

**ACE 427
Spring 2009**

Lecture 4

Statistics Review and Trend Models

**by
Professor Scott H. Irwin**

Required Reading:

Key, N. and M.J. Roberts. “Measures of Trends in Farm Size Tell Differing Stories.” *Amber Waves*, November 2007, pp. 36-37. (427 compass website)

Tannura, M., S. Irwin, and D. Good. “Are Corn Trend Yields Increasing at a Faster Rate?” Marketing and Outlook Brief 2008-02, Department of Agricultural and Consumer Economics, University of Illinois at Urbana-Champaign, February 2008. (427 compass website)

Optional Readings:

Schwager, J.D. “Ch. 16: A Review of Elementary Statistics,” and “Ch. 15: Introduction to Regression Analysis,” *Schwager on Futures: Fundamental Analysis*, New York, NY: John Wiley and Sons, 1995. (427 compass website)

Introduction

- _____ is a measure that is calculated from the values of a variable
- Common usage of “statistic” is different: _____

Univariate (Single Variable) Statistics

- Just looking at the raw data usually is not very enlightening
- Need a way to _____ key _____ of data quickly and easily
- Two characteristics are most important:
 - _____: Where is the _____ of the data?
 - _____: How _____ out are the data?

Measures of Central Tendency

- A statistic that measures the _____ or _____ value of a variable
- Most common measure is the _____:

$$\bar{x} = \frac{1}{T} \sum_{t=1}^T x_t$$

- Another useful measure is the _____:

$$x_{med} = \mathit{middle}(x_{\min}, \dots, x_{\max})$$

where the observations have been ordered from smallest to largest

Measures of Dispersion

- A statistic that measures the _____ or _____ of a variable
- The simplest is the _____:

$$\text{Range} = x_{\max} - x_{\min}$$

- The most widely used measure of dispersion is the _____:

$$s_x^2 = \frac{1}{T-1} \sum_{t=1}^T (x_t - \bar{x})^2$$

What are the units of this statistic?

Standard Deviation

- Standard deviation formula:

$$s_x = \sqrt{\frac{1}{T-1} \sum_{t=1}^T (x_t - \bar{x})^2}$$

- Taking the square root gets back to the original _____ of the variable

Interpretation of Standard Deviation

- _____

- _____ imply about half of the observations have positive deviation from the mean of this size and about half have negative deviations of this size

- _____ imply that about two-thirds of the observations have positive or negative deviations of this size above or below the mean

Bivariate Statistics

- Here, the focus is on _____ between two variables
- Two sets of variables x and y , each with T observations
- T pairs of observations

Covariance

- A measure of the _____ between x and y

$$s_{xy} = \frac{1}{T-1} \sum_{t=1}^T (x_t - \bar{x})(y_t - \bar{y})$$

- Units of s_{xy} are strange: (units x)(units y)

- _____ remedies units problem of covariance statistic:

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

- Effect of dividing by product of standard deviations is to limit the values of r_{xy} to -1 to $+1$

Interpreting Correlation Coefficients

- $+1$ indicates _____ correlation between x and y
- 0 indicates no correlation between x and y
- -1 indicates _____ correlation between x and y

Does Correlation Imply Causation?

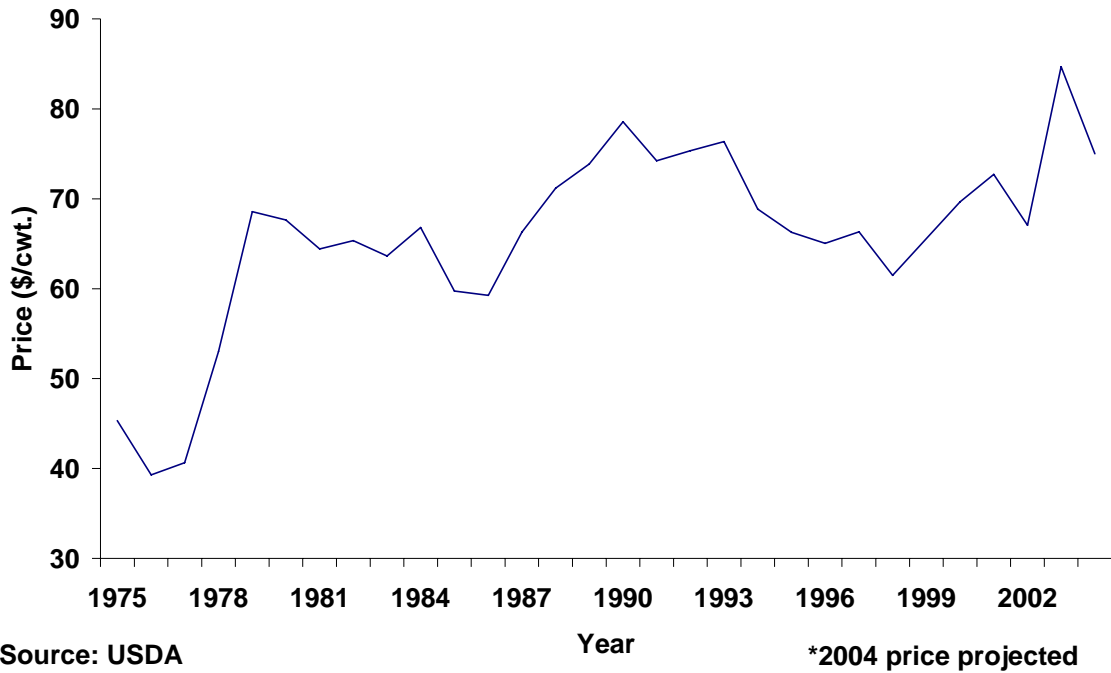
- Many variables are highly correlated, but do not have a _____ relationship
- Two variables may be strongly affected by a _____
- Time series data are plagued by this problem, because most economic variables grow steadily over time
- Economists give this problem a special name: _____

Price of US Slaughter Steers and Hogs, 1975-2004

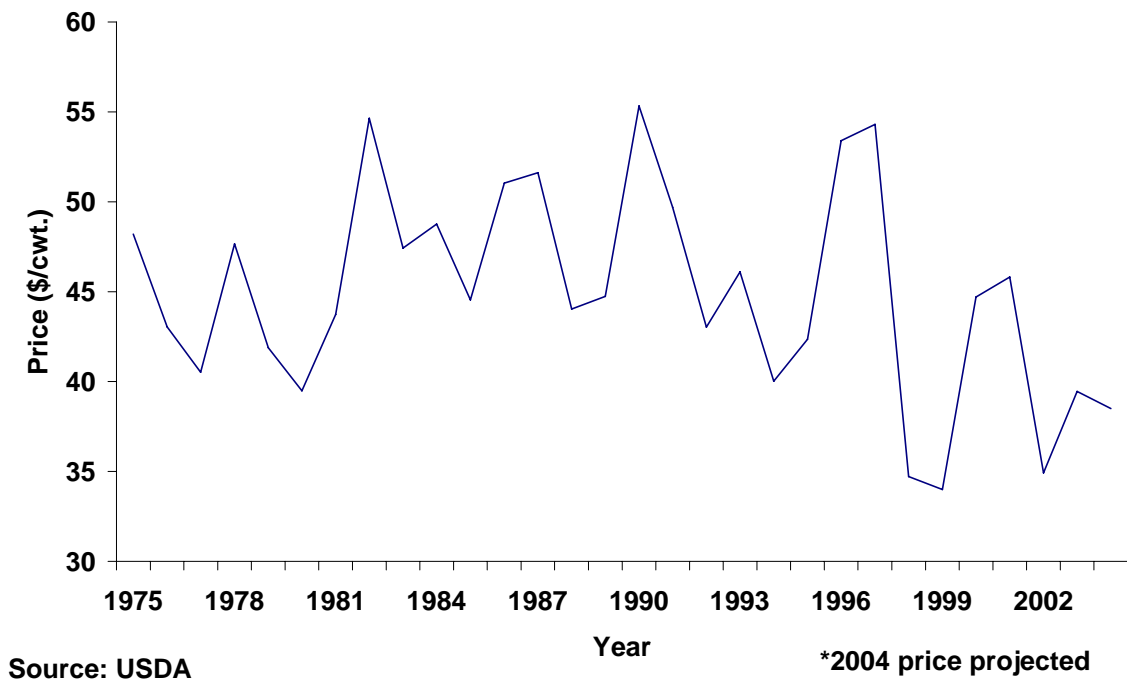
Year	US Slaughter Steer Price	US Hog Price
---\$/cwt.---		
1975	45	48
1976	39	43
1977	41	41
1978	53	48
1979	69	42
1980	68	39
1981	64	44
1982	65	55
1983	64	47
1984	67	49
1985	60	45
1986	59	51
1987	66	52
1988	71	44
1989	74	45
1990	79	55
1991	74	50
1992	75	43
1993	76	46
1994	69	40
1995	66	42
1996	65	53
1997	66	54
1998	61	35
1999	66	34
2000	70	45
2001	73	46
2002	67	35
2003	85	39
2004	75	39

Note: 2004 prices are projected. The source for the data is the USDA.

