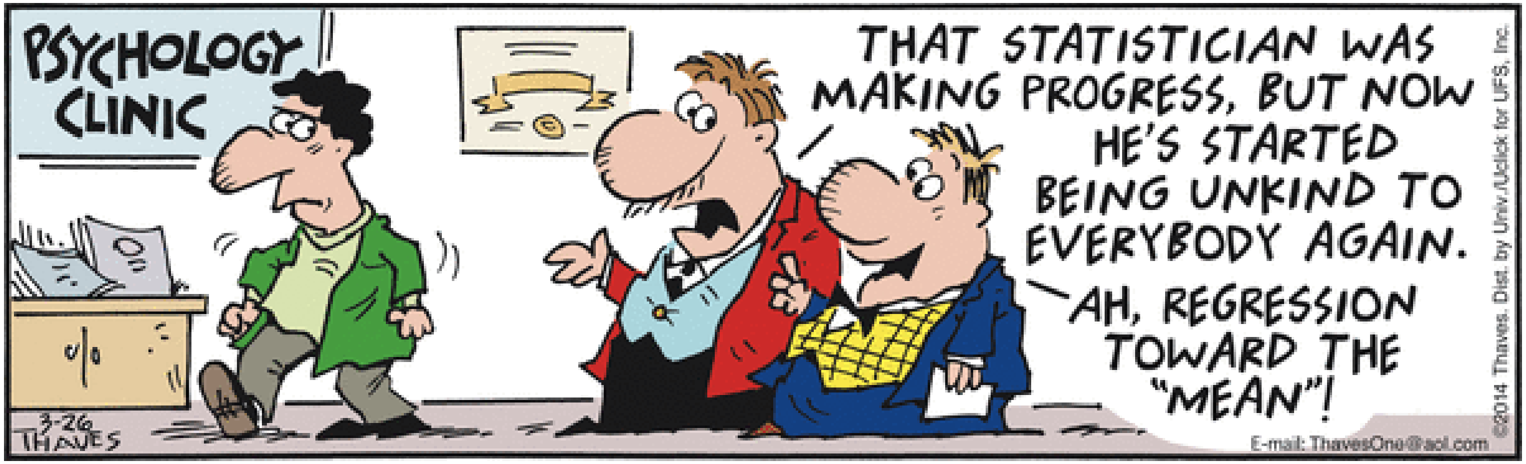
Lecture 2: Bivariate Regression Model



Olivier Deschenes, UCSB, ESM 296, Winter 2018

Lecture 2: Bivariate Regression Model

- Background
- Population regression line
- OLS estimator of the population regression line
- Assumptions of the linear regression model
- Sampling distribution of OLS estimator
  - Law of large numbers
  - Central limit theorem
- Stata examples

Linear Regression Model

- Why do we estimate regressions?

- Linear regression allows us to estimate population slope coefficients, that <u>quantify the association between 2 or more variables</u> and <u>make inferences about them</u> (prediction, hypothesis tests, confidence intervals, etc)

- Ultimately our goal will be to estimate the ***<u>causal effect</u>*** on Y of a unit change in the variable X
  - This will depend crucially on whether the regression errors are uncorrelated with X
  - For now, just think of the problem of fitting a straight line to data on two variables, Y and X

The Population Linear Regression Model

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1,\ldots, n$$

- X is the ***independent variable*** or ***regressor***
- Y is the ***dependent variable***
- β_0 = ***intercept***
- β_1 = ***slope (recall that is $\Delta Y/\Delta X$)***
- The slope measures the change in Y for a 1-unit change in X. The magnitude, sign, and statistical significance of β_1 are important
- u_i = **regression *error***
  - The regression error consists of factors omitted from the model, or possibly measurement error in the measurement of Y. In general, these omitted factors are other factors that influence Y, other than the variable X

The Population Linear Regression Model (ctd)

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1,\ldots, n$$

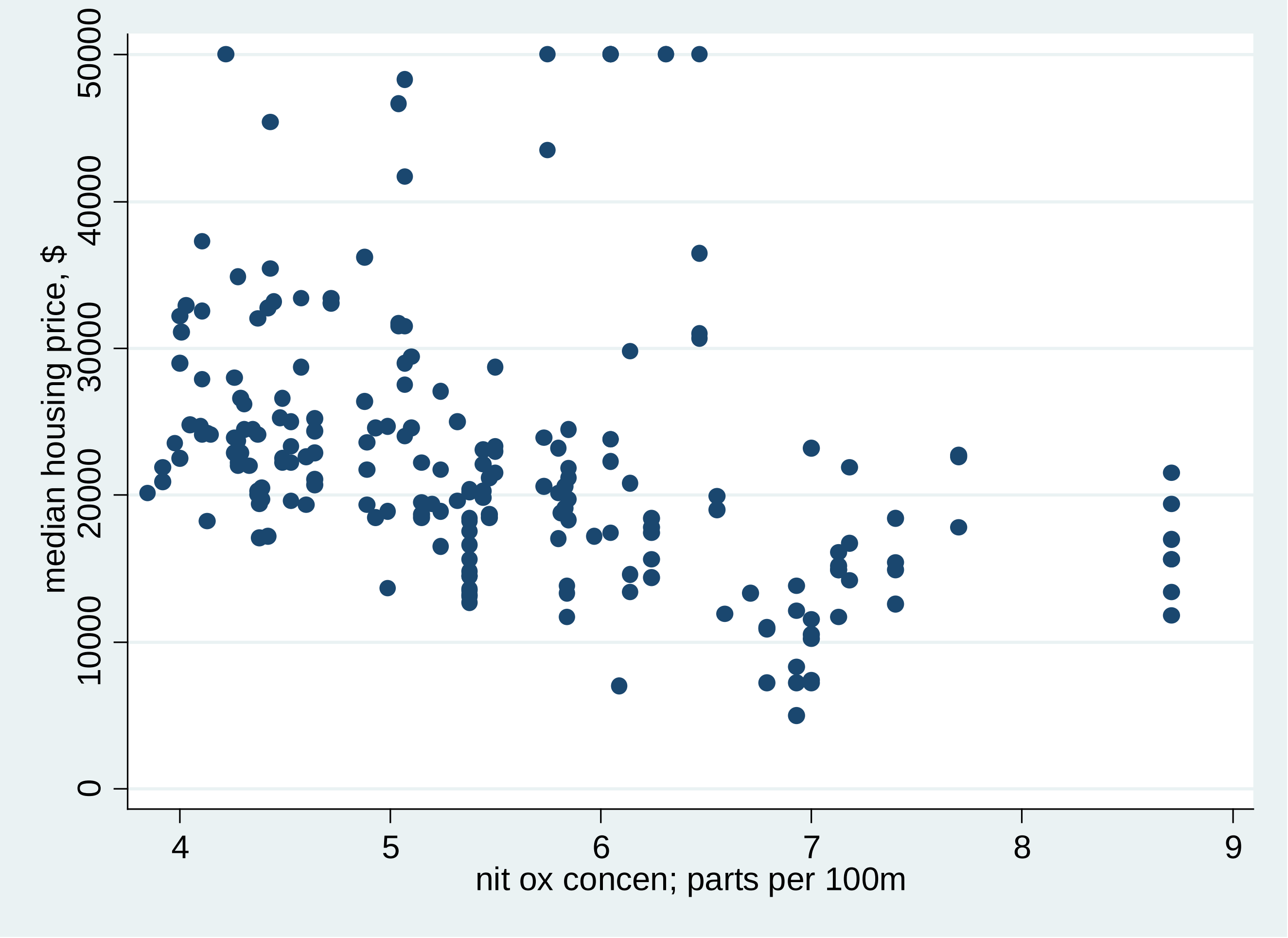
- β_0 = ***intercept***
- β_1 = ***slope***
- $\Rightarrow$ *Both* β_0 and β_1 are unknown parameters

