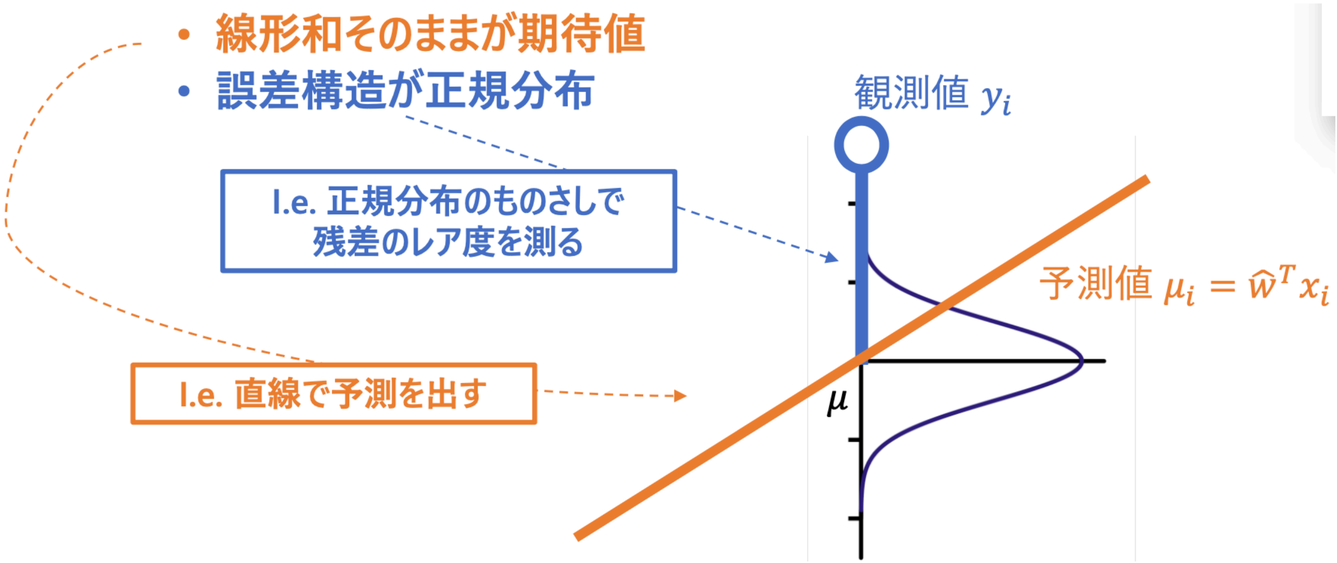
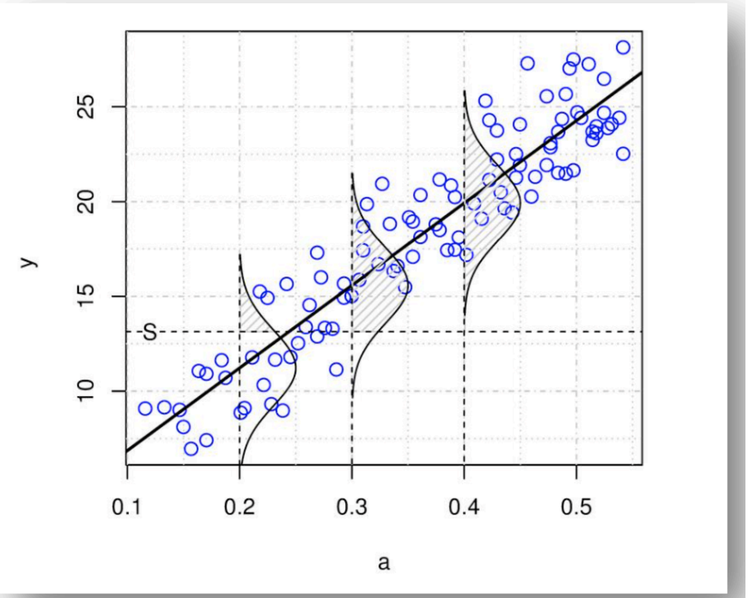
GLM, GAM