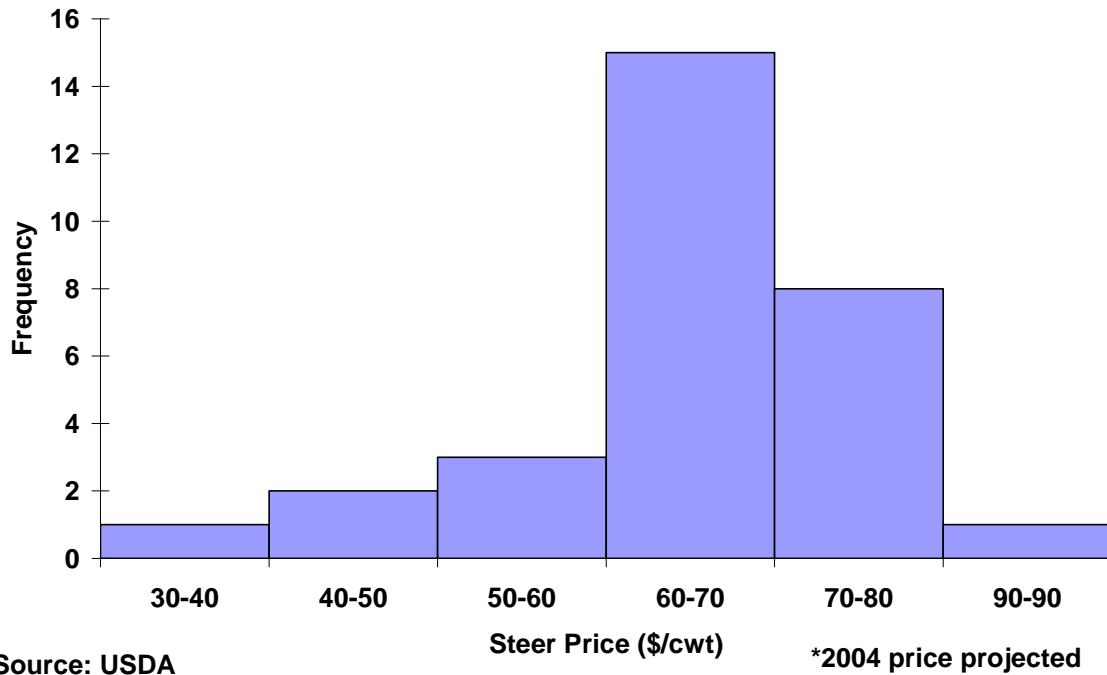
US Slaughter Steer Price, 1975-2004*



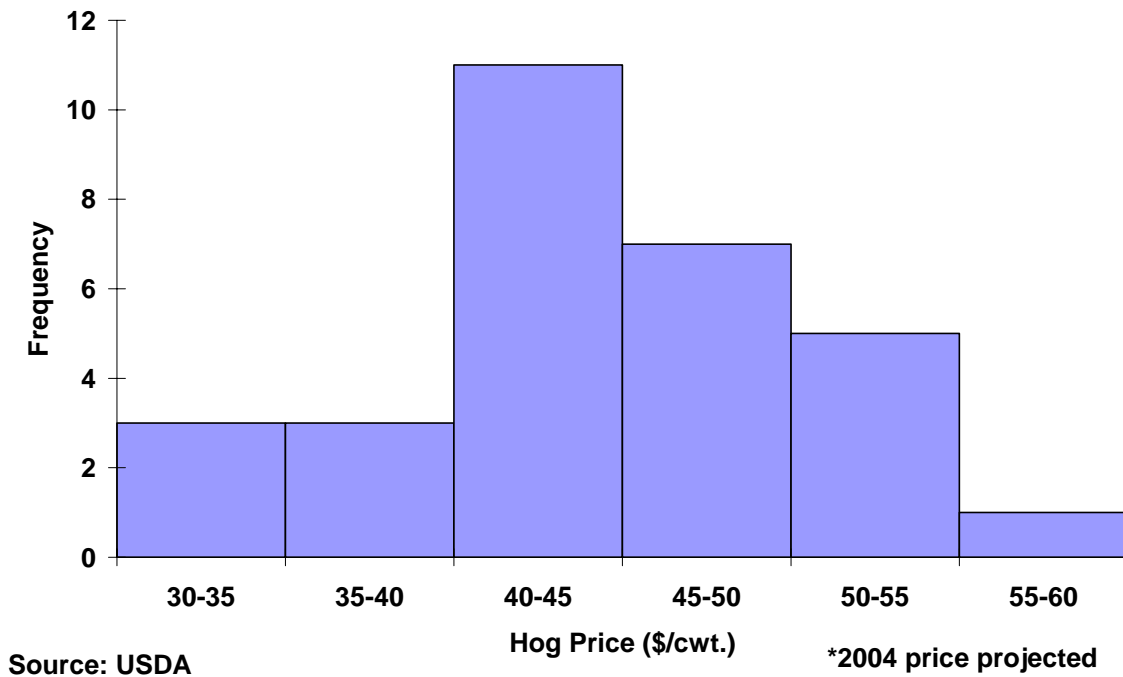
US Price of Hogs, 1975-2004*



Frequency Distribution of US Slaughter Steer Prices, 1975-2004*



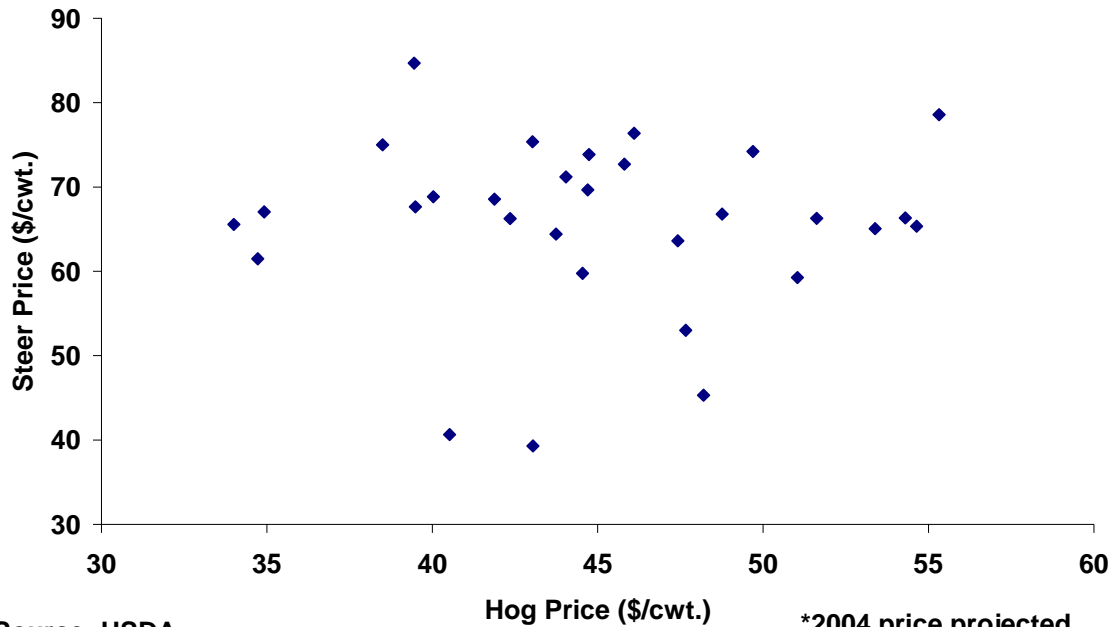
Frequency Distribution of US Hog Prices, 1975-2004*



Excel Output from Descriptive Statistics Option in Data Analysis

	<i>US Slaughter Steer Price</i>		<i>US Hog Price</i>
Mean	65.73	Mean	44.92
Standard Error	1.88	Standard Error	1.07
Median	66.56	Median	44.62
Mode	#N/A	Mode	#N/A
Standard Deviation	10.31	Standard Deviation	5.88
Sample Variance	106.30	Sample Variance	34.53
Kurtosis	1.50	Kurtosis	-0.58
Skewness	-1.07	Skewness	0.00
Range	45.40	Range	21.32
Minimum	39	Minimum	34
Maximum	85	Maximum	55
Sum	1972.04	Sum	1347.64
Count	30	Count	30
Confidence Level(95.0%)	3.85	Confidence Level(95.0%)	2.19

XY Scattergraph of the Relationship Between US Slaughter Steer and Hog Prices, 1975-2004*



Excel Output from Correlation Option in Data Analysis

	<i>US Steer Price</i>	<i>US Hog Price</i>
<i>US Steer Price</i>	1	
<i>US Hog Price</i>	0.01	1

Least Squares Regression and Forecasting

“Regression analysis is probably the single most important tool in interpreting and applying fundamental information.”

