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.

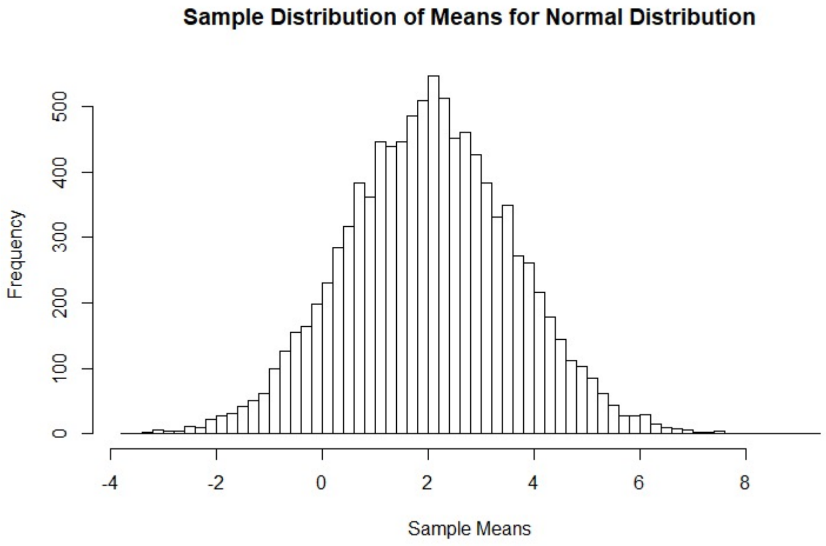
Example – Central Limit Theorem

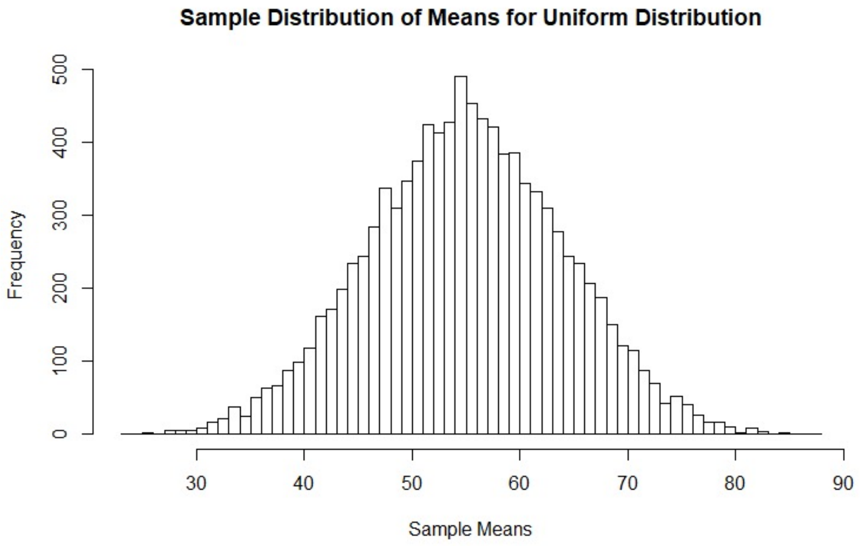
- Assume you do not know the Central Limit Theorem, but you want to understand the sampling distribution of sample means.
- You take samples of size 10, 50, and 100 from the following three population distributions and calculate the sample means:
  1. Normal Distribution
  2. Uniform Distribution
  3.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

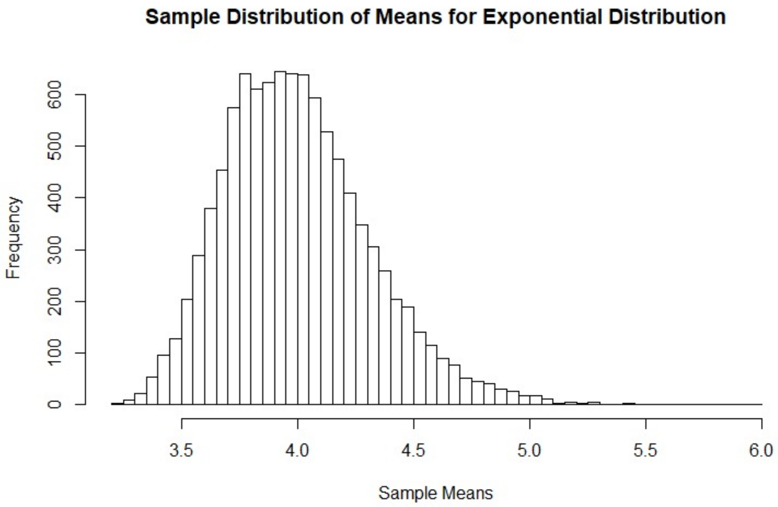
Theory Assessment for CLT – R

```
sample.size <- 10
simulation.size <- 10000

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)
```

Assessment for CLT – R (n = 10)







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 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:

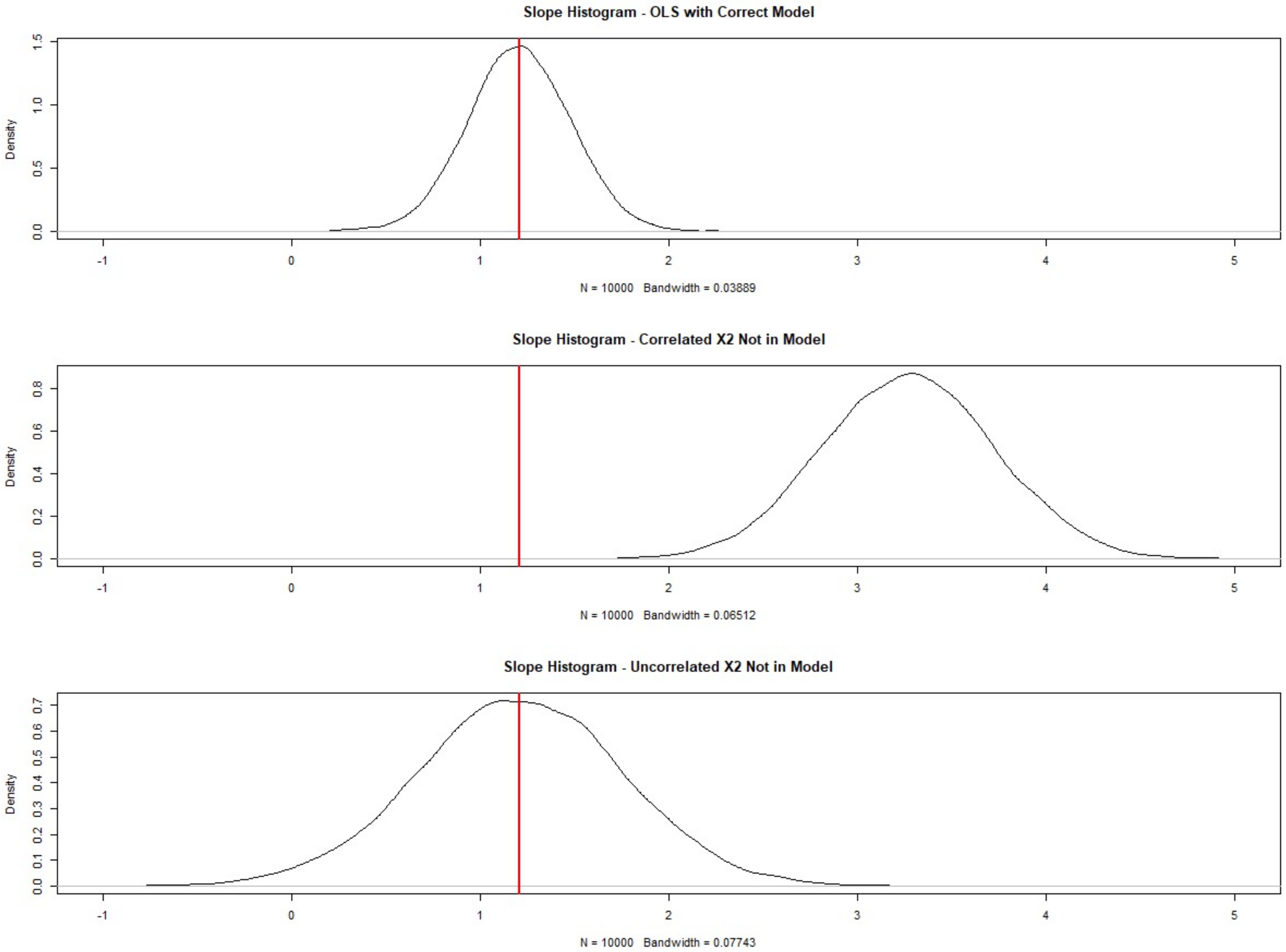
$$Y = -13 + 1.21X_1 + 3.45X_2 + \varepsilon$$