- *We collect a sample of observations on Y and X with the goal of "correctly" estimating* β_0 and β_1

- u_i = **regression *error***
- *Unobserved... Can construct fitted residuals by using estimates for* β_0 and β_1

Data Example (N=206):

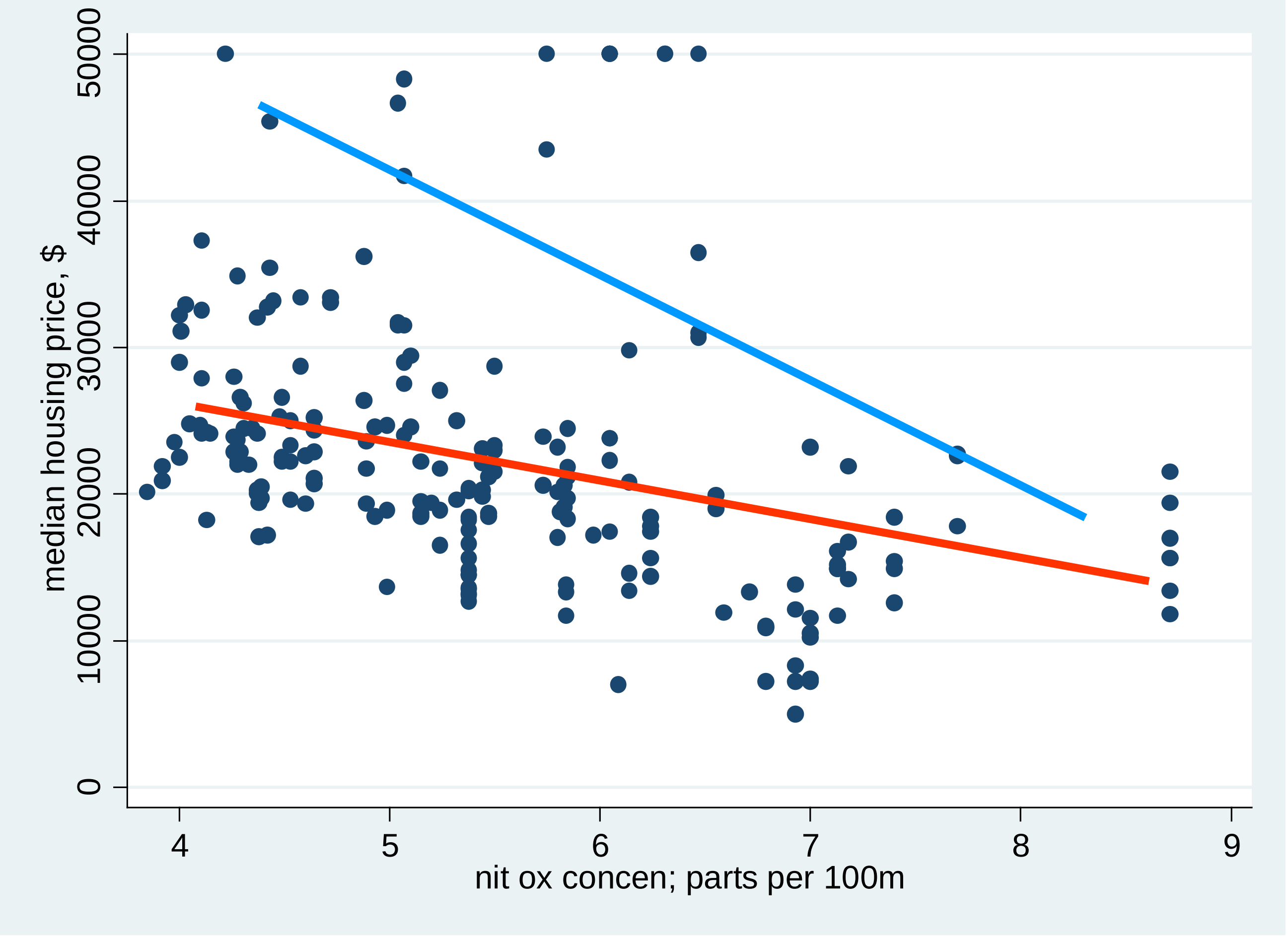
- Data on median housing value by Census tract and NOx concentrations in parts per 100 million (i.e. 100*ppm)
- From Boston MSA in 1970 (Harrison and Rubinfeld, 1978)



In a picture: Observations on Y and X; the fitted regression line; and the regression residuals (the fitted "error term"):

- How to choose the best line? Several criterions exist. Below are 2 possible lines

- OLS chooses the line that minimizes the squared prediction errors (mean-square error). This is the origin of the “least squares” term



- **The OLS estimator** solves: $\min_{b_0, b_1} \sum_{i=1}^{n} [Y_i - (b_0 + b_1 X_i)]^2$

- The OLS estimator minimizes the (average) squared difference between the actual values of Y_i and the prediction ("predicted value") based on the estimated line

- The OLS estimator is denoted by $\hat{\beta}_1$ and $\hat{\beta}_0$

- This minimization problem can be solved using calculus (i.e. solving FOC)

The OLS Estimator, Predicted Values, and Residuals

The OLS estimators of the slope β_1 and the intercept β_0 are

Ratio of sample covariance of Y and X to sample variance of X

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \frac{s_{XY}}{s_X^2} \tag{4.7}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1\overline{X}. \tag{4.8}$$

The OLS predicted values $\hat{Y}_i$ and residuals $\hat{u}_i$ are

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i,\ i = 1, \ldots, n \tag{4.9}$$

$$\hat{u}_i = Y_i - \hat{Y}_i,\ i = 1, \ldots, n. \tag{4.10}$$

The estimated intercept ($\hat{\beta}_0$), slope ($\hat{\beta}_1$), and residual ($\hat{u}_i$) are computed from a sample of n observations of X_i and Y_i, $i = 1, \ldots, n$. These are estimates of the unknown true population intercept (β_0), slope (β_1), and error term (u_i).

OLS Estimator in Bivariate Regression

- From the previous slide, a key formula:

$$\beta_1 = \frac{Cov(Y_i, X_i)}{Var(X_i)} \qquad \hat{\beta}_1 = \frac{S_{XY}}{S_X^2} = Corr(Y, X)\frac{S_Y}{S_X}$$

- In words, $\hat{\beta}_1$ is the sample correlation coefficient between Y and X multiplied by the ratio of the standard deviation of Y to the standard deviation of X

- This is applicable only in the bivariate regression model. We will return to this formula when we discuss multivariate regression...

OLS regression: STATA example

```
regress price nox, robust;
```

```
Linear regression                                      Number of obs =     206
                                                       F(  1,   204) =   44.86
                                                       Prob > F      =  0.0000
                                                       R-squared     =  0.1146
                                                       Root MSE      =    8849

------------------------------------------------------------------------------
             |               Robust
       price |      Coef.   Std. Err.       t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
         nox |  -2775.674   414.4046    -6.70    0.000    -3592.739   -1958.608
       _cons |   38068.27   2222.545    17.13    0.000     33686.17    42450.38
------------------------------------------------------------------------------
```

$$\hat{\text{Price}} = 38068 - 2775.7 \times NOX$$

Estimate of Slope Coefficient "By Hand"

```
. summarize price nox;

    Variable |       Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       price |       206    22723.11    9381.108       5000      50001
         nox |       206    5.528447    1.143977       3.85       8.71

. correlate price nox;
(obs=206)

             |    price      nox
-------------+------------------
       price |   1.0000
         nox |  -0.3385   1.0000
```

- So $\hat{\beta}_1$ = -0.3385*9381.1/1.14 = -2775

Interpretation of the Estimated Slope and Intercept

$$\hat{\text{Price}} = 38068 - 2775.7 \times NOX$$

- The slope means that each additional unit of concentration of NOx per 100 million reduces house values by $2,776
  - i.e. going from 3 pp100m to 4 pp100m lowers house values by 2776 on average (12% of average housing value)

