THEORY AND MODEL ASSESSMENT THROUGH SIMULATION

Dr. Aric LaBarr
Institute for Advanced Analytics

THEORY ASSESSMENT

Central Limit Theorem

Closed Form Solutions?

- In mathematics and statistics, there are popular theories involving distributions of known values.
- The Central Limit Theorem is a classic example.
- Don’t need complicated mathematics for us to approximate distributional assumptions when we use simulations.

Closed Form Solutions?

- This is especially helpful when finding a **closed form solution** is very difficult if not impossible.
- A closed form solution to a mathematical/statistical distribution problem means that you can mathematically calculate the distribution.
- Real world data can be very complicated and changing based on many different inputs which each have their own distribution.
- Simulation can reveal an approximation of these output distributions.

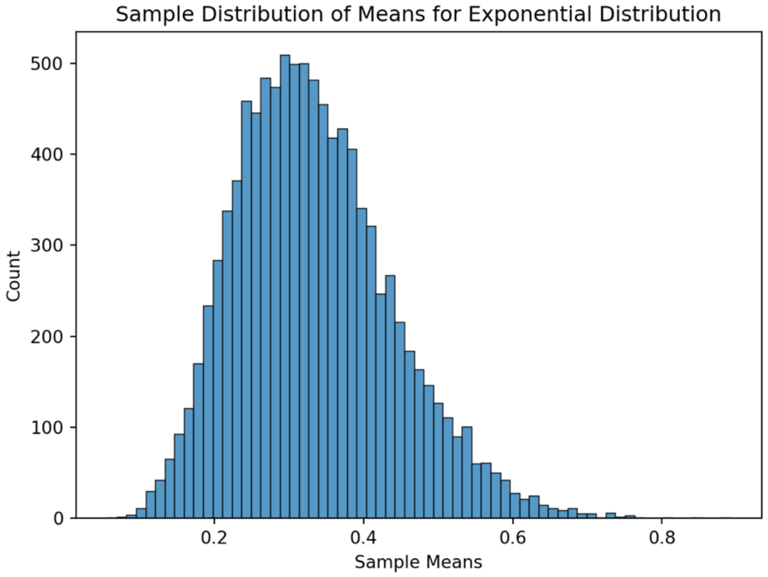
Example – Central Limit Theorem

- Assume you do not know the Central Limit Theorem, but you want to understand the sampling distribution of sample means.
- You take samples of size 10, 50, and 100 from the following three population distributions and calculate the sample means:
  1. Normal Distribution
  2. Uniform Distribution
  3. Exponential Distribution
- What is the sampling distribution of sample means from each of these distributions and sample sizes?

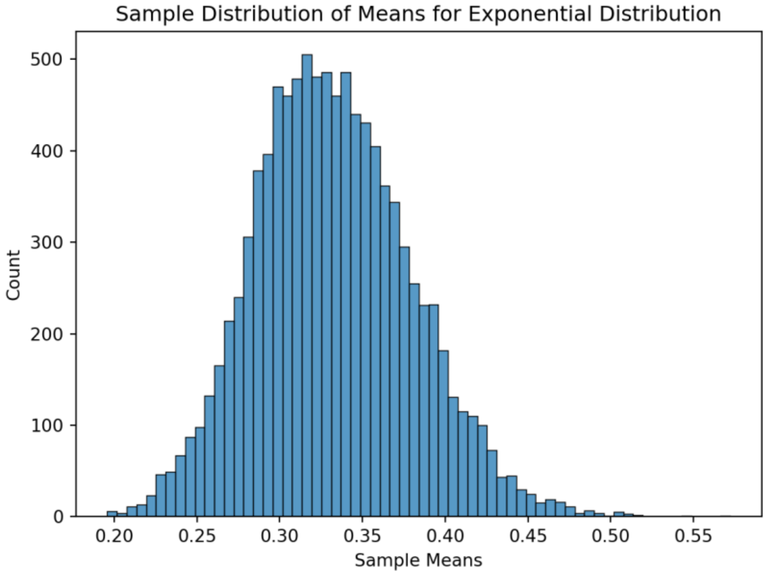
Theory Assessment for CLT – Python

```
import numpy as np

sample_size = 50
simulation_size = 10000

X1 = np.random.normal(loc = 0.04, scale = 0.07, size =
sample_size*simulation_size).reshape(simulation_size, sample_size)
X2 = np.random.uniform(low = 5, high = 105, size =
sample_size*simulation_size).reshape(simulation_size, sample_size)
X3 = np.random.exponential(scale = 0.333, size =
sample_size*simulation_size).reshape(simulation_size, sample_size)

Mean_X1 = X1.mean(axis = 1)
Mean_X2 = X2.mean(axis = 1)
Mean_X3 = X3.mean(axis = 1)
```

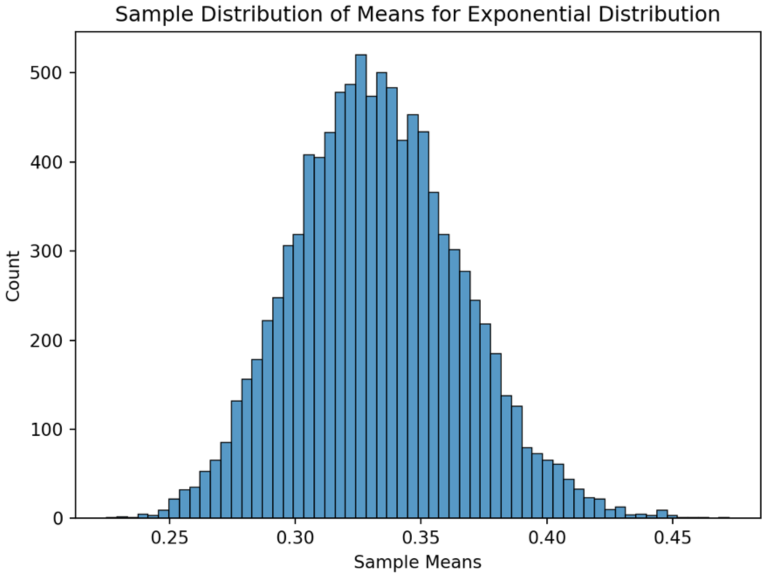
Assessment for CLT – Python



n = 10



n = 50



n = 100

THEORY ASSESSMENT

Omitted Variable Bias

Example – Omitted Variable Bias

- What if you leave out a variable in a linear regression that should have been in the model?
- From the primer we learned that it would change the variance and bias of the coefficients still in the model **depending** on if the variable left out was correlated.
- What if you wanted to know **how bad it could get**?

