Disclosure Risk and Data Utility for Remote Access Servers

Jerome P. Reiter
Institute of Statistics and Decision Sciences
Duke University, Durham NC, USA

Data dissemination: Current practice

- Agency seeks to release microdata.
- Risk of re-identifications from matching to external databases.
- Statistical disclosure limitation applied to data before release.

Data dissemination: The future?

- A “world without microdata.”

- Options for data dissemination in this world:

1. Data summaries only.
2. Restricted access data centers.
3. Synthetic data.

4. Remote Access Server approaches.

Definition of servers

- Server is any system that

  (i) allows users to submit queries for output from statistical analyses of microdata,

  but

  (ii) does not give direct access to microdata.

- Focus on static model servers (not table servers).

Queries and responses

- Queries to model server:

  Users request results from fitting a statistical model to the data.

- Response from model server:

  Answerable query: model output.
  Unanswerable query: no results.

Model output also should include diagnostics.

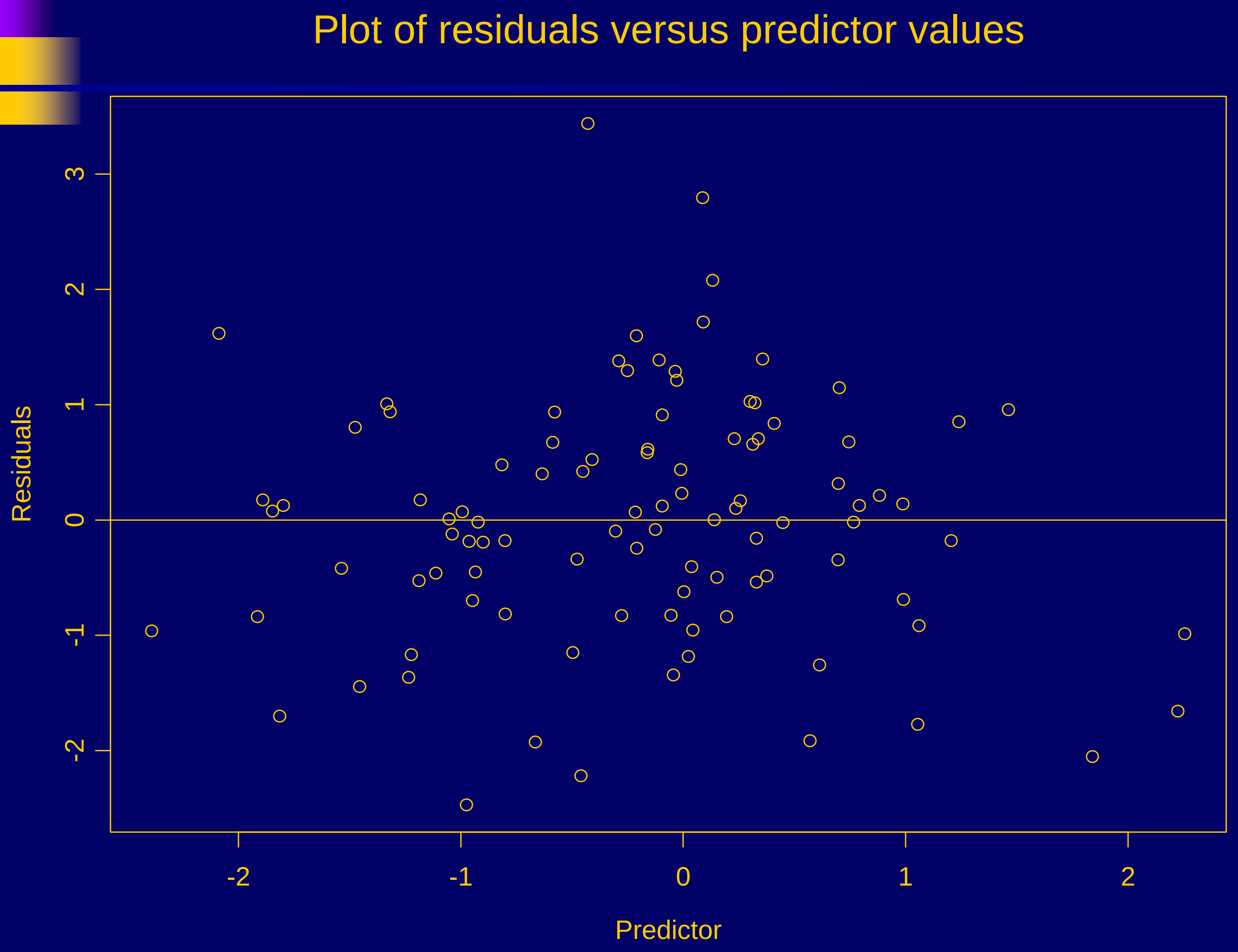
Model diagnostics for servers

- Users need way to assess the fit of their models.
- Standard diagnostics:
  residuals (actual Y minus predicted Y)
- Release may result in disclosures.