- The intercept (taken literally) means that, houses in Census tracks with zero concentrations of NOx are worth $38,068 on average
  - The intercept is of limited use in practice: it often extrapolates the line outside the natural range of the data (here NOx concentrations range from 3.9 and 8.7 pp100m)
  - From now on, we will always include an intercept in the models, but rarely discuss it

The Least Squares Assumptions:

$$Y_i = \beta_0 + \beta_1 X_i + u_i, \qquad i = 1,\ldots, n$$

- **(LSA.1): The conditional distribution of *u* given *X* has mean zero, that is, $E(u_i|X_i) = 0$ [This implies X_i and u_i are uncorrelated]**
  - *This is the key assumption implying that the OLS estimator is consistent (i.e. that OLS unbiased in large samples), that is $\hat{\beta}_1$ is a "correct" estimate of the causal effect of X on Y*
  - ***Untestable*** *assumption without more information*

- **(LSA.2): (X_i, Y_i), $i = 1,\ldots,n$, are i.i.d**
  - *This is true if X_i, Y_i are collected by simple random sampling*
  - *This delivers the sampling distribution of the OLS estimator*
  - *Time-series data not iid*

- **(LSA.3): Large outliers in X_i and/or Y_i are rare**
  - *Technically, (3) means that X_i and Y_i have finite fourth moments. Required to derive the sampling variance of the OLS estimator*

Least Squares Assumption #1: $E(u_i|X_i = x) = 0$

- Recall our model: $Price_i = \beta_0 + \beta_1 NOx_i + u_i$,
- u_i = regression error = other factors that predict house values (besides NOx concentrations)

- What are some of these “other factors”?
  - Other pollutants?
  - Noise? Location close to major roads? Industrial sites?
  - Size of house, age of house, Local amenities, etc

- Is $E(u_i|X_i=x) = 0$ plausible for these other factors?
  - i.e, suppose u_i only composed of 1 factor, another pollutant, PM10
  - Do you believe that $E[PM10_i \mid NOx_i=3] = E[PM10_i \mid NOx_i=9]$

- A benchmark for thinking about LSA #1 assumption is to consider an ideal <u>randomized controlled experiment</u>:
  - X is randomly assigned to subjects (e.g., patients randomly assigned to different medical treatments)
  - Randomization is done following a fixed protocol
  - Because X is assigned randomly, all other individual characteristics – the things that make up the regression error "u" – are independent of X
  - **Thus, in an ideal randomized controlled experiment, $E(u_i|X_i = x) = 0$ (that is, LSA #1 (*very likely*) holds)**
  - **With <u>non-experimental data</u>, we will need to think hard about whether $E(u_i|X_i = x) = 0$ holds**

LSA #2: (X_i, Y_i), i = 1,…,n are i.i.d.

- This arises automatically if the entity "i" (individual, district, etc) is sampled by simple random sampling (SRS): the entity is selected then, for that entity, *X* and *Y* are observed (recorded)

- All the data encountered in this class, while not directly from a SRS will be assumed to be iid
  - Allows us to use simple law of large numbers (LLN) and central limit theorem (CLT)

- Non-i.i.d. sampling occurs for example when data are recorded over time ("time series data"), we will not study these models

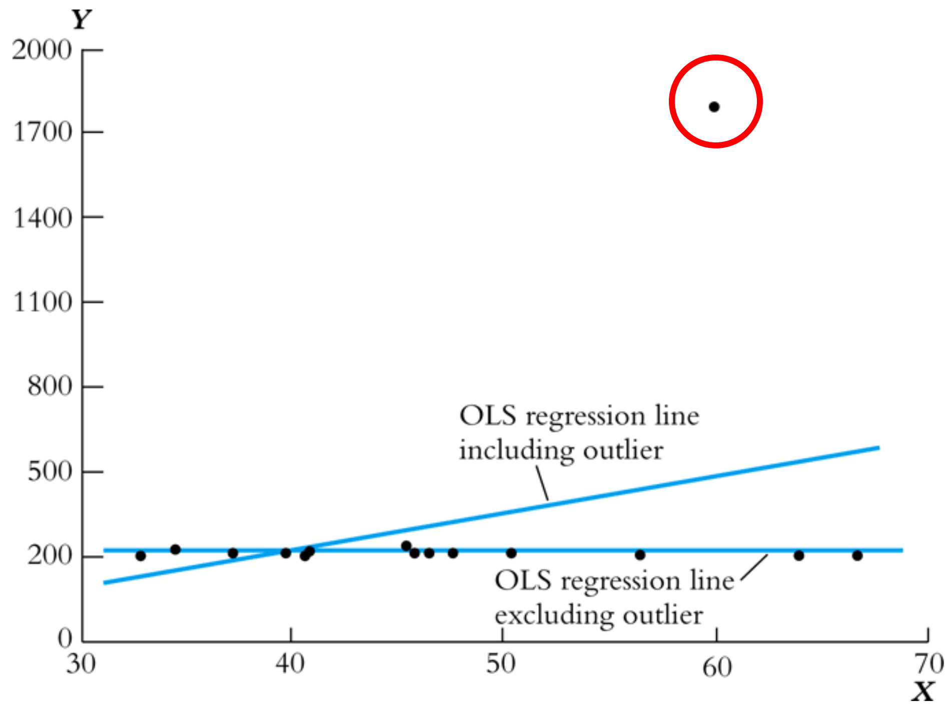
LSA #3: *Large outliers are rare*
Technical statement: $E(X^4) < \infty$ and $E(Y^4) < \infty$

- A large outlier is an extreme value of X or Y

- On a technical level, if X and Y are <u>bounded</u>, then they have finite fourth moments. (All variables considered in this course will satisfy this)

- However, the substance of this assumption is that outliers can strongly influence the results
  - Special estimators other than OLS exist for cases like these

OLS can be sensitive to an outlier:



- In practice, outliers are often data glitches (coding or recording problems). Sometimes they are observations that really shouldn't be in your data set. <u>Scrutinize your data!</u>

The Sampling Distribution of the OLS Estimator

- The OLS estimator is computed from a sample of data $\Rightarrow$ Different samples will give a different value of $\hat{\beta}_1$

- Since the OLS estimator is a function of the data, and that the data is from random sample, the OLS estimator is a <u>random variable</u>, with a probability distribution

- <u>**We want to:**</u>
- 1. Quantify the sampling uncertainty associated with $\hat{\beta}_1$
- 2. Use $\hat{\beta}_1$ to test hypotheses such as $\beta_1 = 0$ or $\beta_1 < 4$
- 3. Construct a confidence interval for β_1
- $\Rightarrow$ **Need <u>sampling distribution</u> of the OLS estimator**

Two Approaches to Derive the Sampling Distribution:

- **(a) Finite sample approach (exact)**
- Show that OLS estimator is unbiased
- Assume u_i is distributed normal to get sampling distribution

- **(b) Large sample approach (approximation)**
- Consider the case where sample size n grows arbitrarily "large"

- Law of large numbers implies that OLS estimator is ***consistent***
  - ***If LSA 1, 2, and 3 are satisfied***

- Central limit theorem implies that the sampling distribution of the OLS estimator is approximately ***normal***
  - ***If LSA 1, 2, and 3 are satisfied***

Law of Large Numbers

Convergence in Probability, Consistency, and the Law of Large Numbers

The sample average $\overline{Y}$ converges in probability to μ_Y (or, equivalently, $\overline{Y}$ is consistent for μ_Y) if the probability that $\overline{Y}$ is in the range $\mu_Y - c$ to $\mu_Y + c$ becomes arbitrarily close to one as n increases for any constant $c > 0$. This is written as $\overline{Y} \xrightarrow{p} \mu_Y$.

