true” value of the coefficient is zero

Rule of thumb, looking for values with absolute value > 1.69

EXCEL SUMMARY OUTPUT

Regression Statistics	
Multiple R	0.93
R Square	0.87
Adjusted R Square	0.85
Standard Error	1.09
Observations	10.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	62.10	62.10	52.29	0.00
Residual	8	9.50	1.19		
Total	9	71.60			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	17.58	1.21	14.54	0.00	14.79	20.37	14.79	20.37
X Variable 1	-1.37	0.19	-7.23	0.00	-1.81	-0.94	-1.81	-0.94

Advanced methods of analysis

Univariate Regression

Ordinary Least Squares (OLS)

Assessing the Model

EXCEL SUMMARY OUTPUT

P-Value

Also known as the significance level

The probability that the result isn’t reliable

(1 - P-Value) is the % of times you’d you’d likely see this result in similar situations

Commonly looking for P-Values < 0.10

Regression Statistics	
Multiple R	0.93
R Square	0.87
Adjusted R Square	0.85
Standard Error	1.09
Observations	10.00

ANOVA

	df	*SS*	*MS*	*F*	*Significance F*
Regression	1	62.10	62.10	52.29	0.00
Residual	8	9.50	1.19		
Total	9	71.60			

	Coefficients	*Standard Error*	*t Stat*	*P-value*	*Lower 95%*	*Upper 95%*	*Lower 95.0%*	*Upper 95.0%*
Intercept	17.58	1.21	14.54	0.00	14.79	20.37	14.79	20.37
X Variable 1	-1.37	0.19	-7.23	0.00	-1.81	-0.94	-1.81	-0.94

Data Considerations / Challenges

Advanced methods of analysis

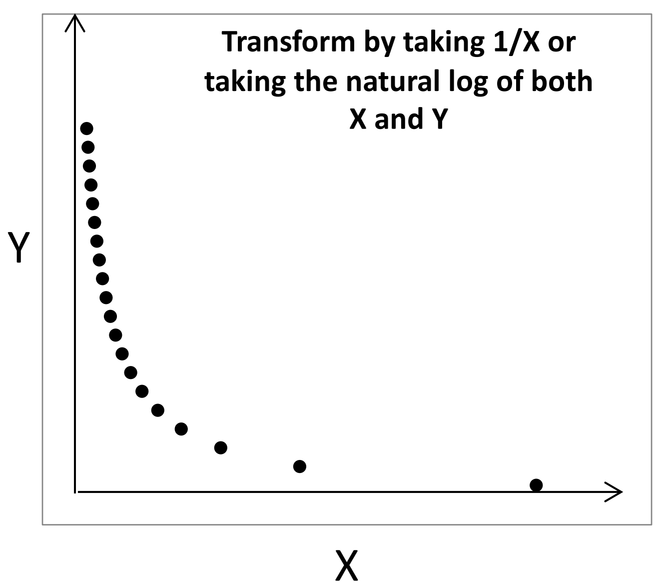
Data Transformations

We noted earlier that in the “real world” data are seldom “perfect”

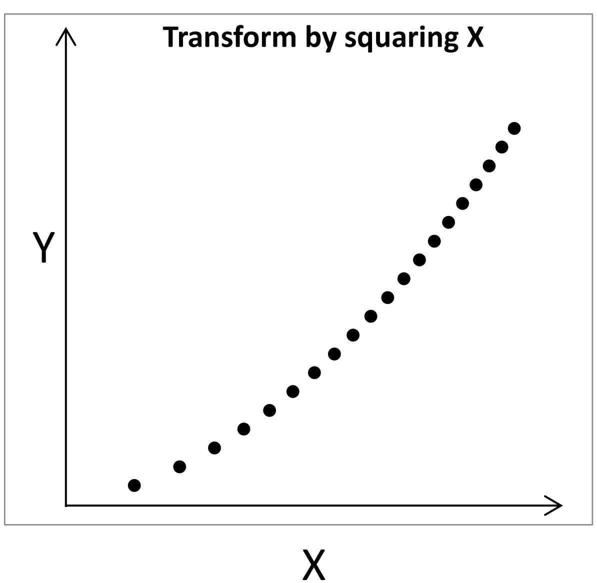
OLS likes data to be as linear as possible…

Sometimes, it is appropriate to “transform” your data to make it more linear…

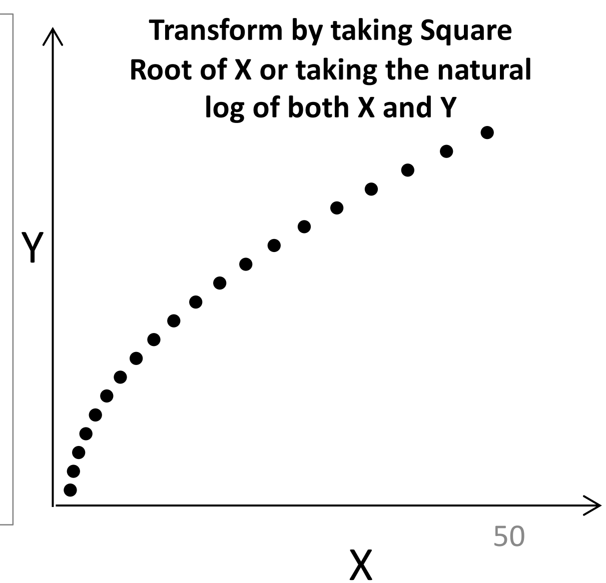
1



2



3



Advanced methods of analysis

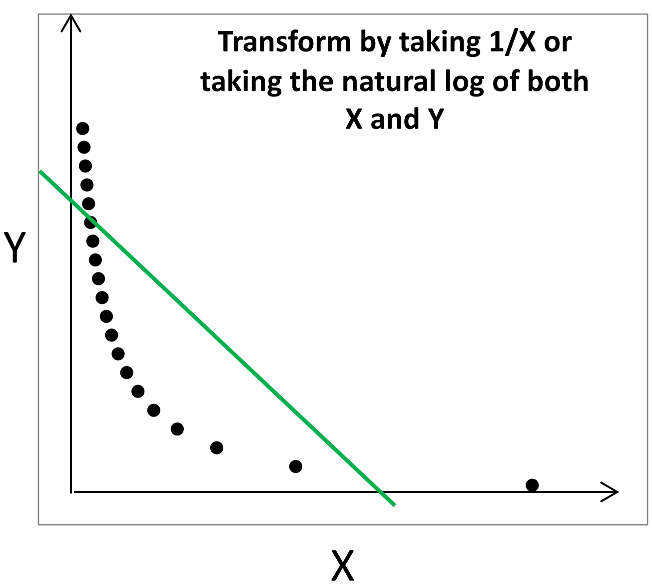
Data Transformations

We noted earlier that in the “real world” data are seldom “perfect”

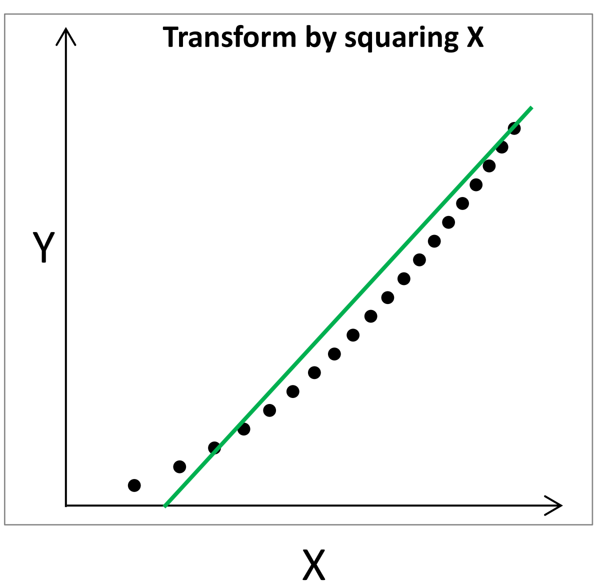
OLS likes data to be as linear as possible…

Sometimes, it is appropriate to “transform” your data to make it more linear…

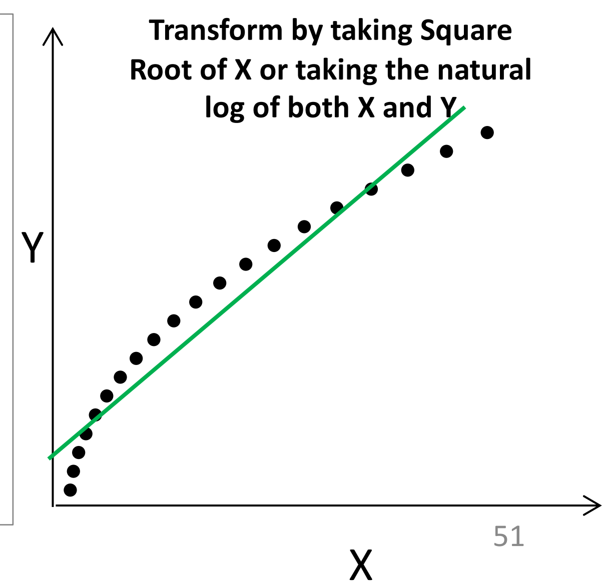
1



2



3



Advanced methods of analysis

Data Transformations

Exercise

Input and Plot the three data sets below and perform the appropriate data transformation.

