Stata VI

Econ 5/Poli 5D Lecture 10

Announcements

- Second homework assignment due Wednesday
- Any questions?

Today's Application: Standard Errors and the Null Hypothesis

- Continue example from last class about movie reviews and box office receipts
- How can we tell if our results are **signficant**
- What does it mean to say our results are **significant**

Today's Software: Stata

- Learn how to test hypotheses and interpret regression output
- Learn new functions like `runiform` along the way

Recap from last week

We **modeled** box office reviews, Y, as a function of rotten tomatoes score, X, assuming the relationship is **linear**

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Our job today is to understand our estimate of $\hat{\beta}_1$

In [1]:

```
* Setup
cd "/Users/Brian/Dropbox/Grad School/Sixth Year/Econ:Poli 5/Lectures/Week 6"

* Load Data
use ./data/movie_ratings_rev.dta, replace
d
```

```
/Users/Brian/Dropbox/Grad School/Sixth Year/Econ:Poli 5/Lectures/Week 6



Contains data from ./data/movie_ratings_rev.dta
  obs:           127
 vars:             7                          21 Dec 2020 15:59
-------------------------------------------------------------------------
------
              storage   display    value
variable name   type    format     label      variable label
-------------------------------------------------------------------------
------
film            str70   %70s                  Title of Movie
rottentomatoes  byte    %8.0g                 Rottent Tomatoes Critic Scor
e
metacritic      byte    %8.0g                 Metacritic Score
genre           str22   %22s
box_office      double  %10.0g                Box Office Revenue in Millio
ns
tickets         double  %10.0g                Tickets Sold in Millions
subsample       float   %9.0g                 Subsample For Illustrative
                                                Purposes
-------------------------------------------------------------------------
```

In [2]:
```
reg box_office rottentomatoes
```

```
      Source |       SS           df       MS      Number of obs   =       127
-------------+----------------------------------   F(1, 125)       =      0.87
       Model |  8234.53875         1  8234.53875   Prob > F        =    0.3531
    Residual |  1184939.29       125  9479.51432   R-squared       =    0.0069
-------------+----------------------------------   Adj R-squared   =   -0.0010
       Total |  1193173.83       126  9469.63357   Root MSE        =    97.363

------------------------------------------------------------------------------
   box_office |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
rottentomat~s |   .2673486   .2868477     0.93   0.353    -.3003586    .8350558
        _cons |   38.37105   18.70257     2.05   0.042     1.356338    75.38576
------------------------------------------------------------------------------
```

```
In [2]: reg box_office rottentomatoes

      Source |       SS           df       MS      Number of obs   =       127
-------------+----------------------------------   F(1, 125)       =      0.87
       Model |  8234.53875         1  8234.53875   Prob > F        =    0.3531
    Residual |  1184939.29       125  9479.51432   R-squared       =    0.0069
-------------+----------------------------------   Adj R-squared   =   -0.0010
       Total |  1193173.83       126  9469.63357   Root MSE        =    97.363

------------------------------------------------------------------------------
   box_office |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
rottentomat~s |   .2673486   .2868477     0.93   0.353    -.3003586    .8350558
        _cons |   38.37105   18.70257     2.05   0.042     1.356338    75.38576
------------------------------------------------------------------------------
```

Interpretation If a rotten tomatoes score goes up by 1, I expect the movie to earn 0.27 million dollars more (or 270,000 more dollars)

Forming predictions

Imagine a studio executive has an idea for a new scene that will surely raise the rotten tomatoes score by 1 point

If the exec asks how much would that increase box office revenue, you could answer?

$$\text{Predicted } Y_i = \hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 \cdot \text{Rotten Tomatoes}_i$$

$$\hat{Y}_i = 38.37 + 0.27 \cdot \text{Rotten Tomatoes}_i$$

Say the RT score goes from 50 to 51. This is predicted to increase box office revenue by:

$$\underbrace{38.37 + 0.27 \cdot 51}_{\text{Predicted value at 51}} - \underbrace{(38.37 + 0.27 \cdot 50)}_{\text{Predicted value at 50}} = 0.27$$

Increasing RT by one is predicted to increase revenue by 0.27 million dollars (270,000) dollars

But are you sure?

You tell the studio executive it is predicted to increase revenue by 270,000 dollars

The studio executive asks you one more question:

- How sure are you?

How to answer this question is the topic of today's class.

The Null Hypothesis

The **null hypothesis** is our default hypothesis

Ex: There is no relationship between rotten tomatoes score and box office revenue ($\beta_1 = 0$)

To reject the null hypothesis implies that we believe there is a relationship between rotten tomatoes score and box office revenue ($\beta_1 \neq 0$)

How do we go about testing $\beta_1 \neq 0$

The Null Hypothesis

To understand the null hypothesis, let's consider an example a simulated example in which the null hypothesis is **true** by construction

We will generate a Y variable and an X variable randomly

Because they are random, we know that there is no relationship between Y and X

In [3]:

```
clear all

*set seed for replication
set seed 42
set obs 1000
gen y = runiform()
gen x = runiform()
```

```
number of observations (_N) was 0, now 1,000
```

In [3]:

```
clear all

*set seed for replication
set seed 42
set obs 1000
gen y = runiform()
gen x = runiform()
```

```
number of observations (_N) was 0, now 1,000
```

`runiform` is like the `RAND` function from Excel

It will generate a random number between 0 and 1

Because I have `set obs 1000`, the code will generate a variable `y` and a variable `x` with each entry a random number between 0 and 1

In [4]:

```
%head y x
```

	y	x
1	.7551555	.054498937
2	.63903141	.92512929
3	.75214517	.6942302
4	.13627268	.41046754
5	.90326899	.16362032
6	.094068311	.41815799
7	.5745703	.33757183
8	.3728877	.30218264
9	.2738741	.69648379
10	.39027089	.52794898

In [5]:
```
reg y x
```

```
      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =      2.31
       Model |  .178048602         1  .178048602   Prob > F        =    0.1292
    Residual |  77.0484397       998  .077202845   R-squared       =    0.0023
-------------+----------------------------------   Adj R-squared   =    0.0013
       Total |  77.2264883       999  .077303792   Root MSE        =    .27785

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.0460968   .0303542    -1.52   0.129    -.1056622    .0134685
       _cons |   .5175435   .0175948    29.41   0.000     .4830166    .5520705
------------------------------------------------------------------------------
```

```
In [5]: reg y x
```

```
      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =      2.31
       Model |  .178048602         1  .178048602   Prob > F        =    0.1292
    Residual |  77.0484397       998  .077202845   R-squared       =    0.0023
-------------+----------------------------------   Adj R-squared   =    0.0013
       Total |  77.2264883       999  .077303792   Root MSE        =    .27785

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |  -.0460968   .0303542    -1.52   0.129    -.1056622    .0134685
       _cons |   .5175435   .0175948    29.41   0.000     .4830166    .5520705
------------------------------------------------------------------------------
```

