THEORY AND MODEL ASSESSMENT THROUGH SIMULATION

Dr. Aric LaBarr

Institute for Advanced Analytics

THEORY ASSESSMENT

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 – SAS

```
data CLT;
    do sim = 1 to &Simulation_Size;
        do obs = 1 to &Sample_Size;
            call streaminit(12345);
            X1 = RAND('Normal', 2, 5);
            X2 = 5 + 100*RAND('Uniform');
            X3 = 3 + RAND('Exponential');

            output;
        end;
    end;
run;

proc means data=CLT noprint mean;
    var X1 X2 X3;
    by sim;
    output out=Means mean(X1 X2 X3) =
                        Mean_X1 Mean_X2 Mean_X3;
run;
```

Theory Assessment for CLT – SAS

```
data CLT;
     do sim = 1 to &Simulation_Size;
          do obs = 1 to &Sample_Size;
               call streaminit(12345);
               X1 = RAND('Normal', 2, 5);
               X2 = 5 + 100*RAND('Uniform');
               X3 = 3 + RAND('Exponential');

               output;
          end;
     end;
run;

proc means data=CLT noprint mean;
     var X1 X2 X3;
     by sim;
     output out=Means mean(X1 X2 X3) =
                                 Mean_X1 Mean_X2 Mean_X3;
run;
```

Theory Assessment for CLT – R

```
X1 <-
matrix(data=rnorm(n=(sample.size*simulation.size),
mean=2, sd=5), nrow=simulation.size,
ncol=sample.size, byrow=TRUE)
X2 <-
matrix(data=runif(n=(sample.size*simulation.size),
min=5, max=105), nrow=simulation.size,
ncol=sample.size, byrow=TRUE)
X3 <-
matrix(data=(rexp(n=(sample.size*simulation.size)) +
3), nrow=simulation.size, ncol=sample.size,
byrow=TRUE)

Mean.X1 <- apply(X1,1,mean)
Mean.X2 <- apply(X2,1,mean)
Mean.X3 <- apply(X3,1,mean)
```

Theory Assessment Target Shuffling

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	Gender	Buy Product?			
25	M	1			
31	F	0			
28	F	1			
42	M	0			
39	M	1			
…	…				
34	F	0			

Target Shuffling

Age	Gender	Buy Product?	Y_1		
25	M	1	0		
31	F	0	1		
28	F	1	1		
42	M	0	0		
39	M	1	0		
…	…				
34	F	0	1		

Target Shuffling

Age	Gender	Buy Product?	Y_1	Y_2	
25	M	1	0	1	
31	F	0	1	1	
28	F	1	1	1	
42	M	0	0	0	
39	M	1	0	0	
…	…				
34	F	0	1	0	

Target Shuffling

Age	Gender	Buy Product?	Y_1	Y_2	…
25	M	1	0	1	…
31	F	0	1	1	…
28	F	1	1	1	…
42	M	0	0	0	…
39	M	1	0	0	…
…	…				…
34	F	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, c, MAPE, etc.)

Target Shuffling

Misclassification Rate from each model!

Age	Gender	Buy Product?	Y_1	Y_2	…
25	M	1	0	1	…
31	F	0	1	1	…
28	F	1	1	1	…
42	M	0	0	0	…
39	M	1	0	0	…
…	…				…
34	F	0	1	0	…