- 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



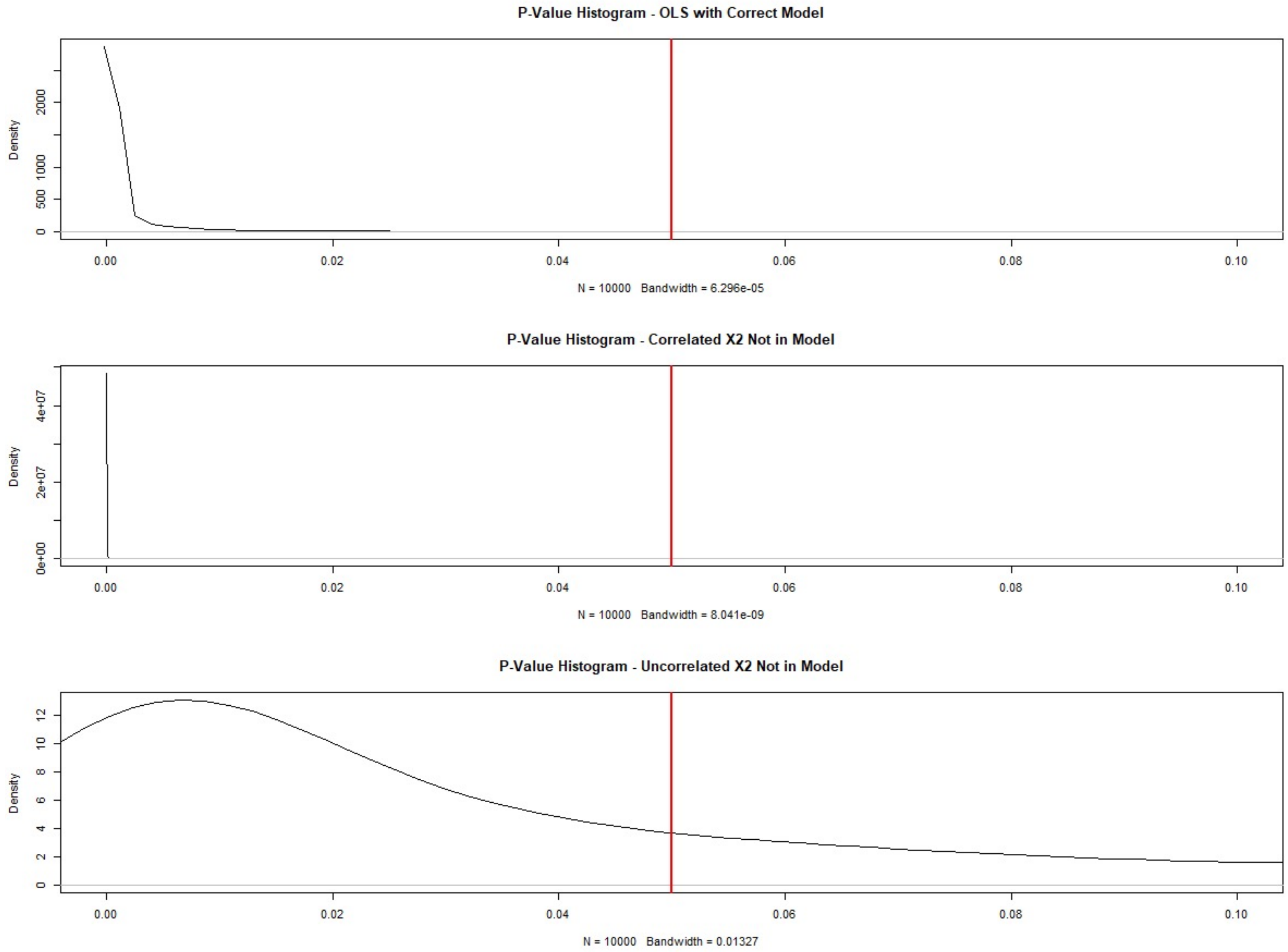
Example – Omitted Variable Bias

- 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



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	1.39%
Correlated X2 Not in Model	0.00%
Uncorrelated X2 Not in Model	40.84%

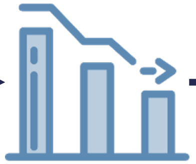
TARGET SHUFFLING

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			



Build Model

Record Model Metric

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		
25	M	1	0		
31	F	0	1		
28	F	1	1		
42	M	0	0		
39	M	1	0		
…	…				
34	F	0	1		