Just by random chance, we found a negative relationship between y and x

If we draw more random numbers we will probably get a different relationship

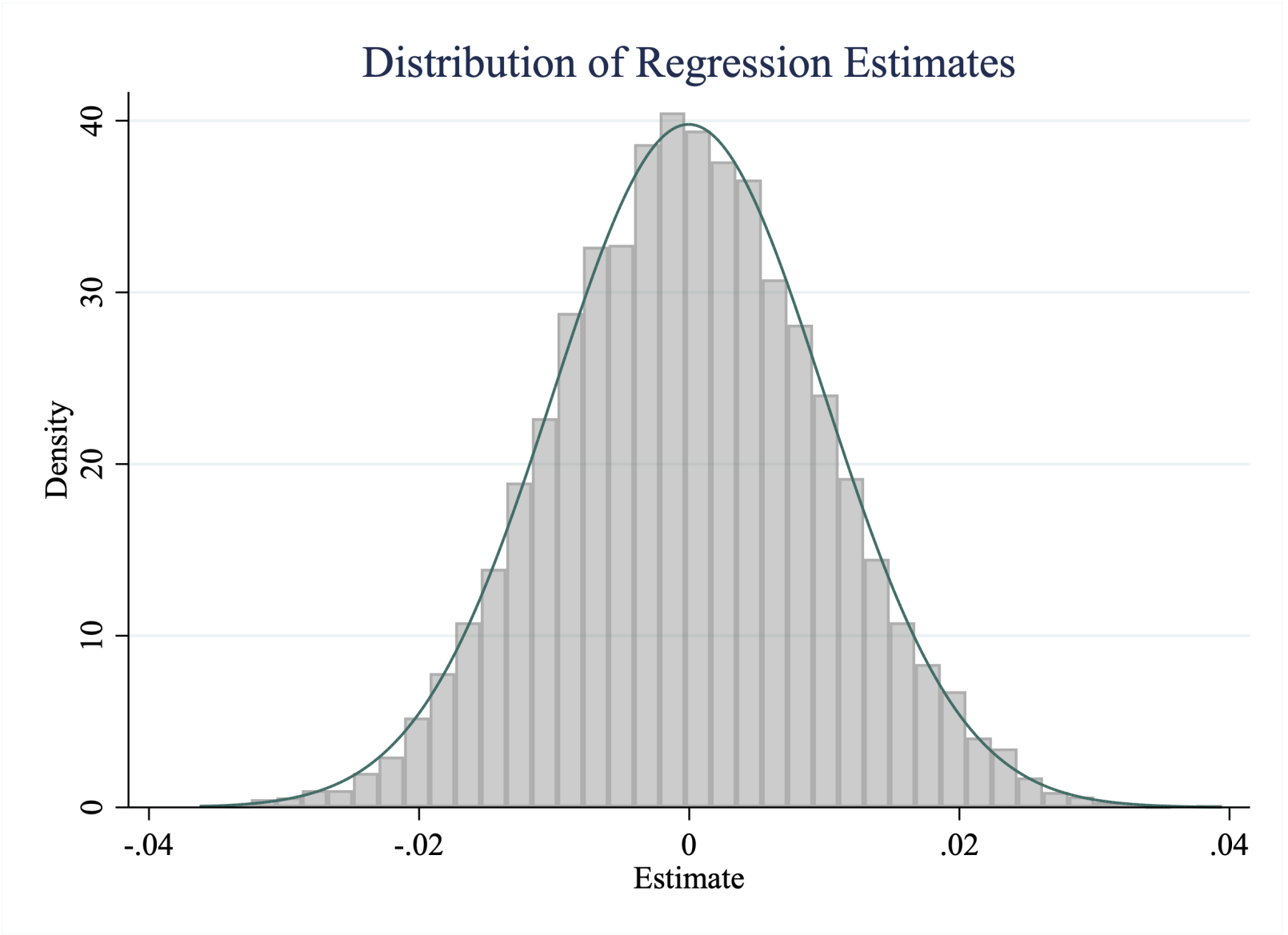
In [6]:
```
*If we draw new random number, we may very well get a different result
replace y = runiform()
replace x = runiform()

reg y x
```

```
(1,000 real changes made)

(1,000 real changes made)


      Source |       SS           df       MS      Number of obs   =     1,000
-------------+----------------------------------   F(1, 998)       =      0.39
       Model |  .032313851         1  .032313851   Prob > F        =    0.5309
    Residual |  82.0703835       998  .082234853   R-squared       =    0.0004
-------------+----------------------------------   Adj R-squared   =   -0.0006
       Total |  82.1026973       999  .082184882   Root MSE        =    .28677

------------------------------------------------------------------------------
           y |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
           x |   -.020153   .0321494    -0.63   0.531    -.0832411    .0429352
       _cons |   .5271859   .0183085    28.79   0.000     .4912585    .5631134
------------------------------------------------------------------------------
```