The law of large numbers says that if Y_i, $i = 1, \ldots, n$ are independently and identically distributed with $E(Y_i) = \mu_Y$ and $\mathrm{var}(Y_i) = \sigma_Y^2 < \infty$, then $\overline{Y} \xrightarrow{p} \mu_Y$.

<u>In words</u>: Sample mean of a random variable converge in probability to population mean

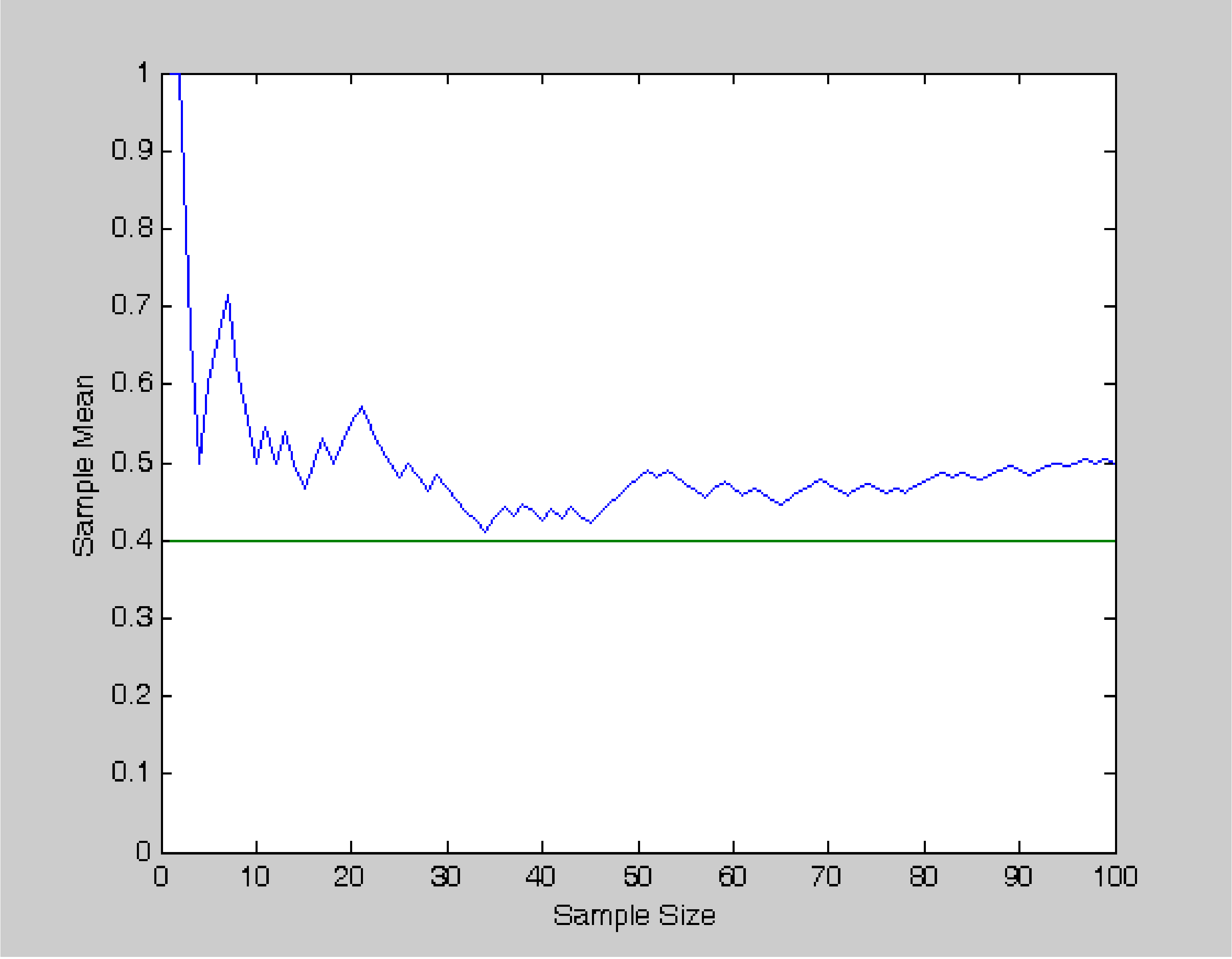
Same logic applies to OLS estimator: Under LSA 1-3, OLS estimator $\hat{\beta}_1$ gets very close to the population value of β_1 (converge in probability) when n grows large

Law of Large Numbers: Example

Suppose X_i~Bernoulli(0.4)

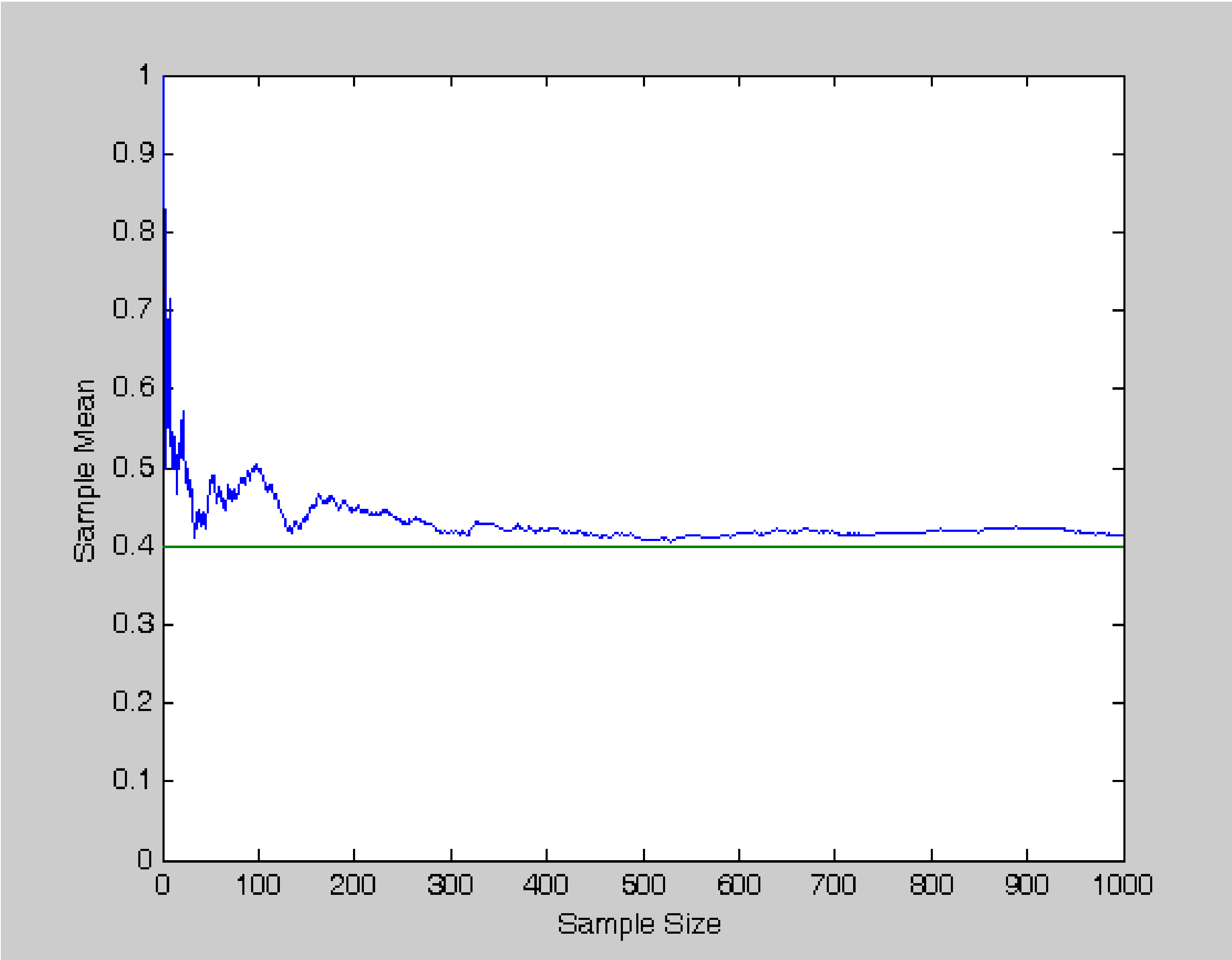
And consider sample mean $Y_n=(1/n)\Sigma X_i$ as a function of "n", for n=1,2,…,10000

