The Power and the Pitfalls of Multiple Regression Analysis (Part 1)

Smita Skrivanek
MoreSteam.com
January 30, 2013

Agenda

- Welcome
- Introduction of MBB Webcast Series
  - Larry Goldman, MoreSteam.com
- Today's Session
  - Smita Skrivanek, MoreSteam.com
- Open Discussion and Questions

MoreSteam.com

- Founded in 2000
- Trained over 380,000 Lean Six Sigma professionals
- Served over 2,000 corporate customers (including 50+% of the F500)
- First firm to offer the complete Black Belt curriculum online and only firm to offer online DfLSS
- Courses reviewed and approved by ASQ and PMI
- Academic Partnership with Ohio State University

Select Customers:

Today's Presenter

Smita Skrivanek

Principal Statistician, MoreSteam.com

- *Develops content & software, reviews projects, and assists students with questions on advanced statistics*
- *Heads research & development for EngineRoom® software*
- *Masters in Applied Statistics from The Ohio State University and an MBA from Indiana University Kelley School of Business*

Discussion Points

- *Multiple Regression overview - uses and application*
- *Types of data that can be analyzed*
- *Alternative approaches to analysis*
- *Some Pitfalls to understand and workarounds to mitigate their effects*

Multiple Regression Defined

A method for estimating the association between a continuous Dependent (outcome) variable and multiple Independent (predictor) variables using a mathematical model.

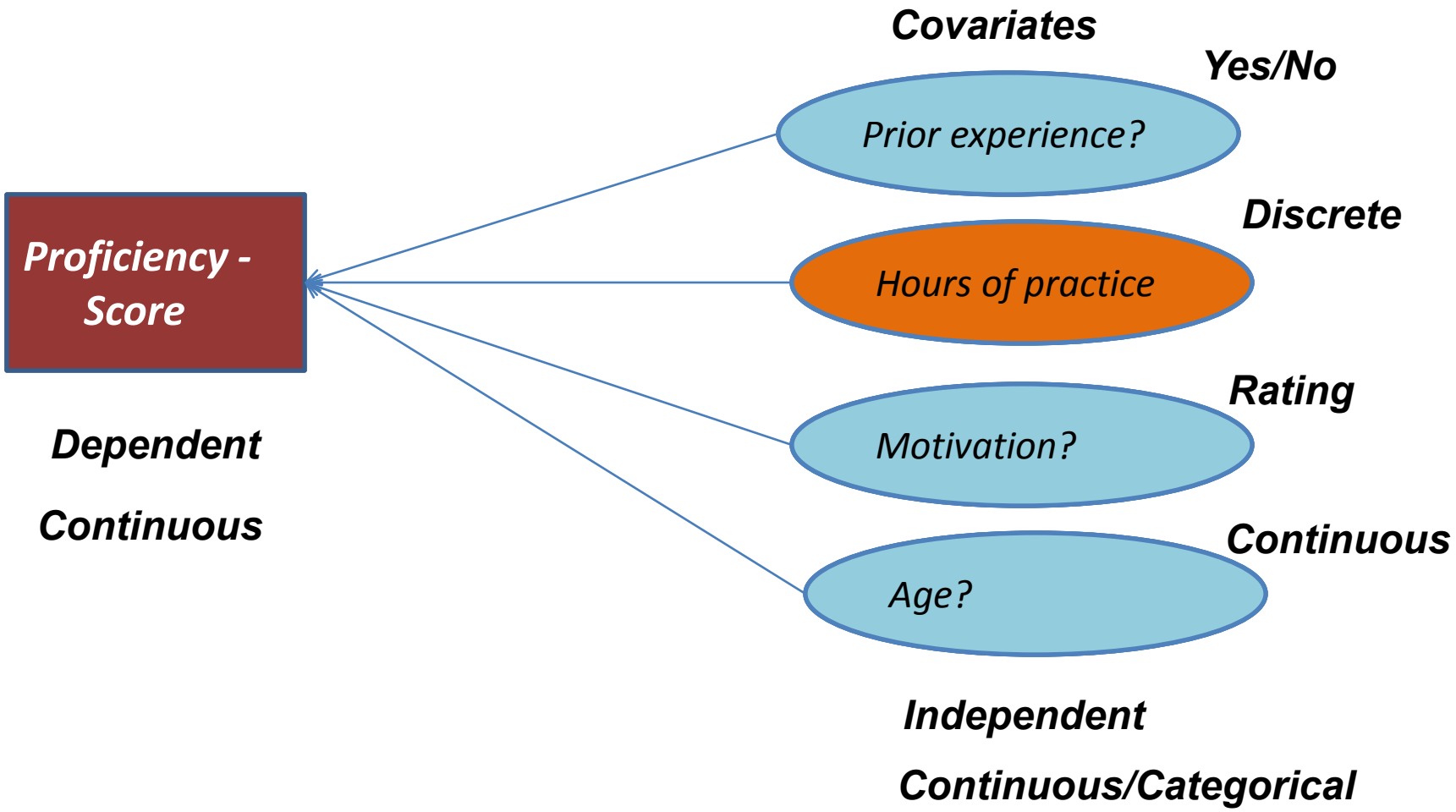
Sir Francis Galton > inheritance in sweet peas (late 19th century)