---Jack Schwager , *Fundamental Analysis*

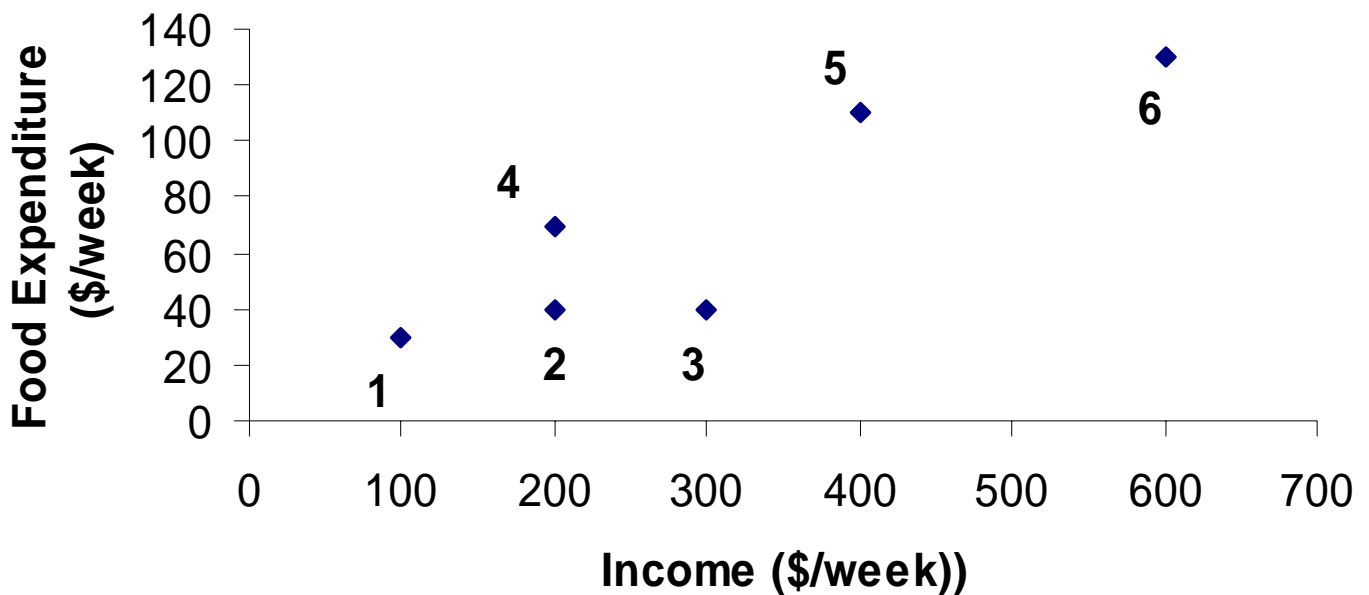
Example: Weekly Food Expenditure For a Family

Week	Food Expenditure
1	\$30
2	40
3	40
4	70
5	110
6	130

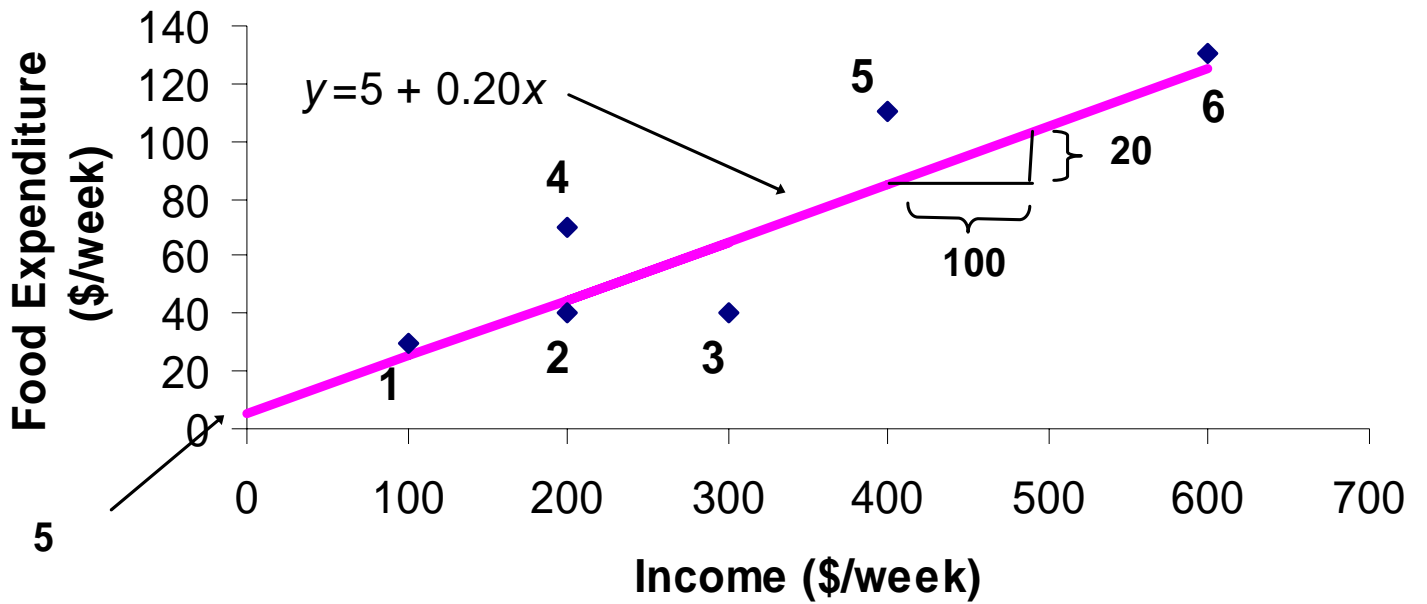
Weekly Food Expenditure and Weekly Income for a Family

Week	Food Expenditure	Income
1	\$30	\$100
2	40	200
3	40	300
4	70	200
5	110	400
6	130	600

Scatter Diagram of Food Expenditure (y) vs. Family Income (x)



“Eyeball” Line for Average Relationship Between Food Expenditures and Family Income



Problems With “Eyeball” Econometrics

- Highly _____
- Different analysts may choose different lines
- Tendency to ignore outliers (extreme observations)
- Works only in two dimensions

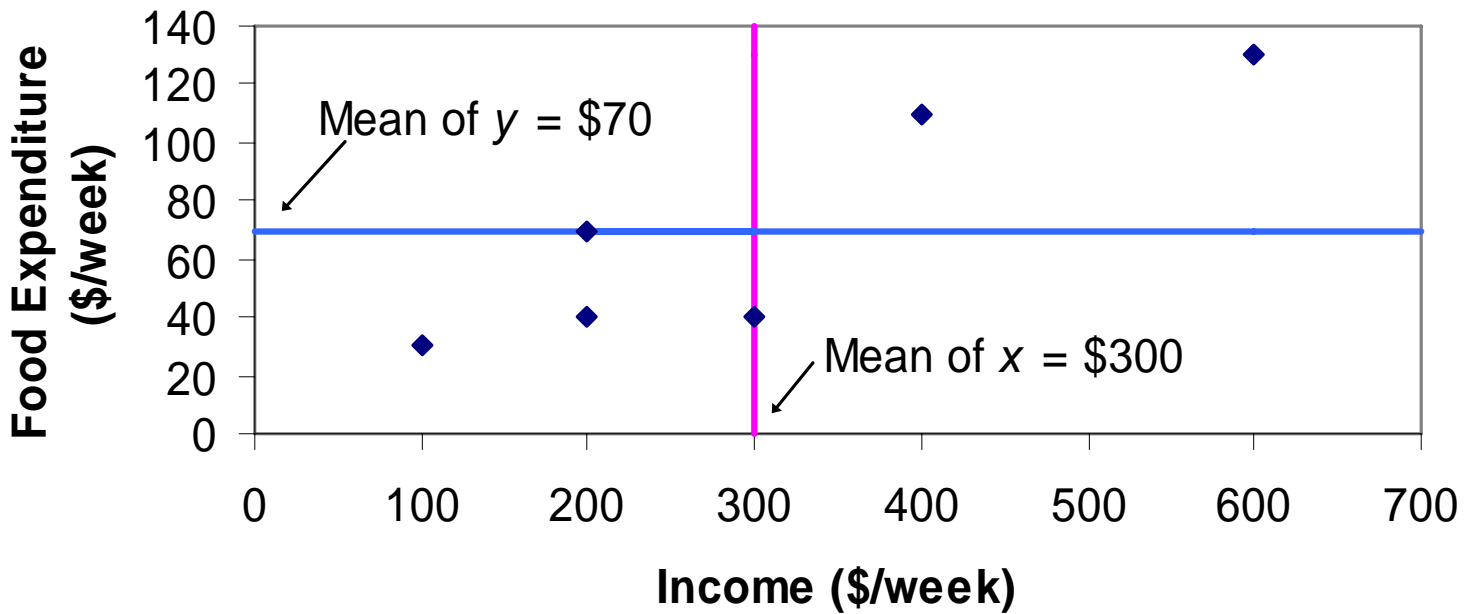
The Method of Least Squares

- Least squares relies on the _____ of mathematics to select a line of _____ relationship
- Often referred to as _____ line
- We will introduce the technique of least squares graphically and then mathematically