We can do this over and over again

Each time we will get a slightly different estimate

I did this 10,000 times and plotted the distribution of the estimates

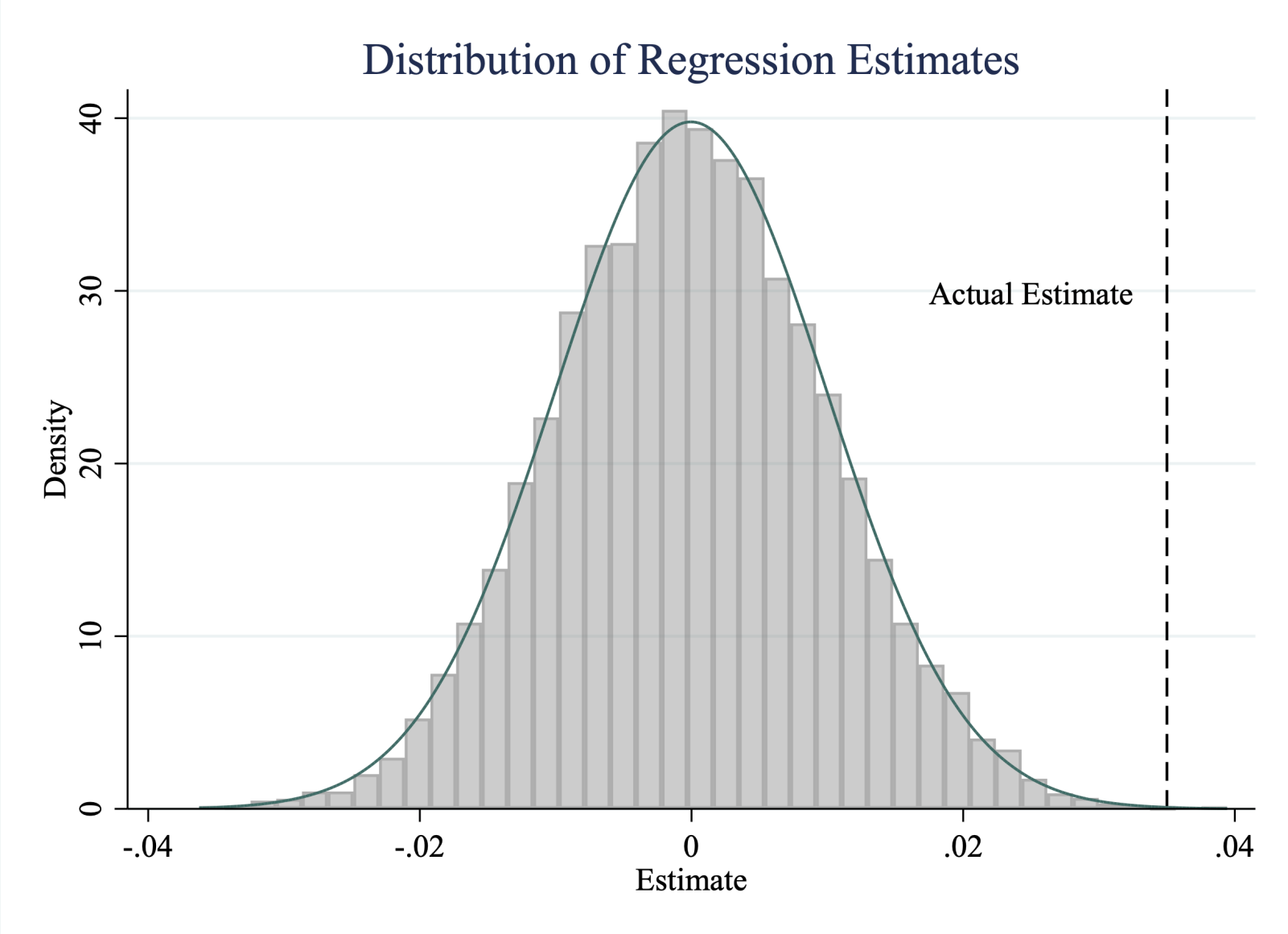

Distribution of Regression Estimates
Density
0
10
20
30
40
-.04
-.02
0
.02
.04
Estimate

Now imagine I have access to real data and I find a coefficient of 0.035

Is this positive estimate likely to be due to chance?

Now imagine I have access to real data and I find a coefficient of 0.035

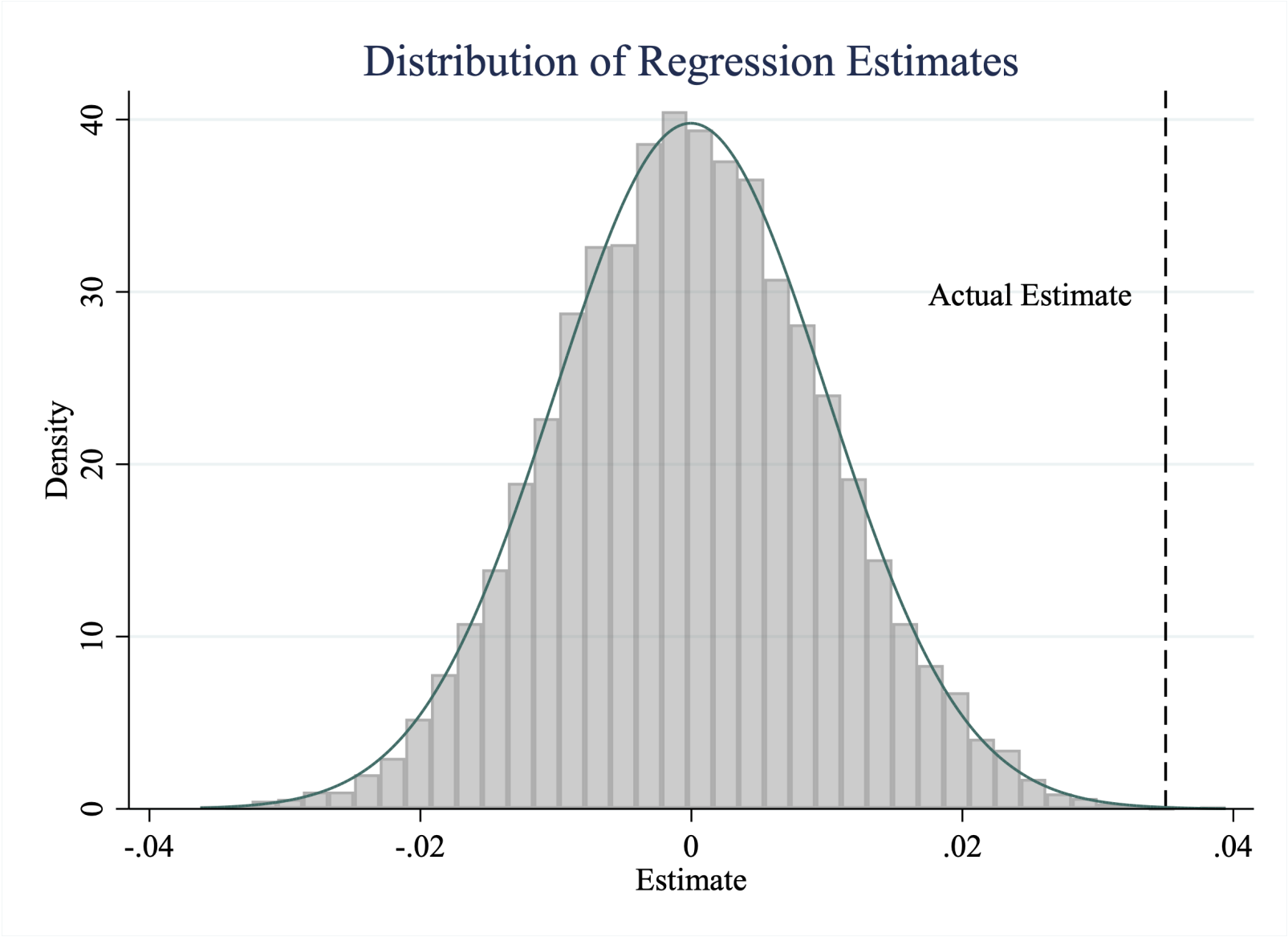
Is this positive estimate likely to be due to chance?



The value of 0.035 for the coefficient is likely significant

Very few of the simulated draws found a value that extreme

Therefore, if the null hypothesis were true it would be very unlikely we would get a value of 0.035