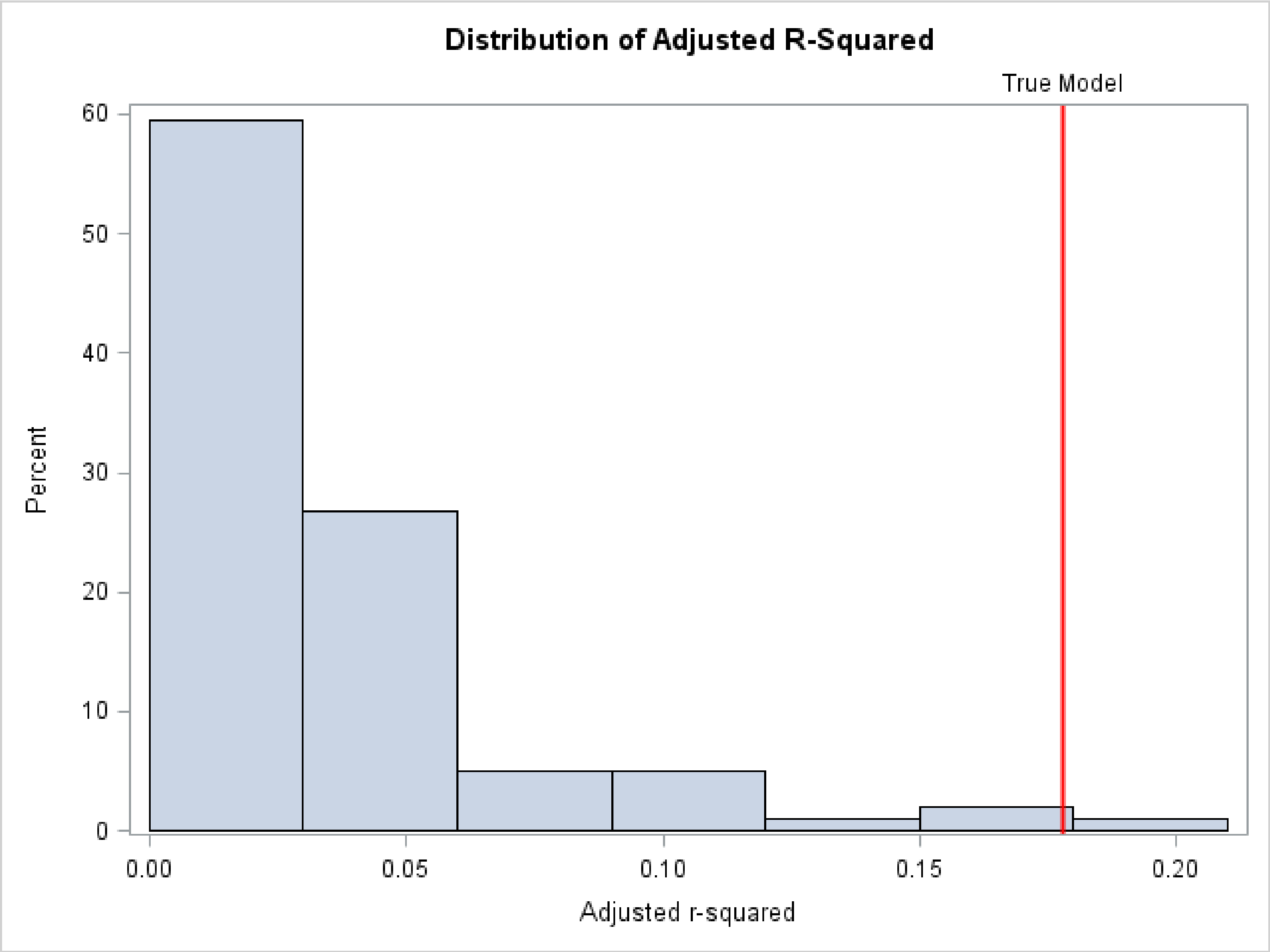
Placebo Effect

- Build model from each of these reshuffled targets and record some measurement of model success (R_A^2, c, 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, c, 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:

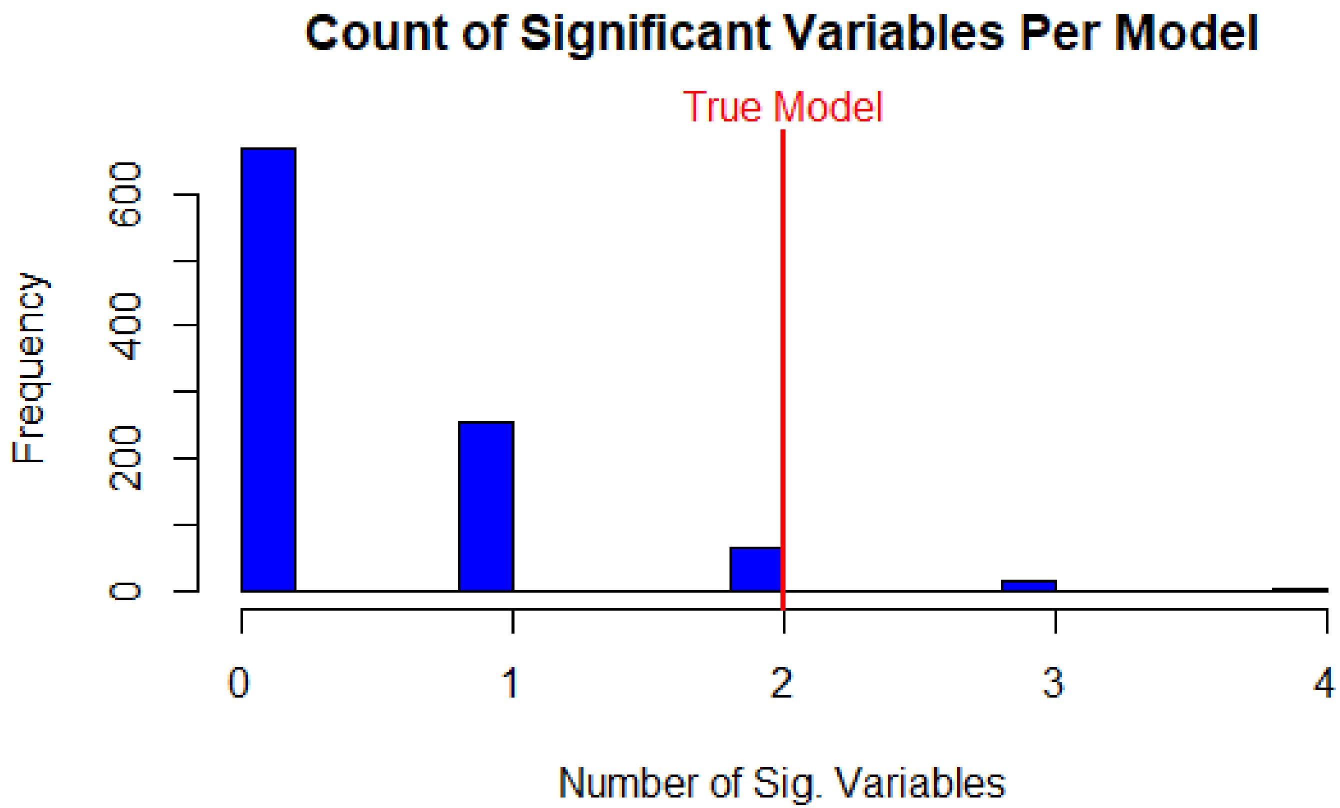
$$y = 5 + 2x_2 - 3x_8 + \varepsilon$$

- Performed target shuffle on the model.

Target Shuffle with 1000 Simulations

Variable	Times Appeared Significant ($p < 0.05$) in a Model
X1	55
X2	62
X3	47
X4	56
X5	50
X6	57
X7	58
X8	40

