Testing Model-Driven Hypotheses with Big Data

Mike W.-L. Cheung
Department of Psychology
National University of Singapore

June 2018

What is Big Data?

- Information companies, e.g., Google, Facebook, and Twitter, store huge amount of data (in terms of exabytes or billions of gigabytes).
- Big data are used in everywhere in making predictions, e.g., fare estimation in Uber, products/friends/twits recommendations in Amazon/Facebook/Twitter.
- Some of these data have also been used in academic research.[1], [2], [3]

---

[1] Lazer, D., Kennedy, R., King, G., & Vespignani, A. (2014). The parable of Google Flu: Traps in big data analysis. *Science*, 343(6176), 1203-1205.

[2] Back, M. D., Stopfer, J. M., Vazire, S., Gaddis, S., Schmukle, S. C., Egloff, B., & Gosling, S. D. (2010). Facebook profiles reflect actual personality, not self-idealization. *Psychological Science*, *21*(3), 372-374.

[3] Broniatowski, D. A., Paul, M. J., & Dredze, M. (2014). Twitter: Big data opportunities. *Science*, *345*(6193), 148.

Open Data Initiatives

- Many governments, e.g., https://govdata.de and https://www.data.gov, make several thousand datasets freely available for download.
- These data sets power many of the mobile Apps.
- There are also open data movements in academic research, e.g., Open Science by ZPID.

Traditional large data sets

- Large (not yet big) data sets in social and behavioral sciences, e.g., World Values Survey, the International Social Survey Programme, the Longitudinal Study of American Youth, the International PISA study, and GESIS, are freely available to researchers.
- These big data sets provide a vast amount of data to researchers.

Main goals of today's talk

- Discuss the roles of psychology and social sciences in the Big Data movement.
- Introduce a model-driven approach to analyze Big Data, which is based on the joint work with Suzanne Jak.[4]
- Extend traditional multivariate techniques, such as regression, path model, and structural equation modeling to Big Data.

[4] Cheung, M. W.-L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology*, *7*(738).

What is Big Data?

- There is no precise definition of what Big Data are.
- It is all relative to the hardware, software, and programming skills available to researchers.
- For example, the specifications of some of my computers:
  - An old notebook: 4 cores of 1.8GHz with 4 GB memory
  - A workstation in my lab: 48 cores of 2.8GHz with 64 GB memory
- If we use High-Performance Computing (HPC) or cloud computing, we may handle huge data sets without any problems.
- This talk focuses on the scenarios that we are interested in analyzing reasonably large data with our existing computers.

Definition of Big Data

- Laney (2001) was probably the first to describe big data with the 3 Vs:[5]
  - **Volume**: size of the data set may lead to problems with storage and analysis;
  - **Velocity**: data that come in at a high rate and have to be processed within a short period;
  - **Variety**: data consisting of many types, often unstructured, such as mixtures of text, photographs, videos, and numbers.

[5] Laney, D. (2001). 3D Data Management: Controling data volume, velocity, and variety

- A fourth V that is often mentioned is **Veracity**, that indicates the importance of the quality (or truthfulness) of the data.
- Other **V**s have also been suggested, e.g., **value** and **validity**. Some say that there are at least 42 **V**s![6]
- In this study, we mainly focus the 2 Vs of *volume* and *veracity*.

[6] https://www.elderresearch.com/blog/42-v-of-big-data

Two traditions of data analysis

- **Statistics** (traditional statistical tools), e.g., regression, structural equation modeling, and multilevel models, emphasize explanations.
- **Computer science** (data analytics), e.g., decision tree, random forests, deep learning, and TensorFlow, solely focus on predictions.
- There are tensions between these two traditions.[7], [8]

[7] Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

[8] Hofman, J. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488.

Are statisticians involved in the Big Data movement?

- Statisticians are worried about their roles in the Big Data movement.
- Here are some titles in the AMStatNews, the membership magazine of the American Statistical Association:
  - "Aren't We Data Science?"
  - "Statistics Losing Ground to Computer Science"
  - "ASA Statement on the Role of Statistics in Data Science"
  - "The Identity of Statistics in Data Science"
  - "Data Science: The Evolution or the Extinction of Statistics?"

How about psychology and social sciences?

- Psychologists are currently not the key players in the Big Data movement.
- Strengths of researchers in social and behavioral sciences:
  - Psychological theories to understand behaviors and cognitive processes;
  - Advanced multivariate techniques, e.g., psychometric theories, meta-analysis, multilevel modeling, and structural equation modeling.
- Weaknesses of researchers in social and behavioral sciences when compared to computer scientists:
  - Lack of computing and programming skills to handle big data;
  - Not well-trained in the concepts of accuracy in predictions (training vs. validation data).

What can we contribute to the development of theories with big data?

- We need to learn more about in statistics and programming!
- However, we should also focus on our strengths:
  - Develop substantive theories (content) with big data;
  - Test theories with multivariate statistics.
- To achieve these goals, we need to extend our tools to big data.

Common approaches to handle big data

- Computer sciences: MapReduce, Apache Hadoop, divide and conquer approach
- R: Split-analyze-apply[9], [10]
- Limitations:
  - These approaches focus on simple tasks;
  - They are not meant to handle complex statistical modelings such as path models and structural equation modeling.

[9] Wickham, H. (2011). The split-apply-combine strategy for data analysis. *Journal of Statistical Software*, *40*(1), 1–29.