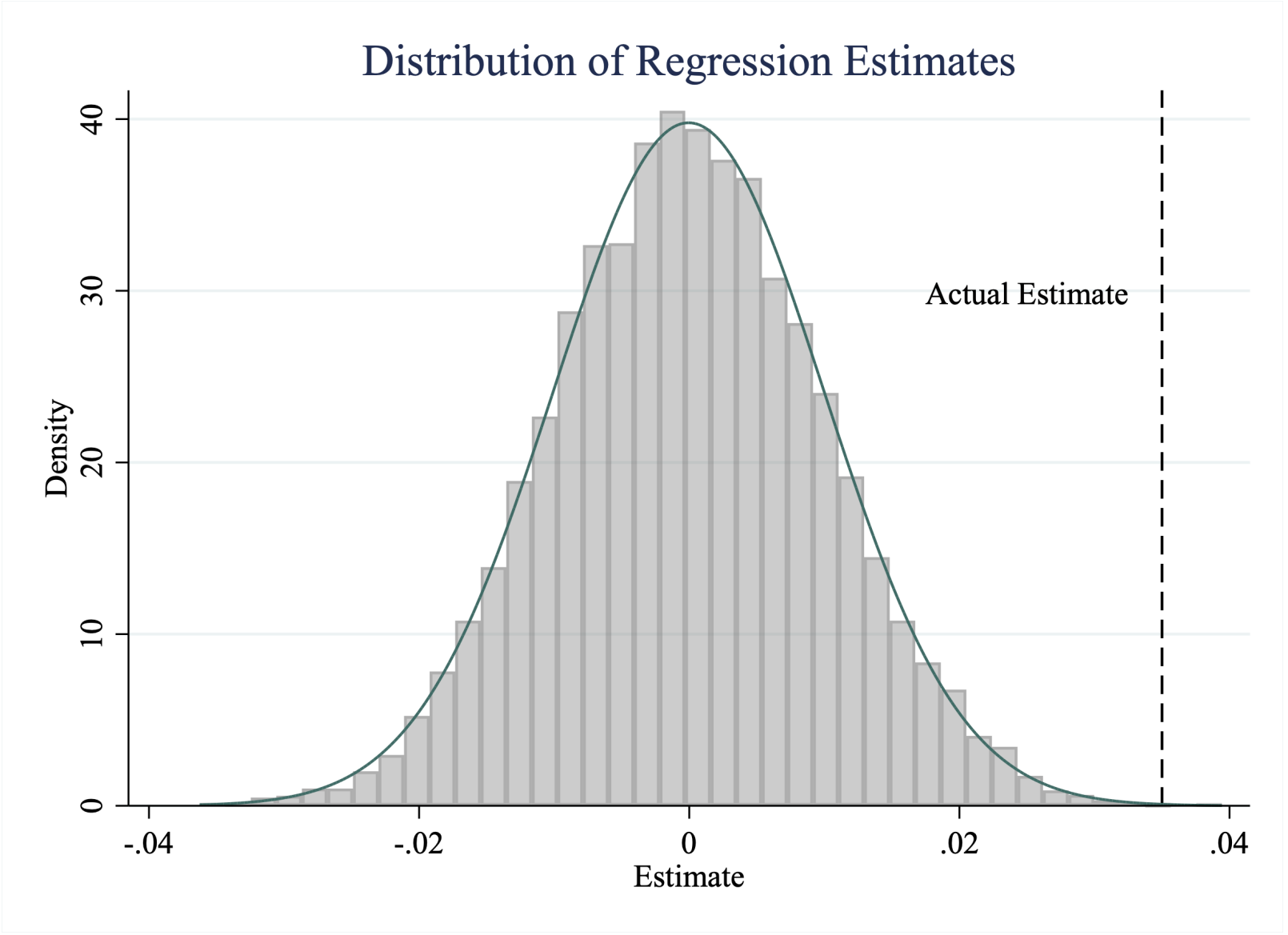

Distribution of Regression Estimates
Density
0
10
20
30
40
Actual Estimate
-.04
-.02
0
.02
.04
Estimate

How unlikely would it be?

In this example only 4 out of 10,000 were as extreme as 0.035

Therefore, the probability of obtaining this result under the null hypothesis, is about 0.4 percent

This probability is referred to as the **p-value**


Distribution of Regression Estimates
Density
0
10
20
30
40
Actual Estimate
-.04
-.02
0
.02
.04
Estimate

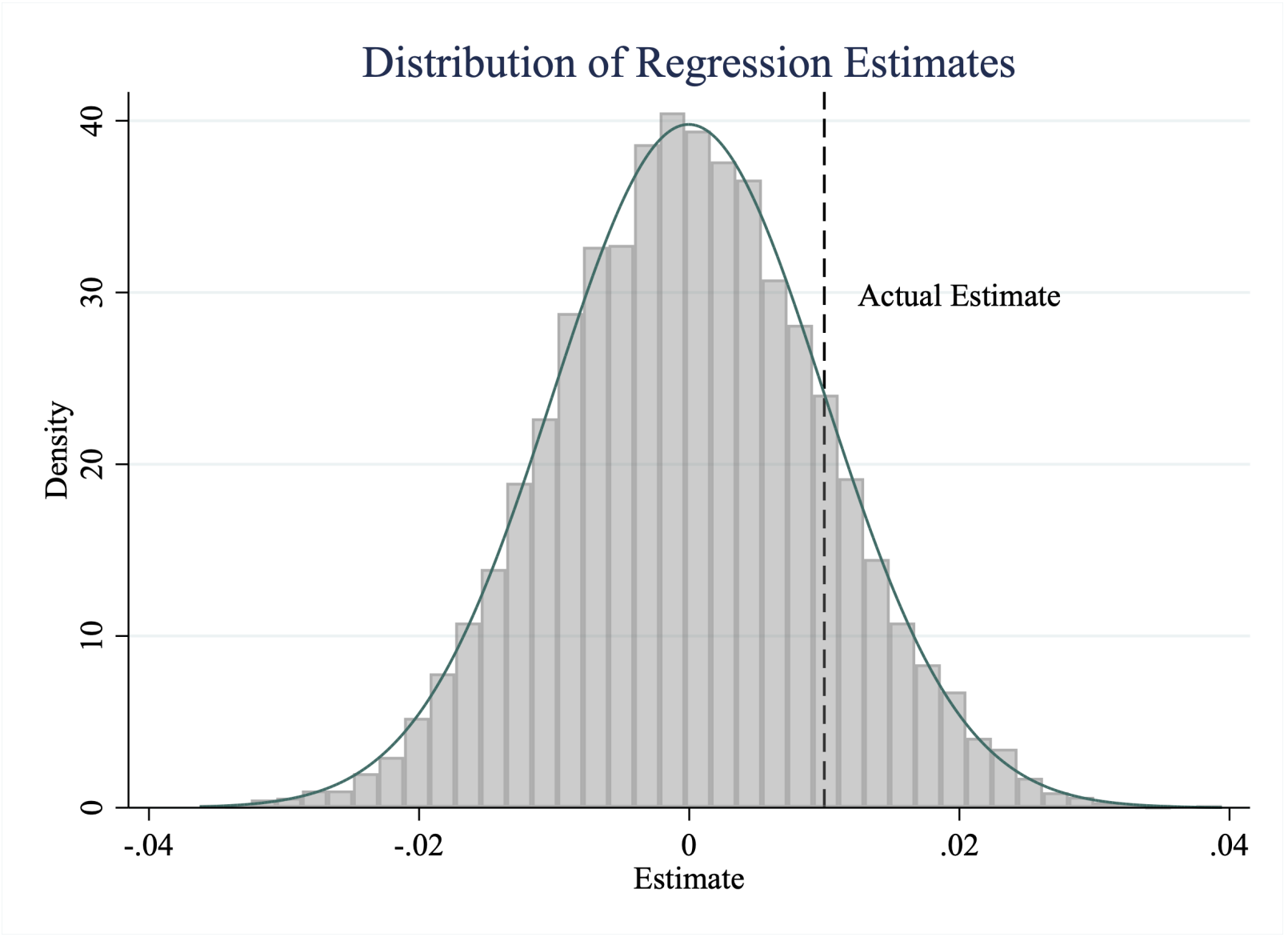
Now consider a different estimate equal to 0.01

