An Application of Differentially Private Linear Mixed Modeling[1]

2012 Quantitative Methods on National Security and Defense

Matthew J. Schneider
Cornell University
(Joint work with John Abowd)

May 7, 2012

[1]We acknowledge NSF grants BCS 0941226, SES 9978093, ITR 0427889, and SES 0922005.

This Research Integrates PPD and SDL

- In the last 10 years, research on statistical privacy was in two academic fields: Computer Science and Statistics.

This Research Integrates PPD and SDL

- In the last 10 years, research on statistical privacy was in two academic fields: Computer Science and Statistics.
- Privacy Preserving Data Mining (PPD) - Computer science research focused on rigorous theoretical definitions and algorithmic aspects of data privacy, but ignored applicability. Previously, cryptography.

This Research Integrates PPD and SDL

- In the last 10 years, research on statistical privacy was in two academic fields: Computer Science and Statistics.
- Privacy Preserving Data Mining (PPD) - Computer science research focused on rigorous theoretical definitions and algorithmic aspects of data privacy, but ignored applicability. Previously, cryptography.
- Statistical Disclosure Limitation (SDL) - Statisticians focused on applicability, but used ad-hoc methods and lacked formal guarantees of privacy. For example, adding random noise to sensitive variables.

63 % of people in the US can be uniquely identified by a combination of their 5-digit zip code, gender, and date of birth (Golle 2006)

- Suppose we have a dataset with 1,000 people where prevalence of some disease is 5.8%.

63 % of people in the US can be uniquely identified by a combination of their 5-digit zip code, gender, and date of birth (Golle 2006)

- Suppose we have a dataset with 1,000 people where prevalence of some disease is 5.8%.
- Remove one person from the study or find him to have a unique combo of zip-gender-DOB: Disease prevalence $\rightarrow$ 5.7%.

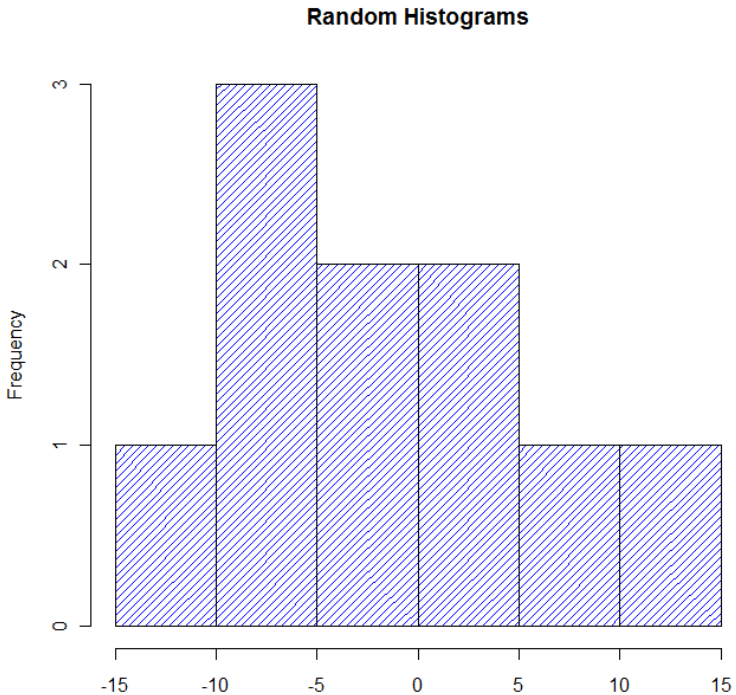
63 % of people in the US can be uniquely identified by a combination of their 5-digit zip code, gender, and date of birth (Golle 2006)

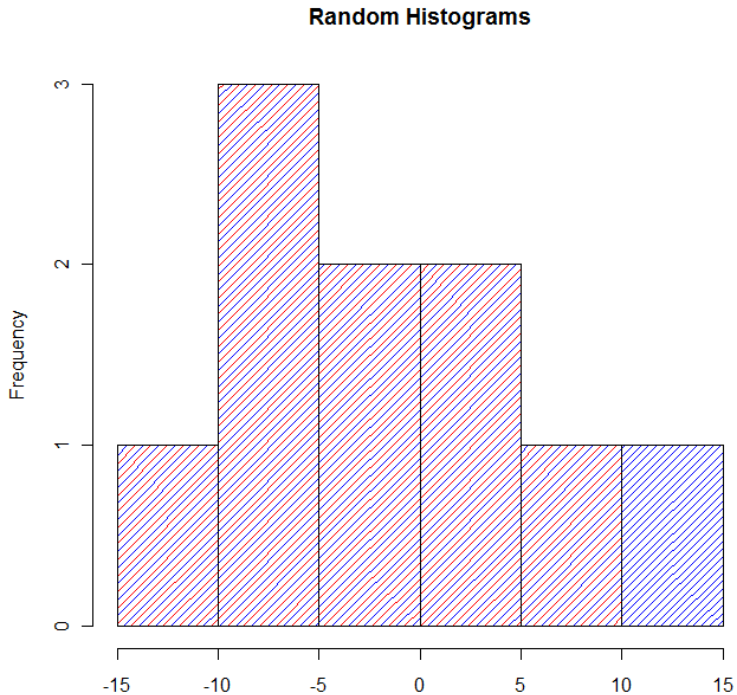
- Suppose we have a dataset with 1,000 people where prevalence of some disease is 5.8%.
- Remove one person from the study or find him to have a unique combo of zip-gender-DOB: Disease prevalence $\rightarrow$ 5.7%.
- We know with 100% certainty that this person has a disease even though he is only $\frac{1}{1000}$th of the dataset.

63 % of people in the US can be uniquely identified by a combination of their 5-digit zip code, gender, and date of birth (Golle 2006)

- Suppose we have a dataset with 1,000 people where prevalence of some disease is 5.8%.
- Remove one person from the study or find him to have a unique combo of zip-gender-DOB: Disease prevalence $\rightarrow$ 5.7%.
- We know with 100% certainty that this person has a disease even though he is only $\frac{1}{1000}$th of the dataset.
- The sensitive variable can also be income, anything from medical records, personnel records, govt or corp data.

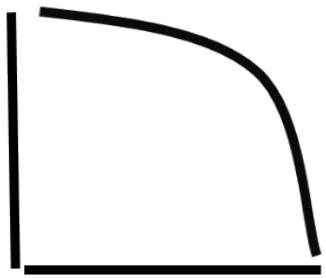
Continuous Data: Remove Max Point
Mean shifts from -2 to -4



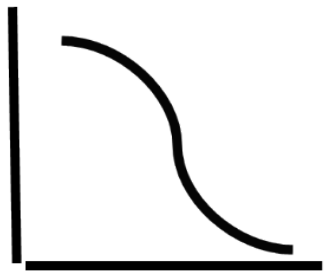


Protection affects the Data Miner AND the Analyst

Protection



Identification Probability



Utility

Practicing Statistical Data Privacy is Supported

- Governmental laws and regulations such as the Privacy Act of 1974 require organizations to invest in privacy.

Practicing Statistical Data Privacy is Supported

- Governmental laws and regulations such as the Privacy Act of 1974 require organizations to invest in privacy.
- The Department of Defense once tried to restrict data mining with the Data Mining Moratorium Act of 2003.

Practicing Statistical Data Privacy is Supported

- Governmental laws and regulations such as the Privacy Act of 1974 require organizations to invest in privacy.
- The Department of Defense once tried to restrict data mining with the Data Mining Moratorium Act of 2003.
- The Healthcare Insurance Portability and Accountability Act has provisions in it for individual privacy.

Practicing Statistical Data Privacy is Supported

- Governmental laws and regulations such as the Privacy Act of 1974 require organizations to invest in privacy.
- The Department of Defense once tried to restrict data mining with the Data Mining Moratorium Act of 2003.
- The Healthcare Insurance Portability and Accountability Act has provisions in it for individual privacy.
- Past, serious breaches in privacy through data mining.

Introduction: We Use a Linear Mixed Model

- A model-based approach formalizes Statistical Disclosure Limitation.

Introduction: We Use a Linear Mixed Model

- A model-based approach formalizes Statistical Disclosure Limitation.
- The linear mixed-effects model is the statistical workhorse of small area estimation, which is an important part of many statistical agencies' publication program.

Introduction: We Use a Linear Mixed Model

- A model-based approach formalizes Statistical Disclosure Limitation.
- The linear mixed-effects model is the statistical workhorse of small area estimation, which is an important part of many statistical agencies' publication program.
- We are interested in an estimate of the extent to which a particular entity (detailed geographical unit or industry) differs from the average.

