データ解析
第十二回「一般化加法モデル」

鈴木　大慈
理学部情報科学科
西八号館 W707 号室
s-taiji@is.titech.ac.jp

今日の講義内容

- 加法モデル
- 一般化加法モデル

構成

スプライン回帰（復習）

$$f(x;\alpha) = \sum_{j=1}^{q} \alpha_j B_j(x).$$

なる関数形で曲線を表現.
3 次スプライン回帰では,

$$\min_{\alpha\in\mathbb{R}^{n+2}} \sum_{i=1}^{n} (Y_i - f(X_i;\alpha))^2 + \lambda \int \left(\frac{\mathrm{d}^2 f(x;\alpha)}{\mathrm{d}x^2} \right)^2 \mathrm{d}x.$$

として f を推定. 3 次スプラインでは B_j は局所的 3 次多項式.

単変量から多変量へ

説明変数が 2 つ以上あったらどうする？

単変量:

$$f(x) = \sum_{j=1}^{q} \alpha_j B_j(x).$$

加法モデル: $x = [x_1, \ldots, x_d]^\top$ に対して,

$$f(x) = \sum_{k=1}^{d} s_k(x_k),$$

$$\text{ただし } s_k(x_k) = \sum_{j=1}^{q_k} \alpha_{k,j} B_{k,j}(x_k).$$

各変数に非線形な関数をかぶせて 足し算.

重回帰との類似

重回帰モデル:

$$f(x) = \sum_{k=1}^{d} \beta_k x_k.$$

加法モデル:

$$f(x) = \sum_{k=1}^{d} s_k(x_k),$$

$$\text{ただし } s_k(x_k) = \sum_{j=1}^{q_k} \alpha_{k,j} B_{k,j}(x_k).$$

$s_k(x_k) = \beta_k x_k$ とすれば線形回帰に帰着される．その意味で一般化になっている．

加法モデルの推定

$$f(x) = \sum_{k=1}^{d} s_k(x_k), \quad (s_k(x_k) = \sum_{j=1}^{q_k} \alpha_{k,j} B_{k,j}(x_k)).$$

各 s_j が三次スプラインの場合:

$$\min_{\alpha_{k,j}} \quad \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{d} s_k(x_{i,k}) \right)^2 + \lambda \sum_{k=1}^{d} \int (s_k''(t))^2 \mathrm{d}t.$$

これも単変量の場合と同様に二次関数の最小化問題として定式化できる.

交互作用も入れたモデル

加法モデルは各変数の関数の和で表すため，変数間の絡みは表現できない．
例：$f(x, y) = xy$.

和で書けない関数 $s_{ab}(x_a, x_b)$ も入れて推定 (2 次の交互作用):

$$f(x) = \sum_{a>b} s_{ab}(x_a, x_b) + \sum_k s_k(x_k).$$

三次以上の交互作用は $s_{abc}(x_a, x_b, x_c)$ のように次数を増やしてゆけば良い．

交互作用も入れたモデルの推定

$\{(x^{(i)}, y_i)\}_{i=1}^n$: サンプル

関数形：

$$f(x) = \sum_{a>b} s_{ab}(x_a, x_b) + \sum_k s_k(x_k).$$

データへの当てはめ：

$$\begin{aligned}
\min_f \quad & \sum_{i=1}^n (y_i - f(x^{(i)}))^2 \\
& + \lambda \sum_{a>b} \int\int \left\{ \left(\frac{\partial^2 s_{ab}}{\partial x_a^2} \right)^2 + 2 \left(\frac{\partial^2 s_{ab}}{\partial x_a \partial x_b} \right)^2 + \left(\frac{\partial^2 s_{ab}}{\partial x_b^2} \right)^2 \right\} \mathrm{d}x_a \mathrm{d}x_b \\
& + \lambda \sum_k \int \left(\frac{\partial^2 s_k}{\partial x_k^2} \right)^2 \mathrm{d}x_k.
\end{aligned}$$

スプラインの多次元化

では，どのようにして $s_{ab}(x_a, x_b)$ を構成するか？

次の２つを紹介：

1. テンソル積スプライン (tensor product spline)
2. 薄板平滑化スプライン (thin plate spline)

基本的には基底 $B_{ab,j}$ の和で表す:

$$s_{ab}(x_a, x_b) = \sum_{j=1}^{q_{ab}} \alpha_{ab,j} B_{ab,j}(x_a, x_b).$$

各スプライン手法の違いはこの基底の構成の仕方の違い.

テンソル積スプライン

一次元のスプライン基底 $b_j(x)$ を用意する (例えば 3 次スプライン).

$$B_{ab,j}(x_a, x_b) = b_{j_a}(x_a) b_{j_b}(x_b)$$

のようにする.

$$s_{ab}(x_a, x_b) = \sum_{j_a=1}^{q_a} \sum_{j_b=1}^{q_b} \alpha_{j_a,j_b} b_{j_a}(x_a) b_{j_b}(x_b).$$

三変数以降も同様.

$$s_{abc}(x_a, x_b, x_c) = \sum_{j_a=1}^{q_a} \sum_{j_b=1}^{q_b} \sum_{j_c=1}^{q_c} \alpha_{j_a,j_b,j_c} \underbrace{b_{j_a}(x_a) b_{j_b}(x_b) b_{j_c}(x_c)}_{\text{テンソル積}}.$$

薄板平滑化スプライン

l 変数，滑らかさのパラメータ $m\ (> l/2)$ に対する薄板平滑化スプライン基底は次で与えられる:

$$\eta_{m,l}(r) := \begin{cases} \dfrac{(-1)^{m+1+l/2}}{2^{2m-1}\pi^{l/2}(m-1)!(m-l/2)!} r^{2m-l}\log(r) & (l \text{ が偶数}), \\ \dfrac{\Gamma(l/2-m)}{2^{2m}\pi^{l/2}(m-1)!} r^{2m-l} & (l \text{ が奇数}), \end{cases}$$

に対し

$$B_{a_1 a_2 \ldots a_m, i}(x_{a_1}, \ldots, x_{a_m}) = \eta_{m,l}(\|[x_{a_1}, \ldots, x_{a_m}] - [x_{a_1}^{(i)}, \ldots, x_{a_m}^{(i)}]\|).$$