Build Model

Record Model Metric

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

Model metric 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	…

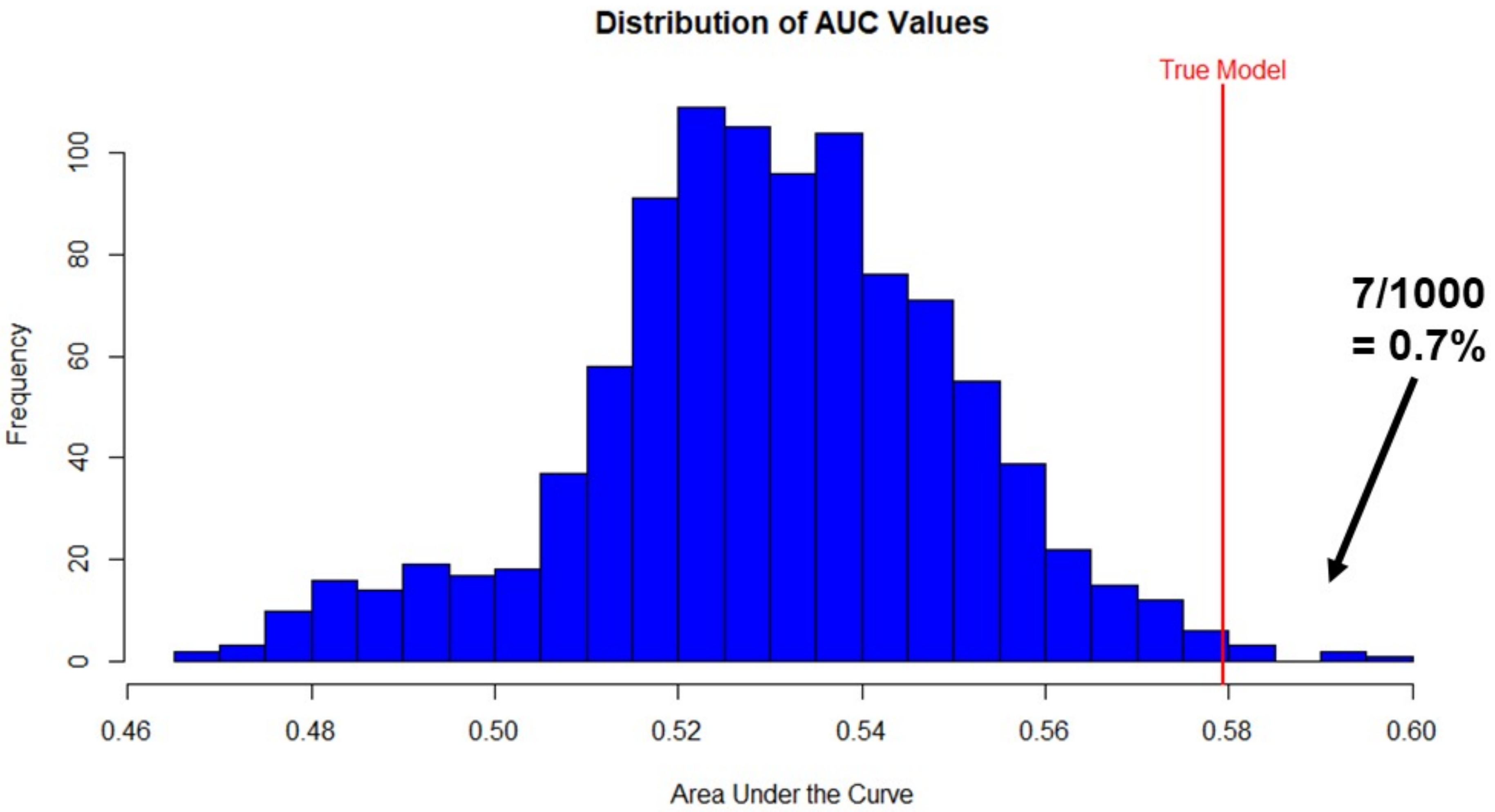
Placebo Effect

- 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.

Fake Data Example

```
Fake <- data.frame(matrix(rnorm(n=(100*8)), nrow=100, ncol=8))
Err <- rnorm(n=100, mean=0, sd=8)
Y <- 5 + 2*Fake$X2 - 3*Fake$X8 + Err
Fake <- cbind(Fake, Err, Y)

sim <- 1000

Y.Shuffle <- matrix(0, nrow=100, ncol=sim)
for(j in 1:sim){
  Uniform <- runif(100)
  Y.Shuffle[,j] <- Y[order(Uniform)]
}

Y.Shuffle <- data.frame(Y.Shuffle)
colnames(Y.Shuffle) <- paste('Y.',seq(1:sim),sep="")

Fake <- data.frame(Fake, Y.Shuffle)

R.sq.A <- rep(0,sim)
for(i in 1:sim){
  R.sq.A[i] <- summary(lm(Fake[,10+i] ~ Fake$X1 + Fake$X2 + Fake$X3 + Fake$X4
                          + Fake$X5 + Fake$X6 + Fake$X7 + Fake$X8))$adj.r.squared
}
True.Rsq.A <- summary(lm(Fake$Y ~ Fake$X1 + Fake$X2 + Fake$X3 + Fake$X4
                         + Fake$X5 + Fake$X6 + Fake$X7 + Fake$X8))$adj.r.squared
```

Fake Data Example

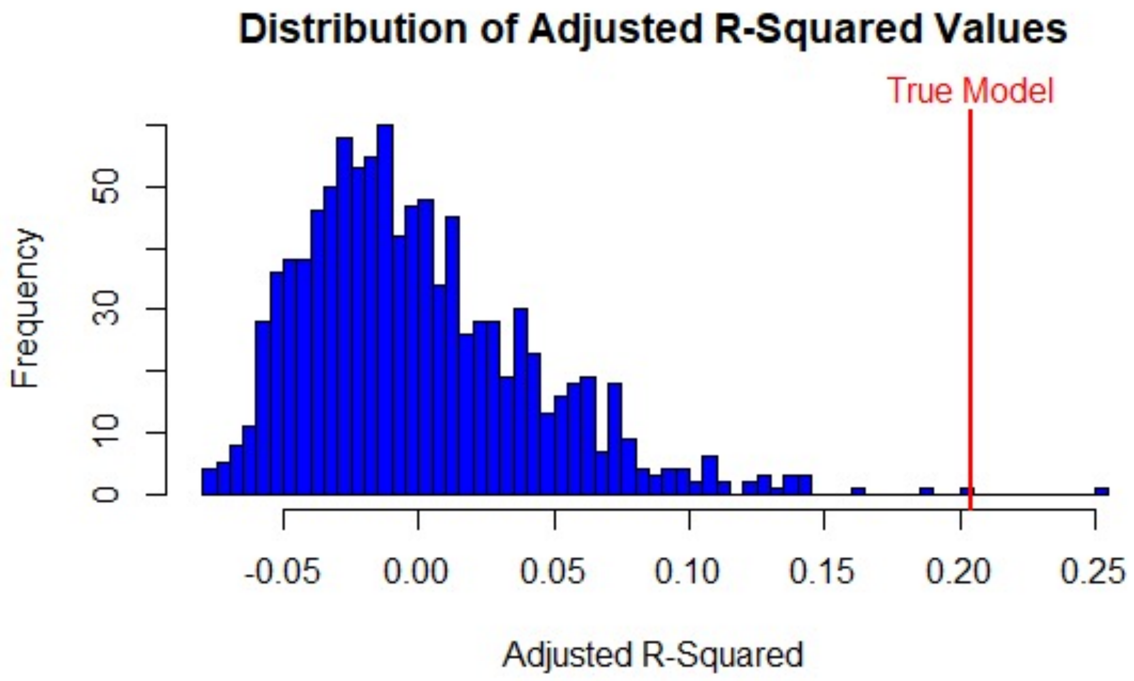
- Randomly generated 8 variables that follow a Normal distribution with mean of 0 and standard deviation of 8.
- Defined relationship with target variable:

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

- Adjusted R^2 from this model: 0.204

Fake Data Example

```r
hist(c(R.sq.A,True.Rsq.A), breaks=50, col = "blue",
     main='Distribution of Adjusted R-Squared Values',
     xlab='Adjusted R-Squared')
abline(v = True.Rsq.A, col="red", lwd=2)
mtext("True Model", at=True.Rsq.A, col="red")
```



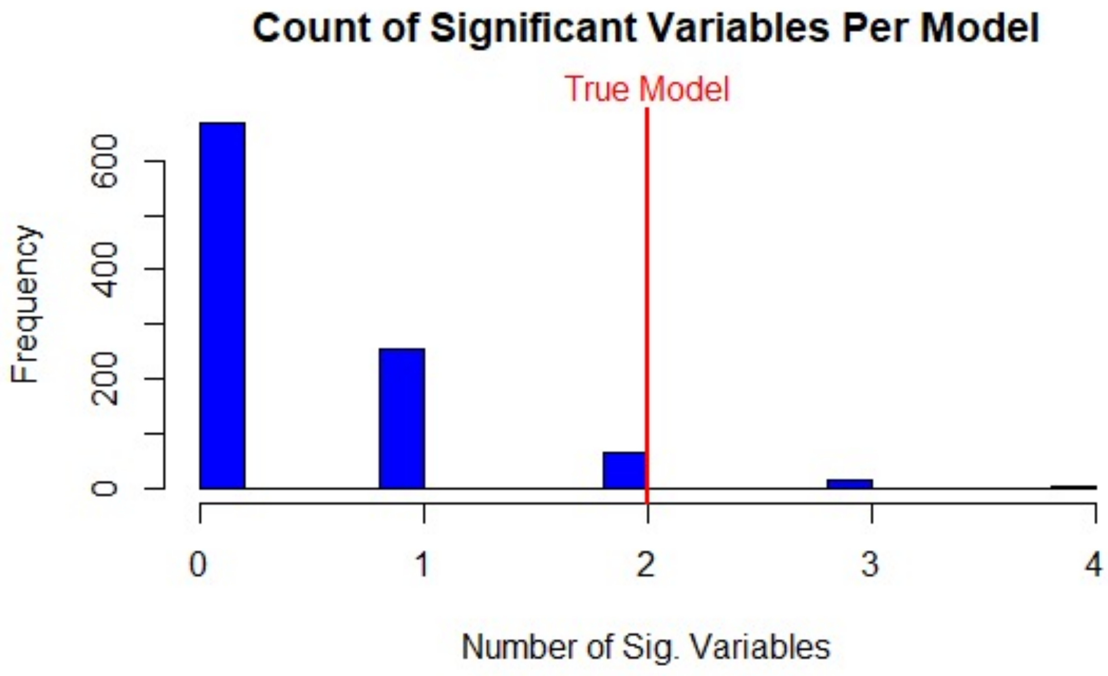
Target Shuffle with 1000 Simulations

```
P.Values <- NULL
for(i in 1:sim){
  P.V <- summary(lm(Fake[,10+i] ~ Fake$X1 + Fake$X2 + Fake$X3 + Fake$X4
                    +                            + Fake$X5 + Fake$X6 +
Fake$X7 + Fake$X8))$coefficients[,4]
  P.Values <- rbind(P.Values, P.V)
}

Sig <- P.Values < 0.05
```

