LEARN MORE BY FUSING MULTIPLE DATASETS TOGETHER

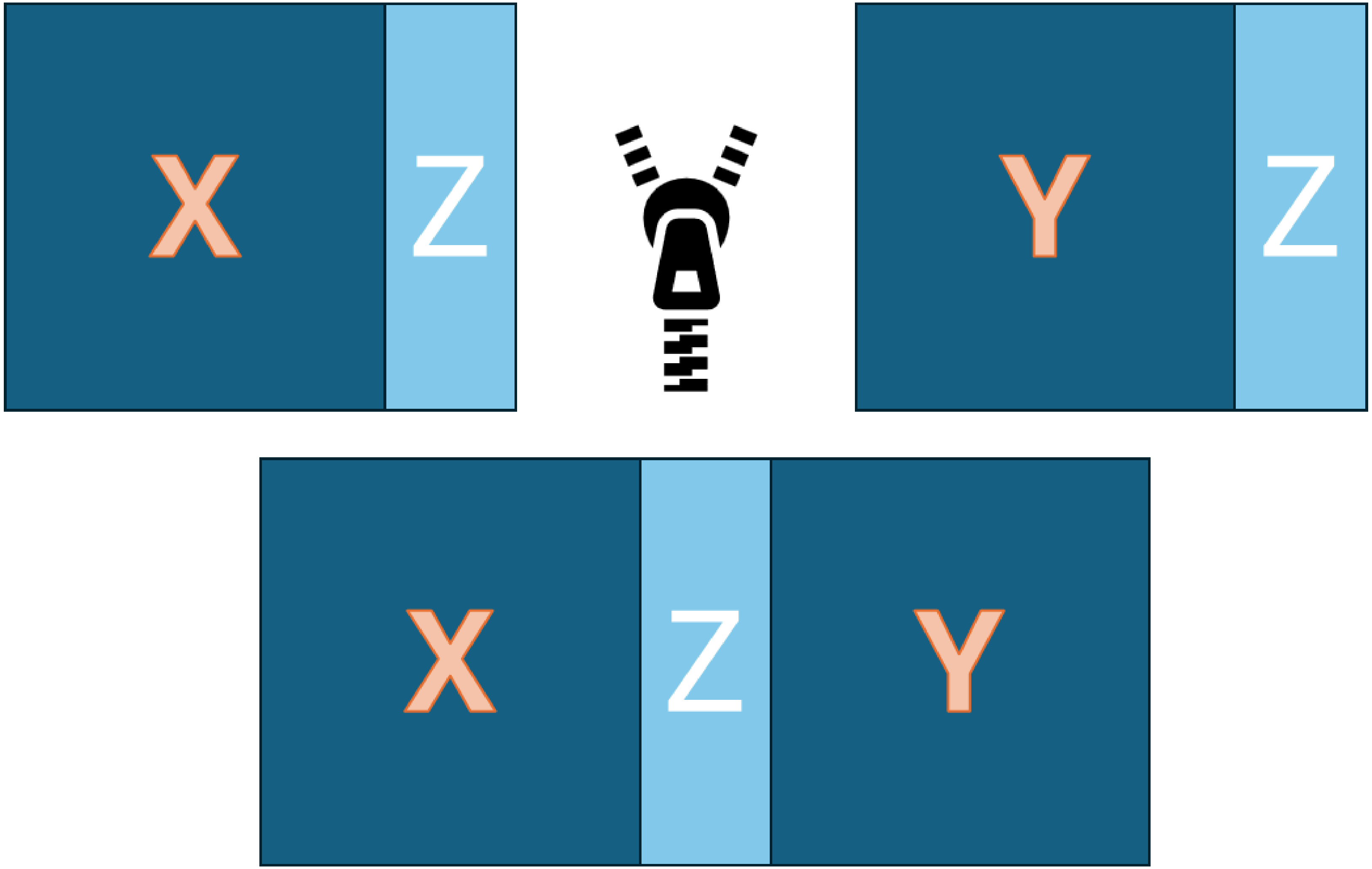
Michael Conklin

56Stats

AGENDA

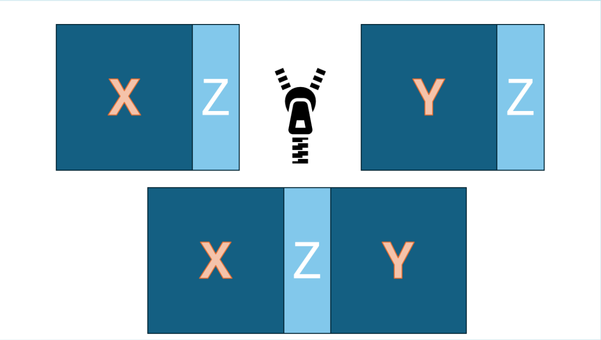
- What is Data Fusion
- How does it work?
- How to measure similarity
- What can we learn from fused data?
- Tips and caveats

WHAT IS DATA FUSION



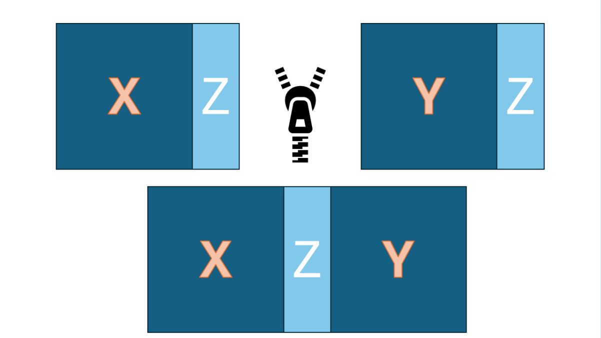
HOW IT WORKS

- By finding records with similar Z variables in both data sets, we combine records to create our new fused data set where we treat the Y data as if it came from the same record as the X data. Even though there is no actual connection between the two initial data sets.
- In terms of two surveys: if we find that John and Mary have similar Z variables then we treat Mary's Y data as if it came from John.
- Is this a good assumption? As always it depends…



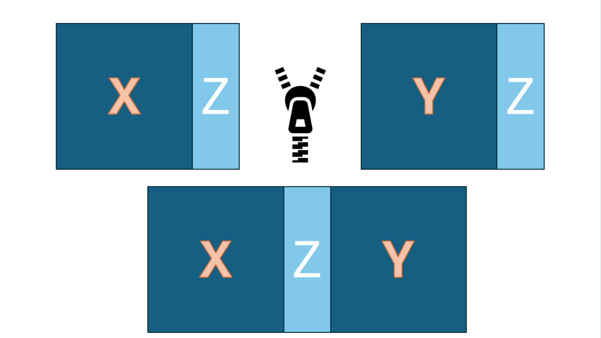
CHOOSING THE Z VARIABLES

- The Z variables need to be related to (predictive of) the X variables in the first data set AND related to (predictive of) the Y variables in the second data set.
- This means it is critically important to have good Z variables. Unfortunately, we are often without the option of choice. We have what we have in the two data sets.
- If the Z variables are poor predictors of the Y or X variables then we are, in effect, randomly putting records together.