全体像

概略

線形回帰モデルは $y = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p + \epsilon$ で定義され、p個の特徴量の重み付き和に正規分布に従う誤差 ε を加算して算出するが、現実世界の多くの問題は単純な重み付き和では解決できない。
（上記式で計算された結果は正規分布に従うと仮定されているが、実際はそうならない事が多い）





問題と解決策

課題	解決策
特徴量が与えられたときの結果yが正規分布に従わない	一般化線形モデル（GLM）
特徴量と結果 y の間の真の関係性が線形ではない。	一般化加法モデル（GAM）

GLM（Generalized Linear Models：一般化線形モデル）

概略

一般化線形モデルでは目的変数が正規分布に従わなくても適用でき、さらに質的変数であってもよい。また、目的変数と説明変数との関係式は簡単な線形式である必要はなく、以下のように表される。

$$g(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

リンク関数

特徴

- 各特徴量が独立 (相互作用を考慮しない)
- **線形和と期待値を繋ぐ「リンク関数 g」**の存在
- g(y)の**誤差構造を、指数型分布族の中から自由にモデリング**可能

GLM（Generalized Linear Models：一般化線形モデル）

リンク関数の気持ち

$$g(y) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$$

リンク関数

・線形和が取りうる値は、すべての実数（マイナスの値にもなるし、ゼロになったりもする） で、

目的変数 y の分布によって、期待値 $E[y]$ の範囲を制御したい（※）時のための関数。

（※）確率値 $0 < E[y] < 1$ になって欲しいとき...

正の値 $0 < E[y]$ になって欲しいとき...

・リンク関数は、逆関数が存在するなめらかな曲線である必要があるが、勝手に決めて良い。

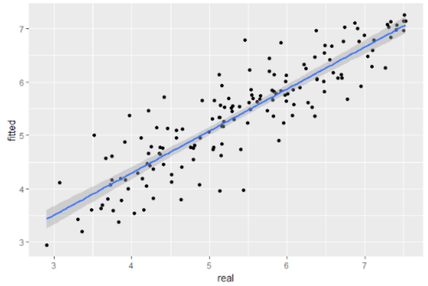
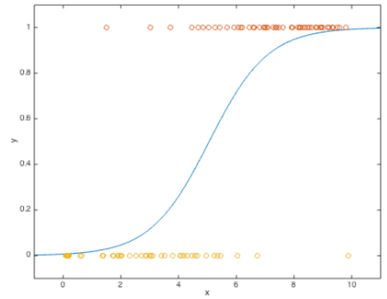
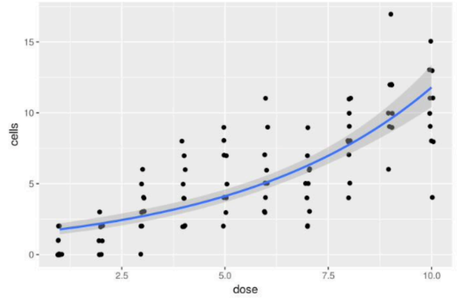
ただし、**一般的に用いられるリンク関数は大体決まっている**

GLM（Generalized Linear Models：一般化線形モデル）

指数型分布族

分布の平均、分散、その他のパラメータを持つ指数が含まれた共通の式によって記述できる分布の集合。

例えば

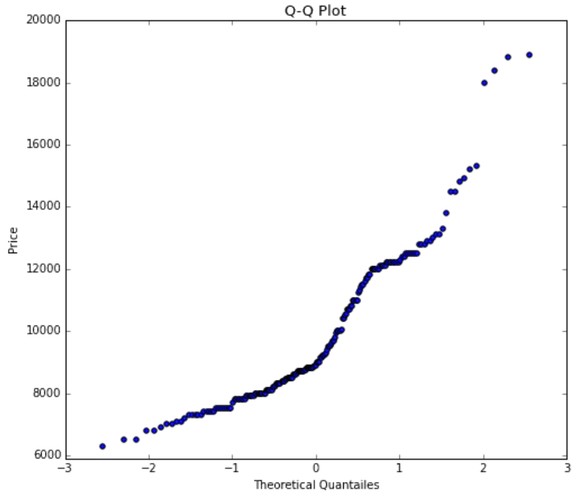
	Linear Regression	Logistic Regression	Poison Regression
			