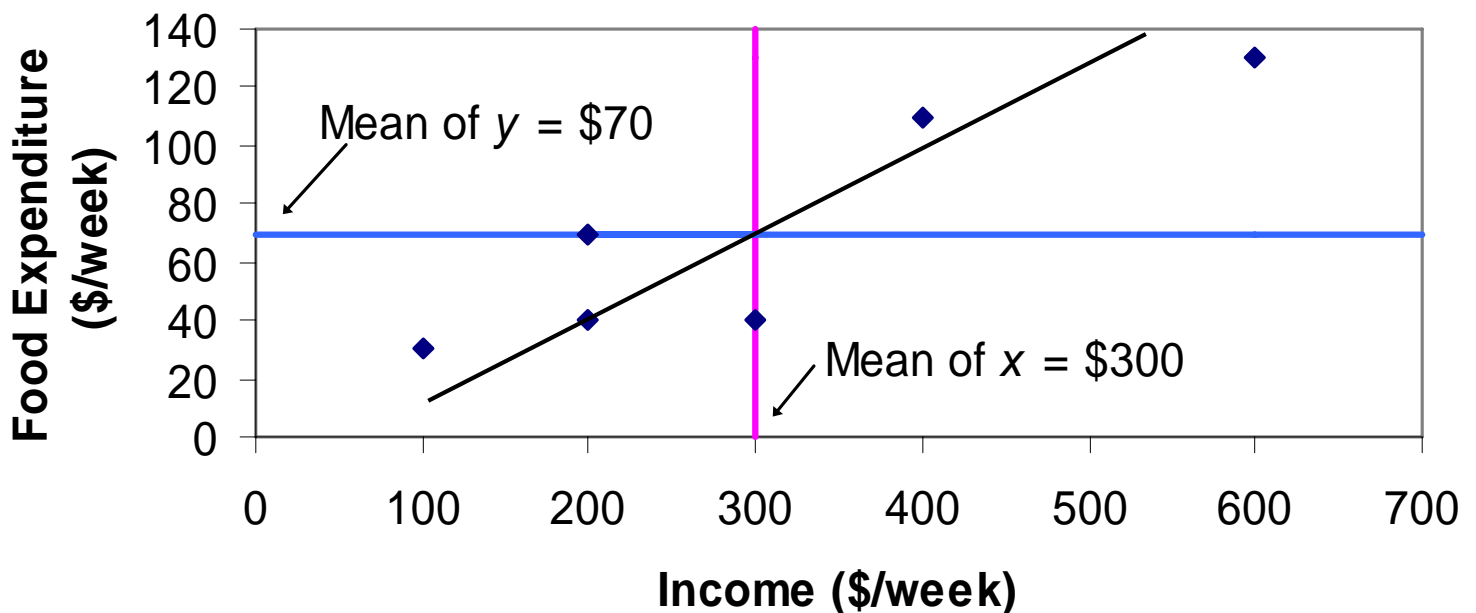
Essential Steps in Least Squares

- Find the _____ of y and x
- _____ line at the means of y and x
- _____ line through “anchor” to find the line that _____ (squared) between the line and data points

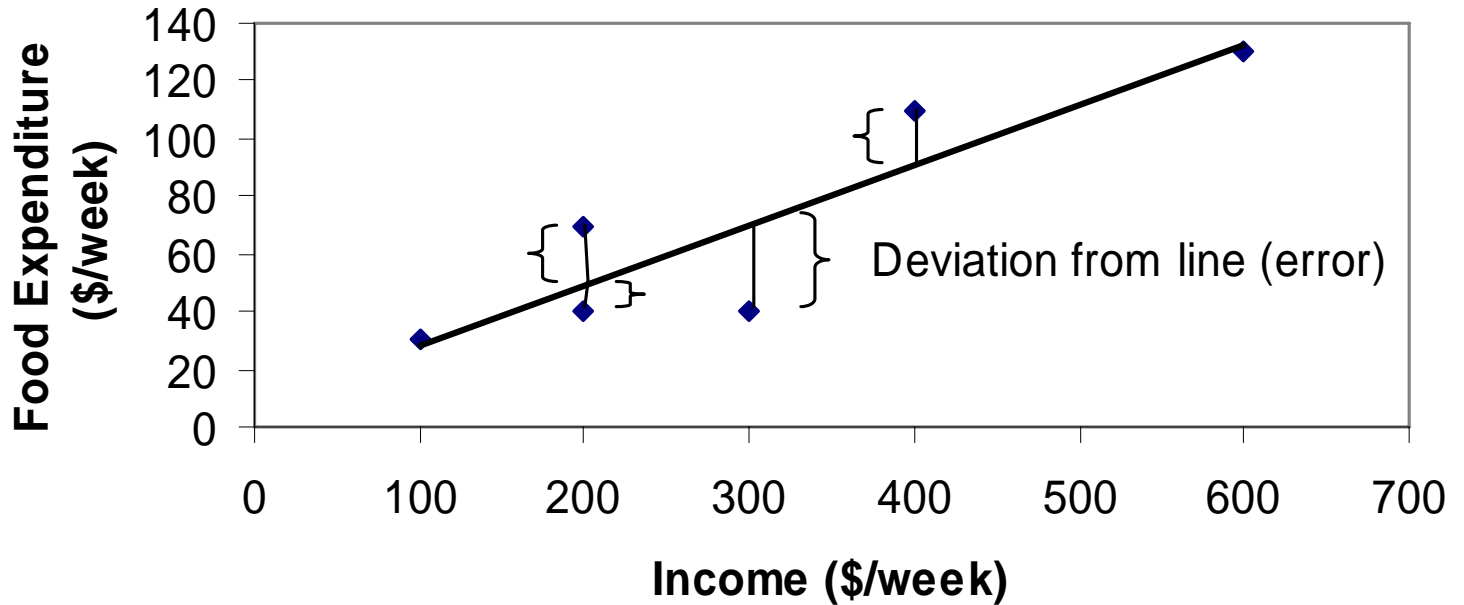
Step 1 of Least Squares: Compute Means of y and x



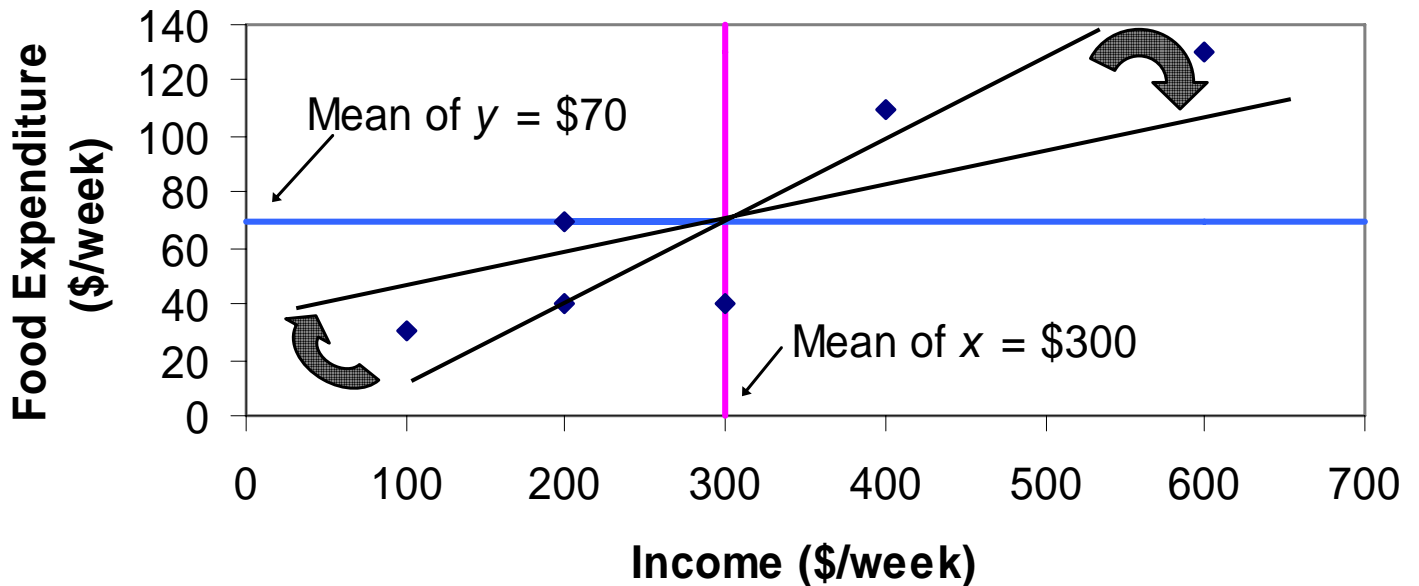
Step 2 of Least Squares: Force Line Through Means of y and x