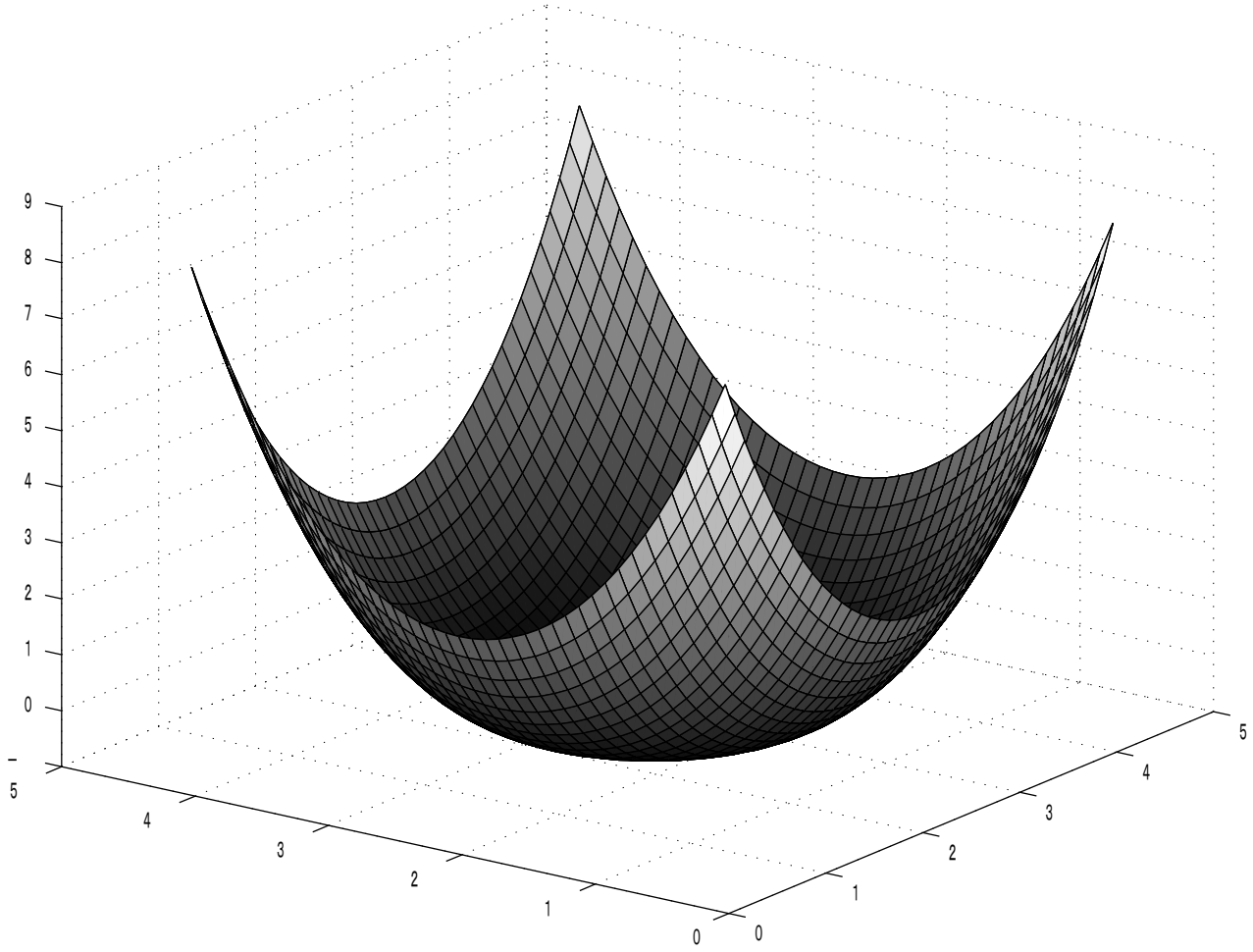
なお，$[x_{a_1}^{(i)}, \ldots, x_{a_m}^{(i)}]$ は節点 (たとえばサンプル点とする).

$m = l = 2$ (2 変数) の時は，

$$B_{a_1 a_2, i}(x_{a_1}, x_{a_2}) = \frac{r_i^2}{8\pi}\log(r_i),$$

ただし，$r_i = \|[x_{a_1}, x_{a_2}] - [x_{a_1}^{(i)}, x_{a_2}^{(i)}]\|$.

薄板平滑化スプライン基底の図



薄板平滑化スプラインの詳細

サンプル: $\{(x^{(i)}, y_i)\}_{i=1}^n$.

実際は,

$$s(x) = \sum_{i=1}^{n} \alpha_i \eta_{m,l}(\|x - x^{(i)}\|) + \sum_{j=1}^{M} \beta_j \phi_j(x),$$

とする. ここで, $M = \binom{m+d-1}{d}$ で ϕ_j は互いに一次独立な m 次以下の多項式, $\alpha = [\alpha_1, \ldots, \alpha_n]^\top$ は $T = (\eta_{m,l}(\|x^{(i)} - x^{(j)}\|))_{i,j=1}^n$ に対し $\alpha^\top T = 0$ を満たすようにする.

これは

$$\min_s \sum_{i=1}^{n} (y_i - s(x^{(i)}))^2 + \lambda \int \cdots \int \sum_{\nu_1 + \cdots + \nu_d = m} \frac{m!}{\nu_1! \ldots \nu_d!} \left(\frac{\partial^m s}{\partial x_1^{\nu_1} \ldots \partial x_d^{\nu_d}} \right)^2 \mathrm{d}x_1 \ldots \mathrm{d}x_d,$$

の最小化元である.

2 変数スプラインの推定

サンプル: $\{(x^{(i)}, y_i)\}_{i=1}^n$.

三次スプラインのテンソル積および $m = 2$ の薄板平滑化スプラインの場合.

$$\min \sum_{i=1}^{n} (y_i - s(x_1^{(i)}, x_2^{(i)}))^2 + \lambda \int\int \left\{ \left(\frac{\partial^2 s}{\partial x_1^2} \right)^2 + \left(\frac{\partial^2 s}{\partial x_1 \partial x_2} \right)^2 + \left(\frac{\partial^2 s}{\partial x_2^2} \right)^2 \right\} \mathrm{d}x_1 \mathrm{d}x_2.$$

多変数スプラインの推定

I 変数，$m+1$ 次スプラインのテンソル積および滑らかさパラメータ m の薄板平滑化スプラインの場合.
各 $s(x_1, \ldots, x_I)$ に対して，正則化項を次のように定める:

$$J(s) = \int \cdots \int \sum_{\nu_1 + \cdots + \nu_I = m} \frac{m!}{\nu_1! \ldots \nu_I!} \left(\frac{\partial^m s}{\partial x_1^{\nu_1} \ldots \partial x_I^{\nu_I}} \right)^2 \mathrm{d}x_1 \ldots \mathrm{d}x_I.$$

これらを足しあわせて回帰:

$$f = \sum_{j_1} s_{j_1}(x_{j_1}) + \sum_{j_2, j_3} s_{j_2 j_3}(x_{j_2}, x_{j_3}) + \sum_{j_4, j_5, j_6} s_{j_4 j_5 j_6}(x_{j_4}, x_{j_5}, x_{j_6}) \ldots$$

に対し，正則化項は

$$\sum_{j_1} J(s_{j_1}) + \sum_{j_2, j_3} J(s_{j_2 j_3}) + \sum_{j_4, j_5, j_6} J(s_{j_4 j_5 j_6}) + \ldots$$