Linear regression → Correlation!

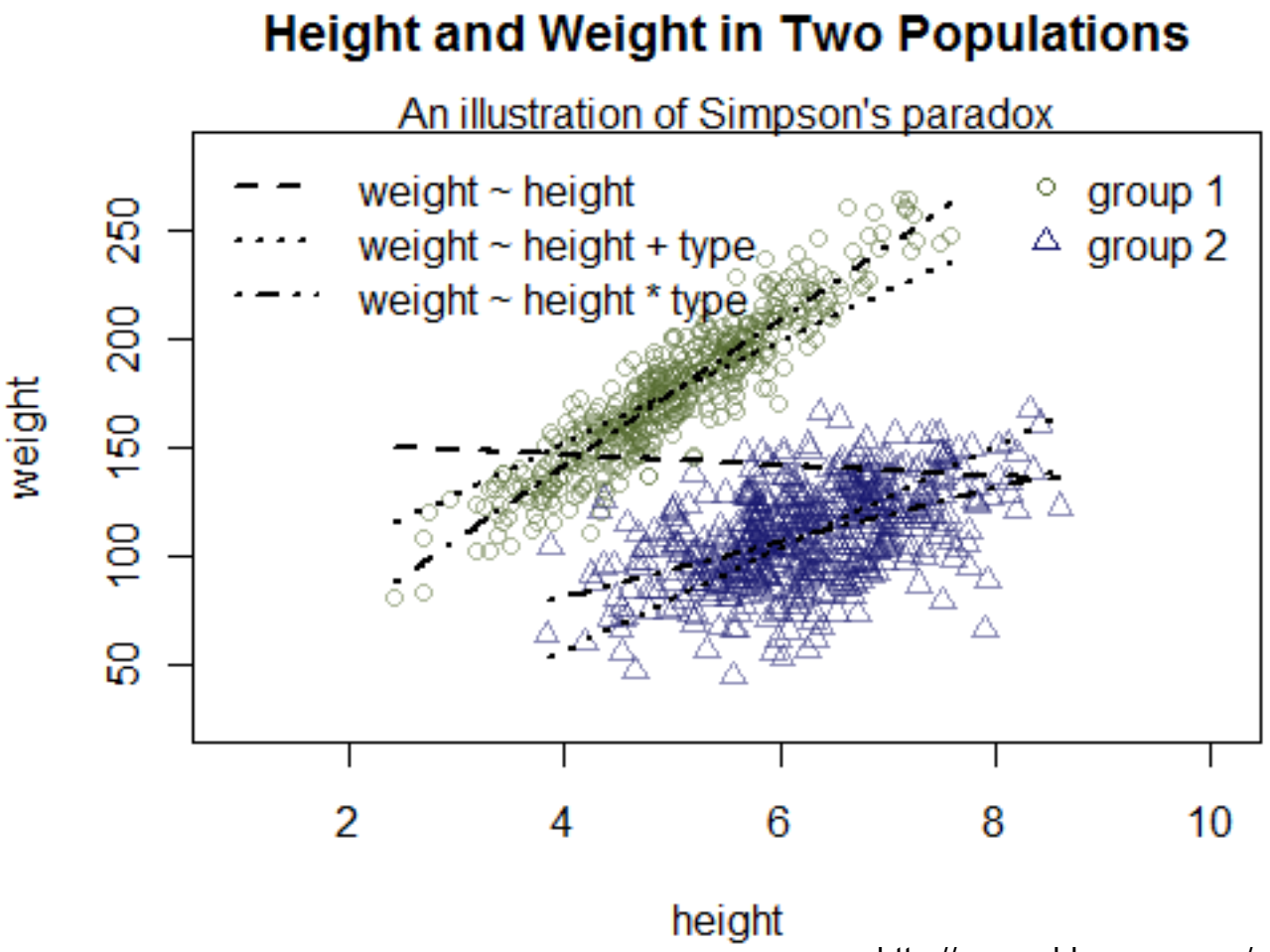
Explain how a system works to produce an observed effect.

