データと予測

広島大学 AI・データイノベーション教育研究センター
稲垣知宏

目標

データとデータ間にある関係に基づいた予測について理解する

この授業で紹介すること

- 因果関係と予測
- ばらつきを最小にする直線

キーワード

因果関係、モデル化、最小２乗法

こんなことはありませんか？

出生数が過去最低を記録して少子化対策を急がないといけないと言われているけど、将来のことなんて分かるのでしょうか？

データ間の関係と予測

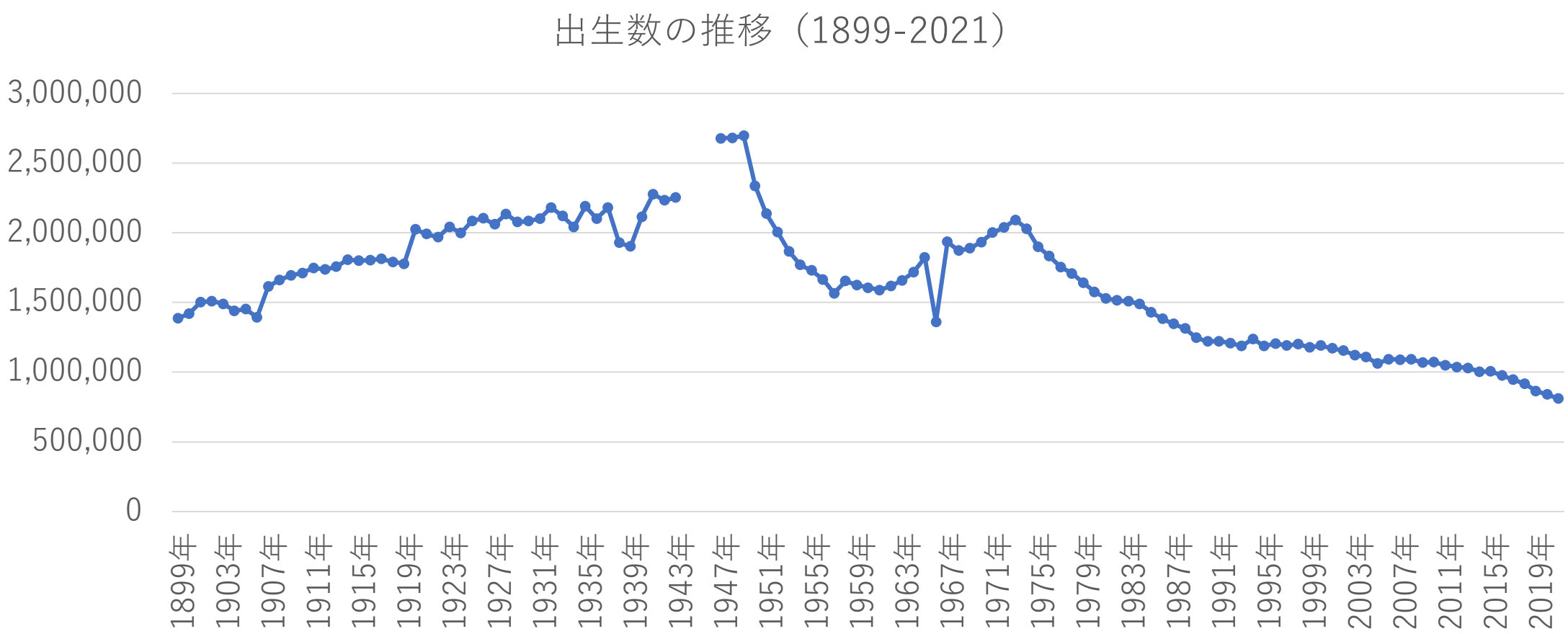
データサイエンスの目的の一つは、対象としている問題に関する情報をデータから読み取ることです。