1

X	Y
10.00	1
5.00	2
3.33	3
2.50	4
2.00	5
1.67	6
1.43	7
1.25	8
1.11	9
1.00	10
0.91	11
0.83	12
0.77	13
0.71	14
0.67	15
0.63	16
0.59	17
0.56	18
0.53	19
0.50	20

2

X	Y
1.00	1
1.41	2
1.73	3
2.00	4
2.24	5
2.45	6
2.65	7
2.83	8
3.00	9
3.16	10
3.32	11
3.46	12
3.61	13
3.74	14
3.87	15
4.00	16
4.12	17
4.24	18
4.36	19
4.47	20

3

X	Y
1	1
4	2
9	3
16	4
25	5
36	6
49	7
64	8
81	9
100	10
121	11
144	12
169	13
196	14
225	15
256	16
289	17
324	18
361	19
400	20

Advanced methods of analysis

Univariate Regression

Ordinary Least Squares (OLS)

Output using a different Statistical Package (E-Views)

Dependent Variable: QUANTITY
Method: Least Squares
Date: 07/20/14 Time: 14:44
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	17.58055	1.209153	14.53956	0.0000
PRICE	-1.373860	0.190000	-7.230843	0.0001

R-squared	0.867297	Mean dependent var	9.200000
Adjusted R-squared	0.850709	S.D. dependent var	2.820559
S.E. of regression	1.089812	Akaike info criterion	3.186744
Sum squared resid	9.501520	Schwarz criterion	3.247261
Log likelihood	-13.93372	Hannan-Quinn criter.	3.120357
F-statistic	52.28509	Durbin-Watson stat	2.619326
Prob(F-statistic)	0.000090		

	Price	Quantity
	X	Y
Obs	$	# Sold
1	10	4
2	5	9
3	6	10
4	3	14
5	6	8
6	6	10
7	8	6
8	5	12
9	7	9
10	5	10

Advanced methods of analysis

Univariate Regression

Problems to watch out for - Heteroskedasticity

$$\text{INCOME} = f(\text{AGE})$$

(+)

Dependent Variable: INCOME
Method: Least Squares
Date: 07/20/14 Time: 17:20
Sample: 1 16
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	15063.69	11228.48	1.341561	0.2011
AGE	762.1433	242.0928	3.148146	0.0071

R-squared	0.414491	Mean dependent var	49312.50
Adjusted R-squared	0.372669	S.D. dependent var	14036.71
S.E. of regression	11117.67	Akaike info criterion	21.58693
Sum squared resid	1.73E+09	Schwarz criterion	21.68350
Log likelihood	-170.6954	Hannan-Quinn criter.	21.59187
F-statistic	9.910821	Durbin-Watson stat	0.435679
Prob(F-statistic)	0.007118		

Model results “look” pretty good

Descent R^2

T-Stat > 1.69

P-Value < 0.10

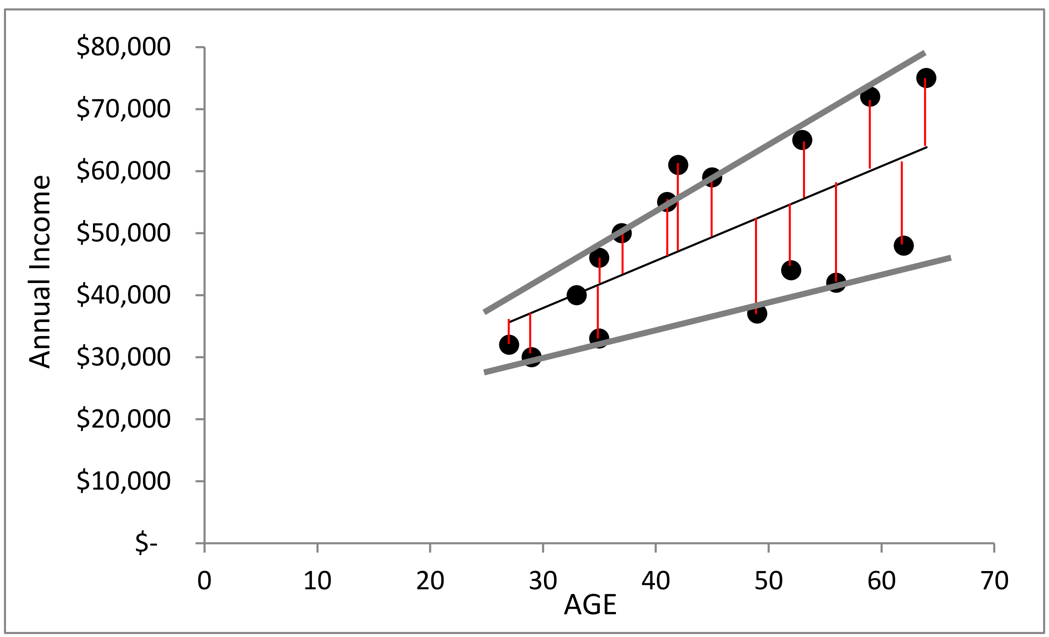
But….

Advanced methods of analysis
Univariate Regression

Problems to watch out for - Heteroskedasticity

Error terms are not randomly distributed
Error terms are correlated the dependent variable (not independent)

INCOME = f(AGE)
(+)



Scatter plot will typically look like a cone or funnel

Residual Error increase as Y (INCOME) and X (AGE) increases
Not independent from each other

