Ordinary Least Squares Regression

May 27, 2010

Nancy Burns (nburns@isr.umich.edu) - University of Michigan

Ordinary Least Squares Regression

Nancy Burns

University of Michigan

Prepared for Presentation at SESRI, Qatar University, Doha, May 2010

Ordinary Least Squares (OLS) Regression

- *Dependent variable*, Y, what we're explaining.
- *Explanatory variable* or *independent variable*, X, what we are using to explain Y.

- When X goes up by a certain amount, on average, what happens to Y? Does it go up, go down, or not change, and by how much? And how certain are we about this effect?

Scatterplots: When X goes up, what happens to Y?

- Source: Shively, 2005

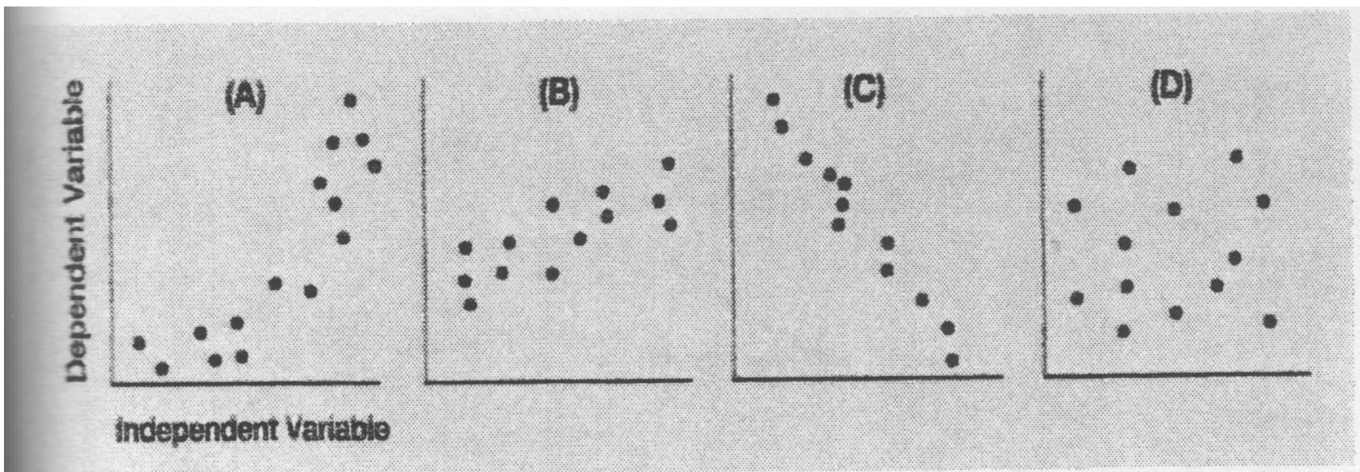


Figure 7–2 Assorted Scattergrams

Scatterplots

- Source: Berry and Sanders

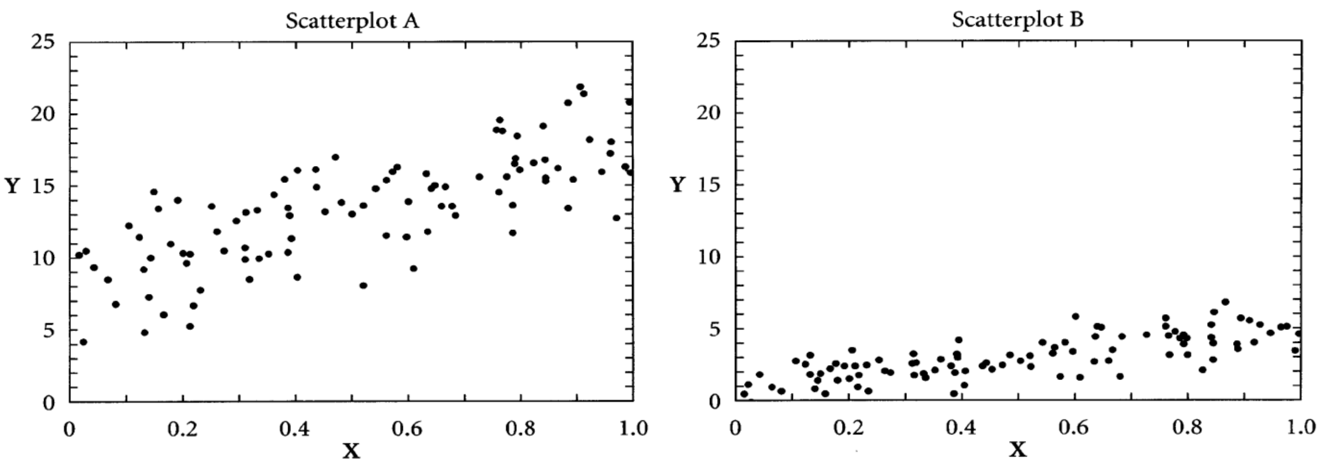


FIGURE 1.3 Two scatterplots with correlation coefficients of +0.75

The predicted value of $Y = a + bX$

What does this equation draw?

The Regression Line

The predicted value of Y =

intercept + slope * X

Y is the dependent variable

X is the explanatory variable

The Regression Line

Source: Wonnacott and Wonnacott, 1990.