Values this extreme happen quite frequently under the null hypothesis

Therefore this could be due to random chance

Therefore we **fail to reject the null hypothesis**

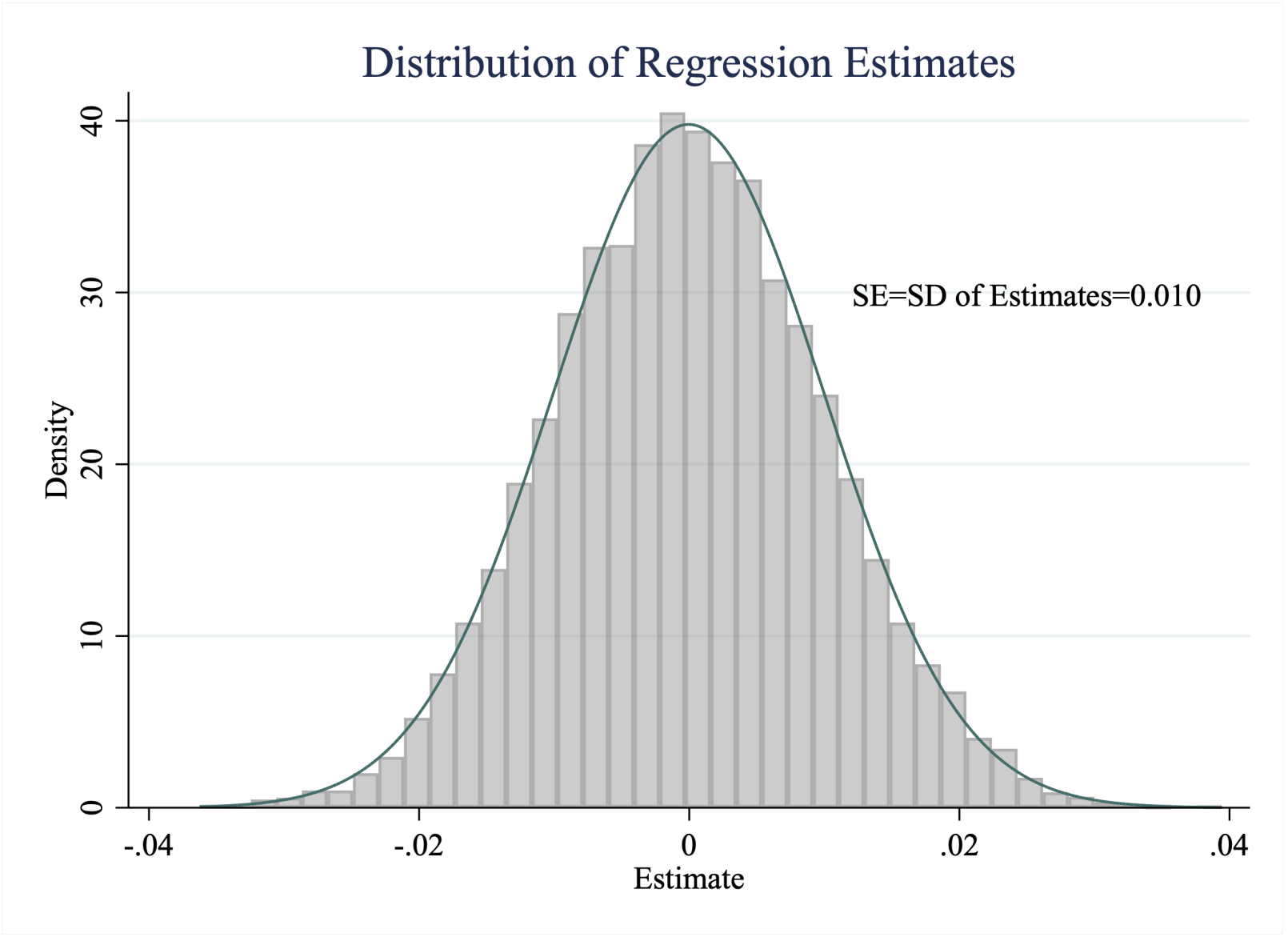
This is often referred to as **the estimate is not signficant**


Distribution of Regression Estimates
Density
0
10
20
30
40
Actual Estimate
-.04
-.02
0
.02
.04
Estimate

The **standard error** (SE) of the estimate is the standard deviation of the sampling distribution

If SE is very small, then even slight deviations from zero will be signficant

If SE is large, then we need very large effects to conclude that the relationship is signficant


Distribution of Regression Estimates
SE=SD of Estimates=0.010
Density
0
10
20
30
40
-.04
-.02
0
.02
.04
Estimate

95 percent confidence interval

95 percent confidence interval for an estimate is equal to:

$$[\hat{\beta}_1 - 1.96 \times SE, \hat{\beta}_1 + 1.96 \times SE]$$

Exact Interpretation (subtle): Were we to repeat this process many, many times, the fraction of confidence intervals that contain the true parameter β_1 would be 95 percent

Loose Interpretation: We are reasonably confident that the true parameter falls within this interval

The magic of statistics

We simulated data to give you intuition about the null hypothesis, p-values and significance

But in practice, the distribution of β under the null hypothesis is known **theoretically**

- It is not a coincidence that our simulated estimates resembled a normal curve
- The distribution of β is normal under the null hypothesis
- This is a mathematical theorem

What this means for you is that Stata automatically calculates all these statistics every time you run a regression

Let's go back to the regression and see if our understanding has changed

In [7]:
```
use ./data/movie_ratings_rev.dta, replace
reg box_office rottentomatoes
```

```
      Source |       SS           df       MS      Number of obs   =       127
-------------+----------------------------------   F(1, 125)       =      0.87
       Model |  8234.53875         1  8234.53875   Prob > F        =    0.3531
    Residual |  1184939.29       125  9479.51432   R-squared       =    0.0069
-------------+----------------------------------   Adj R-squared   =   -0.0010
       Total |  1193173.83       126  9469.63357   Root MSE        =    97.363

------------------------------------------------------------------------------
   box_office |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
rottentomat~s |   .2673486   .2868477     0.93   0.353    -.3003586    .8350558
        _cons |   38.37105   18.70257     2.05   0.042     1.356338    75.38576
------------------------------------------------------------------------------
```

Now if the studio executive asks you how sure you are that a 1 unit increase in rotten tomatoes score increases box office by 267,000 dollars you can give a very precise answer

Answer

- The standard error of the estimate is 0.287 (or 287,000 dollars)
- I cannot reject the null hypothesis that there is no relationship between box office revenue and rotten tomatoes score
- The 95 percent confidence interval ranges between [-300,000,835,000]