誤差構造 (指数型分布族)	正規分布	ベルヌーイ分布	ポアソン分布
リンク関数のきもち	特になし ($E[y]$は何でも良い)	$0 \leq E[y] \leq 1$ にしたい	$0 \leq E[y]$ の範囲にしたい
リンク関数	恒等関数 (identity)	ロジット関数、プロビット関数	対数関数 log

GLM（Generalized Linear Models：一般化線形モデル）

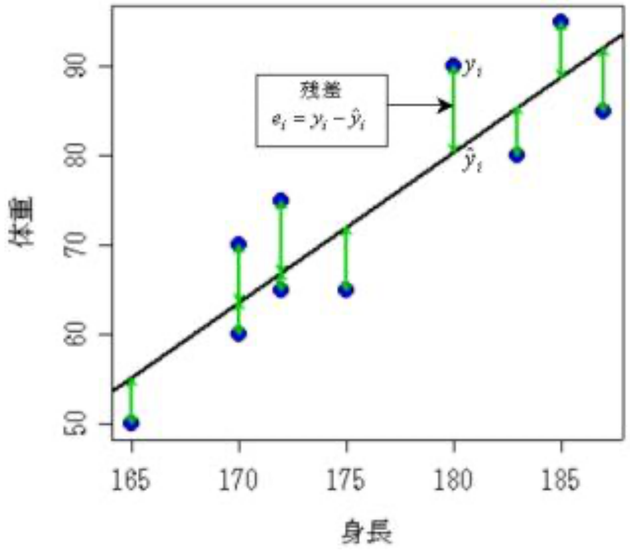
可視化その他

誤差の正当性

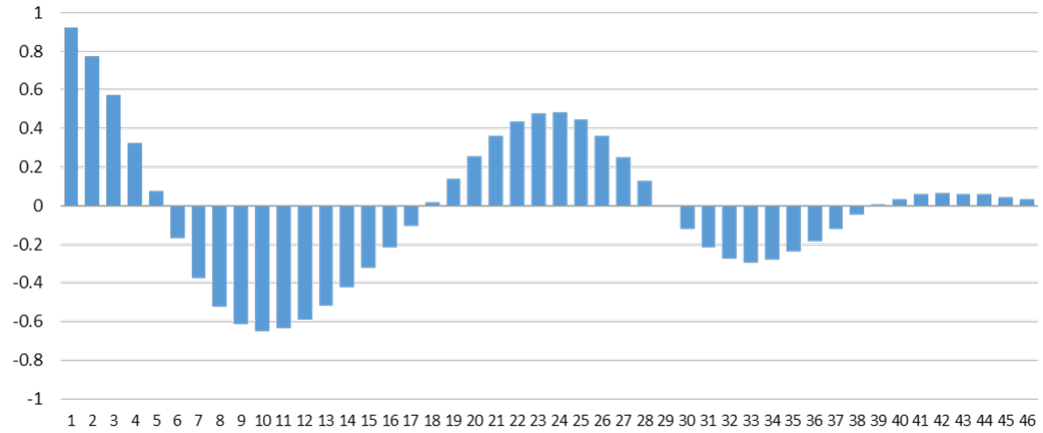
（qqプロットで分布確認）



残差の確認



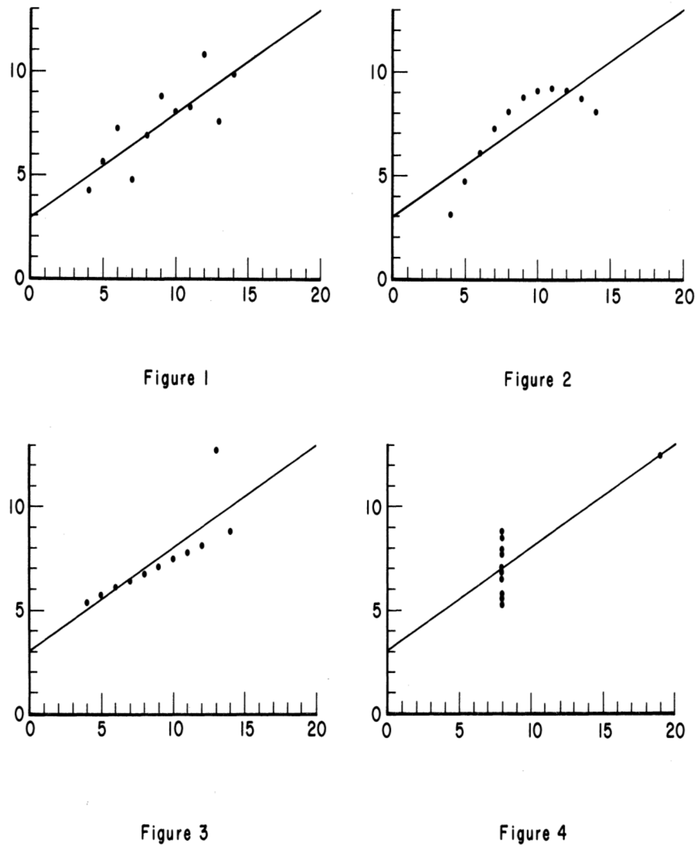
係数のプラス、マイナス表記



GAM（Generalized Additive Model：一般化加法モデル）

GLMの課題

- 説明変数と目的変数の関係は、本当に線形 なのか?
- リンク関数で変形されているものの、本質的に**GLMは 「説明変数の重み付き総和」**で関係を表現するモデル
- 右図**異なるデータに対して同じ線形モデルが学習されてしまう**可能性がある (Anscombe 1973)



http://www.sjsu.edu/faculty/gerstman/StatPrimer/anscombe1973.pdf

解決策

① 説明変数の非線形変換（logで変換、カテゴリ化等の特徴量エンジニアリング）

② **GAMを活用する**

