Math, Stats, and Mathstats Review

ECONOMETRICS (ECON 360)

BEN VAN KAMMEN, PHD

Outline

These preliminaries serve to signal to students what tools they need to know to succeed in ECON 360 and refresh their familiarity with these tools. These are things you would be expected to learn in introductory college courses in:

- Algebra,
- Calculus,
- Probability and Statistics.

They are also addressed in the appendices (particularly A, B, C) of the Wooldridge text (which is required for this class).

The summation operator

Notation short hand for representing the sum of a set of observations.

- Consists of an index (usually *i*) that uniquely identifies the observations,
- the sum operator, Σ,
- and the terms that must be added for all together for a given number of observations (*n*).

I.e.,

$$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots x_n.$$

The summation operator (continued)

The sum operator has some convenient properties:

1. Summing constants (anything not indexed with an *i*) does not really require a sum operator.

$$\sum_{i=1}^{n} c = c + c + \ldots + c = c(1 + 1 + \ldots 1) = cn.$$

The summation operator (continued)

2. Constants can be "pulled through" or factored out of a term using a sum operator.

$$\sum_{i=1}^{n} c x_i = c \sum_{i=1}^{n} x_i \,.$$

3. Sums are commutative.

$$\sum_{i=1}^{n} (a x_i + b y_i) = a \sum_{i=1}^{n} x_i + b \sum_{i=1}^{n} y_i \,.$$

The summation operator (concluded)

4. The sum of a fraction does not equal the fraction of the sums.

$$\sum_{i=1}^{n} \frac{x_i}{y_i} \neq \frac{\sum_{i=1}^{n} x_i}{\sum_{i=1}^{n} y_i}.$$

There are other clever “tricks” that involve simplifying the terms inside a sum operator that are useful and will be demonstrated later.

Functions

Function: a rule that associates a value of (variable) x with exactly one value of (variable) y.

When y is a function of x, write $y = f(x)$.

Economics applications:

- *demand* and *supply* functions that associate a price with a quantity demanded or supplied,
- production functions (associate quantities of inputs with quantities of outputs),
- cost functions (associate quantities of outputs with costs incurred),
- profit functions (quantity and profit),
- utility functions (quantity and utility).

Functions (continued)

Functions can involve:

- the basic operations: "+", "-", "×","÷",
- exponents and roots,
  - e.g., $f(x) = x^2$. The superscript "2" (the *exponent*) indicates the power to which x is raised.
  - e.g., $f(x) = \sqrt{x}$. This is the *square root of* x. Can be expressed as : $f(x) = \sqrt{x} = x^{\frac{1}{2}}$.
- constants,
  - e.g., $f(x) = x + 5$. In this example 5 is a constant.
- and coefficients.
  - e.g., $f(x) = 10x$. In this example 10 is the coefficient.

Kinds of functions

Only variables, constants, and coefficients (no exponents) → linear function.

- Form is: $y = f(x) = mx + b$, where m is a coefficient and b is a constant.

Form: $y = f(x) = ax^2 + bx + c$, is a quadratic function.

Power function: $y = f(x) = ax^b$.

Exponential function: $y = f(x) = ab^x$.

- b is the base for the exponent x.

Logarithmic function: $y = f(x) = \log_b x$.

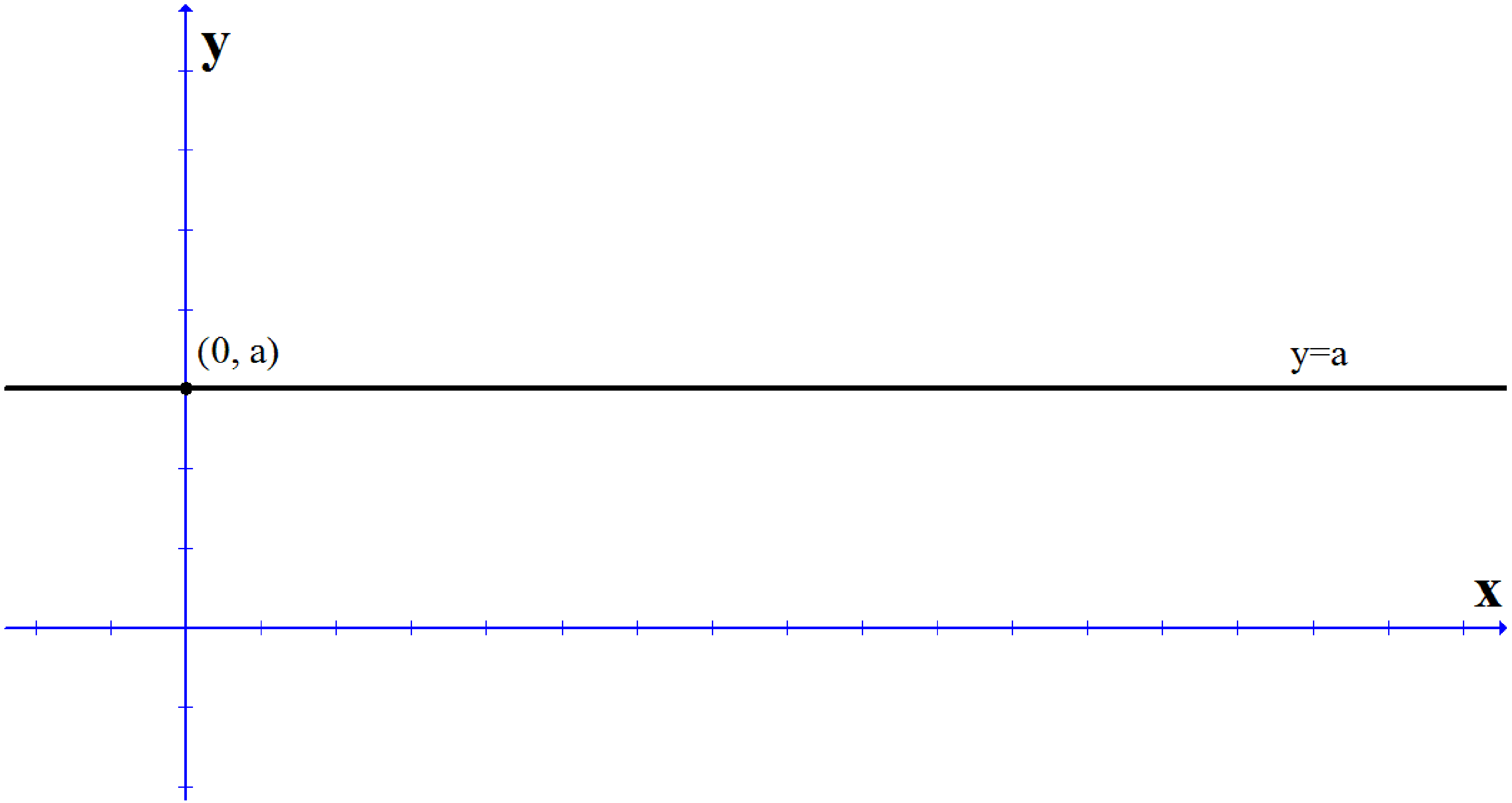
- Again b is the base.
- y is the power to which you must raise x to get b.
- The base in many applications is “Euler’s number”: $e \approx 2.718$.