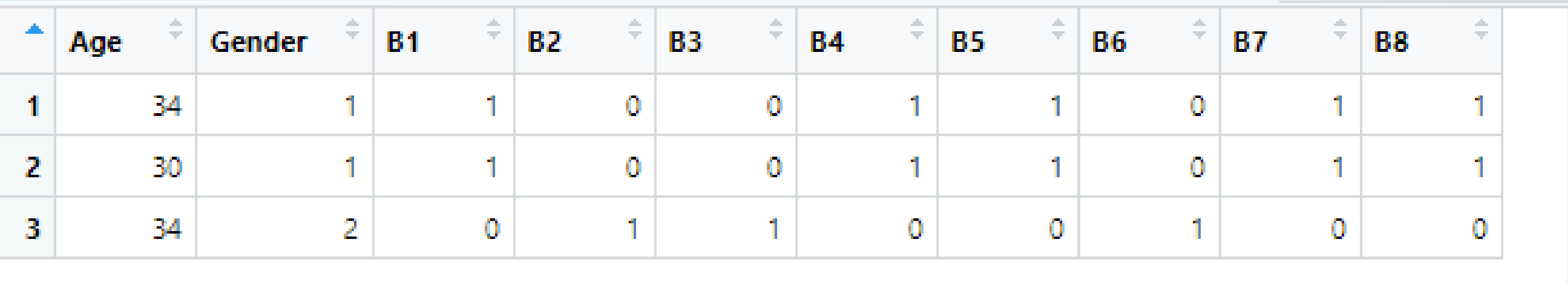
CANDIDATES FOR GOOD Z VARIABLES

- Frequently, the Z variables are demographics. Unfortunately, these are rarely sufficient.
- Specific behaviors are better, especially if they are related to the X and Y variables of interest.
- Attitudes, if available are very useful
- Don’t let anyone tell you there is a small set of “golden questions” that always work to fuse a data set.



HOW TO MEASURE SIMILARITY

- To decide if two records are similar we compare the similarity of their vectors of Z variables.
- WARNING: Many programs default to using Euclidean distance as the measure of similarity.



	Age	Gender	B1	B2	B3	B4	B5	B6	B7	B8
1	34	1	1	0	0	1	1	0	1	1
2	30	1	1	0	0	1	1	0	1	1
3	34	2	0	1	1	0	0	1	0	0

- In this example, Euclidean distance effectively matches records 1 and 3 solely on the basis of age

```
  1 2 3
1 0 4 3
2 4 0 5
3 3 5 0
```

MY FAVORITE SIMILARITY MEASURE

- To mitigate the problems of different variables having different units of measure and variance, sometimes using Mahalanobis distance is recommended. Effectively a standardized distance.

```
         1        2        3
1 0.000000 1.333333 1.333333
2 1.333333 0.000000 1.333333
3 1.333333 1.333333 0.000000
```

- In this case a random match would occur
- But – my favorite is the Gower distance metric.

GOWER DISTANCE

- Handles different types of data (metric,categorical,logical) with ease
- All distances on individual variables range from 0 to 1
- Logical variables (Important when measuring behaviors) are matched only with positive matchs. That way we only match records based on action instead of inaction.

```
    1   2   3
1 0.0 0.1 0.9
2 0.1 0.0 1.0
3 0.9 1.0 0.0
```

- In our example, records 1 and 2 will be matched, as we would like.

WE FUSED THE DATA...NOW WHAT?

- Now we can look at the relationship between the X variables and the Y variables.

$$cor(x, y) > cor(x, \hat{y})$$

- The correlation (or any measure of association) between an x variable and a y variable will always be weaker when the y variable is matched via fusion. This is because the fusion always adds some random error.
- Therefore, it is very difficult to determine if the relationship between two x variables is stronger than the relationship between an x and a y variable.

ASSESSING "GOODNESS" OF FUSION

- Data Fusion introduces error because the Z variables do not perfectly predict the Y or the X variables.
- And generally we have no examples where the X and Y variables are jointly observed . . . our two data sets are completely different sources.
- Is there a way to estimate the amount of error that a fusion introduces, and if so, is that useful?