Target Shuffle with 1000 Simulations



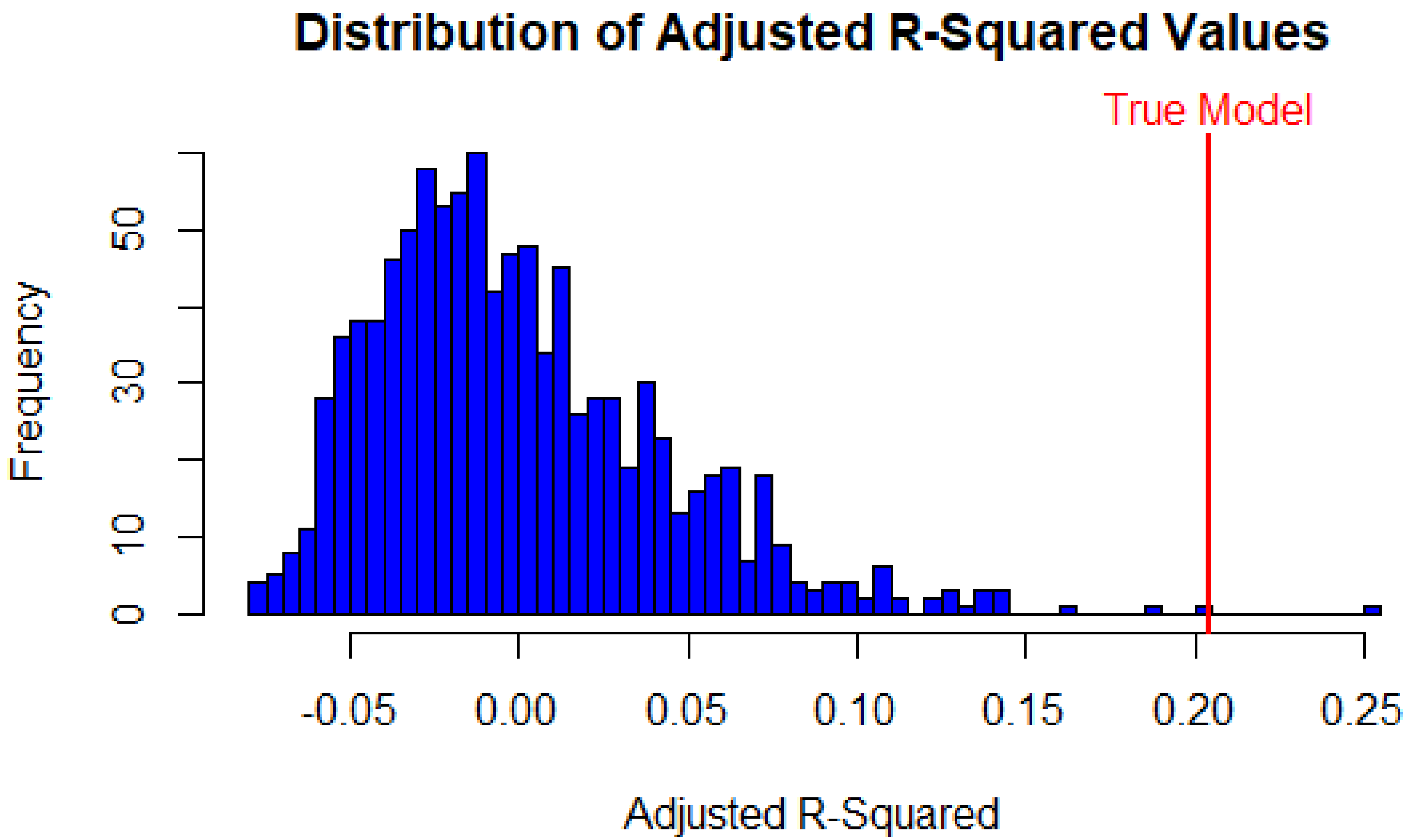
Fake Data Example

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