Introduction: We Use a Linear Mixed Model

- A model-based approach formalizes Statistical Disclosure Limitation.
- The linear mixed-effects model is the statistical workhorse of small area estimation, which is an important part of many statistical agencies' publication program.
- We are interested in an estimate of the extent to which a particular entity (detailed geographical unit or industry) differs from the average.
- That deviation is modeled as the realization of a random process, and is estimated conditional on the actual values of a few entities with the particular level of the detailed factor under study.

Data Sources: QWI Data is Hierarchical/Mixed, Temporal, Bounded, and Categorical

- We use the Census Bureau's Quarterly Workforce Indicators (QWI) as our application of linear mixed model estimation to small area and industrial detail data protection.

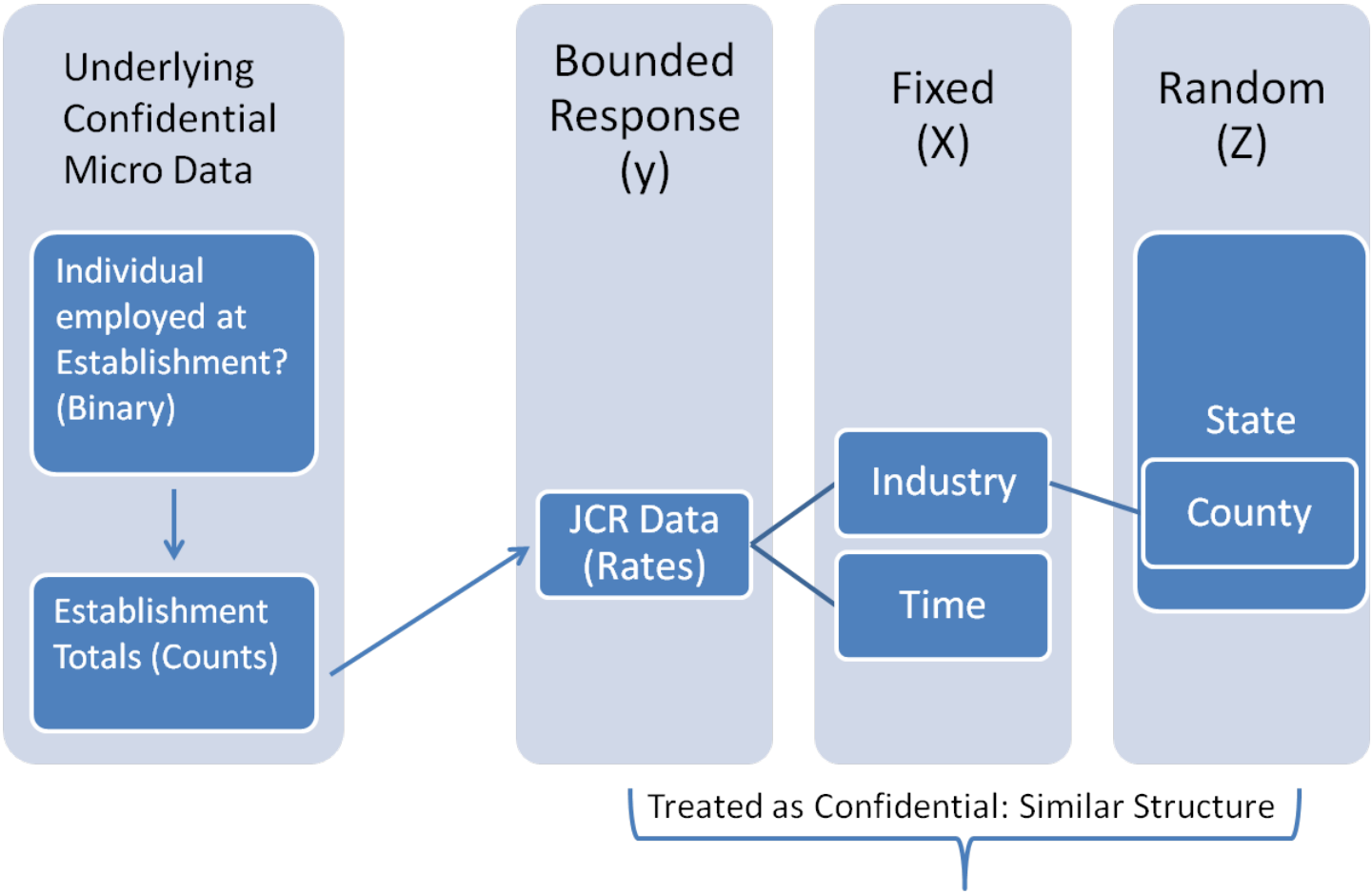
Data Sources: QWI Data is Hierarchical/Mixed, Temporal, Bounded, and Categorical

- We use the Census Bureau's Quarterly Workforce Indicators (QWI) as our application of linear mixed model estimation to small area and industrial detail data protection.
- The QWI data contain employment counts, accessions, separations, and explanatory variables of interest, namely, industry (20 levels), state (48 levels), county (3,113 unique levels within state), and quarter (80 levels, 1990:Q2 to 2010:Q1).

Data Sources: QWI Data is Hierarchical/Mixed, Temporal, Bounded, and Categorical

- We use the Census Bureau's Quarterly Workforce Indicators (QWI) as our application of linear mixed model estimation to small area and industrial detail data protection.
- The QWI data contain employment counts, accessions, separations, and explanatory variables of interest, namely, industry (20 levels), state (48 levels), county (3,113 unique levels within state), and quarter (80 levels, 1990:Q2 to 2010:Q1).
- The dependent variable of interest is job creation rate (JCR). We model rates instead of levels because the differentially private MLE requires a bounded parameter space, and these rates are naturally bounded ($JCR \in (0, 2)$).

Data Sources: Tailored Toward Small Area Estimates



Differential Privacy

- Smith defines differential privacy for a randomized algorithm T() as being ϵ-differentially private if for all neigboring pairs of databases x and x', and for all measureable subsets of events S:

$$Pr(T(x) \in S) \leq exp(\epsilon) \times Pr(T(x') \in S)$$

Differential Privacy

- Smith defines differential privacy for a randomized algorithm T() as being ϵ-differentially private if for all neigboring pairs of databases x and x', and for all measureable subsets of events S:

  $$Pr(T(x) \in S) \leq exp(\epsilon) \times Pr(T(x') \in S)$$

- Our application:
  (y, X, Z)

Differential Privacy

- Smith defines differential privacy for a randomized algorithm T() as being ϵ-differentially private if for all neigboring pairs of databases x and x', and for all measureable subsets of events S:

  $$Pr(T(x) \in S) \leq exp(\epsilon) \times Pr(T(x') \in S)$$

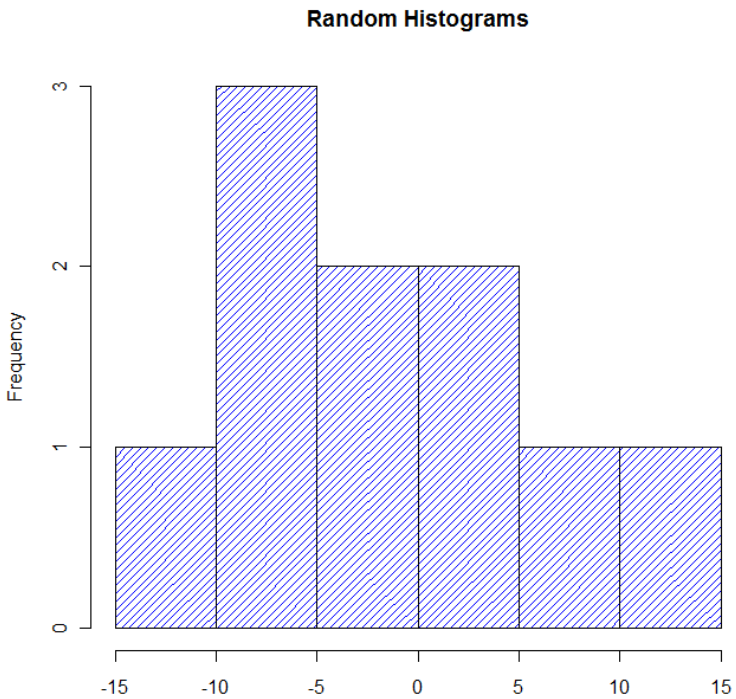
- Our application:
  (y, X, Z)
  MLE(β)=$\hat{\beta}$= $(X^T V^{-1} X)^{-1} X^T V^{-1} y$ and $V = ZGZ^T + R$