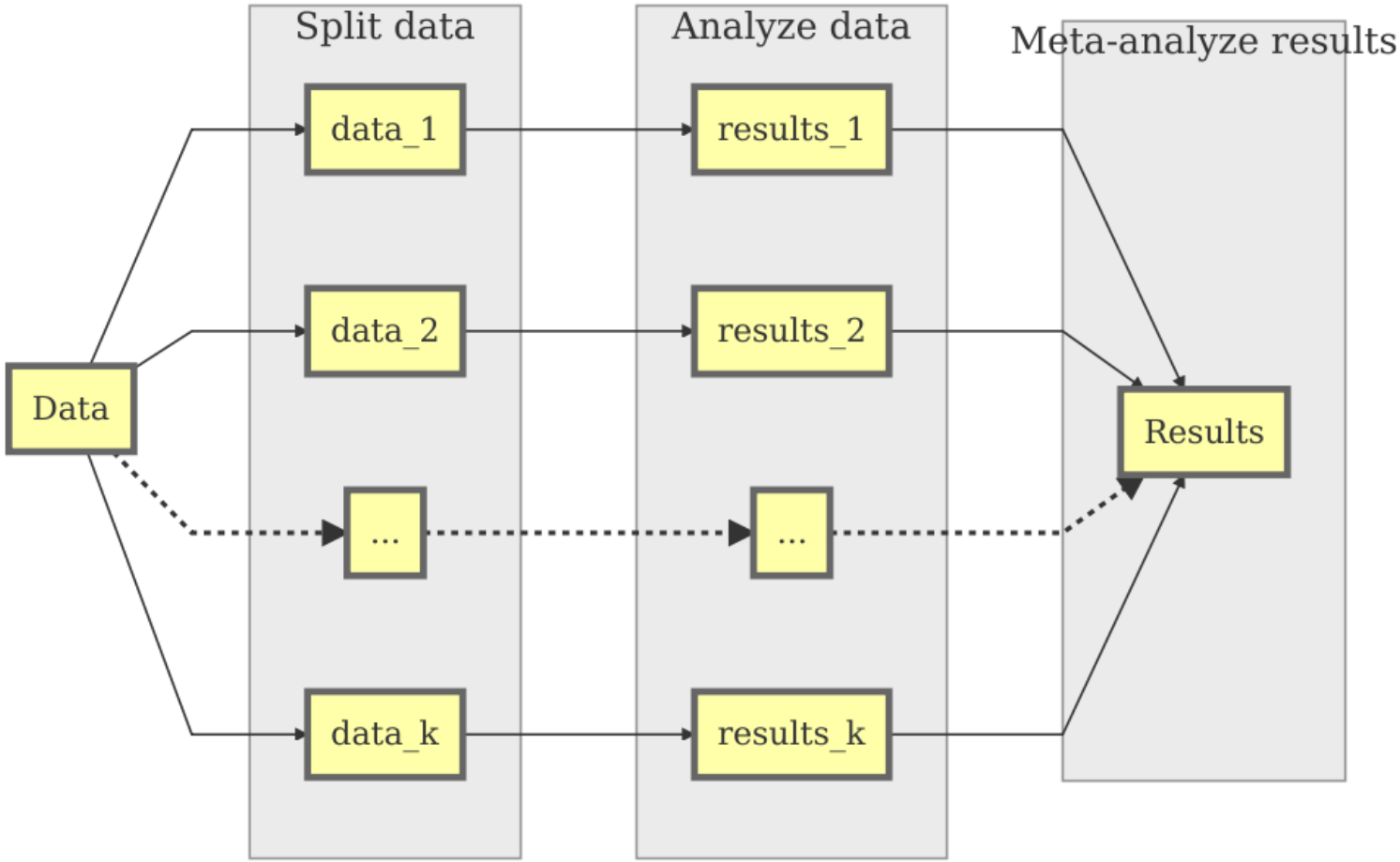
[10] Matloff, N. (2016). Software Alchemy: Turning complex statistical computations into embarrassingly-parallel ones. *Journal of Statistical Software*, *71*(4), 1–15.

An alternative approach

- A Split/Analyze/Meta-analyze (SAM) Approach:[11]
  - It is conceptually similar to the conventional approaches to handling big data;
  - The main difference is that we utilize **meta-analysis** in the last step;
- Advantages:
  - Conventional multivariate techniques can be applied to big data.
  - Researchers can analyze big data from a theory testing approach.

[11] Cheung, M. W.-L., & Jak, S. (2016). Analyzing big data in psychology: A split/analyze/meta-analyze approach. *Frontiers in Psychology*, *7*(738).

Conceptual model



Step 1: Splitting data

- **Random split**:
  - Randomly split the data into k pseudo "studies";
  - Pseudo "studies" are direct replicates of each other;
  - Pseudo "studies" are only different due to the sampling error;
  - A fixed-effects meta-analysis is used to combine the results.
- **Stratified split**:
  - Split the data according to some existing characteristics, e.g., geographic locations, time, into k pseudo "studies";
  - Pseudo "studies" may be different beyond sampling error;
  - Study characteristics may be used to explain these differences;
  - A random- and mixed-effects meta-analyses may be used to combine the results.

Step 2: Analyzing data as separate "studies"

- For example, regression, reliability analysis, path analysis, multilevel models, confirmatory factor analysis, or even structural equation modeling.
- Suppose there are p parameter estimates; we obtain the $p \times 1$ effect sizes $\boldsymbol{y}_i$ in the ith study with its $p \times p$ sampling variance-covariance matrix $\boldsymbol{V}_i$.

Step 3: Meta-analyzing the results (1)

- Apply univariate and multivariate meta-analyses, and meta-analytic structural modeling (MASEM) to synthesize the results;[12]
- **Random split**: Fixed-effects meta-analysis for data with only the sampling error:
  - $\boldsymbol{y}_i = \beta_{Fixed} + \boldsymbol{e}_i$ with $\boldsymbol{e}_i \sim \mathcal{N}(0, \boldsymbol{V}_i)$
  - β_{Fixed} is the vector of the common population effect size under a fixed-effects model;
  - $\boldsymbol{V}_i$ is the conditional sampling covariance matrix of $\boldsymbol{y}_i$;

[12] Cheung, M. W.-L. (2015). *Meta-analysis: A structural equation modeling approach*. Chichester, West Sussex: John Wiley & Sons.