関係：あるデータと別のデータに関係があるのかないのか

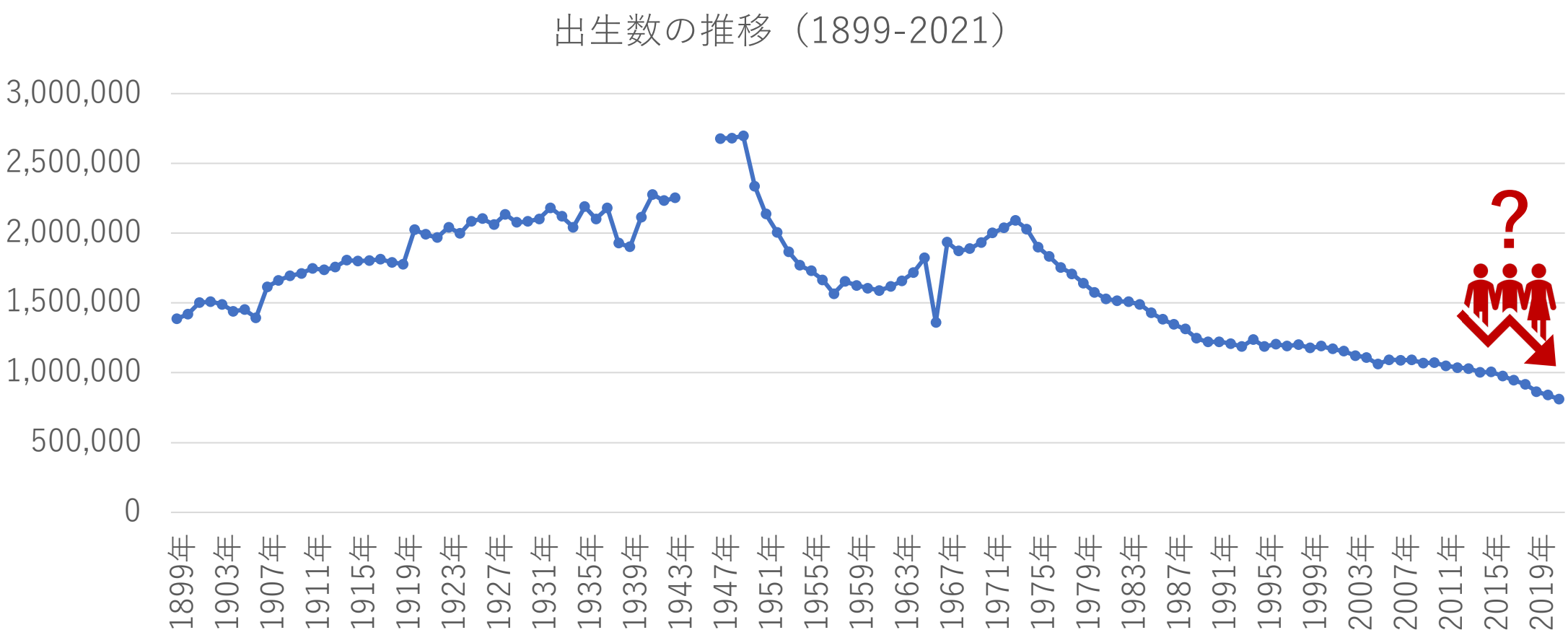
予測：あるデータの変化が、別のデータをどう変えるのか

といったことを、データに基づいて分析します。

出生数の推移

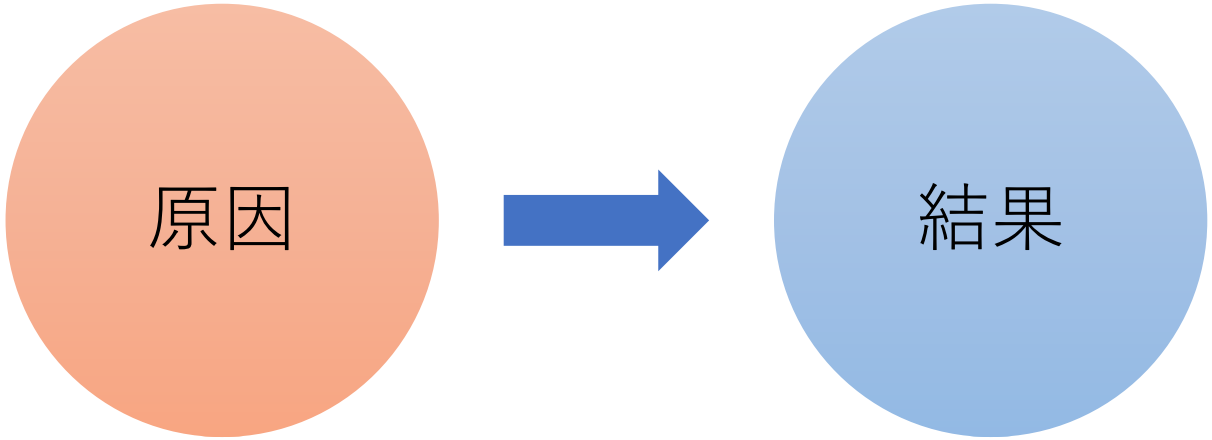

出生数の推移（1899-2021）
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
1899年
1903年
1907年
1911年
1915年
1919年
1923年
1927年
1931年
1935年
1939年
1943年
1947年
1951年
1955年
1959年
1963年
1967年
1971年
1975年
1979年
1983年
1987年
1991年
1995年
1999年
2003年
2007年
2011年
2015年
2019年

出生数の推移


出生数の推移（1899-2021）
3,000,000
2,500,000
2,000,000
1,500,000
1,000,000
500,000
0
1899年
1903年
1907年
1911年
1915年
1919年
1923年
1927年
1931年
1935年
1939年
1943年
1947年
1951年
1955年
1959年
1963年
1967年
1971年
1975年
1979年
1983年
1987年
1991年
1995年
1999年
2003年
2007年
2011年
2015年
2019年
?

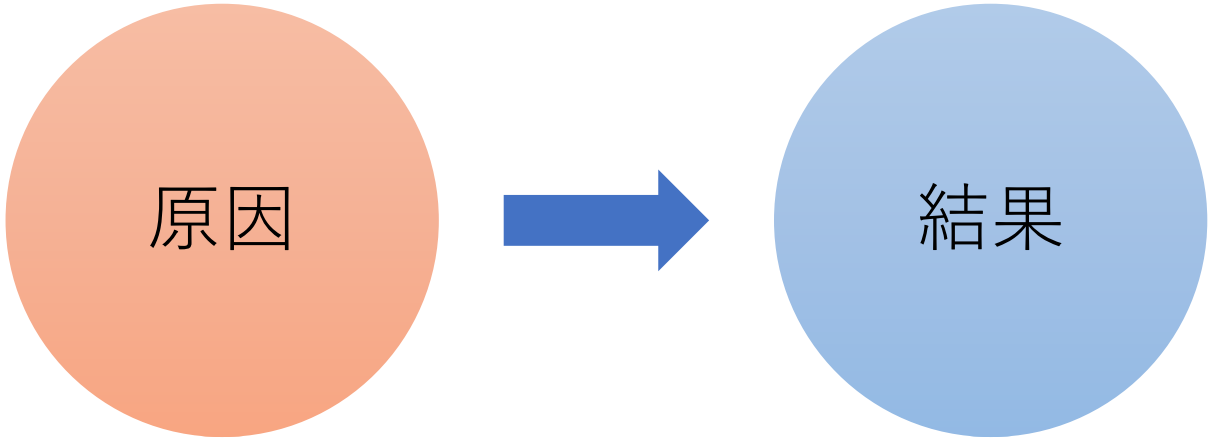
因果関係

どうして出生数は変化するのでしょうか？
出生数変化の原因は何なのでしょうか？
原因とそれによって生じる結果との関係を因果関係と呼びます。



因果関係

どうして出生数は変化するのでしょうか？
出生数変化の原因は何なのでしょうか？
原因とそれによって生じる結果との関係を因果関係と呼びます。



社会現象では、原因は一つとは限らず、さまざまな原因が絡み合うことで複合的にある結果を引き起こすことが一般的です。

例題

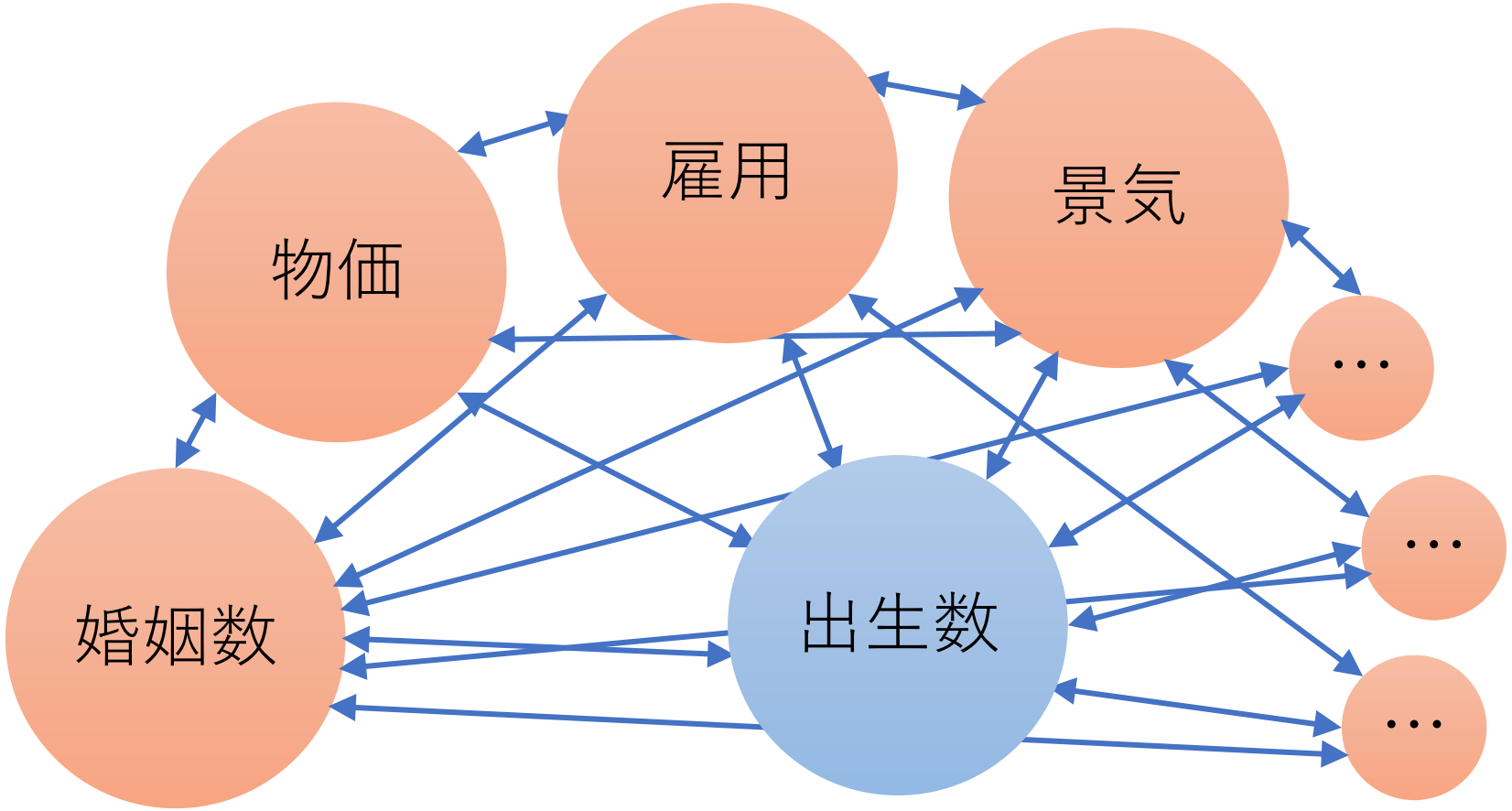
出生数に影響しそうな要因（出会いの機会、安定した就職先など）を、思いつく限り挙げてください。

