DataScientist Society

変数変換による線形加工

～ロジスティック回帰(または線形回帰モデル)での前処理～

ロジスティック回帰モデルは何を前提にしているのか？

- ロジスティック回帰モデル

pは$y = 1$となる確率.

$$\log\left(\frac{p}{1-p}\right) = a_1x_1 + a_2x_2 + \cdots + a_nx_n + b$$

→ 対数オッズを線形和と切片で表現している

- 参考：線形回帰モデル

$$y = a_1x_1 + a_2x_2 + \cdots + a_nx_n + b$$

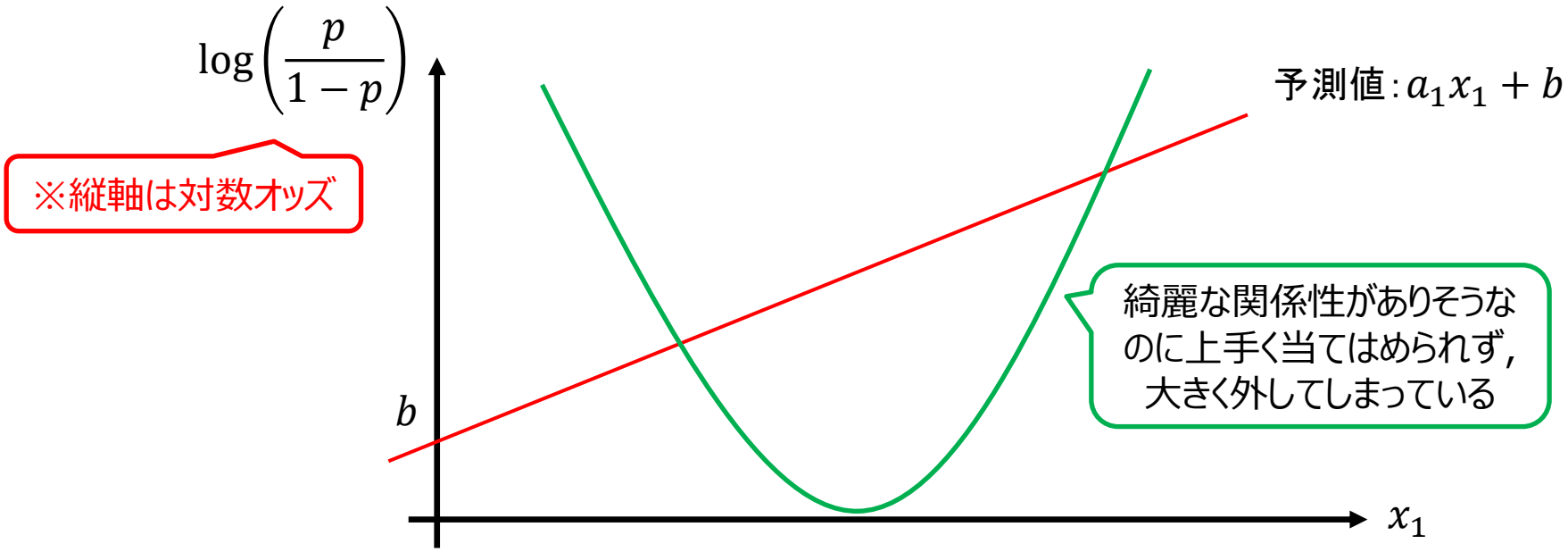
→ 目的変数を線形和と切片で表現している

上手く予測できないケース

簡単のため, 1変数の場合について考える.

$$\log\left(\frac{p}{1-p}\right) = a_1 x_1 + b$$

対数オッズと説明変数x_1との関係性は1次関数であれば当てはまりが良いが, そうでない場合(例えば2次関数), 上手く当てはめることができず, 予測精度も悪くなる.