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

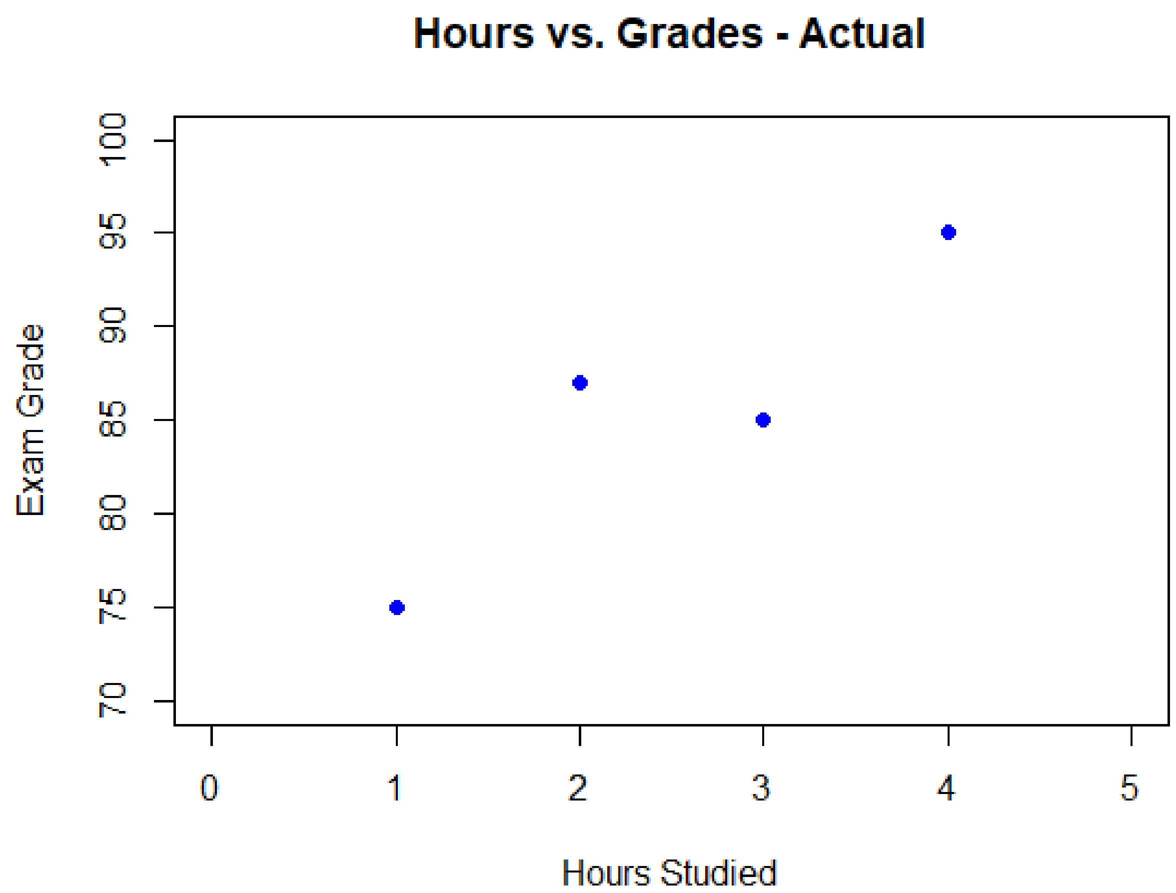
- Adjusted R^2 from this model: 0.204

Target Shuffle with 1000 Simulations