のようにする．いずれにせよ，二次関数最小化で解ける．

構成

一般化加法モデル

これまでのモデル:

$$y_i = f(x_i) + \epsilon_i,$$

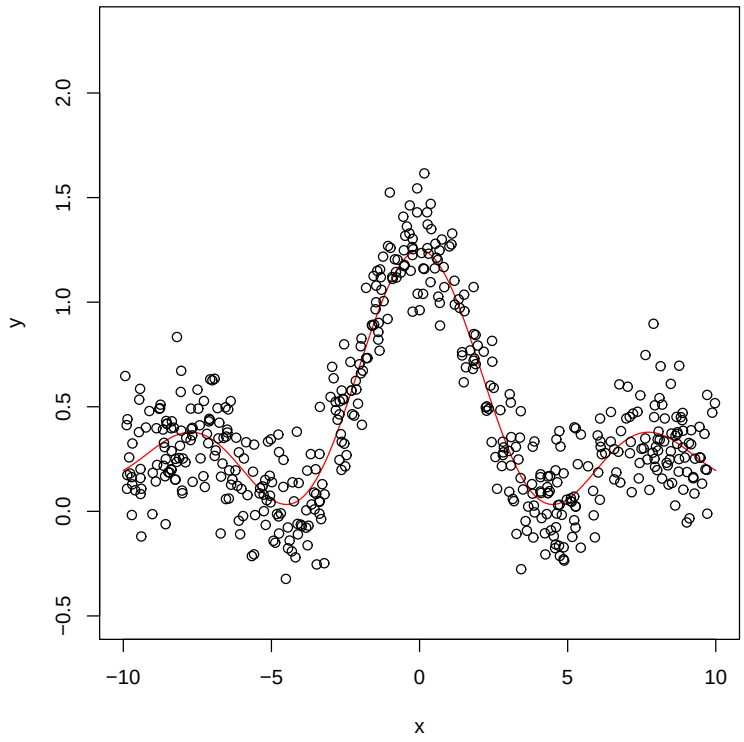
ϵ_i は平均 0 の正規分布.

$\Longrightarrow$ 一般化:

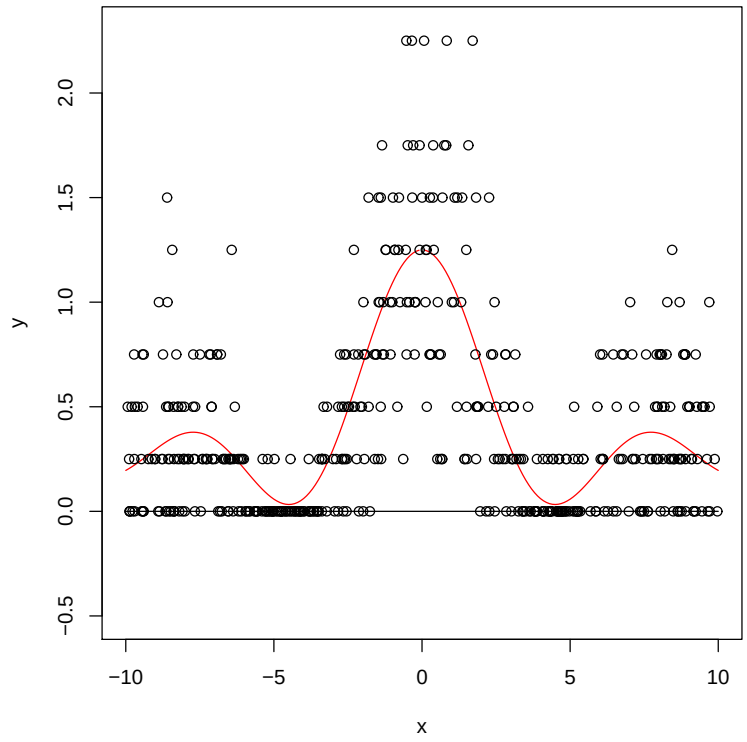
$$y_i \sim P_{\theta=g(f(x_i))}(Y).$$

1. $P_\theta(Y)$ あるパラメトリックモデル.
2. g^{-1} はリンク関数 (固定).

f をノンパラメトリックに推定したい.



(a) 正規分布



(b) ポアソン

リンク関数の例

- ポアソン分布と対数リンク関数:
  $\mathrm{Po}_\theta(Y) = \frac{\theta^Y e^{-\theta}}{Y!}$ $(\theta > 0, Y = 0, 1, 2, \dots)$.

$$y_i \sim \mathrm{Po}(\theta = \exp(f(x)))$$

  $g(f(x)) = \exp(f(x))$, $g^{-1}(\theta) = \log(\theta) = f(x)$: 対数リンク関数.
  $f(x)$ が負の値をとっても大丈夫！
- 二項分布とロジットリンク関数:
  $\mathrm{Bin}_{\theta|N}(Y) = \binom{N}{Y}\theta^Y(1-\theta)^{N-Y}$ $(\theta \in (0,1),\ Y = 0, 1, \dots, N)$.

$$y_i \sim \mathrm{Bin}\left(\theta = \frac{1}{1+\exp(-f(x))}|N\right)$$

  $g(f(x)) = \frac{1}{1+\exp(-f(x))}$, $f(x) = g^{-1}(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$: ロジット関数.

一般化加法モデルの推定

負の対数尤度:

$$\ell(y, u) = -\log(p_{\theta=g(u)}(y)).$$

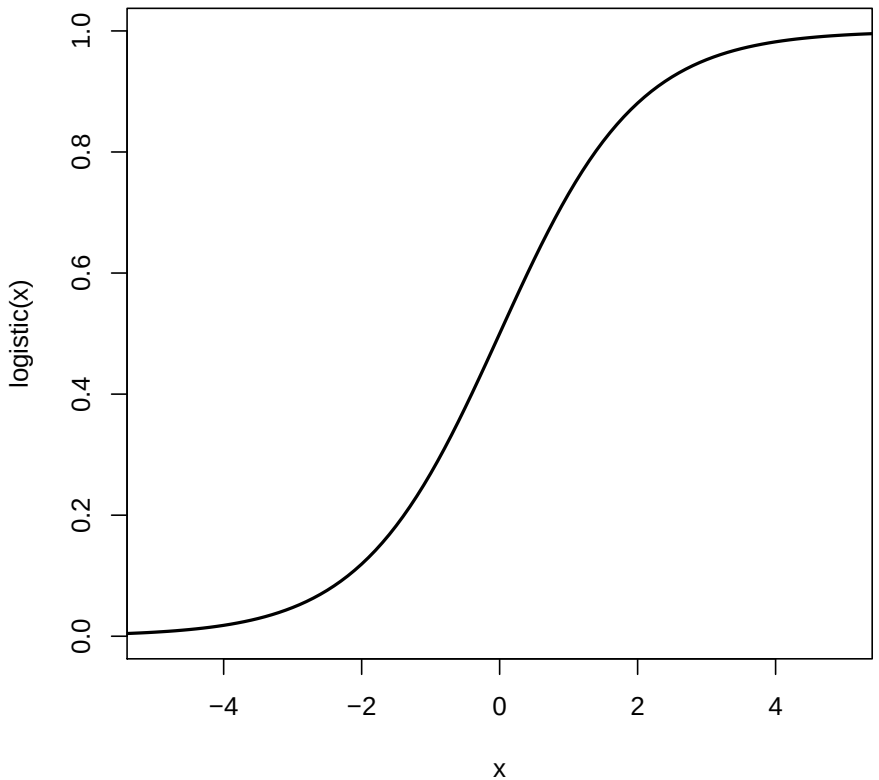
正則化項付き尤度最大化:

$$\min_{s_{j_1}, s_{j_2,j_3}, \ldots} \sum_{i=1}^{n} \ell(y_i, f(x_i)) + \lambda \left\{ \sum_{j_1} J(s_{j_1}) + \sum_{j_2,j_3} J(s_{j_2 j_3}) + \sum_{j_4,j_5,j_6} J(s_{j_4 j_5 j_6}) + \ldots \right\},$$

$$\text{ただし,} \quad f = \sum_{j_1} s_{j_1}(x_{j_1}) + \sum_{j_2,j_3} s_{j_2 j_3}(x_{j_2}, x_{j_3}) + \ldots.$$

$\Longrightarrow$ (大体の場合) 凸最適化で解ける.