Graphical illustration of example X_i~Bernoulli(0.4), $Y_n = (1/n)\Sigma X_i$



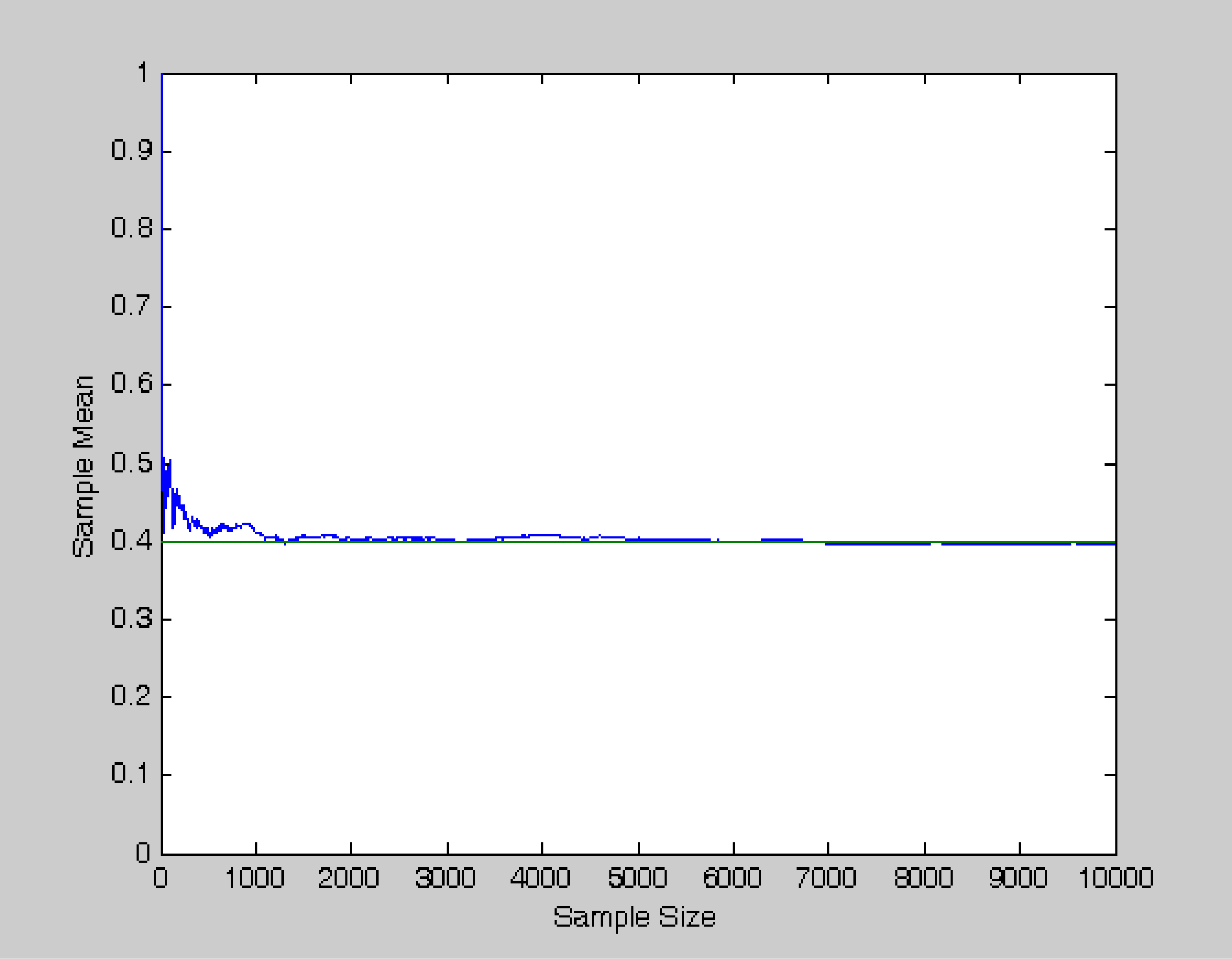
n=1 - 100

Graphical illustration of example
$X_i \sim$Bernoulli(0.4), $Y_n = (1/n)\Sigma X_i$



n=1 - 1000

Graphical illustration of example
X_i~Bernoulli(0.4), $Y_n = (1/n)\Sigma X_i$



n=1 - 10000

Central Limit Theorem

The Central Limit Theorem

Suppose that $Y_1, \ldots, Y_n$ are i.i.d. with $E(Y_i) = \mu_Y$ and $\text{var}(Y_i) = \sigma_Y^2$, where $0 < \sigma_Y^2 < \infty$. As $n \longrightarrow \infty$, the distribution of $(\overline{Y} - \mu_Y)/\sigma_{\overline{Y}}$ (where $\sigma_{\overline{Y}}^2 = \sigma_Y^2/n$) becomes arbitrarily well approximated by the standard normal distribution.

- Or put in another way, as sample size n gets arbitrarily large:

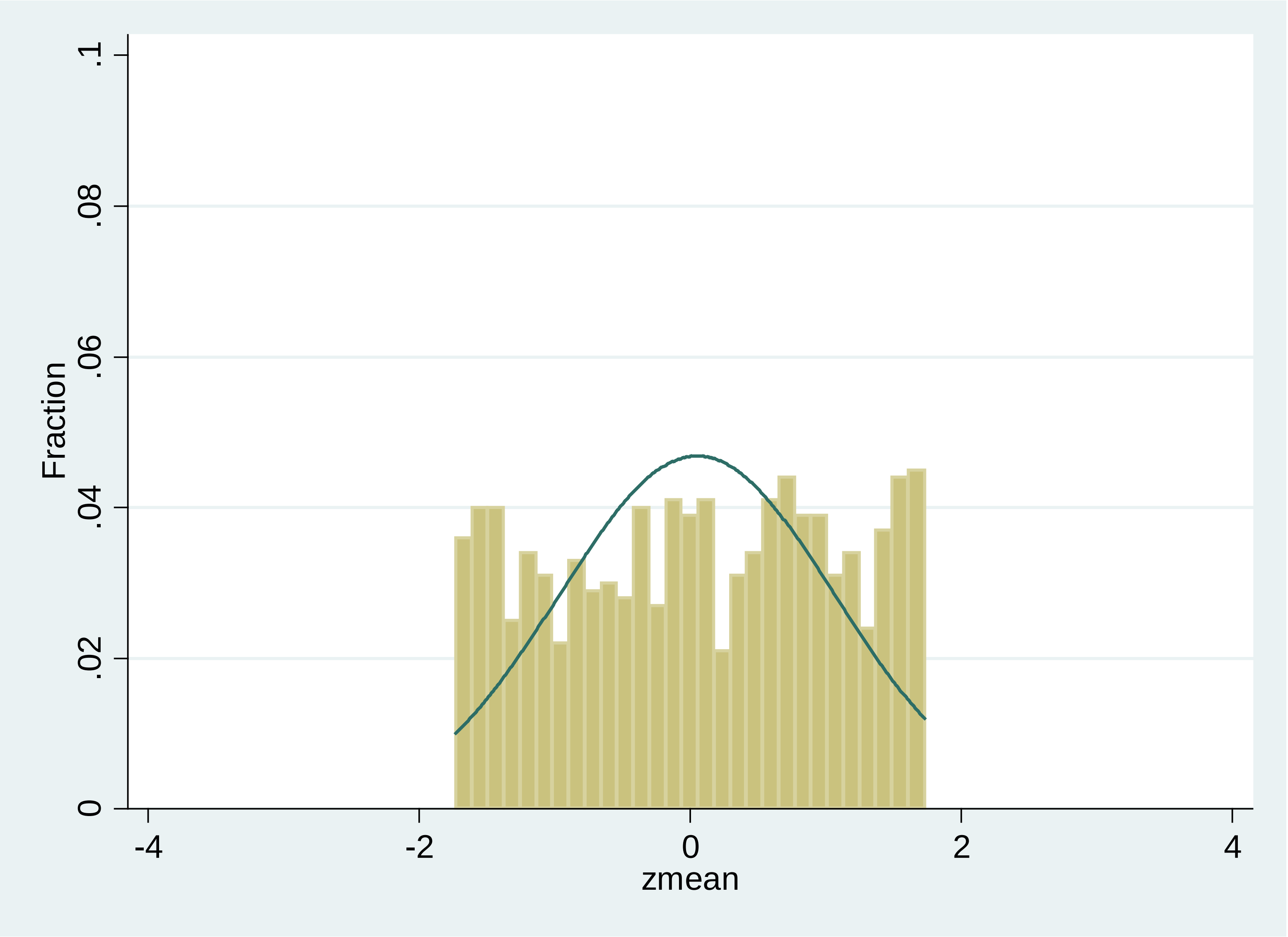
$$\overline{Y} \overset{A}{\approx} N\left(\mu_Y, \frac{\sigma_Y^2}{n}\right)$$