Time Series

Advanced methods of analysis

Time Series Data

Dice Exercise

Take 3 Dice

Roll the dice 15 different times at 10 second intervals

Create a data series in sequential order (1,2…15) of or our rolls

Each data point is the sum of the values of the 3 dice

1st roll = Observation 1,

2nd roll = Observation 2, etc

Using Excel, plot our results

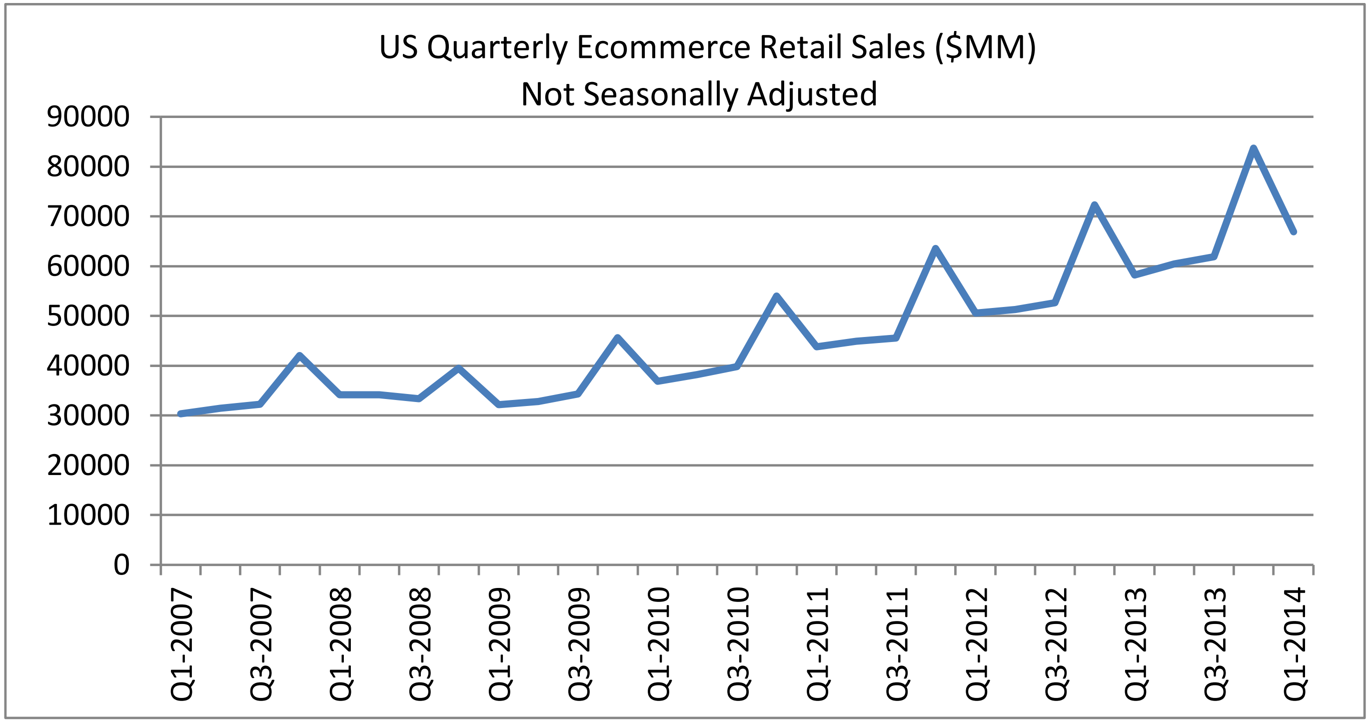
Discussion

Advanced methods of analysis

Time Series Data

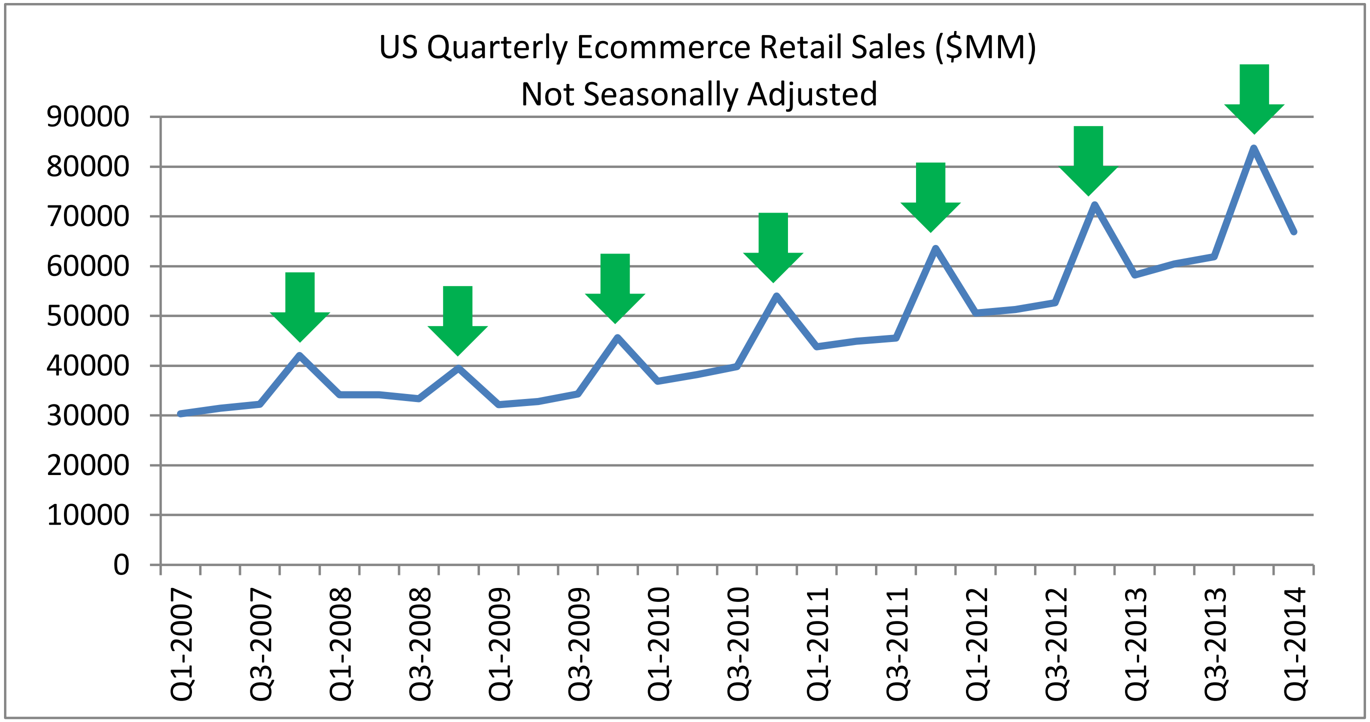
Data series over time

Seconds
Hours
Days
Weeks
Months
Quarters
Years
etc.



