取得したデータが偏っており限られている場合の是正方法の検討

松香研究室 20L1035Z 毛利太一

1. はじめに

研究で何らかの成果を得るためには、実験参加者の協力が必要となることが多い。実験参加者が偏っている場合や、実験参加者の総数が少ない場合などで実験結果の分析を行うと、研究で本来得たかった結果が得れないのではないかという疑問から本研究を行った。

本研究では、何らかの 2 変数の関係性を推定することを主題とした研究を研究対象とした。y=0.5x に従い乱数を標準化して生成し、全てのデータを母集団と見たてた。母集団の中の特定の範囲(1<x<2)を定め、該当範囲から 30 データ無作為抽出したものを、実際に取得することができたサンプルとし、該当範囲の 30 データから母集団の回帰直線に近づけることが本研究の主題である。

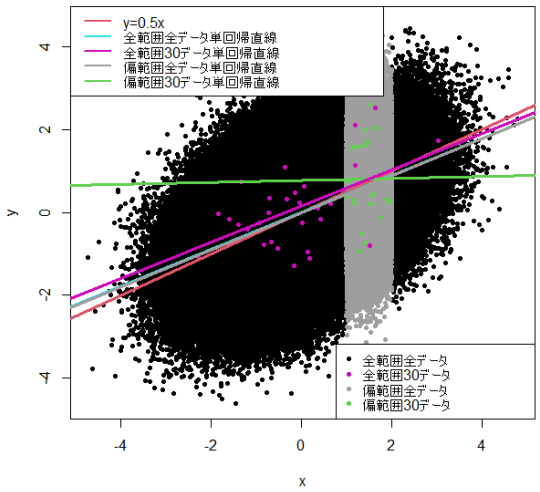


図 1　本研究で用いたモデル

2. シミュレーション 1

2.1. 目的

得られたデータのみで母集団の 2 変数の関係性を推定することが本シミュレーションの目的である。

本研究のシナリオは、「千葉大学生 30 人の実験参加者から、知能指数(IQ 値)とひと月の読書量の関係性を分析する。」であり、説明変数には IQ 値、目的変数にはひと月の読書量とした。

2.2. 方法

MCMC 法を用いたベイズ推定を行った。MCMC 法を用いたベイズ推定は、乱数を用いた試行を繰り返すことで近似解を算出する手法であり、制約を設けることが可能な推定方法であるため採用した。モデルで抽出した 30 データから引き得る回帰線の切片と傾きの組み合わせを制約条件の下乱数によって 4000 組生成した。4000 組の切片と傾きの平均値となる直線の傾きを、30 データの単回帰直線の傾きに近づけることができるかを検証した。