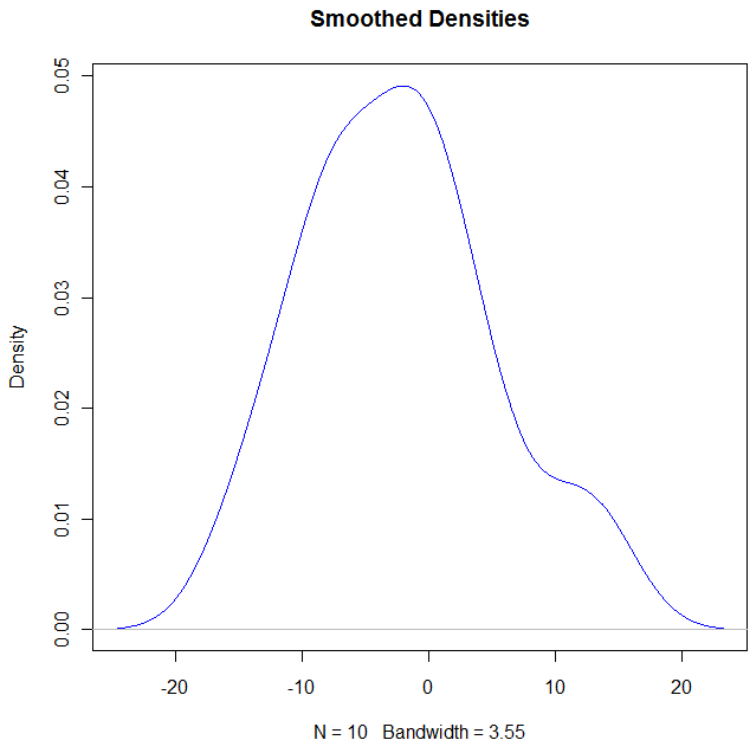
Differential Privacy

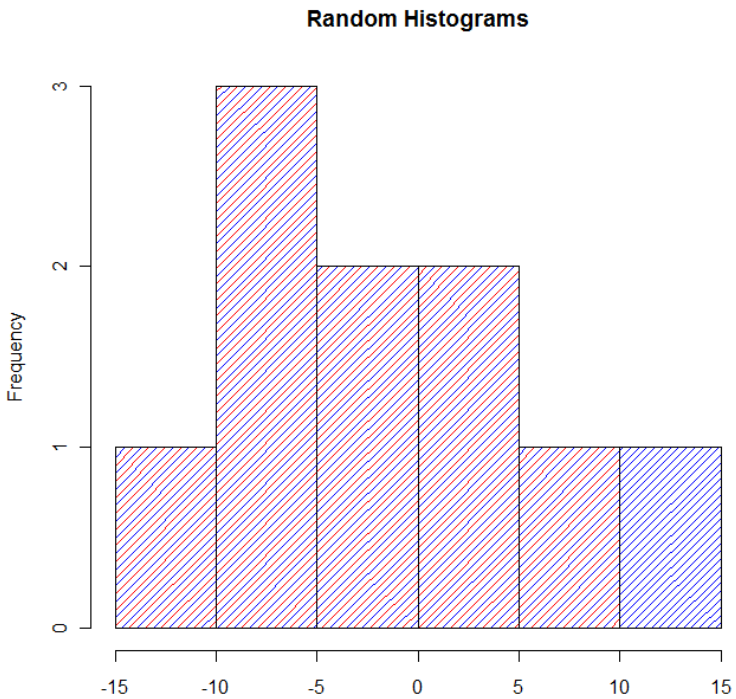
- Smith defines differential privacy for a randomized algorithm T() as being ϵ-differentially private if for all neigboring pairs of databases x and x', and for all measureable subsets of events S:

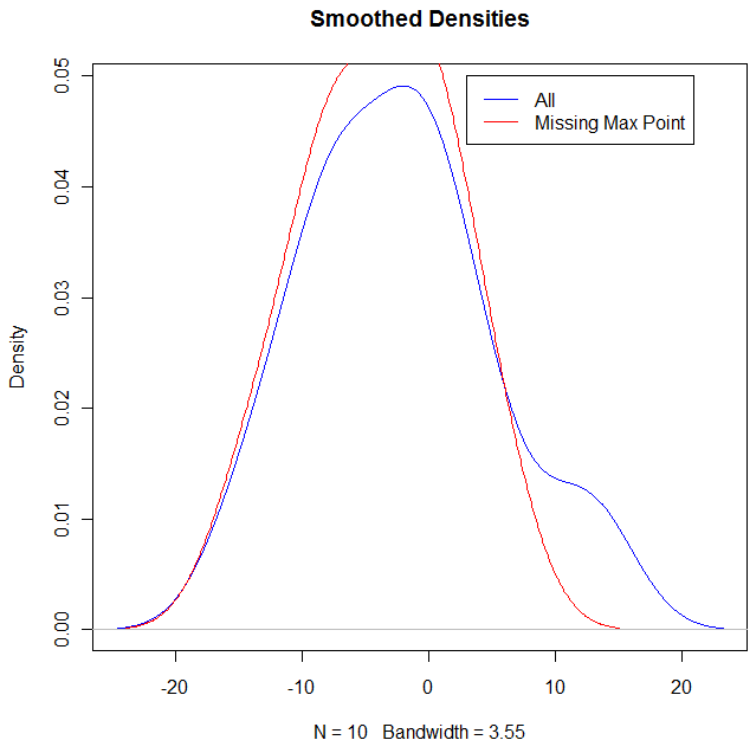
  $$Pr(T(x) \in S) \leq exp(\epsilon) \times Pr(T(x') \in S)$$

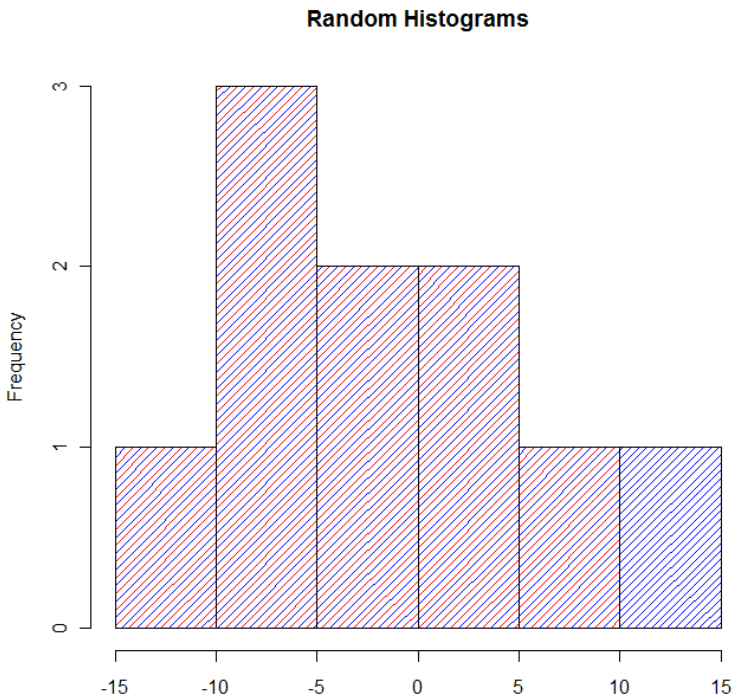
- Our application:
  (y, X, Z)
  $\text{MLE}(\beta)=\hat{\beta}= (X^T V^{-1} X)^{-1} X^T V^{-1} y$ and $V = ZGZ^T + R$
  $EBLUP(u) = \hat{u}= \hat{G} Z^T \hat{V}^{-1} (y - X\hat{\beta})$