解説

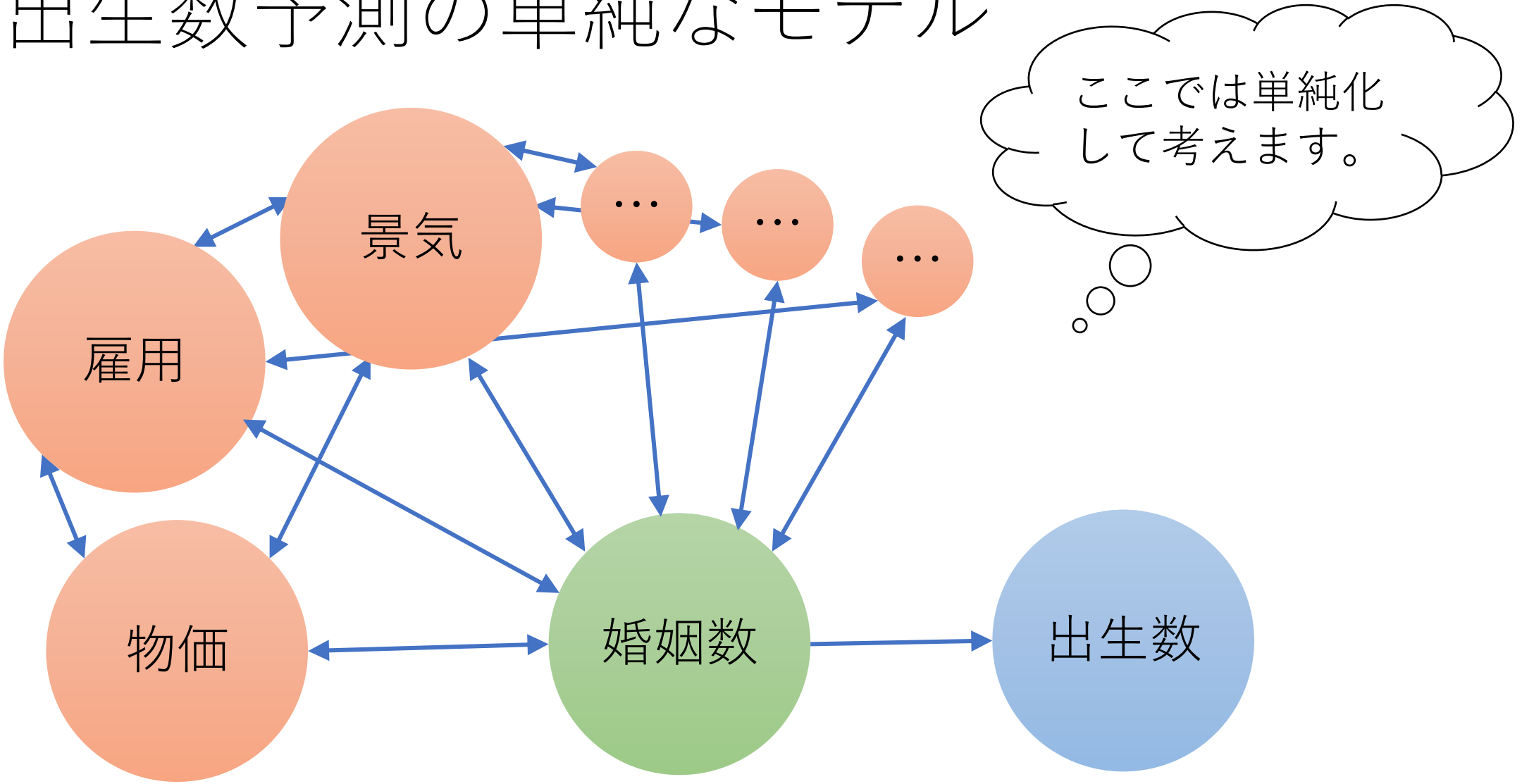
どの様な要因を挙げましたか。以下のいずれかに分類できますか。

- 景気に関する要因
  - ✓物価、雇用、収入など
- 人口に関する要因
  - ✓人口構成、婚姻数、離婚数など
- 政策に関する要因
  - ✓子育て支援、医療費の補助、奨学金など
- その他の要因
  - ✓干支、感染症、災害など