Example: Central Limit Theorem

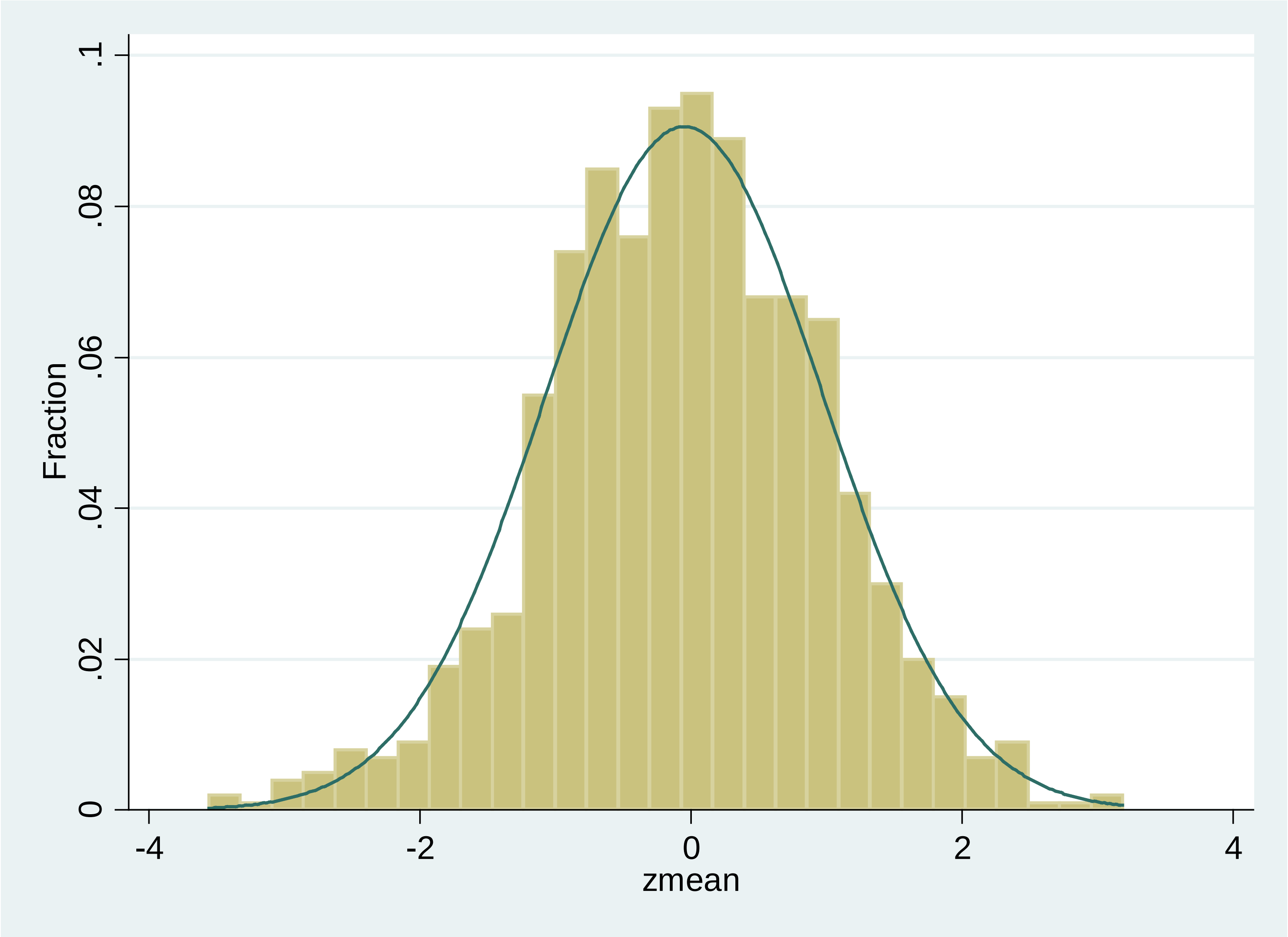
- Consider the following experiment:
- Draw a sample of size “n” from the Uniform [0,1] distribution and calculate the sample mean
- Rescale by subtracting population mean (0.5), multiplying by √n and dividing by standard deviation (1/12)
- CLT says that as “n” grows large, distribution of rescaled sample mean should get closer and closer to a Normal distribution (here a N(0,1))