This may not be true for all types of movies

Let's look at the genre's we have in our dataset

This may not be true for all types of movies

Let's look at the genre's we have in our dataset

In [8]:
```
tab genre
```

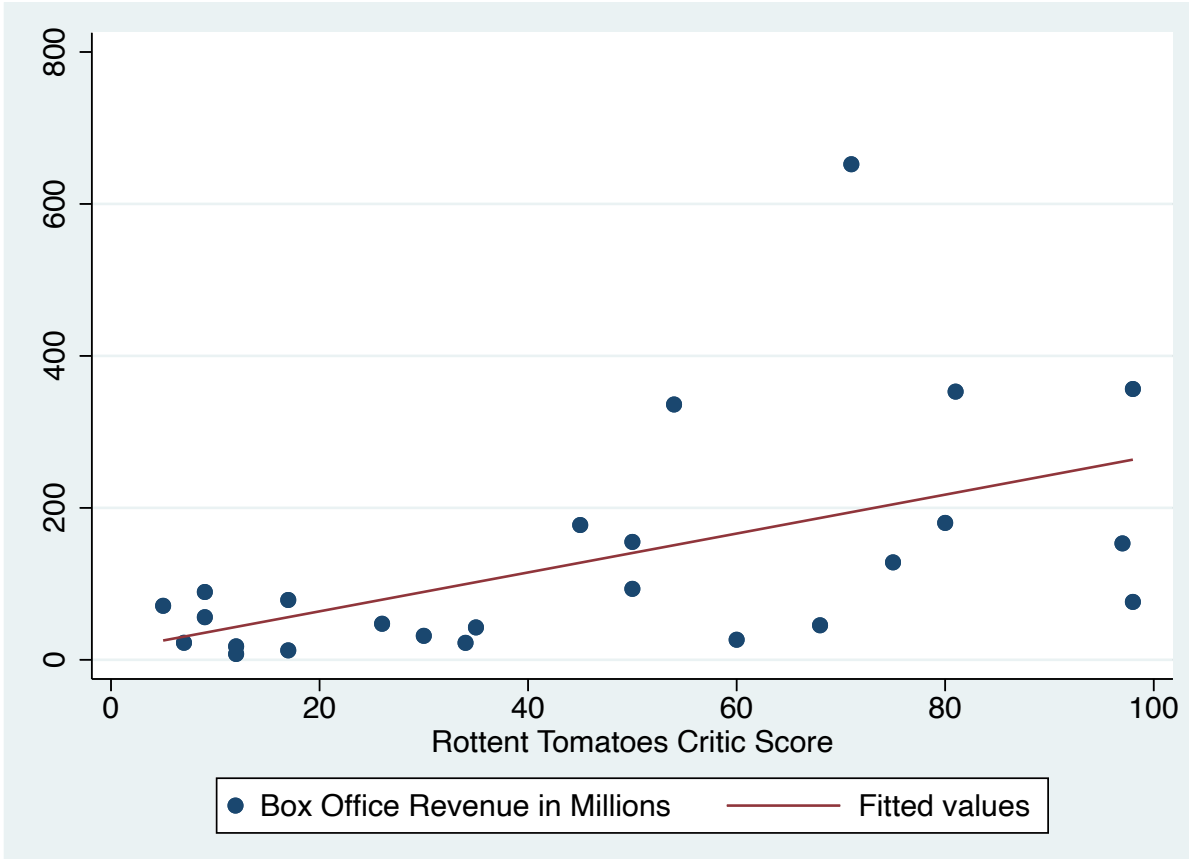
```
                genre |      Freq.     Percent        Cum.
----------------------+-----------------------------------
       ActionAdventure |         25       22.32       22.32
               Comedy |         25       22.32       44.64
          Documentary |          8        7.14       51.79
                Drama |         38       33.93       85.71
        HorrorThriller |         16       14.29      100.00
----------------------+-----------------------------------
                Total |        112      100.00
```

In [9]:
```
reg box_office rottentomatoes if genre=="ActionAdventure"
```

```
      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(1, 23)        =      9.02
       Model |  153910.013         1  153910.013   Prob > F        =    0.0063
    Residual |   392238.19        23  17053.8344   R-squared       =    0.2818
-------------+----------------------------------   Adj R-squared   =    0.2506
       Total |  546148.203        24  22756.1751   Root MSE        =    130.59

------------------------------------------------------------------------------
  box_office |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rottentomat~s |   2.560044   .8521675     3.00   0.006      .797201    4.322887
        _cons |   12.59101   46.82054     0.27   0.790    -84.26465    109.4467
------------------------------------------------------------------------------
```

In [17]:

```
twoway (scatter box_office rottentomatoes if genre == "ActionAdventure") ///
(lfit box_office rottentomatoes if genre == "ActionAdventure")
```

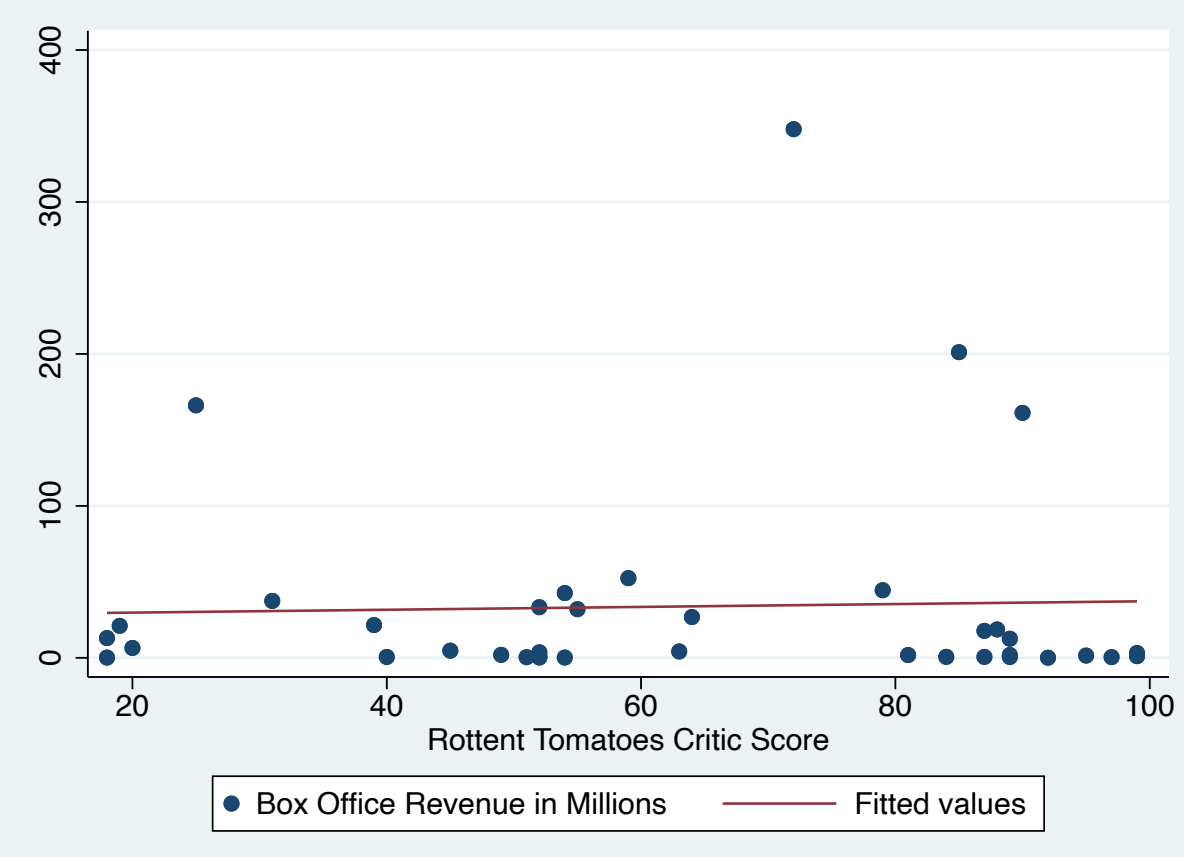


In [10]:
```
reg box_office rottentomatoes if genre=="Drama"
```

```
      Source |       SS           df       MS      Number of obs   =        38
-------------+----------------------------------   F(1, 36)        =      0.04
       Model |  216.266225         1  216.266225   Prob > F        =    0.8383
    Residual |  184212.212        36  5117.00589   R-squared       =    0.0012
-------------+----------------------------------   Adj R-squared   =   -0.0266
       Total |  184428.478        37  4984.55347   Root MSE        =    71.533

------------------------------------------------------------------------------
  box_office |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
rottentomat~s |    .093912    .456809     0.21   0.838    -.8325396    1.020364
        _cons |   27.86077   31.25353     0.89   0.379    -35.52433    91.24587
------------------------------------------------------------------------------
```