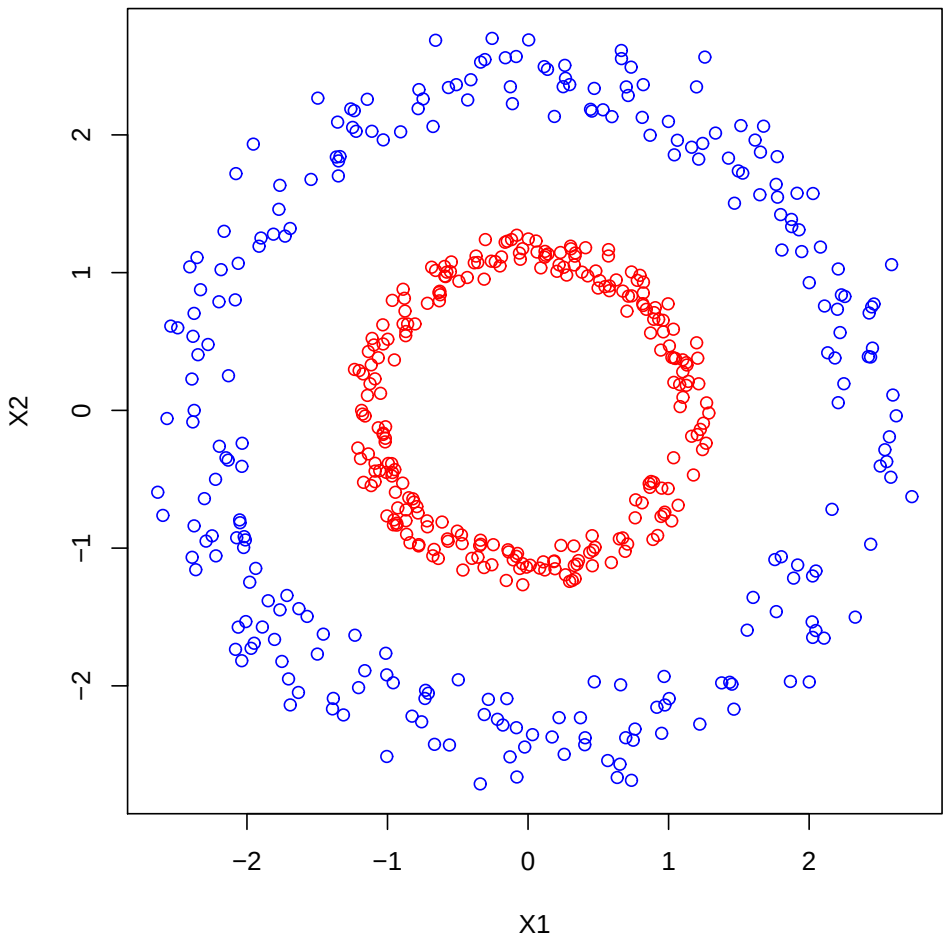
判別分析



ロジット関数 $g(u) = \frac{1}{1+\exp(-u)}$

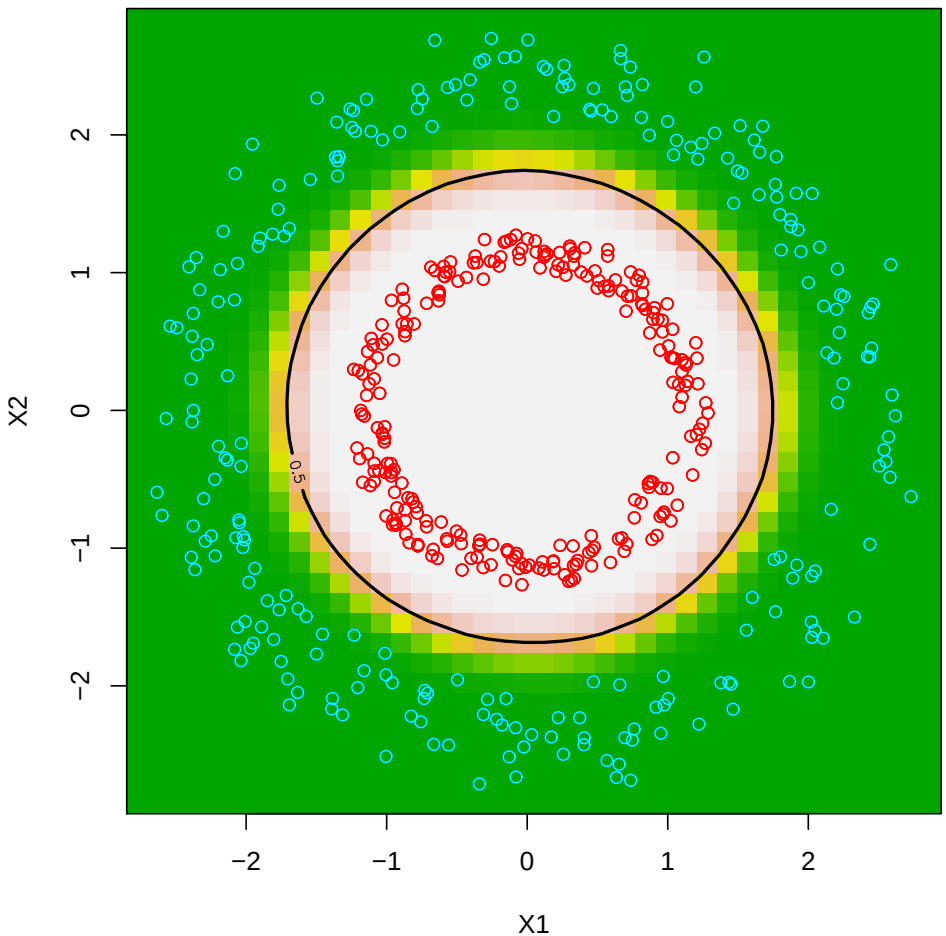
非線形判別分析

```
fit <- gam(Y~s(X1,X2),data=artdata,family=binomial(link=logit))
```



