

統計解析特論 講義メモ 2 回帰分析

- データ： $(\boldsymbol{x}_1, y_1), \dots, (\boldsymbol{x}_n, y_n)$, $\boldsymbol{x}_i \in \mathcal{X}$, $y_n \in \mathbb{R}$.
- y を \boldsymbol{x} の関数で説明する
- 用語の説明
 - \boldsymbol{x} ：入力（独立変数・説明変数）
 - y ：出力（従属変数・目的変数）

線形モデル

x と y の関係：基底関数 $\phi_1(\mathbf{x}), \dots, \phi_d(\mathbf{x})$ の線形和でを説明.

$$\text{線形モデル： } y_i = \sum_{k=1}^d \phi_k(\mathbf{x}_i) \theta_k + \varepsilon_i = \boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\theta} + \varepsilon_i.$$

$$\boldsymbol{\phi}(\mathbf{x}) = (\phi_1(\mathbf{x}), \dots, \phi_d(\mathbf{x}))^T.$$

データからパラメータ $\boldsymbol{\theta} = (\theta_1, \dots, \theta_d)^T$ を推定.

線形モデル：「関数部分 $\boldsymbol{\phi}(\mathbf{x})^T \boldsymbol{\theta}$ がパラメータ $\boldsymbol{\theta}$ について線形」という意味.

- **1変数 (or 単回帰) モデル：**

$$y_i = \theta_0 + \theta_1 x_i + \varepsilon_i, \quad x_i \in \mathbb{R}, \quad \boldsymbol{\theta} = (\theta_0, \theta_1)^T.$$

- **多項式回帰モデル：**

$$y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \dots + \theta_d x_i^d + \varepsilon_i, \quad x_i \in \mathbb{R}, \quad \boldsymbol{\theta} = (\theta_0, \theta_1, \dots, \theta_d)^T.$$

最小2乗法

パラメータ θ の推定法：データとモデルの間の2乗誤差を最小化.

$$\underbrace{\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}}_Y = \underbrace{\begin{pmatrix} \phi_1(\mathbf{x}_1) & \cdots & \phi_d(\mathbf{x}_1) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_n) & \cdots & \phi_d(\mathbf{x}_n) \end{pmatrix}}_{\Phi} \underbrace{\begin{pmatrix} \theta_1 \\ \vdots \\ \theta_d \end{pmatrix}}_{\theta} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$\text{関数 } \phi(\mathbf{x})^T \theta \text{ の2乗誤差} = \sum_{i=1}^n (y_i - \theta^T \phi(\mathbf{x}_i))^2 = \|Y - \Phi\theta\|^2 \longrightarrow \min_{\theta}$$

$$\text{極値条件： } \frac{\partial}{\partial \theta} \|Y - \Phi\theta\|^2 = -2\Phi^T(Y - \Phi\hat{\theta}) = \mathbf{0}$$

$$\implies \text{最小2乗推定量： } \hat{\theta} = (\Phi^T \Phi)^{-1} \Phi^T Y$$

note: 多変数関数の微分

関数 $f : \mathbb{R}^k \rightarrow \mathbb{R}$ に対して, $\Phi \in \mathbb{R}^{k \times d}$, $\boldsymbol{\theta} \in \mathbb{R}^d$, $g(\boldsymbol{\theta}) = f(\Phi\boldsymbol{\theta})$ とすると

$$\frac{\partial g}{\partial \theta_j}(\boldsymbol{\theta}) = \sum_k \frac{\partial f}{\partial z_k}(\Phi\boldsymbol{\theta}) \frac{\partial (\Phi\boldsymbol{\theta})_k}{\partial \theta_j} = \sum_k \frac{\partial f}{\partial z_k} \Phi_{kj}$$