Graphing linear functions

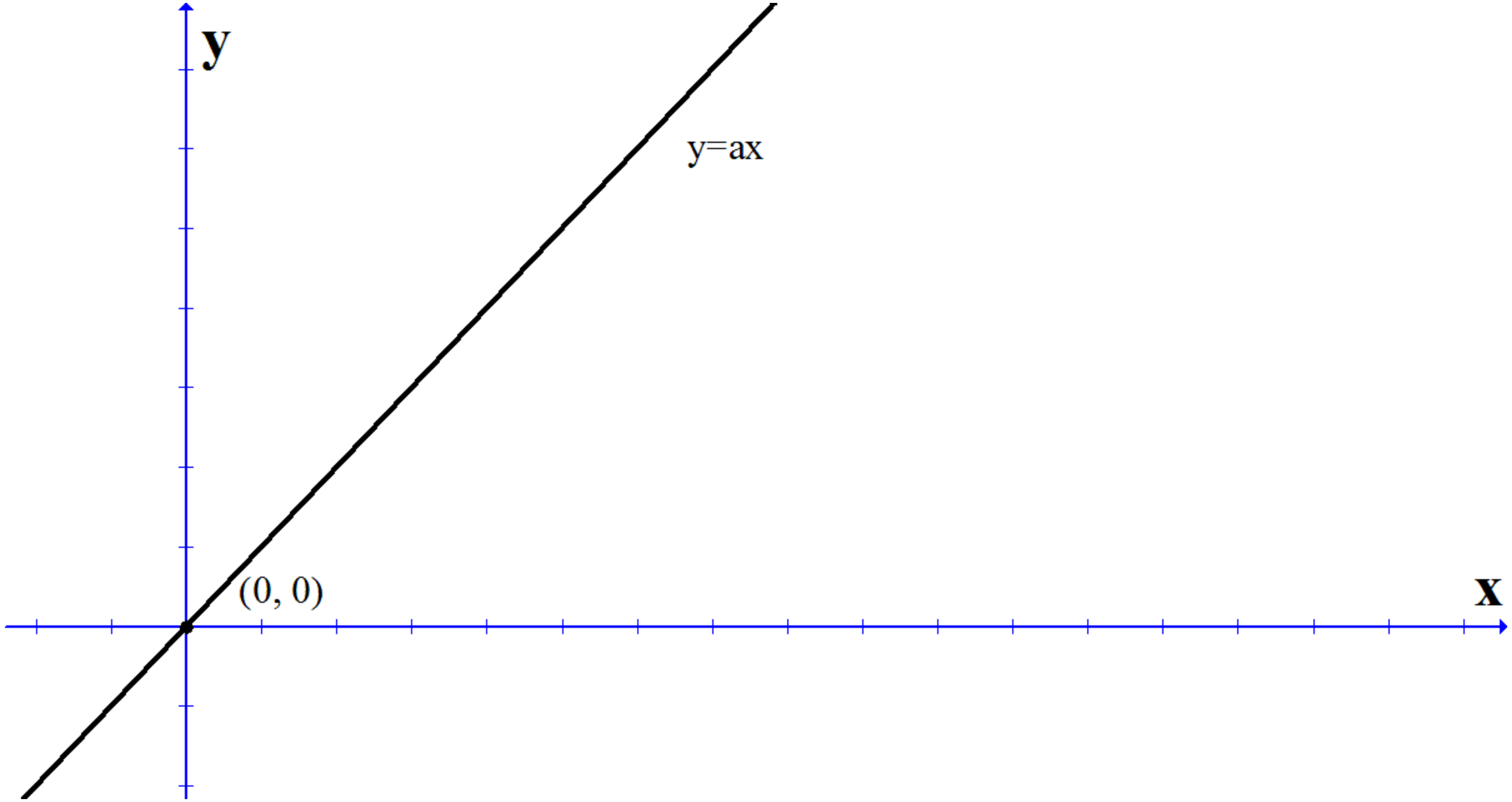
2 known points or the slope-intercept form.

