

Conducting Empirical Legal Scholarship

Lee Epstein¹ Andrew D. Martin²

¹Northwestern University
School of Law & Department of Political Science

²Washington University in St. Louis
School of Law & Department of Political Science

The Course

- 1 Overview of Empirical Research
- 2 Designing Research
- 3 Collecting & Coding Data
- 4 Statistical Software
- 5 The Logic of Statistical Inference
- 6 Data Analysis

The Course

- 1 Overview of Empirical Research
- 2 Designing Research
- 3 Collecting & Coding Data
- 4 Statistical Software
- 5 The Logic of Statistical Inference
- 6 Data Analysis

More Data Analysis

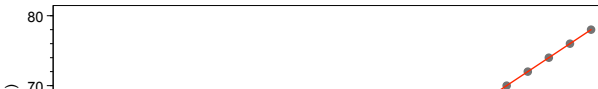
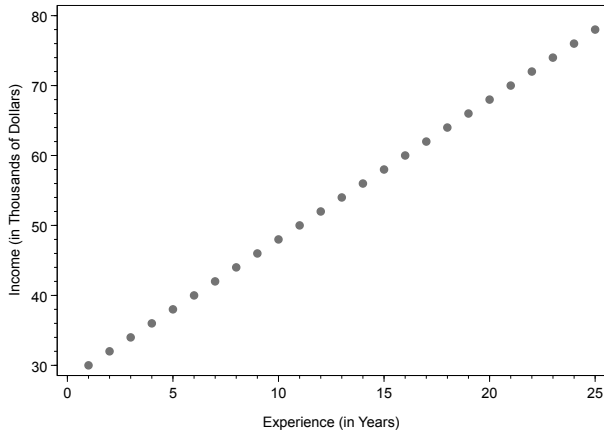
- Simple Linear Regression: A Conceptual Overview
- Simple Linear Regression: Statistical Control, Inference
- Multiple Regression
- Logit/Probit

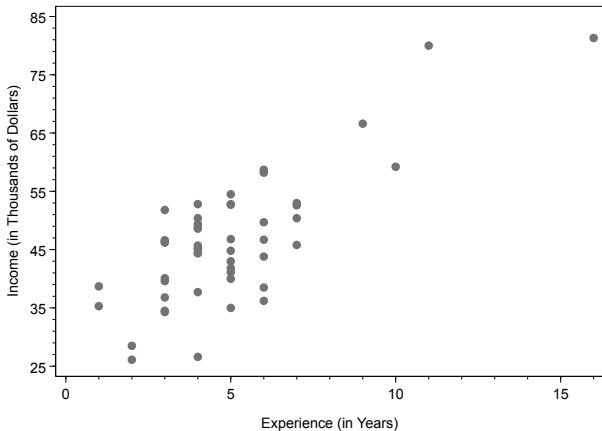
What is regression?

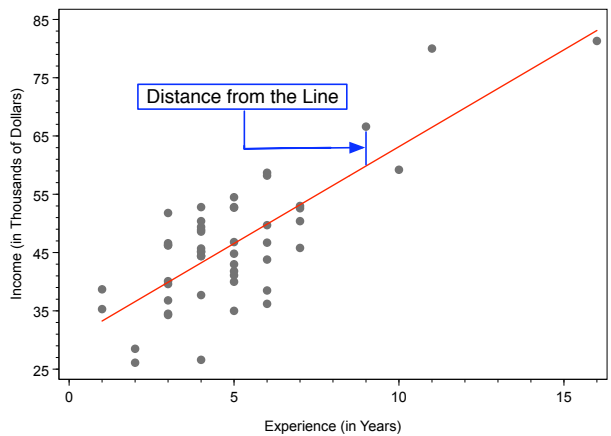
A tool to understand (“make an inference about”) the relationship between a dependent variable (Y) and one or (typically) more independent variables (X). The general idea is that we can explain Y in terms of the variation in the values of X . In its strongest form, $X \rightarrow Y$, that is, X causes Y .

What is simple linear regression?

Linear regression helps us understand the relationship between Y and X when the two are linearly related and Y is measured at the interval level.