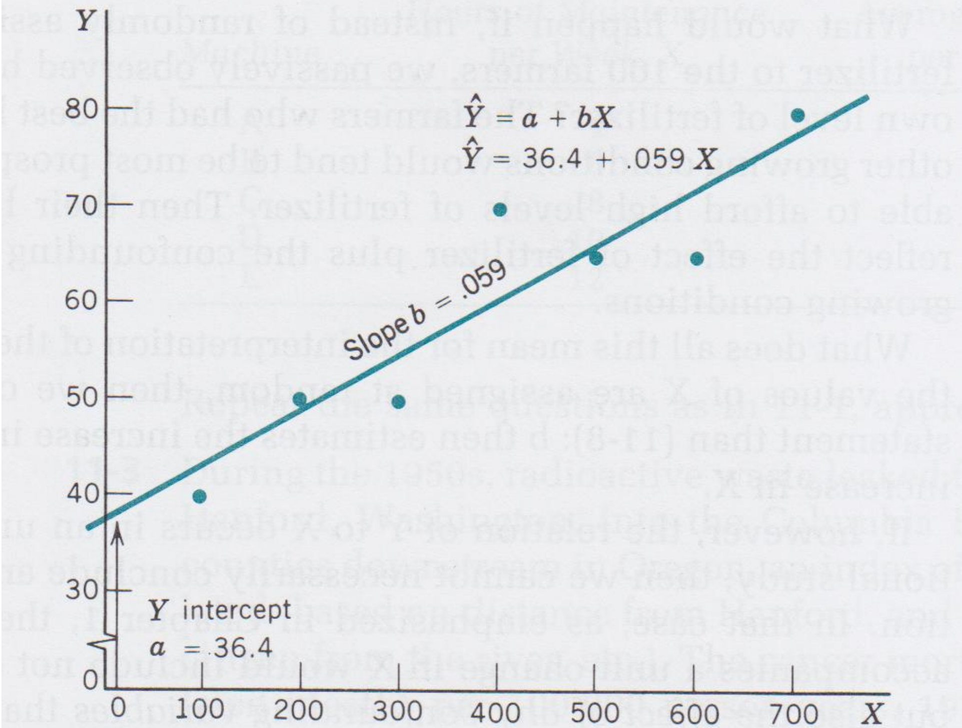


FIGURE 11-4

The least squares line fitted to the data of Example 11-1.

The Regression Line

- Source: Shively, 2005.

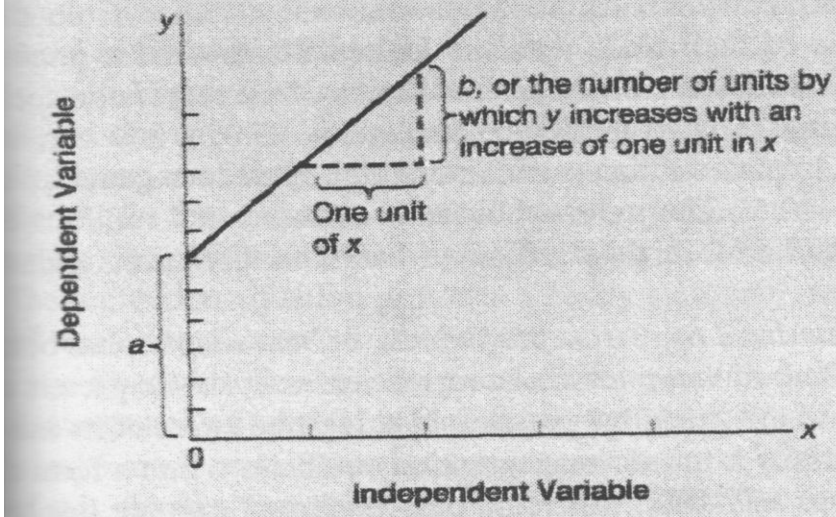
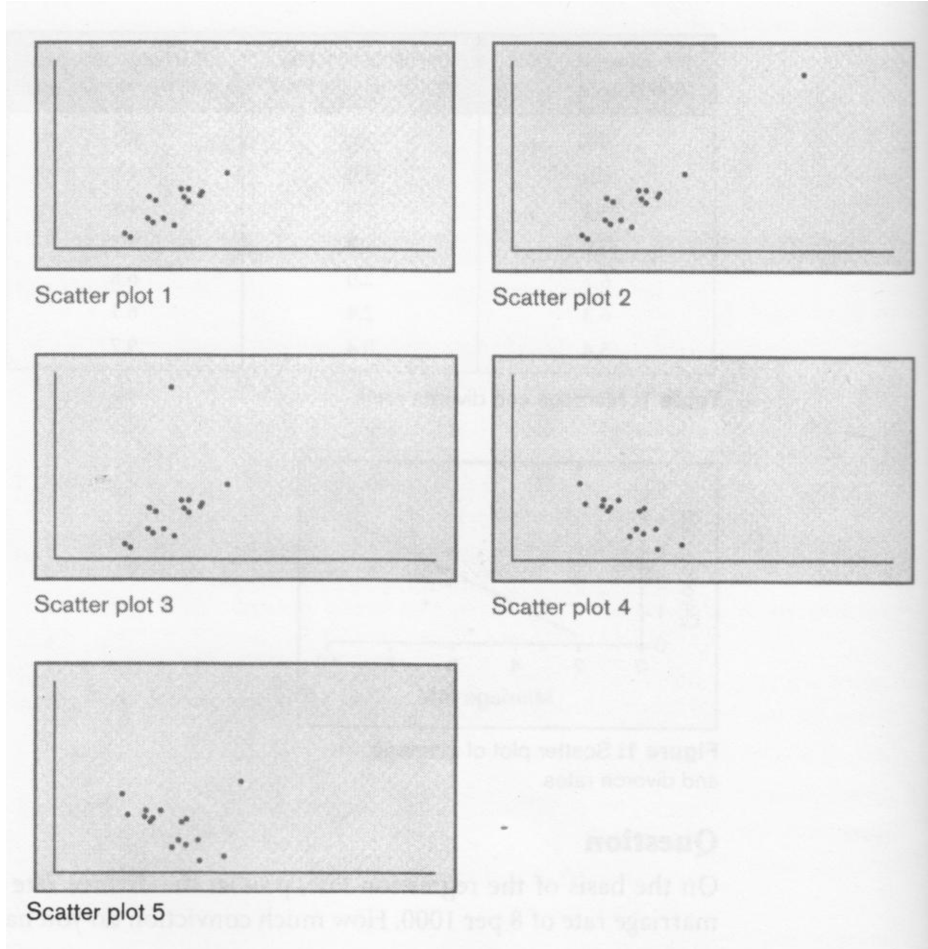


Figure 7–4 The Regression Equation

The equation of this line is $y = 6 + 3x$. The predicted value of y when x is 4, for instance, is $6 + (4 \times 3)$, or 18.