Predict how an outcome will change if you introduce a change in one of the factors that influence it, keeping all the other model factors constant.

Overview



Solving Simpson's Paradox via Regression



http://www.r-bloggers.com/simpsons-paradox/

Uses and Applications

Business/Industry

Public Policy

Healthcare

Courts

Education and Social Sciences

Discussion Points

- *Multiple Regression overview - uses and application*
- *Types of data that can be analyzed*
- *Alternative approaches to analysis*
- *Some Pitfalls to understand and workarounds to mitigate their effects*

Types of Data (Y) and MR Analyses

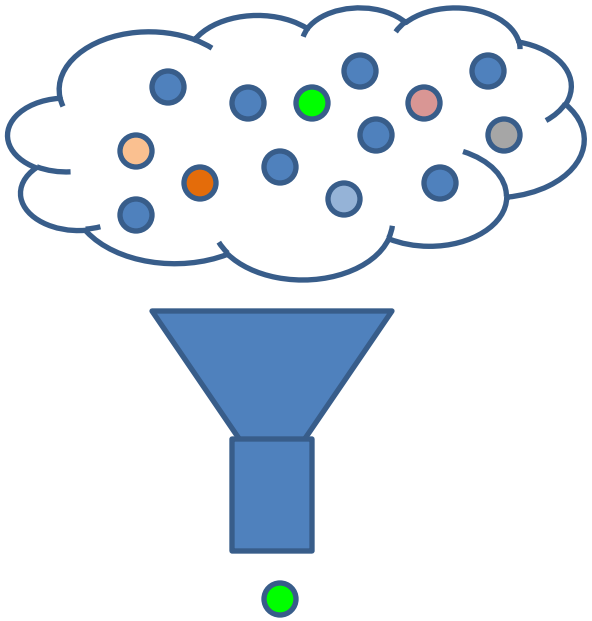
Dependent, Y	MR Analysis
Continuous	Least Squares
Continuous (Elapsed time)	Proportional Hazards (Survival Analysis)
Discrete Counts/Rates	Poisson/Negative binomial /Log-linear
Binary	Logistic
Nominal	Multinomial Logistic

Discussion Points

- *Multiple Regression overview - uses and application*
- *Types of data that can be analyzed*
- *Alternative approaches to analysis*
- *Some Pitfalls to understand and workarounds to mitigate their effects*

