

---

# Fitting Equations to Data

---

URBPL 5/6010: Urban Research  
University of Utah  
Pam Perlich  
Rev. 10/19/2006

---

# Why Fit Equations to Data?

- To establish whether or not there is:
  - A time trend in time series data
  - A relationship between two or more different sets of data
- Given the existence of a relationship, fitting equations enables us to identify the mathematical function that best captures the relationship

---

# Meanings of Coefficients

- $0 \leq R^2 \leq 1$ 
  - $0 \rightarrow$  No relationship
  - $1 \rightarrow$  Perfect fit
- $R$  : correlation coefficient
  - Square root of  $R^2$  and signed according to the direction of the relationship
  - $-1 \leq R \leq 1$ 
    - $1 \rightarrow$  Perfect fit, positive relationship
    - $-1 \rightarrow$  Perfect fit, inverse relationship
    - $0 \rightarrow$  No relationship

---

# Relationship Between Data Series

- Two sets of data may be related.
- These may be interpreted as a cause-effect relationship.
- This relationship may be captured using a mathematical formula.
- This function may be linear or nonlinear.

---

## Data as X-Y Pairs

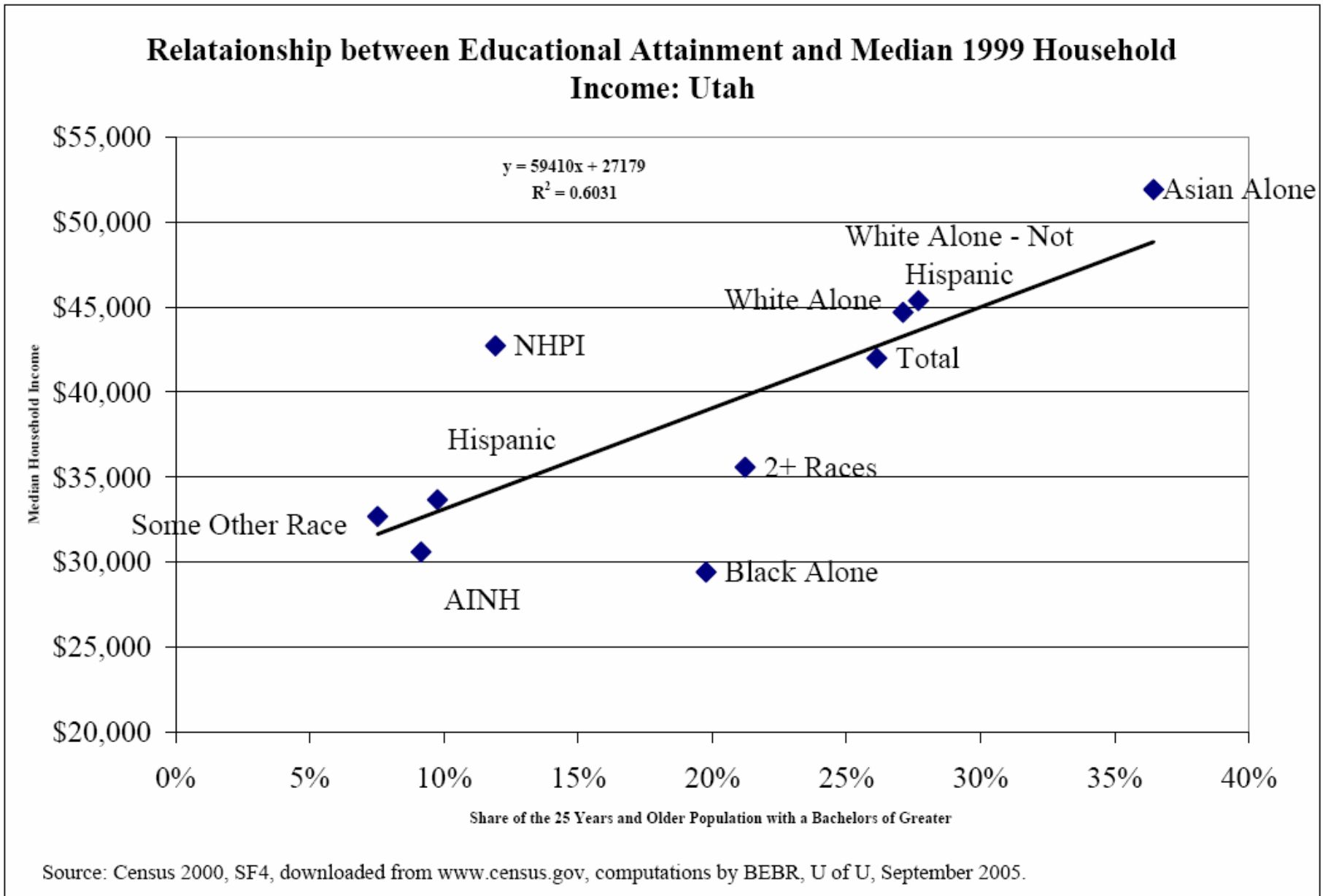
- Data may be represented in two dimensional space (on an  $x - y$  axis)
- These pairs of data can be expressed as data points:  $x_1, y_1$
- In Excel, these are graphed using the “X-Y Scatter” type.

