Please get in groups of 6! We’re passing out one sheet of white paper and one of PURPLE. Fold each piece of paper so there are 6 sections. Each person should write their age in months on the white piece of paper (in a quadrant) and their height in inches on the PURPLE piece of paper (in a quadrant). Once everyone’s height and age are on the pieces of paper, rip them up to create smaller pieces of paper.

|  |  |
| --- | --- |
| 217 | 71 |
| 218 | 58 |
| 122 | 80 |
| 204 | 62 |
| 220 | 72 |
| 216 | 60 |

Line up your pairs of observations in two columns (Height and Age) so that each person’s height is next to their age. This is your data! It’s even in tidy format!

1. Use R to create a dataset which looks like our dataset here. Estimate a linear model which predicts height based on age. Change the code below to correspond to your data.

data <- data.frame(height = c(71, 58, 80, 62, 72, 60), age = c(217, 218, 122, 204, 220, 216))

1. Below is the output from estimating the linear model . Where is height, and is age. Circle in the output where we find and write a one sentence interpretation of based on the output (Interpret for a family member, not your stats teacher!).

Call:

lm(formula = height ~ age, data = data)

Coefficients:

(Intercept) age

409.285 -3.123

Interpretation:

|  |  |
| --- | --- |
| Age | Height |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. Now try with your data. **Fill in the table** to the right, estimate a linear model predicting height from age in your dataset, and **provide the estimate of with an interpretation** of what it means in your data. Check out page 3, there is a graph that you’ll fill out shortly. Draw a big thick vertical line to represent where the b1 from the original sample is.
2. When thinking about model comparison we want to imagine a world where the simple model is true. Write out the GLM equation for the simple model here:
3. Notice that the simple model is a version of the complex model, where . We can interpret this as a situation where knowing a person’s X **does not** help us guess their Y. Can you apply this interpretation to the current situation?

Knowing a person’s \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ does not help us guess their \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_.

If knowing a person’s age does not help us guess their height, then it should not matter that Person A’s height is attached to Person A’s age, rather it just as likely could have been Person B’s height. We can use this understanding to generate a sampling distribution of s which all come from samples where height and age are unrelated. We can do this by shuffling height and re-estimating our model!

1. Give it a try! Pick up your pile of 6 height cards, shuffle them up, and they lay them down again in a column next to the age variables. Fill in the table to the right with your new dataset.

|  |  |
| --- | --- |
| Age | Height |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

1. Estimate a linear model predicting height from age using this new shuffled dataset. What is in this new dataset?
2. You’re going to do step 6 and 7 a few times. When you do fill in from each shuffled sample on the sampling distribution on the next page. Do it at least 10 times! Break up the task within your team for efficiency, have people doing different parts: shuffle, put data in R and run linear model, add b1 to histogram.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |  |

**-4 -3 -2 -1 0 1 2 3 4**

**COUNT**

**b1**

1. Is your sampling distribution centered around the original estimate from the original sample? Explain why this distribution is or is not centered around the original estimate?
2. Compare where the shuffled samples ended up as compared to the original sample estimate (marked on your graph with a vertical line). Does it look like your data could have been generated from the simple model? Explain why or why not.

**YOU CAN WORK AT YOUR OWN PACE FROM HERE ON OUT, BUT I RECOMMEND CONTINUING TO CHECK IN WITH YOUR GROUP**

1. Let’s start to do this using R, since shuffling by hand can be pretty arduous. Try running the command shuffle(data$height). Describe what you get? What is this equivalent to doing with your cards?
2. Now try estimating a model with shuffled height values and pulling out the b1 coefficient! Below is the command for the original analysis, how would we change the code to use shuffled height values instead (edit the text below)?

b1(height~age, data = data)

1. What if we shuffle many many times, each time pulling out a b1 to represent a random sample from a population where the simple model is true. Do this using a do()\* statement! Draw a histogram of the sampling distribution you get (no need to be perfect, just the relative shape), make sure to label the X and Y axis. Draw a vertical line on this plot as well, at the place of the original value just like on the previous graph.
2. Compare where the shuffled samples ended up as compared to the original sample estimate (marked on your graph with a vertical line). Does it look like your data could have been generated from the simple model? Explain why or why not.
3. What if we looked at the F-value instead of . Below is the supernova output from my original data. Run supernova on your original data, and draw a vertical line on the graph below where the original F is. Use a do statement in combination with the fVal function to create a sampling distribution of F-values under the assumption that the simple model is true. Draw the distribution of F-values you find on the graph below.

> supernova(lm(height~age, data = data))

Analysis of Variance Table (Type III SS)

Model: height ~ age

SS df MS F PRE p

----- ----------------- -------- - -------- ----- ------ -----

Model (error reduced) | 3559.050 1 3559.050 3.738 0.4831 .1253

Error (from model) | 3808.450 4 952.112

----- ----------------- -------- - -------- ----- ------ -----

Total (empty model) | 7367.500 5 1473.500

**F**

**0 1 2 3 4 5 6**

1. Use R to calculate the probability that a shuffled sample had an F which was greater than or equal to the F observed in your original sample.
2. Write an interpretation of this probability (also called a p-value). Remember to interpret for a family member, not your stats teacher!
3. How does this probability compare to the p-value from your supernova output, is it similar? Should it be?
4. People often get confused between bootstrapping and shuffling. Shuffling is a method for creating samples from a population where the simple model is true. Bootstrapping is a method for creating samples from a population which looks like our original sample. Describe at least 2 similarities and 2 differences between bootstrapping and shuffling.
5. If we had bootstrapped the sampling distribution of F or do you think it would have looked similar or very different from the shuffled distributions you drew on the previous pages. Will this always be the case?
6. We shuffled height, but we could have shuffled age. Do you think this would make a difference in what we find? Try it with age and reflect on whether the results are different.