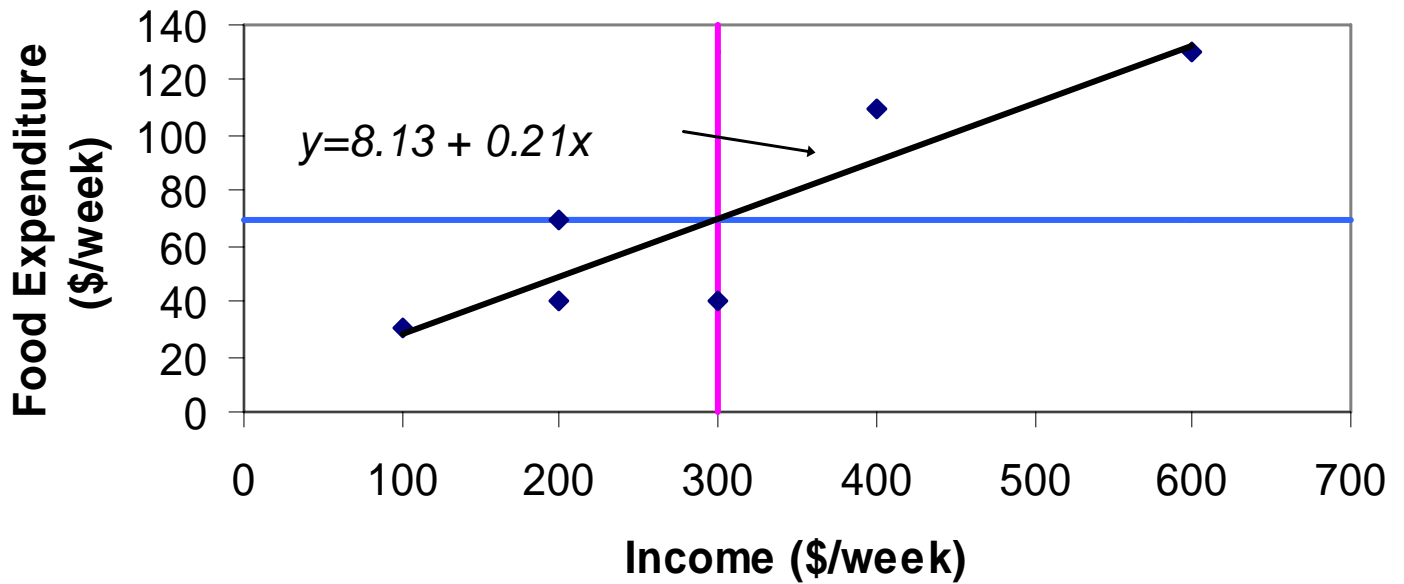
Step 3 of Least Squares: Compute Deviations from Line



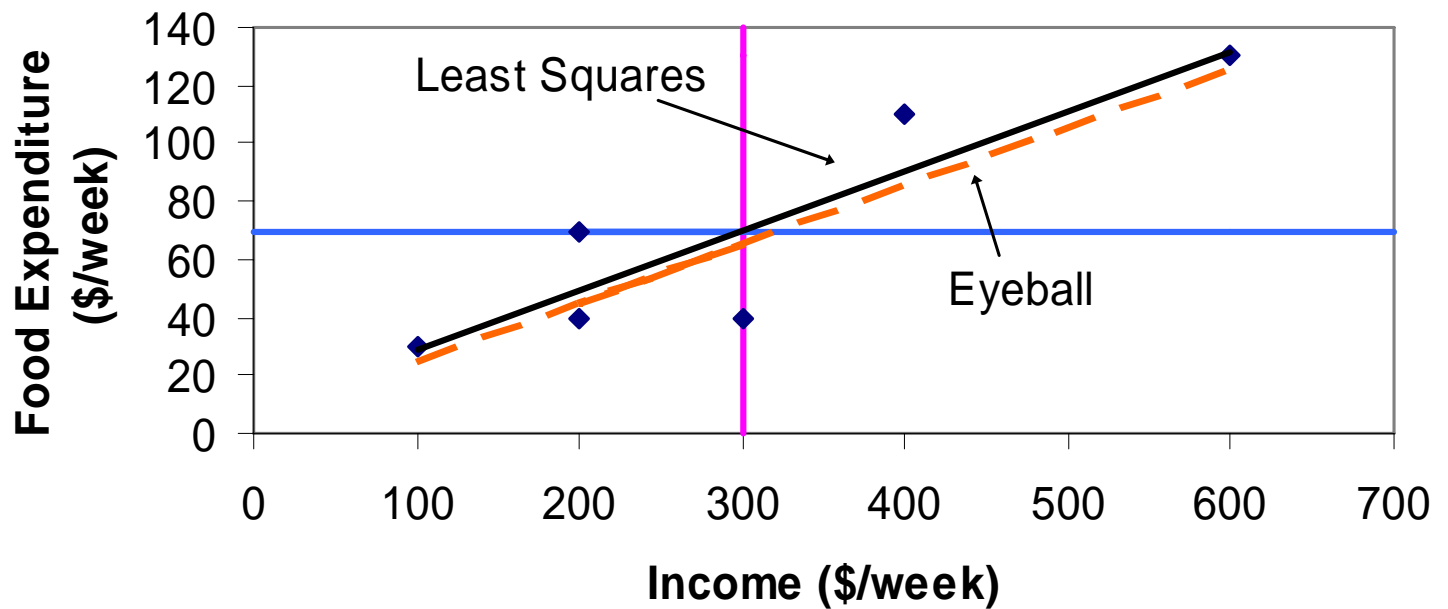
Step 4 of Least Squares: Rotate Line to Find Minimum Sum of Squared Deviations from Line



Estimated Least Squares Line



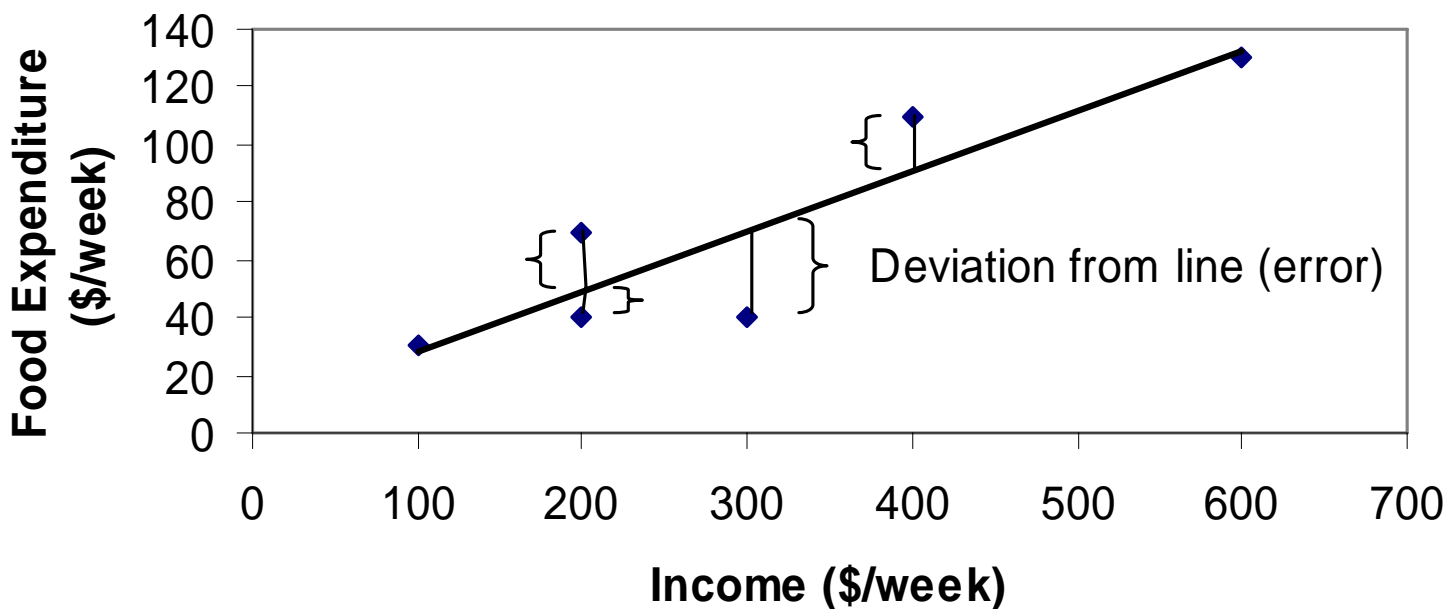
Comparison of Least Squares and Eyeball Lines



Mathematical Development of Least Squares Line

- The graphical approach is useful to develop intuitive understanding of least squares approach to fitting a line to data
- Mathematical development is needed to actually compute needed quantitative estimates of intercept and slope

Regression Error Term



Why Does the Error Term Exist in the First Place?