---

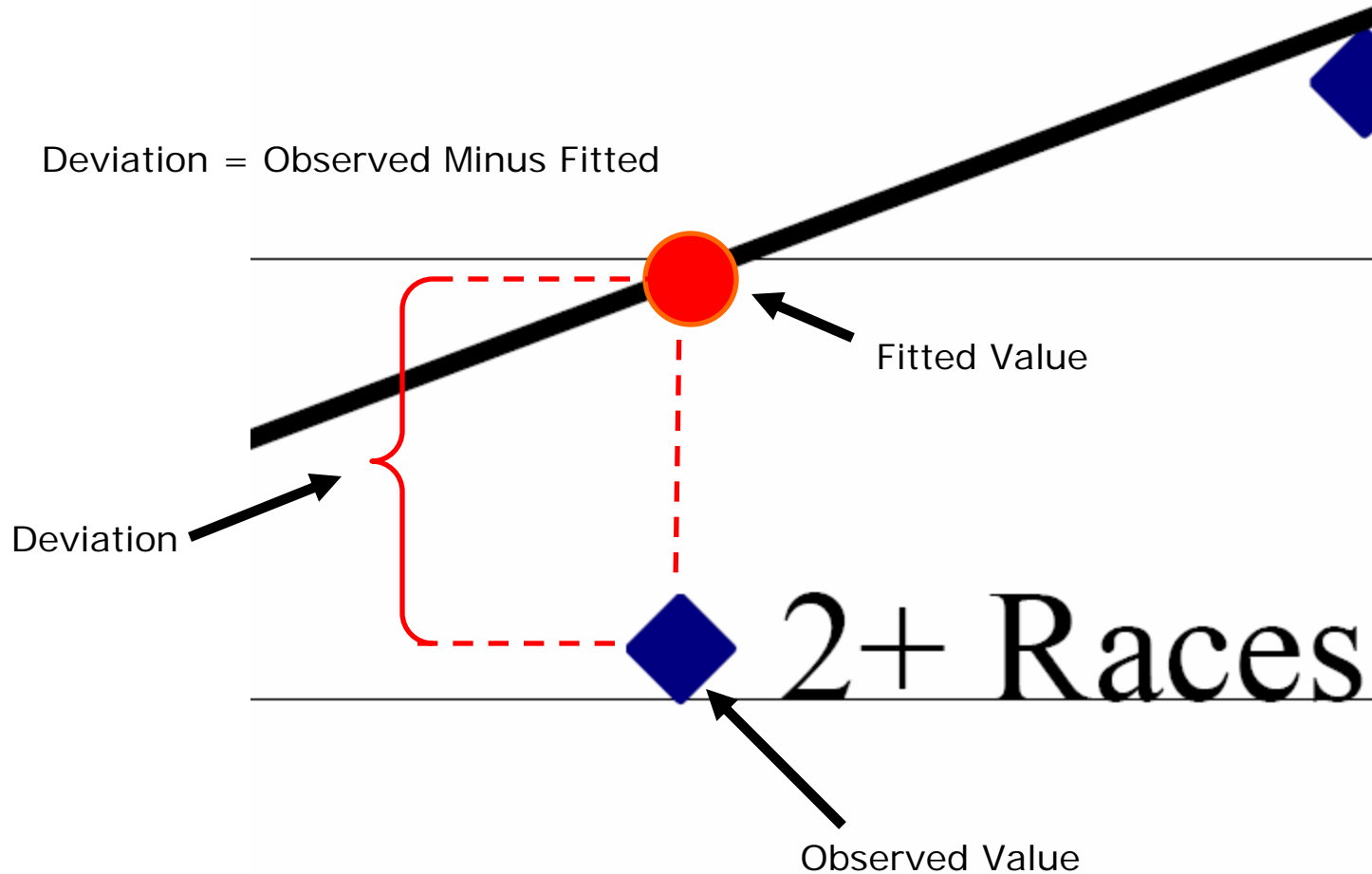
# Examples

- Horsepower of automobiles and miles per gallon
- Number of vehicles on a road and amount of emissions generated
- Regional economic growth rates and net immigration rates
- Amounts of emission and incidence of respiratory disease