GAM（Generalized Additive Model：一般化加法モデル）

参照元：一般化線形モデル(GLM) & 一般化加法モデル(GAM)
山口 順也 日本マイクロソフト株式会社

GAMとは

- 一般化加法モデル (Generalized Additive Model; GAM) は、1990 年 に Hastie と Tibshirani によって提案された統計モデル
- GLM の線形和という制約を緩和
- より柔軟な曲線 (3 次スプライン関数) で各変数の期待値への寄与を計算する

$$g(\mu) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$$

$$g(\mu) = w_0 + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d)$$

説明変数の重み付き和から、各説明変数の一般的な変換の総和に拡張

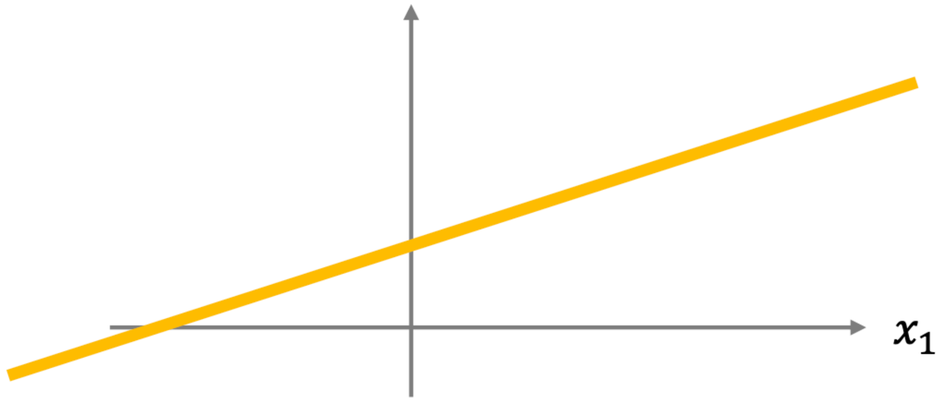
GAM（Generalized Additive Model：一般化加法モデル）

参照元：一般化線形モデル(GLM) & 一般化加法モデル(GAM)
山口 順也 日本マイクロソフト株式会社

GAMとは

$$g(\mu) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d \quad \Rightarrow \quad g(\mu) = w_0 + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d)$$