Example – Omitted Variable Bias

- Build the following regression model:

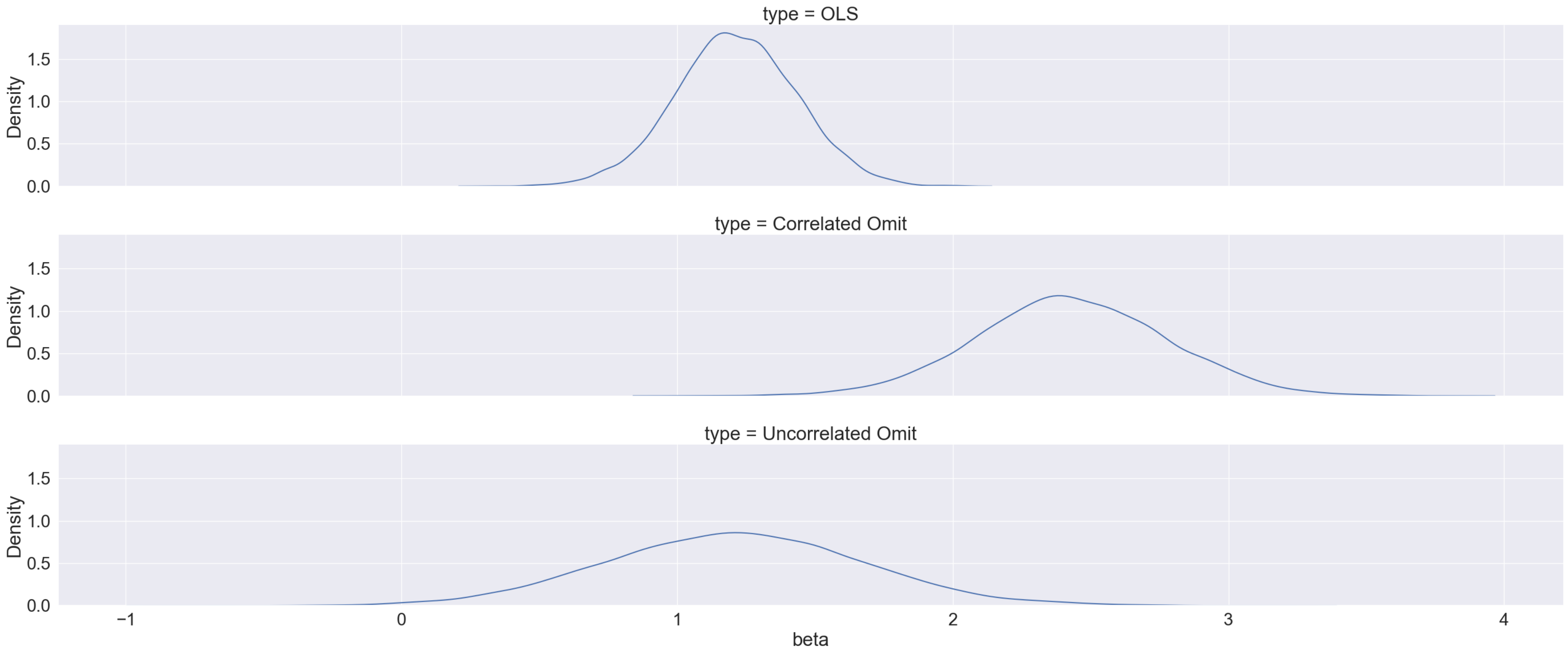
$$Y = -13 + 1.21X_1 + 3.45X_2 + \varepsilon$$

- Assume the errors are normally distributed with mean of 0 and standard deviation of 1.5.
- Assume the predictors follow standard normal distributions.

Example – Omitted Variable Bias

- Build 10,000 linear regressions (each of sample size 50) and record the coefficients from the regression model when one of the variables is omitted. Look at the following:
  1. Distribution of coefficient in the model
     - What if the omitted variable isn’t correlated with the others?
     - What if the omitted variable is correlated with the others?

Example – Omitted Variable Bias



Example – Omitted Variable Bias

- Build 10,000 linear regressions (each of sample size 50) and record the coefficients from the regression model when one of the variables is omitted. Look at the following:
  1. Distribution of coefficient in the model
     - What if the omitted variable isn't correlated with the others? **UNBIASED, MORE VARIANCE**
     - What if the omitted variable is correlated with the others? **BIASED, MORE VARIANCE**

Example – Omitted Variable Bias

- Build 10,000 linear regressions (each of sample size 50) and record the coefficients from the regression model when one of the variables is omitted. Look at the following:
  2. How many times did you incorrectly NOT reject the null hypothesis on the coefficient in each of these scenarios?