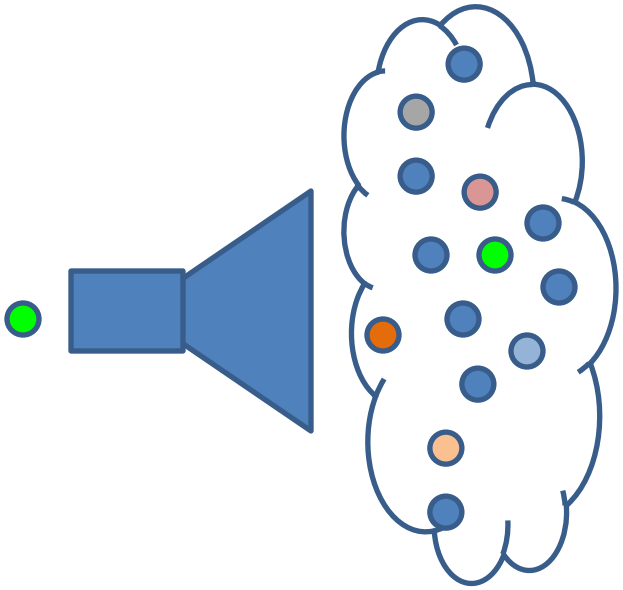
Reasoning

Deductive



Laws, rules, principles

Inductive



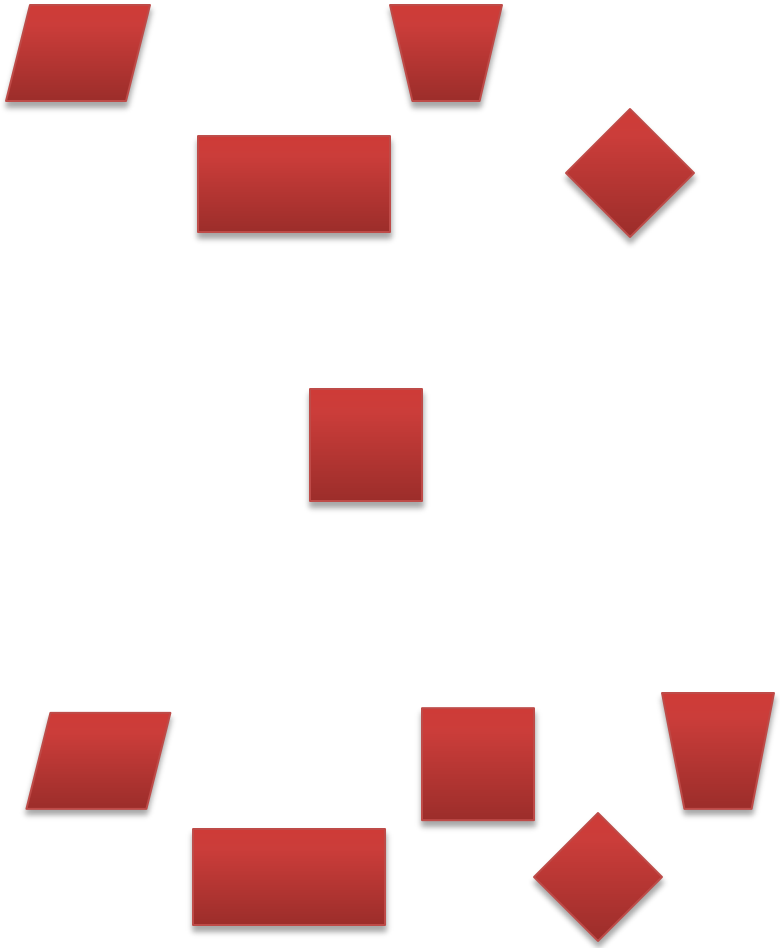
Experience, observation, pattern recognition

Deduction Example

All quadrilaterals have four sides.

A Square has four sides

A Square is a quadrilateral.



Induction Examples

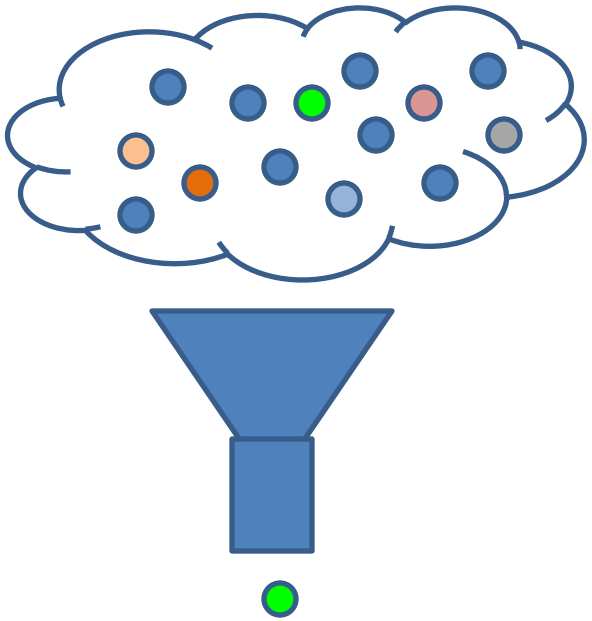
10, 20, 30, 40, 50

Every time I eat shrimp, I get ill. → I am allergic to shrimp.

Everyone I've met at this company is friendly. → Everyone at this company is friendly.