Step 3: Meta-analyzing the results (2)

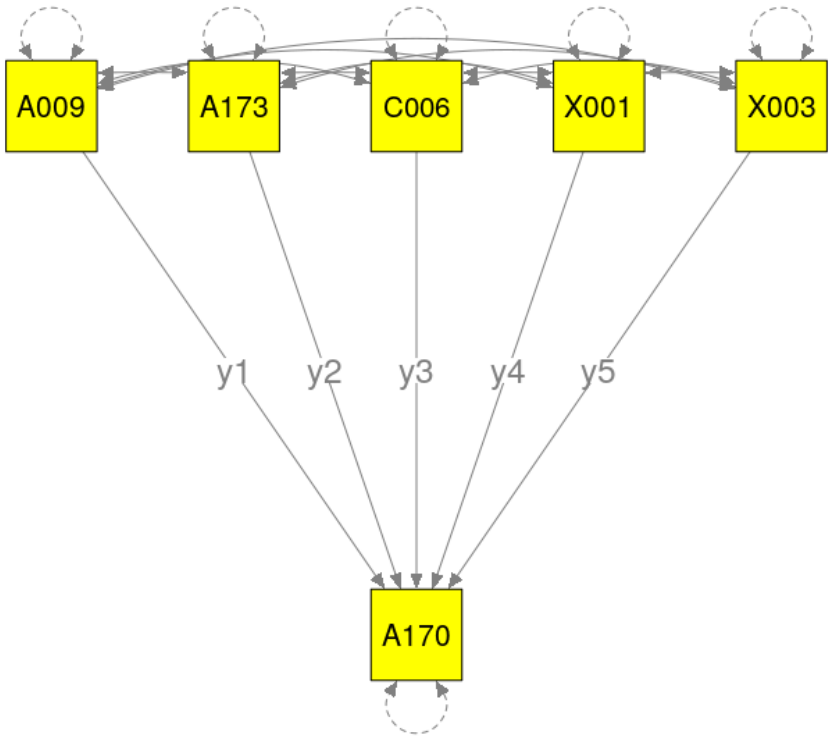
- **Stratefied split**: Random-effects meta-analysis for data with both sampling error and true differences on the population parameters:
  - $\boldsymbol{y}_i = \beta_{Random} + \boldsymbol{u}_i + \boldsymbol{e}_i$ with $\boldsymbol{e}_i \sim \mathcal{N}(0, \boldsymbol{V}_i)$ and $\boldsymbol{u}_i \sim \mathcal{N}(0, \boldsymbol{T}^2)$
  - β_{Random} is the vector of the average population effect size under a random-effects model;
  - $\boldsymbol{V}_i$ is the conditional sampling covariance matrix of $\boldsymbol{y}_i$;
  - $\boldsymbol{T}^2$ is the heterogeneity variance of the random effects.

Example 1: World Values Survey

- 343,309 participants on 1,377 variables spanning 100 regions with 6 waves.
- This is only a *large*, but not *big*, data set.
- The illustrations show how the SAM approach can be used to analyze similar large data sets in psychology and social sciences.
- The data were split by countries and waves. There were 239 *studies* with 240 to 6,025 respondents per study.
- We will focus on illustrating the proposed approach rather than on interpreting the results substantively.

Multiple regression (1)

- **Dependent variable**: life satisfaction (A170)
- **Independent variables**: subjective state of health (A009), freedom of choice and control (A173), financial satisfaction (C006), sex (X001), and age (X003).
- **Effect sizes**: Regression coefficients (y1 to y5).



Multiple regression (2)

- There is a total of 239 *studies* with some missing data.
- Here are the first few rows of the effect sizes after running the regression analysis in each "study".

```
##   Wave        y1        y2        y3        y4         y5          v11
## 1    1 0.3173139 0.3279715 0.3087195 0.1167210 0.06480454 0.004959965
## 2    1 0.2875391 0.2462064 0.2125849 0.2149695 0.07156565 0.002218773
## 3    1        NA        NA        NA        NA         NA           NA
## 4    1        NA        NA        NA        NA         NA           NA
## 5    1 0.2398536 0.1521864 0.4860760 0.3525533 0.09317681 0.003821318
## 6    1        NA        NA        NA        NA         NA           NA
##
##              v21           v31          v41           v51          v22
## 1 -0.0001420895 -0.0002418505 6.136374e-04 0.0005859175 0.0008974672
## 2 -0.0001046037 -0.0001085599 6.129034e-05 0.0002894008 0.0004612145
## 3             NA            NA           NA            NA           NA
## 4             NA            NA           NA            NA           NA
## 5 -0.0001484672 -0.0002370722 6.355607e-04 0.0003563672 0.0005892315
## 6             NA            NA           NA            NA           NA
##
##              v32           v42           v52          v33           v43
## 1 -2.836507e-04 -0.0001485246 -5.196734e-05 0.0007291665  1.310907e-04
## 2 -8.993028e-05 -0.0000151603 -1.167143e-05 0.0004244703  4.266898e-05
## 3             NA            NA           NA           NA           NA
## 4             NA            NA           NA           NA           NA
## 5 -1.055706e-04  0.0000329972  1.897503e-04 0.0005549081 -5.228161e-05
## 6             NA            NA           NA           NA           NA
```

Multiple regression (3)