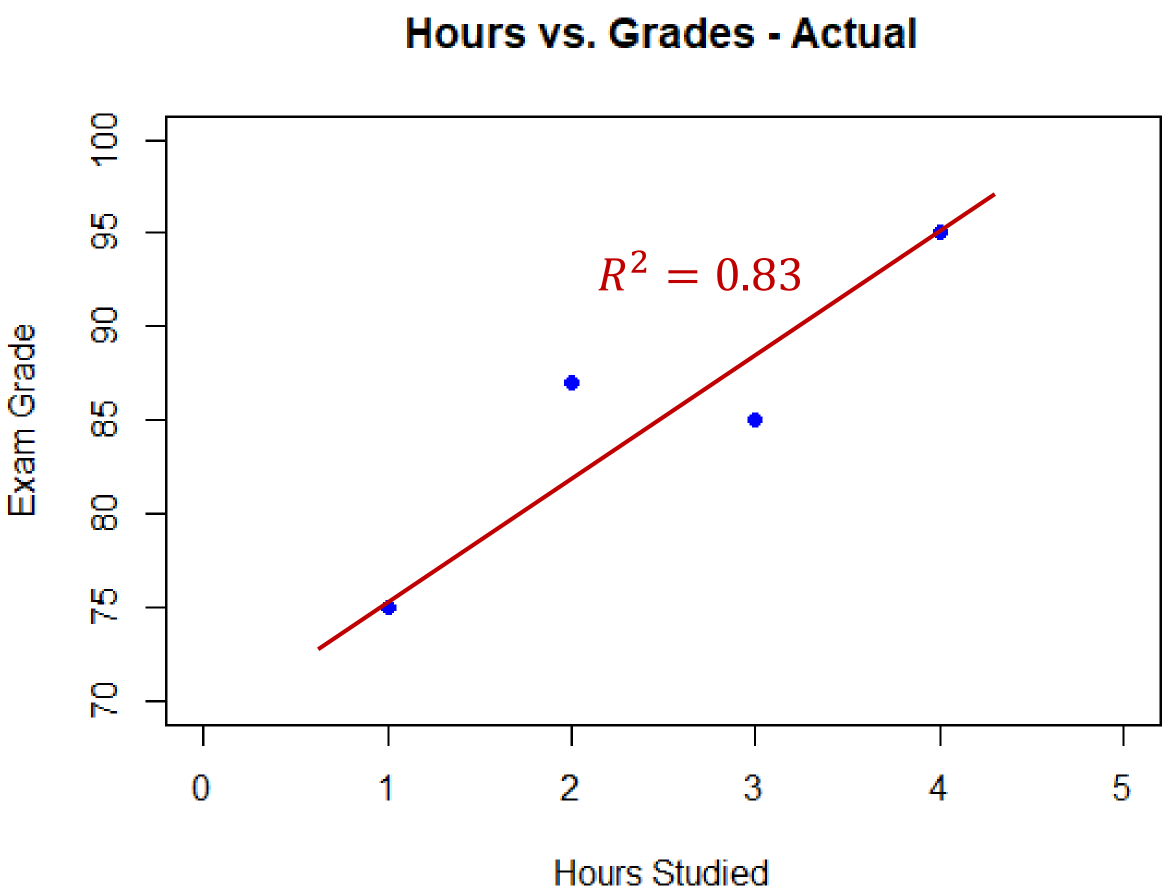


Student Grade Analogy

Student Grade Analogy



Student Grade Analogy



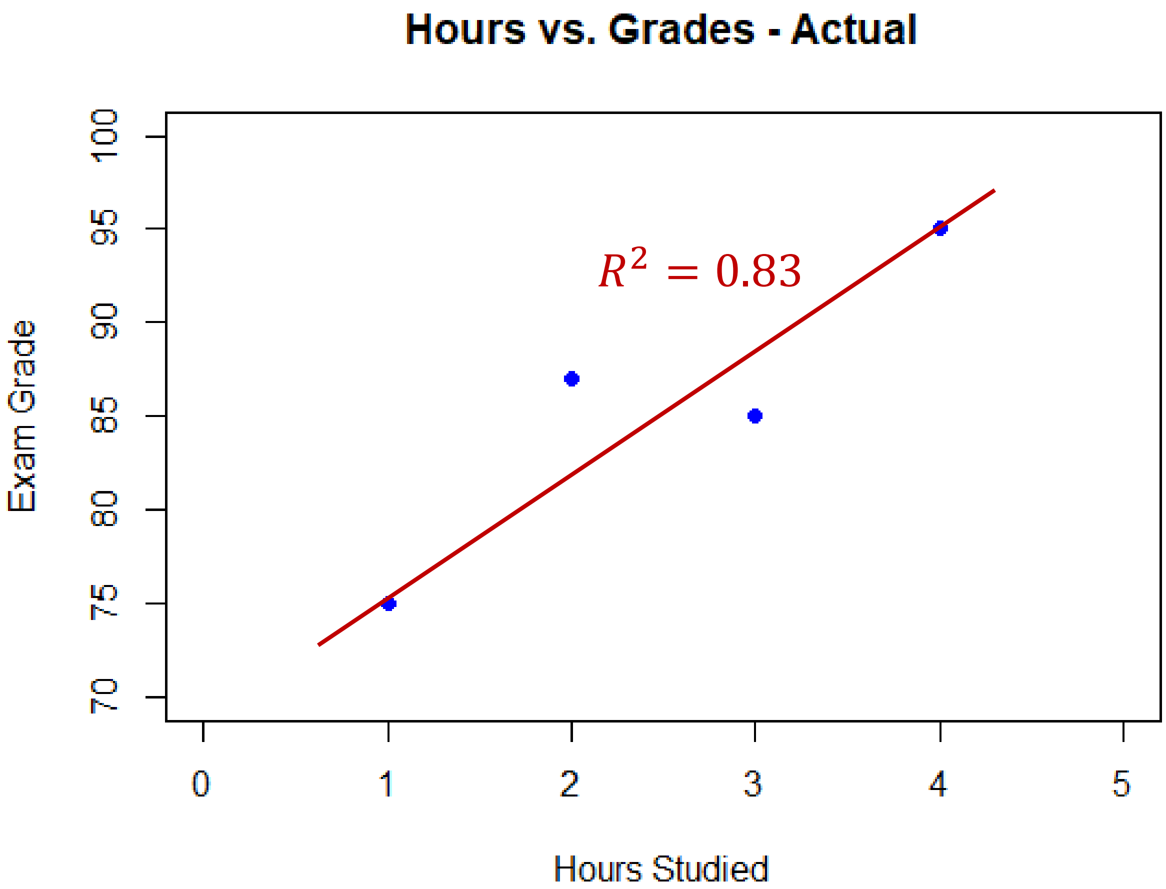
Permutations?