- Since these functions are always a straight line, you just need to know two points that the line goes through; you can then connect them with a line.
- If you know that the coefficient is the slope and the constant is the y intercept, you can graph the line using that information.

A linear function without slope

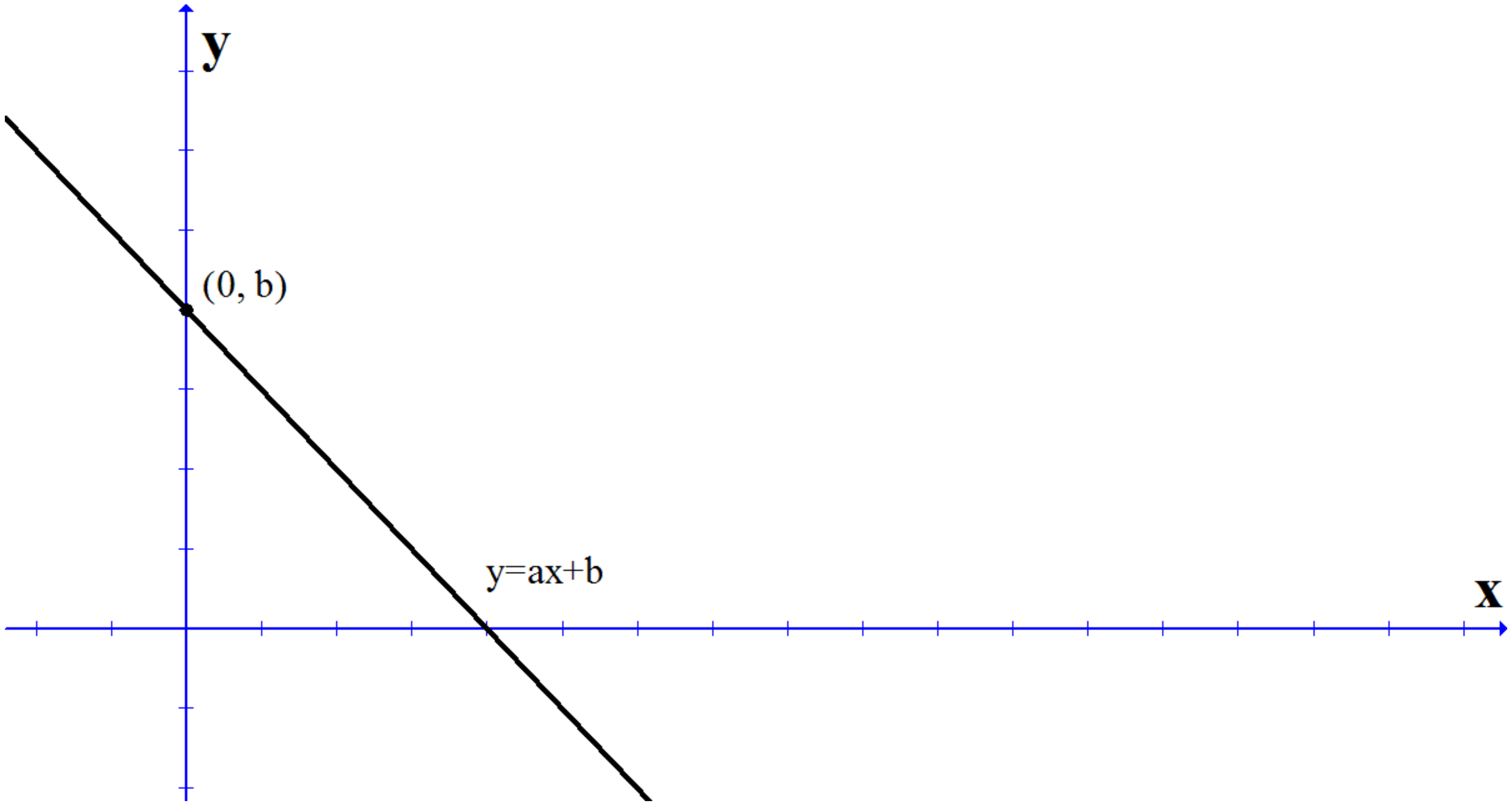


A linear function with slope only



Note that the slope here is positive (a>0); slope can be negative, too.

A linear function with slope and constant



This is a case where the slope of the line is negative (a<0).

Slope of a linear function

The slope of a linear function is the same along its entire length.

- The slope is constant.

The function, $f(x) = 4x$, has a *slope of 4*—which is a constant.

The simplicity of constant slope is what makes linear functions so useful for our purposes in economics classes.

As you will see, the slope of a function has very powerful intuitive implications for econometric analysis.

- More on the other functional forms as necessary, later in the course.

Differential calculus

Non linear functions have <u>variable</u> slope.

To calculate slope, <u>differentiate</u> the function instead of differencing (“rise over run”).

- Result is called the <u>first derivative</u>.