Scatterplot Exercise

- Source: Schaeffer…



Some questions

- When x goes up by one unit, for which of these slides does y go down?
- If you were drawing a line to describe the points for the graphs where y goes down when x goes up, which would have the steeper slope? For which one would y go down more as x goes up?
- Three of these have exactly the same coefficient on x. Which three?
- The three have different correlations between x and y; which is higher, and which is lower?

Scatterplot Exercise, for later

- Source: Scheaffer, …

Activity

1. Match each of the five scatter plots to the description of its regression line and correlation coefficient. The scales on the axes of the scatter plots are the same.

a. $r = 0.83, y = -2.1 + 1.4x$

b. $r = -0.31, y = 7.8 - 0.5x$

c. $r = 0.96, y = -2.1 + 1.4x$

d. $r = -0.83, y = 11.8 - 1.4x$

e. $r = 0.41, y = -1.4 + 1.4x$

Example

- Working in groups, graph the following results.

  Earnings = -84078 + 1563*Height.

Graph the regression line. Including all of the elements of the graph in the last PowerPoint slide.

Source: Gelman and Nolan, “Teaching Statistics”.

Hints to get started

When will y = 0? When x is about 54 inches.

Why? 0 = -84078 + 1563*x

0+84078 = -84078 + 84078 + 1563*x

84078=0 + 1563*x.

84078= 1563*x.

84078/1563=(1563/1563)*x.

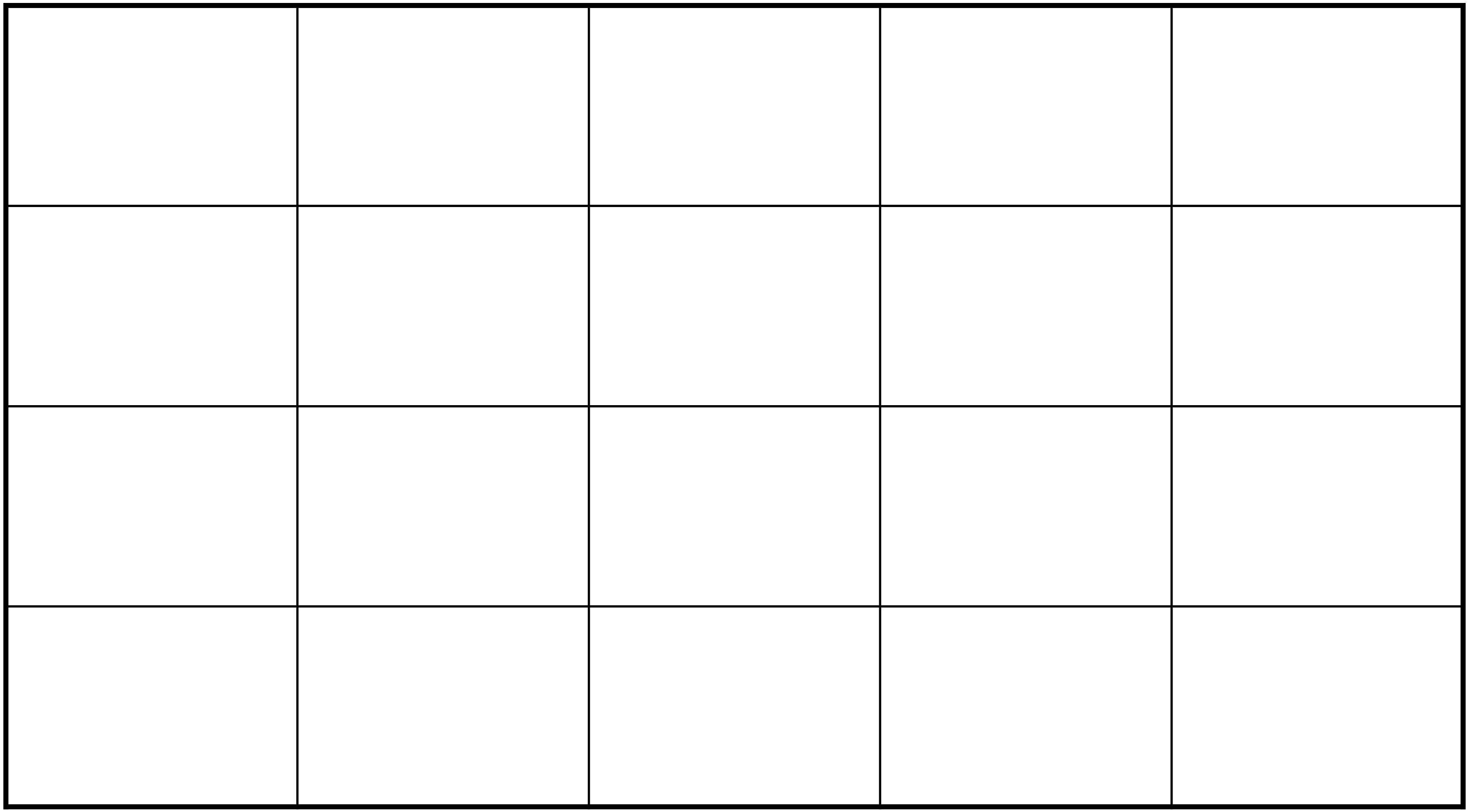
53.79 = x

Y is in US dollars. X is in inches.