出生数に関わるさまざまな要因



出生数予測の単純なモデル



使えるデータをダウンロード

政府統計のポータルサイト

e-Stat：https://www.e-stat.go.jp/

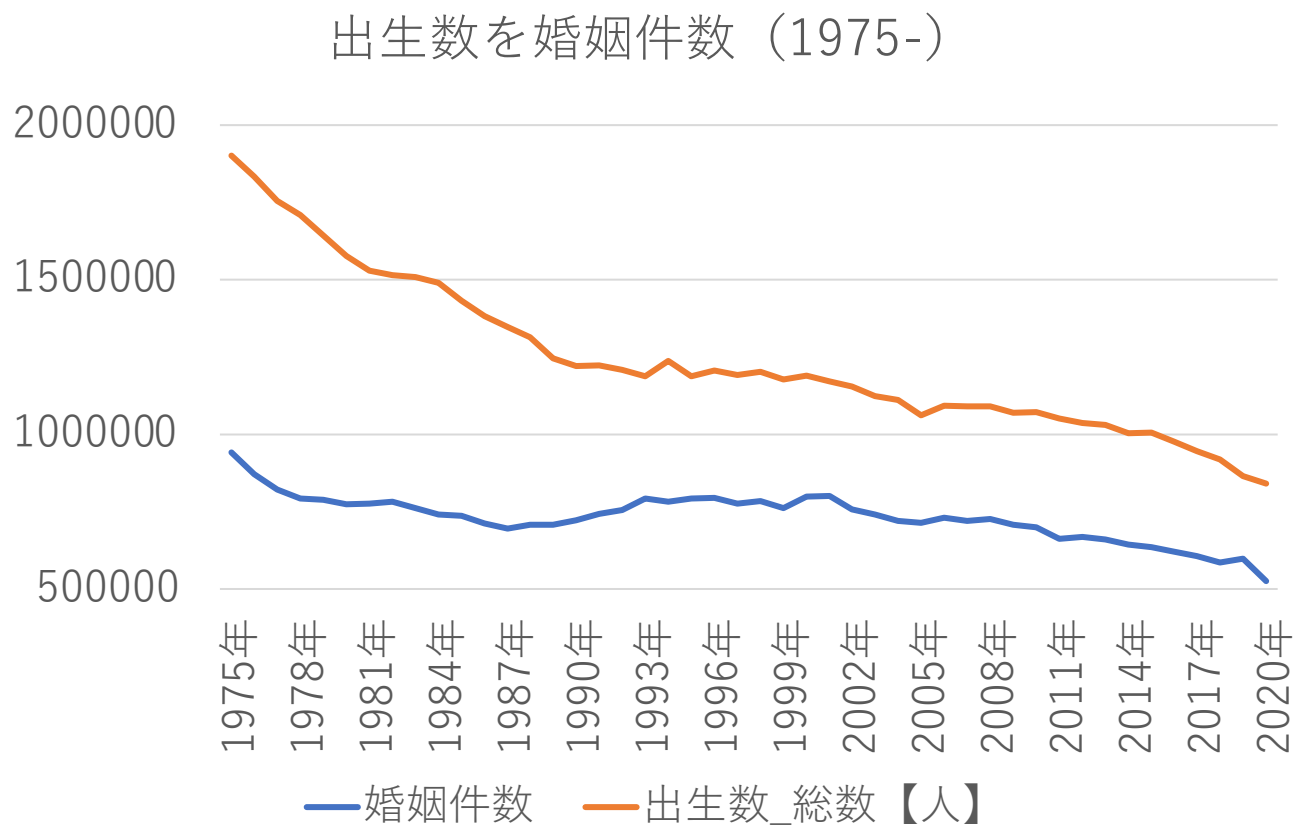
からダウンロードできる、出生数と婚姻数のデータを利用することにします。

政府統計の総合窓口(e-Stat)
（https://www.e-stat.go.jp/）

出生数予測の単純なモデル

出生数を予測するのに、何年か前の婚姻数と関係しているのではないかと考えたとします。

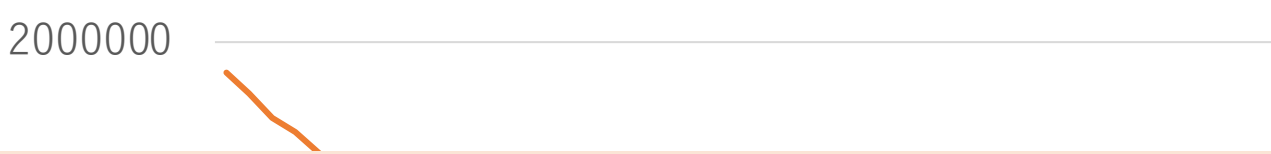
何年前とするかは、ここでは、相関係数が大きくなるかどうかで決めることにします。



出生数予測の単純なモデル

出生数を予測するのに、
何年か前の婚姻数と関係
しているのではないかと
考えたとします。

何年前
では、
なるた
とにし



相関係数は、-1から1の値を取り、２つの異なる尺度間の関係（相関の強さ）を示す。

- 相関係数が-1に近いとき：２つの尺度の一方を増やすともう一方は減る
- 相関係数が1に近いとき：２つの尺度の一方を増やすともう一方も増える

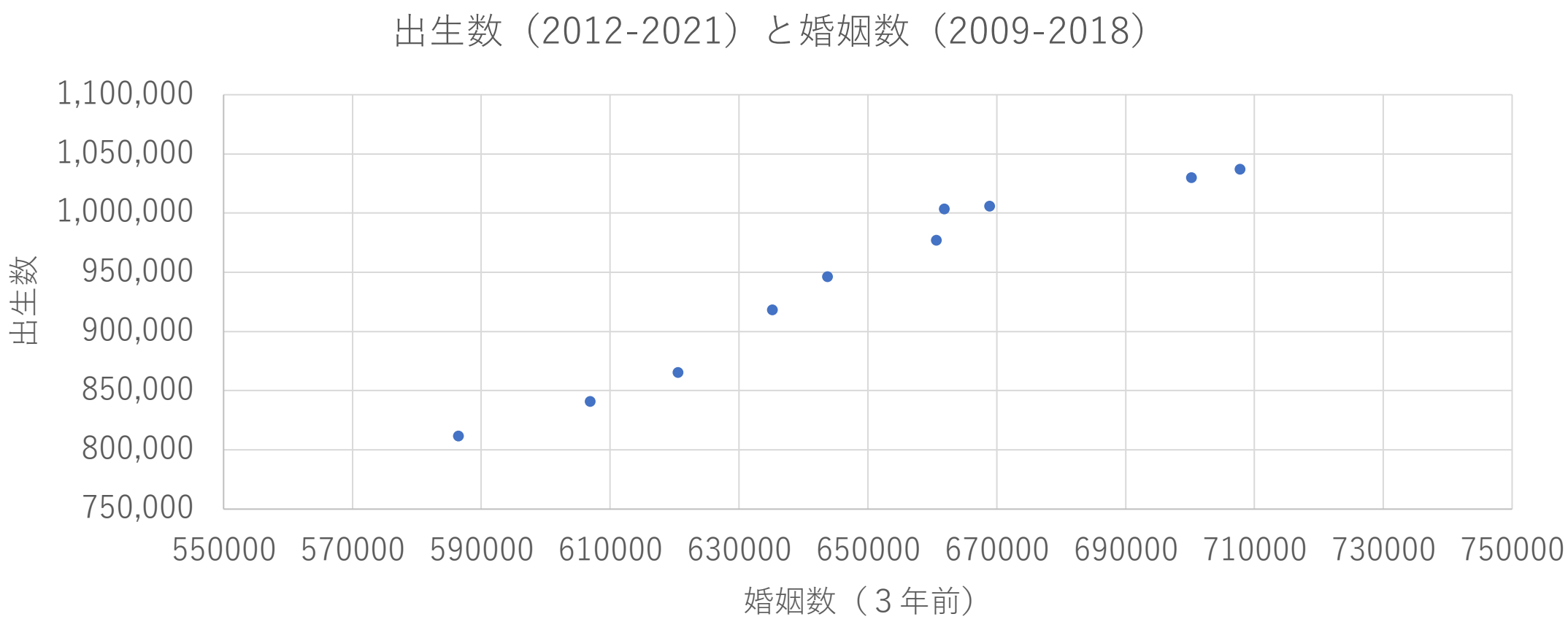
相関が強いのは？

ここで、10年間の出生数とそれと同じ年、1つ前の年、…、７つ前の年の婚姻数で相関係数を計算してみました。下表からは出生数と3年前の婚姻数の相関係数が大きい様です。