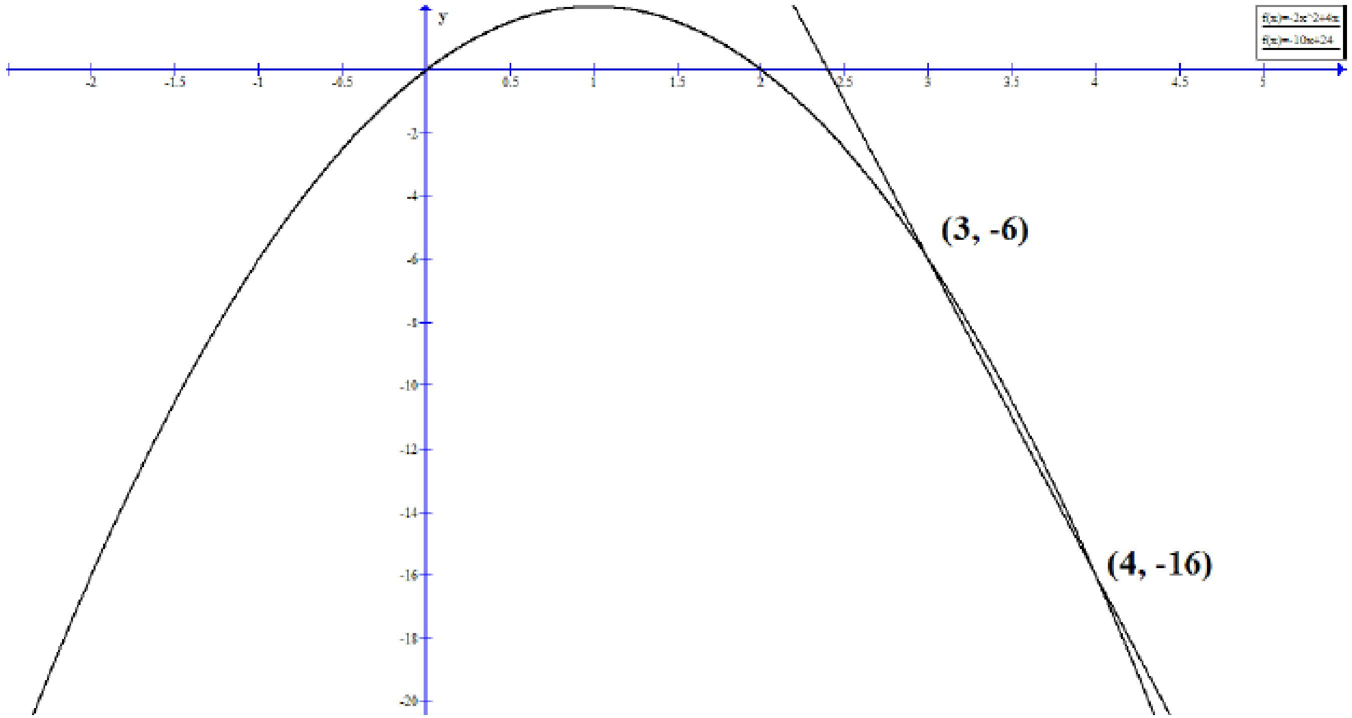
Example: $y = f(x) = -2x^2 + 4x$.

- Rise over run at $x = 3$ is:

$$\frac{f(4) - f(3)}{4 - 3} = -16 - (-6) = -10.$$

- Gives you the slope of the <u>secant</u> line connecting the 2 points.

Differential calculus (continued)



Differential calculus (continued)

Differencing in the limit as $\Delta x \rightarrow 0$.

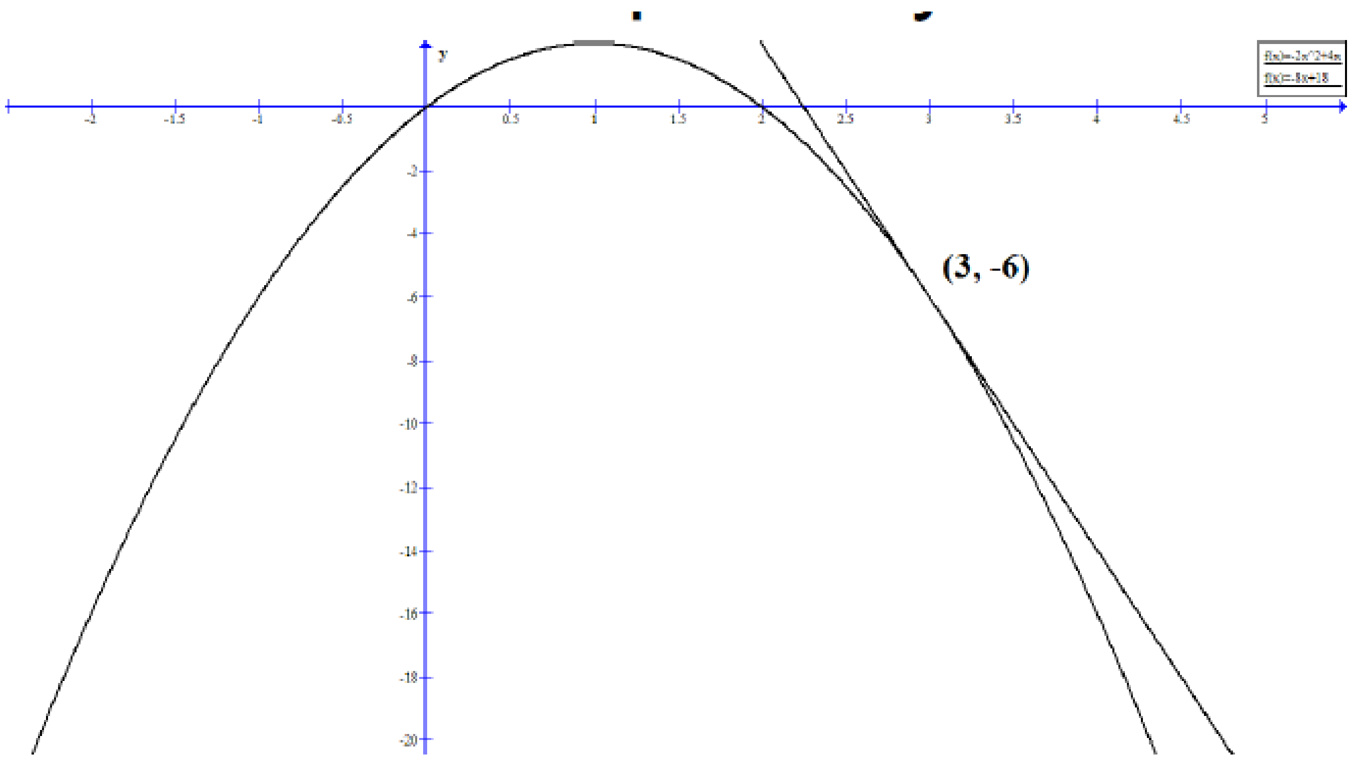
- Moving the 2 points infinitely close together.
- Replace Δx with notation ε.
- Again beginning at 3:

$$slope = \frac{\Delta f(x)}{\varepsilon} = \frac{-2(9 + 6\varepsilon + \varepsilon^2) + 4(3 + \varepsilon) - (-6)}{\varepsilon} = -2\varepsilon - 8$$

Taking the limit as change goes to zero simply gives you,

$$\frac{\delta y}{\delta x} = -8.$$

First derivative: the slope of the line tangent to the function



The tangent line just barely touches the function at the chosen point. At that specific point the line has the same slope as the parabola (negative 8).

Differentiation in practice

"Cookbook" procedure for differentiation.

- Multiply the coefficient on each term by the exponent on that term.
- Reduce the exponent on each term by one.
- For this example,

$$f(x) = -2x^2 + 4x,$$

the first derivative is,

$$\frac{\delta f(x)}{\delta x} = -2(2)x^{2-1} + 4(1)x^{1-1} = -4x + 4.$$

Partial differentiation

When functions have more than 1 argument, they have more than 1 derivative.

- Just have to specify the variable with which you are differentiating, e.g.,

$$y = f(x, z).$$

Notation:

- $\frac{\partial y}{\partial x}$, or
- f_x, or
- y_k.

Economics examples:

- Marginal utility,
- Marginal product of labor.

Partial differentiation (continued)

Differentiate with respect to 1 independent variable at a time.

- Treat other arguments as constants.
- Extra easy when there are no interaction terms (in which 2 variables multiply together).

Example: $Q(L, K) = AK^{a}L^{1-a}$

$$\frac{\partial Q}{\partial K} = aAK^{a-1}L^{1-a}$$

and

$$\frac{\partial Q}{\partial L} = (1-a)AK^{a}L^{-a}.$$

Tabular descriptive statistics: cross tabulation

A tabular summary of data for two variables.

Cross tabulation can be used when:

- one variable is nominal/ordinal and the other is ordinal/interval/ratio,
- both variables are ordinal/interval/ratio, or
- both variables are nominal/ordinal.

Cross tabulation (continued)

Cross Tabulation of Nights and Kids	Number of Kids										
Nights Business Travel (last 3 mo.)	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**Grand Total**
0-9	888	469	690	381	152	61	34	20	8	3	2706
10-19	34	9	26	9	4	2	1	1			86
20-29	11	4	7	2	3		1				28
30-39	8	4	6	4	5		1	1			29
40-49	4	1	3	2							10
50-59			2	1	1						4
60-69	3	2	2	1	1		1				10
70-79		2	3								5
80-90	1		1								2
Grand Total	**949**	**491**	**740**	**400**	**166**	**63**	**38**	**22**	**8**	**3**	**2880**

Cross tabulation (continued)

Recall what a frequency distribution for one variable does.

Cross tabulation as multiple frequency distributions:

- one for each value of a second variable.

More information by expressing frequencies relative to either the column or row totals.

- Analogous to the relative (or percent) frequency distribution.

Cross tabulation, column percentages

Cross Tabulation of Nights and Kids	Number of Kids									
Nights Business Travel (last 3 mo.)	0	1	2	3	4	5	6	7	8	9
0-9	93.57%	95.52%	93.24%	95.25%	91.57%	96.83%	89.47%	90.91%	100.00%	100.00%
10-19	3.58%	1.83%	3.51%	2.25%	2.41%	3.17%	2.63%	4.55%	0.00%	0.00%
20-29	1.16%	0.81%	0.95%	0.50%	1.81%	0.00%	2.63%	0.00%	0.00%	0.00%
30-39	0.84%	0.81%	0.81%	1.00%	3.01%	0.00%	2.63%	4.55%	0.00%	0.00%
40-49	0.42%	0.20%	0.41%	0.50%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
50-59	0.00%	0.00%	0.27%	0.25%	0.60%	0.00%	0.00%	0.00%	0.00%	0.00%
60-69	0.32%	0.41%	0.27%	0.25%	0.60%	0.00%	2.63%	0.00%	0.00%	0.00%
70-79	0.00%	0.41%	0.41%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
80-90	0.11%	0.00%	0.14%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%
Grand Total	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**	**100.00%**

Cross tabulation, row percentages

Cross Tabulation of Nights and Kids	Number of Kids										
Nights Business Travel (last 3 mo.)	0	1	2	3	4	5	6	7	8	9	Grand Total
0-9	32.82%	17.33%	25.50%	14.08%	5.62%	2.25%	1.26%	0.74%	0.30%	0.11%	100.00%
10-19	39.53%	10.47%	30.23%	10.47%	4.65%	2.33%	1.16%	1.16%	0.00%	0.00%	100.00%
20-29	39.29%	14.29%	25.00%	7.14%	10.71%	0.00%	3.57%	0.00%	0.00%	0.00%	100.00%
30-39	27.59%	13.79%	20.69%	13.79%	17.24%	0.00%	3.45%	3.45%	0.00%	0.00%	100.00%
40-49	40.00%	10.00%	30.00%	20.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
50-59	0.00%	0.00%	50.00%	25.00%	25.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
60-69	30.00%	20.00%	20.00%	10.00%	10.00%	0.00%	10.00%	0.00%	0.00%	0.00%	100.00%
70-79	0.00%	40.00%	60.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%
80-90	50.00%	0.00%	50.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100.00%

Cross tabulation (concluded)

Column percentages:

- each column sums to 100%.
- Note that the rows do not sum to 100%.
- Vice versa for row percentages.