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

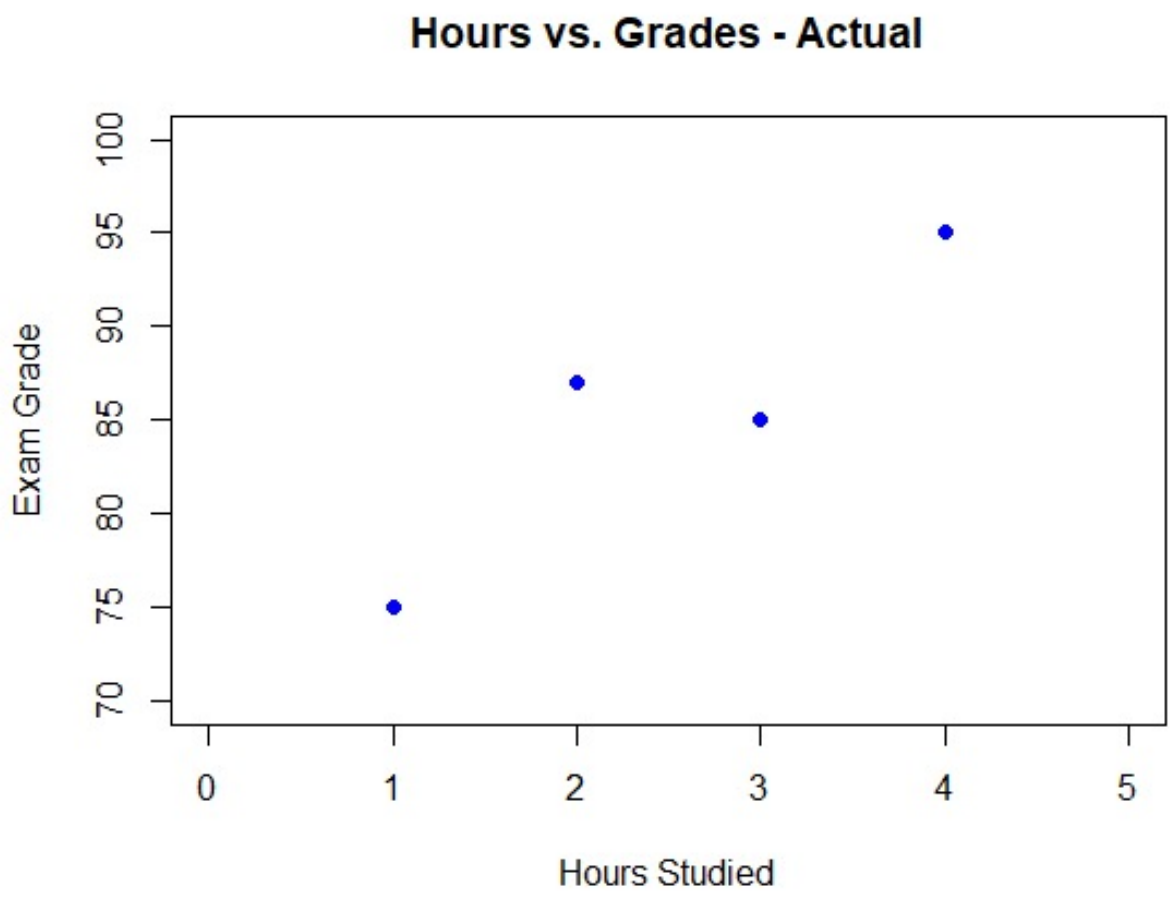
Fake Data Example

```
hist(rowSums(Sig)-1, breaks=25, col = "blue",
     main='Count of Significant Variables Per Model',
     xlab='Number of Sig. Variables')
abline(v = 2, col="red", lwd=2)
mtext("True Model", at=2, col="red")
```