Can diagnostics be released?

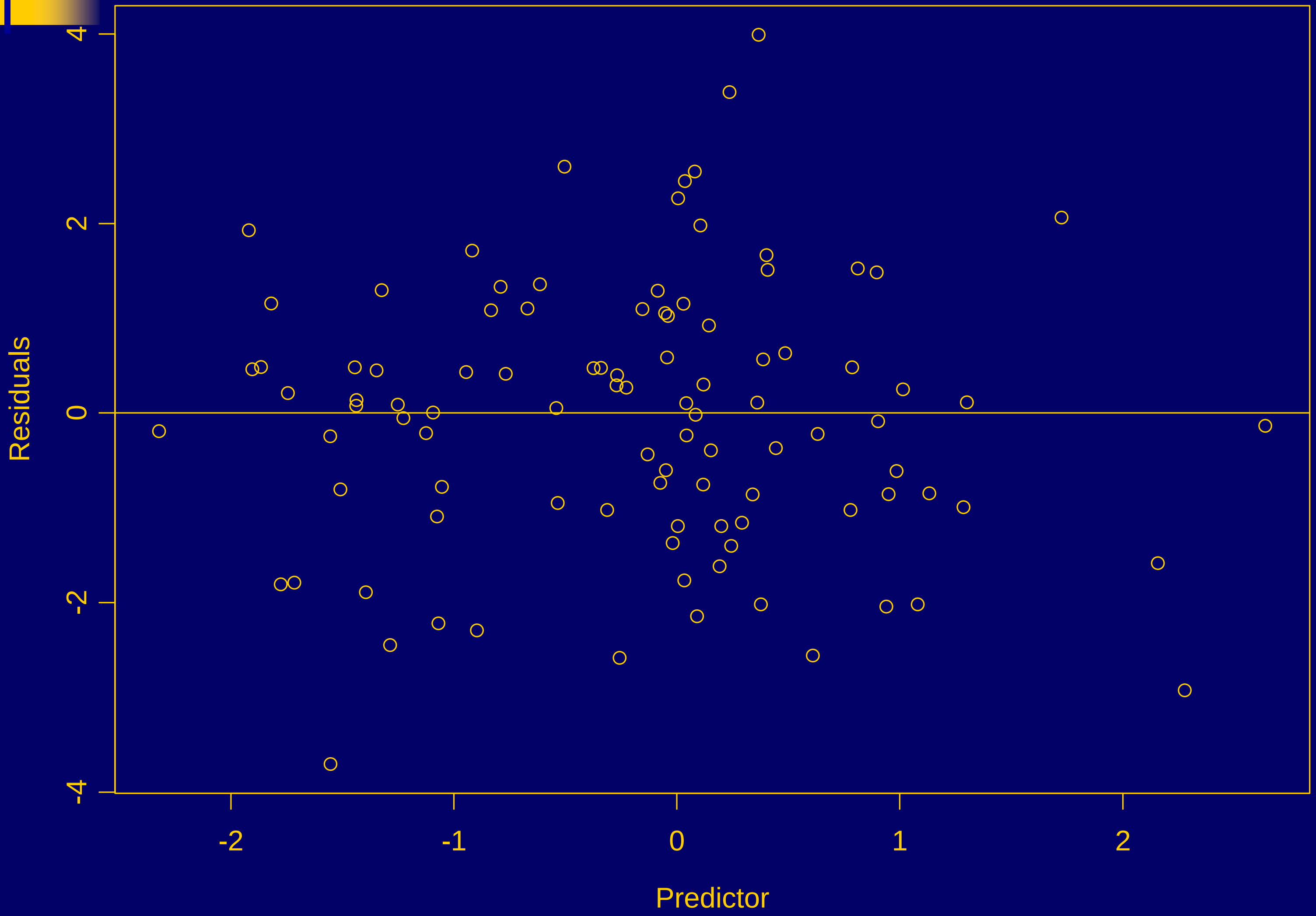
- Release synthetic (simulated) diagnostics.
- Mimic patterns in the real-data diagnostics.
- Users can interpret synthetic diagnostics as they would actual ones.