条件は、制約なしの条件、切片に制約を設ける条件、傾きに制約を設ける条件、切片と傾きの両方に制約を設ける条件の 4 条件を行った。

2.3. 結果

切片に制約を設ける条件、傾きに制約を設ける条件、切片と傾きの両方に制約を設ける条件の 3 条件で、母集団の回帰直線に近づけることができた。

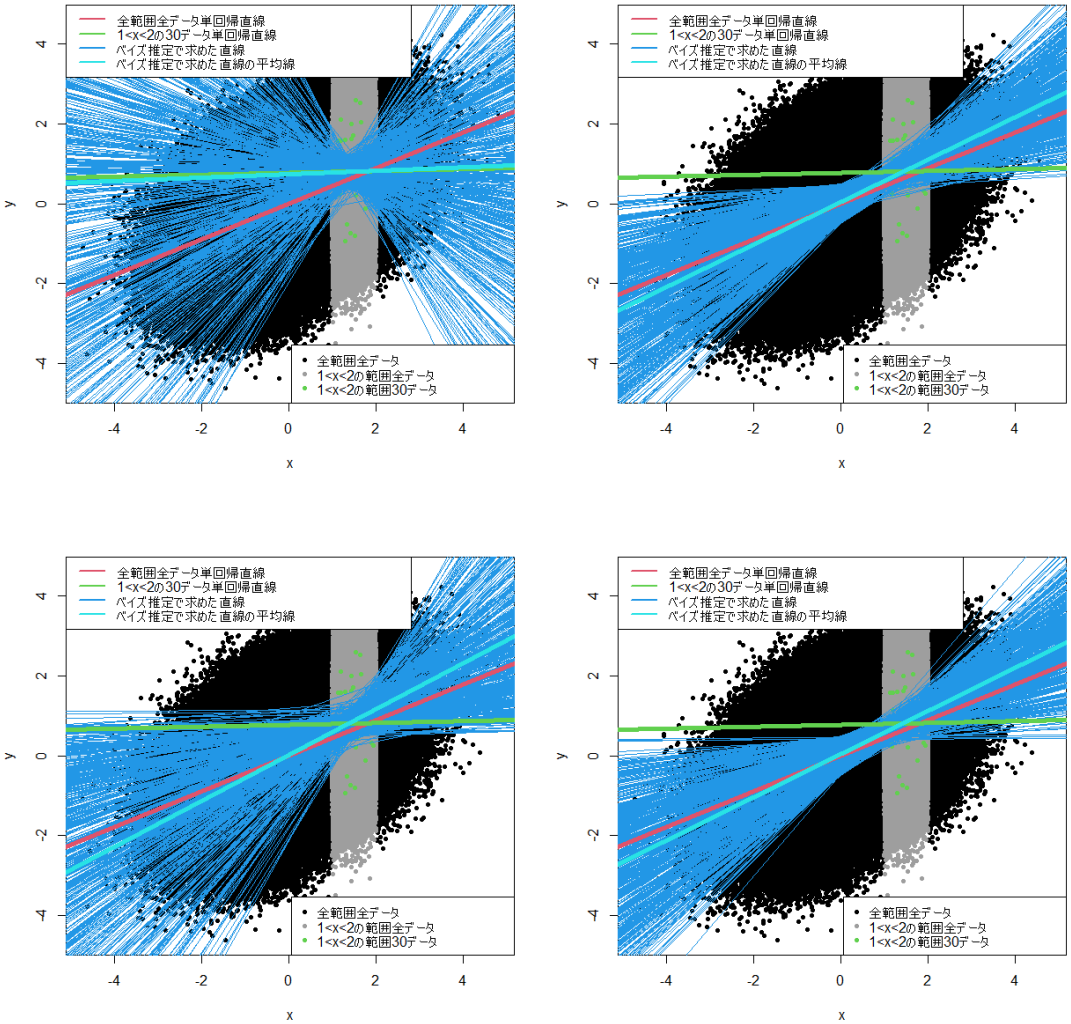


図 2

左上:制約なし条件　右上:切片に制約を設けた条件

左下:傾きに制約を設けた条件　右下:切片と傾きの両方に制約を設けた条件

3. シミュレーション 2

3.1. 目的

実験参加者が偏っており限られている場合を想定して、実験方法を調整することで母集団の関係性を推定することが本シミュレーションの目的である。

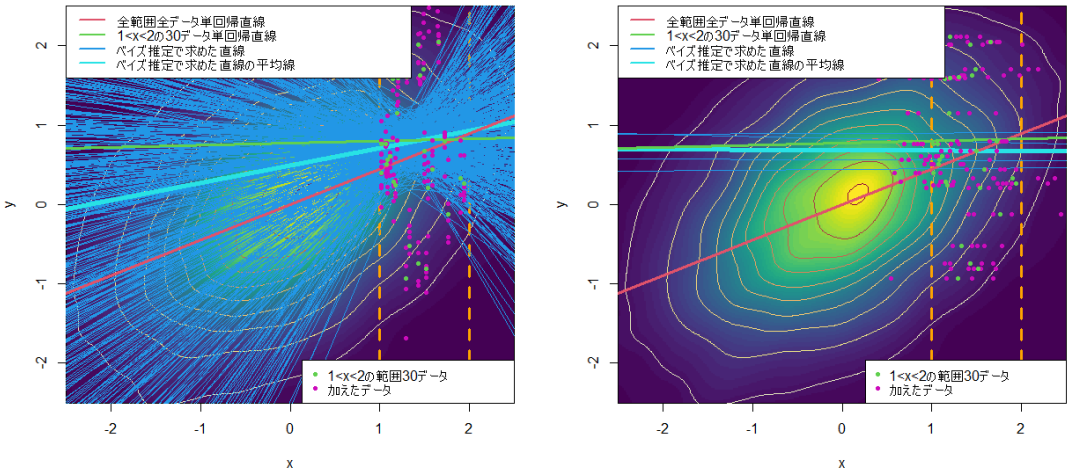
本研究のシナリオは、「千葉大学生 30 人の実験参加者から、2 変数の関係性を分析する。」であった。