Example: Central Limit Theorem

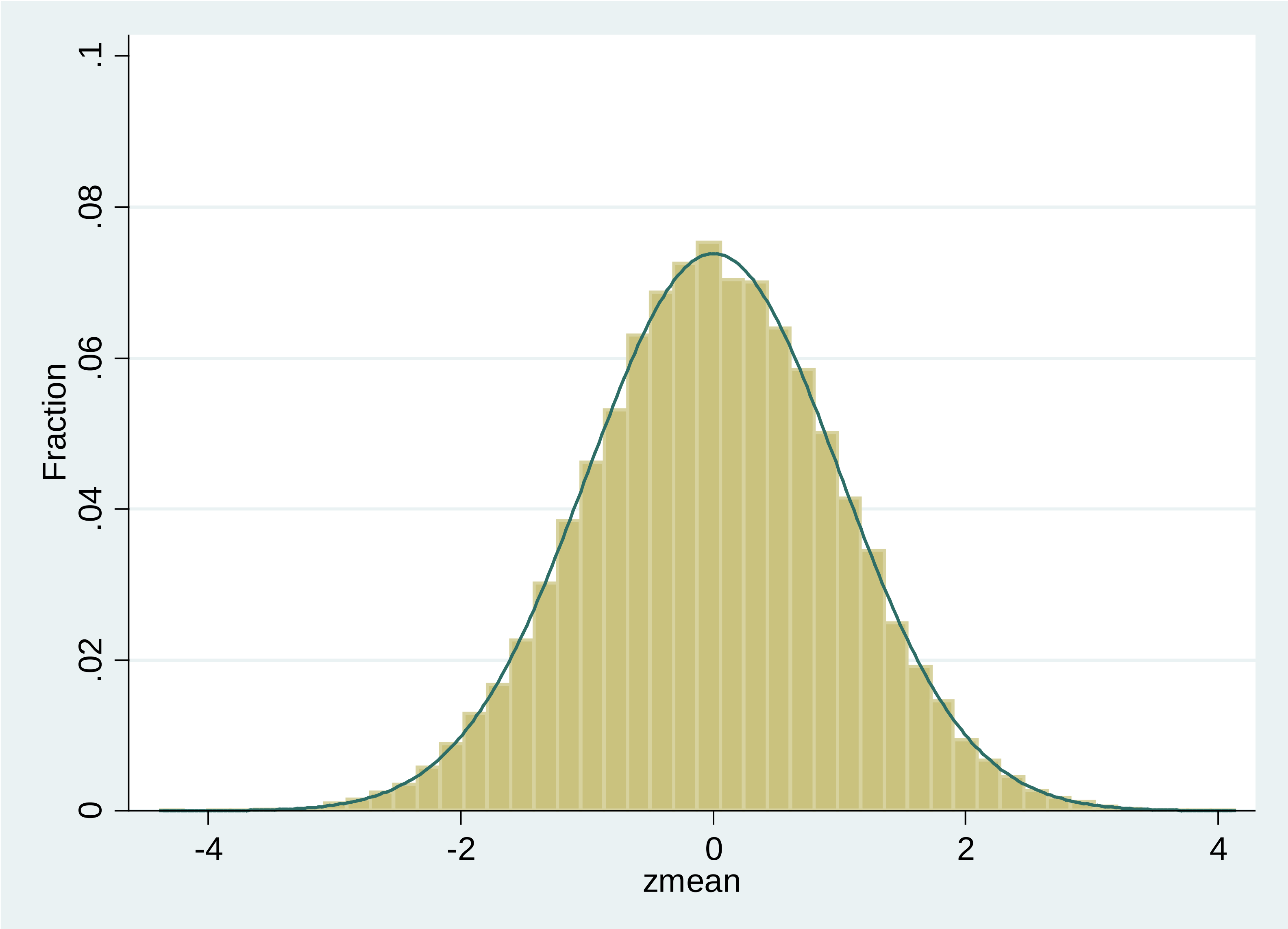
- Here "n" =1 (1000 samples)



“n” =30 (1000 samples)



“n” =3000 (1000 samples)



Establishing the Consistency of the OLS Estimator using Monte Carlo Simulation

- Fix population values: $\beta_0=3$ and $\beta_1=0.5$
- Generate values for $X_i=\{0,1\}$
- Generate random regression error terms $u_i \sim N(0,\sigma^2)$
- Note that by construction $E[u_i|X_i]=0$, so LSA#1 satisfied

- Construct $Y_i = 3 + 0.5*X_i + u_i$ $\quad\quad$ $i=1,\ldots n$

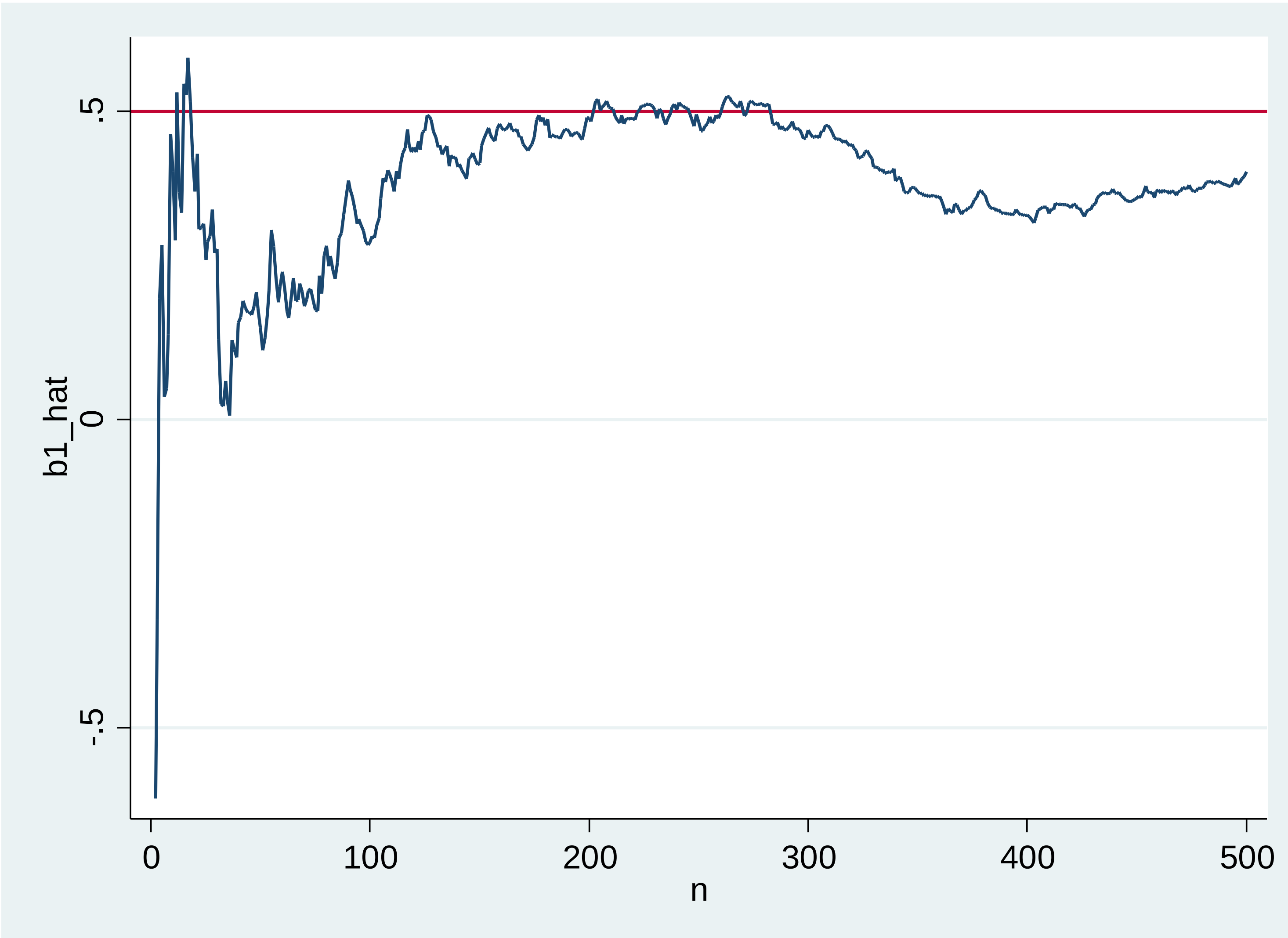
- Consider sample size n=50,000

- Estimate β_0 and β_1 by OLS for samples of size n=2, 3, 4, ..., 50,000, ... (here focus on β_1)