Linear regression: Good fit (actual residuals)



Linear regression:
Good fit (synthetic residuals)

Plot of synthetic residuals versus synthetic predictor values



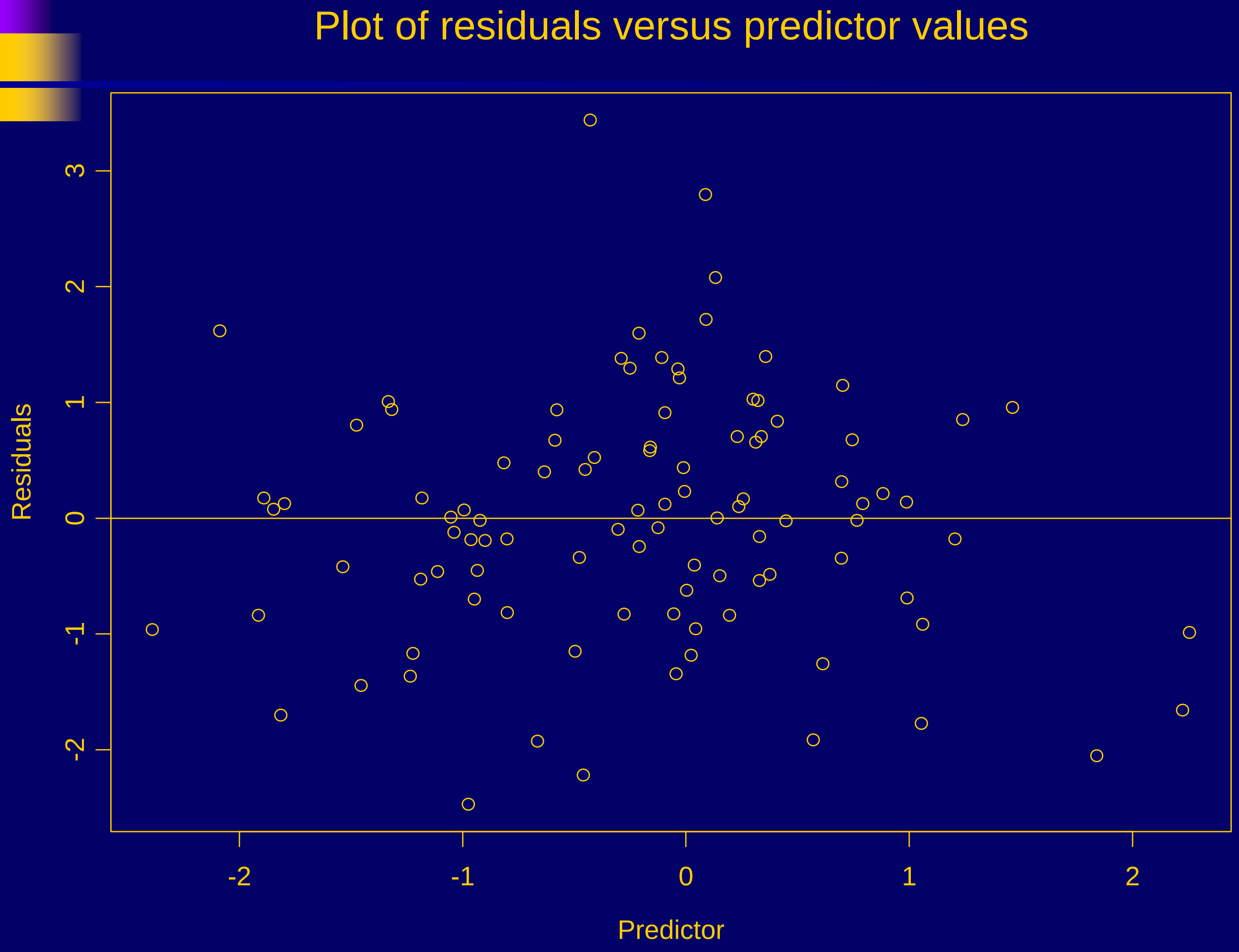
Synthetic diagnostics: Linear regression

- Generate synthetic values, x_{kp}^{s}

- For submitted regression, generate synthetic (standardized) residuals for each x_{kp}