# Fit a Function to the Data



# Fit a Function to the Data



---

# Least Squares Method

- Find the line that minimizes the squared differences between the observed dependent variable and the calculated dependent variable.
- $y=ax+b$  is the linear function
- $(x',y')$  is the observed pair.
- Minimize the sum of all  $(y'-y)^2$

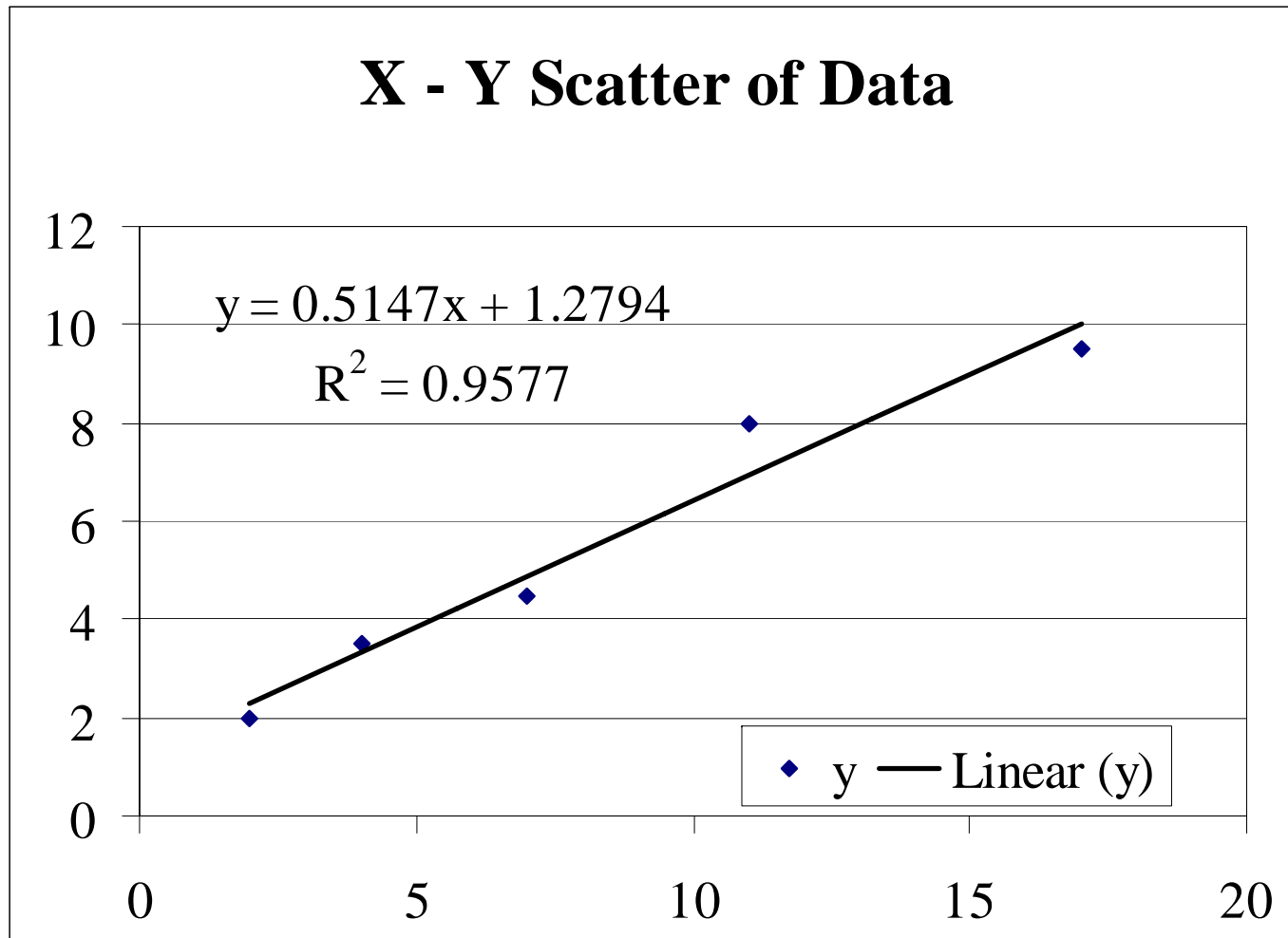
---

## Solve these Simultaneous Equations

$$a \sum_{i=1}^n x_i + bn = a \sum_{i=1}^n y_i$$

$$a \sum_{i=1}^n x_i^2 + b \sum_{i=1}^n x_i = \sum_{i=1}^n x_i y_i$$

# Example: Scatter Plot



From: Gottfried, page 106

# Solution in Excel Matrix Algebra

In Matrix Form						
	41	5		a		27.5
	479	41	*	b	=	299
	a			-0.0574	0.0070	27.5
	b	=		0.6709	-0.0574	299
	a			0.514706		
	b	=		1.279412		