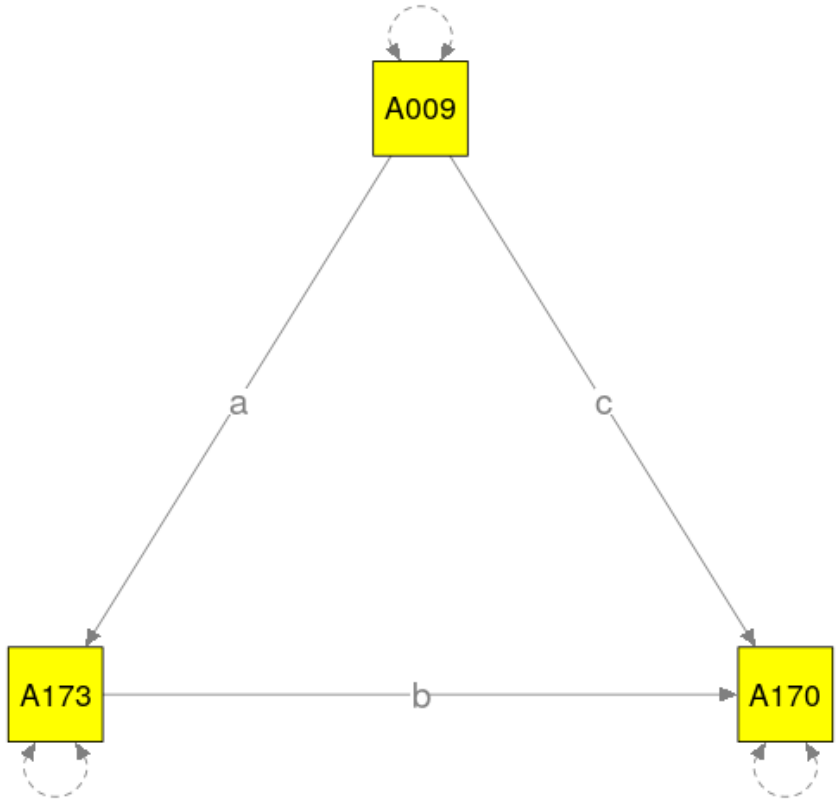
- The effect sizes were used to fit a multivariate meta-analysis.
- The average regression coefficients and their 95% confidence intervals (CIs):
- $$\begin{bmatrix} \beta_{Subjective\ state\ of\ health} \\ \beta_{Freedom\ of\ choice\ and\ control} \\ \beta_{Financial\ satisfaction} \\ \beta_{Sex} \\ \beta_{Age} \end{bmatrix} = \begin{bmatrix} 0.40(0.38, 0.42) \\ 0.21(0.19, 0.22) \\ 0.38(0.36, 0.40) \\ 0.13(0.11, 0.15) \\ 0.03(0.03, 0.04) \end{bmatrix}.$$
- The parameter estimates are quite precise because of the large sample size (N=343,309).

Multiple regression (3)

- There are variations on the regression coefficients across countries and waves. We include *Wave* as a moderator:
  - $$\begin{bmatrix} \beta_{\textit{Subjective state of health}} \\ \beta_{\textit{Freedom of choice and control}} \\ \beta_{\textit{Financial satisfaction}} \\ \beta_{\textit{Sex}} \\ \beta_{\textit{Age}} \end{bmatrix} = \begin{bmatrix} \mathbf{0.03(0.02, 0.05)} \\ -0.01(-0.02, -0.00) \\ -0.02(-0.04, -0.01) \\ \mathbf{0.00(-0.01, 0.02)} \\ -0.01(-0.01, -0.00) \end{bmatrix} \times \textit{Wave}.$$
  - R^2 explained by the *wave*: $\begin{bmatrix} \mathbf{0.10} \\ 0.02 \\ 0.04 \\ \mathbf{0.00} \\ 0.03 \end{bmatrix}$.
  - The regression coefficient of *Subjective state of health* gets larger over wave whereas the regression coefficient of *Sex* is stable over time (no linear relationship).

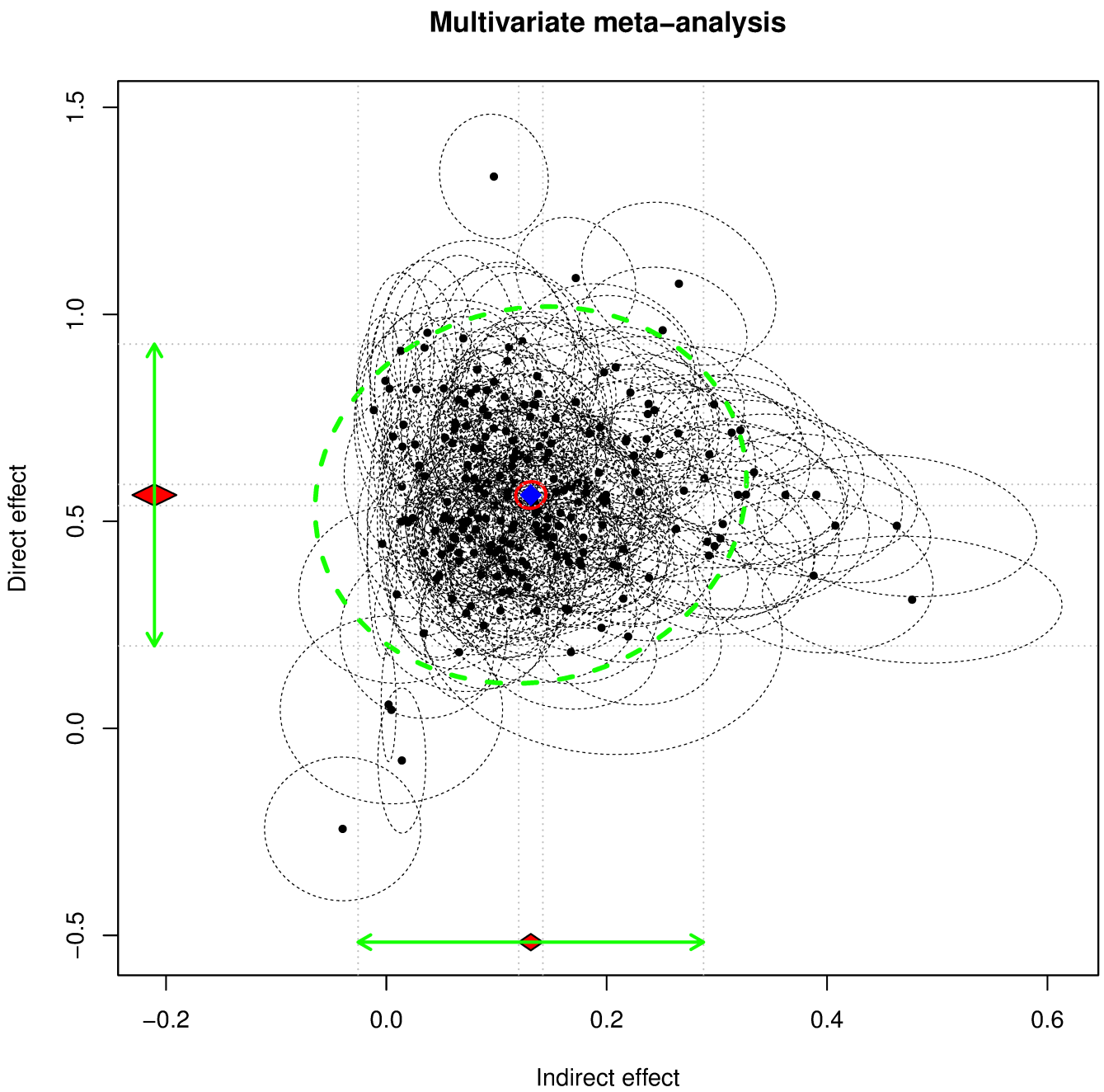
Mediation analysis (1)