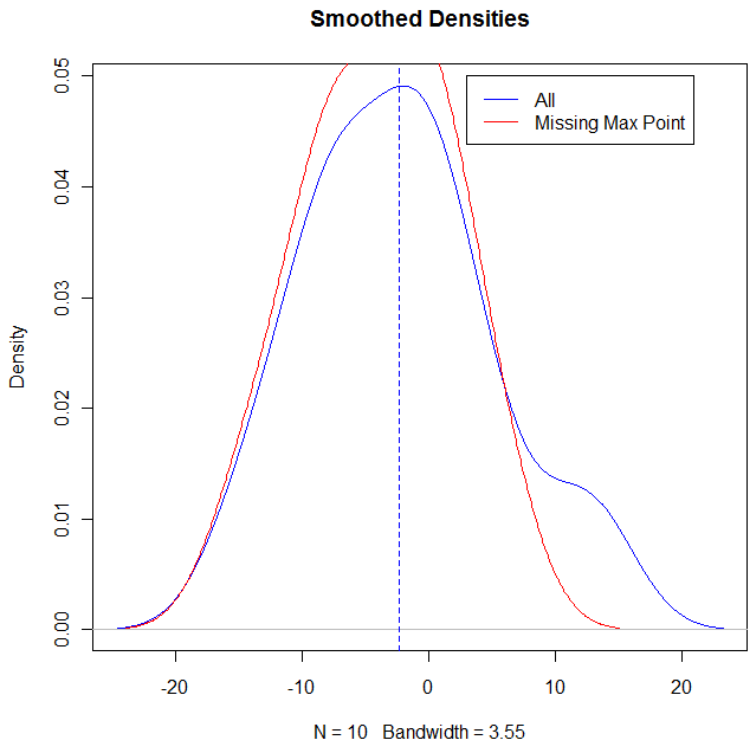

Random Histograms
Frequency
0
1
2
3
-15
-10
-5
0
5
10
15

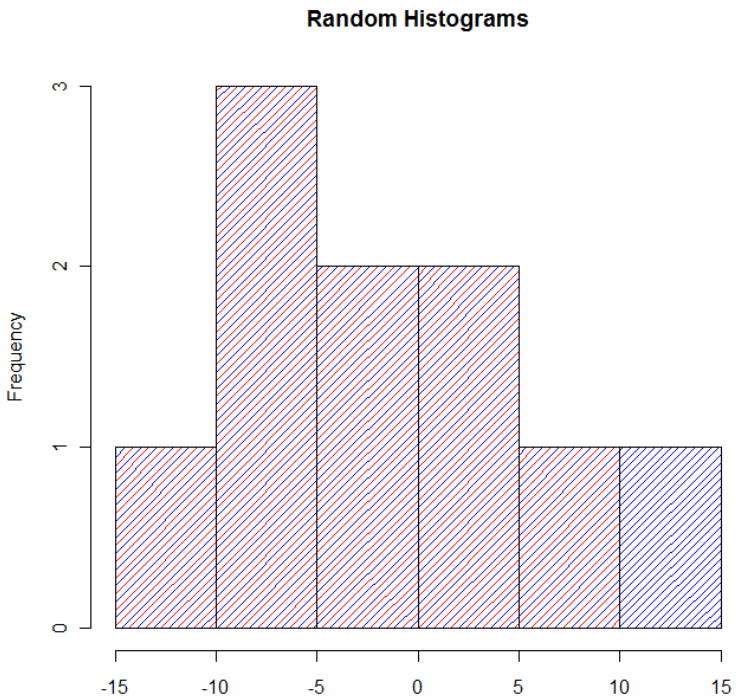

Smoothed Densities
Density
0.00
0.01
0.02
0.03
0.04
0.05
-20
-10
0
10
20
N = 10 Bandwidth = 3.55

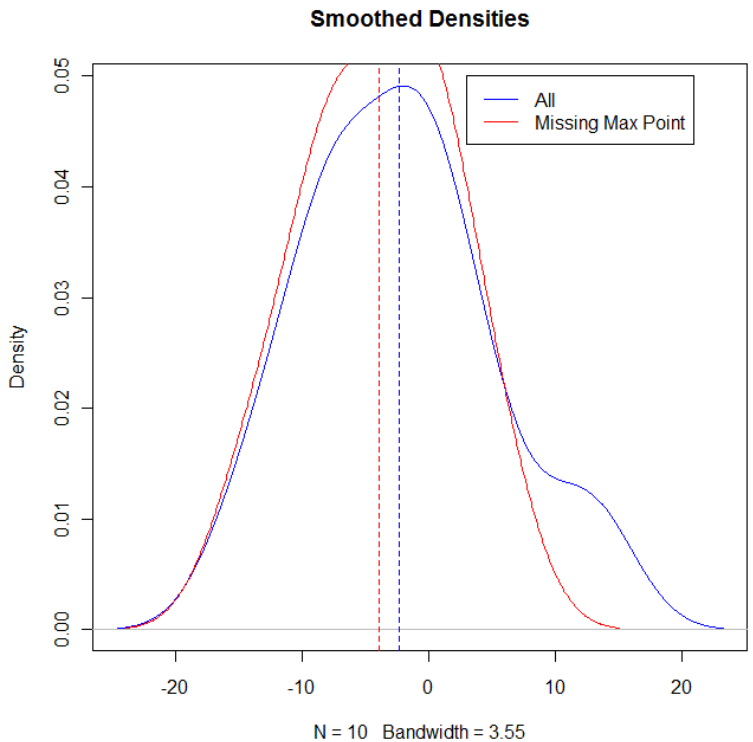

Random Histograms
Frequency
0
1
2
3
-15
-10
-5
0
5
10
15

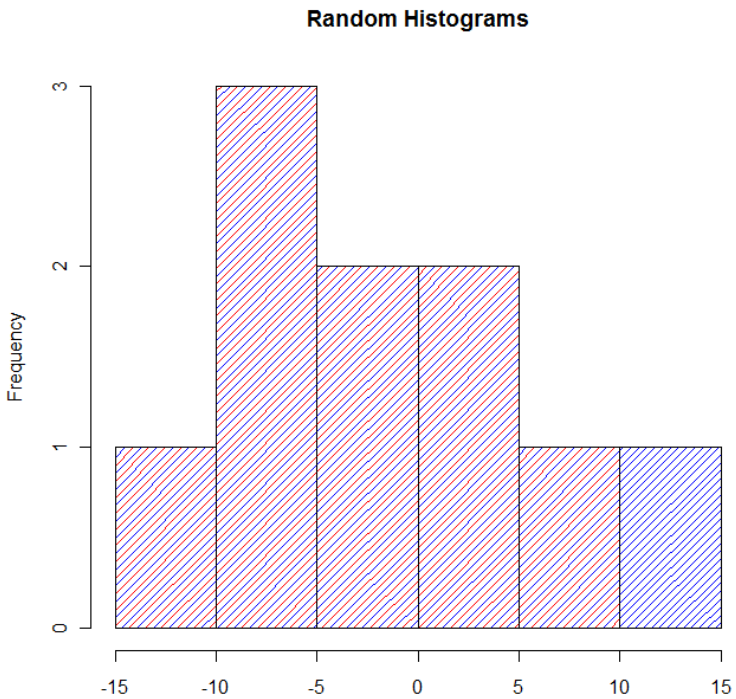

Smoothed Densities
Density
0.00
0.01
0.02
0.03
0.04
0.05
All
Missing Max Point
-20
-10
0
10
20
N = 10 Bandwidth = 3.55

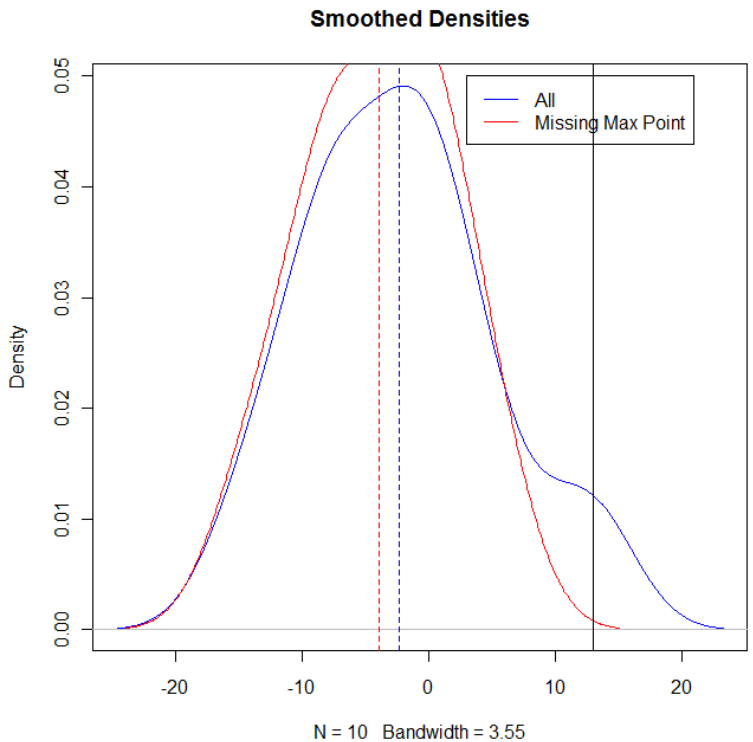

Random Histograms
Frequency
0
1
2
3
-15
-10
-5
0
5
10
15


Smoothed Densities
Density
0.00
0.01
0.02
0.03
0.04
0.05
All
Missing Max Point
-20
-10
0
10
20
N = 10 Bandwidth = 3.55


Random Histograms
Frequency
0
1
2
3
-15
-10
-5
0
5
10
15


Smoothed Densities
Density
0.00
0.01
0.02
0.03
0.04
0.05
All
Missing Max Point
-20
-10
0
10
20
N = 10 Bandwidth = 3.55