- Trends
- Patterns
- Changes / Shocks in either trends or patterns

Advanced methods of analysis
Time Series Data

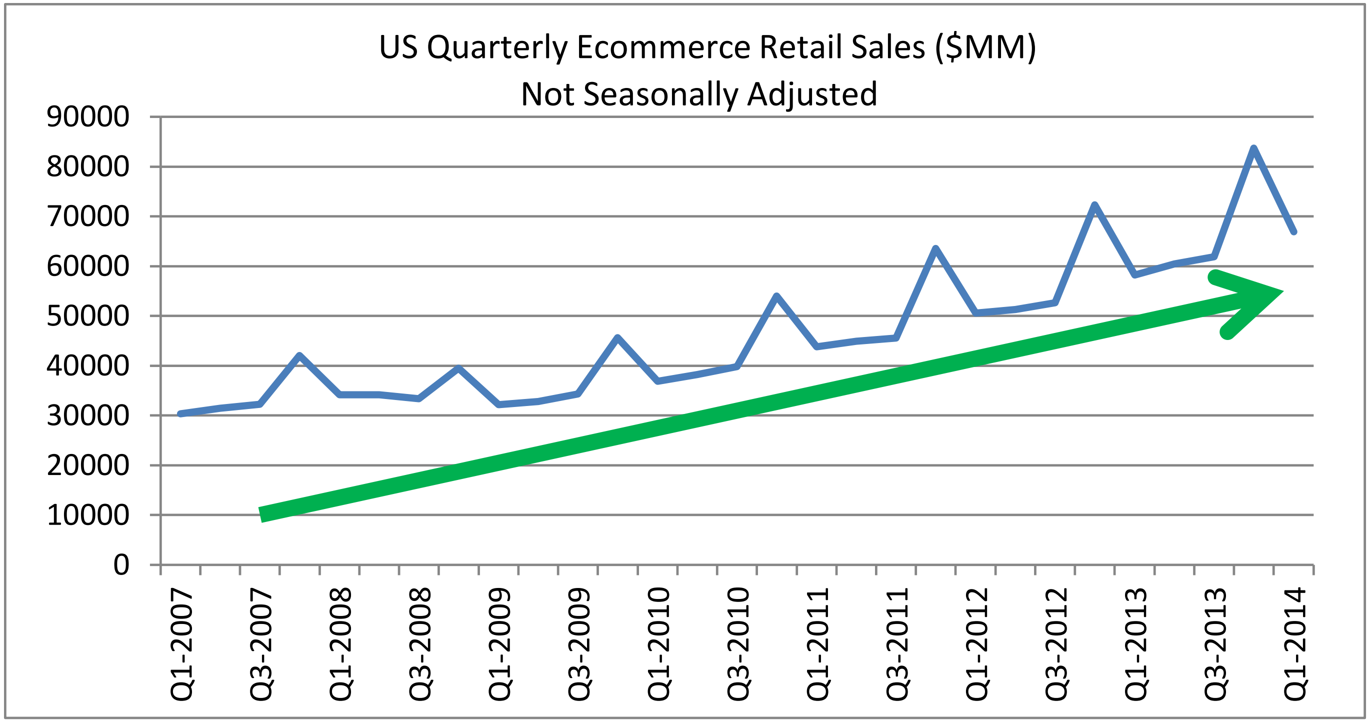


Seasonality

Periodic, repetitive, and predictable movement around a trend

Advanced methods of analysis

Time Series Data



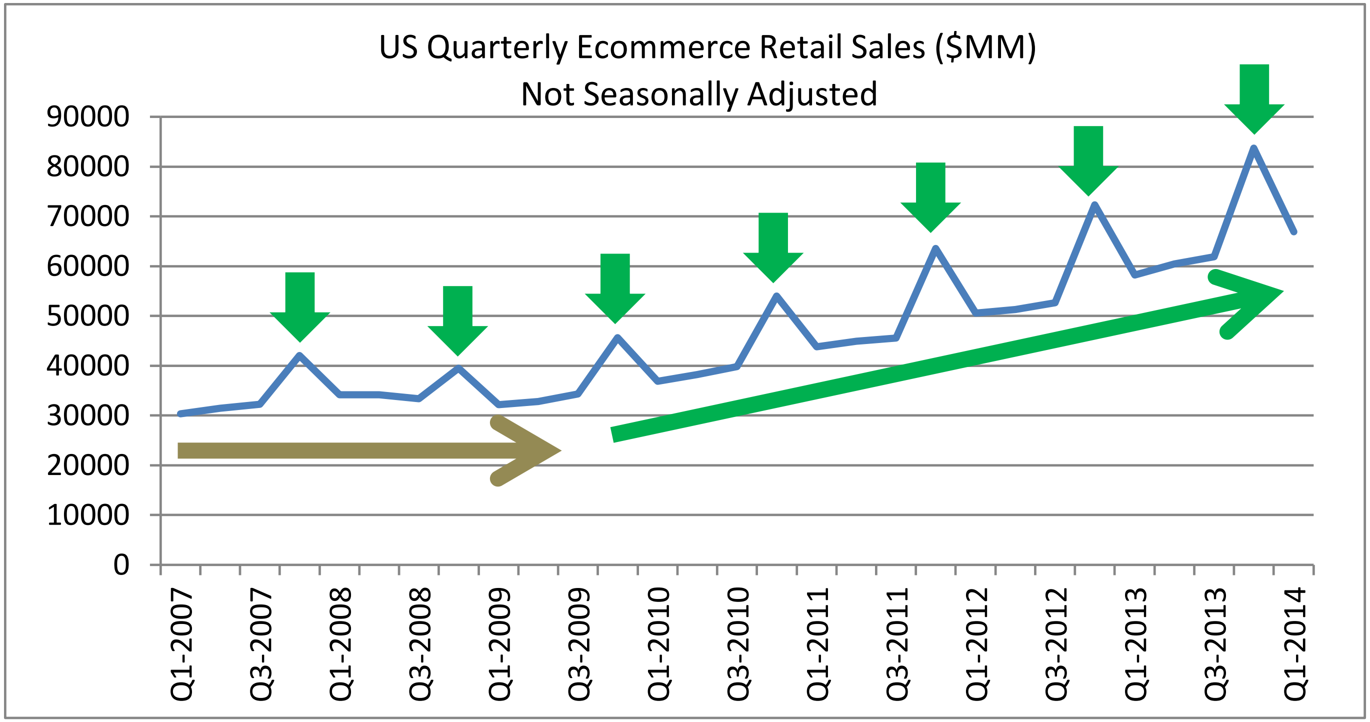
Seasonality

Trend

Long-term movement when things like ***seasonality*** *have been filtered out*

Advanced methods of analysis

Time Series Data



Seasonality

Trend

Shocks

Observed breaks or deviations in seasonality and / or trend

Advanced methods of analysis

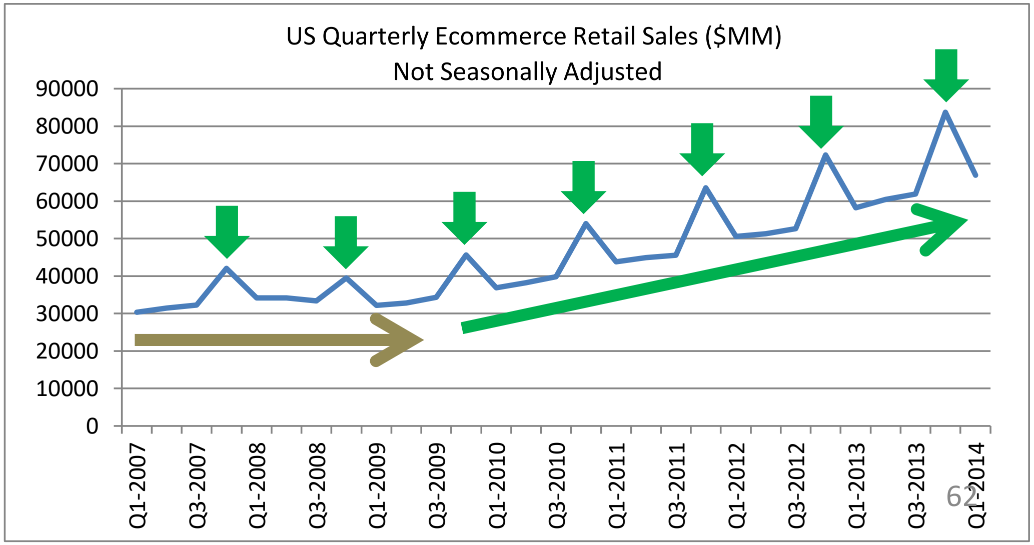
Time Series Data – US Ecommerce Sales

	Ecommerce Sales	Q4 Dummy	TIME	Recession Dummy
Q1-2007	30,322	0	1	0
Q2-2007	31,493	0	2	0
Q3-2007	32,248	0	3	0
Q4-2007	42,063	1	4	1
Q1-2008	34,169	0	5	1
Q2-2008	34,170	0	6	1
Q3-2008	33,396	0	7	1
Q4-2008	39,498	1	8	1
Q1-2009	32,185	0	9	1
Q2-2009	32,789	0	10	1
Q3-2009	34,317	0	11	0
Q4-2009	45,617	1	12	0
Q1-2010	36,842	0	13	0
Q2-2010	38,230	0	14	0
Q3-2010	39,809	0	15	0
Q4-2010	54,014	1	16	0
Q1-2011	43,841	0	17	0
Q2-2011	44,948	0	18	0
Q3-2011	45,545	0	19	0
Q4-2011	63,549	1	20	0
Q1-2012	50,589	0	21	0
Q2-2012	51,285	0	22	0
Q3-2012	52,643	0	23	0
Q4-2012	72,361	1	24	0
Q1-2013	58,215	0	25	0
Q2-2013	60,498	0	26	0
Q3-2013	61,857	0	27	0
Q4-2013	83,709	1	28	0
Q1-2014	66,917	0	29	0

Seasonality Created a “dummy variable” to control for the spike in Q4 each year

Trend Simple linear time variable that increments by one each period

Shock Created a “dummy variable” to control for the recession between Q4-2007 and Q2-2009



Advanced methods of analysis

Time Series Data – US Ecommerce Sales

Let's do some regressions and discuss our results

Dependent Variable: ECOM_SALES
Method: Least Squares
Date: 07/22/14 Time: 17:37
Sample: 2007Q1 2014Q1
Included observations: 29

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	43014.00	2715.331	15.84116	0.0000
Q4_DUMMY	14244.71	5526.788	2.577395	0.0157

R-squared	0.197455	Mean dependent var	46452.38
Adjusted R-squared	0.167731	S.D. dependent var	13960.54
S.E. of regression	12736.03	Akaike info criterion	21.80873
Sum squared resid	4.38E+09	Schwarz criterion	21.90303
Log likelihood	-314.2266	Hannan-Quinn criter.	21.83826
F-statistic	6.642963	Durbin-Watson stat	0.082249
Prob(F-statistic)	0.015736		

Coefficient = 14,244 MM

Meaning that all things equal, Q4 is typically $14.2B larger than the other quarters

t-Stat = 2.577 and P-Value = 0.01

OK results – definitely significant

R2 = 0.19

Not great, but we're only controlling for one piece of information, so not bad

Advanced methods of analysis

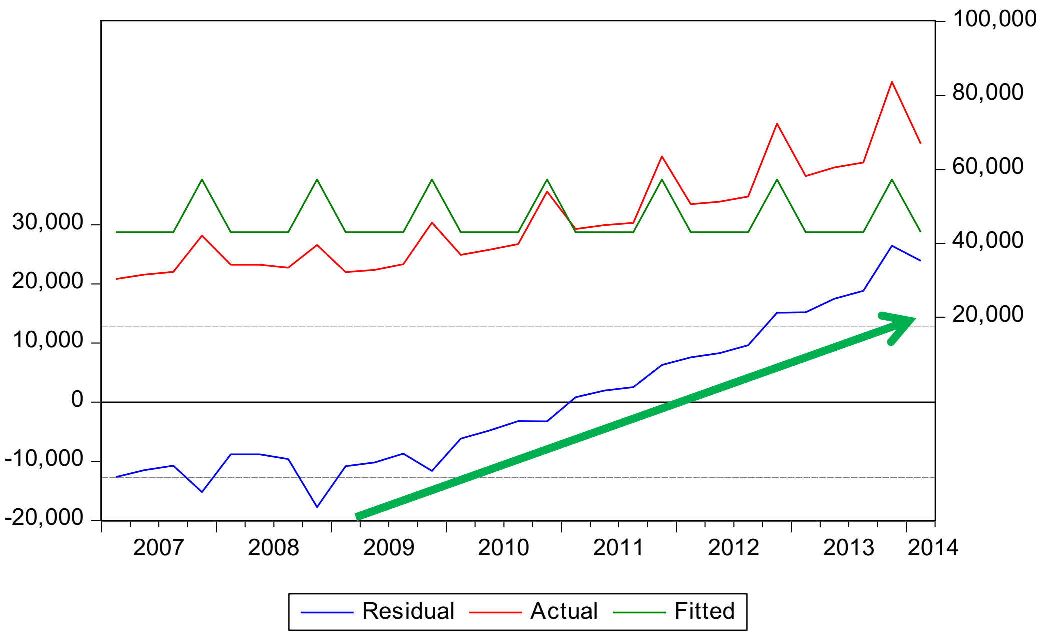
Time Series Data – US Ecommerce Sales

Dependent Variable: ECOM_SALES
Method: Least Squares
Date: 07/22/14 Time: 17:37
Sample: 2007Q1 2014Q1
Included observations: 29

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	43014.00	2715.331	15.84116	0.0000
Q4_DUMMY	14244.71	5526.788	2.577395	0.0157

R-squared	0.197455	Mean dependent var	46452.38
Adjusted R-squared	0.167731	S.D. dependent var	13960.54
S.E. of regression	12736.03	Akaike info criterion	21.80873
Sum squared resid	4.38E+09	Schwarz criterion	21.90303
Log likelihood	-314.2266	Hannan-Quinn criter.	21.83826
F-statistic	6.642963	Durbin-Watson stat	0.082249
Prob(F-statistic)	0.015736		

Assessing the residual errors



There is a time trend in our residual error (blue line) at the bottom of the graph

Sometimes called a "drift"

We want the blue line to essentially bounce (randomly) around the "Zero Line" with roughly equal amount of points above and below the zero line (Zero Mean)

Advanced methods of analysis

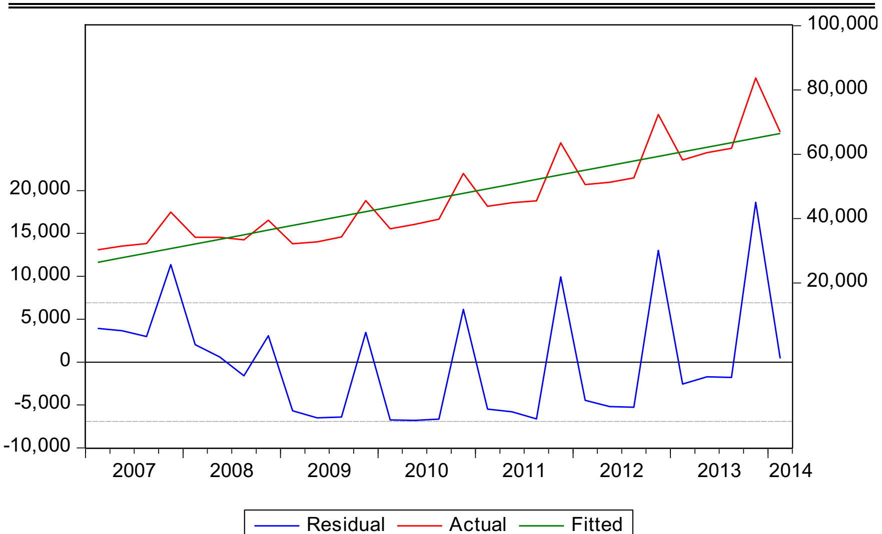
Time Series Data – US Ecommerce Sales

Looking at only the time variable (ignoring the Q4 spikes)