- _____ variables
- _____ error if relationship between y and x is not exactly a perfectly linear relationship.
- Strictly _____ behavior that may be unique to observation

Objective in Least Squares Regression: Minimize Sum of Squared Errors

Difference between actual values of y and values of y predicted by line:

$$e_t = y_t - b_1 - b_2 x_t$$

(error can be positive or negative value)

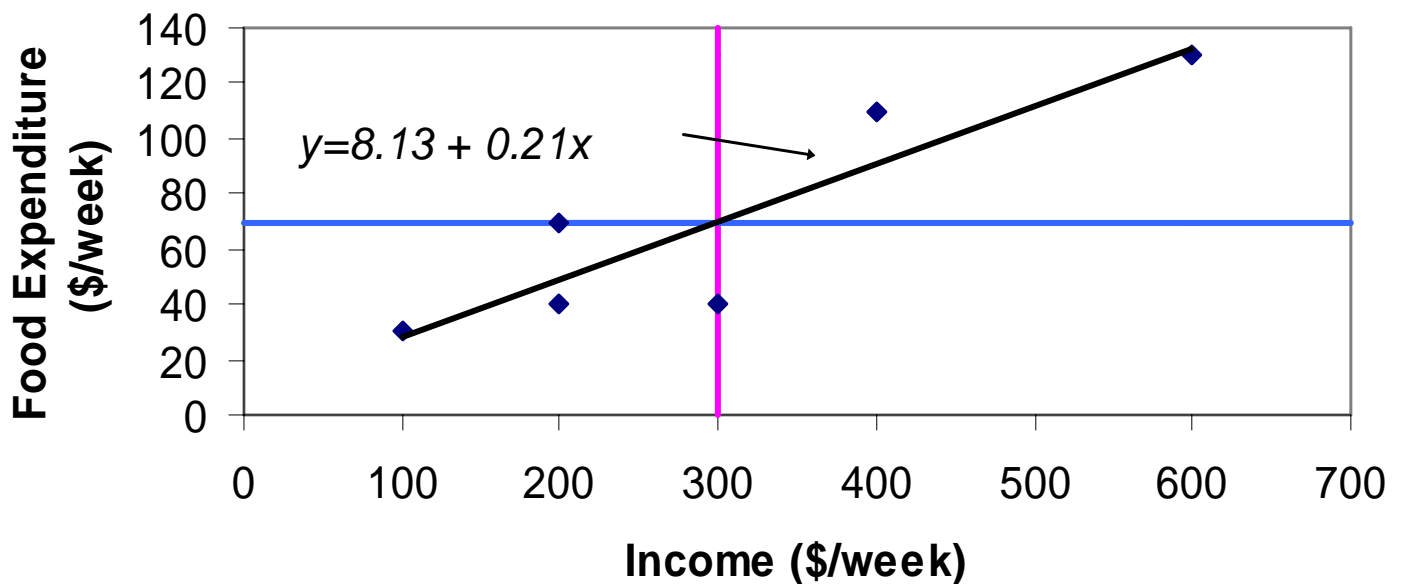
$$\text{Min SSE} = \sum_{t=1}^T e_t^2 = \sum_{t=1}^T (y_t - b_1 - b_2 x_t)^2$$

Formulas for Regression Coefficients

$$b_1 = \bar{y} - b_2 \bar{x}$$

$$b_2 = \frac{T \sum_{t=1}^T y_t x_t - \sum_{t=1}^T x_t \sum_{t=1}^T y_t}{T \sum_{t=1}^T x_t^2 - \left(\sum_{t=1}^T x_t \right)^2}$$

Estimated Least Squares Line for Food Expenditure Example



Sample Regression Output from Excel

<i>Regression Statistics</i>	
Multiple R	0.89
R Square	0.79
Adjusted R Square	0.74
Standard Error	21.18
Observations	6

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	6806.25	6806.25	15.18	0.02
Residual	4	1793.75	448.44		
Total	5	8600			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	8.13	18.08	0.45	0.68	-42.08	58.33
X Variable 1	0.21	0.05	3.90	0.02	0.06	0.35

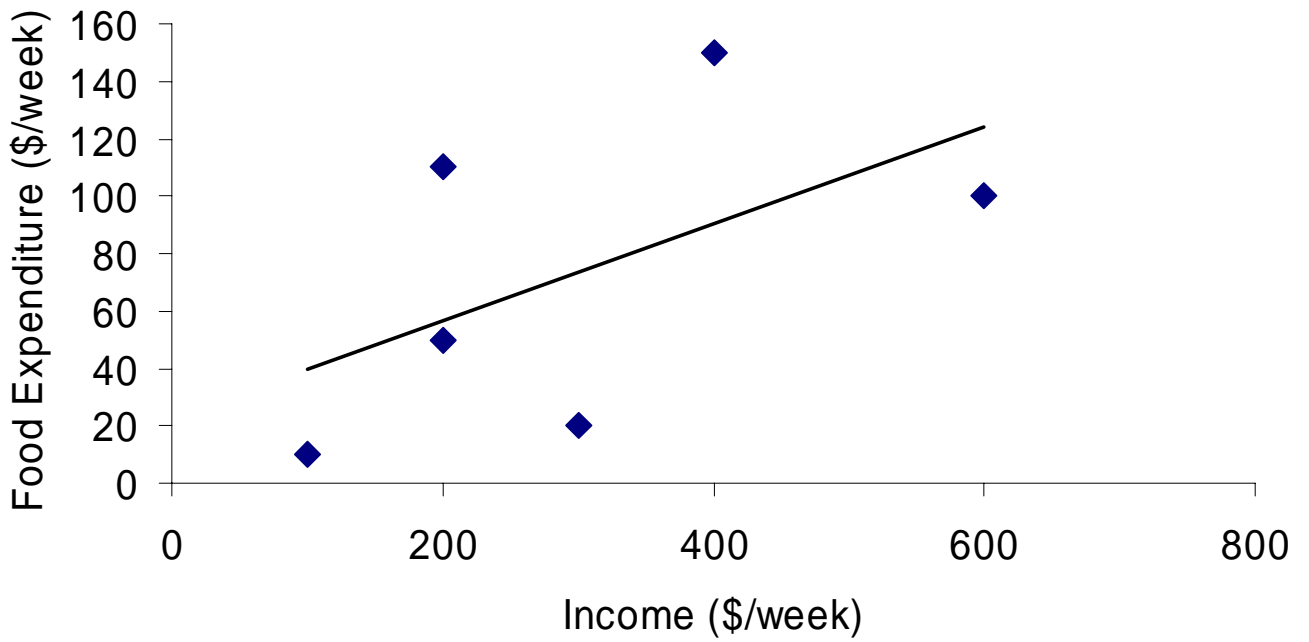
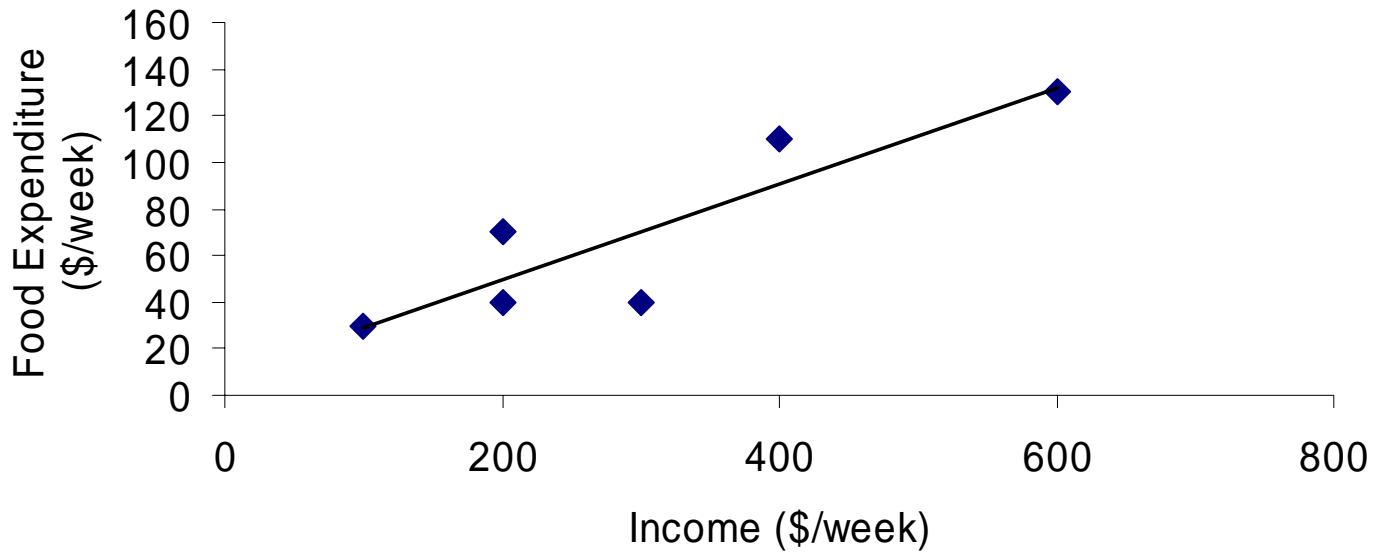
Interpreting the Slope