Random Histograms
Frequency
0
1
2
3
-15
-10
-5
0
5
10
15


Smoothed Densities
Density
0.00
0.01
0.02
0.03
0.04
0.05
All
Missing Max Point
-20
-10
0
10
20
N = 10 Bandwidth = 3.55

Model Specification: LMM with Two Mixed-Effects

$$y = X\beta + Zu + \xi \tag{1}$$

Model Specification: LMM with Two Mixed-Effects

$$y = X\beta + Zu + \xi \tag{1}$$

$$\xi \sim N(0, \sigma_\xi^2 I_N) = N(0, R), R = \sigma_\xi^2 I_N$$

$$u_s \sim N(0, \sigma_s^2 I_{48}), u_c \sim N(0, \sigma_c^2 I_{3113})$$

$$u = (u_s^T, u_c^T)^T \sim N(0, G)$$

$$G = \begin{bmatrix} \sigma_s^2 I_{48} & 0 \\ 0 & \sigma_c^2 I_{3113} \end{bmatrix}$$

Model Specification: LMM with Two Mixed-Effects

$$y = X\beta + Zu + \xi \tag{1}$$

$$\xi \sim N(0, \sigma_\xi^2 I_N) = N(0, R), R = \sigma_\xi^2 I_N$$
$$u_s \sim N(0, \sigma_s^2 I_{48}), u_c \sim N(0, \sigma_c^2 I_{3113})$$
$$u = (u_s^T, u_c^T)^T \sim N(0, G)$$
$$G = \begin{bmatrix} \sigma_s^2 I_{48} & 0 \\ 0 & \sigma_c^2 I_{3113} \end{bmatrix}$$

$$E[y|X, Z] = X\beta, y \sim N(X\beta, ZGZ^T + R)$$

and given random effects due to state and county:

$$E[y|X, Z, u] = X\beta + Zu, (y|u) \sim N(X\beta + Zu, \sigma_\xi^2 I_N)$$

Model Specification: 1000s of Estimates, 2.4 Million N

We calculate initial global estimates ($\hat{\beta}^{global}$,$\hat{u}^{global}$,$\hat{\sigma}^{global}$) using REML as a benchmark for the differentially private methods in this paper that use sub-sampling and Laplace noise.

Estimate	Dimension	Description
$\hat{\beta}_1,\ldots,\hat{\beta}_{20}$	20	Industry (n) MLEs
$\hat{\beta}_{21}$	1	Quarter (t) MLE
$\hat{u}_1,\ldots,\hat{u}_{48}$	48	State (s) EBLUPs
$\hat{u}_{49},\ldots,\hat{u}_{49+3112}$	3113	County (c) EBLUPs
$\hat{\sigma}^2_\xi$	1	Residual Variance
$\hat{\sigma}^2_s$	1	State (s) Variance
$\hat{\sigma}^2_c$	1	County (c) Variance

Table : Estimate Descriptions

We Estimate Thousands of LMMs

We Estimate Thousands of LMMs

We apply Smith's method of differential privacy with sub-sampling.

We Estimate Thousands of LMMs

We apply Smith's method of differential privacy with sub-sampling.

1. Choose the number of subsamples k

We Estimate Thousands of LMMs

We apply Smith's method of differential privacy with sub-sampling.

1. Choose the number of subsamples k
2. Divide the input (y, X, Z) into k disjoint blocks, *i.e.* sub-samples by rows, $B_1, ..., B_{(i)}, ..., B_k$ of $n_k = \lfloor \frac{N}{k} \rfloor$ points each where $B_{(i)}$ denotes the i^{th} disjoint subset and N is the total number of observations. The complete dataset for each of the three models is denoted by $(y, X, Z) = \bigcup (y_1, X_1, Z_1), ..., (y_{(i)}, X_{(i)}, Z_{(i)}), ..., (y_n, X_n, Z_n)$.

We Estimate Thousands of LMMs

We apply Smith's method of differential privacy with sub-sampling.

1. Choose the number of subsamples k
2. Divide the input (y, X, Z) into k disjoint blocks, *i.e.* sub-samples by rows, $B_1, ..., B_{(i)}, ..., B_k$ of $n_k = \lfloor \frac{N}{k} \rfloor$ points each where $B_{(i)}$ denotes the i^{th} disjoint subset and N is the total number of observations. The complete dataset for each of the three models is denoted by $(y, X, Z) = \bigcup (y_1, X_1, Z_1), ..., (y_{(i)}, X_{(i)}, Z_{(i)}), ..., (y_n, X_n, Z_n)$.
3. Using `lmer()`, calculate k sets of estimates from the previous table using the data of each block only.

As k Increases, Biases of $\hat{\beta}_{(i)}$, $\hat{u}_{(i)}$ Increase

We found empirical evidence that our estimates of $\hat{\beta}_{(i)}$ and $\hat{u}_{(i)}$ are more biased for increased values of subsamples k

As k Increases, Biases of $\hat{\beta}_{(i)}$, $\hat{u}_{(i)}$ Increase

We found empirical evidence that our estimates of $\hat{\beta}_{(i)}$ and $\hat{u}_{(i)}$ are more biased for increased values of subsamples k

-
- $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ and $BLUP = GZ^T V^{-1}(y - X\hat{\beta})$, where $V = ZGZ^T + R$.

As k Increases, Biases of $\hat{\beta}_{(i)}$, $\hat{u}_{(i)}$ Increase

We found empirical evidence that our estimates of $\hat{\beta}_{(i)}$ and $\hat{u}_{(i)}$ are more biased for increased values of subsamples k

-
- $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ and $BLUP = GZ^T V^{-1}(y - X\hat{\beta})$, where $V = ZGZ^T + R$.
- Our estimate of u is $EBLUP(u) = \hat{u} = \hat{G} Z^T \hat{V}^{-1}(y - X\hat{\beta})$.

As k Increases, Biases of $\hat{\beta}_{(i)}$, $\hat{u}_{(i)}$ Increase

We found empirical evidence that our estimates of $\hat{\beta}_{(i)}$ and $\hat{u}_{(i)}$ are more biased for increased values of subsamples k

-
- $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ and $BLUP = GZ^T V^{-1}(y - X\hat{\beta})$, where $V = ZGZ^T + R$.
- Our estimate of u is $EBLUP(u) = \hat{u} = \hat{G} Z^T \hat{V}^{-1}(y - X\hat{\beta})$.
- $EBLUP$s are more biased as we increase the number of subsamples k. Additionally, the estimated variance components become larger as k increased. We regress biased estimates on biases of variances to correct bias.

As k Increases, Biases of $\hat{\beta}_{(i)}$, $\hat{u}_{(i)}$ Increase

We found empirical evidence that our estimates of $\hat{\beta}_{(i)}$ and $\hat{u}_{(i)}$ are more biased for increased values of subsamples k

-
- $\hat{\beta} = (X^T V^{-1} X)^{-1} X^T V^{-1} y$ and $BLUP = GZ^T V^{-1}(y - X\hat{\beta})$, where $V = ZGZ^T + R$.
- Our estimate of u is $EBLUP(u) = \hat{u} = \hat{G} Z^T \hat{V}^{-1}(y - X\hat{\beta})$.
- $EBLUP$s are more biased as we increase the number of subsamples k. Additionally, the estimated variance components become larger as k increased. We regress biased estimates on biases of variances to correct bias.
- As ϵ ranges from 0.1 (very private) to 4.6 (less private), the optimal k^*ranges from about 22,470 to 4,858. A value of $k^* > 9,000$ is not feasible within the REML computation because the low sample size ($n_k = 151$) does not permit any estimation at all.