One more hint

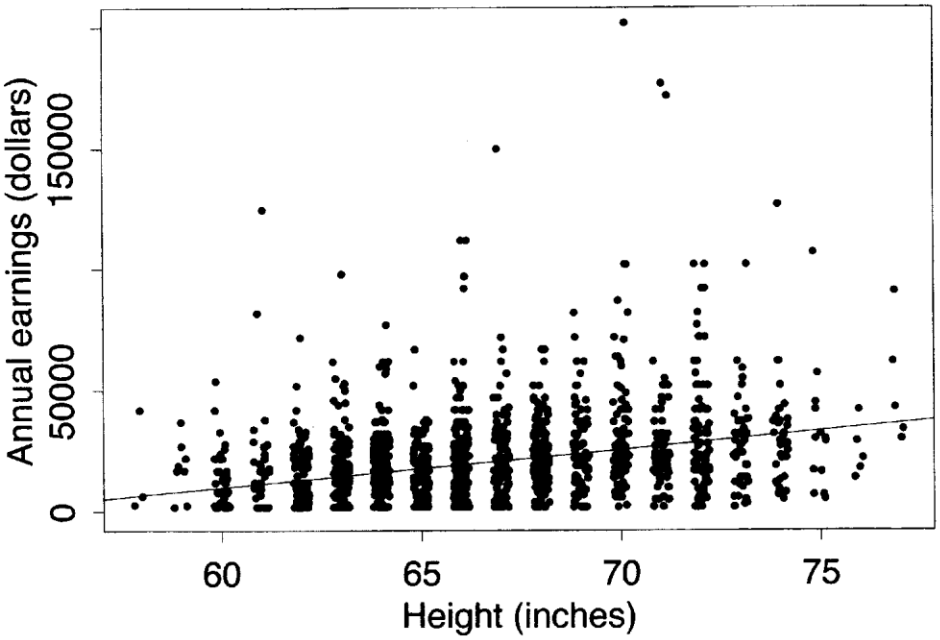
What does one extra inch in height yield in terms of dollars of income?

Small piece of graph paper



Scatterplot of height and earnings

- Source: Gelman and Nolan



How do we calculate a and b, the intercept (or constant) and the slope?

Minimize the sum of squared residuals.

Would use calculus and calculate partial derivatives with respect to a and b.

"But of all these principles, least squares is the most simple: by the others we would be led into the most complicated calculations."

Gauss, 1809

Residual

A residual is the difference between our observation, y, and the predicted value of y from our model.

We want the difference to be small.

Minimizing the Sum of Squared Residuals

- Source: Shively, 2005.