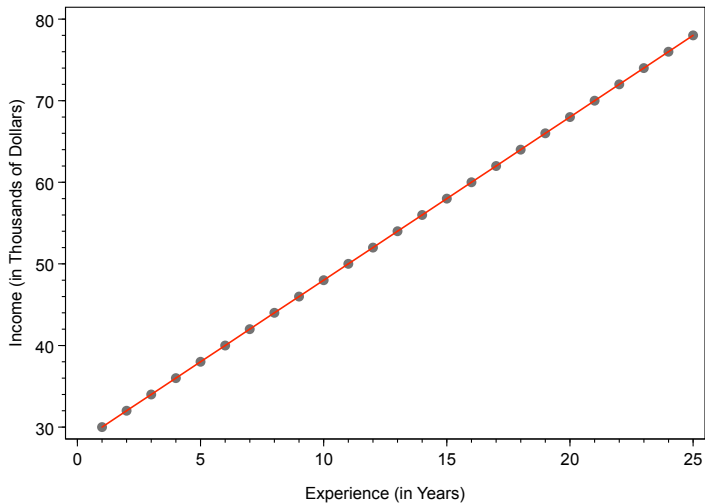






Think About a Straight Line: $Y = a + bX$

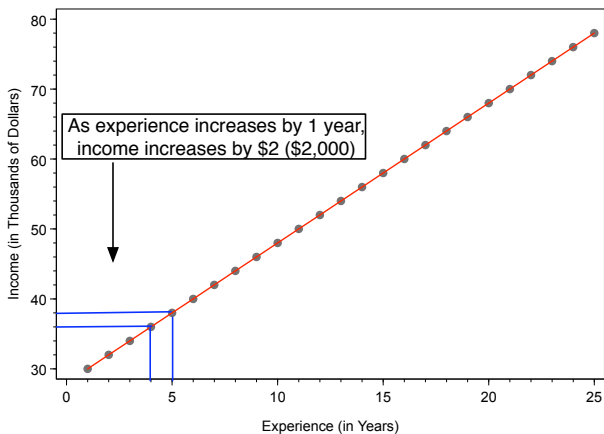
- Y = dependent variable (on Y axis)
- X = independent variable (on X axis)
- a = intercept (predicted value of Y when $X=0$)
- b = slope (the change in Y when X changes by 1 unit)



```
. regress fakeinc fakeexp
```

Source	SS	df	MS			
Model	5200	1	5200	Number of obs =	25	
Residual	0	23	0	F(1, 23) =	.	
Total	5200	24	216.666667	Prob > F =	.	
				R-squared =	1.0000	
				Adj R-squared =	1.0000	
				Root MSE =	0	

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
fakeinc	2
fakeexp	28
_cons					



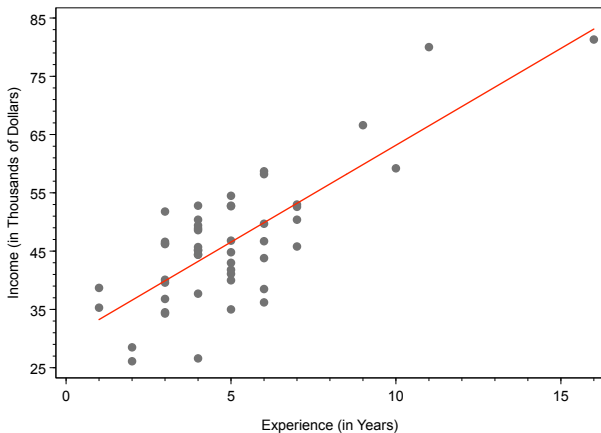
$$\text{Income} = 28 + (2 \times \text{Years of Experience})$$

Income = $28 + (2 \times \text{Years of Experience})$

- To calculate the expected income for a worker with four years of experience: $28 + (2 \times 4) = \$36$ (\$36,000)
- To calculate the expected income for a worker with five years of experience: $28 + (2 \times 5) = \$38$ (\$38,000)

Note the increase by \$2 (\$2,000). For every 1-unit change in the independent variable (experience), there's a \$2 (\$2,000) increase in the dependent variable (income).

[SOFTWARE DEMO]



$$\text{Income} = 29.9 + (3.3 \times \text{Years of Experience})$$

Income = $29.9 + (3.3 \times \text{Years of Experience})$

- To calculate the expected income for a worker with four years of experience: $29.9 + (3.3 \times 4) = \43.1 (\$43,100)
- To calculate the expected income for a worker with five years of experience: $29.9 + (3.3 \times 5) = \46.4 (\$46,400)

The State Dataset

- The independent variable is **hsdip** (% of population that has a high school diploma)
- The dependent variable is **turnout** (% voters casting ballots in races for the U.S. House)
- Exercises
 - 1 Generate a scatterplot with a regression line
 - 2 Estimate regression of **turnout** on **hsdip**

Linear Regression I

- The linear regression model can also be used to perform inference about the relationship between two variables.
- Y is the dependent variable.
- X is the independent variable.
- We are interested in the population relationship:

$$Y = \alpha + \beta X + \varepsilon$$

Linear Regression II

- α is the intercept. It is the expected value of Y when X equals zero.
- β is the slope. For a one unit increase in X , it is the expected increase in Y .
- Note that the slope captures the nature of the relationship. It could be positive, negative, or zero.