Example – Omitted Variable Bias

- Build 10,000 linear regressions (each of sample size 50) and record the coefficients from the regression model when one of the variables is omitted. Look at the following:

  2. How many times did you incorrectly NOT reject the null hypothesis on the coefficient in each of these scenarios?

Model	Percentage of Time NOT Rejecting Null
Correct Model – OLS	1.18%
Correlated X2 Not in Model	0.00%
Uncorrelated X2 Not in Model	41.33%

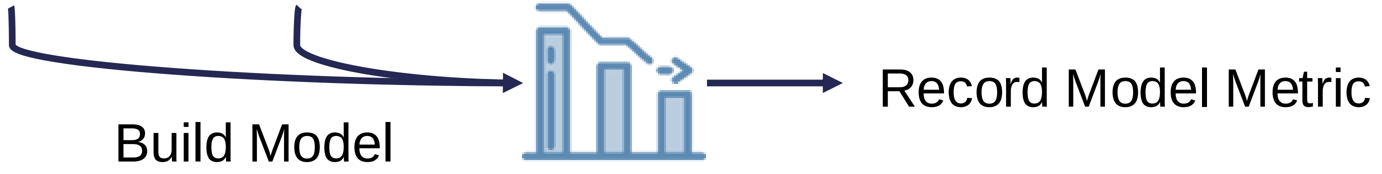
TARGET SHUFFLING

Target Shuffling

- Target shuffling has been around for a long time but has recently been brought back into popularity by John Elder.
- **Target shuffling** is when you randomly reorder the target variable values among the sample, while keeping the predictor variable values fixed.

Target Shuffling

Age	Loyalty Program	Buy Product?			
25	Y	1			
31	N	0			
28	N	1			
42	Y	0			
39	Y	1			
…	…				
34	N	0			

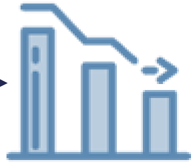


Target Shuffling

Age	Loyalty Program	Buy Product?	Y_1		
25	Y	1	0		
31	N	0	1		
28	N	1	1		
42	Y	0	0		
39	Y	1	0		
…	…				
34	N	0	1		

Target Shuffling

Age	Loyalty Program	Buy Product?	Y_1		
25	Y	1	0		
31	N	0	1		
28	N	1	1		
42	Y	0	0		
39	Y	1	0		
…	…				
34	N	0	1		



Record Model Metric

Build Model

Target Shuffling

Age	Loyalty Program	Buy Product?	Y_1	Y_2	
25	Y	1	0	1	
31	N	0	1	1	
28	N	1	1	1	
42	Y	0	0	0	
39	Y	1	0	0	
…	…				
34	N	0	1	0	

Target Shuffling

Age	Loyalty Program	Buy Product?	Y_1	Y_2	…
25	Y	1	0	1	…
31	N	0	1	1	…
28	N	1	1	1	…
42	Y	0	0	0	…
39	Y	1	0	0	…
…	…				…
34	N	0	1	0	…

Target Shuffling

- Target shuffling has been around for a long time, but has recently been brought back into popularity by John Elder.
- **Target shuffling** is when you randomly reorder the target variable values among the sample, while keeping the predictor variable values fixed.
- Build model from each of these reshuffled targets and record some measurement of model success (R_A^2, AUC, MAPE, etc.)

Target Shuffling

Model metric from each model!

Age	Loyalty Program	Buy Product?	Y_1	Y_2	…
25	Y	1	0	1	…
31	N	0	1	1	…
28	N	1	1	1	…
42	Y	0	0	0	…
39	Y	1	0	0	…
…	…				…
34	N	0	1	0	…