- **Dependent variable**: life satisfaction (A170)
- **Mediator**: freedom of choice and control (A173)
- **Predictor**: subjective state of health (A009)
- **Effect sizes**: Indirect effect ($a * b$) and direct effect (c)



Mediation analysis (2)

- The estimated coefficients and their 95% CIs:
  - $\begin{bmatrix} \textit{Indirect effect} \\ \textit{Direct effect} \end{bmatrix} = \begin{bmatrix} 0.13(0.12, 0.14) \\ 0.57(0.54, 0.59) \end{bmatrix}$,
  - $T^2 = \begin{bmatrix} 0.01 & \\ 0.00 & 0.04 \end{bmatrix}$, and
  - $I^2 = \begin{bmatrix} 0.96 \\ 0.90 \end{bmatrix}$.
- The direct effect is much larger than the indirect effect.


Multivariate meta−analysis
Direct effect
Indirect effect
1.5
1.0
0.5
0.0
−0.5
−0.2
0.0
0.2
0.4
0.6

Confirmatory factor analysis (1)

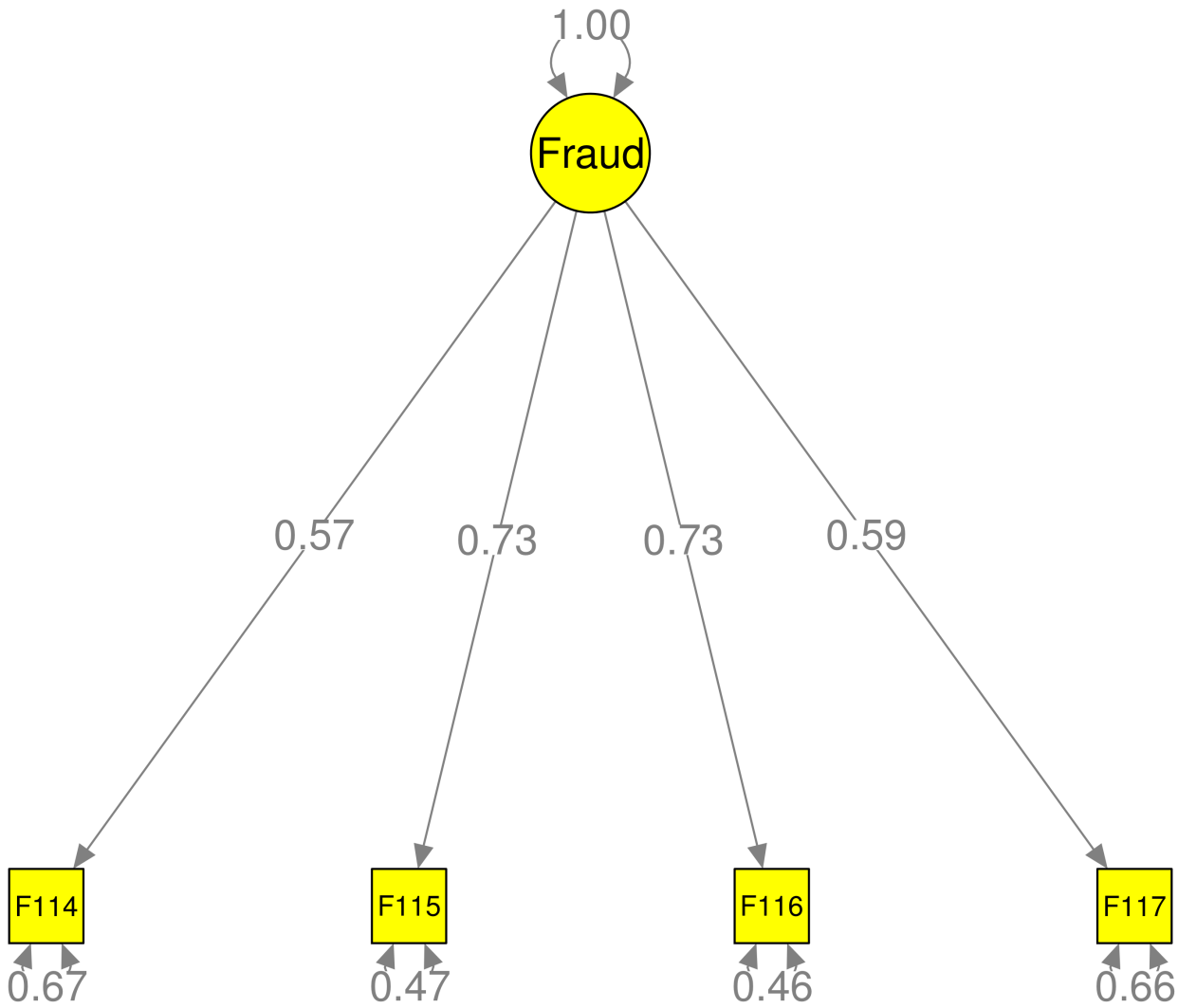
- The SAM approach also allow us to verify the quality of data.
- Four items on the attitude of **fraud**, about whether it is justifiable to:
  - claim government benefits to which you are not entitled (F114);
  - avoid paying a fare on public transport (F115);
  - cheat on taxes (F116);
  - accept a bribe from someone in the course of carrying out one's duties (F117).
- The random-effects two-stage SEM (TSSEM) was used to test the proposed model.[13]