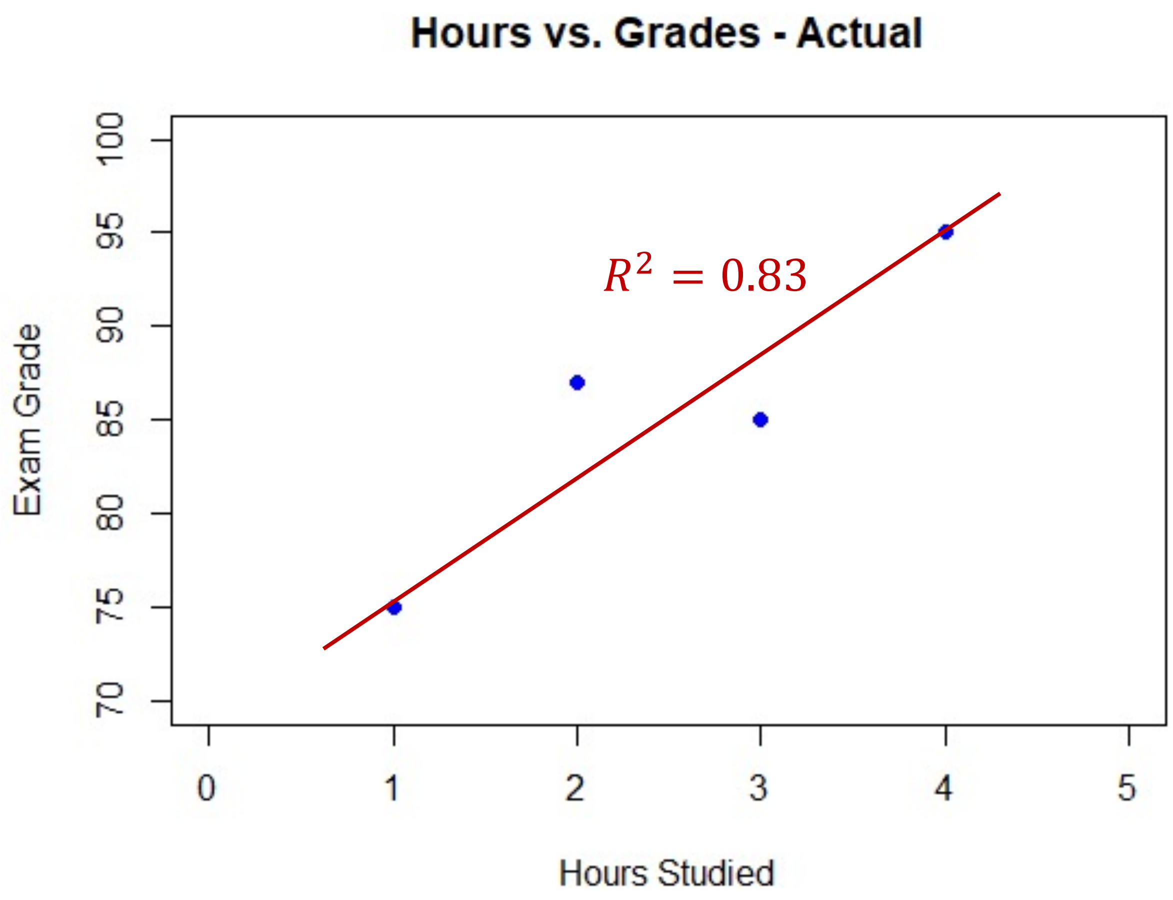


Student Grade Analogy

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
75	85	87	95

1	2	3	4
75	95	85	87

1	2	3	4
85	87	75	95

1	2	3	4
87	75	85	95

1	2	3	4
87	95	75	85

1	2	3	4
95	85	75	87

1	2	3	4
75	87	85	95

1	2	3	4
75	95	87	85

1	2	3	4
85	87	95	75

1	2	3	4
87	75	95	85

1	2	3	4
87	95	75	85

1	2	3	4
95	87	75	85

1	2	3	4
75	85	95	87

1	2	3	4
85	75	87	95

1	2	3	4
85	95	75	87

1	2	3	4
87	85	75	95

1	2	3	4
95	75	85	87

1	2	3	4
95	85	87	75

1	2	3	4
75	87	95	85

1	2	3	4
85	75	95	87

1	2	3	4
85	95	87	75

1	2	3	4
87	85	95	75

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

1	2	3	4
75	95	85	87

1	2	3	4
85	87	75	95

1	2	3	4
87	75	85	95

1	2	3	4
87	95	75	85

1	2	3	4
95	85	75	87

1	2	3	4
75	87	85	95

1	2	3	4
75	95	87	85

1	2	3	4
85	87	95	75

1	2	3	4
87	75	95	85

1	2	3	4
87	95	75	85

1	2	3	4
95	87	75	85

1	2	3	4
75	85	95	87

1	2	3	4
85	75	87	95

1	2	3	4
85	95	75	87

1	2	3	4
87	85	75	95

1	2	3	4
95	75	85	87