Inverse

From: Gottfried, page 106

---

# Goodness of Fit Measure

Sum of Squared Errors

$$SSE = \sum_{i=1}^n [y_i - f(x_i)]^2$$

---

# Goodness of Fit Measure

r-squared

$$r^2 = 1 - \frac{SSE}{SST}$$

Where:

$$SST = \sum_{i=1}^n [y_i - \bar{y}]^2$$

---

# Values of $r^2$

- $0 \leq r^2 \leq 1$
- As  $r^2$  approaches 1, the fit is better
- As  $r^2$  approaches 0, the fit is worse



---

# “Add a Trend Line” in Excel

- Plot the data in an x-y scatter
- Right click the series on the graph
- Select “Add a Trend Line”
- Select “Linear” from the Type tab
- Select “Display Equation” and “Display r squared” from the “Options” tab
- Note: for other applications you may select “Forecast” as well

---

# Using the Analysis Tool Pack

- Enter the data into the worksheet
- From “Tools” menu, select “Data Analysis/Regression Tool.
- Complete the required selections.

C6	=	x														
A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q

x	y
2	2
4	3.5
7	4.5
11	8
17	9.5

### Regression

**Input**

Input y Range:

Input x Range:

Labels  Constant is Zero

Confidence Level  %

**Output options**

Output Range:

New Worksheet Ply:

New Workbook

**Residuals**

Residuals  Residual Plots

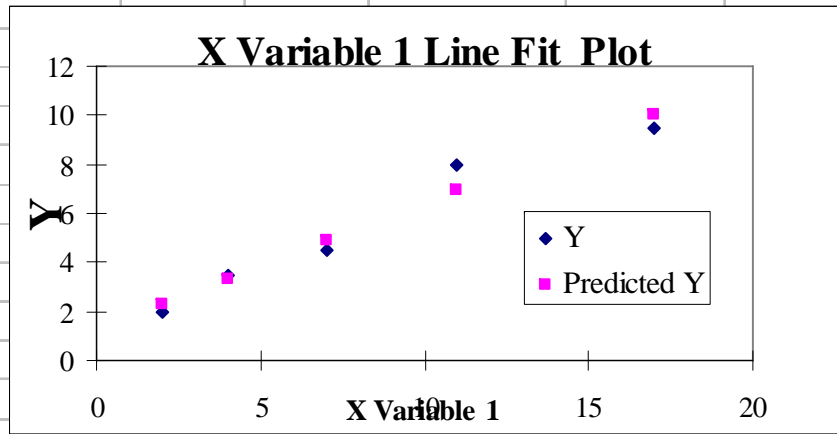
Standardized Residuals  Line Fit Plots

**Normal Probability**

Normal Probability Plots

# Results

x	y	SUMMARY OUTPUT								
2	2									
4	3.5									
7	4.5	<i>Regression Statistics</i>								
11	8	Multiple R	0.978643887							
17	9.5	R Square	0.957743857							
		Adjusted R Square	0.943658476							
		Standard Error	0.745903847							
		Observations	5							
		<i>ANOVA</i>								
		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>				
		Regression	37.83088235	37.83088	67.99559	0.003734402				
		Residual	1.669117647	0.556373						
		Total	39.5							
		<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>	<i>Lower 95.0%</i>	<i>Upper 95.0%</i>	
		Intercept	1.279411765	0.610944132	2.094155	0.127272	-0.664886955	3.223710484	-0.664886955	3.223710484
		X Variable 1	0.514705882	0.062419278	8.245944	0.003734	0.316059694	0.71335207	0.316059694	0.71335207



---

# Summary

- Simple linear regression fits a line to a set of  $x,y$  coordinates.
- This procedure minimizes squared errors.
- $r^2$  is a measure of goodness of fit
  - The better the fit, the closer  $r^2$  is to 1
- There are multiple ways to compute linear regression in Excel.

---

# Extensions

- You can fit other (nonlinear) functions to data.
  - Shape of the function tell you about the series.
  - Polynomials is many degrees have so many “bends” that you can fit a series very well.  
Interpretation may be difficult.
- Multivariate regression (linear or nonlinear) incorporates many independent variables.