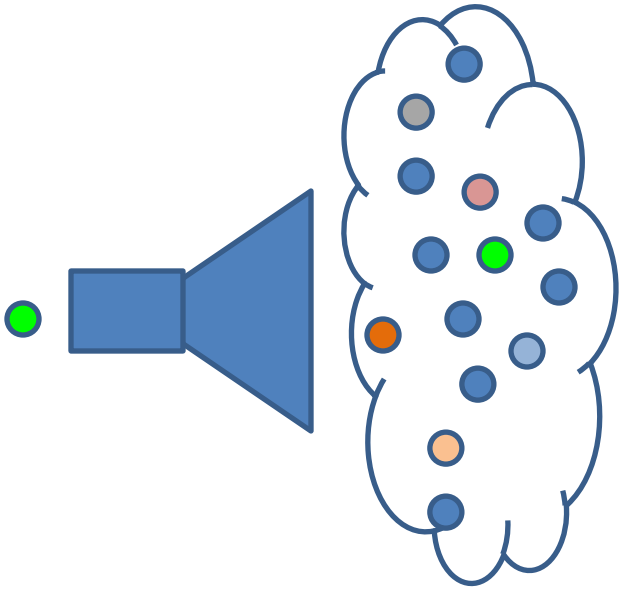
Reasoning

Deductive



Laws, rules, principles

Regression!



Experience, observation, pattern recognition

QUESTIONS?

Pitfalls

- Observational data
- Measurement Error in Variables
- Model misspecification
- Assumption violation
- Multicollinearity
- Extrapolation

Pitfall 1: Observational data

Early language skills reduce preschool tantrums, study finds

People with higher IQs make wiser economic choices

Sincere smiling promotes longevity

Church attendance boosts immunity

Don't be a Super Bowl statistic: Stress of watching the big game can be hazardous to heart, research suggests

(unless your team wins): ***Winning World Cup lowers heart attack deaths***

Source: http://jfmueller.faculty.noctrl.edu/100/correlation_or_causation.htm

Pitfall 1: Observational data

The only way to find out what will happen when a complex system is disturbed is to disturb the system, not merely to observe it passively.

- George Box

Pitfall 2: Variables Measured with Error

Explained variance is incorrectly allocated among the independent variables

Biased regression coefficients (betas)

Inflated Type I and II error rates

What to do:

- Errors-in-Variables Regression:

$$Y = \alpha + \beta F_X + E_Y$$

$$X = F_X + E_X$$

Fx = latent variable
X = proxy of Fx
Ex = Measurement error in X

- Only use the model for prediction; do NOT to try to interpret the regression coefficients!

Pitfall 3: Model Misspecification

Missing influential variables → Simpson's Paradox, ↑ beta weights, type II error

Including non-influential variables → ↓ Precision of results

Overfitting → *Unreliable predictions, ↑ type I error, multicollinearity*

What to do:

1. Ensure association of included X's with Y have sound underpinnings.
2. Collect representative samples.
3. Run multiple models – compare results.

Pitfall 4: Violations of Assumptions

Non-linearity

Non-normality

Non-constant variance

Non-independence

→ *Unreliable standard errors*

→ *Inflated type 1 error*

→ *Unreliable F tests*

What to do:

1. Non-normality, non-constant variance = fairly robust
2. Non-linearity, correlated errors = include higher order terms, time series model.
3. Transform (but make sure the transformed data are interpretable.)/robust methods.