1	2	3	4
95	85	87	75

1	2	3	4
75	87	95	85

1	2	3	4
85	75	95	87

1	2	3	4
85	95	87	75

1	2	3	4
87	85	95	75

1	2	3	4
95	75	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
75	95

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

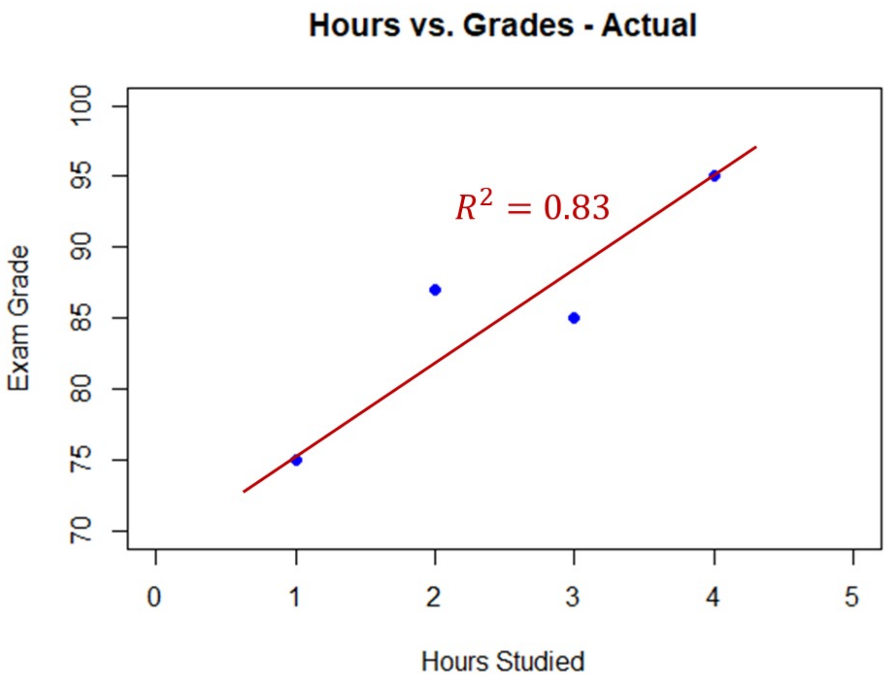
1	2
75	95

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

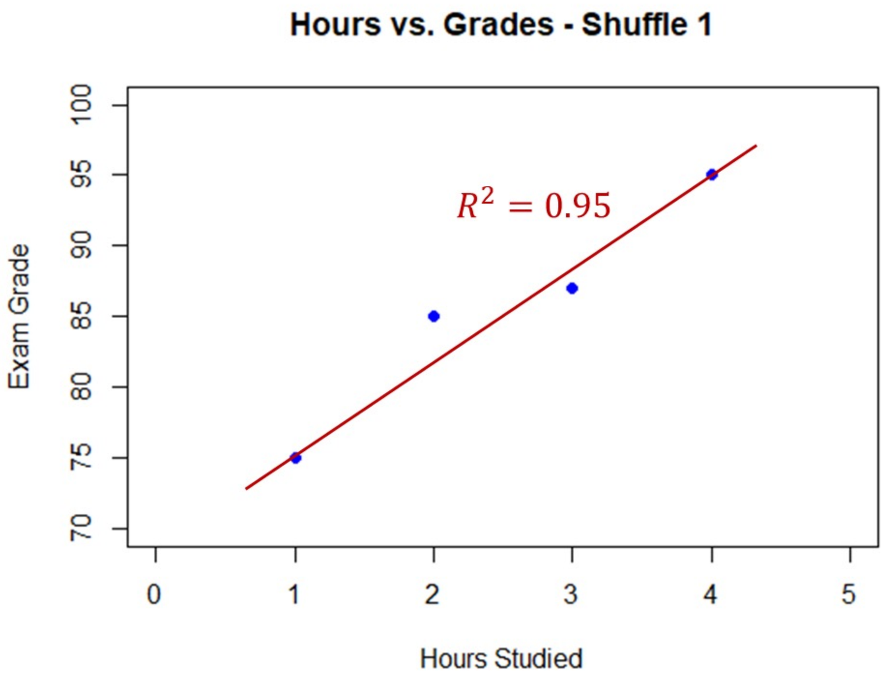
1	2	3	4
75	87	95	85

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!

1	2	3	4
75	85	87	95

1	2
75	95

1	2	3	4
75	87	85	95

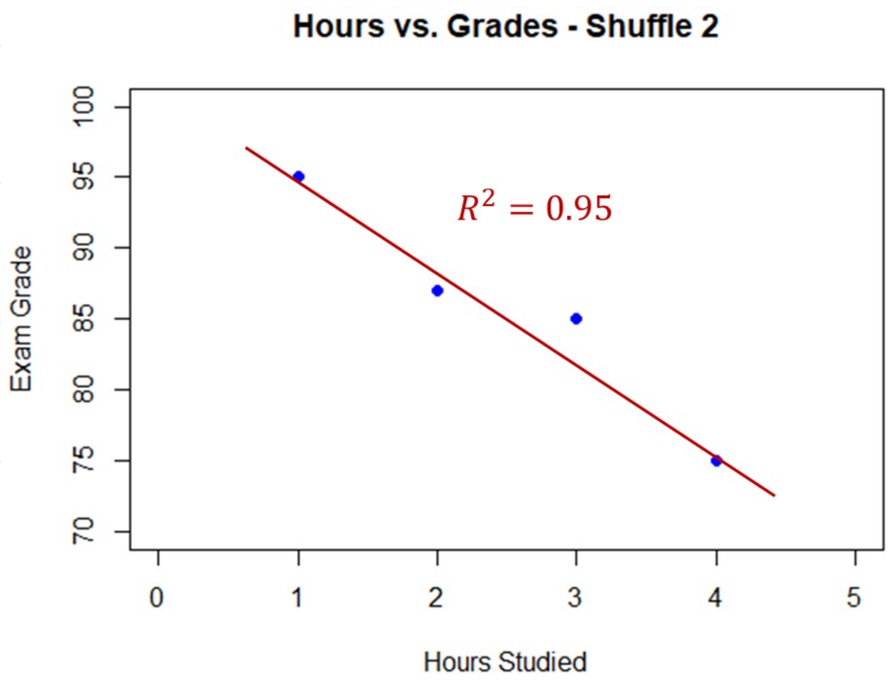
1	2
75	95

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!
- 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!!!

Student Grade Example

```
x <- c(75, 85, 87, 95)

y.all <- data.frame(t(permutations(4,4,x)), input = 1:4)

my_lms <- lapply(1:24, function(x) lm(y.all[,x] ~ y.all$input))
summaries <- lapply(my_lms, summary)
rsq <- sapply(summaries, function(x) c(r_sq = x$r.squared))
```