[13] Cheung, M. W.-L. (2014). Fixed- and random-effects meta-analytic structural equation modeling: Examples and analyses in R. *Behavior Research Methods*, *46*(1), 29–40.

Confirmatory factor analysis (2)

- Stage 1 model: The correlation matrices are heterogeneous with $I^2 \approx .98$.
- Stage 2 model: The proposed model fits the data reasonable well with $\chi^2(df = 2) = 333.92, p < .001$, CFI=.9334, RMSEA=.0230, and SRMR=.0472.

- All the factor loadings are very high.



Reliability generalization

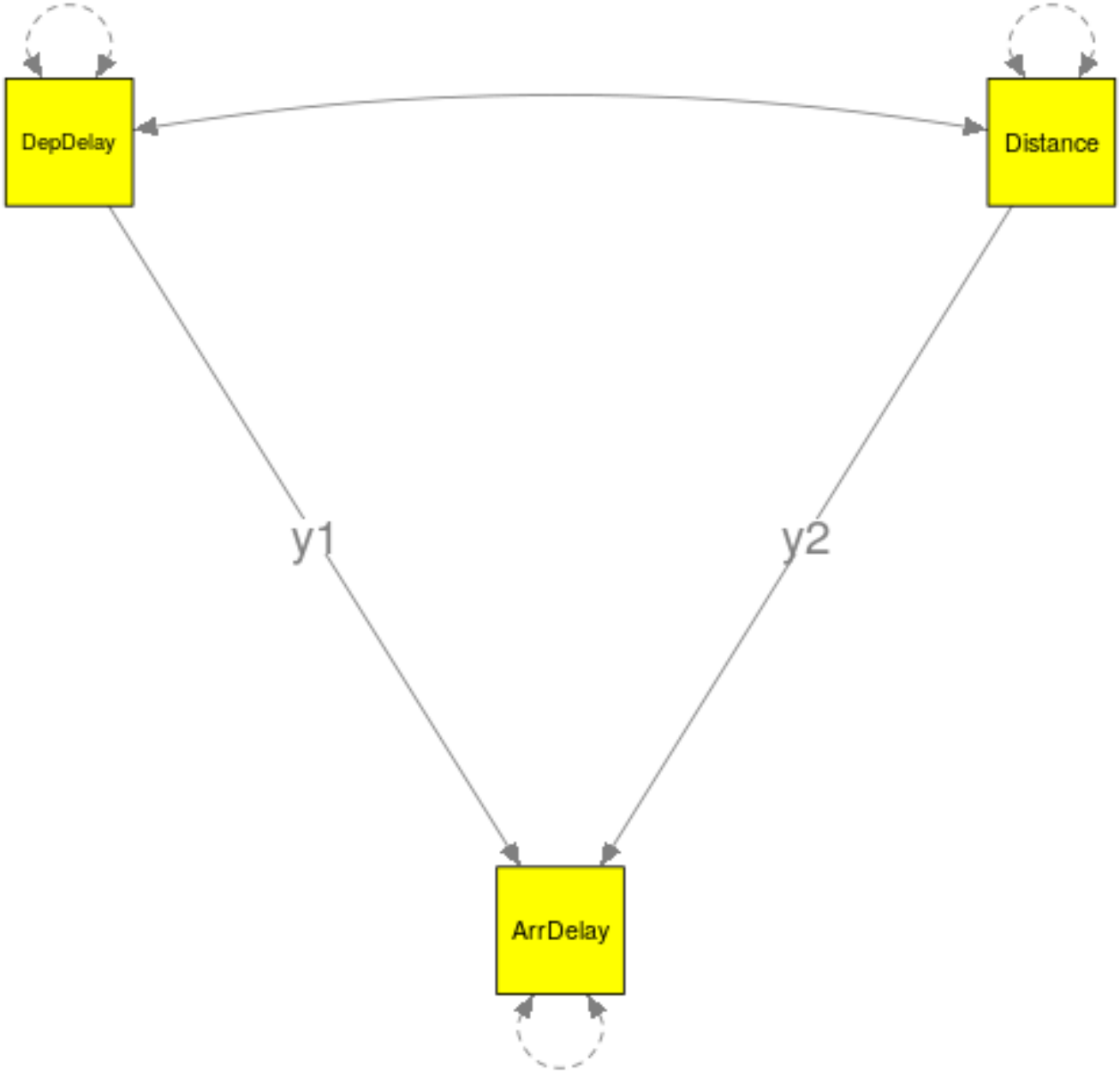
- If we are going to use the **fraud** scale, we may want to check whether the reliability can be generalized across countries and waves.
- The estimated reliability coefficient (coefficient alpha) and its 95% CIs:
  - Average reliability coefficient= 0.68 (0.66, 0.70),
  - $T^2 = 0.02$, and
  - $I^2 = 1.00$.