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

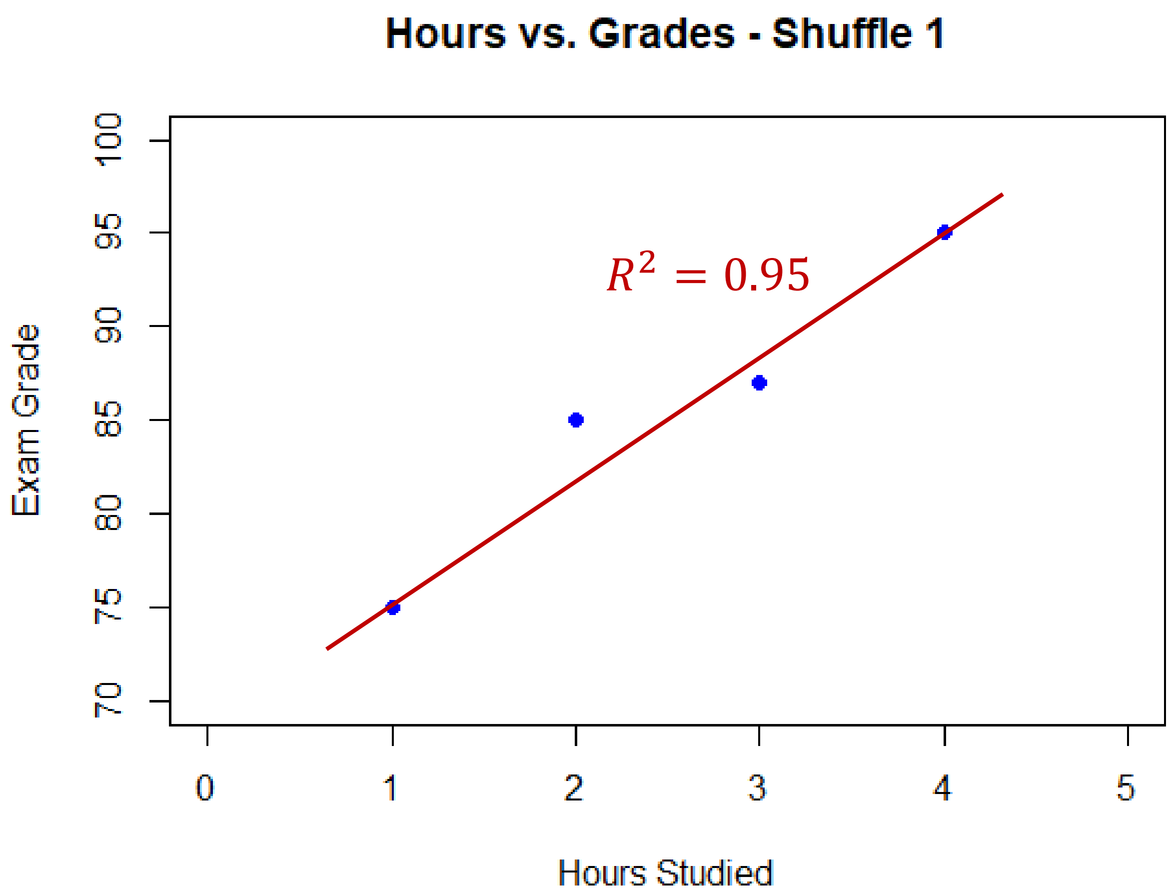
Permutations?