対処法：変数変換

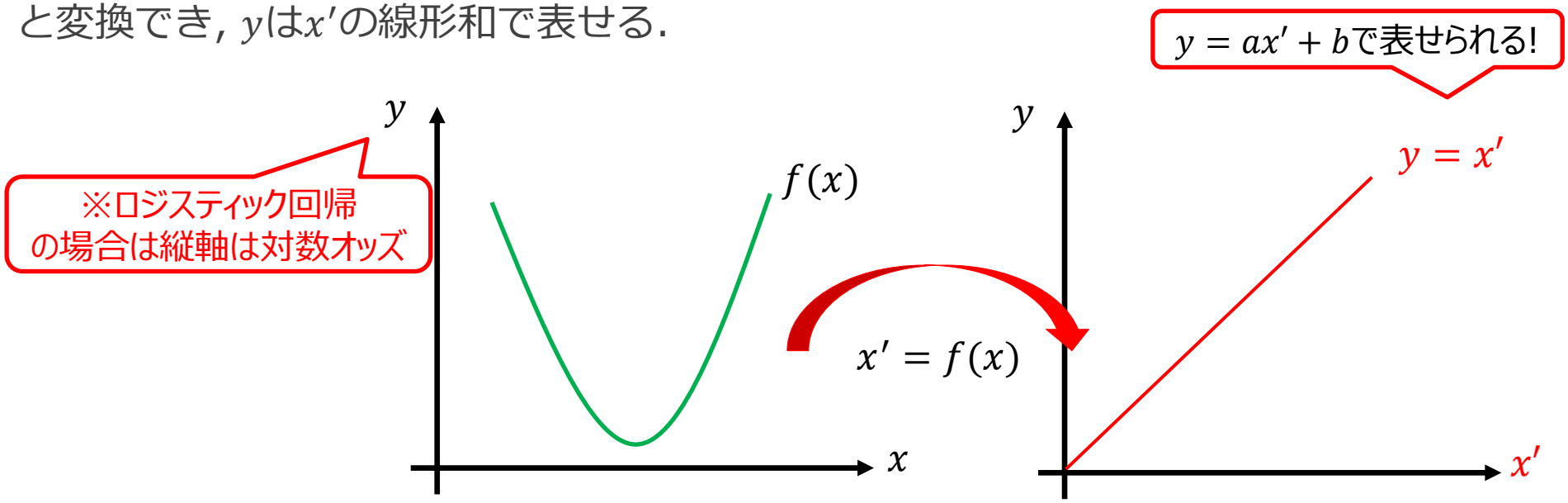
ロジスティック回帰だけでなく，線形回帰も同様のため，この後は$y = ax + b$として話を進めていく．近似したい(予測したい)任意の関数を$f(x)$とすると

$$y = f(x)$$

となるので，$x' = f(x)$と変数変換すると，

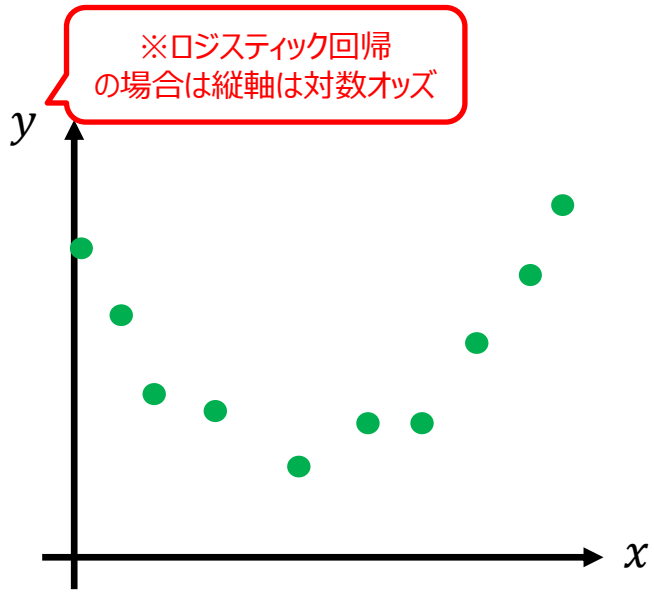
$$y = f(x) \Leftrightarrow y = x'$$

と変換でき，yはx'の線形和で表せる．



どうやって$f(x)$を見つけるのか？

学習データを用いて，y(対数オッズ)と説明変数の関係をグラフとして可視化し，人間がグラフを見て判断する．



グラフを見て，$y = x^2$っぽいなと考え

- $x' = x^2$と変数変換する
- x'でモデルを作成する
- 検証データで精度を確認する

上記ステップで精度向上が確認できたらその加工を採用する

y(対数オッズ)と説明変数の関係を可視化したけど，複雑な関数だった場合，その関数を思いつかない，どうすれば…

→　多くは次頁の関数パターンとなる場合が多い，またあまりにも複雑な関係性のものはロジスティック回帰や線形回帰で予測することが適していない

対数オッズとの関係をどうやって可視化するのか？

対数オッズは目的変数である$y = 1$となる確率pを使用している. このpはたいていの場合, データとしては存在しない. したがって, 擬似的にpを計算する必要がある.