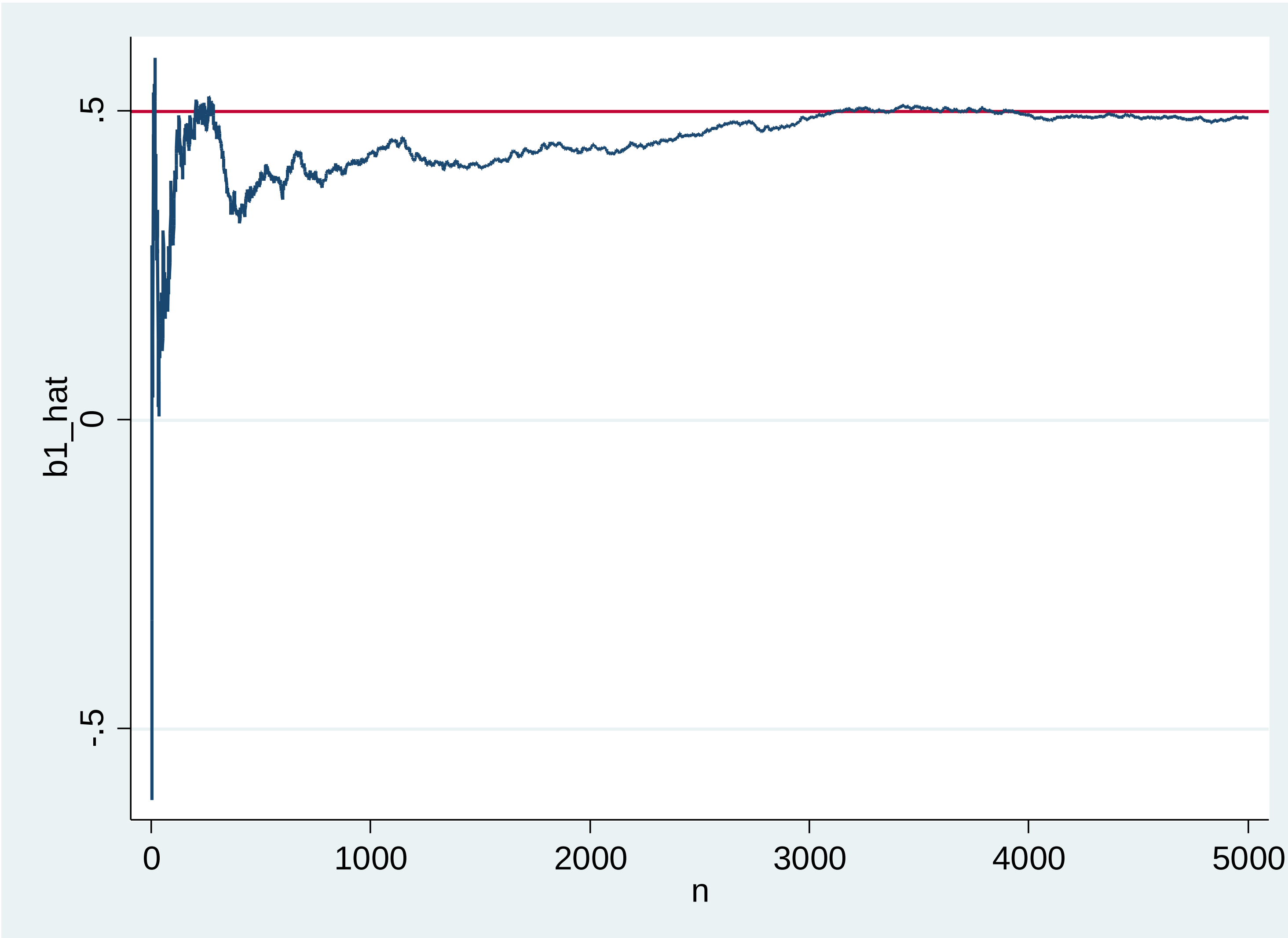
-

- $\Rightarrow$ **As we look further and further away in the sequence (n increases), the estimates of $\boldsymbol{\beta_1}$ should get closer and closer to 0.5**

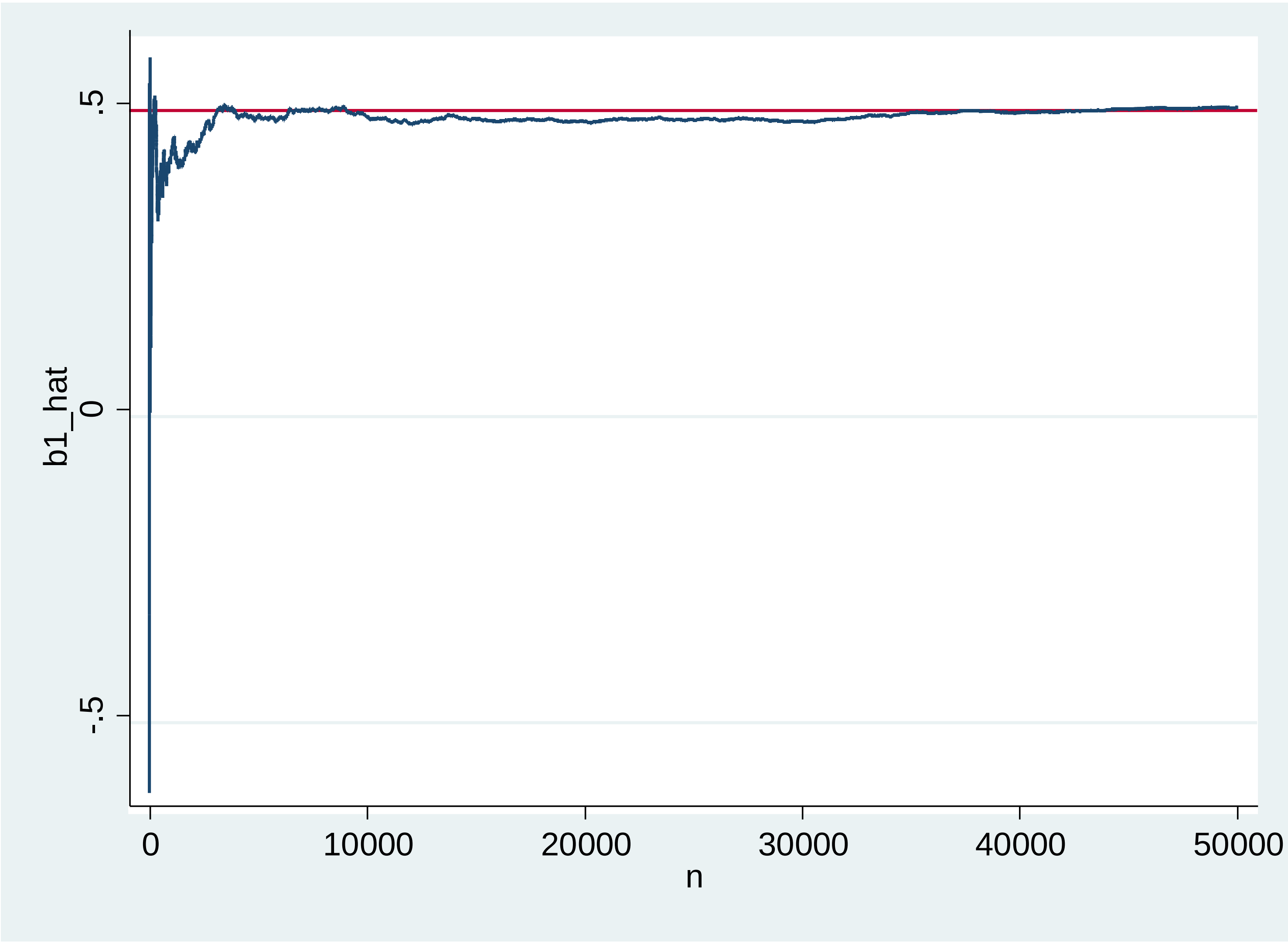
n=1, 2, ..., 500



n=1, 2 ,…, 5000



n = 1, 2, …, 50,000



Conclusion: Large Sample Distribution of OLS Estimator

- **Two key results**: Under LSA#1, #2, #3, and when sample size "n" grows large

- **1. OLS estimator is consistent, i.e.** $\hat{\beta}_1 \xrightarrow{p} \beta_1$

- **2. OLS estimator is approximately distributed as a normal random variable:**

$$\hat{\beta}_1 \overset{A}{\approx} N\left(\beta_1, \boxed{\frac{Var[(X_i - \mu_X)u_i]}{nVar(X_i)^2}}\right)$$

Note: this is the heteroskedasticity-robust estimate of the sampling variance

The standard errors reported by STATA under the "regress y x, <u>robust</u>" command is an estimate of the square root of the sampling variance of the OLS estimator (i.e. term in red box)