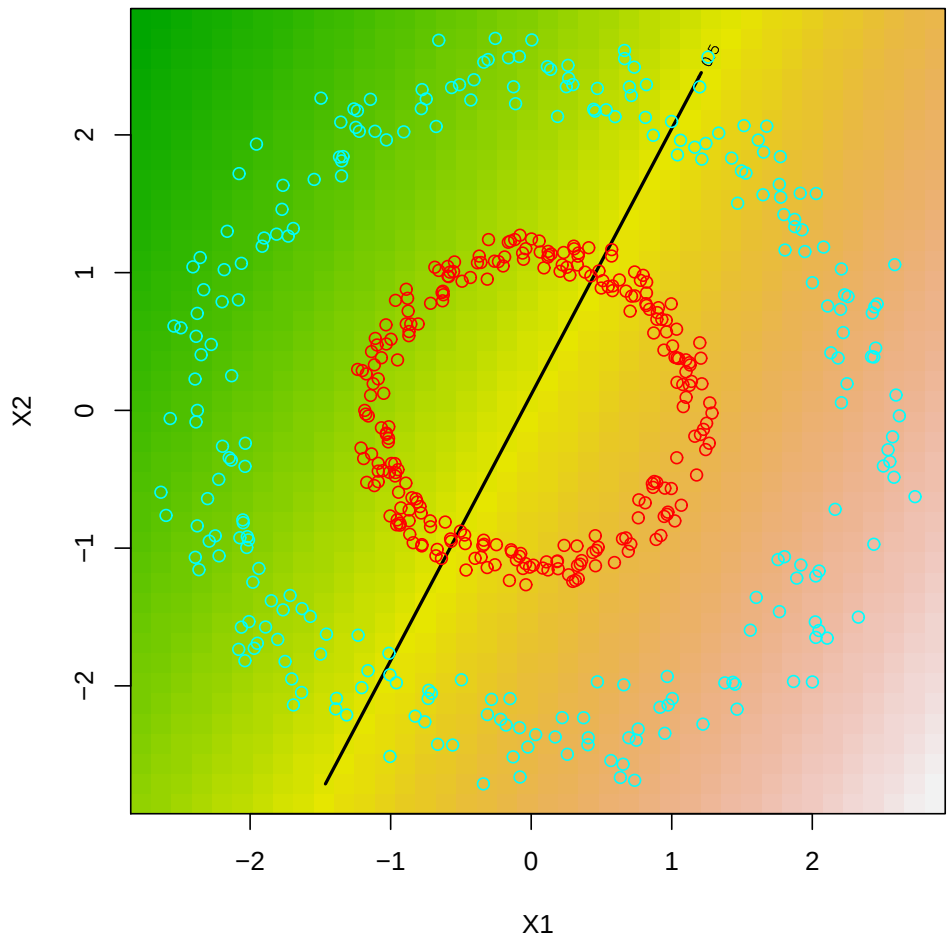
非線形判別分析

```
fit <- gam(Y~s(X1,X2),data=artdata,family=binomial(link=logit))
```



非線形判別分析

```
fit <- gam(Y~s(X1,X2),data=artdata,family=binomial(link=logit))
```



gam の使い方

`Y~s(X1)+s(X2)` のようにするだけで，glm と使い方は同じ．
link 関数，分布族ともに glm と同じものが使える．

構成

カリフォルニア住宅価格データ

```
chouse <- read.csv("cal_house.csv",header=TRUE)
```

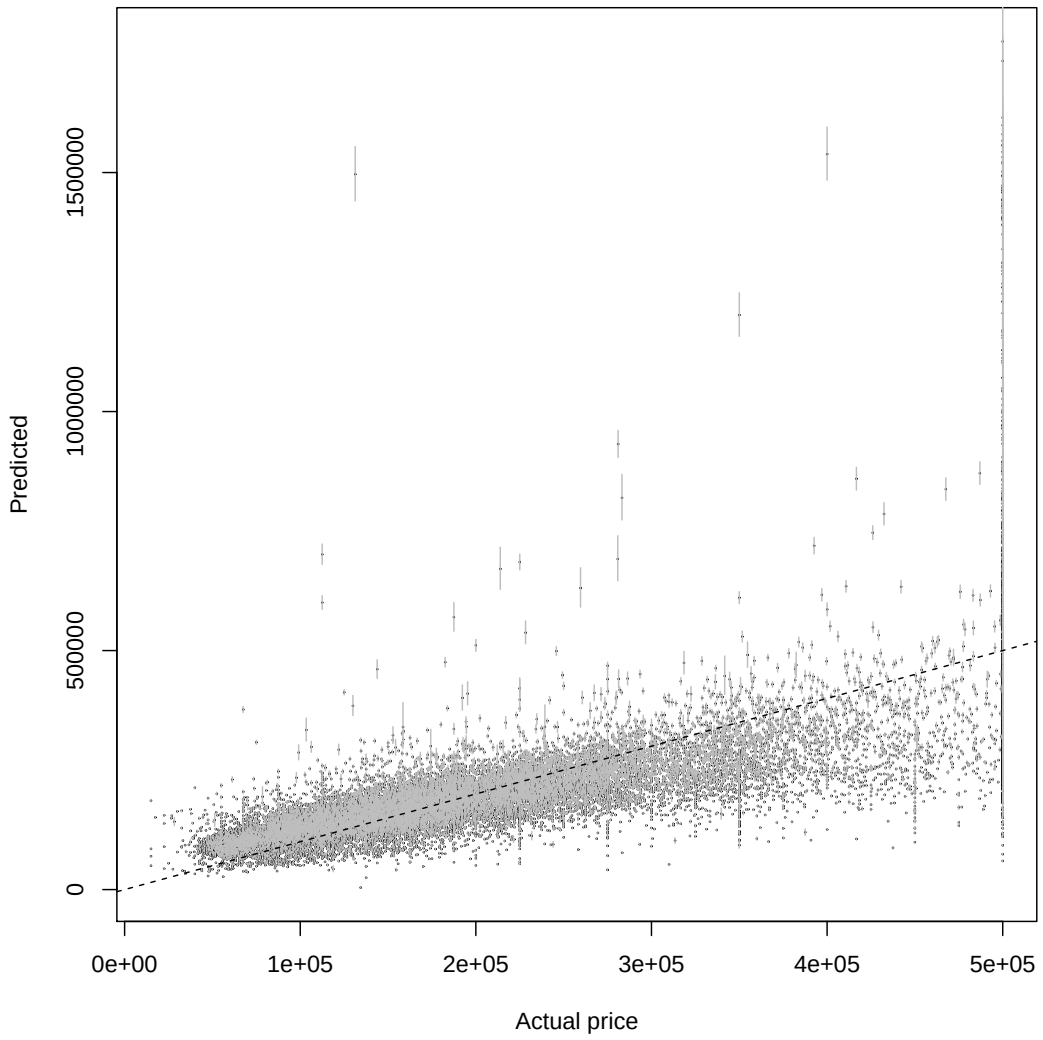
20640 サンプル， 9 変数, カリフォルニア州の各地区の住宅価格データ

- "MedHouseValue": その地区の住宅価格の中央値 (これを説明したい)
- "MedIncome": 収入の中央値
- "MedHouseAge": 築年数の中央値
- "TotalRooms": 部屋数の中央値
- "TotalBedrooms": ベッドルーム数の中央値
- "Population": 人口
- "Households": 家持人の数
- "Latitude": 緯度
- "Longitude": 経度

線形回帰

```
> lmfit <- gam(log(MedHouseValue)~MedIncome+...+Longitude,
>              data=chouse)
Parametric coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.180e+01  3.059e-01 -38.570  < 2e-16 ***
MedIncome      1.782e-01  1.639e-03 108.753  < 2e-16 ***
MedHouseAge    3.261e-03  2.111e-04  15.446  < 2e-16 ***
TotalRooms    -3.186e-05  3.855e-06  -8.265  < 2e-16 ***
TotalBedrooms  4.798e-04  3.375e-05  14.215  < 2e-16 ***
Population    -1.725e-04  5.277e-06 -32.687  < 2e-16 ***
Households     2.493e-04  3.675e-05   6.783 1.21e-11 ***
Latitude      -2.801e-01  3.293e-03 -85.078  < 2e-16 ***
Longitude     -2.762e-01  3.487e-03 -79.212  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

R-sq.(adj) =  0.643   Deviance explained = 64.3%
GCV score = 0.11568  Scale est. = 0.11563    n = 20640
```

線形回帰の結果



あまり当てはまりが良くない.