Dependent Variable: ECOM_SALES
Method: Least Squares
Date: 07/22/14 Time: 17:53
Sample: 2007Q1 2014Q1
Included observations: 29

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24973.52	2639.852	9.460195	0.0000
TIME	1431.924	153.6981	9.316470	0.0000

R-squared	0.762735	Mean dependent var	46452.38
Adjusted R-squared	0.753947	S.D. dependent var	13960.54
S.E. of regression	6924.950	Akaike info criterion	20.59012
Sum squared resid	1.29E+09	Schwarz criterion	20.68442
Log likelihood	-296.5568	Hannan-Quinn criter.	20.61965
F-statistic	86.79662	Durbin-Watson stat	1.987252
Prob(F-statistic)	0.000000		



Coefficient = 1,431.9 MM

Meaning Ecommerce sales are growing by ~ $1.4B each quarter (between 2007 – present)

t-Stat = 9.31 and P-Value = 0.00

OK Strong – definitely significant

R2 = 0.76

Pretty good!

Removed most of the drift in the residual error

See these periodic "shocks" which we know always happen in Q4

Looks like something was happening from 2007 – mid-2009

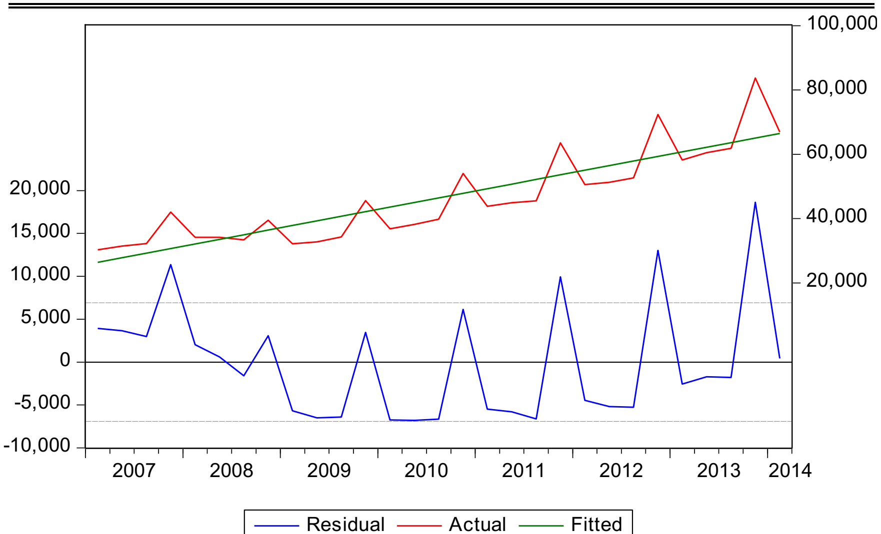
Advanced methods of analysis
Time Series Data – US Ecommerce Sales

Looking at only the time variable (ignoring the Q4 spikes)

Dependent Variable: ECOM_SALES
Method: Least Squares
Date: 07/22/14   Time: 17:53
Sample: 2007Q1 2014Q1
Included observations: 29

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	24973.52	2639.852	9.460195	0.0000
TIME	1431.924	153.6981	9.316470	0.0000

R-squared	0.762735	Mean dependent var	46452.38
Adjusted R-squared	0.753947	S.D. dependent var	13960.54
S.E. of regression	6924.950	Akaike info criterion	20.59012
Sum squared resid	1.29E+09	Schwarz criterion	20.68442
Log likelihood	-296.5568	Hannan-Quinn criter.	20.61965
F-statistic	86.79662	Durbin-Watson stat	1.987252
Prob(F-statistic)	0.000000		



Coefficient = 1,431.9 MM

Meaning Ecommerce sales are growing by ~ $1.4B each quarter (between 2007 – present)

t-Stat = 9.31   and P-Value = 0.00

OK Strong – definitely significant

R2 = 0.76

Pretty good!

Removed most of the drift in the residual error

See these periodic “shocks” which we know always happen in Q4

Looks like something was happening from 2007 – mid-2009

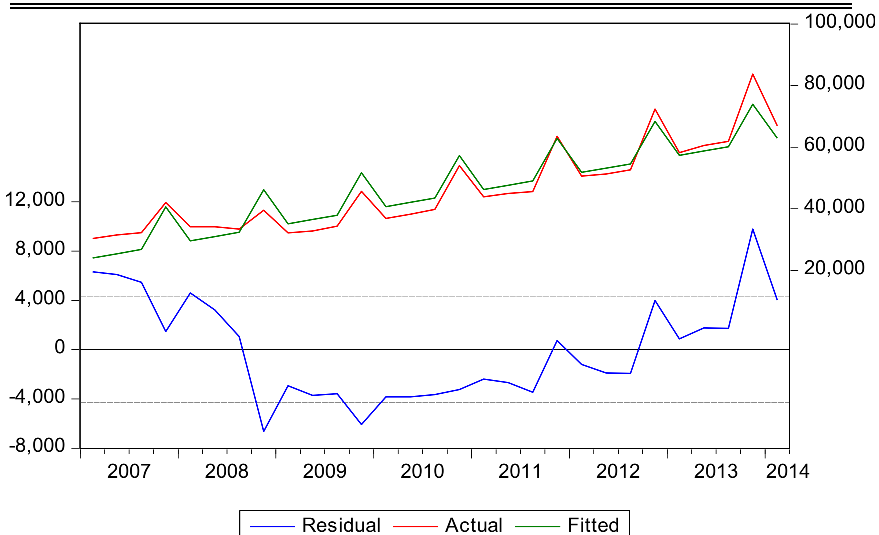
Advanced methods of analysis
Time Series Data – US Ecommerce Sales

Now also bringing in the Q4 Spikes into the model

Dependent Variable: ECOM_SALES
Method: Least Squares
Date: 07/22/14 Time: 17:59
Sample: 2007Q1 2014Q1
Included observations: 29

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	22619.21	1676.178	13.49452	0.0000
TIME	1389.119	95.60087	14.53040	0.0000
Q4_DUMMY	12413.60	1869.169	6.641243	0.0000

R-squared	0.912006	Mean dependent var	46452.38
Adjusted R-squared	0.905237	S.D. dependent var	13960.54
S.E. of regression	4297.547	Akaike info criterion	19.66717
Sum squared resid	4.80E+08	Schwarz criterion	19.80862
Log likelihood	-282.1740	Hannan-Quinn criter.	19.71147
F-statistic	134.7377	Durbin-Watson stat	0.593286
Prob(F-statistic)	0.000000		



Coefficients changed a little from our prior specifications

Meaning Ecommerce sales are growing by ~ $1.4B each quarter (between 2007 – present)

t-Stats and P-Values are very strong definitely significant

R2 = 0.91

Getting better!

Removed most of the drift in the residual error – but it is still present

We are capturing the Q4 Seasonal Sales Spikes nicely

Looks like something was happening from 2007 – mid-2009

Advanced methods of analysis

Time Series Data – US Ecommerce Sales

Recap

Analysis of data series over time

- Trends
- Patterns
- Changes / Shocks in either trends or patterns

Time Trend

Seasonality

Dummy Variable

Semi- A Theoretical Approach

We've Addressed What – Not Why

Advanced methods of analysis

Time Series Regression

Plot the data

- Over time and vs other data with possible relationships
- What patterns / relationships do you observe?
- Are there recurring patterns

Run a regression with “time” as the independent variable

- What patterns / trends (if any) do you see in the residuals

Discussion

Advanced methods of analysis
Multivariate Time Series Regression

ACME Widget Company

- 52 Weeks of Data
- Units Sold
- Net Price of Widgets Sold
- Supply of Widgets on Hand
- Knowledge of Promotional Event Timing
- Major Weather Event Mid Year Caused Significant Disruption

1) How much does Price affect Units Sold?
2) Does my supply on hand impact sales?
3) Are my Promotional Events working?
4) How much did the weather event cost me?
5) Anything else I need to be aware of?