3.2. 方法

シミュレーション 1 と同様、MCMC 法を用いたベイズ推定を行った。モデルで抽出した 30 データと、各条件のもと追加されるデータ 120 データから引き得る回帰線の切片と傾きの組み合わせを乱数によって 4000 組生成した。4000 組の切片と傾きの平均値となる直線の傾きを、30 データの単回帰直線の傾きに近づけることができるかを検証した。

3.3. 結果

目的変数を複数回取得する条件と、目的変数と説明変数の両方を複数回取得する条件で、母集団の回帰直線に近づけることができた。



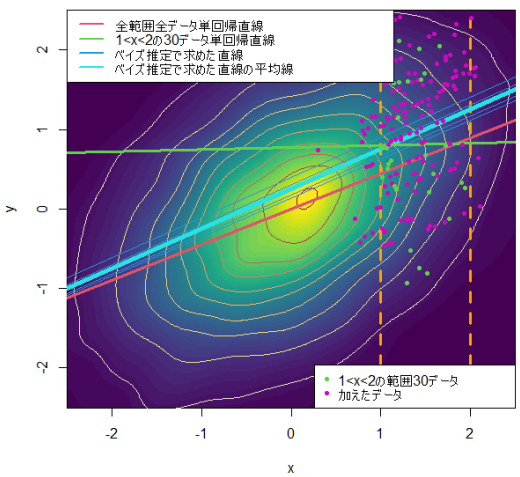


図 3
左上:目的変数を複数回取得した条件　右上: 説明変数を複数回取得した条件
下: 目的変数と説明変数の両方を複数回取得した条件

4. シミュレーション 3

4.1. 目的

シミュレーション 2 の説明変数を複数回取得する場合で、母集団の単回帰直線に近づけることができなかったことから、説明変数を増やすことで母集団の単回帰直線に近づけることが目的である。

4.2. 方法

複数個の説明変数と 1 つの目的変数で傾きを推定し、それが真の傾きの値と近くなるかを誤差と決定係数の観点から検討した。複数個の説明変数は 2 から 10 個まで検討した。

4.3. 結果

説明変数を多く取得すればするほど、真の傾きとの誤差は小さくなり、決定係数が大きくなるという結果が示された。

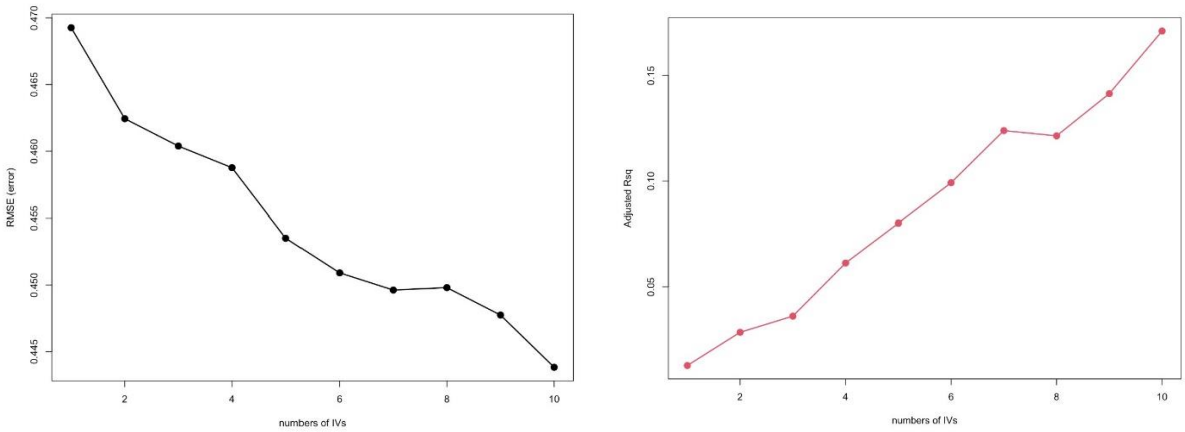


図 3 左:説明変数の数と真の傾きとの誤差　右:説明変数の数と決定係数の値

5. 総合考察

実験参加者が偏っており限られている場合でも、分析方法や実験方法を調整することで、本来得たい母集団の 2 変数の関係性を示すことができた。もちろん、実験参加者の偏りがないように参加を募ることや、1 人でも多くの実験参加者を集めることで、本研究のような検討を行う必要が無い場合も考えられるが、大学生の研究等では難しいことも考えられる。また、十分に実験参加者を集められているならば、本研究は参考にならない可能性がある。しかし本研究によって、分析方法を正しく理解して、場面に応じて使い分けることが必須であることが示された。