加法モデルの推定

```
addfit <- gam(log(MedHouseValue) ~ s(MedIncome)+ s(MedHouseAge)
  + s(TotalRooms) + s(TotalBedrooms) + s(Population)
  + s(Households) + s(Latitude) + s(Longitude), data=chouse)
```

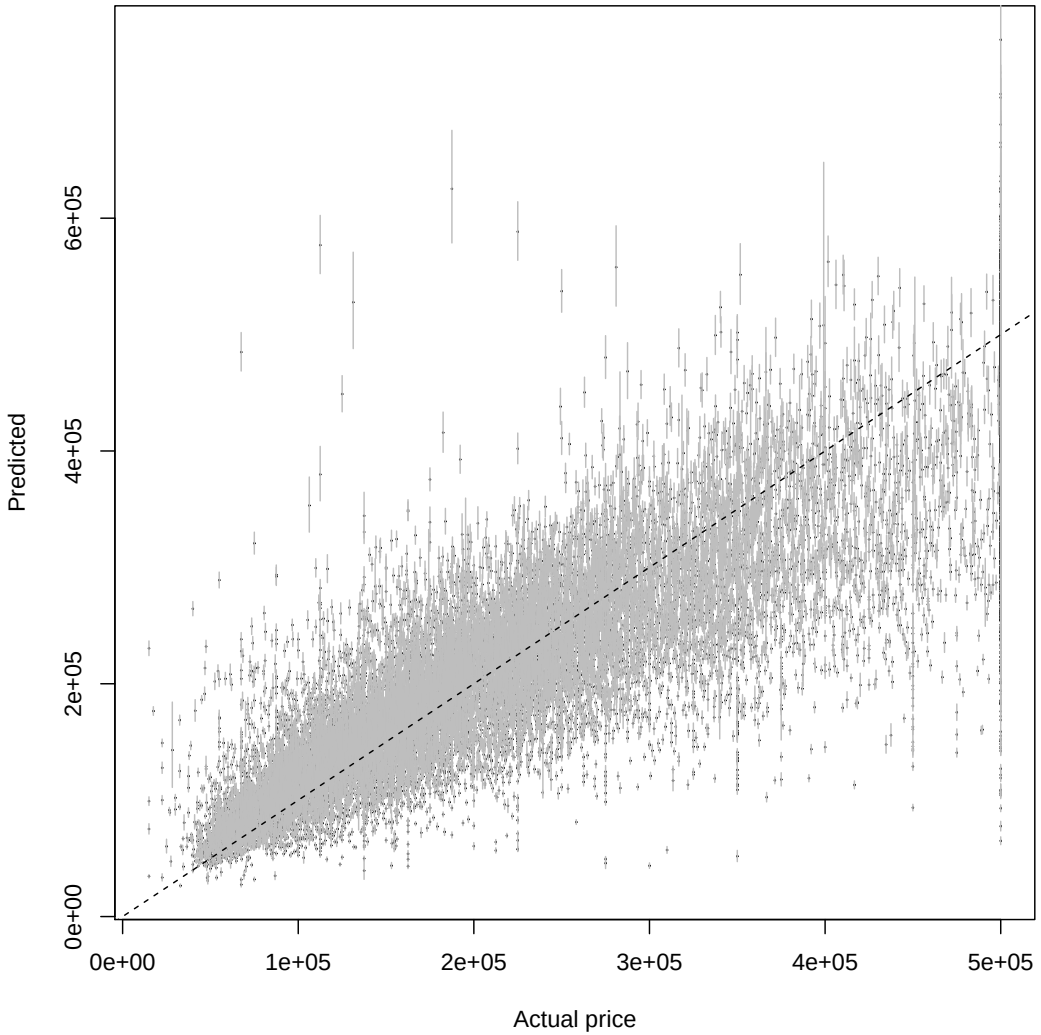
s(·) でスプライン平滑化を当てはめ.

加法モデルの要約

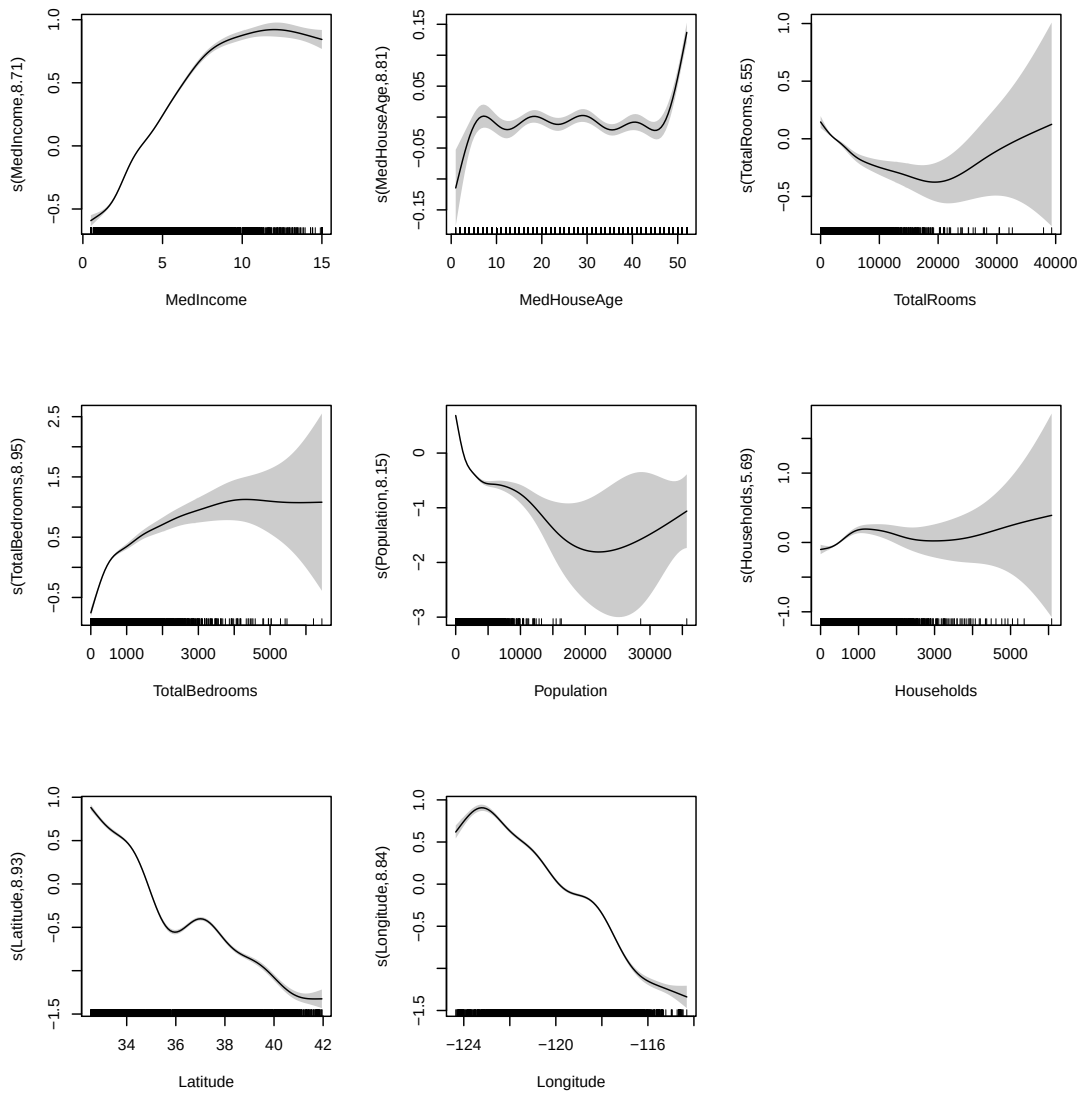
```
> summary(addfit)
Approximate significance of smooth terms:
                    edf Ref.df       F  p-value
s(MedIncome)      8.708  8.975 950.298  < 2e-16 ***
s(MedHouseAge)    8.807  8.987  35.131  < 2e-16 ***
s(TotalRooms)     6.545  7.833  11.040 3.09e-15 ***
s(TotalBedrooms) 8.955  8.987  61.631  < 2e-16 ***
s(Population)     8.148  8.722 286.264  < 2e-16 ***
s(Households)     5.692  6.898   9.329 2.29e-11 ***
s(Latitude)       8.928  8.998 819.245  < 2e-16 ***
s(Longitude)      8.844  8.992 732.237  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

R-sq.(adj) =  0.747   Deviance explained = 74.8%
GCV score = 0.08206  Scale est. = 0.081799  n = 20640
```