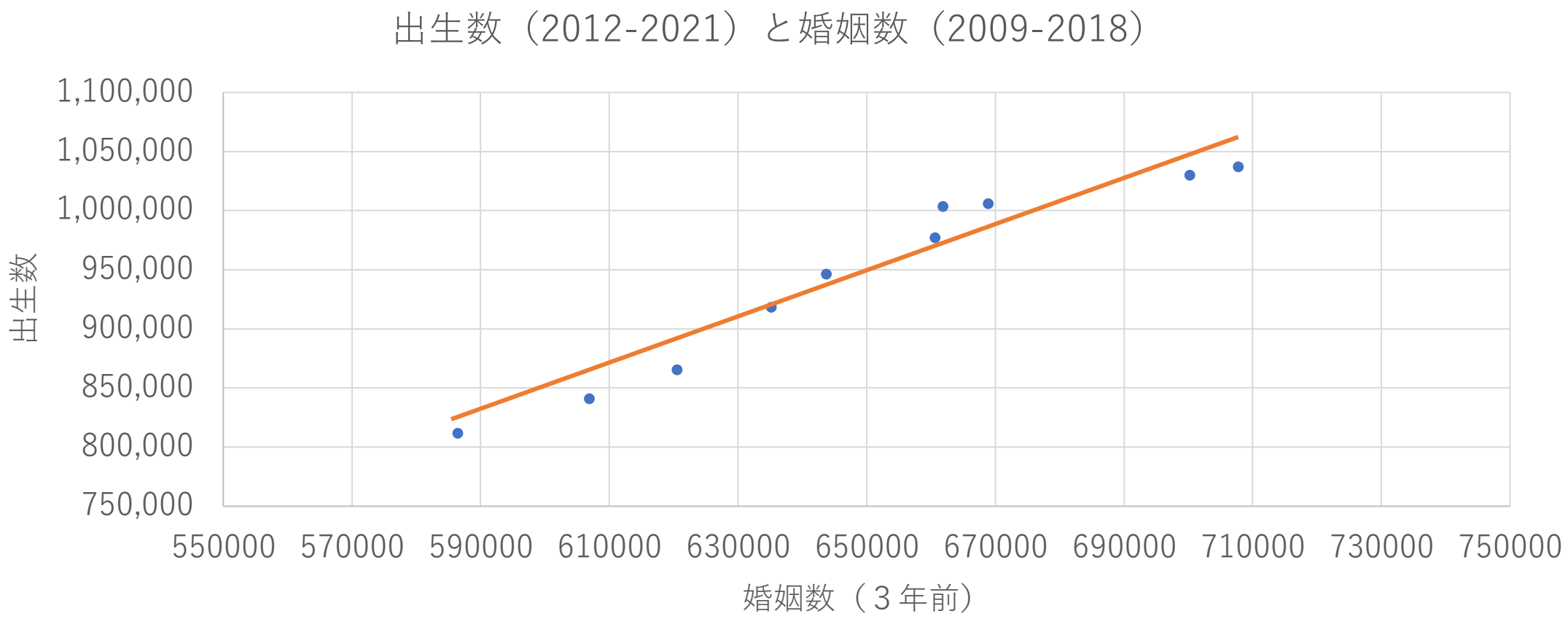
婚姻数

出生数	同年	1年前	2年前	3年前	4年前	5年前	6年前	7年前
2012-2021		0.94068	0.94623	0.96916	0.93642	0.95329	0.95347	0.94797
2011-2020	0.93626	0.93685	0.95606	0.94677	0.92798	0.947	0.93026	0.92006
2010-2019	0.92331	0.96308	0.9403	0.94525	0.92313	0.91938	0.89082	0.91062
2009-2018	0.97257	0.94888	0.95179	0.95696	0.91269	0.88982	0.88187	0.89309
平均	0.94405	0.94737	0.9486	0.95454	0.92505	0.92737	0.91411	0.91794

散布図で確認

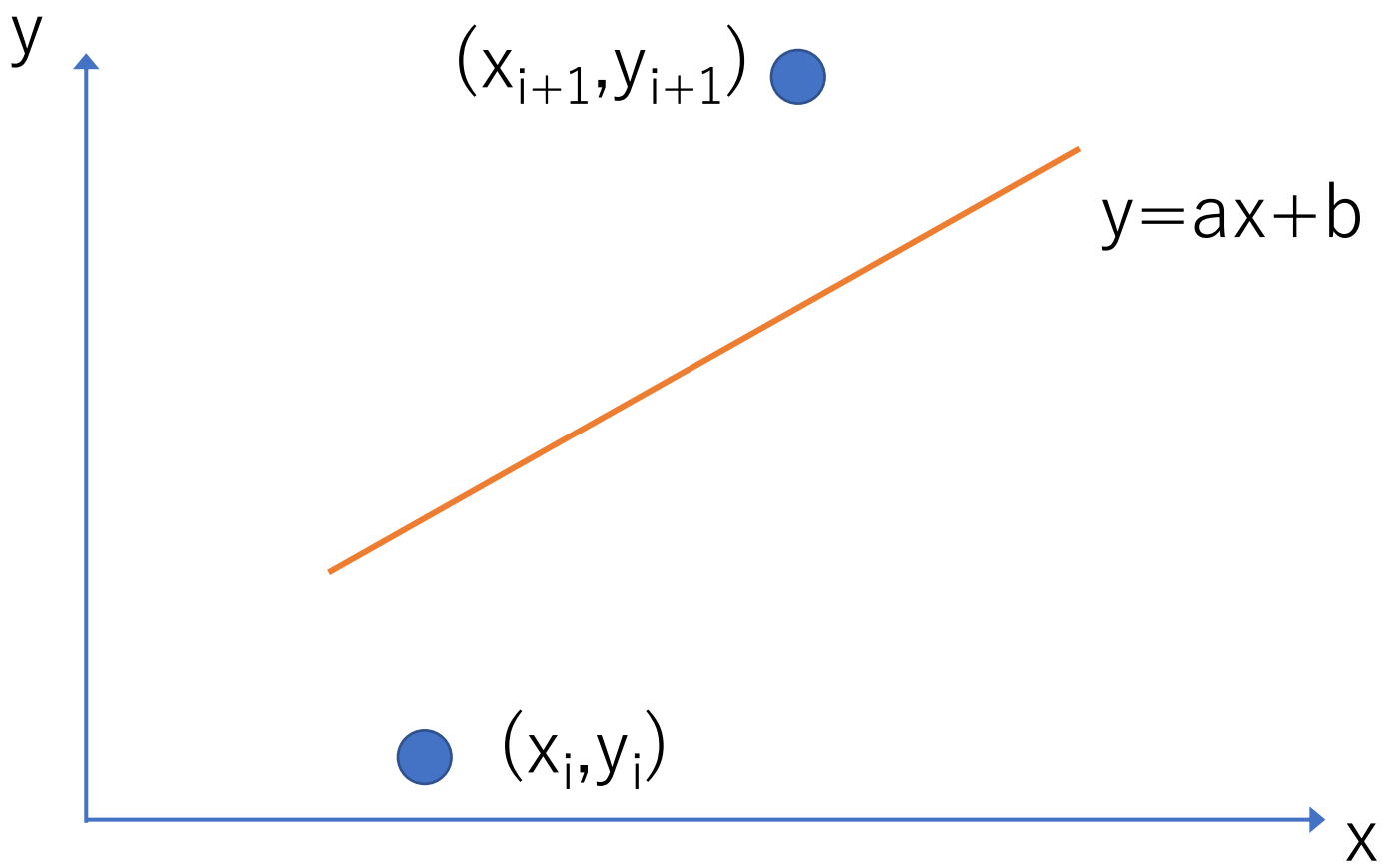

出生数（2012-2021）と婚姻数（2009-2018）
出生数
1,100,000
1,050,000
1,000,000
950,000
900,000
850,000
800,000
750,000
550000
570000
590000
610000
630000
650000
670000
690000
710000
730000
750000
婚姻数（３年前）

直線で近似


出生数（2012-2021）と婚姻数（2009-2018）
出生数
1,100,000
1,050,000
1,000,000
950,000
900,000
850,000
800,000
750,000
550000
570000
590000
610000
630000
650000
670000
690000
710000
730000
750000
婚姻数（３年前）