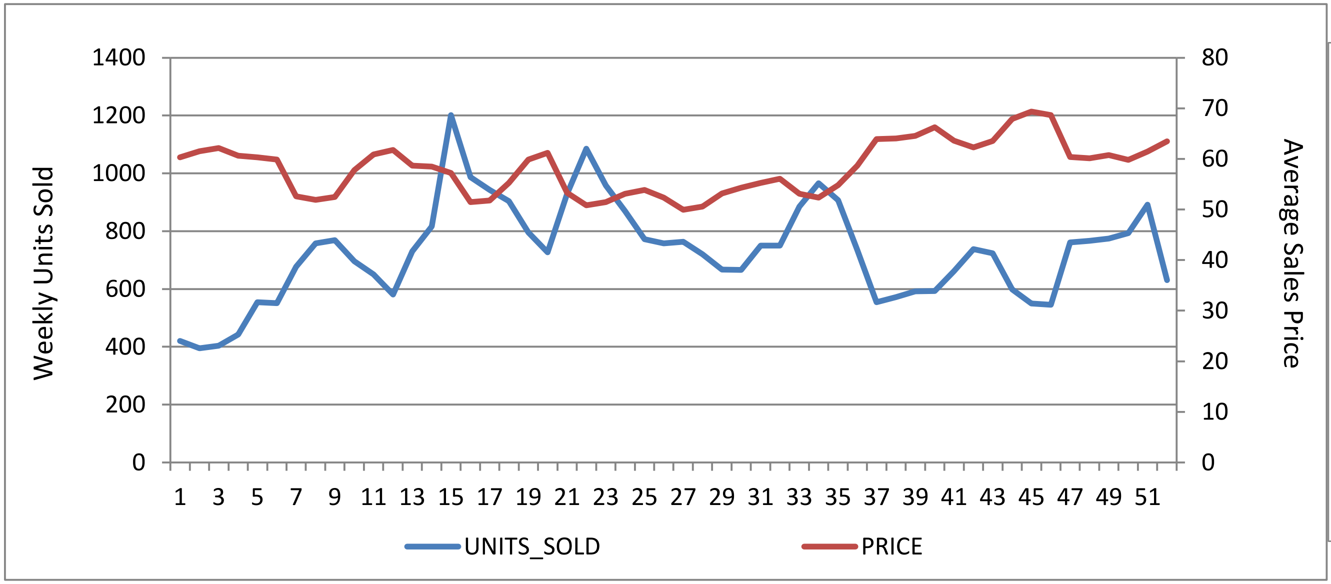
Advanced methods of analysis
Multivariate Time Series Regression

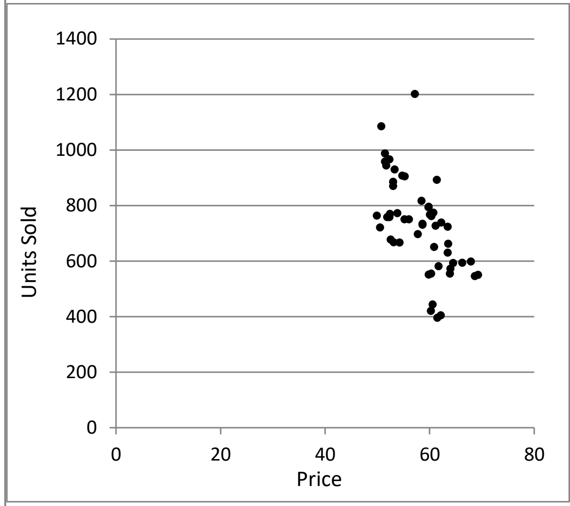
ACME Widget Company

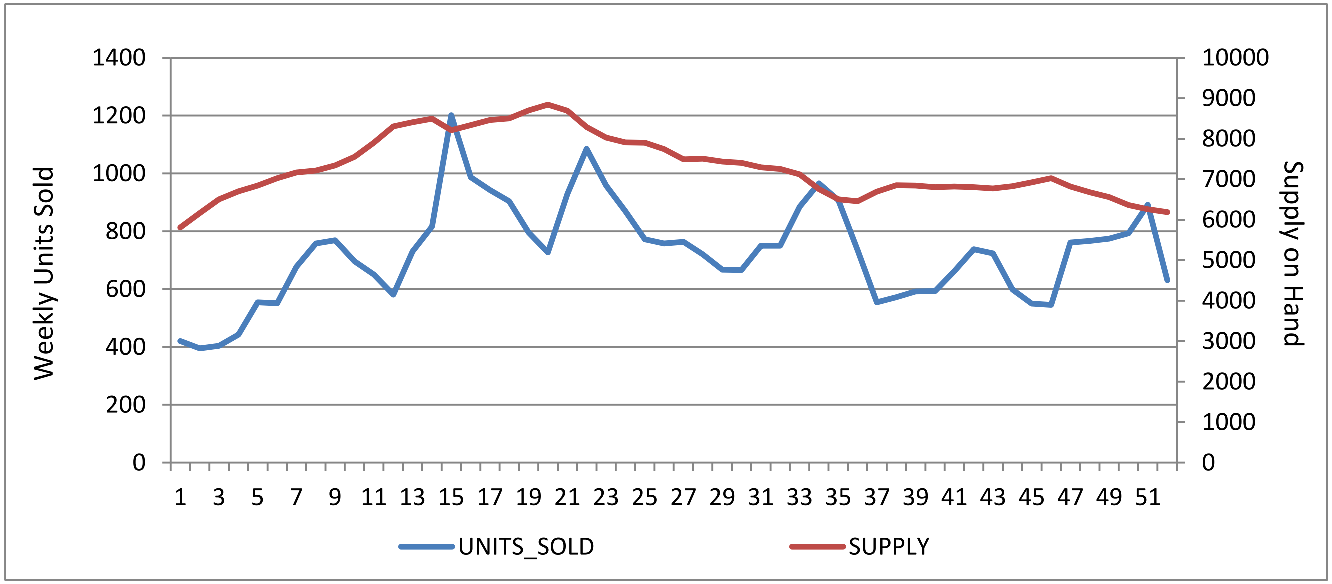
Answer questions by estimating a weekly demand model

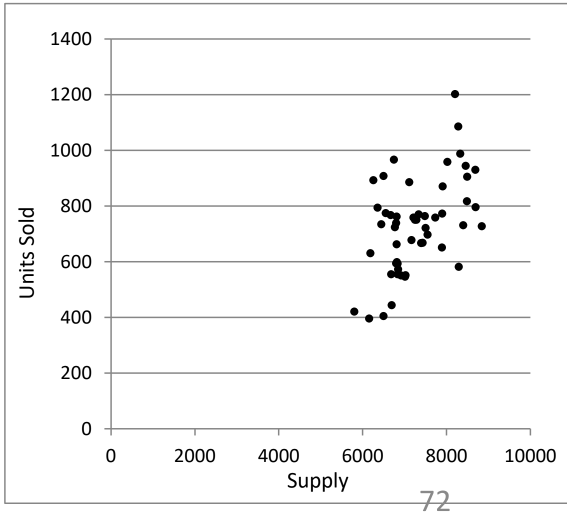
Steps:
1) Look at (plot) the data
2) Look at Correlation Matrix
3) Iteratively build the Weekly Demand Model   UNITS = f (Price, etc…)
4) Assess Output with each iteration
5) Review the residual errors of each iteration

Advanced methods of analysis
Multivariate Time Series Regression



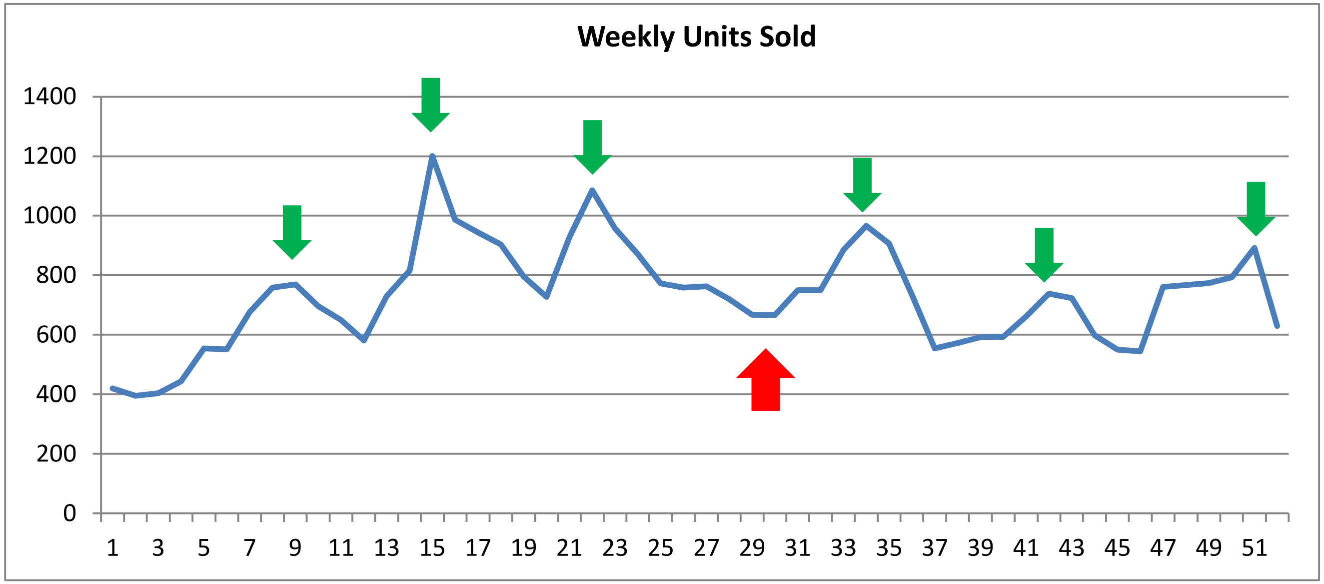


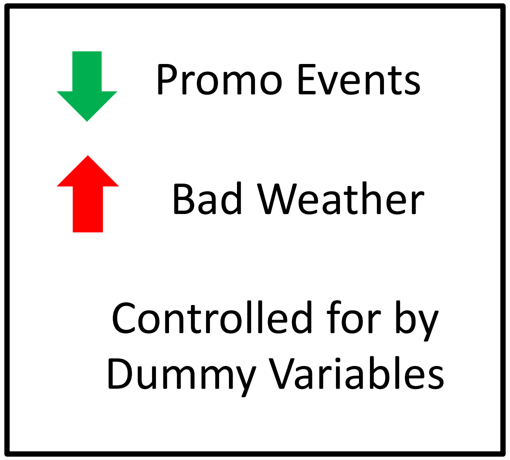




Advanced methods of analysis

Multivariate Time Series Regression





Independent Variable
Correlation Matrix

	PRICE	SUPPLY	PROMO_EVENT	BAD_WEATHER
PRICE	1.00	-0.44	-0.15	-0.18
SUPPLY	-0.44	1.00	-0.01	0.03
PROMO_EVENT	-0.15	-0.01	1.00	-0.07
BAD_WEATHER	-0.18	0.03	-0.07	1.00