The two tables have slightly different interpretations.

- In the first one, we're comparing the distributions of workers' travel behavior based on how many kids they have.
- In the second, we're comparing the distributions of their parenthood responsibilities based on how often they travel.

Graphical descriptive statistics: scatterplot

A graphical presentation of the relationship between two quantitative variables.

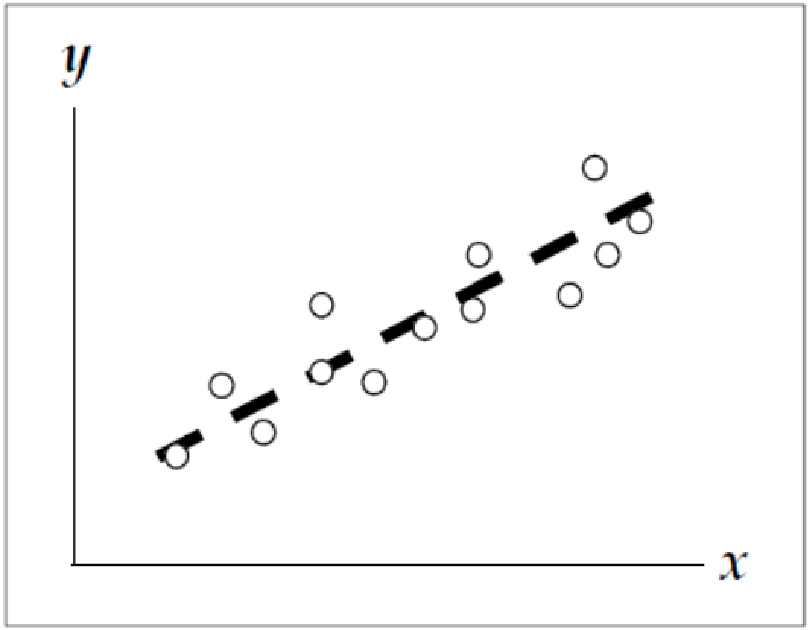
One variable is shown on the horizontal axis and the other variable is shown on the vertical axis.

The general pattern of the plotted points suggests the overall relationship between the variables.

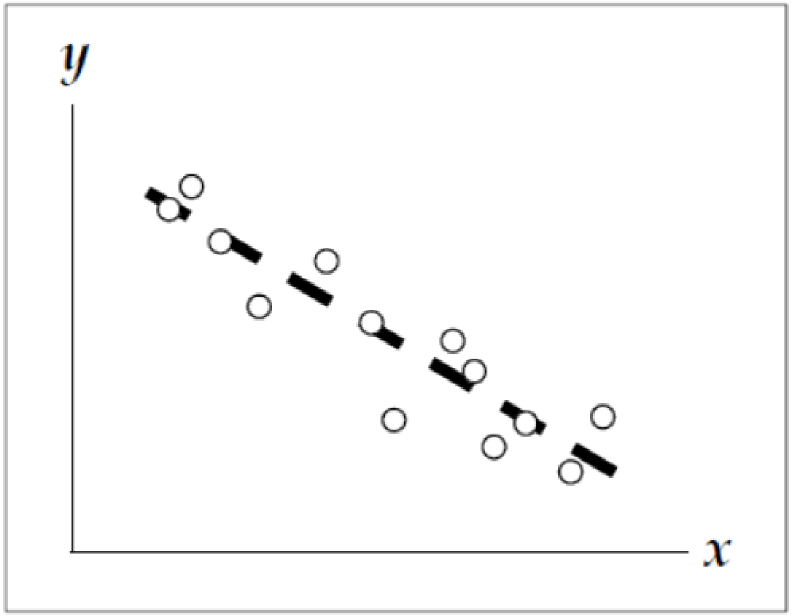
A trend line can be added as an approximation of the relationship.

Scatterplot (continued)