Estimating Linear Regression Parameters I

- If the population regression were true, we would only observe fundamental variability.
- We will also encounter sampling variability which we will also have to take account of.
- We will assume that the distribution of the residuals is Normal (bell-shaped).
- The linear regression can thus be written:

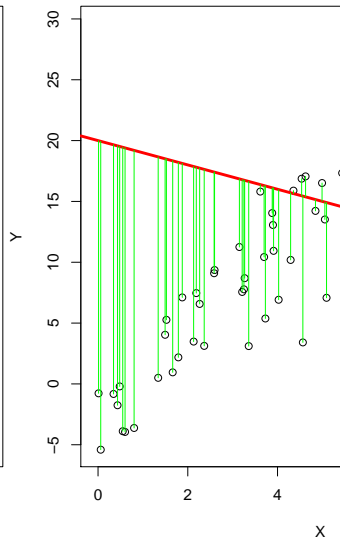
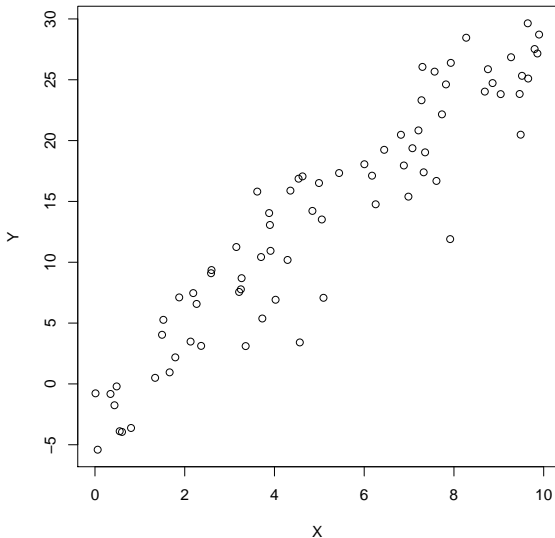
$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \sigma^2)$$

Estimating Linear Regression Parameters II

- This is a model with three parameters.
- How do we estimate them? Ordinary Least Squares.
- This results in our estimates (the formulas you can look up in any textbook):

$$a, b, \text{ and } \hat{\sigma}^2$$

Finding the Ordinary Least Squares (OLS) Line



Inference for the Linear Regression Model

- What would be the key hypothesis to test for the linear regression model?

$$H_0 : \beta = 0$$

- If we could reject the null, we could conclude that there is a statistically significant relationship between the dependent variable and the independent variable.
- Statistical software reports these p-values, which are also computed from the Student-t distribution.

Model Fit

- We also might want to know, in general, how well a model fits.
- One quantity is the *standard error of the estimate* (SEE): $\hat{\sigma}$.
- Another quantity is called the *R-squared*.
- *R-squared* is the percentage of explained variance, and is the correlation coefficient to the second power.
- When a model is perfect, *R-squared* goes to one, and *SEE* goes to zero.

Guidelines

- Remember that *inference* is the goal.
- Lines are oftentimes good approximations, but sometimes relationships are non-linear. Make sure to visualize the relationship.
- Linear regression, in general, is non-robust. Make sure to check for outliers.
- A significant relationship does not imply a causal relationship.
- Resist the urge to extrapolate beyond the range of the observed data.

Simple Linear Regression in Stata [`state.dta`]

Simple Linear Regression in Stata [`state.dta`]

In-Class Assignment [nom.dta]

In-Class Assignment [nom.dta]

Statistical Control

- Experiments – The Gold Standard
- Observational Data
- Potential Outcomes
- Simpson's Paradox

Simpson's Paradox

- Suppose we are interested in how counsel affects outcomes in criminal DUI cases.
- Here are some hypothetical data:

	Public Def.	Private
Win	20	24
Lose	20	16
Percent Win	50%	60%

- But maybe the sex of the defendant is relevant.

Simpson's Paradox

- We will thus control for *both* independent variables at the same time

- For men:

	Public Def.	Private
Win	12	3
Lose	18	7
Percent Win	40%	30%

- For women:

	Public Def.	Private
Win	8	21
Lose	2	9
Percent Win	80%	70%

Implications

- The punch-line is that if we do not control for the right variables, we may not only get no results, but we might get wrong results.
- How do we know whether our models are properly specified?
 - Theory
 - Robustness Checking

Multiple Regression

- We would like to expand the simple linear regression model to allow for statistical control.
- Now we have two (or more) independent variables: X_1 and X_2 .
- To find the effect of X_2 while controlling for X_1 , and vice versa, consider the following:

$$Y_i = \alpha + \beta_1 X_{1,i} + \beta_2 X_{2,i} + \varepsilon_i$$

Notes on Multiple Regression

- The intercept coefficient is now the expected value of Y when both X s are zero.
- There are now two partial slope coefficients, which capture the effect of a one unit change in one X while holding the other constant.
- The measures of model fit are the same.
- The model can be extended to include as many explanatory variables as you want.
- This model is the workhorse of empirical social science (along with its cousins for other types of dependent variables).

Multiple Regression in Stata [nom.dta]

Multiple Regression in Stata [nom.dta]

Independent Variables

- Dichotomous (yes/no) variable
- Nominal or ordinal variable
- Interactions
- Non-linearities

Things to Learn in a Regression Class

- Diagnostics for outliers and leverage
- Diagnostics for non-linearity
- Diagnostics for incorrect distributional assumptions (non-constant error variance, or time-correlated errors)
- Multi-collinearity

Models for Other Dependent Variables

- Dichotomous – logit or probit regression
- Ordinal – ordinal logit or ordinal probit regression
- Nominal – multinomial (or conditional) logit or multinomial probit
- All of these models are non-linear, which makes interpretation a bit more difficult

Conclusion

Things I Have Always Wanted to Know
About Statistics But Have Been Afraid to
Ask...