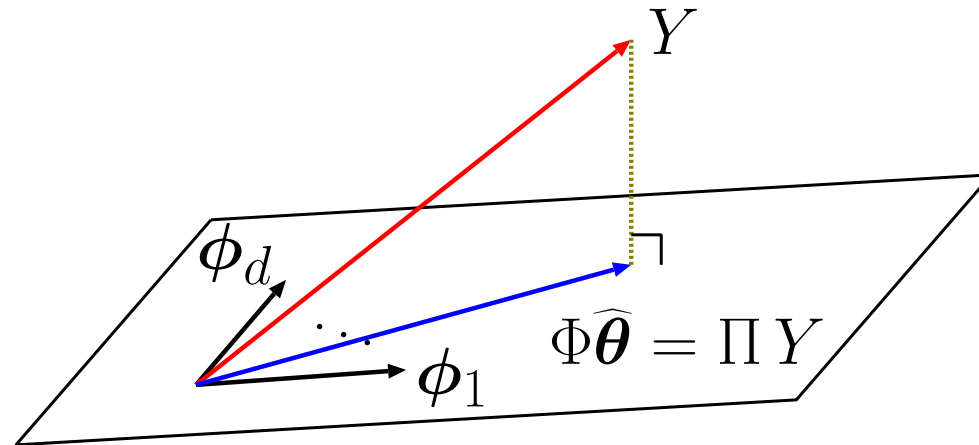
したがって $\frac{\partial f}{\partial \mathbf{z}} = \left(\frac{\partial f}{\partial z_1}, \dots, \frac{\partial f}{\partial z_k} \right)^T$ とすると

$$\frac{\partial g}{\partial \boldsymbol{\theta}}(\Phi\boldsymbol{\theta}) = \Phi^T \frac{\partial f}{\partial \mathbf{z}}(\Phi\boldsymbol{\theta}).$$

最小2乗推定量の幾何的解釈

$\Phi = (\phi_1, \dots, \phi_d) \in \mathbb{R}^{n \times d}$ とする： $\min_{\theta} \|Y - \Phi\theta\|^2 \rightarrow \theta = \hat{\theta}$.
 $\Rightarrow \Phi\hat{\theta}$ ： ϕ_1, \dots, ϕ_d で張られる \mathbb{R}^n の部分空間への $Y \in \mathbb{R}^n$ の射影.

最小2乗法： $\hat{\theta} = (\Phi^T\Phi)^{-1}\Phi^TY$ なので $\Phi\hat{\theta} = \Phi(\Phi^T\Phi)^{-1}\Phi^TY$.
 $\Pi = \Phi(\Phi^T\Phi)^{-1}\Phi^T \in \mathbb{R}^{n \times n}$ は $\text{span}\{\phi_1, \dots, \phi_d\}$ への直交射影.



射影の補足

- 射影演算子： $\Pi = \Pi^T$, $\Pi^2 = \Pi$ を満たす正方行列 Π .
- 射影演算子 Π の性質
 - Π の固有値は 0 か 1.
 - $S = \text{range}(\Pi)$ とすると, $\Pi\mathbf{y}$ は $\mathbf{y} \in \mathbb{R}^n$ の S への直交射影.
 $\mathbf{a} \in S$ に対して $\Pi\mathbf{a} = \mathbf{a}$, $\mathbf{b} \in S^\perp$ に対して $\Pi\mathbf{b} = \mathbf{0}$.
 - $\mathbf{y} = \Pi\mathbf{y} + (I - \Pi)\mathbf{y}$ は, S と S^\perp 方向への \mathbf{y} の直交分解.

最適な回帰関数

仮定：データ (\boldsymbol{x}, y) の確率密度を $p(\boldsymbol{x}, y)$ とする。
期待 2 乗誤差のもとで最適な関数は

$$f^*(\boldsymbol{x}) = E[y|\boldsymbol{x}] = \int yp(y|\boldsymbol{x})dy.$$

Proof. 任意の関数 $g(\boldsymbol{x})$ に対して以下が成立.

$$\mathbb{E}[(y - g(\boldsymbol{x}))^2] = \mathbb{E}[(y - f^*(\boldsymbol{x}))^2] + \mathbb{E}[(f^*(\boldsymbol{x}) - g(\boldsymbol{x}))^2].$$



近似誤差・推定誤差

- 線形モデル： $g_{\theta}(\mathbf{x}) = \phi(\mathbf{x})^T \boldsymbol{\theta}$.
- 期待 2 乗誤差ものとして最適なパラメータ $\boldsymbol{\theta}^*$ ：以下を満たす.

$$\min_{\boldsymbol{\theta}} \mathbb{E}[(f^*(\mathbf{x}) - g_{\boldsymbol{\theta}}(\mathbf{x}))^2] = \mathbb{E}[(f^*(\mathbf{x}) - g_{\boldsymbol{\theta}^*}(\mathbf{x}))^2].$$