In [20]:

```
twoway (scatter box_office rottentomatoes if genre == "Drama") ///
(lfit box_office rottentomatoes if genre == "Drama")
```

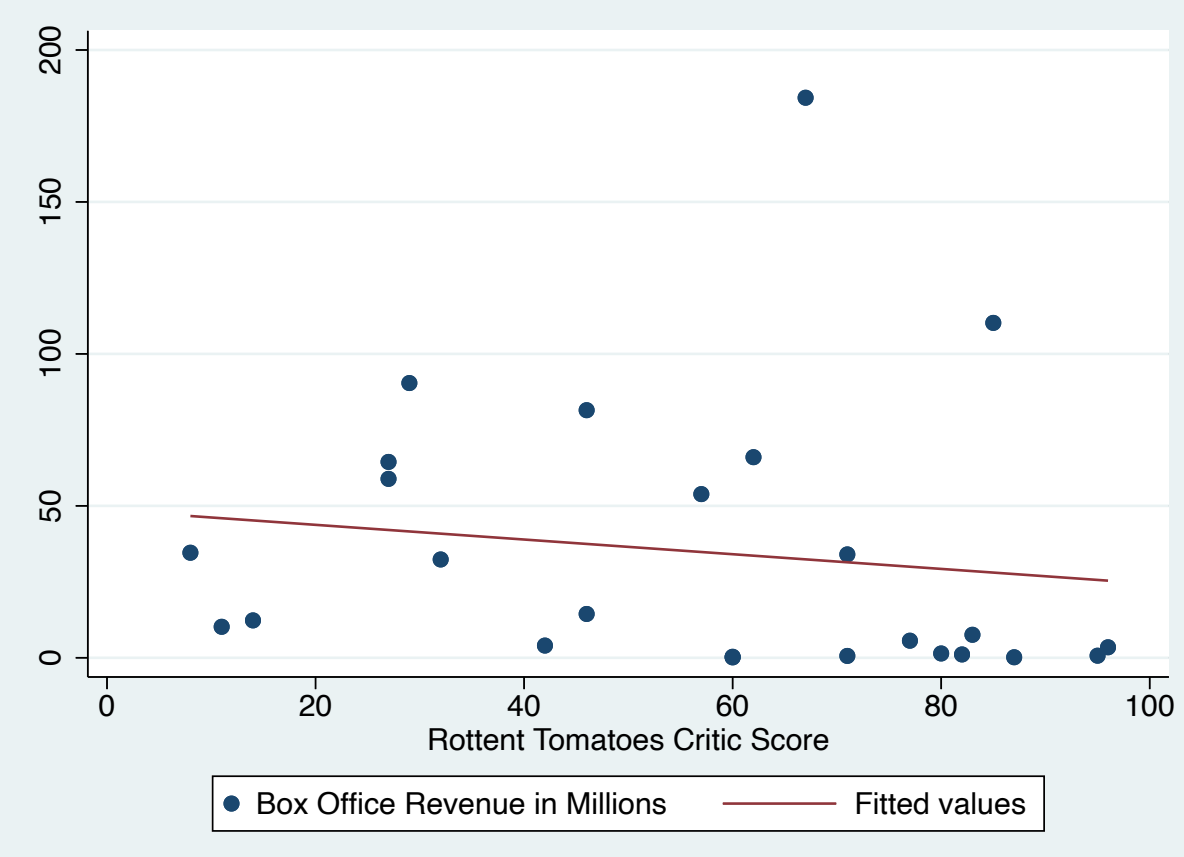


In [11]:
```
reg box_office rottentomatoes if genre=="Comedy"
```

```
      Source |       SS           df       MS      Number of obs   =        25
-------------+----------------------------------   F(1, 23)        =      0.49
       Model |  1019.29426         1  1019.29426   Prob > F        =    0.4930
    Residual |  48300.7387        23  2100.03212   R-squared       =    0.0207
-------------+----------------------------------   Adj R-squared   =   -0.0219
       Total |   49320.033        24  2055.00137   Root MSE        =    45.826

------------------------------------------------------------------------------
   box_office |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
rottentomat~s |  -.2416725   .3468889    -0.70   0.493    -.9592668    .4759218
        _cons |   48.59209   21.66776     2.24   0.035     3.768915    93.41526
------------------------------------------------------------------------------
```

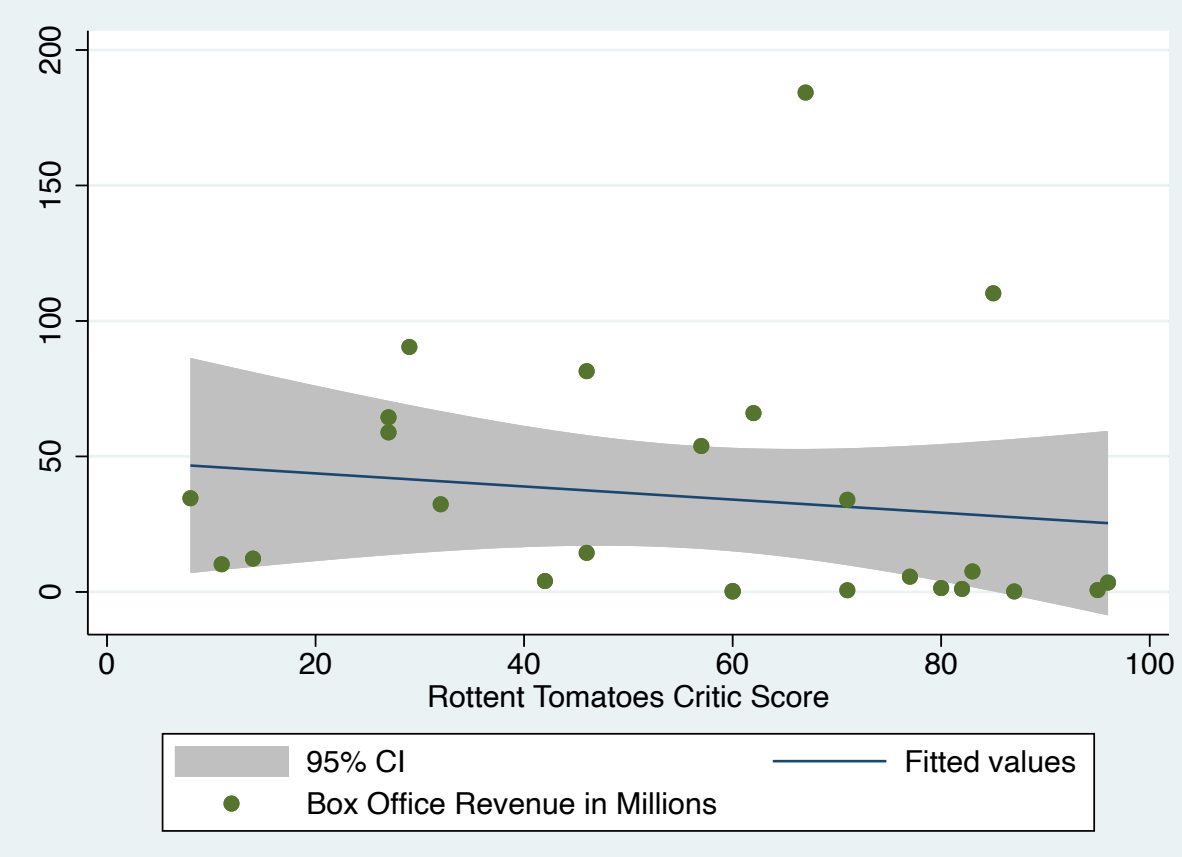
In [23]:

```
twoway (scatter box_office rottentomatoes if genre == "Comedy") ///
(lfit box_office rottentomatoes if genre == "Comedy")
```