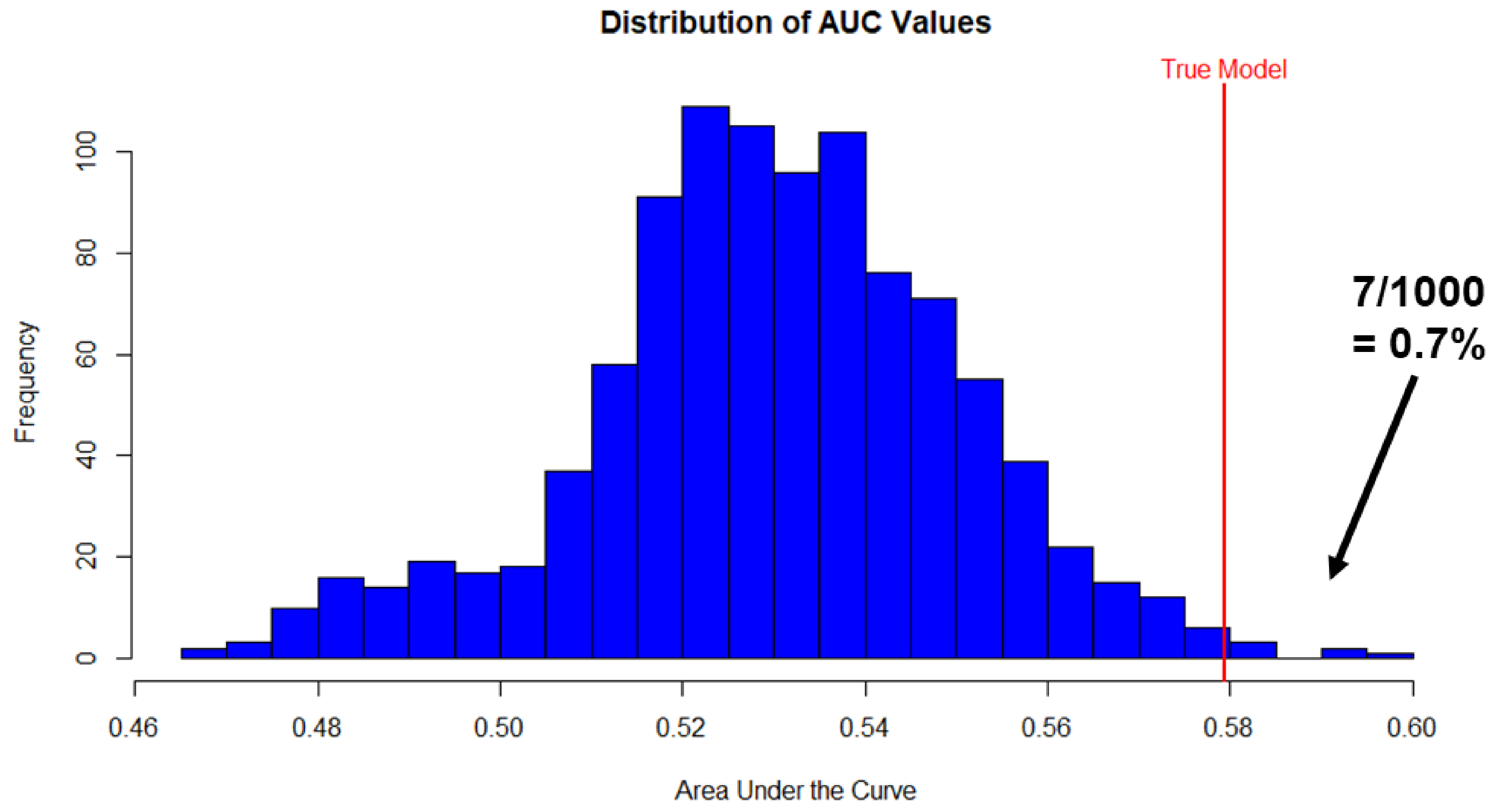
Placebo Effect

- Build model from each of these reshuffled targets and record some measurement of model success (R_A^2, AUC, MAPE, etc.)
- This should remove the pattern from the data, but **some pattern may exist due to randomness**.
- Look at distribution of all measurements of model success and find your value from the true model!

Placebo Effect

- Build model from each of these reshuffled targets and record some measurement of model success (R_A^2, AUC, MAPE, etc.)
- This should remove the pattern from the data, but **some pattern may exist due to randomness**.
- Look at distribution of all measurements of model success and find your value from the true model!
- What is probability your model would have occurred due to randomness?

Target Shuffling



Fake Data Example

- Randomly generated 8 variables that follow a Normal distribution with mean of 0 and standard deviation of 8.
- Defined relationship with target variable:

$$y = 5 + 2x_2 - 3x_8 + \varepsilon$$

Fake Data Example

- Defined relationship with target variable:

$$y = 5 + 2x_2 - 3x_8 + \varepsilon$$

- Performed target shuffle on the model.

Fake Data Example

```
np.random.seed(12345)

Fake = np.random.normal(loc = 0, scale = 1, size = 8*100).reshape(100, 8)
Fake = pd.DataFrame(Fake)
Fake.columns = ['X1', 'X2', 'X3', 'X4', 'X5', 'X6', 'X7', 'X8']

Fake['Y'] = 5 + 2*Fake['X2'] - 3*Fake['X8'] + np.random.normal(loc = 0, scale = 6,
size = 100)

sim = 1000

iteration = 1
for i in range(sim):
  data = Fake['Y']
  data = pd.DataFrame(data)
  data['uni'] = np.random.uniform(size = 100)
  data = data.sort_values(by = ['uni'])
  data = data.reset_index()

  col_name = 'Y' + str(iteration)
  Fake.loc[:, col_name] = data['Y']
  iteration = iteration + 1
```

Fake Data Example

```
import statsmodels.api as sm

X = Fake.iloc[:, range(8)]
X = sm.add_constant(X)

R_sq_A = []
for i in range(1000):
  Y = Fake.iloc[:, i + 9]
  model = sm.OLS(Y, X).fit()
  rsa = model.rsquared_adj
  R_sq_A.append(rsa)

Y = Fake.iloc[:, 8]
model = sm.OLS(Y, X).fit()
rsa = model.rsquared_adj
R_sq_A.append(rsa)

R_sq_A = pd.DataFrame(R_sq_A)
R_sq_A.columns = ['R_sq_A']
```