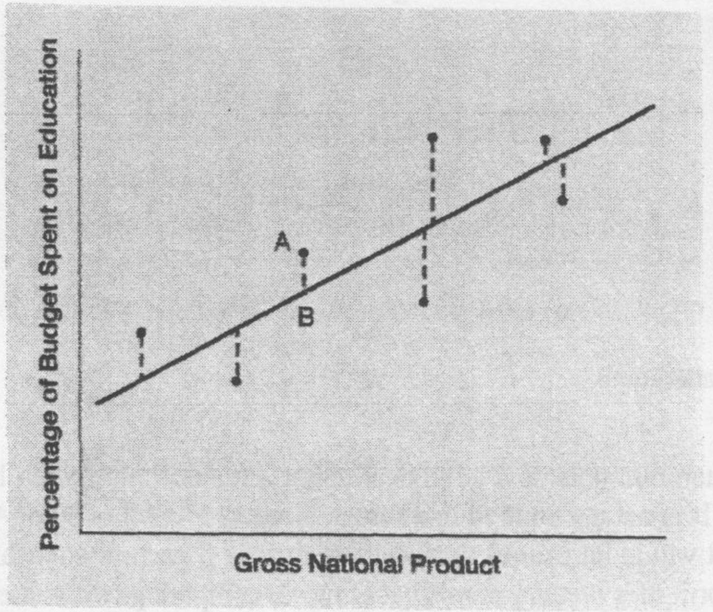


Figure 7–3 The Regression Line

Minimizing the vertical distance

- Source: Berry and Sanders

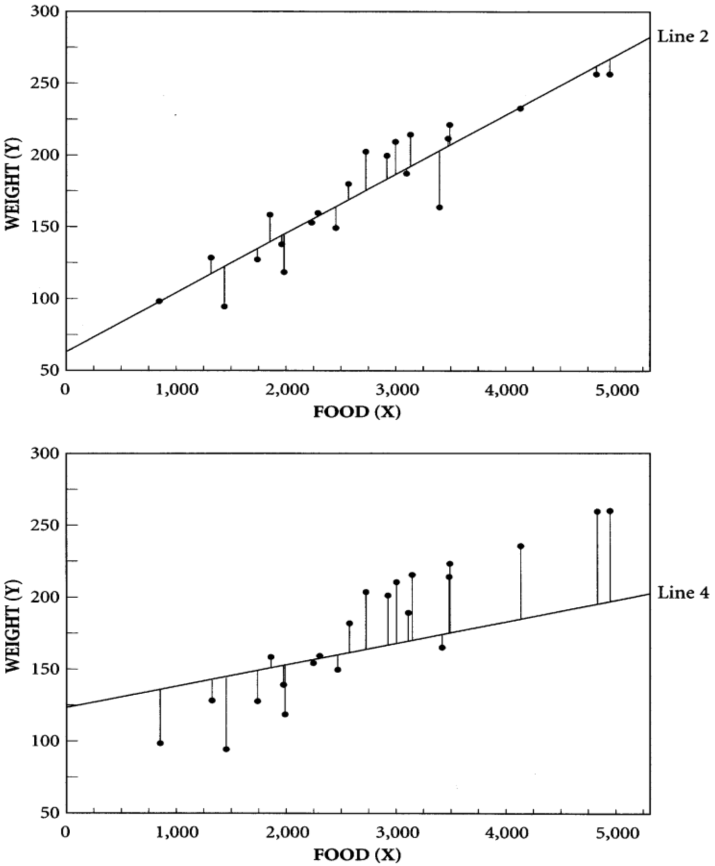


FIGURE 2.3 Vertical distances between points and two lines

After a lot of math, in the bivariate case, we would find that

a=the mean of Y – b*the mean of X

And

$$b = \frac{\text{the covariance of X and Y}}{\text{the variance of X}}$$

Variance and Covariance

- Covariance describes how X and Y vary together.
- Variance describes how much a variable varies around its mean.

  The variance of X, then, describes how much spread there is away from the mean of X. We divide the squared distances by n or n-1 to get the average distance of the data from the mean.

We could do the math and calculate the coefficients, but we wouldn’t yet have the tools to draw inferences to data we don’t have.

Without one more tool, all we have is a way to describe our data. **Without one more tool, we do not have a way to say how certain we are about that description.**

Inference

Our challenge is that we are not describing a full population.

Instead, we are drawing an inference from a sample to describe a population.

We need assumptions and tools from probability to allow us to draw these inferences.

Inference from Samples

The tools from probability and the assumptions we will make allow us to say how certain we are about the estimates we calculate with our sample.

Assumptions That Allow Us to Get to Inference, that Allow Us to Say How Certain We Are

Most important assumptions

- All of the observations come from distributions with the same variance.
- Knowing something about one Y does not give us information about another Y.

Observations coming from distributions with the same variance?

What if some observations are highly predictable and others are less predictable?

The classic case involves asking how expenditures on meals depend on income. We would have smaller residuals for low income people than for high-income people.

There are solutions to this problem – a problem called **"heteroskedasticity"**, the most straightforward of which is to use White's standard errors or heteroskedastic-consistent standard errors. I include citations that cover this method.

Knowing something about one observation of y tells us something about another observation of Y

Examples of this:

Time
Space – countries, neighborhoods
Families
Schools

There are solutions for this problem -- a problem called **“autocorrelation**”-- and I’ll include references that cover these solutions.

Can you think of other examples from your own work of…

- Times when the observations do not come from distributions with the same variance?

- Times when knowing something about one Y gives us information about another Y?

The key to inference is the

sampling distribution.

The Sampling Distribution

- Source: Wonnacott and Wonnacott, 1990.

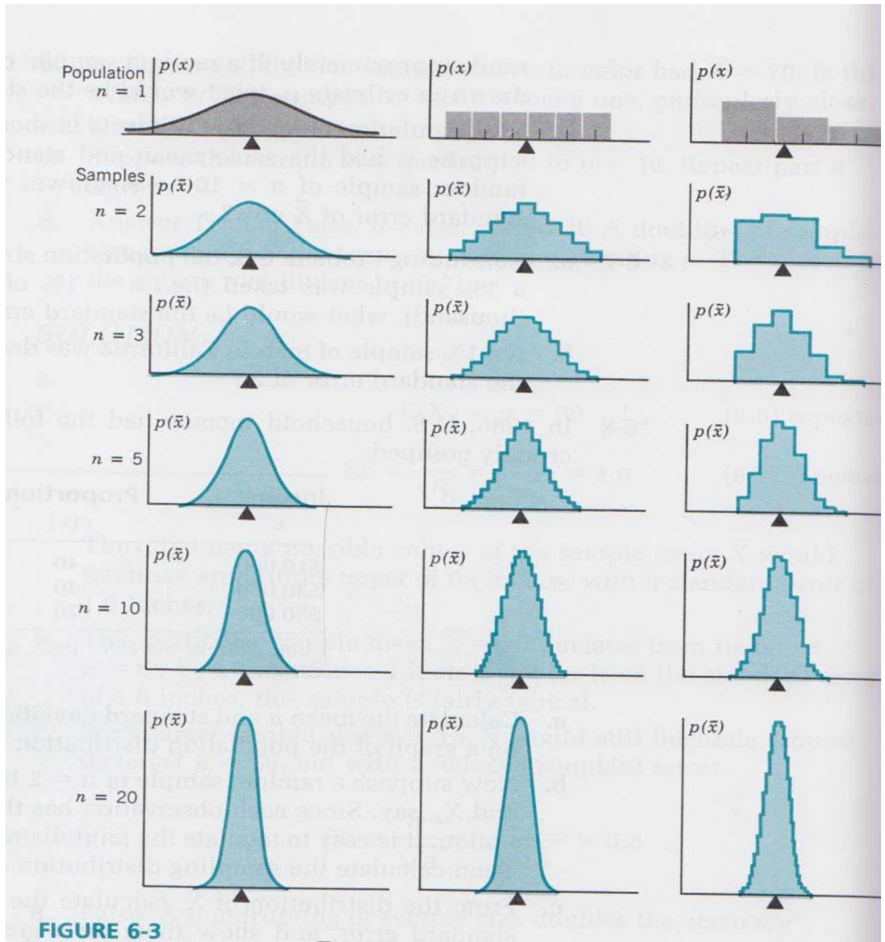


FIGURE 6-3

Sampling Distribution of the Mean

- Source: Wonnacott and Wonnacott, 1990.