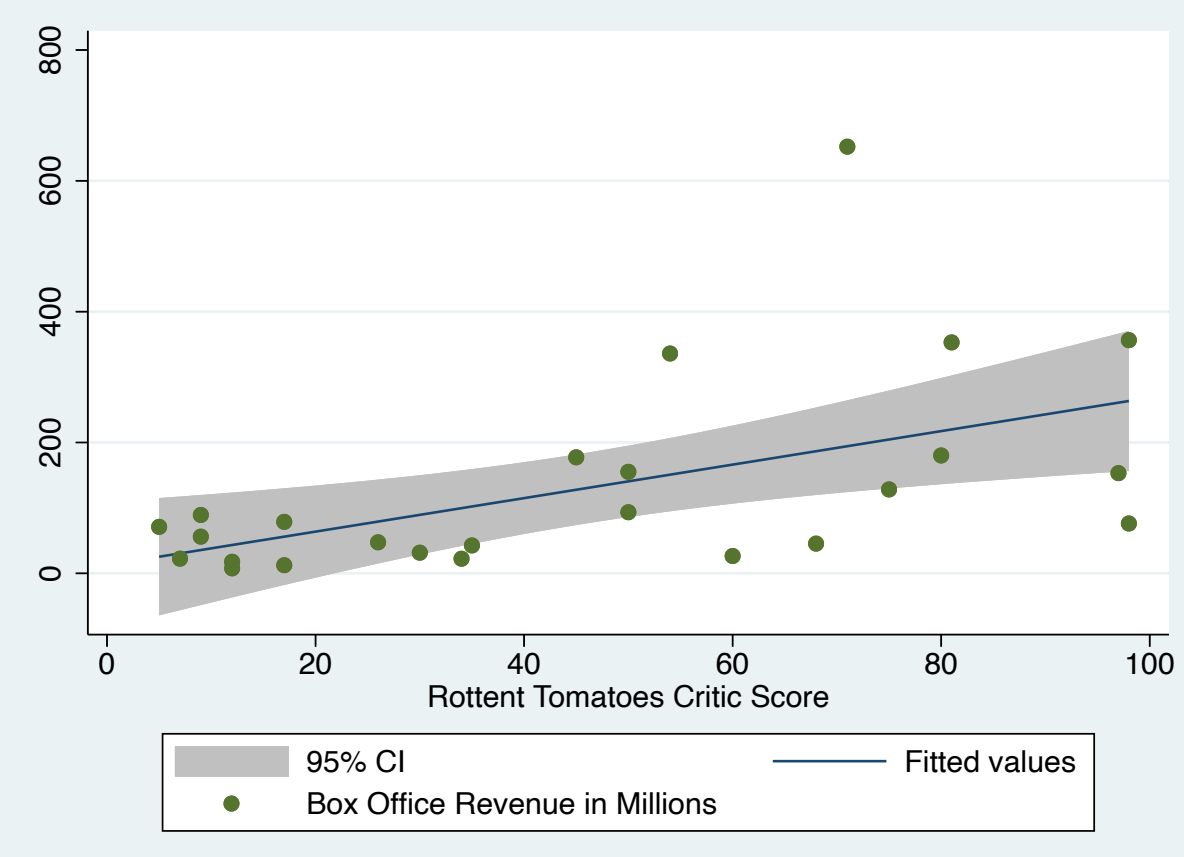
In [22]:

```
twoway (lfitci box_office rottentomatoes if genre == "Comedy") ///
(scatter box_office rottentomatoes if genre == "Comedy")
```



In [25]:

```
twoway (lfitci box_office rottentomatoes if genre == "ActionAdventure") ///
(scatter box_office rottentomatoes if genre == "ActionAdventure")
```



Stata will store coefficient and standard error estimates for you

`_b[varname]` stores the coefficient

`_se[varname]` stores the standard error

so imagine I want to construct the confidence interval myself

One thing I could do is:

Stata will store coefficient and standard error estimates for you

`_b[varname]` stores the coefficient

`_se[varname]` stores the standard error

so imagine I want to construct the confidence interval myself

One thing I could do is:

In [12]:
```
* best fit line
gen beta = _b[rottentomatoes] if _n==1
* lower bound of confidence interval
gen lb_ci = _b[rottentomatoes]-1.96*_se[rottentomatoes] if _n==1
* upper bound of confidence interval
gen ub_ci = _b[rottentomatoes]+1.96*_se[rottentomatoes] if _n==1
```

```
(126 missing values generated)

(126 missing values generated)

(126 missing values generated)
```

In [13]:
```
%head beta ub_ci lb_ci
```

	beta	ub_ci	lb_ci
1	-.24167247	.43822971	-.92157465
2	.	.	.
3	.	.	.
4	.	.	.
5	.	.	.
6	.	.	.
7	.	.	.
8	.	.	.
9	.	.	.
10	.	.	.

Today's Application: Standard Errors and the Null Hypothesis

- We are often interested in not only estimating a relationship, but understanding how certain we are about our estimates
- Higher rotten tomatoes scores are associated with higher box office revenue, but the results are **not significant**

Today's Software: Stata

- Learn how to test hypotheses and interpret regression output
- Learn new functions like `runiform` along the way

Final Projects

- Assignment description and examples posted on Canvas
- Link to suggested datasets, you can use some other data but check with me first
- Ask an interesting question, come up with an argument and hypotheses, present descriptive and visual evidence in favor or opposition to your argument
- Should submit a final report and any associated code
- Should use tools from class, but no expectation that you will use R vs. Stata vs. Excel
- Any questions?