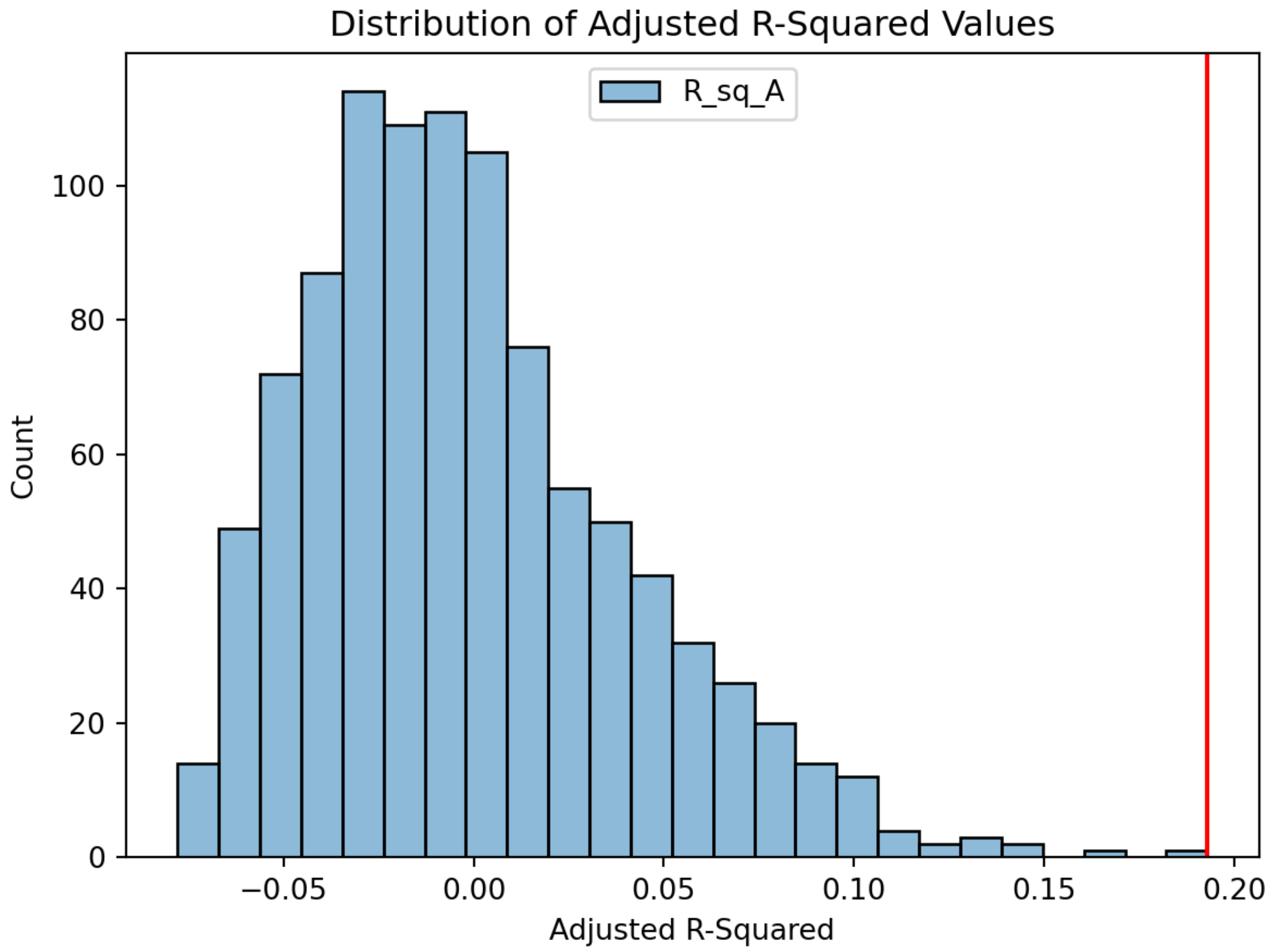
Fake Data Example

- Randomly generated 8 variables that follow a Normal distribution with mean of 0 and standard deviation of 8.
- Defined relationship with target variable:

$$y = 5 + 2x_2 - 3x_8 + \varepsilon$$

- Adjusted R^2 from this model: 0.204

Fake Data Example

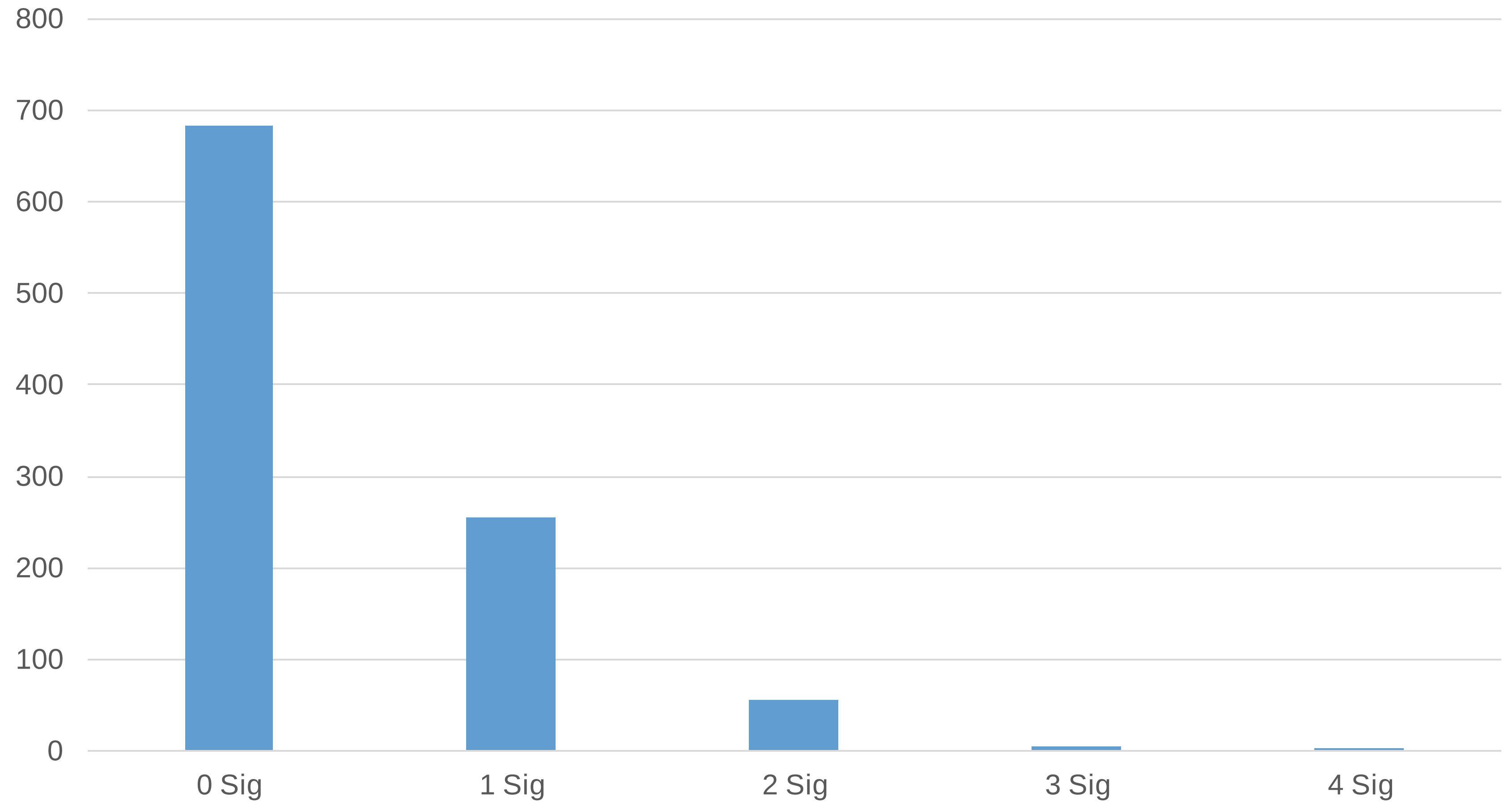


Target Shuffle with 1000 Simulations

```
Y = Fake.iloc[:, 9]
model = sm.OLS(Y, X).fit()
Pval = pd.DataFrame(model.pvalues)

for i in range(999):
  Y = Fake.iloc[:, i + 10]
  model = sm.OLS(Y, X).fit()
  Pval2 = model.pvalues
  Pval = pd.concat([Pval, Pval2], axis = 1)

Pval = Pval < 0.05

Pval.sum(axis = 1)
```

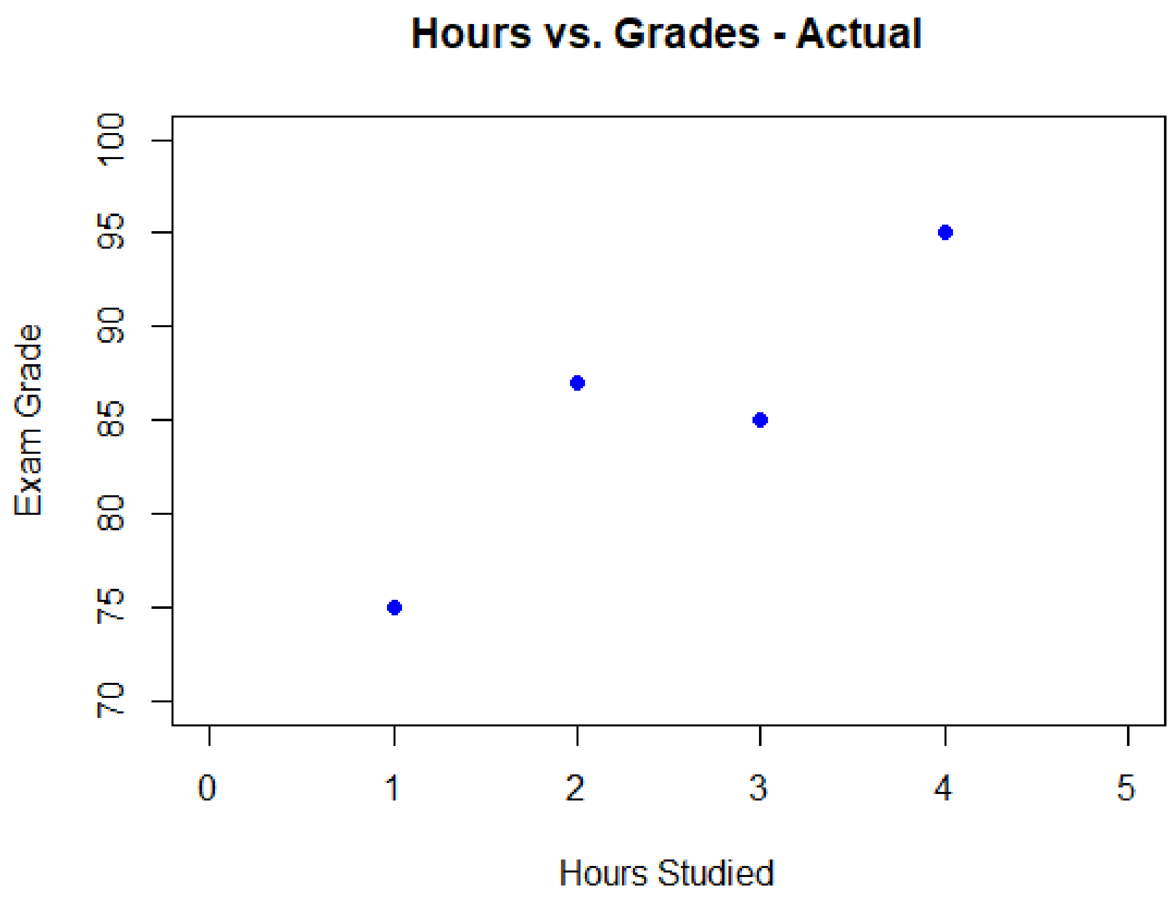
Variable	Times Appeared Significant ($p < 0.05$) in a Model
X1	47
X2	41
X3	48
X4	39
X5	51
X6	66
X7	52
X8	41