GCV が改善 (0.11568 から 0.08206 へ).

加法モデル回帰の結果


Predicted
0e+00
2e+05
4e+05
6e+05
0e+00
1e+05
2e+05
3e+05
4e+05
5e+05
Actual price

```
plot(addfit,scale=0,se=2,shade=TRUE,pages=1)
```



加法モデル: 交互作用有り

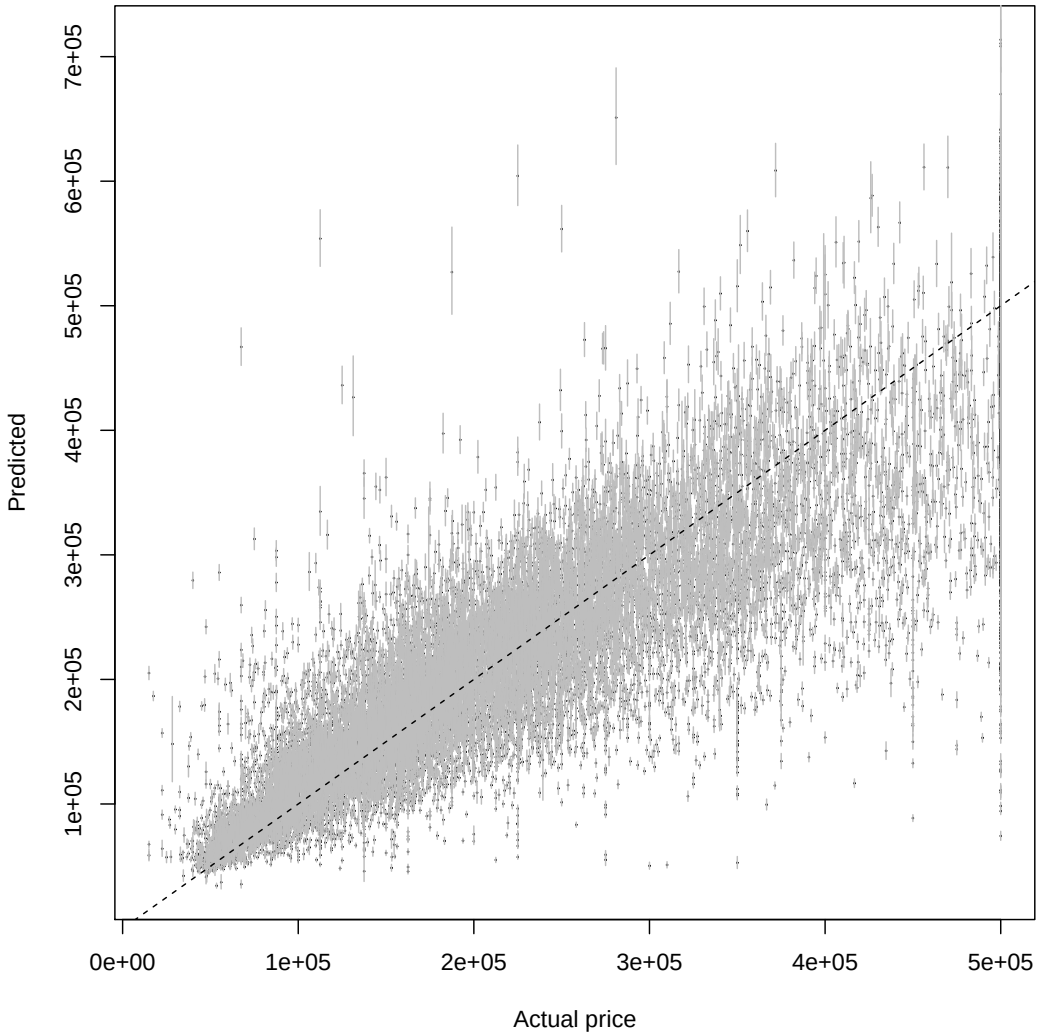
```
addfit2 <- gam(log(MedHouseValue) ~ s(MedIncome)
 + s(MedHouseAge) + s(TotalRooms) +s(TotalBedrooms)
 + s(Population) + s(Households)
 + s(Longitude,Latitude),
 data=chouse)
```

s(Longitude,Latitude) で，緯度・経度の交互作用を取り込む.