Pitfall 5: Multicollinearity

BAD

Poor interpretation of the model

Unstable Predictions

Large standard errors, inflated Type 2 errors

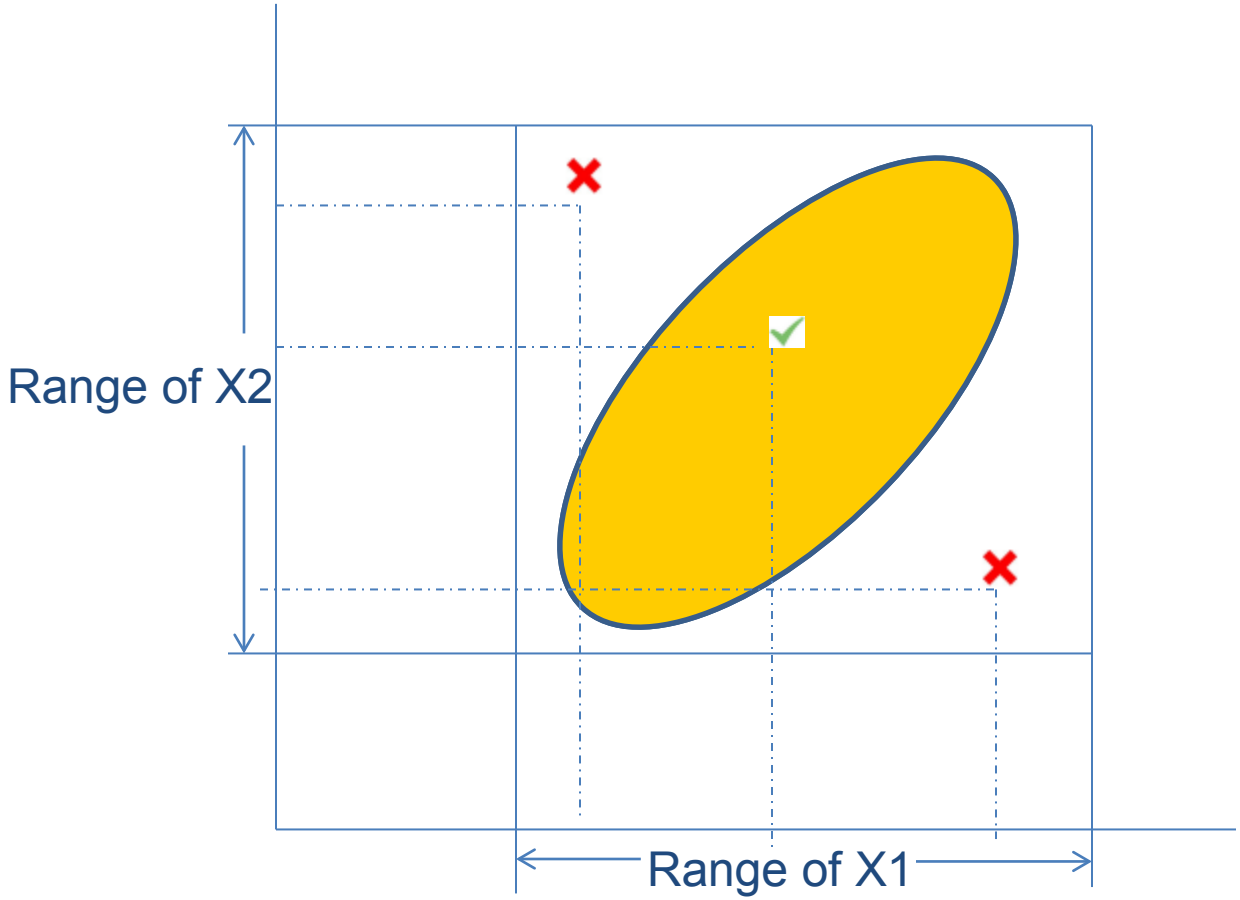
GOOD

Helps simplify the model

Indicates presence of underlying “construct” or “latent” variable

Pitfall 6: Extrapolation

Do not predict too far outside the bounds!



Good Habits

- *Use sound theoretical and empirical principles to drive the selection of variables for study.*
- *Aim for model parsimony (keep it simple!) – complex models are unreliable.*
- *Is the analysis robust? Overfitting and applying regression analysis to a not-truly random sample will result in poor predictions.*
- *The acid test in statistical modeling is prediction. Is it verifiable? Always make sure to cross validate your results on a different set of data.*
- *Keep in mind that a prediction gives the average response value for the given combination of predictor values – don't expect it to be true!*

References

- *Data Analysis And Regression – a second course in statistics*
  - Mosteller and Tukey