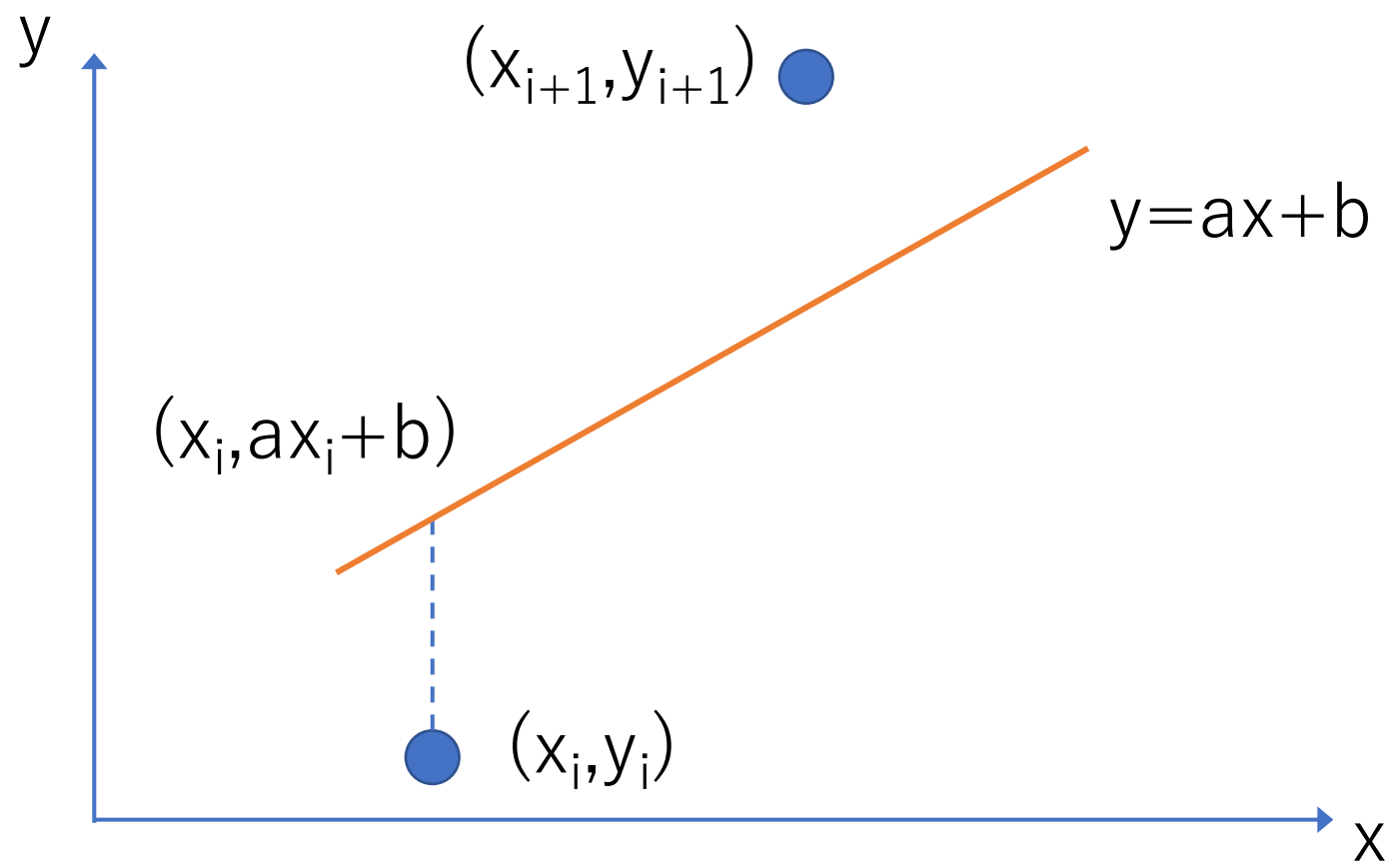
データ点の分布を直線で近似

データ点の分布を近似する直線(y=ax+b)？



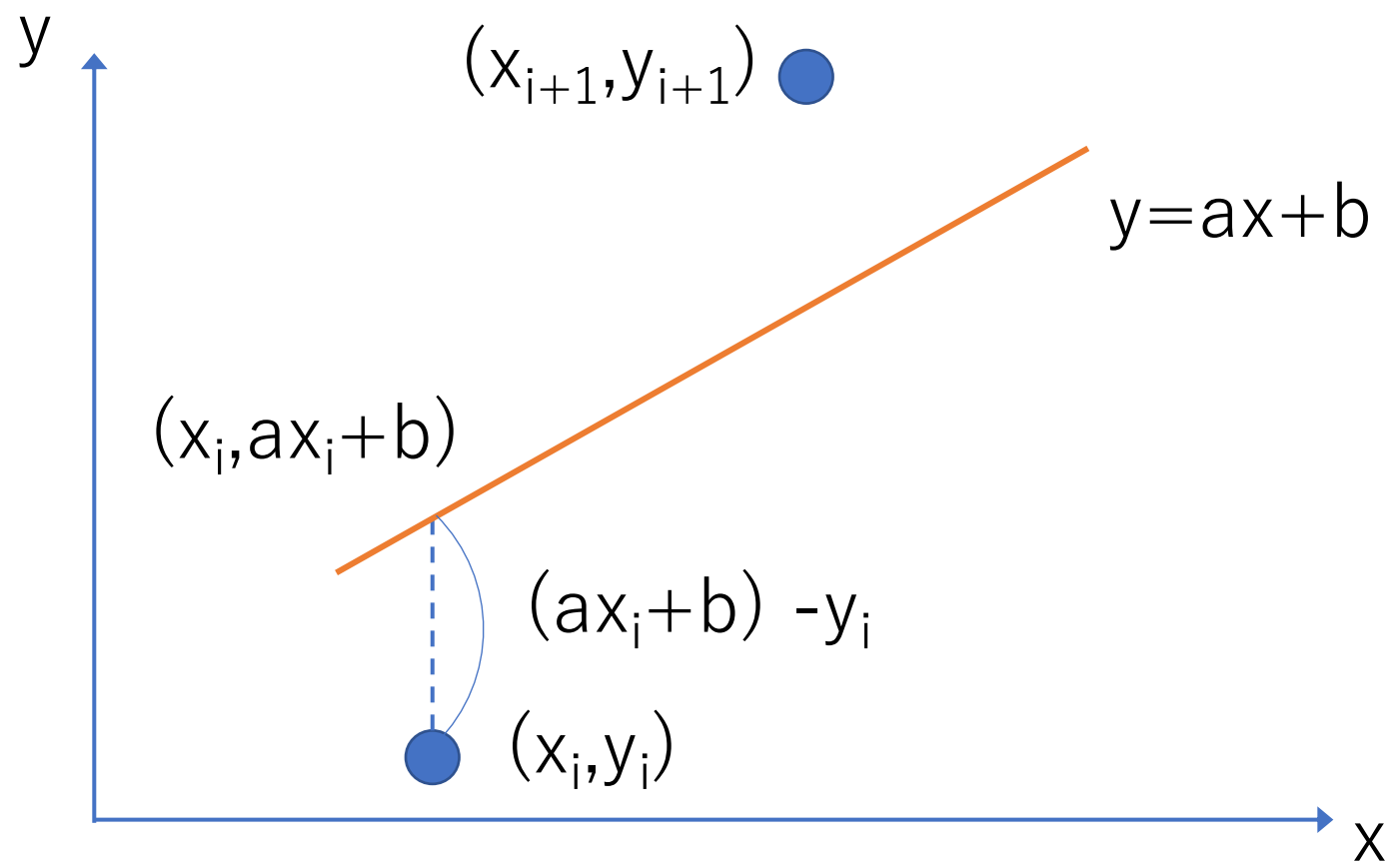
データ点の分布を直線で近似

直線(y=ax+b)とデータ点のばらつきを求める。



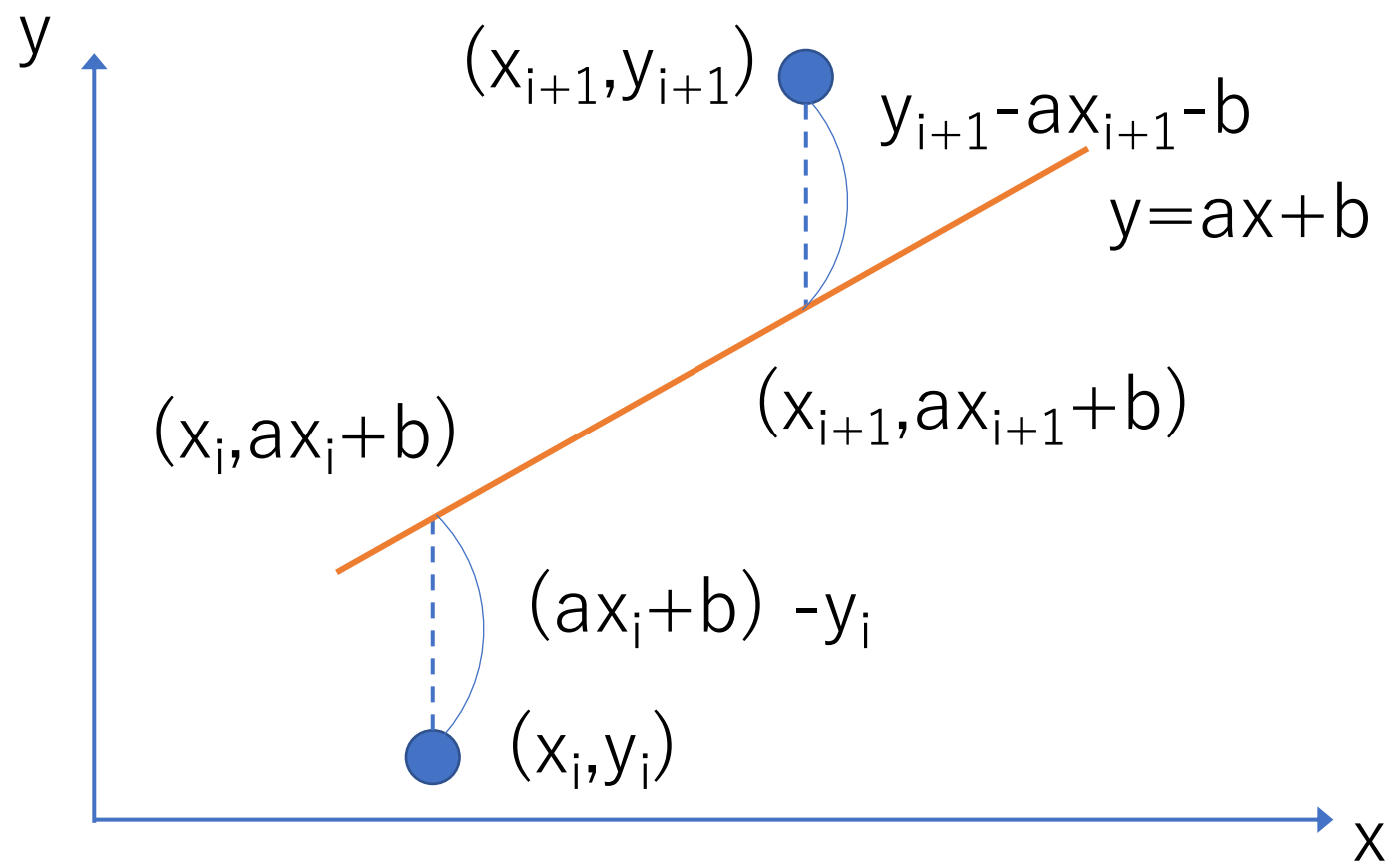
データ点の分布を直線で近似

直線(y=ax+b)とデータ点のばらつきを求める。



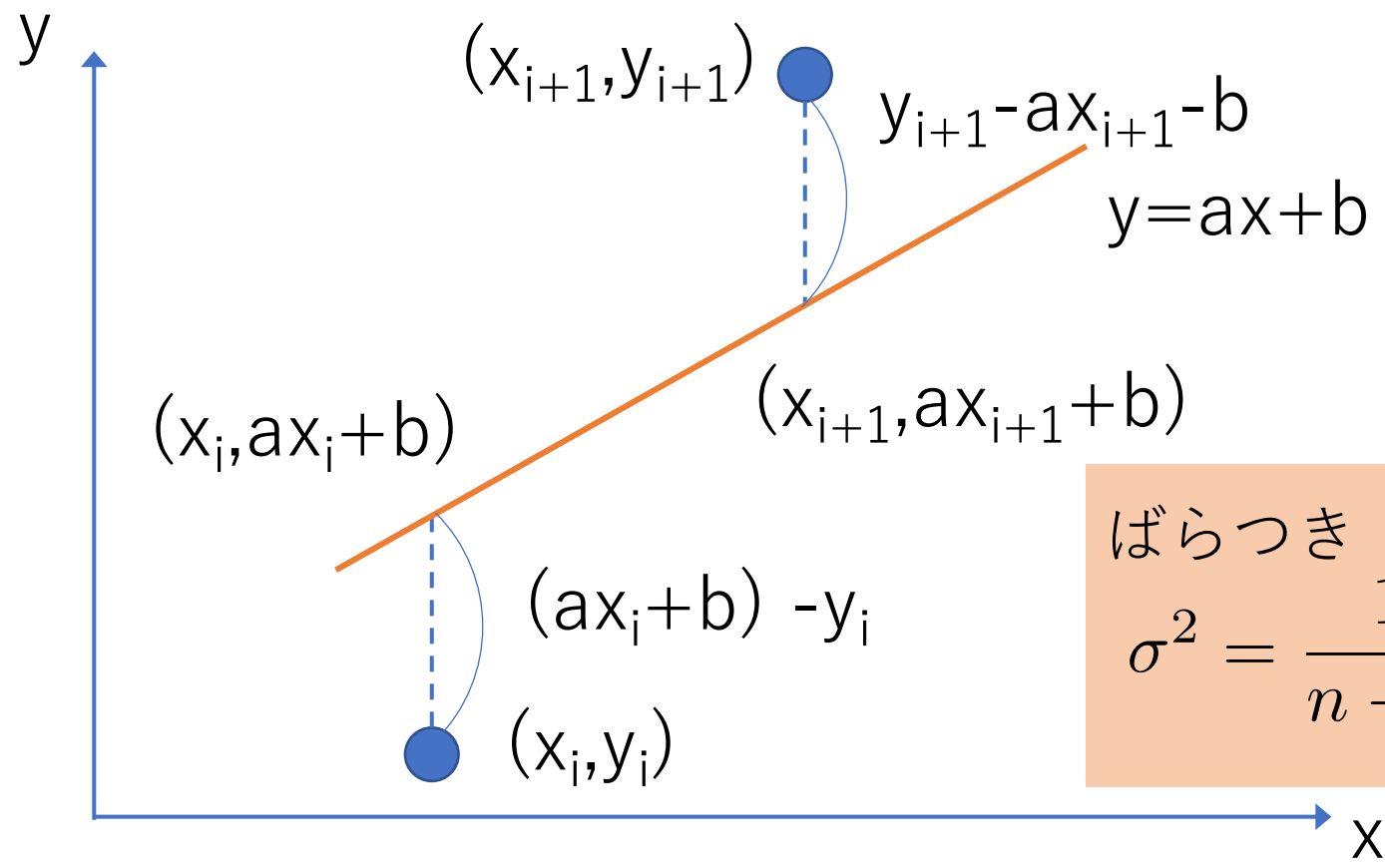
データ点の分布を直線で近似

直線(y=ax+b)とデータ点のばらつきを求める。



データ点の分布を直線で近似

直線(y=ax+b)とデータ点のばらつきを求める。



ばらつき

$$\sigma^2 = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - ax_i - b)^2$$

最小２乗法：ばらつきを最小にする

傾きa:

$$a = \frac{(x\text{と}y\text{の共分散})}{(x\text{の分散})}$$

切片b:

$$b = (y\text{の平均値}) - a\times(x\text{の平均値})$$

最小２乗法：ばらつきを最小にする

傾きa:

$$a = \frac{(x\text{と}y\text{の共分散})}{(x\text{の分散})}$$

切片b

xの分散：$(x_i - (x\text{の平均値}))^2$の平均
yの分散：$(y_i - (y\text{の平均値}))^2$の平均
xとyの共分散：
$(x_i - (x\text{の平均値}))\,(y_i - (y\text{の平均値}))$の平均