1番目の説明変数のスコア計算



GAM（Generalized Additive Model：一般化加法モデル）

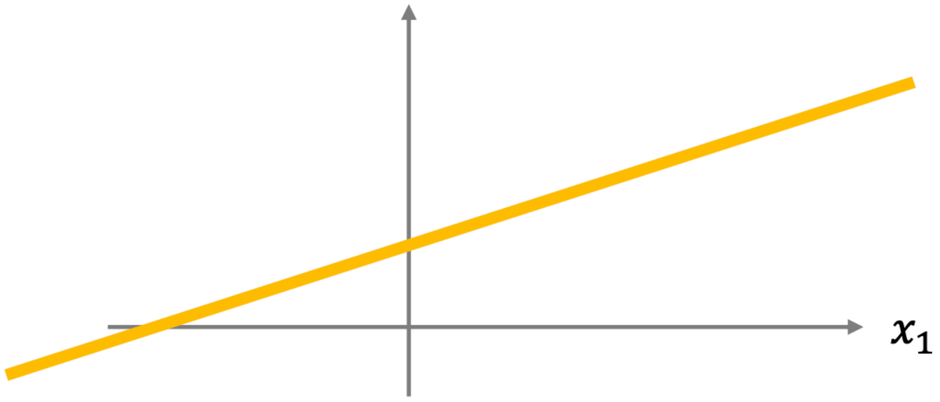
参照元：一般化線形モデル(GLM) & 一般化加法モデル(GAM)
山口 順也 日本マイクロソフト株式会社

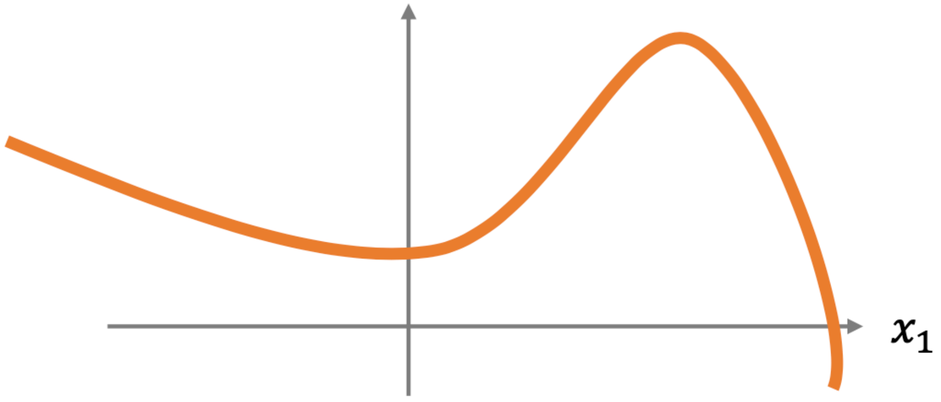
GAMとは

$$g(\mu) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_d x_d$$

$$g(\mu) = w_0 + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d)$$

1番目の説明変数のスコア計算





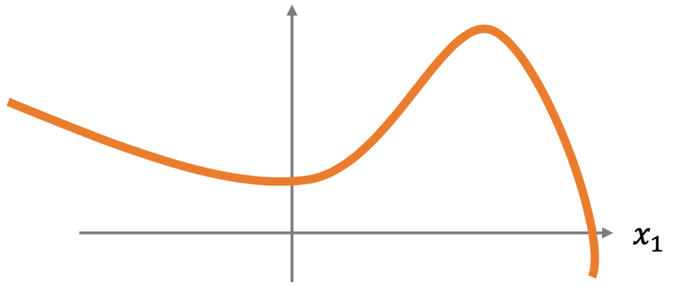
x_1 の場所によってスコアが柔軟に変動する

GAM（Generalized Additive Model：一般化加法モデル）

参照元：一般化線形モデル(GLM) & 一般化加法モデル(GAM)
山口 順也 日本マイクロソフト株式会社

疑問点

右記のような曲線をどうやって書くか？



各データ点を線で結ぶ ⇒ 線が「カクカク」するからダメ

各データ点で連続になるようにする ⇒ 異常値に影響を受けやすいモデルになる

GAM（Generalized Additive Model：一般化加法モデル）

参照元：一般化線形モデル(GLM) & 一般化加法モデル(GAM)
山口 順也 日本マイクロソフト株式会社

解決策

RidgeやLassoのように正則化項を導入してモデルが過度に複雑になることを防いではどうか。正則化と同じように、回帰関数の曲率がなるべく小さくなるように関数を求めてみる。