線形モデルに対して次式が成立：

$$\begin{aligned} \mathbb{E}[(y - g_{\boldsymbol{\theta}}(\mathbf{x}))^2] &= \mathbb{E}[(y - f^*(\mathbf{x}))^2] + \mathbb{E}[(f^*(\mathbf{x}) - g_{\boldsymbol{\theta}}(\mathbf{x}))^2] \\ &= \underbrace{\mathbb{E}[(y - f^*(\mathbf{x}))^2]}_{\text{ノイズ}} + \underbrace{\mathbb{E}[(f^*(\mathbf{x}) - g_{\boldsymbol{\theta}^*}(\mathbf{x}))^2]}_{\text{線形モデルの誤差}} + \underbrace{\mathbb{E}[(g_{\boldsymbol{\theta}^*}(\mathbf{x}) - g_{\boldsymbol{\theta}}(\mathbf{x}))^2]}_{\text{パラメータの誤差}} \end{aligned}$$

- 大数の法則： $\frac{1}{n} \sum_{i=1}^n (y_i - g_{\boldsymbol{\theta}}(\mathbf{x}_i))^2 \xrightarrow{p} \mathbb{E}[(y - g_{\boldsymbol{\theta}}(\mathbf{x}))^2]$.
 - 最小 2 乗法：「パラメータの誤差」を近似的に最小化.
 - データ数 $\rightarrow \infty$ のとき $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$.

最小 2 乗推定量の性質：

- 一致性： $\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^*$ ($n \rightarrow \infty$)
- 不変性： $f^*(\boldsymbol{x}) = g_{\boldsymbol{\theta}^*}(\boldsymbol{x})$ のとき, $\mathbb{E}[\hat{\boldsymbol{\theta}}] = \boldsymbol{\theta}^*$

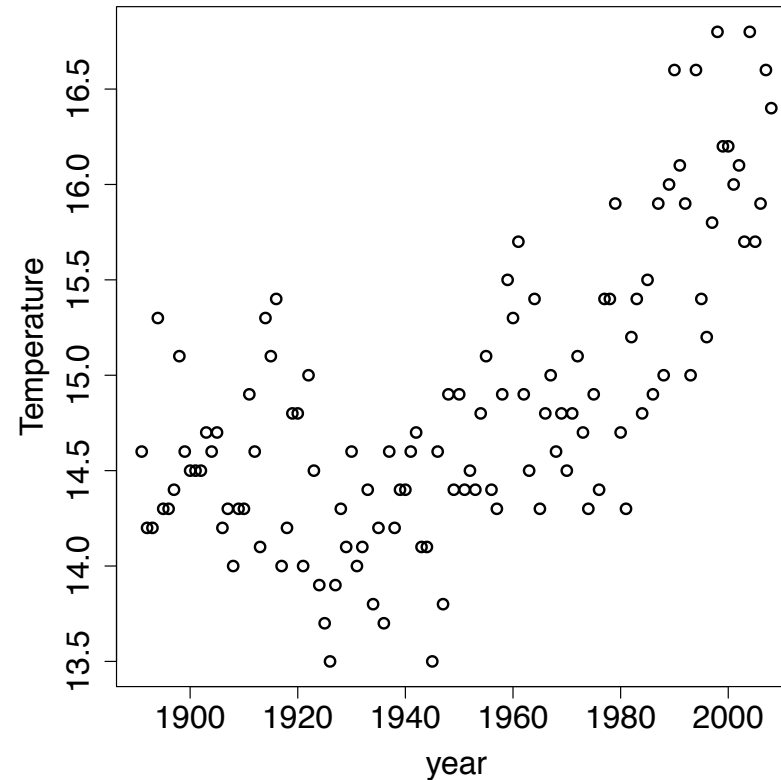
$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\theta}}] &= \int (\Phi^T \Phi)^{-1} \Phi^T Y p(y_1 | \boldsymbol{x}_1) p(\boldsymbol{x}_1) \cdots p(y_n | \boldsymbol{x}_n) p(\boldsymbol{x}_n) dy dx \\ &= \int (\Phi^T \Phi)^{-1} (\Phi^T \Phi) \boldsymbol{\theta}^* p(\boldsymbol{x}_1) \cdots p(\boldsymbol{x}_n) dx = \boldsymbol{\theta}^*\end{aligned}$$