$$t_{kp}^{s} = b_{kp} + v_{kp} + n_{kp}$$

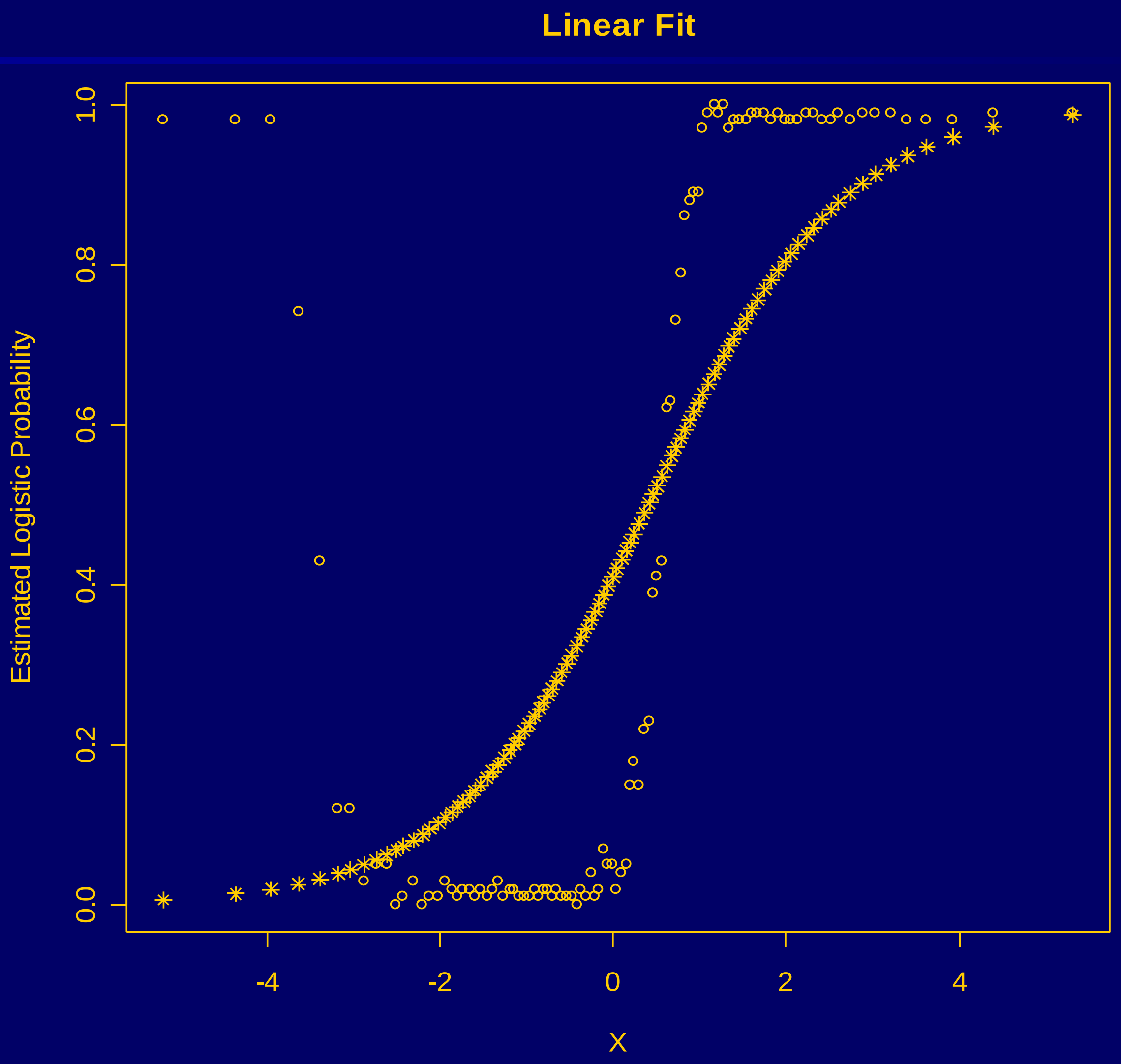
Linear regression: Good fit (actual residuals)