カテゴリ変数は, そのままカテゴリごとに

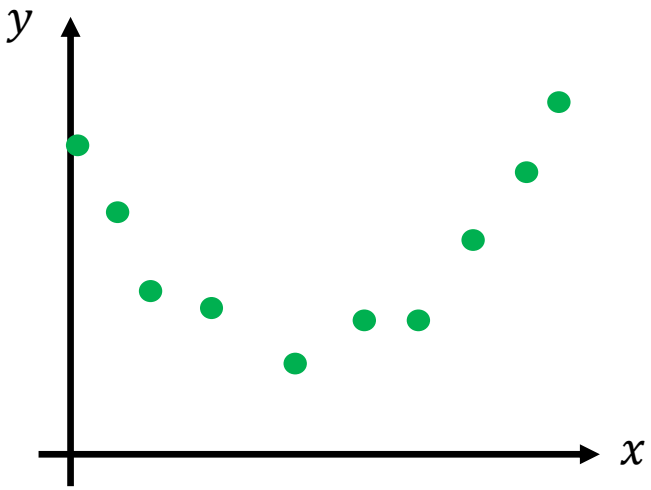
$$p = (\text{該当カテゴリの} y = 1 \text{となる割合})$$

とし, 連続変数の場合はある程度カテゴリ化して同様にpを計算する.

→ 年齢を年代にカテゴリ化

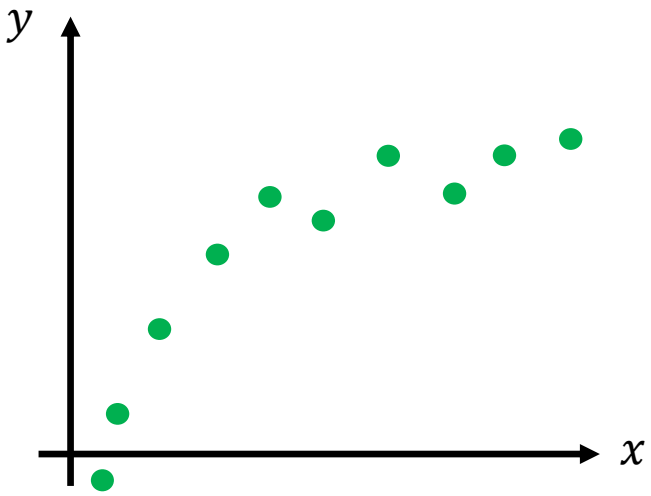
→ 身長を150cm台, 160cm台, … と10cm刻みでカテゴリ化

良く出てくる関係性



二次関数(または絶対値)型:$x' = (x-c)^2+b, x' = |x-c|+b$

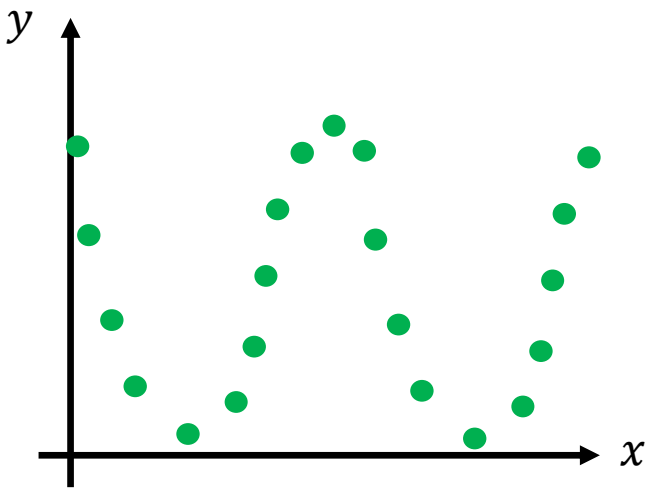
- 線形回帰:電気料金と気温の関係
  - 冬になると暖房で電気料金が上がり, 夏になると冷房で電気料金が上がる
- ロジスティック回帰:保険料の発生確率の対数オッズと年齢
  - 年齢が低いとケガや病気をしやすく, 成人するとあまりケガや病気をしにくくなるが, 高齢者になるとまた病気やケガをしやすくなる



対数型:$x' = \log(x+c)+b$

- 線形回帰:アイスの売り上げと気温の関係
  - 気温が上がるほど, アイスの売り上げは上がるが, 35度を超えると頭打ちになる
- ロジスティック回帰:ある住宅購入確率の対数オッズと年収
  - 年収が多くなるほど, ある住宅(5000万円)を購入する意欲は高くなるが, 年収1億円超えた所から変わらなくなる

良く出てくる関係性



周期関数型：使用する変数を再考してみる

- 線形回帰：電気料金と年月の関係
  - 冬になると暖房で電気料金が上がり，夏になると冷房で電気料金が上がる

  → こちらは使用する説明変数が下手で，本質は気温との関係