別の書き方：

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\theta}}] &= \mathbb{E}_x \mathbb{E}_{y|x} [(\Phi^T \Phi)^{-1} \Phi^T Y] = \mathbb{E}_x [(\Phi^T \Phi)^{-1} \Phi^T \mathbb{E}_{y|x} [Y | \boldsymbol{x}]] \\ &= \mathbb{E}_x [(\Phi^T \Phi)^{-1} \Phi^T \Phi \boldsymbol{\theta}^*] = \mathbb{E}_x [\boldsymbol{\theta}^*] = \boldsymbol{\theta}^*\end{aligned}$$

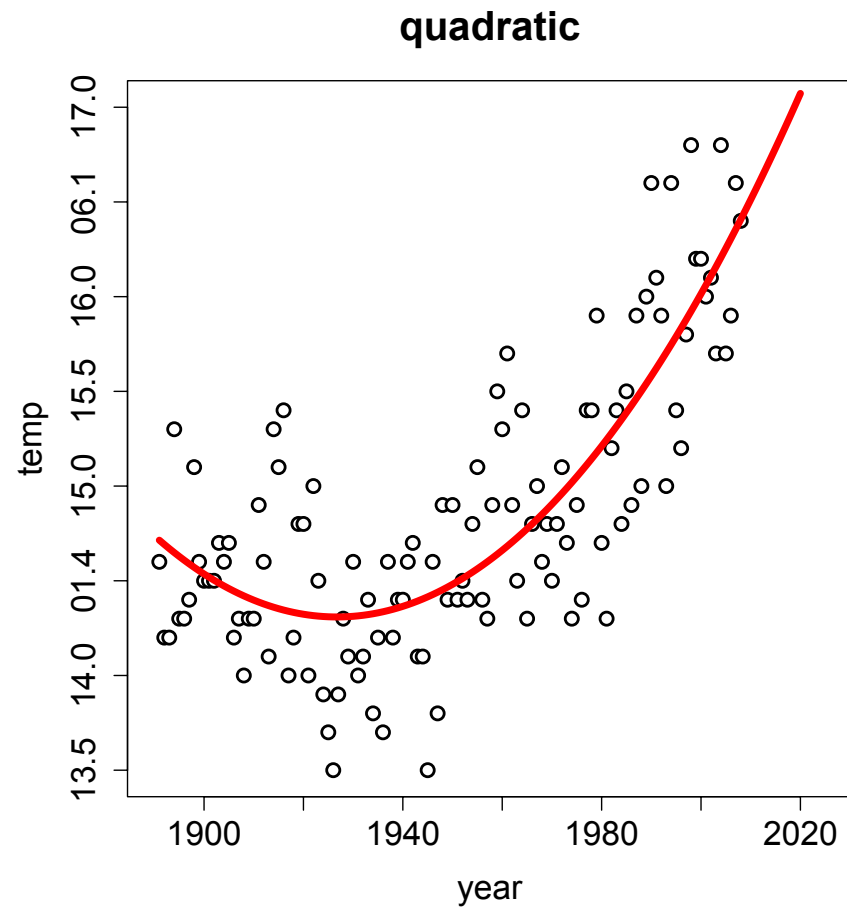
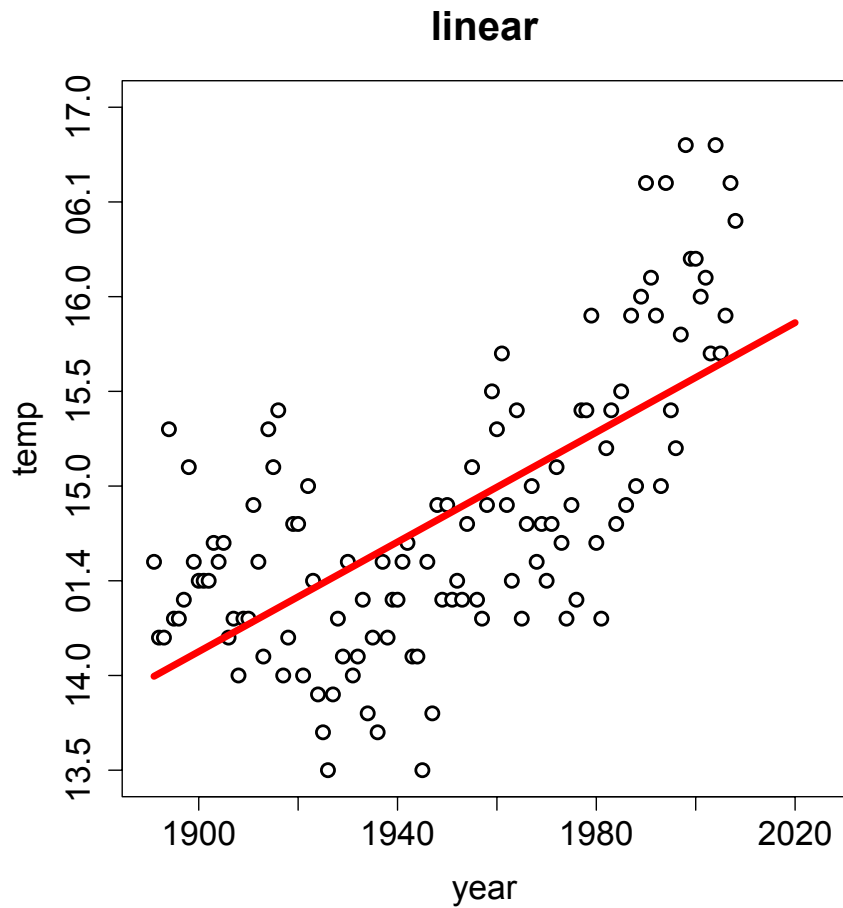
データ解析の例：名古屋の平均気温