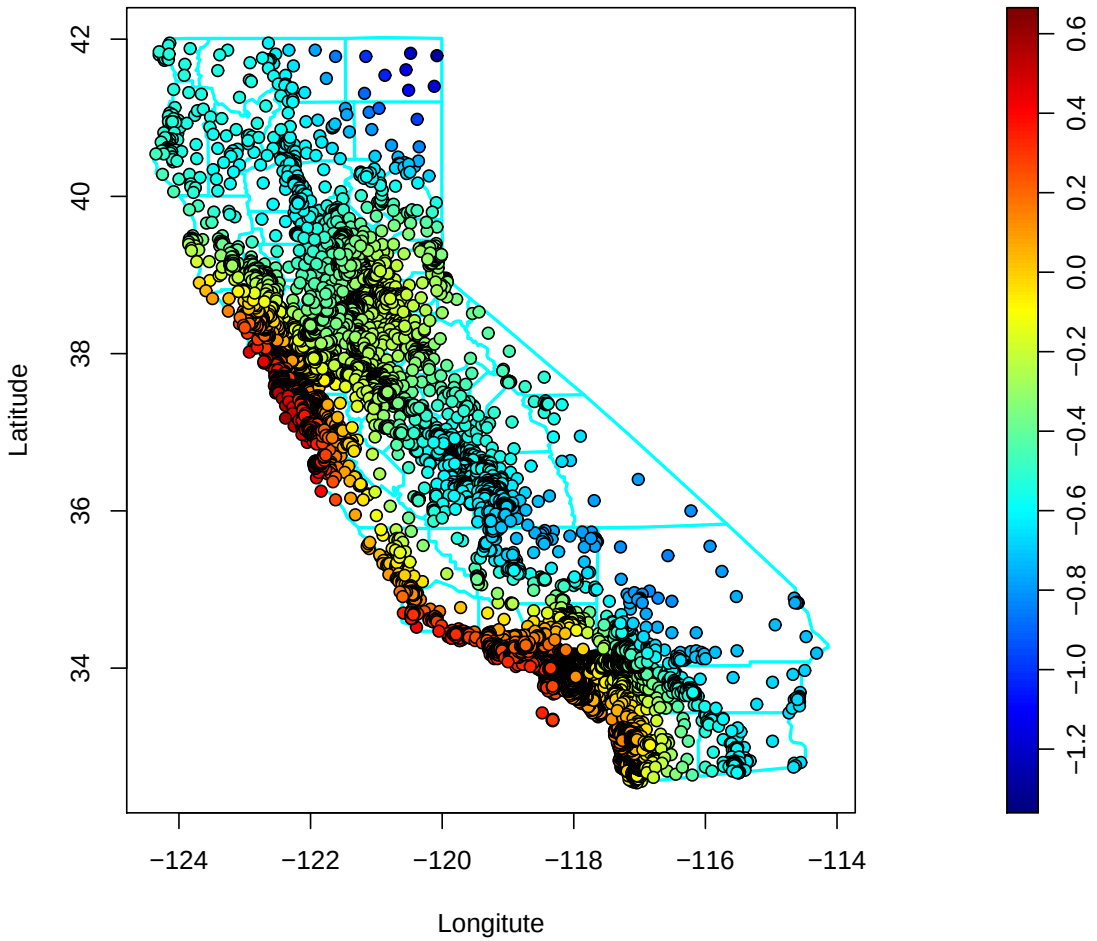
```
> summary(addfit2)

Approximate significance of smooth terms:
                           edf Ref.df       F p-value
s(MedIncome)             8.702  8.974 887.632 < 2e-16 ***
s(MedHouseAge)           8.835  8.990  24.006 < 2e-16 ***
s(TotalRooms)            5.045  6.396   3.606 0.00113 **
s(TotalBedrooms)         5.052  6.116  24.541 < 2e-16 ***
s(Population)            9.000  9.000 288.392 < 2e-16 ***
s(Households)            5.278  6.296  28.730 < 2e-16 ***
s(Longitude,Latitude) 28.855 28.998 520.236 < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’
1

R-sq.(adj) =  0.778   Deviance explained = 77.8%
GCV score = 0.072298  Scale est. = 0.072047  n = 20640
```

GCV が改善 (0.11568 → 0.08206 → 0.072298).

交互作用有り加法モデル: 結果

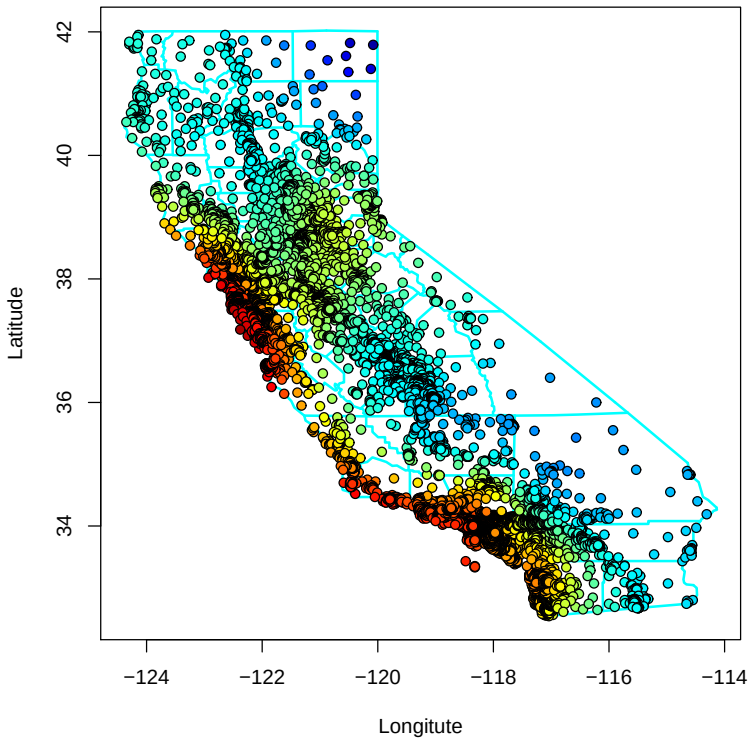


緯度経度のみから価格を予測

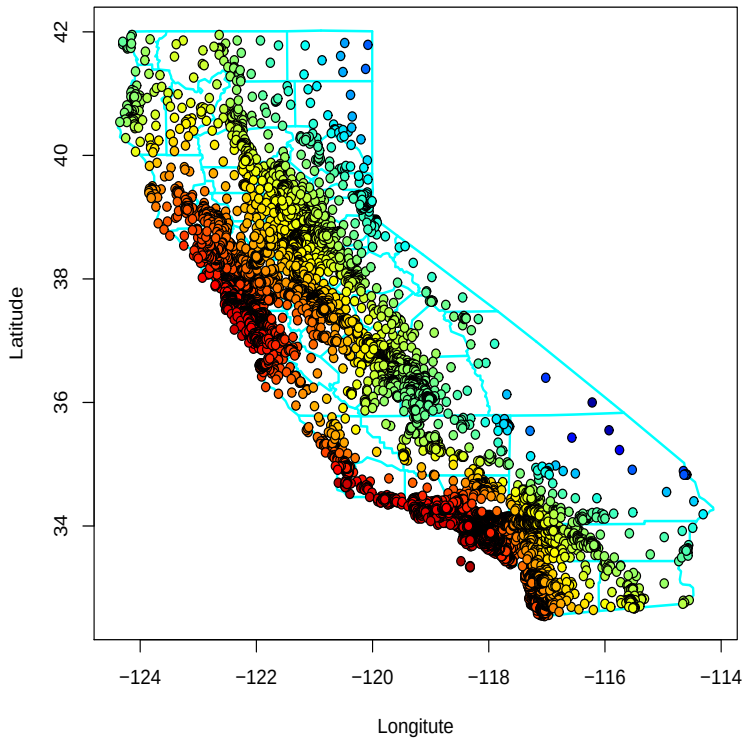


交互作用なしとありとの違い

緯度経度のみから価格を予測



(c) 交互作用あり (MSE= 0.180)



(d) 交互作用なし (MSE= 0.201)

まとめ

ノンパラメトリック回帰を多変量・非正規に拡張.

1. 加法モデル: 各成分の非線形関数の和
2. 交互作用モデル: ２つ以上の変数を用いて非線形関数を構成
3. 一般化加法モデル: 判別分析やポアソン回帰なども可能

これらのモデルを用いることでデータ解析の幅もかなり広がる.

講義情報ページ

http://www.is.titech.ac.jp/~s-taiji/lecture/dataanalysis/dataanalysis.html