A Positive Relationship

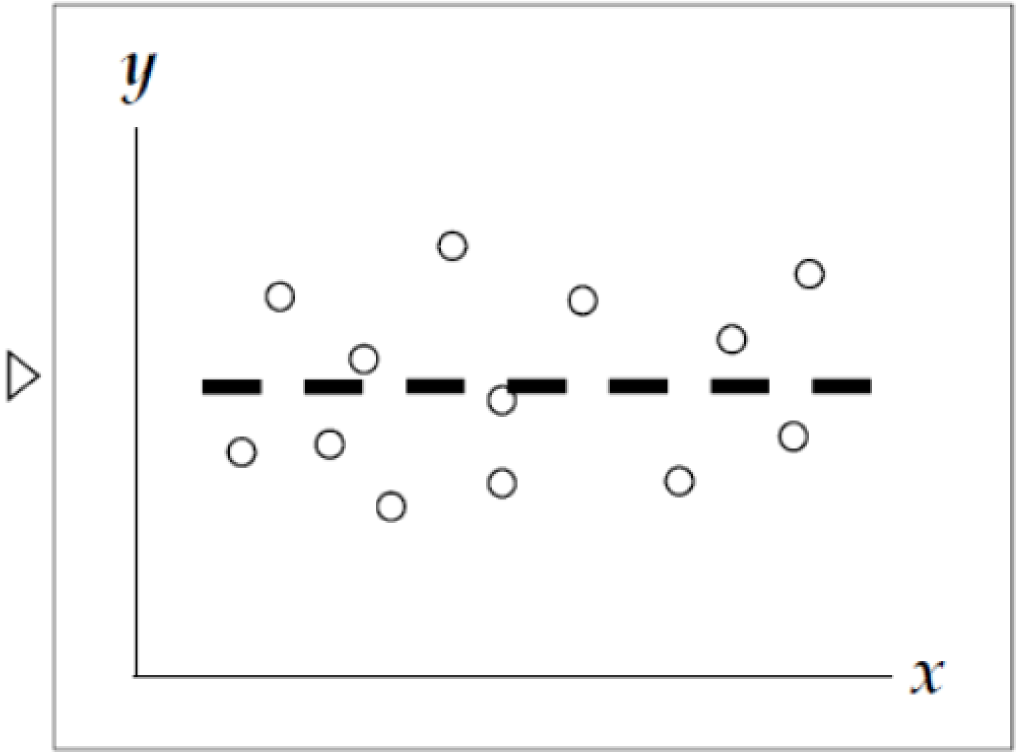


A Negative Relationship



Scatterplot (continued)

No Apparent Relationship



Numerical multivariate descriptive statistics

Measures of association between two variables

- quantify visual evidence of a relationship.

They are covariance (s_{xy}) and the correlation coefficient (r_{xy}).

- Note these are sample statistics that estimate the true underlying relationship in the population:

$$E(s_{xy}) = \sigma_{xy} \text{ and } E(r_{xy}) = \rho_{xy} \text{ (population covariance and correlation).}$$

Measures of association (continued)

Positive values indicate a positive relationship, and negative values indicate a negative relationship.

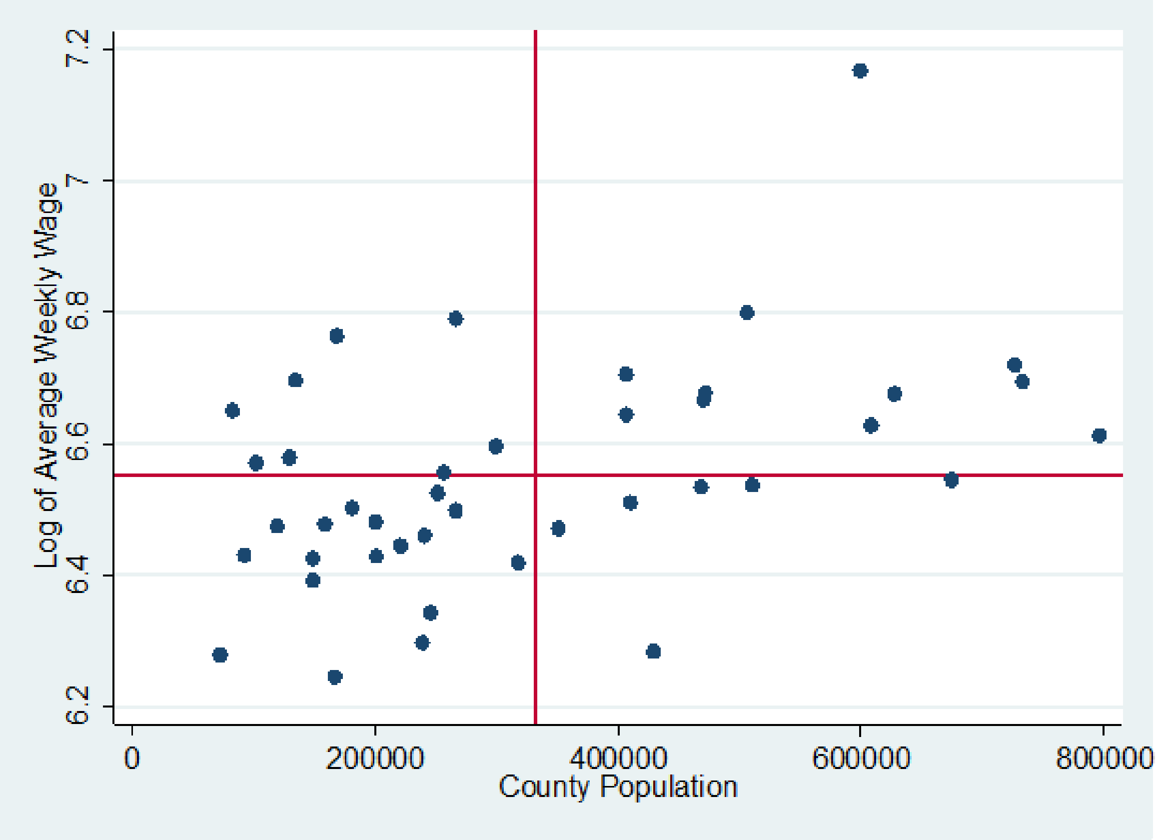
$$s_{xy} \equiv \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) \text{ and,}$$

$$r_{xy} \equiv \frac{s_{xy}}{s_x s_y}\text{ ; } s_x \text{ and } s_y \text{ are sample standard deviations of } x \text{ and } y.$$

To demonstrate how these definitions imply positive and negative signs, consider a scatterplot and the point, $(\bar{x}, \bar{y})$.

- Variables with a “bar” over them usually indicate the <u>sample mean</u>.

Measures of association (continued)



The sample means are the intersection of the red lines in the graph.

Measures of association (continued)

The point $(\bar{x}, \bar{y})$ divides the plot into four regions:

- The "north-east": in which $x_i > \bar{x}$ and $y_i > \bar{y}$.
- The "south-east": in which $x_i > \bar{x}$ and $y_i < \bar{y}$.
- The "south-west": in which $x_i < \bar{x}$ and $y_i < \bar{y}$.
- The "north-west": in which $x_i < \bar{x}$ and $y_i > \bar{y}$.