データ：名古屋の平均気温 (1891～2008年). 気象庁のホームページにデータがある



1次式： $y_i = \theta_0 + \theta_1 x_i + \varepsilon_i,$

2次式： $y_i = \theta_0 + \theta_1 x_i + \theta_2 x_i^2 + \varepsilon_i$



- 1次式よりも2次式のほうが良くあてはまっている
- どちらが良いかを定める客観的な方法：クロスバリデーションなど

§ 最小2乗法の性質, 交差検証法

- 天気データ：1次式と2次式のどちらが良いか？
→ 予測誤差が小さいほうがよい。

関数 $f(\boldsymbol{x})$ の予測誤差 (テスト誤差) :

$$e(f) = \mathbb{E}_{\boldsymbol{x}, y}[(y - f(\boldsymbol{x}))^2] = \int (y - f(\boldsymbol{x}))^2 p(\boldsymbol{x}, y) d\boldsymbol{x} dy$$

(\boldsymbol{x}, y) : 将来のデータ

- 学習誤差 (**training error**) : 観測データに対する誤差

$$\hat{e}(f) = \frac{1}{n} \sum_{i=1}^n (y_i - f(\boldsymbol{x}_i))^2.$$

最小2乗推定量の学習誤差 : $\hat{e}(g_{\hat{\theta}}) = \min_{\theta} \hat{e}(g_{\theta})$

過学習

統計モデル $M_1 \subset M_2 \subset \dots \subset M_\ell$.

例: M_j は j 次多項式モデル.