- In all cases, the slope of a least squares regression is the _____

- Slope = 0.21 in previous example
- If income per week increases one unit, food expenditure per week will increase _____,
or
- If income increases \$1 per week, food expenditure will increase _____

Regression Fit

- Least squares _____ that regression line is best fitting for a _____ of data
- Does not tell how well regression “fits” the observations in an _____
- Need a measure of the “goodness-of-fit” of the regression line



Definition of R^2

The proportion of _____ in y_t _____ by the regression

$$0 \leq R^2 \leq 1$$

<i>Regression Statistics</i>	
Multiple R	0.89
R Square	0.79
Adjusted R Square	0.74
Standard Error	21.18
Observations	6

ANOVA

	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	6806.25	6806.25	15.18	0.02
Residual	4	1793.75	448.44		
Total	5	8600			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	8.13	18.08	0.45	0.68	-42.08	58.33
X Variable 1	0.21	0.05	3.90	0.02	0.06	0.35

Interpreting R^2

- Correct interpretations:
 - The _____ in x explains ___ percent of the _____ in y
 - The variation in the independent variable explains 79 percent of the variation in the dependent variable
 - The regression model explains 79 percent of the variation in y

Precision of the Estimated Line

- Regressions typically are estimated using only a _____ of all possible observations (a sample)
- As a result, the estimated lines are likely to vary from one sample to the next
- Standard errors provide estimates of _____
 - Intercept: $se(b_1)$
 - Slope: $se(b_2)$

<i>Regression Statistics</i>	
Multiple R	0.89
R Square	0.79
Adjusted R Square	0.74
Standard Error	21.18
Observations	6

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	6806.25	6806.25	15.18	0.02
Residual	4	1793.75	448.44		
Total	5	8600			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	8.13	18.08	0.45	0.68	-42.08	58.33
X Variable 1	0.21	0.05	3.90	0.02	0.06	0.35

Standard Errors

- Intercept: $se(b_1)=18.08$
- Slope: $se(b_2)=0.05$
- Interpretation:
 - The “typical” error in estimating the intercept, without regard to sign, is 18.08
 - The _____

95% Confidence Interval

$$b_2 \pm t_{0.05/2, T-1} \cdot se(b_2)$$

$$0.21 \pm 2.776 \cdot 0.05$$

$$0.06 \text{ to } 0.35$$

- Interpretation:

We are _____

<i>Regression Statistics</i>	
Multiple R	0.89
R Square	0.79
Adjusted R Square	0.74
Standard Error	21.18
Observations	6

ANOVA					
	<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Significance F</i>
Regression	1	6806.25	6806.25	15.18	0.02
Residual	4	1793.75	448.44		
Total	5	8600			

	<i>Coefficients</i>	<i>Standard Error</i>	<i>t Stat</i>	<i>P-value</i>	<i>Lower 95%</i>	<i>Upper 95%</i>
Intercept	8.13	18.08	0.45	0.68	-42.08	58.33
X Variable 1	0.21	0.05	3.90	0.02	0.06	0.35

Hypothesis Testing

- Hypotheses:
 - Null: Slope equals zero
 - Alternative: Slope does not equal zero

- Decision rules:
 - _____ null if hypothesized value (0) falls in 95% confidence interval
 - _____ the null and accept the alternative if the hypothesized value (0) falls outside the 95% confidence interval

Regression Application: Trend Analysis

- Identification of long-term _____ in data
- Trend can be up or down
- Trends arise because of gradual changes in _____ conditions
 - New technology
 - Changing tastes and preferences
 - New uses

Estimating Trends

- Key question is how to provide a quantitative estimate of trend
- Need a quantitative estimate to be able to provide _____ forecasts
- To illustrate the process we will estimate the trend in real hog prices

Specification of a Trend Regression

$$y_t = b_1 + b_2 x_t$$

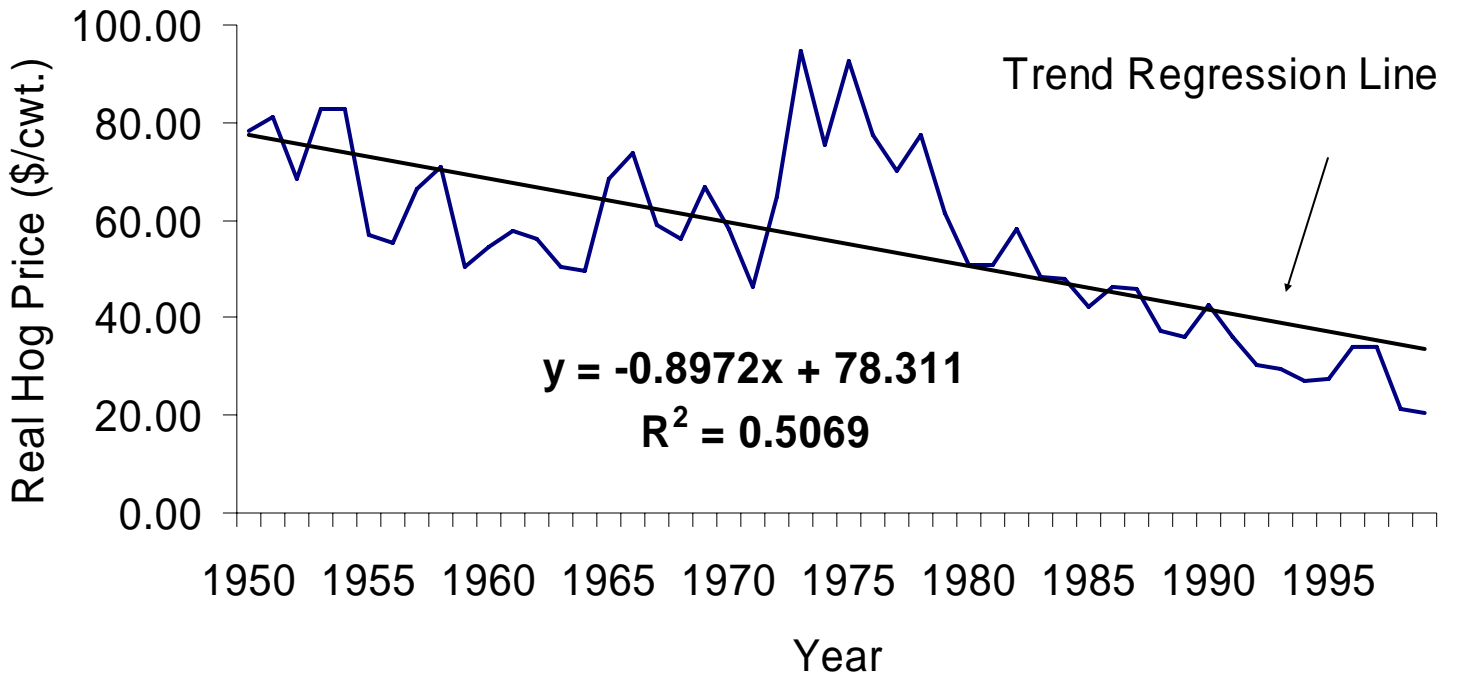
- y is the economic variable of interest
 - Real (inflation-adjusted) price of hogs in the US