Fake Data Example – # Significant Variables

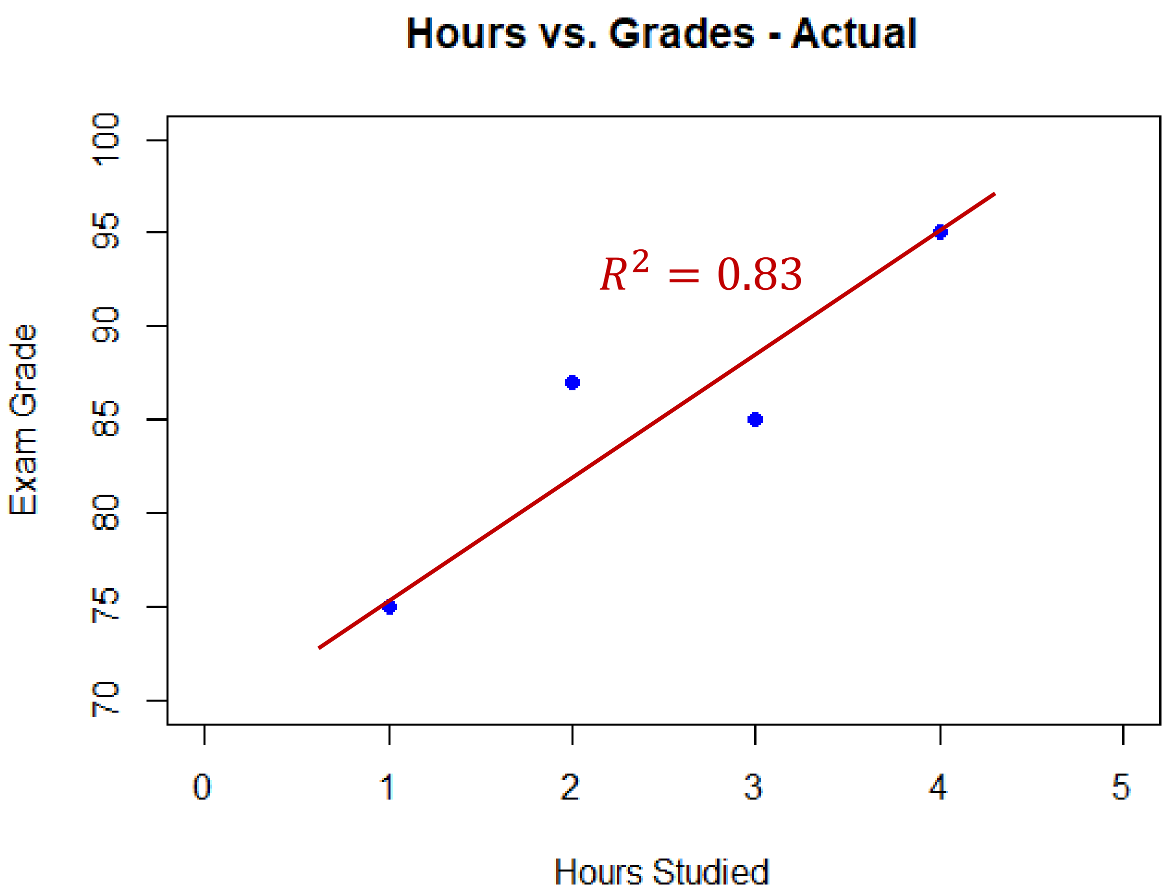


Student Grade Analogy

Student Grade Analogy



Student Grade Analogy



Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
75	85	87	95	75	95	85	87	85	87	75	95	87	75	85	95	87	95	75	85	95	85	75	87

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
75	87	85	95	75	95	87	85	85	87	95	75	87	75	95	85	87	95	75	85	95	87	75	85

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
75	85	95	87	85	75	87	95	85	95	75	87	87	85	75	95	95	75	85	87	95	85	87	75

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
75	87	95	85	85	75	95	87	85	95	87	75	87	85	95	75	95	75	87	85	95	87	85	75

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!
- There are 3 possible combinations that produce a regression with an R^2 that is greater than or equal to our actual data.

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
75	85	87	95	75	95	85	87	85	87	75	95	87	75	85	95	87	95	75	85	95	85	75	87

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
75	87	85	95	75	95	87	85	85	87	95	75	87	75	95	85	87	95	75	85	95	87	75	85

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
75	85	95	87	85	75	87	95	85	95	75	87	87	85	75	95	95	75	85	87	95	85	87	75

1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4	1	2	3	4
75	87	95	85	85	75	95	87	85	95	87	75	87	85	95	75	95	75	87	85	95	87	85	75