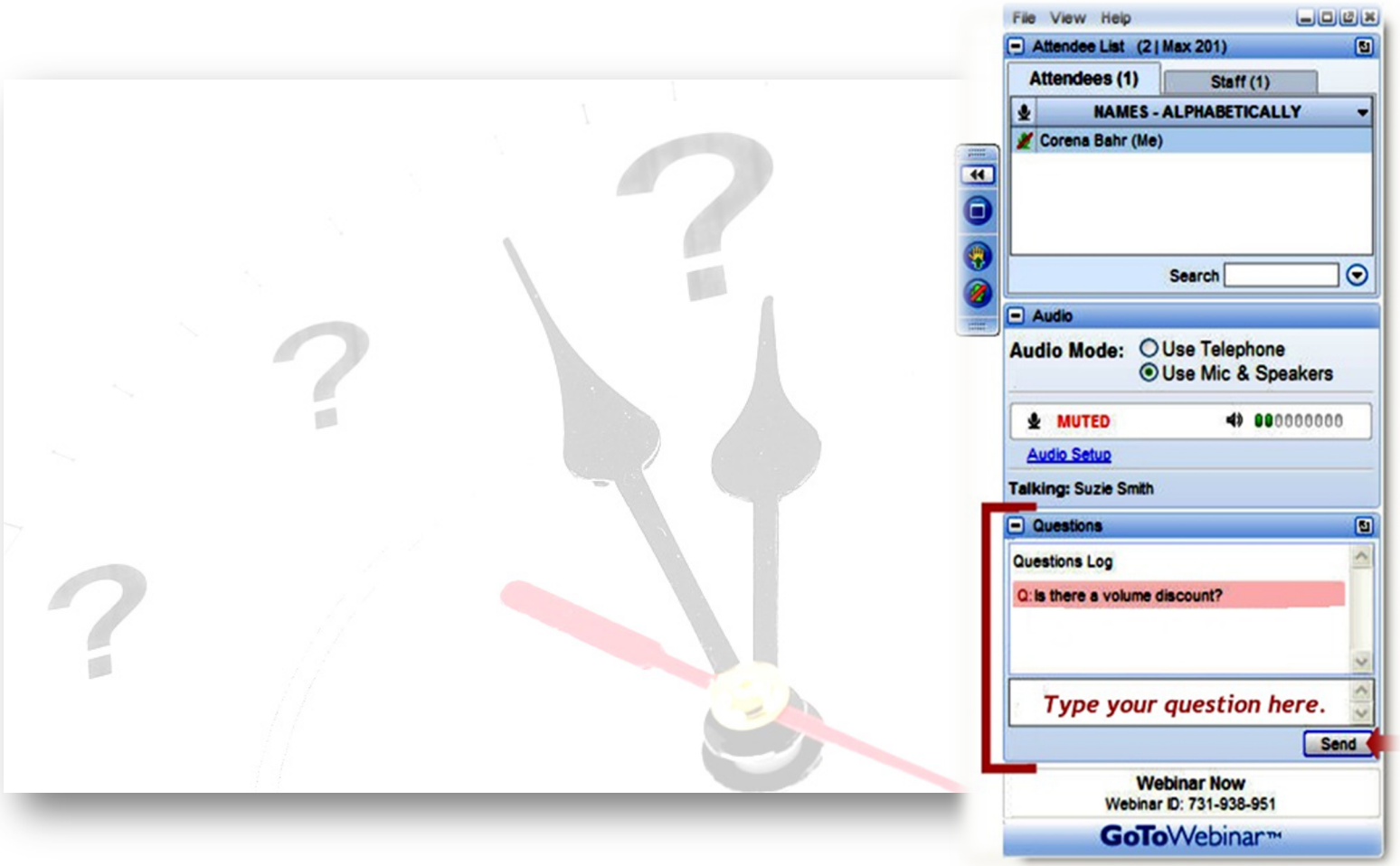
- *Latent Variables (Variates) and Multicollinearity - "Good or Bad?"*
  - Melinda K. Higgins, Ph.D.

- *The role of causal reasoning in understanding Simpson's paradox, Lord's paradox, and the suppression effect - Onyebuchi A Arah*
  http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2266743/

- *Correlation or Causation? – Jon Mueller,* Professor of Psychology
  http://jfmueller.faculty.noctrl.edu/100/correlation_or_causation.htm

- *Simpson's Paradox Illustrated*
  http://www.r-bloggers.com/simpsons-paradox/

Thank You for Joining Us



Master Black Belt Program

- Offered in partnership with Fisher College of Business at The Ohio State University
- Employs a Blended Learning model with world-class instruction delivered in both the classroom and online
- Covers the MBB Body of Knowledge, topics ranging from advanced *DOE* to *Leading Change* to *Finance for MBBs*

Resource Links and Contacts

Questions? Comments? We'd love to hear from you.

Smita Skrivanek, Senior Statistician – MoreSteam.com
sskrivanek@moresteam.com

Larry Goldman, Vice President Marketing – MoreSteam.com
lgoldman@moresteam.com

Join us for our next Webcast on February 13th:

"Into the Trenches of Regression Analysis (Part 2)" – Smita Skrivanek, MoreSteam.com

Archived presentations and other materials:
http://www.moresteam.com/presentations/