Example 2: Airlines Data

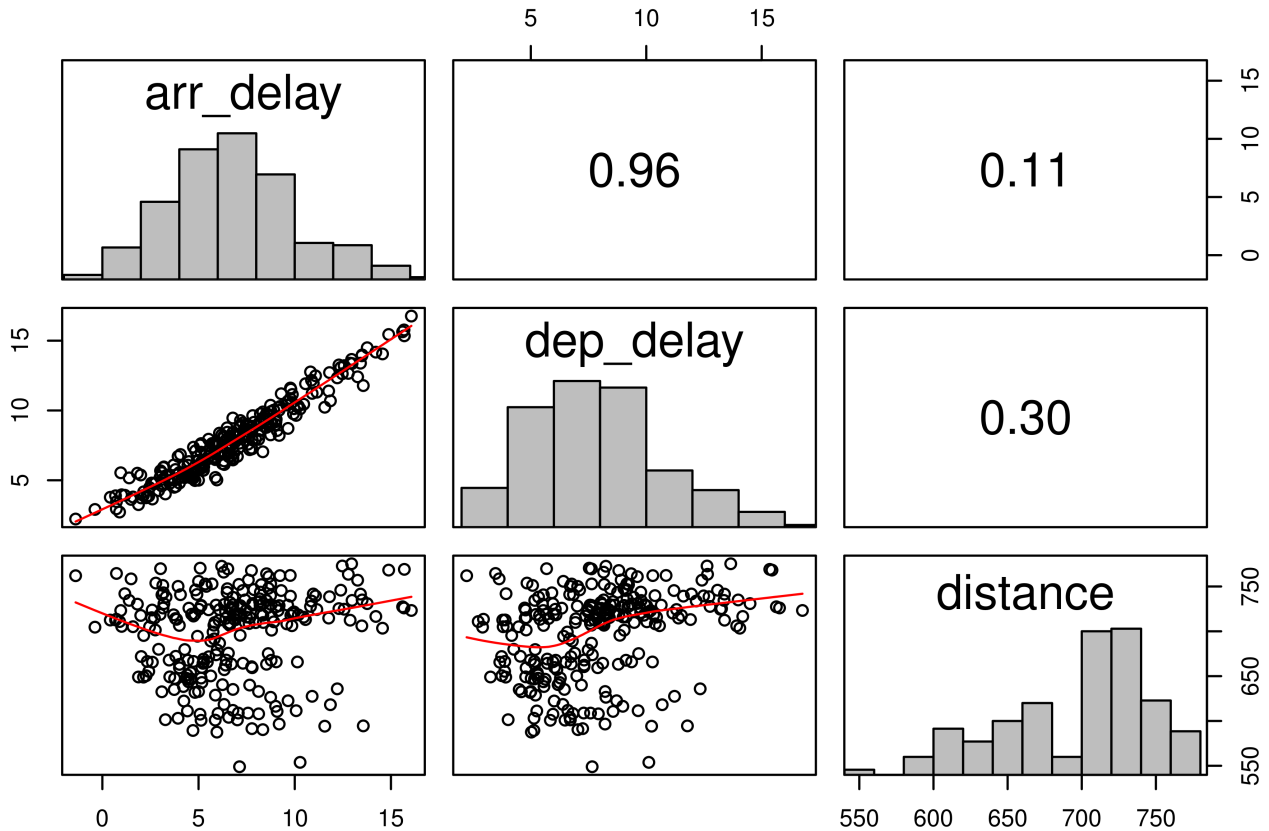
- The airlines dataset contains data on 29 variables from more than 123 million flight records for almost all arrivals and departures at airports in the USA from 1987 to 2008.
- The sizes of the compressed files and the uncompressed files are 1.7 GB and 12 GB, respectively.
- It is usually used as a sample data set for big data analysis.
- The data set was split by years. There were a total of 22 pseudo "studies," with sample sizes ranging from 1,311,826 flights in 1987 to 7,453,215 flights in 2007.

Research questions

- **Dependent variable**: Arrival delay time
- **Independent variables**: Departure delay time and distance
- **Research hypotheses**:
  - The departure delay time was positively related to arrival delay time;
  - The distance between the airports was negatively related to arrival delay time.
- A random-effects model is fitted to account for the nested structure of the data. The seasonal variation is approximately accounted for by considering the data nested within **Month**, **Day of Month**, **Day Of Week**, while geographical differences is approximately accounted for by considering the data nested within **origin** and **destination** airports.