Advanced methods of analysis
Multivariate Time Series Regression

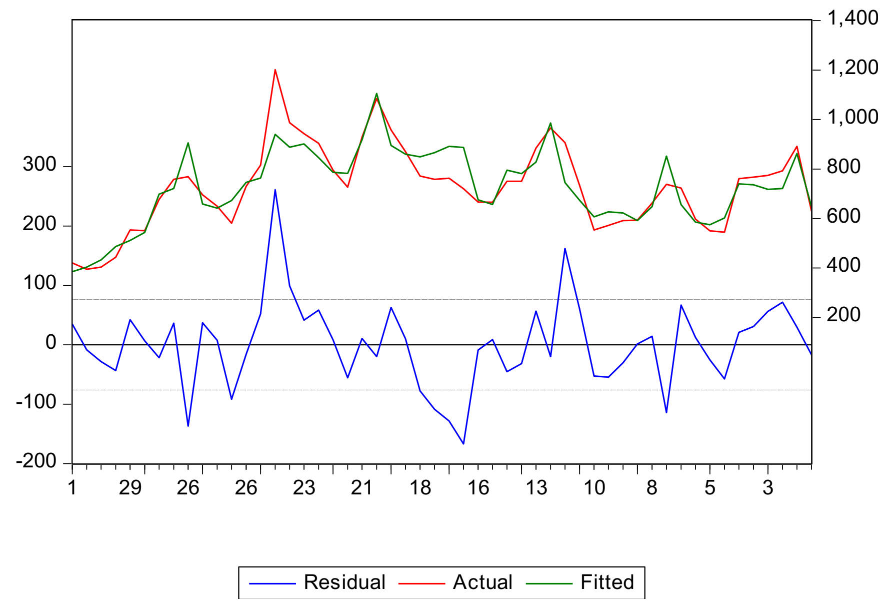
Final Model Output

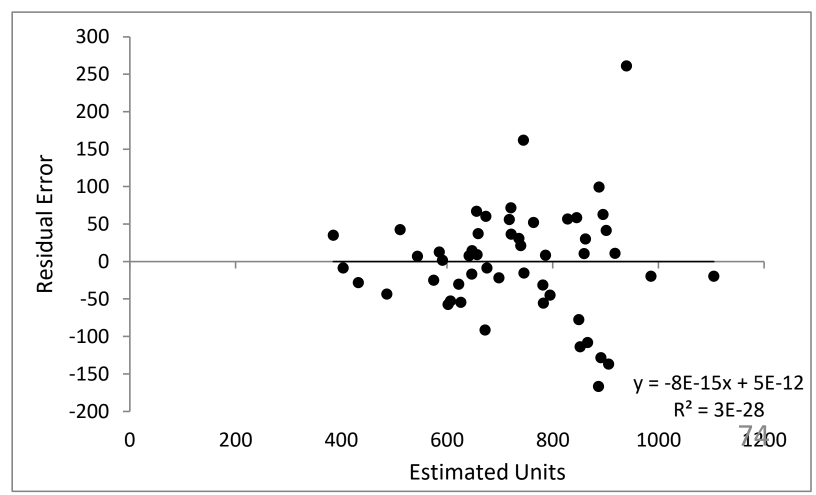
Dependent Variable: UNITS_SOLD
Method: Least Squares
Date: 07/23/14 Time: 12:59
Sample: 1/01/2014 12/24/2014
Included observations: 52

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	901.3092	215.9840	4.173037	0.0001
PRICE	-18.47527	2.480796	-7.447315	0.0000
SUPPLY	0.101883	0.015892	6.411067	0.0000
PROMO_EVENT	175.7728	34.08523	5.156860	0.0000
BAD_WEATHER	-161.6135	56.72894	-2.848872	0.0065
TIME	5.536277	0.793918	6.973357	0.0000

R-squared	0.816102	Mean dependent var	729.3269
Adjusted R-squared	0.796113	S.D. dependent var	169.0745
S.E. of regression	76.34355	Akaike info criterion	11.61653
Sum squared resid	268103.5	Schwarz criterion	11.84167
Log likelihood	-296.0298	Hannan-Quinn criter.	11.70285
F-statistic	40.82783	Durbin-Watson stat	1.372811
Prob(F-statistic)	0.000000		

Residual Error Structure





Multivariate Regression

Advanced methods of analysis

Multivariate Regression

Regression analysis with more than one variable

$Y = f(X_1, X_2, \ldots)$

Quantity = f (Price, Income, Price of Substitutes, etc.)

Can be “cross section” or “time series” data

Advanced methods of analysis

Multivariate Regression

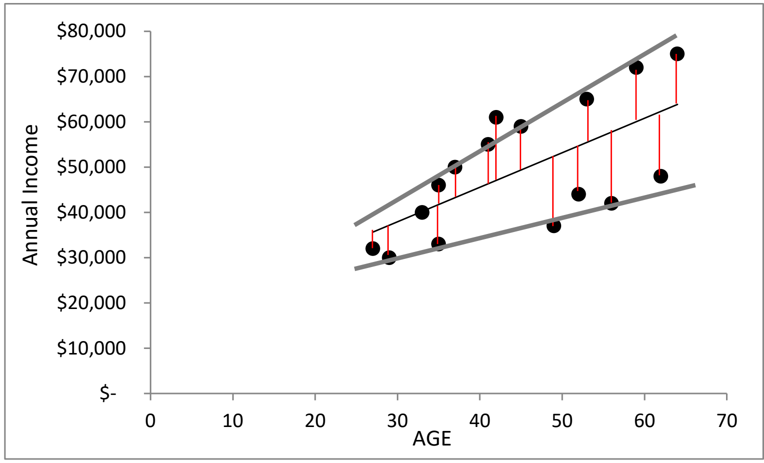
Revisiting our model that had a Heteroskedasticity problem

$$\text{INCOME} = f(\underset{(+)}{\text{AGE}})$$

Dependent Variable: INCOME
Method: Least Squares
Date: 07/20/14 Time: 17:20
Sample: 1 16
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	15063.69	11228.48	1.341561	0.2011
AGE	762.1433	242.0928	3.148146	0.0071

R-squared	0.414491	Mean dependent var	49312.50
Adjusted R-squared	0.372669	S.D. dependent var	14036.71
S.E. of regression	11117.67	Akaike info criterion	21.58693
Sum squared resid	1.73E+09	Schwarz criterion	21.68350
Log likelihood	-170.6954	Hannan-Quinn criter.	21.59187
F-statistic	9.910821	Durbin-Watson stat	0.435679
Prob(F-statistic)	0.007118		



Is often a sign that something important is missing from the model

Advanced methods of analysis

Multivariate Regression

Revisiting our model that had a Heteroskedasticity problem

INCOME = f(AGE, Education, Gender)

(+) (+) (+ / -)

Obs	INCOME	AGE	GENDER	EDUCATION
1	30,000	29	0	14
2	32,000	27	1	16
3	33,000	35	0	16
4	37,000	49	0	12
5	40,000	33	1	14
6	42,000	56	0	12
7	44,000	52	0	18
8	46,000	35	1	12
9	48,000	62	0	14
10	50,000	37	1	12
11	55,000	41	1	14
12	59,000	45	1	16
13	61,000	42	1	20
14	65,000	53	1	18
15	72,000	59	1	18
16	75,000	64	1	14

Dataset has 16 people in it

Gender Variable is either a "1" or "0"

1 = Male

0 = Female

This is called a "dummy" variable

Education is number of years of school completed

12 = High School

14 = Associates Degree

16 = Bachelors Degree

18 = Masters Degree

20 = PhD or Professional Degree (i.e., Law)