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

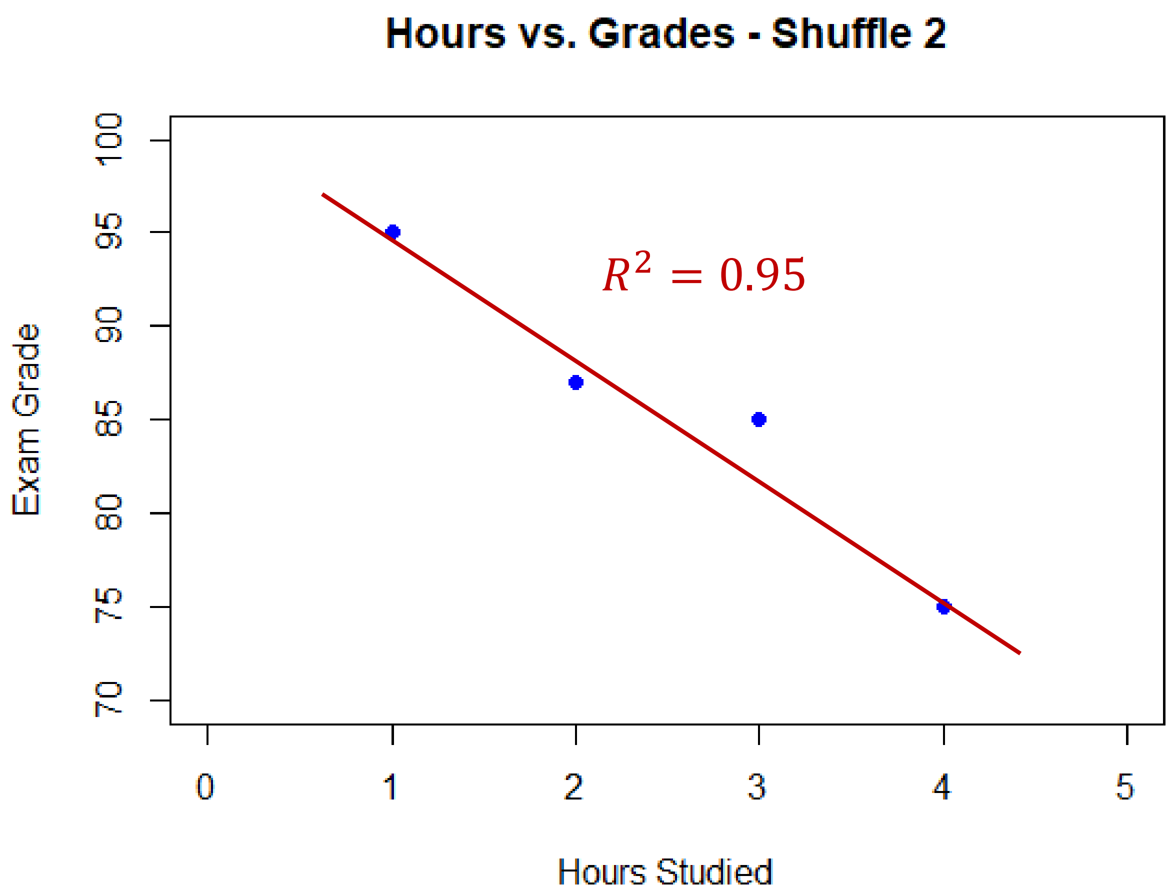
Student Grade Analogy



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!
- There are 3 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!!!

Theory Assessment **Target Shuffling**