M_j のもとでの最小 2 乗推定量: $\hat{g}_j(\mathbf{x})$, $j = 1, \dots, \ell$

$$\hat{e}(\hat{g}_j) = \min_{g \in M_j} \hat{e}(g).$$

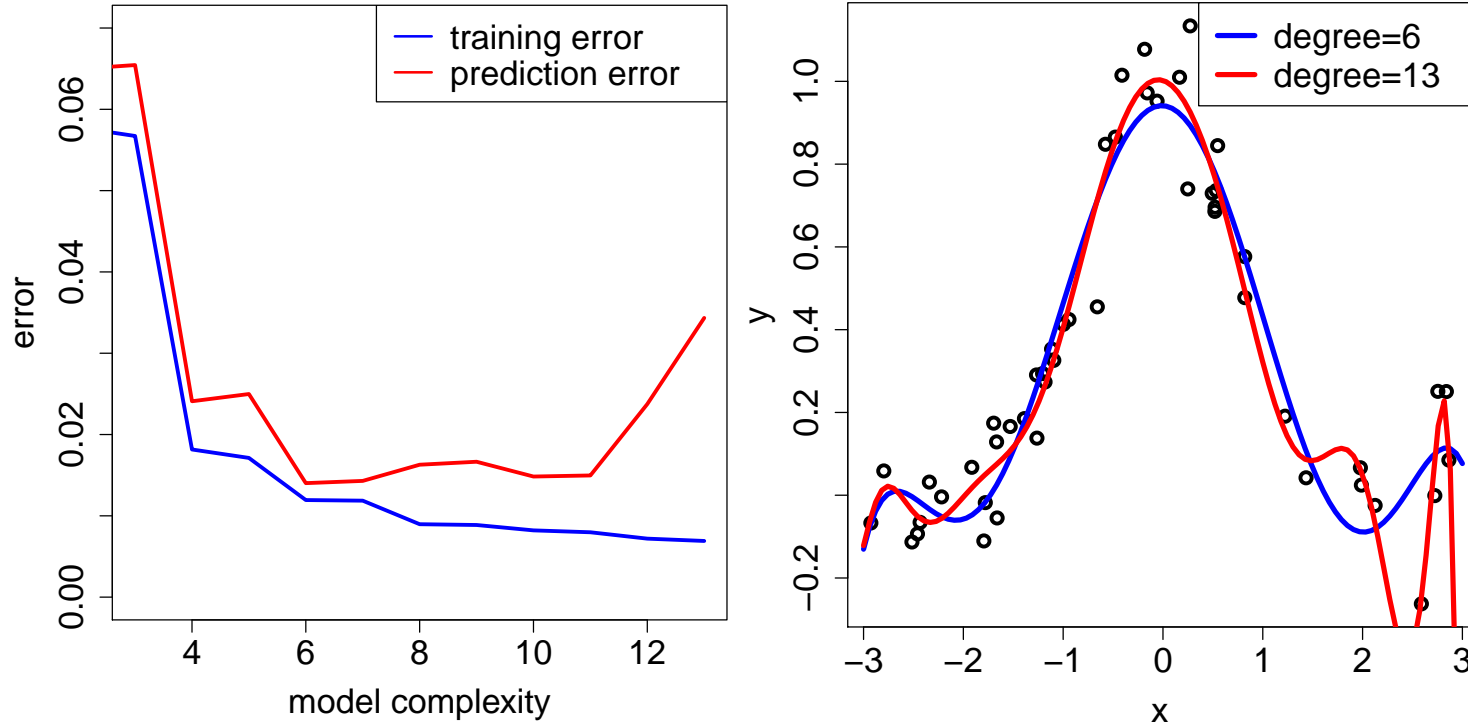
以下が成立:

$$\hat{e}(\hat{g}_1) \geq \hat{e}(\hat{g}_2) \geq \dots \geq \hat{e}(\hat{g}_\ell).$$

どの \hat{g}_j を用いるのがよいか?

- $e(\hat{g}_j)$ を小さくする \hat{g}_j を使いたい.
- 注意: $\hat{e}(\hat{g}_j)$ は $e(\hat{g}_j)$ のよい推定値にはならない.
(とくに j が大きいとき)