$$\hat{y} = y + e?$$

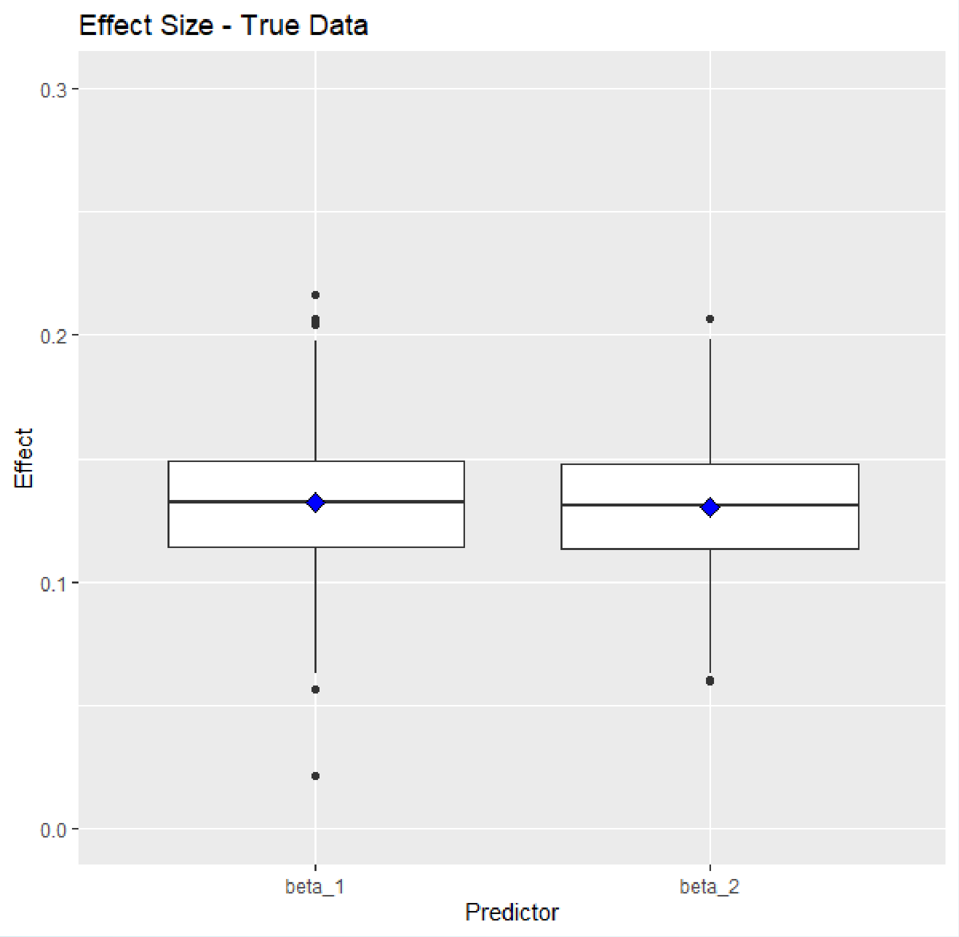
LET'S START WITH A SIMULATION

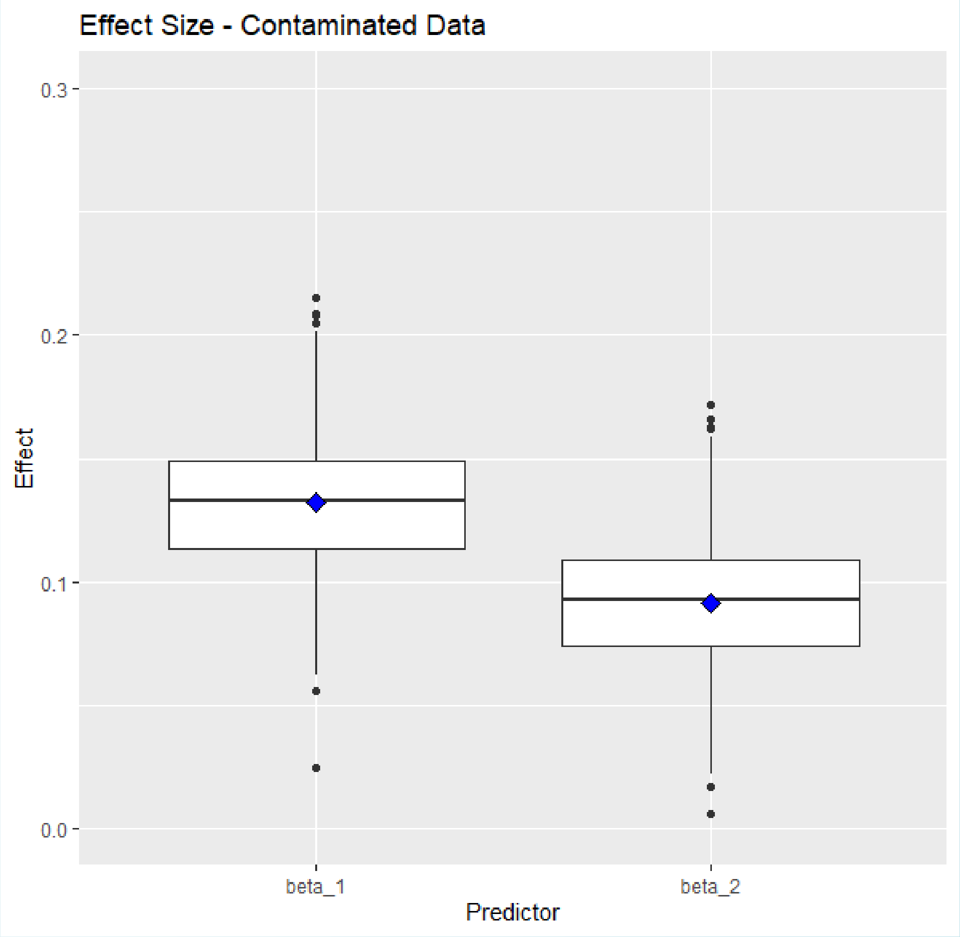
- This is a fairly simple set up

$$Pr(X_1) = a + b_1X_2 + b_2Y_1 + e$$

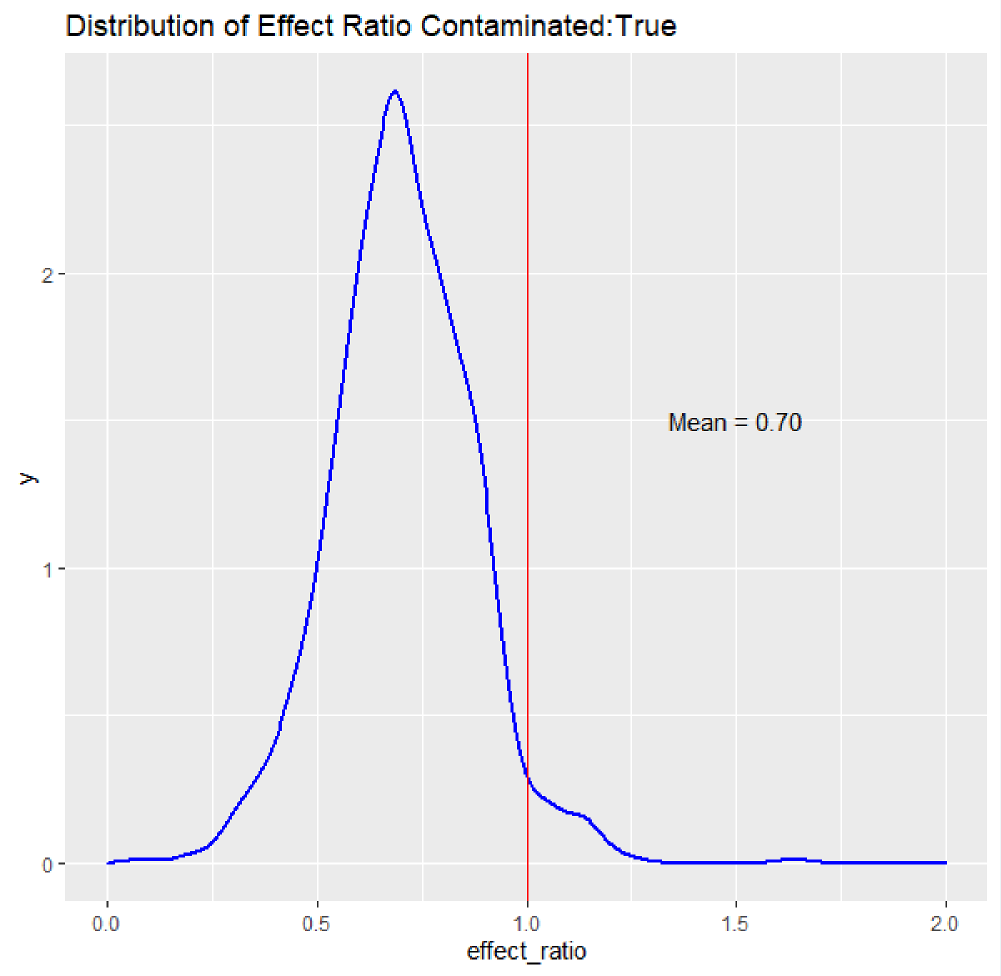
- All of the variables (X_1, X_2, Y_1) are binary variables and $b_1 = b_2$
- For each simulation we generate X_2 and Y_1 and then sample X_1 from a bernouli distribution with the resultant probability.
- We then run 1000 simulations that use logit models to estimate b_1 and b_2
- We then create $\hat{Y}_1$ where we replace 30% of the Y_1 data points with new values drawn from an independent distribution.
- We run the same models with the $\hat{Y}_1$ data. We can compare the effect sizes of the b_2 in both cases to see the impact of the introduced error.