Measures of association (continued)

Consider the product, $(x_i - \bar{x})(y_i - \bar{y})$ for an observation in each of the four regions.

- For number 1, it is a positive times a positive.
- For number 3, it is a negative times a negative.
- For the other two, it's a negative times a positive.

For the odd-numbered regions, the product will be positive, and for the even-numbered regions, it will be negative.

Covariance

Recall the appearance of a positive relationship:

- it should have a lot of observations in the “north-east” and “south-west” parts of the scatterplot.
- Similarly a negative relationship should have a lot of points in the “north-west” and “south-east”.

If the data set has a lot of observations in the odd-numbered regions on the scatterplot, they will count positively toward the covariance.

And observations in the even-numbered regions count negatively toward covariance.

If there are more of the former, covariance will be positive, and if there are more of the latter, it will be negative.

- For the variables on the graph above, the covariance is roughly 15,234, a positive number—a positive relationship between population and average wages across counties.

Correlation coefficient

Estimates the strength of the relationship

Divide covariance by the product of standard deviations.

- Standard deviation is always positive → preserves the sign of covariance.
- Standardizes the scale: measures with larger scales can be compared purely on the basis of how strong the relationship is.

An “r” of 1 represents a perfect positive linear relationship, and an “r” of -1 is a perfect negative linear relationship.

Zero is the dividing point between positive and negative association and represents zero covariance and zero relationship between the two variables.

- The correlation coefficient for the population and wage (above) is 0.4583—a modestly strong positive relationship.

Random variables

Numerical descriptions for the outcome of a probabilistic experiment.

If *x* is a random variable (r.v.), it takes on different values with a specific probability distribution.

- Discrete: a finite number of values or an infinite series of values.
- Continuous: can assume any (infinitely many) value on one or more interval(s).

Probabilities summarized by a probability function:

- $\Pr(x = x_i) = f(x_i) \in [0,1]$
- $\sum_{i \in I} f(x_i) = 1$

“That there, Clark, is an r.v.” – Cousin Eddie . . . different kind of r.v. though

Describing a random variable

Mean or expected value or first moment of the distribution.

- A probability-weighted average of all the values the r.v. takes.

$$E(x) \equiv \sum_{i \in I} x_i f(x_i)$$

Variance or second moment.

- A probability weighted average of the squared deviations from the mean.

$$\sigma^2 = Var(x) \equiv \sum_{i \in I} \left(x_i - E(x)\right)^2 f(x_i)$$

Expected value

Properties.

- Expected value of a constant, c, is $E(c) = c$.
- Constants "pull through" expectations: $E(ax + b) = aE(x) + b$.
- Expected value is commutative: $E(\sum_{i=1}^{n} a_i x_i) = \sum_{i=1}^{n} a_i * E(x_i)$.

Variance

Properties.

- The properties of expected value also apply because variance can be phrased as the expectation of:

$$Var(x) = E\big(x - E(x)\big)^2 = E(x^2) - [E(x)]^2.$$

- Additionally, iff $\Pr(x = c) = 1$, the variance of x is zero.
- Constants "pull through", a la,

$$Var(ax + b) = E[ax + b - E(ax + b)]^2 = E\big[a\big(x - E(x)\big)\big]^2, \text{so}$$
$$Var(ax + b) = E\left[a^2\big(x - E(x)\big)^2\right] = a^2 * Var(x).$$

Variance (continued)

More properties.

- Generalizing to the variance of a sum of random variables.

$$Var(ax + by) = E\big[a\big(x - E(x)\big) + b\big(y - E(y)\big)\big]^2$$

$$\Leftrightarrow Var(ax + by) = E\left[a^2\big(x - E(x)\big)^2 + b^2\big(y - E(y)\big)^2 + 2ab\big(x - E(x)\big)\big(y - E(y)\big)\right]$$

$$\Leftrightarrow Var(ax + by) = a^2Var(x) + b^2Var(y) + 2ab * Cov(x, y)$$

- For pairwise uncorrelated variables (none is correlated with any other r.v.).

$$Var\left(\sum_{i \in I} a_i x_i\right) = \sum_{i \in I} a_i^2 Var(x_i)\,,$$

because all the covariance terms are zero.

Conditional expectation

Same notion as expected value, but re-weighted with information generated by knowing the value/range of a r.v., e.g.,

- $E(y|x = 6)$,
- $E(y|x \geq 6)$,
- $E(y|y \geq 6)$.

Properties.

- Functions of r.v. behave like constants when you condition on the r.v.

$$E(c(x)|x) = c(x), \text{for any function}, c$$

- Unconditional and conditional expectations are related by:

$$E\left[\left(a(x)y + b(x)\right)|x\right] = a(x) * E(y|x) + b(x)$$

- And

$$E(y|x) = E(y) \text{ iff } x \text{ and } y \text{ are independent.}$$

Conditional expectation (continued)

Law of iterated expectations.

$$E[E(y|x)] = E(y)$$

- Take the expectation of the value in brackets across all different values of *x* and you arrive at the unconditional expectation.

If x and y are independent, their correlation and covariance are zero.

Conclusion

Other things you learn in earlier classes (not specifically covered here) are used in ECON 360, too.

- Distributions and their properties,
- Performing statistical inference,
- Optimization using differential calculus.

These topics are central to performing regression analysis and will be covered in more detail in subsequent lectures.

The topics in this set of notes is so fundamental that I expect students to be familiar with them prior to beginning to turning their attention to regression.

- Hopefully after this review, that is the case.