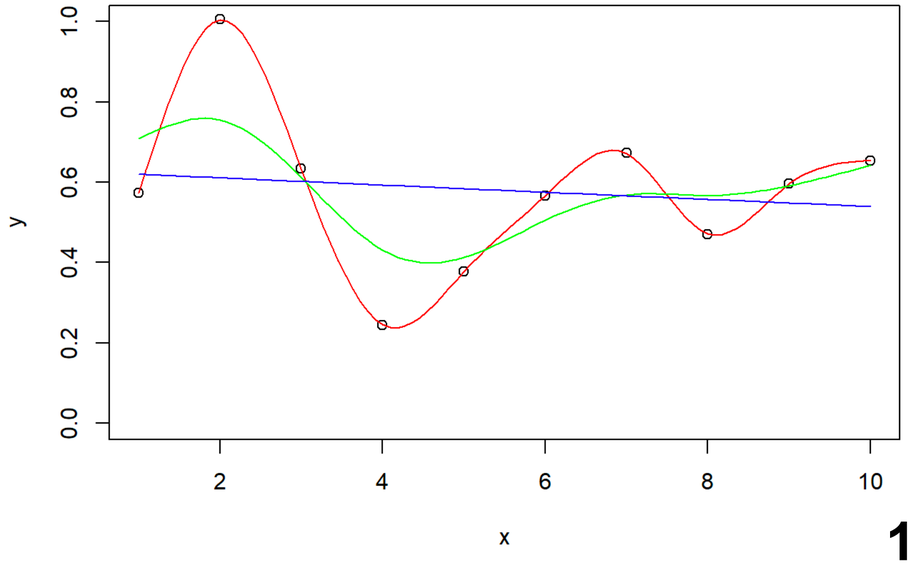
曲線推計問題

スプライン（区分的に定義された多項式）曲線を使ってフィッティングしていく　＝　平滑化スプライン曲線による当てはめ

$$\min_{f \in C^2} \sum_{i=1}^{n} (y_i - f(x_i))^2 + \lambda \int_a^b (f''(t))^2 dt, \quad \lambda > 0$$

λは、スプライン曲線を調整するパラメータ。
λの値を小さくすると、目的関数における誤差平方和の比重が大きくなり、
データに曲線をあてはめようとする傾向が強くなる。（＝曲線のカーブが多くなる）
λの値を大きくすると、あてはめが硬くなり直線に近付く。

・平滑化スプライン曲線は、Xに対するYの期待値を調べるのに役立つ。
・曲線の各部分に近い点ほど、形状に大きな影響を及ぼしており、λの値を小さくすると、点が曲線に与える影響が大きくなり、曲線が柔軟になる。



GAM（Generalized Additive Model：一般化加法モデル）

参照元：一般化線形モデル(GLM) & 一般化加法モデル(GAM)
山口 順也 日本マイクロソフト株式会社

GAM整理

- GAMの式は右記のとおり　$g(\mu) = w_0 + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d)$
- 特徴量の数、曲線が必要
- 同時に推定することは難しい
- よって、収束するまで以下のイテレーションを 繰り返す

  $f_2, \ldots, f_d$ は学習済みと仮定して f_1 を学習

  $f_1, f_3 \ldots, f_d$は学習済みと仮定して f_2 を学習

$$g(\mu) = w_0 + f_1(x_1) + f_2(x_2) + \cdots + f_d(x_d)$$

相互作用項を上手く GAM に組み込んだ GA2M もある

※ ランダムフォレストを凌駕する性能を示すことも