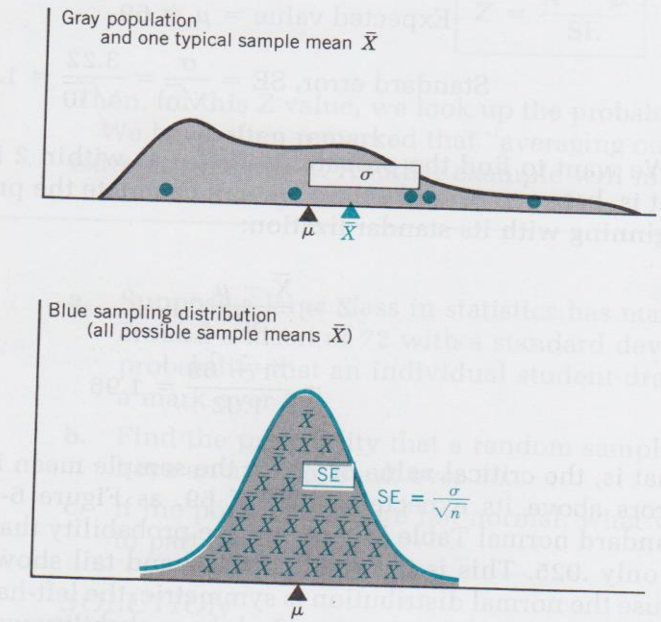


FIGURE 6-4

The sampling distribution of $\bar{X}$ is approximately normal, and more concentrated around μ than is the population. It is colored blue to distinguish it from the population in gray.

Sampling distribution

- The slope and intercept are draws from a distribution. That distribution comes from estimates calculated on repeated samples.

- Central Limit Theorem, or the Normal Approximation Rule, helps us describe those distributions.

Central Limit Theorem

If we take random samples of size *n* from a population with a given mean and a given standard deviation, then, as *n* gets large, the mean of X approaches the Normal distribution, with the same mean as the population and with a standard deviation of the population standard deviation over the square root of *n*.

Sampling Distributions and Sample Sizes

- Hanushek and Jackson.

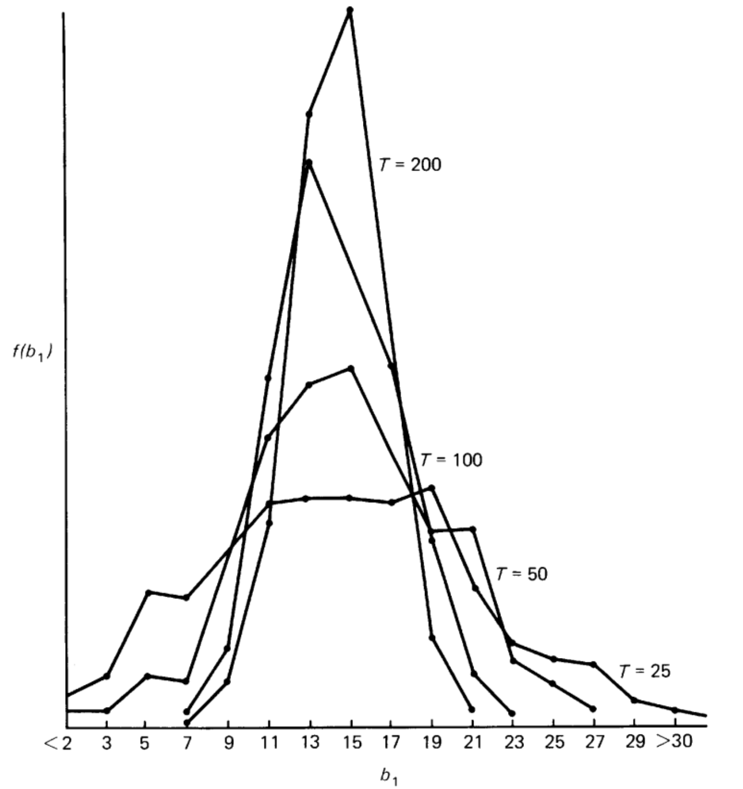


FIGURE 3.1 Distributions of simulated b_1. Horizontal axis indicates midpoint of range of values for estimated coefficient.

We'll estimate the standard errors of our regression coefficients.

These standard errors are our measures of the variability of b and a. They are a function of the variability of y and x and of the sample size.

For example, when there's little variance in x, we have little certainty about b, and our estimates of the variability of b will be quite large. When our samples are small, our estimate of the variability of our coefficients will be larger.

The Value of Small Standard Errors

These standard errors describe our estimate of the sampling distribution of b and a.

They give us the ability to describe a 95% confidence interval around b.

They let us say how certain we are about the estimates we’ve calculated from our sample.

The t-distribution

In practice, because our sample sizes won’t be exceedingly large and because we won’t know sigma in advance, we will assume that our estimates are drawn from a **t-distribution** instead of from a Normal distribution. With large sample sizes, t is not distinguishable from a Normal distribution.

In Practice

We will often measure our certainty about the coefficient with a t-statistic, based on a t distribution.

$$t = \frac{\text{our estimate of the coefficient}}{\text{the standard error of the coefficient.}}$$

When t is large, it suggests we are more certain that our estimate is different from 0.

The value of t

The value of t gives us a measure of whether there’s a lot of variability or a little in the relationship between y and x in our sample. Of course, the larger our sample size, the more certain we can be about our inferences.

Critical Values of t

For a 95% confidence interval.

Sample Size	**Critical Value of t**
10	2.23
100	1.98
Infinity	1.96

When the **absolute value** of t for our coefficient is greater than or equal to the critical value of t at a particular level of confidence, we have a measure of how certain we are about the coefficient at hand.

Conventionally, we use a 95% confidence interval, or a .05 level of statistical significance. Often, we also report the level of statistical significance.