最小２乗法：ばらつきを最小にする

傾きa:

$$a = \frac{(x\text{と}y\text{の共分散})}{(x\text{の分散})}$$

切片b:

$$b = (y\text{の平均値}) - a\times(x\text{の平均値})$$

出生数を予想するモデル

モデル：(出生数) $= a\times$(3年前の婚姻数) $+ b$

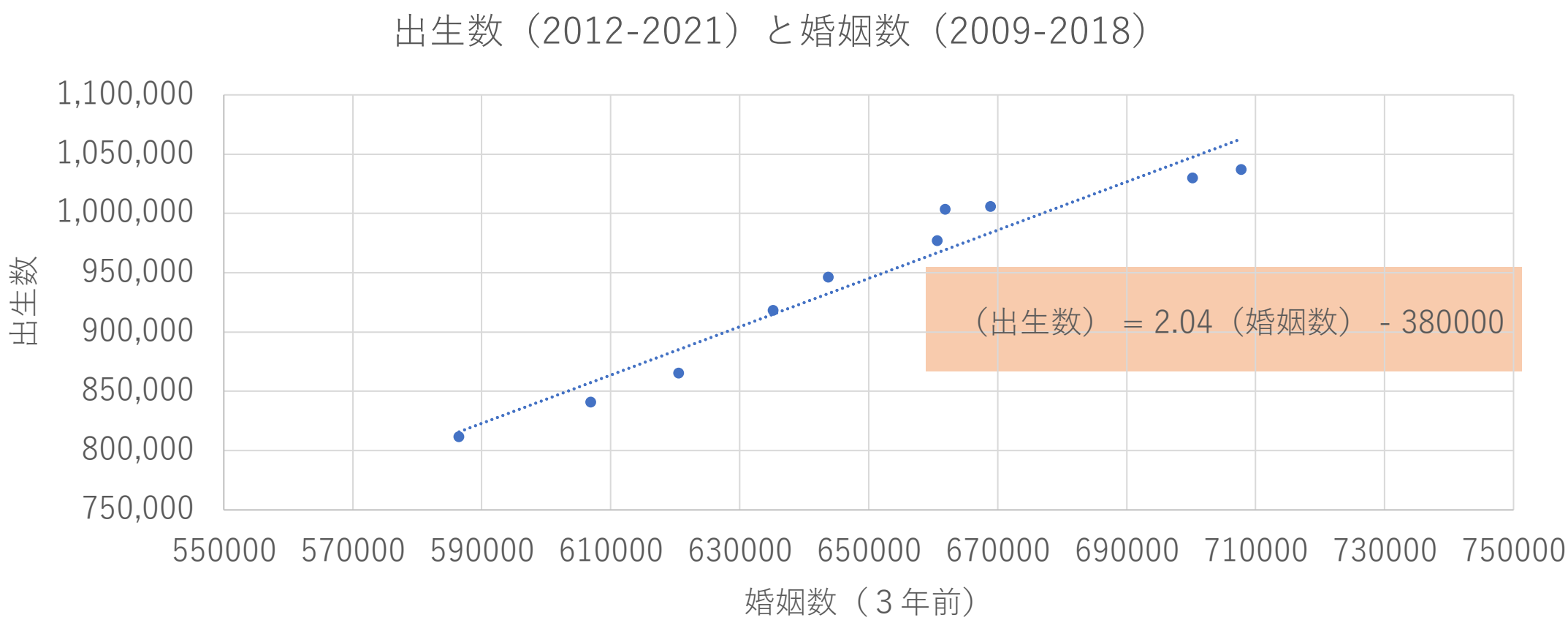
出生数の平均値（2012-2021）：944,000
婚姻数の平均値（2009-2018）：649,000
婚姻数の分散（2009-2018）：1,350,000,000
出生数と婚姻数の共分散：2,750,000,000

出生数の標準偏差（分散の平方根）が77,000程度ですので、有効数字３桁で計算しています。

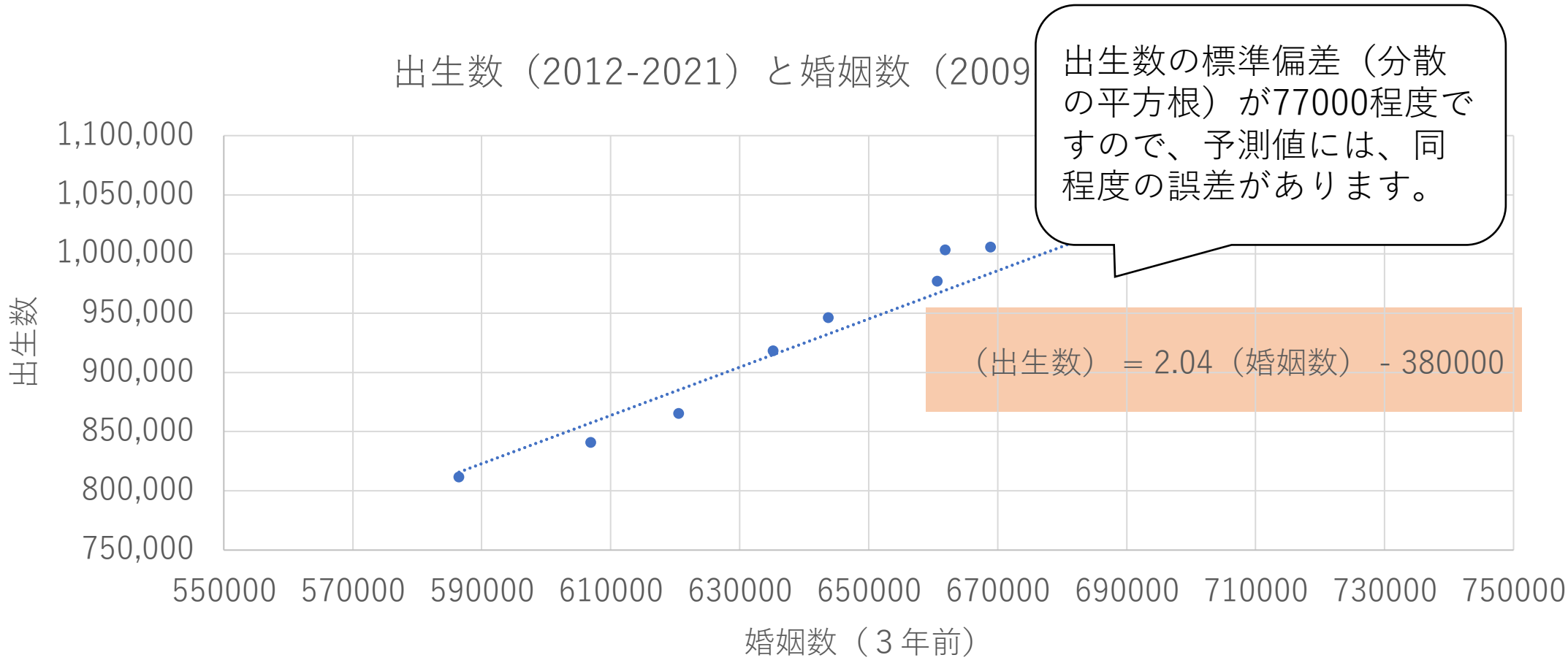
$$a = \frac{2{,}750{,}000{,}000}{1{,}350{,}000{,}000} = 2.04$$

$$b = 944{,}000 - 2.04 \times 649{,}000 = -380{,}000$$

出生数を予想するモデル



出生数を予想するモデル



問題

最新の婚姻数を調べて、その3年後の出生数を予想しなさい。