SIMULATION RESULTS





RATIO OF CONTAMINATED EFFECTS TO TRUE EFFECTS



- The average effect in the contaminated data is 70% of the effect in the true data.
- Remarkably this is the equivalent to the percent of data that was correct – i.e. we contaminated 30% of the data

BUT WE DON’T KNOW THE ERROR RATE

- The main issue is that we have two completely separate data sets. We would like to know the error that the fusion induces for potentially many different Y variables.
- Fortunately there is a simple solution.
- We can split the data in the donor data set and fuse one part to the other. This gives us the ability to see how often our fused data matches the original data on any given Y variables.
- This gives us an estimate of how much error we are introducing due to the fusion and therefore tells us how much our correlations or estimates are likely to be attenuated.
- It also gives us a way of fine tuning the fusion by evaluating different groups of Z variables as the matching variables by minimizing the error on certain key Y variables or all the Y variables taken together.

TIPS

- I use the R package StatMatch to do the fusion. It has a variety of algorithms that allow you to fine tune the fusion and easily incorporates the Gower distance metric.
- One feature of StatMatch is the specification of donor classes. These are groups of Z variables that are forced to be exact matches.

– For example, specifying age group and gender as donor classes forces any match between recipient and donor to be in the same gender and age group.

– This has two advantages – even though demographics are rarely great matching choices for fusion, end users like the idea of ensuring age and gender matches. More importantly, by limiting the matches to subgroups of donors and recipients, the size of the distance matrix that needs to be calculated is greatly reduced, which speeds up the process dramatically.

TIPS(2)

- Can we fuse data sets where the Z variables are not really the same?

– Examples might be TV metered data and a Survey, or Website metered data and a Survey, or GeoLocation Data and a Survey.

– You cannot ask someone in a survey how many seconds they were tuned to CNN in the past 24 hours and expect to get a good answer. You can ask them if they watched CNN in the past month and get a reasonably accurate answer. In your TV data you can sum the seconds of CNN watched in 30 days and select some minimum threshhold that should mean a user actually would remember watching the channel.

– Do this for a variety of channel types and you will have a pretty good correspondence between survey respondents and metered televisions.

TIPS(3)

- What if my two data sets are dramatically different in size?

– Hopefully you can be using the smaller data set as your recipients. Then you can limit the donor records to be used only once.

– This may not be the case. Your donor data set may be much smaller than your recipient data. This means that donor records may be used more than once which will have a tendency to reduce the variance of the Y variables in the fused data set. StatMatch has methods to limit the reuse in the matching.

– In any case, by fusing the donor data to itself as described earlier, in similar proportions, you can evaluate the impact on the accuracy of your final fused data set.

SUMMARY

- Data Fusion can be a viable tool for performing inference using variables from different data sets.
- Fusion is not magic. It introduces error into the combined data set and you would be wise to evaluate and account for that increased error.
- Your Z (or matching) variables need to be carefully thought out. Poor choices lead to random matching which leads to spurious results.
- But - if you have been careful, you can learn things that were impossible to learn from separate data sets.

QUESTIONS?

- For a copy of the slides email me at Mike@56stats.com
- or go to my website https://56stats.com

THANK YOU!

SCAN ME