P value

- The p value helps us understand how likely it is to get the sample estimate we got if the null hypothesis (often of no effect) is true.
- It gives us a sense of the probability of seeing results as or more extreme than those we actually observe if the effect is actually 0.

Drawing Inferences

Predicting Earnings, Ordinary Least Squares

Variable	Coefficient	S.E.	t
Height	1563.138	133.448	11.713
Constant	-84078.32	8901.098	-9.446

N = 1379

R-squared = .09

Source: Gelman and Nolan 2002.

Questions to ask

- On what scales are our variables measured?
- Are our coefficients statistically significant?
- Are our coefficients substantively significant?
- Are there omitted variables that will affect our estimates of the coefficients at hand?

A Multivariate Model

Predicting Earnings in US Dollars,
Ordinary Least Squares

Variable	Coefficient	S.E.	t	p-value
Height in inches	550.5448	184.57	2.983	.003
Woman	-11254.57	1448.892	-7.768	.000
Constant	-84078.32	8901.098	-9.446	.908

N = 1379
R-squared = .13
Source; Gelman and Nolan, 2002.

Another Multivariate Model

Predicting Hours Working, Ordinary Least Squares Regression

	Women	**Men**
Education	4.26***	1.92***
	(.60)	(.47)
Marriage	-0.53*	1.17***
	(.25)	(.24)
Pre-school Children	-2.25***	1.54***
	(.33)	(.32)
School-aged Children	-0.14	1.65***
	(.29)	(.28)
N	1288	1177
Adjusted R-Squared	.30	.44

Source: Burns, Schlozman, and Verba.

* p<.05; ** p<.01; *** p < .001. Controlling for other variables.

Predicting Free Time

	Women	**Men**
Marriage	-0.86***	-.32***
Pre-school Children	-2.29***	-.53***
School-aged Children	-0.88***	-.53***

Source: Burns, Schlozman, and Verba 2001.

Controlling for education, activity in high school, race or ethnicity, age, hours on the job, job level, and citizenship

*Coefficient significant at < .05.

**Coefficient significant at < .01.

***Coefficient significant at < .001.

Predicting Level of Education (US, from GSS data)

	1972	**2006**
Parents' Education	**.379***	**.415***
	(.028)	(.015)
Rural	**-.029***	**-.029***
	(.013)	(.013)
Age 26-35	.024	**.035***
	(.019)	(.011)
Age 36-45	.001	**.025***
	(.015)	(.011)
Age 46-55	-.006	**.028***
	(.018)	(.011)
Age 56-65	**-.040***	**.042***
	(.019)	(.012)
Age 66 and older	**-.074***	**.058***
	(.020)	(.010)
Female	-.005	-.011
	(.010)	(.007)
R-squared	.312	.293
N	597	1379

Interpreting coefficients

Ask the question:

Compared to what?

Another Multivariate Model

- Source: Neuenschwander, Vida, Garrett, and Eccles. 2007.

Predicting how good a student believes she or he is at math, ordinary least squares (n=528 children)

	Coefficient
Parents' Educational Expectations for Their Child	.22*
The Child's Prior Math Grades	.28*

A diagram of a series of coefficients in a model

Source: Neueschwander, Vica, Garett, and Eccles 2007.

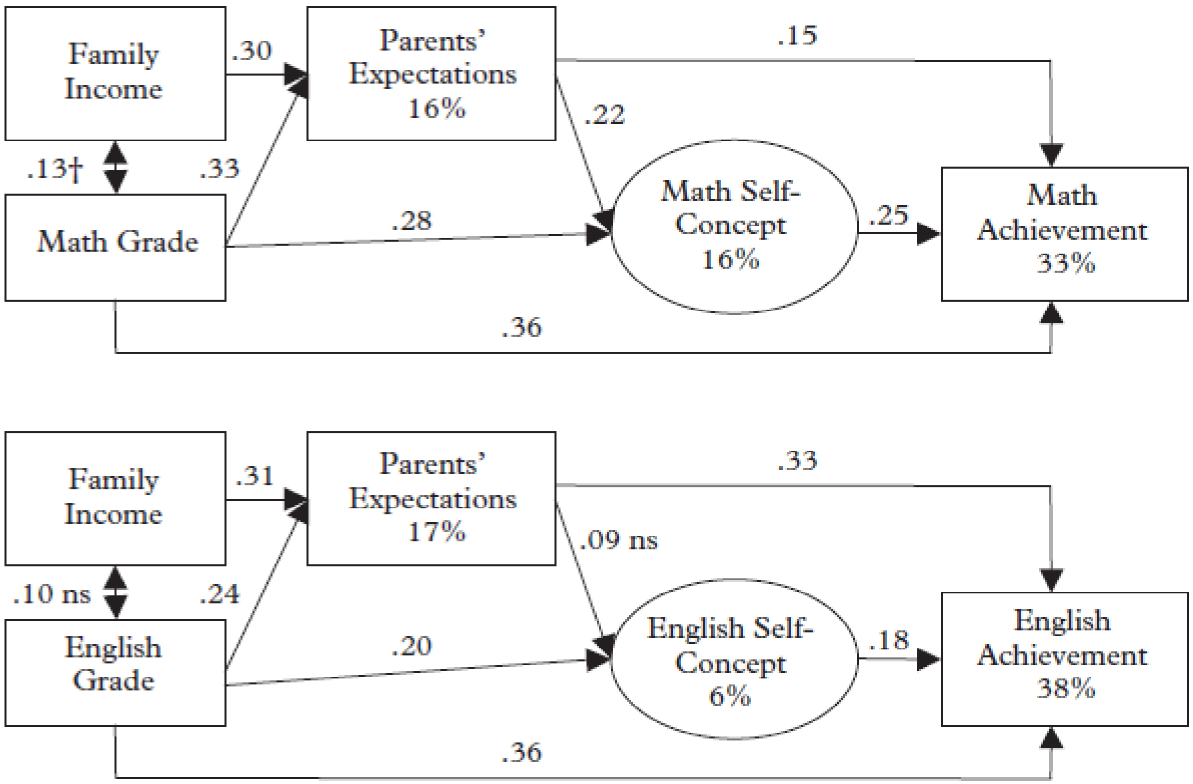


Figure 3. CAB explanation of school achievements in math and in English (6th grades, $^{\dagger}p < .10$, ns: non significant).