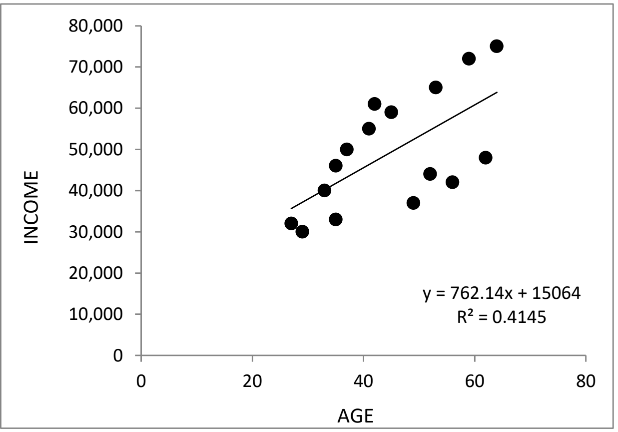
Advanced methods of analysis

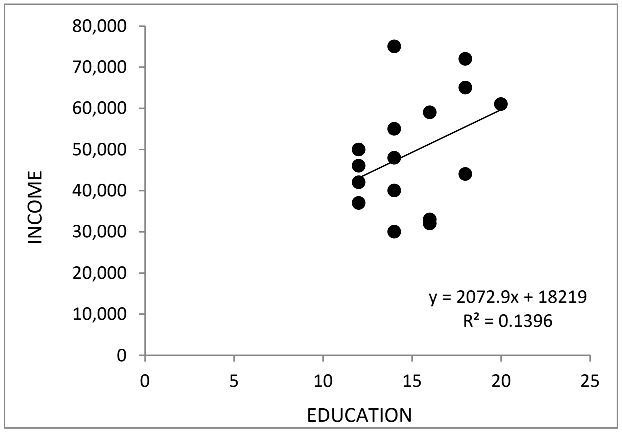
Multivariate Regression

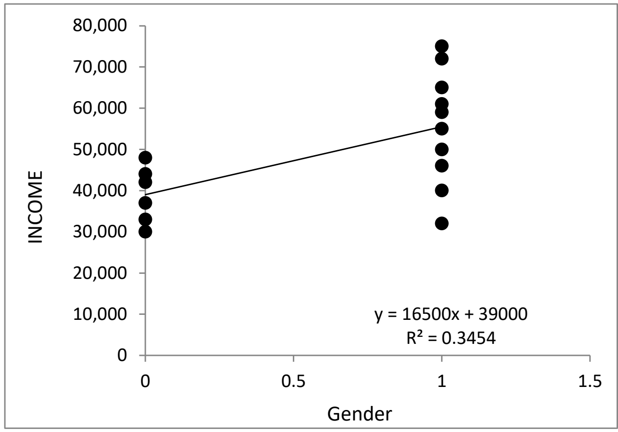
INCOME = f(AGE, Education, Gender)

(+) (+) (+ / -)

Plot out the data







Each univariate plot and excel trend line regression suggests there is a positive relationship between the independent and dependent variables

What about the relationship between the independent variables?

Advanced methods of analysis
Multivariate Regression

$$\text{INCOME} = f(\text{AGE, Education, Gender})$$

(+) (+) **(+ / -)**

Look at the Correlation Matrix

	AGE	EDUCATION	GENDER
AGE	1.00	0.11	-0.15
EDUCATION	0.11	1.00	0.21
GENDER	-0.15	0.21	1.00

Correlation is the measure the strength of association (relationship) between two variables

A high correlation coefficient indicates two variables are not (or are not sufficiently) independent from one another

- High correlation of the independent variables gives rise to Multicolinearity

Loose rule of thumb to test for model inclusion is ~ < 0.75

Advanced methods of analysis
Multivariate Regression

INCOME = f(AGE, Education, Gender)
(+) (+) **(+ / -)**

Dependent Variable: INCOME
Method: Least Squares
Date: 07/21/14 Time: 13:41
Sample: 1 16
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-13864.13	7989.759	-1.735238	0.1083
AGE	860.5635	101.1059	8.511506	0.0000
EDUCATION	856.3410	479.2658	1.786777	0.0992
GENDER	18655.91	2438.227	7.651426	0.0000

R-squared	0.916257	Mean dependent var	49312.50
Adjusted R-squared	0.895322	S.D. dependent var	14036.71
S.E. of regression	4541.444	Akaike info criterion	19.89220
Sum squared resid	2.47E+08	Schwarz criterion	20.08534
Log likelihood	-155.1376	Hannan-Quinn criter.	19.90209
F-statistic	43.76531	Durbin-Watson stat	0.681418
Prob(F-statistic)	0.000001		

Regression Checklist

- ✓ R2 and Adj R2 are high and “close” to one another
- ✓ Coefficients all have the expected signs
- ✓ t-Statistics are all > 1.69 (Absolute value)
- ✓ P-Values are all 0.10 or smaller

What about the residual errors?

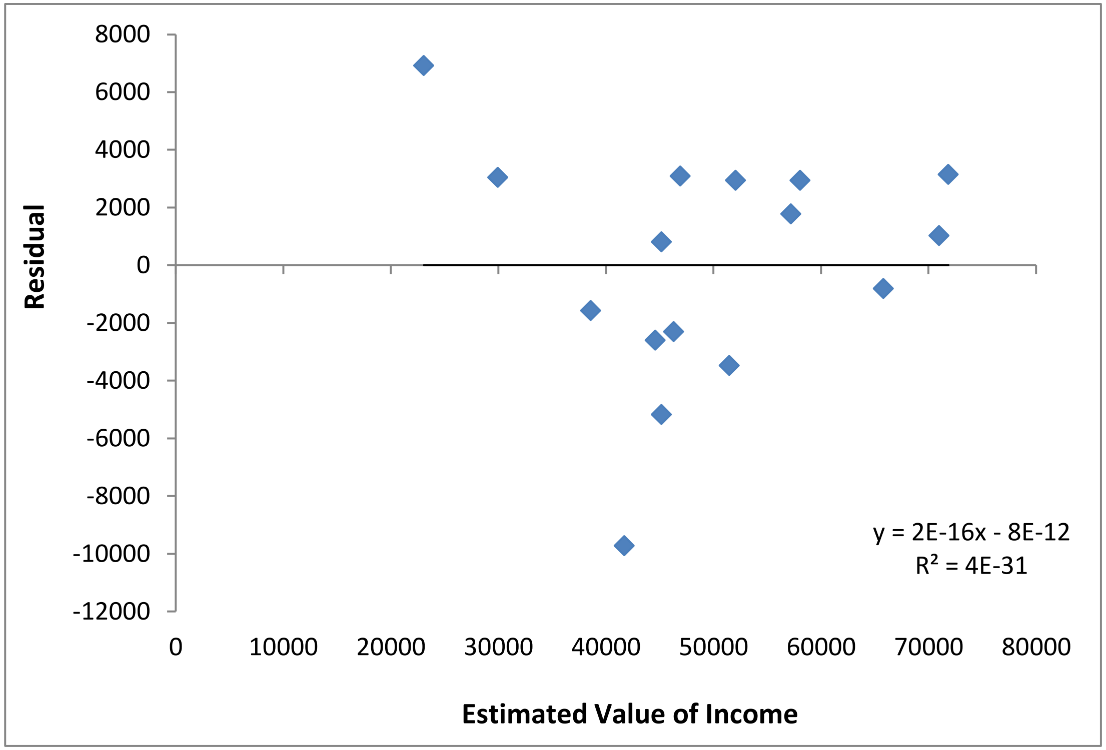
Advanced methods of analysis
Multivariate Regression

INCOME = f(AGE, Education, Gender)
(+) (+) (+ / -)

Evaluating the model residual error structure

OBS	INCOME	Estimate	Residual
1	30,000	23,081	6,919
2	32,000	41,728	-9,728
3	33,000	29,957	3,043
4	37,000	38,580	-1,580
5	40,000	45,179	-5,179
6	42,000	44,604	-2,604
7	44,000	46,299	-2,299
8	46,000	45,188	812
9	48,000	51,480	-3,480
10	50,000	46,909	3,091
11	55,000	52,064	2,936
12	59,000	57,219	1,781
13	61,000	58,062	2,938
14	65,000	65,816	-816
15	72,000	70,979	1,021
16	75,000	71,857	3,143
			0.00 Avg

Correl
Est vs. Resid
0.00000



Advanced methods of analysis

Multivariate Regression

A word on Multicollinearity

We noted earlier that this is a situation where two or more of the independent variables are highly correlated with one another

- Imagine two people in a hallway yelling something really important at the top of their lungs – hard to understand what is being said

Let's introduce another variable (average time in years in their current job)

Dependent Variable: INCOME
Method: Least Squares
Date: 07/21/14 Time: 14:33
Sample: 1 16
Included observations: 16

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-88501.46	75280.57	-1.175622	0.2646
AGE	-868.7334	1737.290	-0.500051	0.6269
YRS IN JOB	23030.39	23097.61	0.997090	0.3401
EDUCATION	809.4847	481.6798	1.680545	0.1210
GENDER	18579.48	2440.022	7.614472	0.0000

R-squared	0.923199	Mean dependent var	49312.50
Adjusted R-squared	0.895271	S.D. dependent var	14036.71
S.E. of regression	4542.544	Akaike info criterion	19.93067
Sum squared resid	2.27E+08	Schwarz criterion	20.17210
Log likelihood	-154.4453	Hannan-Quinn criter.	19.94303
F-statistic	33.05663	Durbin-Watson stat	0.924386
Prob(F-statistic)	0.000005		

- Notice that the sign on AGE changed
- Also AGE is no longer significant
  Std. Error (Coefficient) went way up!
- AGE and YRS IN JOB are 0.998 correlated

Rotate variables in and out to see which specification yields best overall fit