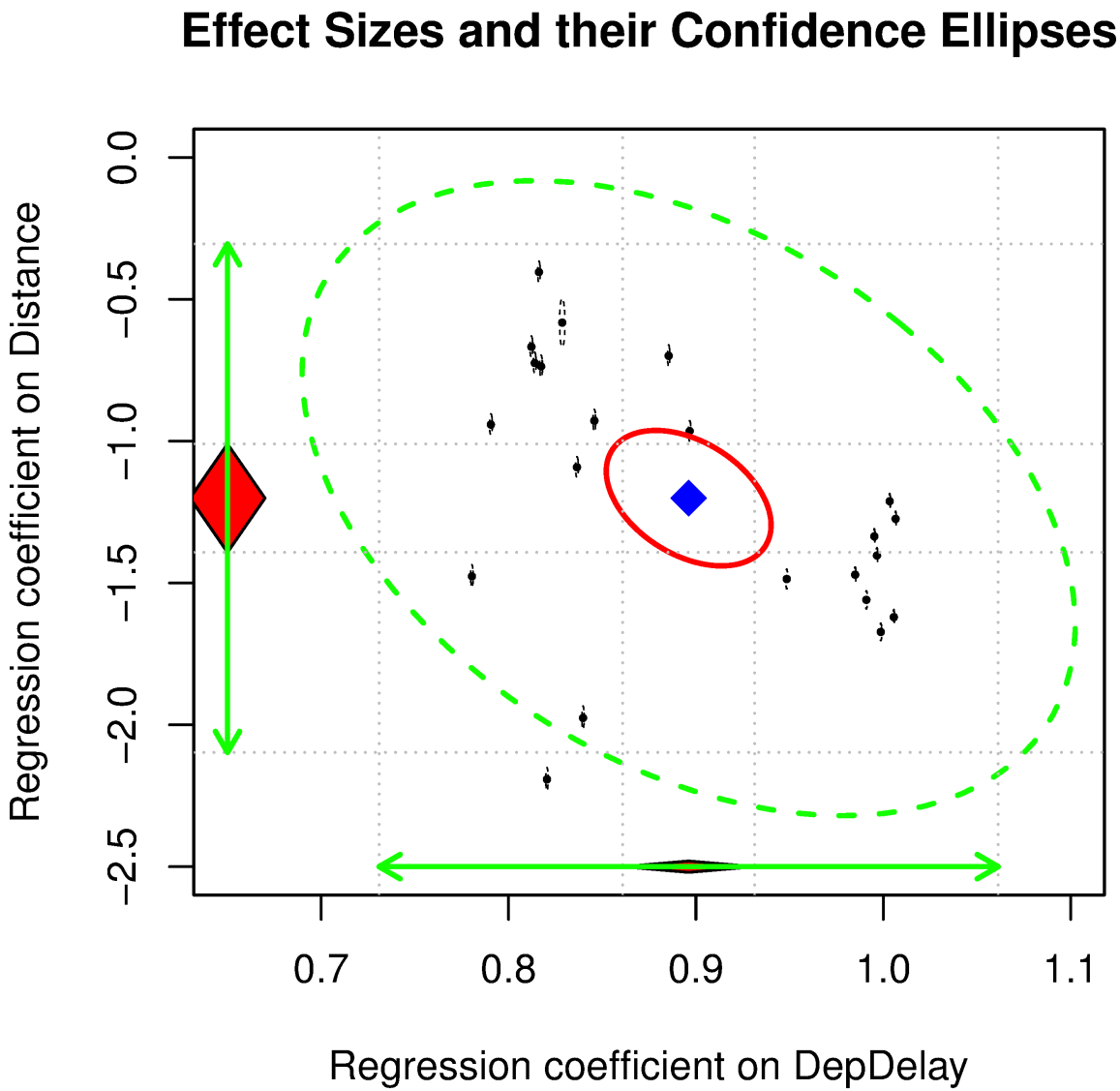

DepDelay
Distance
y1
y2
ArrDelay

Summary of the data



Results

- The estimated coefficients and their 95% CIs:
  - $\begin{bmatrix}\beta_{Dep}\\ \beta_{Dist}\end{bmatrix} = \begin{bmatrix}0.90(0.86, 0.93)\\ -1.20(-1.39, -1.00)\end{bmatrix}$,
  - $T^2 = \begin{bmatrix}0.01 & \\ -0.02 & 0.21\end{bmatrix}$, and
  - $I^2 = \begin{bmatrix}1.00\\ 1.00\end{bmatrix}$.
- The findings are consistent with our hypotheses.


Effect Sizes and their Confidence Ellipses
Regression coefficient on Distance
0.0
−0.5
−1.0
−1.5
−2.0
−2.5
0.7
0.8
0.9
1.0
1.1
Regression coefficient on DepDelay

How many pseudo "studies" do we need in the random split (fixed-effects meta-analysis)?

- The total no. of data $\approx$ no. of "studies" x sample size per study
- Although we do not have a formal proof here, results based on different numbers of studies are comparable.

TABLE 1 | Comparisons between analysis of raw data, and analysis based on a fixed-effects meta-analysis with random splits.

Numbers of studies	Raw data ($k = 1$)	$k = 5$	$k = 10$	$k = 50$	$k = 100$	$k = 500$	$k = 1000$
REGRESSION COEFFICIENTS							
Subjective state of health (A009)	0.4333	0.4333	0.4333	0.4333	0.4334	0.4330	0.4336
Freedom of choice and control (A173)	0.2313	0.2313	0.2313	0.2313	0.2314	0.2315	0.2322
Financial satisfaction (C006)	0.4243	0.4243	0.4243	0.4244	0.4245	0.4257	0.4259
Sex (X001)	0.1708	0.1707	0.1708	0.1708	0.1705	0.1701	0.1698
Age (X003)	0.0580	0.0580	0.0580	0.0580	0.0581	0.0579	0.0575
STANDARD ERRORS							
Subjective state of health (A009)	0.0043	0.0043	0.0043	0.0043	0.0043	0.0043	0.0043
Freedom of choice and control (A173)	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015	0.0015
Financial satisfaction (C006)	0.0015	0.0015	0.0015	0.0015	0.0015	0.0014	0.0014
Sex (X001)	0.0070	0.0070	0.0070	0.0070	0.0070	0.0069	0.0069
Age (X003)	0.0023	0.0023	0.0023	0.0023	0.0023	0.0022	0.0022

Dependent variable is life satisfaction (A170).

Conclusions

- More and more big data are available to researchers.
- On the one hand, we need to learn more machine learning techniques and programming so that we can communicate with computer scientists.
- On the other hand, we need to demonstrate our unique contributions to big data analysis.
- The Split/Analyze/Meta-analyze (SAM) approach allows researchers to analyze large data sets with the conventional multivariate statistics.

Thank you!

- Comments and questions are welcome!