- x is a time index
 - Index can be specified as _____:
 - 1980, 1981, 1982, 1983,....., 1999
 - Index can also be specified as the ____ of years in sample:
 - 1, 2, 3, 4, 5,....., 20

First 10 Observations for a Trend Regression

	(y)	(x)
	Real	Time
Year	Hog Price	Index
1950	78.38	1
1951	80.94	2
1952	68.42	3
1953	82.67	4
1954	82.71	5
1955	56.89	6
1956	55.30	7
1957	66.27	8
1958	70.80	9
1959	50.48	10

Real Price of Hogs in the US, 1950-1999 (1982-1984=100)



Trend Regressions in Excel

- Can be estimated using the Regression option of the Tools/Data Analysis menu
- Can be estimated directly within a chart using the Trendline option of the Chart menu
- Options allow printing of equation and line

Trend Forecast of Real Hog Price in 2000

- Simply plug the forecast year index value into estimated regression
- 2000 Forecast ($x=51$)
 - $y_{2000} = -0.8972(51) + 78.311$
 - $y_{2000} = \$32.55/\text{cwt.}$ (1982-1984 dollars)

Multiple Variable Regression

- In this case, _____ is regressed on y
- Basic concepts are the same as when only one x is used
- Move to three or more dimensions to demonstrate visually!

An Example of Multiple Variable Regression

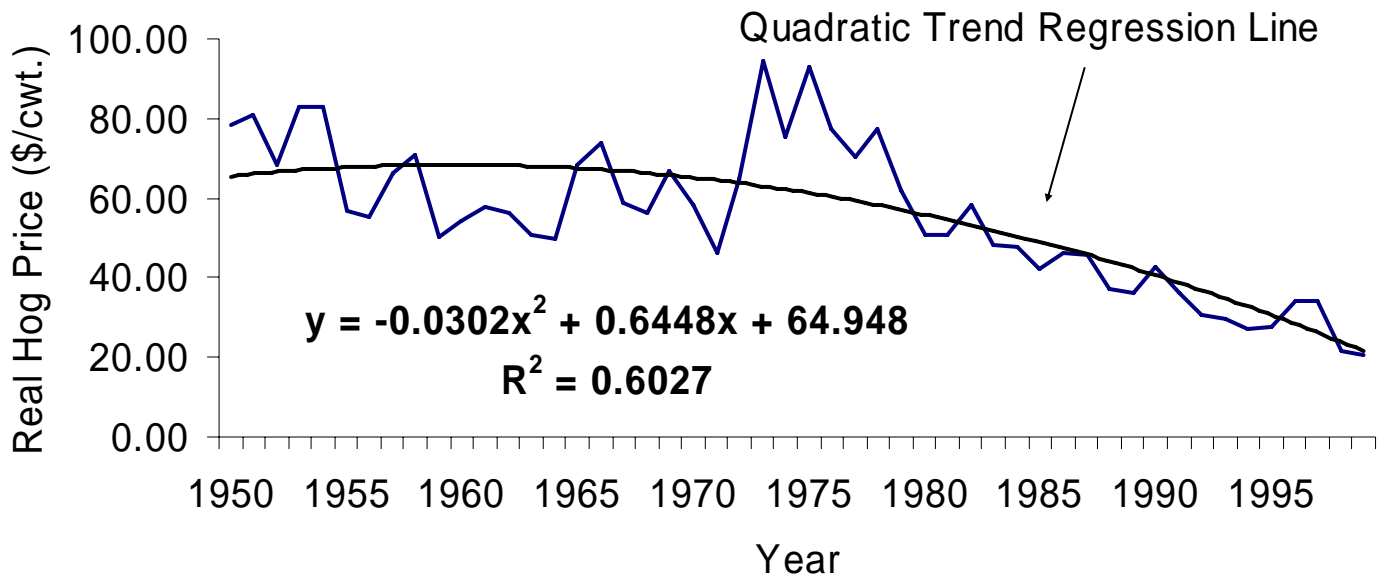
- What if we think the trend in a variable is non-linear instead of linear?
- One solution is to propose a _____ trend model:

$$y_t = b_1 + b_2 x_t + x_t^2$$

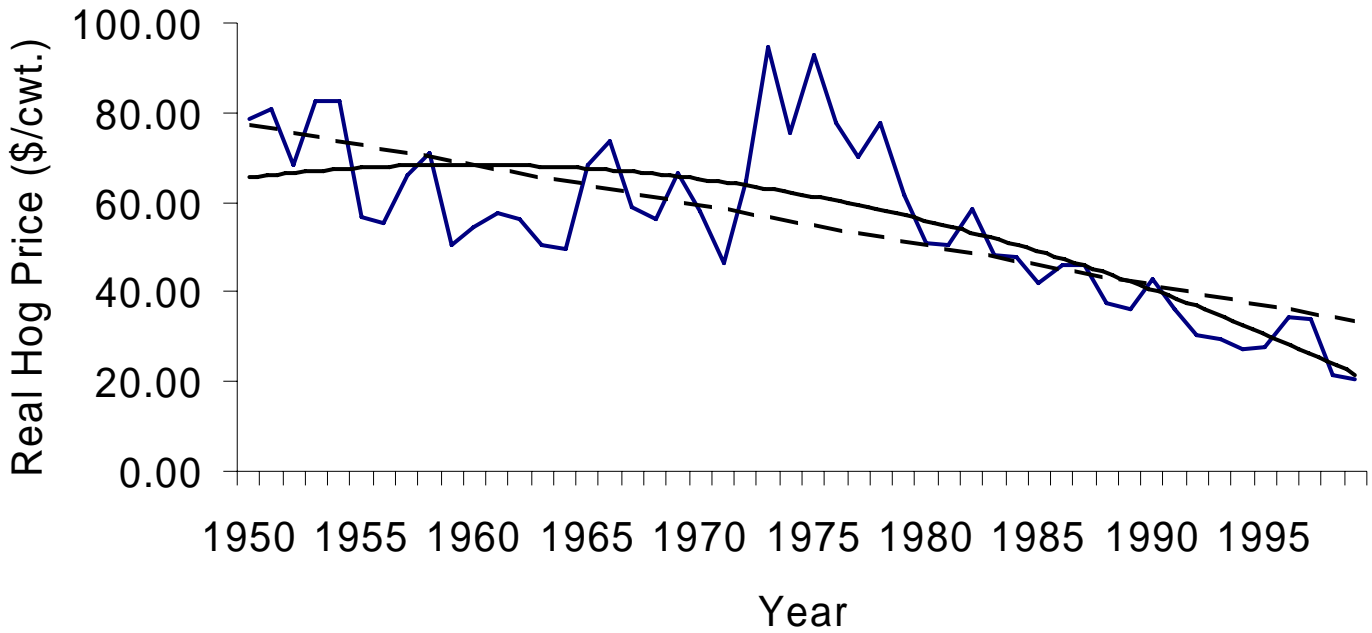
First 10 Observations for a Quadratic Trend Regression

	(y)	(x)	(x ²)
			Time
	Real	Time	Index
Year	Hog Price	Index	Squared
1950	78.38	1	1
1951	80.94	2	4
1952	68.42	3	9
1953	82.67	4	16
1954	82.71	5	25
1955	56.89	6	36
1956	55.30	7	49
1957	66.27	8	64
1958	70.80	9	81
1959	50.48	10	100

Real Price of Hogs in the US, 1950-1999 (1982-1984=100)



Real Price of Hogs in the US, 1950-1999 (1982-1984=100)



Quadratic Trend Forecast of Real Hog Price in 2000

- Again, simply plug the forecast year index value into estimated regression
- 2000 Forecast ($x=51$)
 - $y_{2000} = -0.0302(51^2) + 0.6448(51) + 64.948$
 - $y_{2000} = \$19.28/\text{cwt.}$ (1982-1984 dollars)
- *Linear trend:* $y_{2000} = \$32.55/\text{cwt.}$ (1982-1984 dollars)

Points to Remember

- Number of x variables is only limited (in theory) by the _____
- Economic theory used to guide the specification of relationships

$$PH_t = a + b_1PP_t + b_2BP_t + b_3CP_t + b_4DI_t$$

where

PH_t is the price of hogs for quarter t

PP_t is per person production of pork in quarter t

BP_t is per person production of beef in quarter t

CP_t is per person production of chicken in quarter t

DI_t is per person US disposable consumer income in quarter t