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

1	2	3	4
75	85	87	95

1	**2**	**3**	**4**
75	**87**	**85**	**95**

1	2	3	4
75	85	95	87

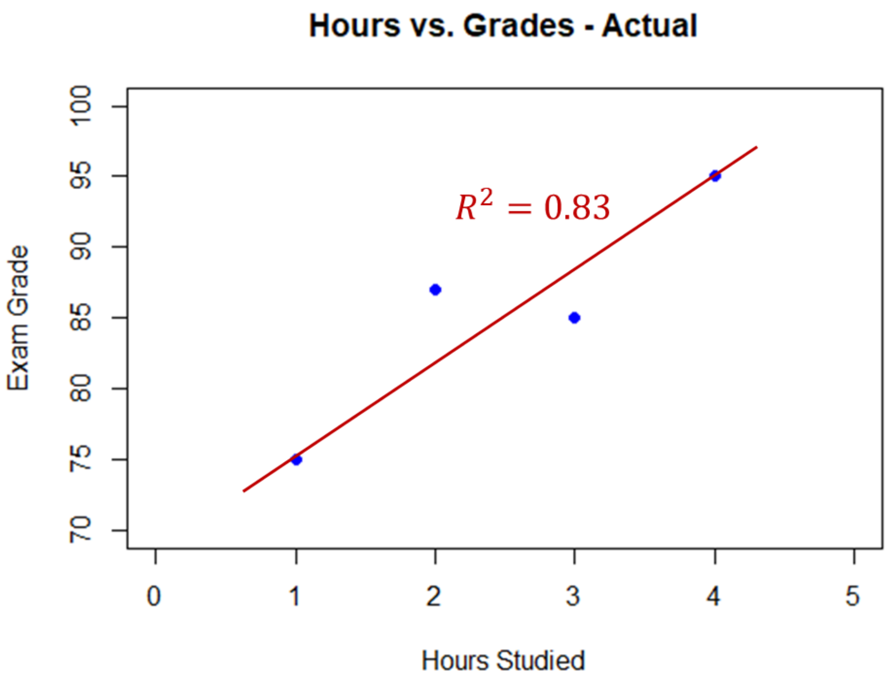
1	2	3	4
75	87	95	85

1	2
75	95

1	2
75	95

1	2
85	75

1	2
85	75



3	4
75	85

3	4
75	85

3	4
85	87

3	4
87	85

1	2	3	4
95	85	75	87

1	2	3	4
95	87	75	85

1	2	3	4
95	85	87	75

1	2	3	4
95	87	85	75

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

1	**2**	**3**	**4**
75	**85**	**87**	**95**

1	2
75	95

1	2	3	4
75	87	85	95

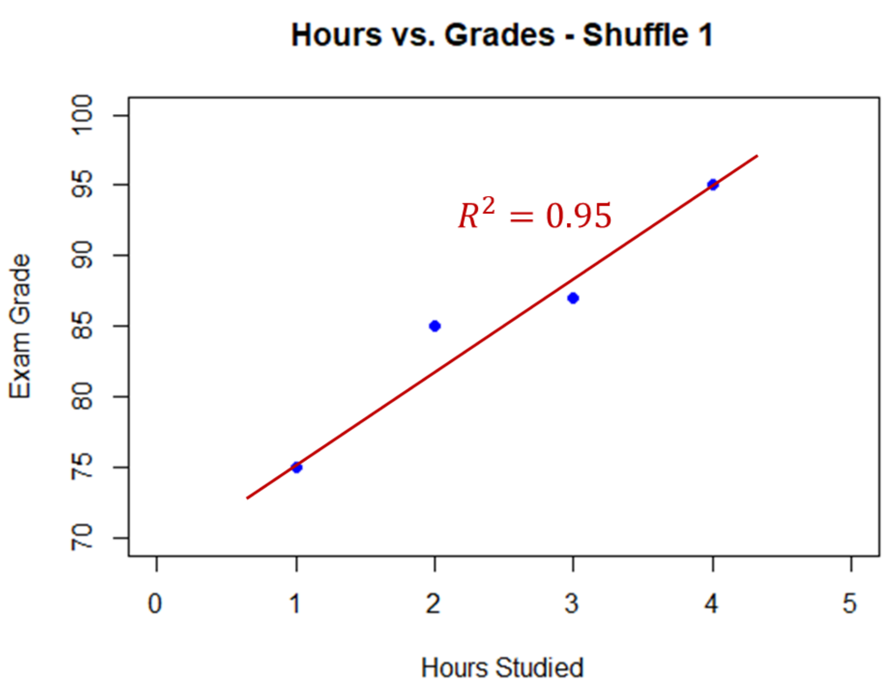
1	2
75	95

1	2	3	4
75	85	95	87

1	2
85	75

1	2	3	4
75	87	95	85

1	2
85	75



3	4
75	85

1	2	3	4
95	85	75	87

3	4
75	85

1	2	3	4
95	87	75	85

3	4
85	87

1	2	3	4
95	85	87	75

3	4
87	85

1	2	3	4
95	87	85	75

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!

1	2	3	4
75	85	87	95

1	2	3	4
75	87	85	95

1	2	3	4
75	85	95	87

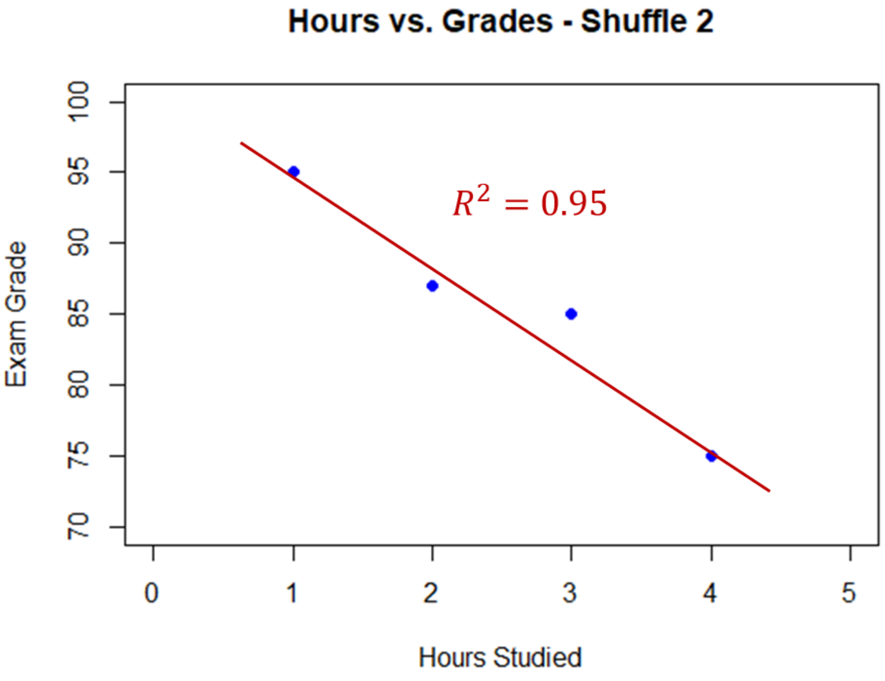
1	2	3	4
75	87	95	85

1	2
75	95

1	2
75	95

1	2
85	75

1	2
85	75



3	4
75	85

3	4
75	85

3	4
85	87

3	4
87	85

1	2	3	4
95	85	75	87

1	2	3	4
95	87	75	85

1	2	3	4
95	85	87	75

1	**2**	**3**	**4**
95	**87**	**85**	**75**

Permutations?

- How many different ways can four students get the grades 75, 85, 87, and 95?
- 24 possible ways this happens!
- There are 4 possible combinations that produce a regression with an R^2 that is greater than or equal to our actual data.

$$\frac{4}{24} = \frac{1}{6} = 16.67\%$$

Permutations vs. Target Shuffling

- 4 possible test grades:

$$4! = 24$$

- 40 possible test grades:

$$40! = 8.16 \times 10^{47}$$

Permutations vs. Target Shuffling

- 4 possible test grades:

$$4! = 24$$

- 40 possible test grades:

$$40! = 8.16 \times 10^{47}$$

- NEED TO SAMPLE!!!