Confidential Estimates are Averages of Sub-Sampled Estimates

Average the estimates over k blocks:

$$\hat{\beta}^{**} = \frac{\sum_{i=1}^{k} \hat{\beta}_{(i)}}{k}, \hat{u}^{**} = \frac{\sum_{i=1}^{k} \hat{u}_{(i)}}{k},$$

$$\hat{\sigma}_s^{2**} = \frac{\sum_{i=1}^{k} \hat{\sigma}_{s(i)}^2}{k}, \hat{\sigma}_c^{2**} = \frac{\sum_{i=1}^{k} \hat{\sigma}_{c(i)}^2}{k},$$

and

$$\hat{\sigma}_\xi^{2**} = \frac{\sum_{i=1}^{k} \hat{\sigma}_{\xi(i)}^2}{k}$$

Natural Bounds of Parameters Limit Noise Required for Released Estimates

Estimate	JCR Max Range	JCR Λ
$\hat{\beta}_1,\ldots,\hat{\beta}_{20}$	<2.49	2
$\hat{\beta}_{21}$	0.003	.01
$\hat{u}_1,\ldots,\hat{u}_{48}$	1.67	2
$\hat{u}_{49},\ldots,\hat{u}_{3161}$	2.04	2
$\hat{\sigma}^2_\xi$	0.20	0.25
$\hat{\sigma}^2_s$	0.26	0.25
$\hat{\sigma}^2_c$	0.19	0.25

Table : Max Empirical Ranges

$$\text{Released Estimate} = \text{Confidential Estimate} + \text{Laplace}\left(\frac{\Lambda}{k\epsilon}\right)$$

Fitted Values are Sums of Allocated Privacy Budgets

$$JCR^{DP_\epsilon} = \hat{y}^{DP_\epsilon} = ?X\hat{\beta}^{DP_\epsilon} + Z\hat{u}^{DP_\epsilon}$$

Fitted Values are Sums of Allocated Privacy Budgets

$$JCR^{DP_\epsilon} = \hat{y}^{DP_\epsilon} = ?X\hat{\beta}^{DP_\epsilon} + Z\hat{u}^{DP_\epsilon}$$

Lemma

For any choice of the number of sub-samples k, a fitted value for any row is $C\epsilon$-differentially private where C is 2, the assumed number of non-zero entries in Z and U for an added or deleted row r.

Fitted Values are Sums of Allocated Privacy Budgets

$$JCR^{DP_\epsilon} = \hat{y}^{DP_\epsilon} = ?X\hat{\beta}^{DP_\epsilon} + Z\hat{u}^{DP_\epsilon}$$

Lemma

For any choice of the number of sub-samples k, a fitted value for any row is $C\epsilon$-differentially private where C is 2, the assumed number of non-zero entries in Z and U for an added or deleted row r.

0.1ϵMLE$+0.9\epsilon$EBLUP$= (0.1 + 0.9)\epsilon$-differentially private

In our models, ϵ changes to generate the $R - U$ curve

- Duncan [3] states that "in its most basic form, an $R - U$ confidentiality map is the set of paired values (R, U) of disclosure risk and data utility that correspond to various strategies for data release."

In our models, ϵ changes to generate the $R - U$ curve

- Duncan [3] states that "in its most basic form, an $R - U$ confidentiality map is the set of paired values (R, U) of disclosure risk and data utility that correspond to various strategies for data release."
- Low disclosure risk has good differential privacy, which says that "any possible outcome of an analysis should be "almost" equally likely, independent of whether any individuals opts into or opts out of the data set" [4].

In our models, ϵ changes to generate the $R - U$ curve

- Duncan [3] states that "in its most basic form, an $R - U$ confidentiality map is the set of paired values (R, U) of disclosure risk and data utility that correspond to various strategies for data release."
- Low disclosure risk has good differential privacy, which says that "any possible outcome of an analysis should be "almost" equally likely, independent of whether any individuals opts into or opts out of the data set" [4].
- Lower values of ϵ correspond to lower levels of risk and higher levels of privacy.

Correlations of Fitted Values are Plotted over ϵ

- For all values of ϵ, calculate the predicted rates:

$$JCR^{DP} = \hat{y}^{DP_\epsilon} = X\hat{\beta}^{DP_{.51\epsilon}} + Z\hat{u}^{DP_{.49\epsilon}}$$

Correlations of Fitted Values are Plotted over ϵ

- For all values of ϵ, calculate the predicted rates:

$$JCR^{DP} = \hat{y}^{DP_\epsilon} = X\hat{\beta}^{DP_{.51\epsilon}} + Z\hat{u}^{DP_{.49\epsilon}}$$

- For $k = 1$ or all of the data, calculate the predicted rates:

$$JCR^{global} = \hat{y}^{global} = X\hat{\beta}^{global} + Z\hat{u}^{global}$$

Correlations of Fitted Values are Plotted over ϵ

- For all values of ϵ, calculate the predicted rates:

$$JCR^{DP} = \hat{y}^{DP_\epsilon} = X\hat{\beta}^{DP_{.51\epsilon}} + Z\hat{u}^{DP_{.49\epsilon}}$$

- For $k = 1$ or all of the data, calculate the predicted rates:

$$JCR^{global} = \hat{y}^{global} = X\hat{\beta}^{global} + Z\hat{u}^{global}$$

- Calculate the correlations between y and $\hat{y}^{global}$, $\hat{y}^{DP_\epsilon}$.

Correlations of Fitted Values are Plotted over ϵ

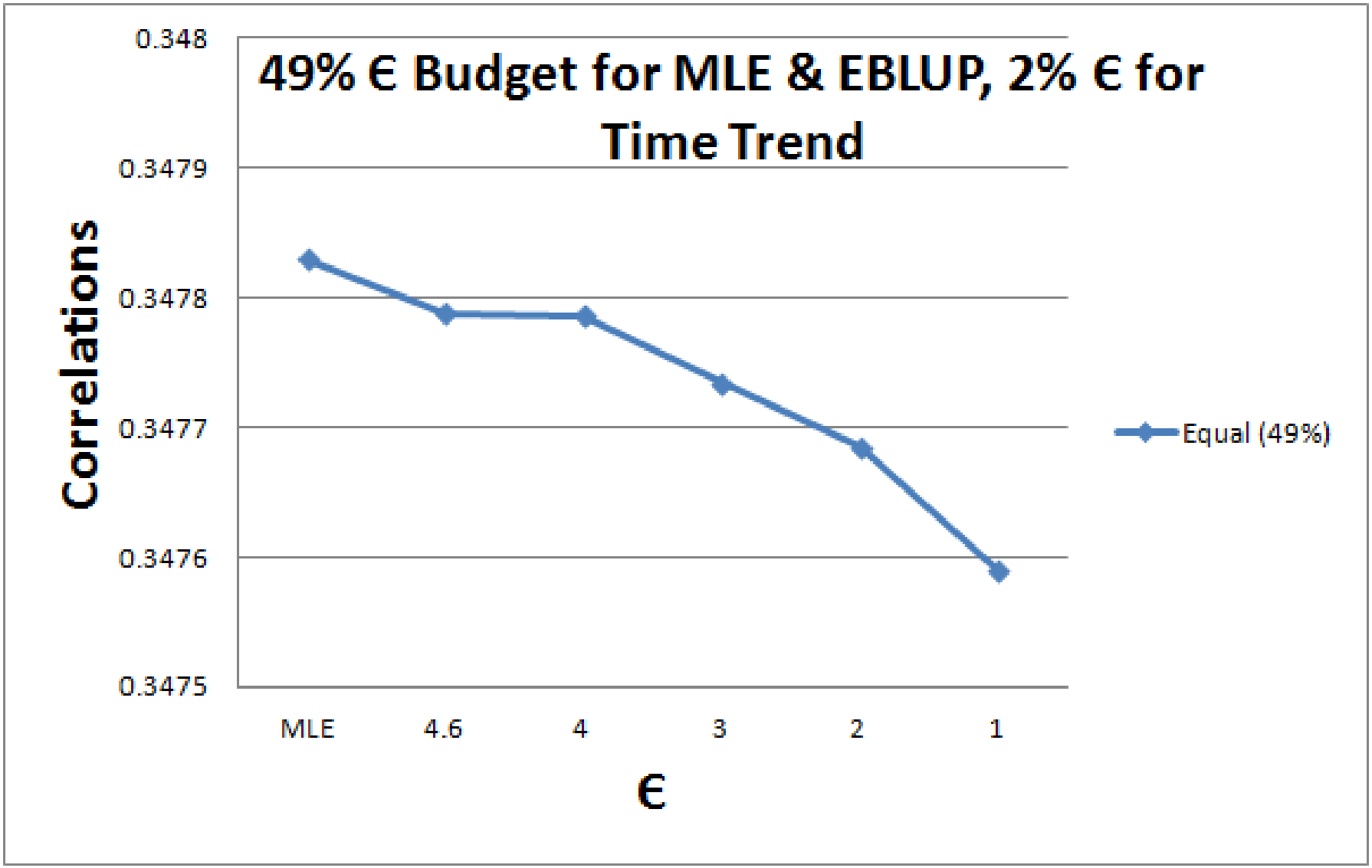
- For all values of ϵ, calculate the predicted rates:

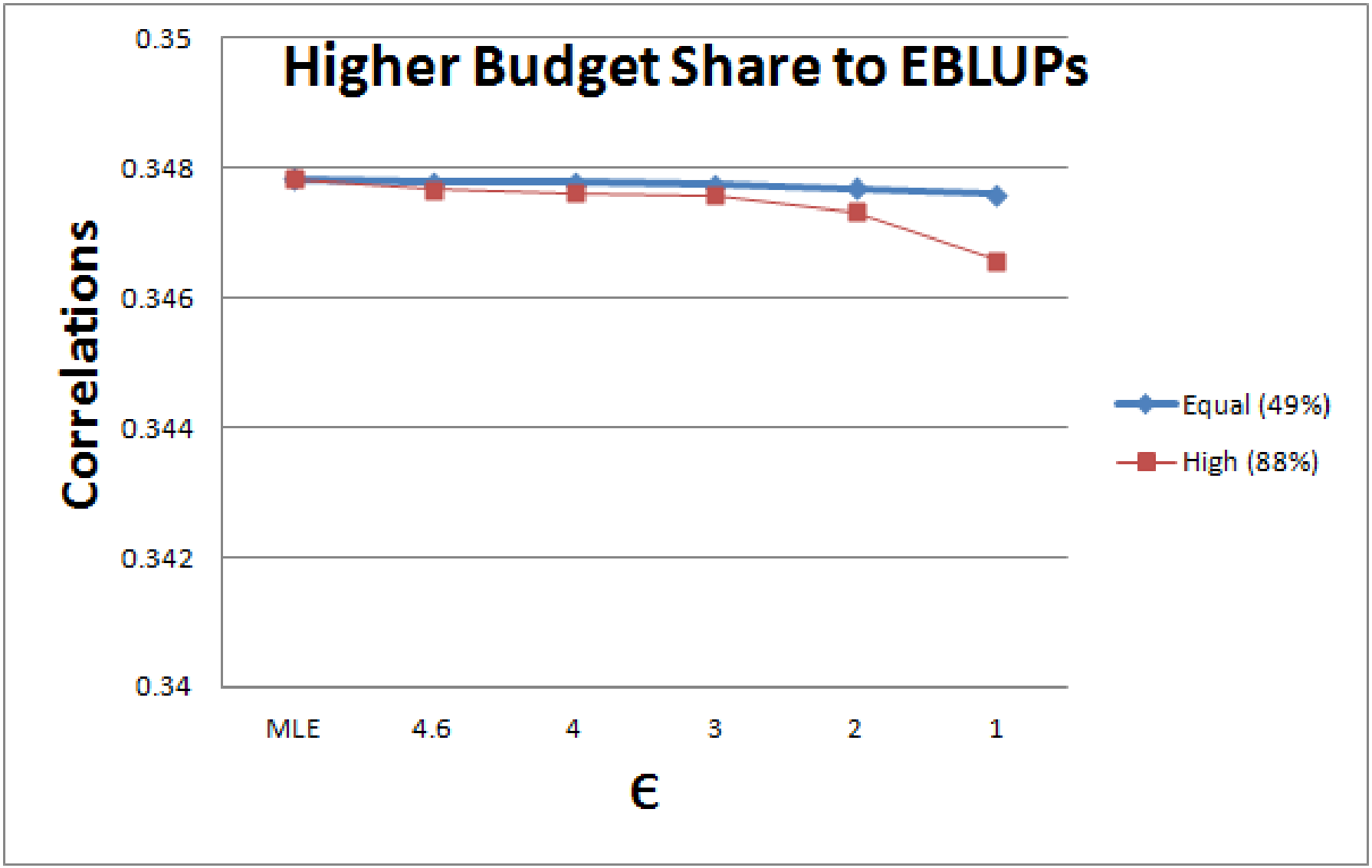
$$JCR^{DP} = \hat{y}^{DP_\epsilon} = X\hat{\beta}^{DP_{.51\epsilon}} + Z\hat{u}^{DP_{.49\epsilon}}$$

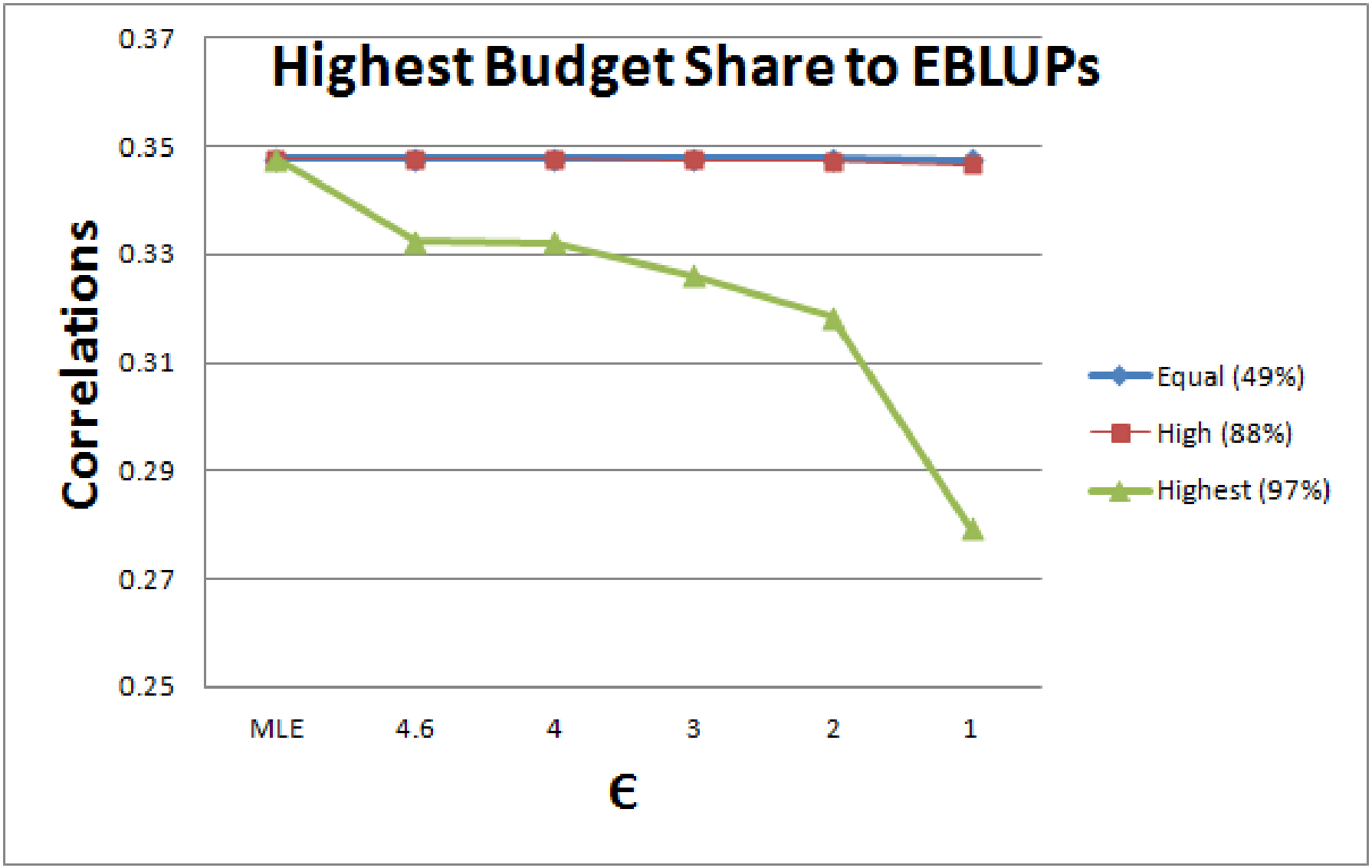
- For $k = 1$ or all of the data, calculate the predicted rates:

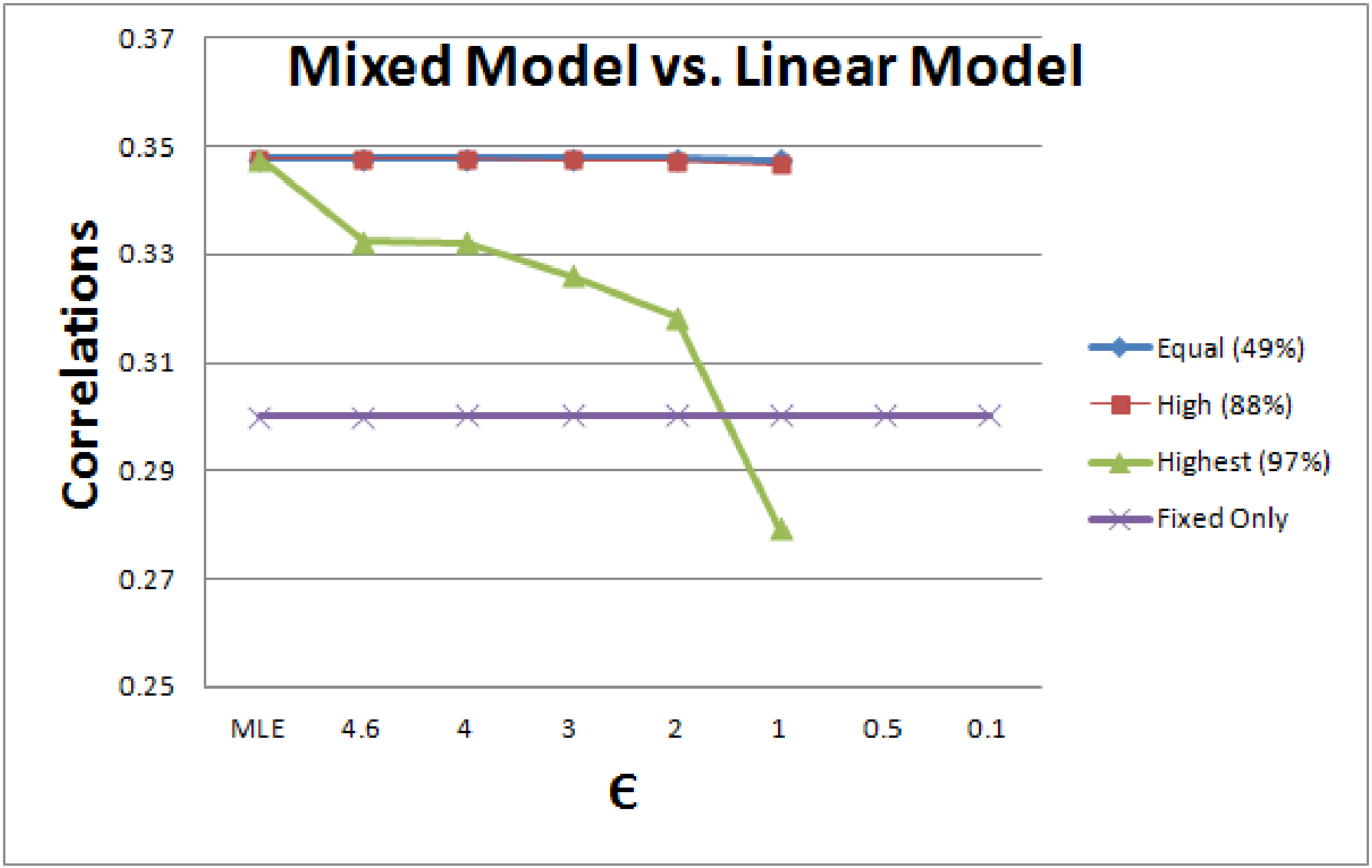
$$JCR^{global} = \hat{y}^{global} = X\hat{\beta}^{global} + Z\hat{u}^{global}$$

- Calculate the correlations between y and $\hat{y}^{global}$, $\hat{y}^{DP_\epsilon}$.
- Finally, plot the correlations as a function of ϵ.


49% Є Budget for MLE & EBLUP, 2% Є for Time Trend
Correlations
0.348
0.3479
0.3478
0.3477
0.3476
0.3475
MLE
4.6
4
3
2
1
Є
Equal (49%)


Higher Budget Share to EBLUPs
Correlations
0.35
0.348
0.346
0.344
0.342
0.34
MLE
4.6
4
3
2
1
ϵ
Equal (49%)
High (88%)


Highest Budget Share to EBLUPs
Correlations
0.37
0.35
0.33
0.31
0.29
0.27
0.25
MLE
4.6
4
3
2
1
ϵ
Equal (49%)
High (88%)
Highest (97%)


Mixed Model vs. Linear Model
Correlations
0.37
0.35
0.33
0.31
0.29
0.27
0.25
MLE
4.6
4
3
2
1
0.5
0.1
ϵ
Equal (49%)
High (88%)
Highest (97%)
Fixed Only

Conclusions

- The estimator is feasible in realistic problems when the random effects are limited to one high-dimensional factor, county in our case, but the non-private MLE doesn't fit well.

Conclusions

- The estimator is feasible in realistic problems when the random effects are limited to one high-dimensional factor, county in our case, but the non-private MLE doesn't fit well.
- For the protection levels that are feasible, the difference between the differentially private estimator and the MLE increases as the protection increases, as shown in our R-U plots.

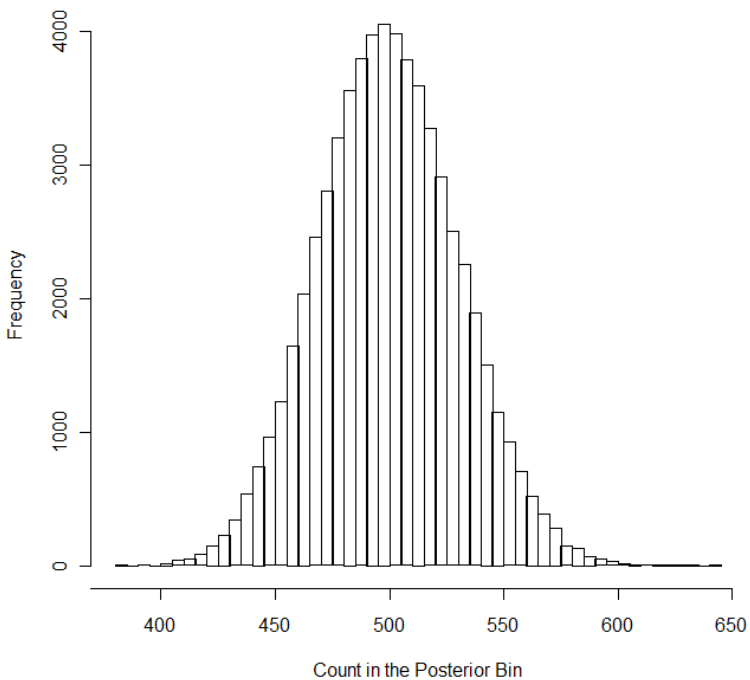
Conclusions

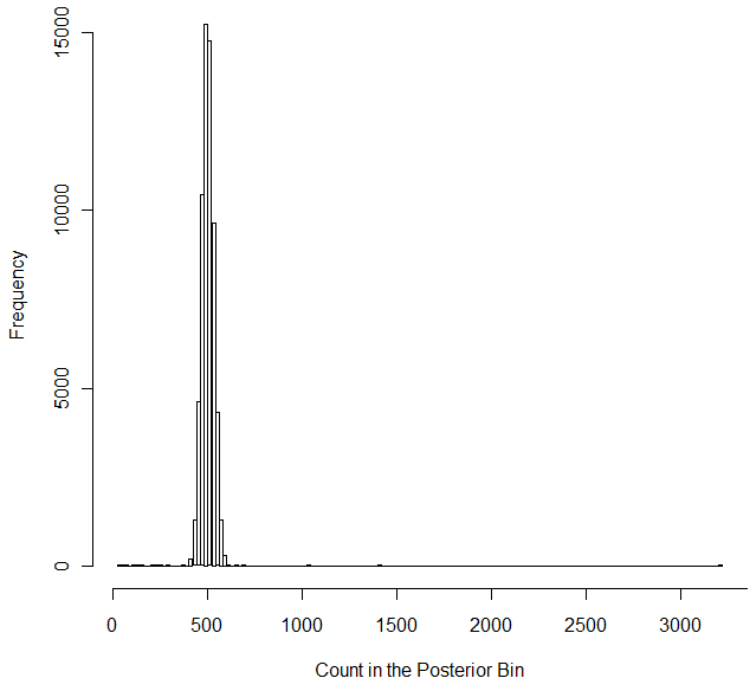
- The estimator is feasible in realistic problems when the random effects are limited to one high-dimensional factor, county in our case, but the non-private MLE doesn't fit well.
- For the protection levels that are feasible, the difference between the differentially private estimator and the MLE increases as the protection increases, as shown in our R-U plots.
- Our problem was chosen to give the differentially private estimator a reasonable chance of success. In particular, the dependent variable was bounded, which is not usually the case in detailed tabulations of continuous data.

Conclusions

- The estimator is feasible in realistic problems when the random effects are limited to one high-dimensional factor, county in our case, but the non-private MLE doesn't fit well.
- For the protection levels that are feasible, the difference between the differentially private estimator and the MLE increases as the protection increases, as shown in our R-U plots.
- Our problem was chosen to give the differentially private estimator a reasonable chance of success. In particular, the dependent variable was bounded, which is not usually the case in detailed tabulations of continuous data.
- The estimator is not likely to work well for cases where there are several factors with many levels, as would be the case in our example if we used both county and detailed industry effects.

Bayesian Model:Posterior Counts in Bins of 500 Normal Variation vs. Missing Most Influential Observation





Future Directions