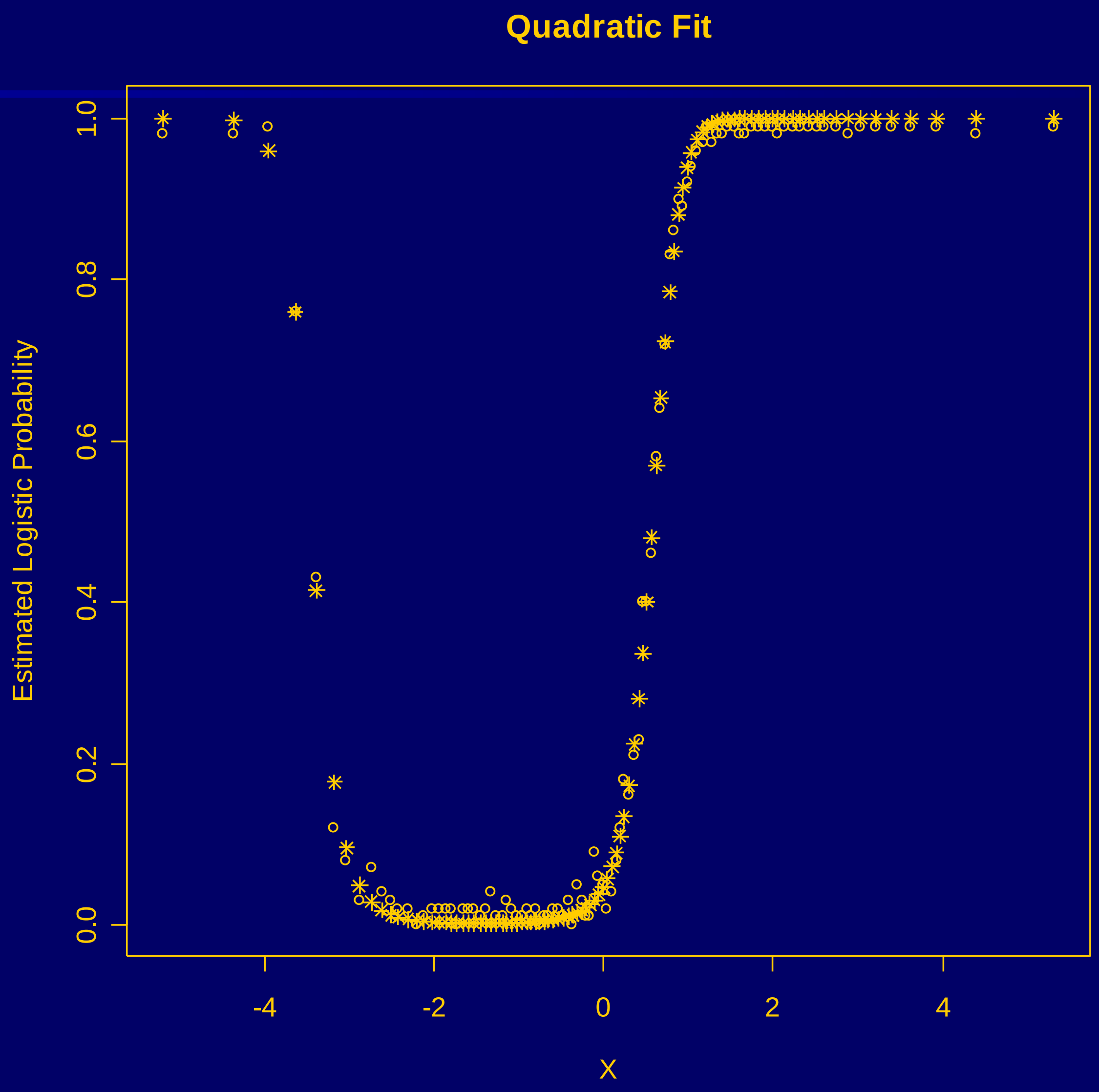
Synthetic diagnostics: Logistic regression

- Partition X_p into categories with 100 units.
- For submitted regression, calculate % of “successes” in each category.
- Add random noise to above %s.
- Plot perturbed %s versus averages of predicted probabilities in categories.

Logisitic regression:
Fit linear in X, true quad. in X



Logisitic regression: Fit and true quadratic in X



Are servers always safe from disclosures?

- Two main types of disclosure:

  (i) Re-identification/attribute disclosures

  (ii) Inferential disclosures

Examples of identity or attribute disclosures in servers.

- Transformation attacks to attempt re-identification or attribute disclosures

  1) Fit dummy variable equal to one for a particular value of a predictor, say *x*.

  2) Transform predictors to have super-high leverage:

$$f(X) = \frac{1}{(X - x) + \varepsilon}$$

Examples of inferential disclosures in servers

- Disallow relationships from being estimated exactly.

  Example: *Y* sensitive, *X* highly correlated with *Y*.
  No response to queries for *Corr* (*X, Y*).

- May be possible to reconstruct suppressed relationships from queries for other relationships.

  Example: No response for regression of *Y* on (*W,Z).*
  Response to *Y* on *W* and *Y* on *Z*
  results in full knowledge of *Y* on (*W,Z).*

Examples of inferential disclosures in servers, continued

- May be possible to obtain bounds on suppressed relationships from queries for other relationships.

  Example: Linear regression.

  Positive definiteness ->
  bounds on unreleased coefficients.

Risk and Utility Measures for Static Model Servers

- Risk Measures:

  In sample prediction risk.
  Out of sample prediction risk.

- Utility Measures:

  Volume of release.
  Statistical usefulness of release.

Next steps in developing model servers

- Limit transformation attacks without undue compromise of data utility.
- Formulate risk and utility measures for complicated models.
- Work with agencies to implement server ideas, including diagnostics.