Omitted Variables

Number of hours of TV watching per day

	B (s.e.)	t	p value
Education	-2.02 (.22)	-9.29	.0000
Age	.14 (.21)	.66	.5124
Age over 65	.75 (.19)	4.04	.0001

Adjusted R-squared .06

N 2494

Omitted variables

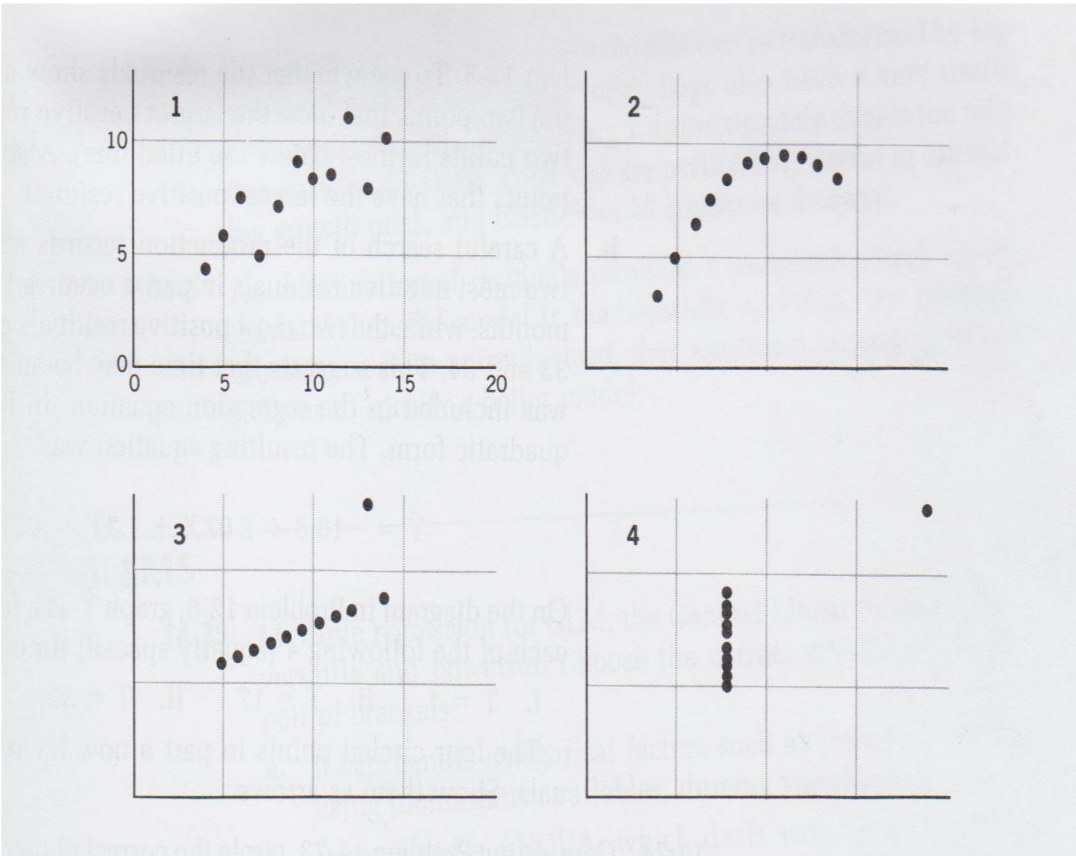
Number of hours of TV watching per day

	B (s.e.)	t	p value
Education	-1.55 (.21)	-7.35	.0000
Age	-.15 (.20)	-.74	.4588
Age over 65	-.06 (.19)	-.008	.7633
In the workforce	-1.57 (.11)	-14.466	.0000

Adjusted R-squared	.13
N	2494

Class Exercise

- Source: Wonnacott and Wonnacott.



Know your data

Know how it was collected.

Know what the data look like.

Know how the variables are distributed.

Know what the residuals look like.

Explore the difference between observations your model predicts well and cases your model doesn’t predict well.

Regression is Resilient

We are fortunate to have a tool like ordinary least squares regression. It is a sturdy tool, and yields useful estimates even in the face of mild deviations from assumptions.

Not all tools of estimation have this resilience.

Challenges to Regression's Resilience

Three things that make the Central Limit Theorem not hold:

- Non-random samples.
- Heteroskedasticity, varying variances of Y.
- Autocorrelation, the lack of independence of our observations of Y.

Without the Central Limit Theorem

We cannot estimate the standard errors well that allow us to move from our samples to describe the population.

The other challenge to regression's resilience

Model specification, or what variables we choose to include in our models and why.

We Have Tools

- We have tools to use to detect the problems we just talked about.
- And we have tools to use to correct those problems.
- We just have to know what sorts of problems we could encounter.

For further reading

Wonnacott and Wonnacott. 1990. Introductory Statistics for Business and Economics, 4th edition. John Wiley and Sons.

For those comfortable with more mathematics:

William H. Greene. 2008. Econometric Analysis, 6th edition. Prentice-Hall.

Further Reading

Gelman and Nolan. 2002. **Teaching Statistics**. Oxford: Oxford University Press.