例



- 過学習

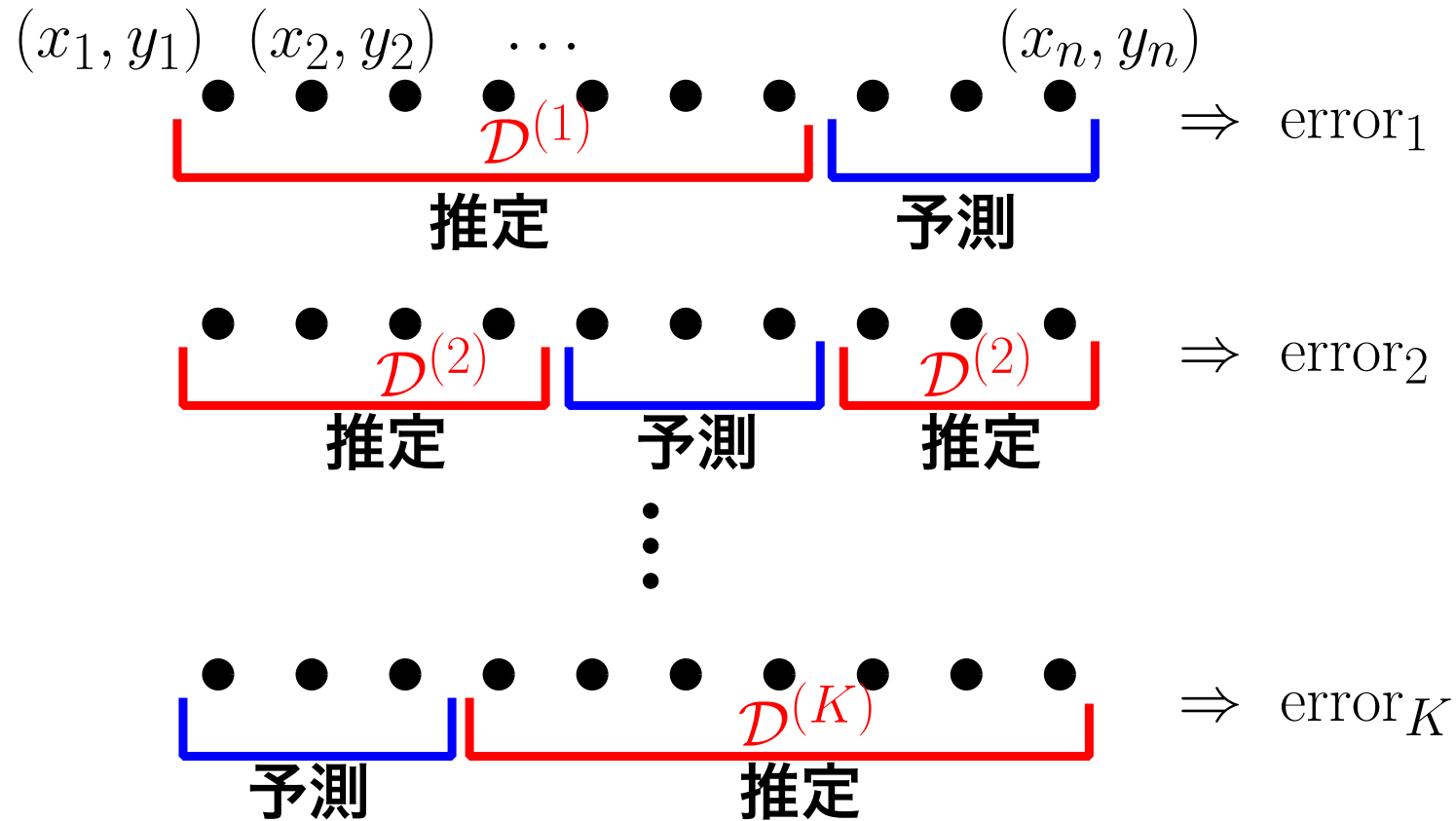
- モデルの次元が高いほど学習誤差は小さくなる。
一方、予測誤差は大きくなる傾向がある。
- 学習誤差：学習に使うデータで評価も行う → 評価が甘くなる

交差検証法 (cross validation, cv)

- 目標：推定量の予測誤差を推定
→ 適切なモデル M_j を選択
- ポイント：データを分ける
 - \hat{g}_j の学習に使うデータ
 - $e(\hat{g}_j)$ の推定に使うデータ

K重交差検証法 (K-fold cross validation)

推定方法のイメージ：データをK個のグループに分割



平均 $\hat{e}_{cv} = \frac{1}{K} \sum_{\ell=1}^K \text{error}_\ell$ を予測誤差の推定値とする.

K 重交差検証法のアルゴリズム

1. データ $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ をほぼ同じサイズの K 個のグループ $\mathcal{D}_1, \dots, \mathcal{D}_K$ に分割. $\mathcal{D}^{(k)}$ を以下のように定める.

$$\mathcal{D}_i \cap \mathcal{D}_j = \emptyset \quad (i \neq j), \quad \bigcup_{i=1}^K \mathcal{D}_i = \mathcal{D}, \quad \mathcal{D}^{(k)} := \mathcal{D} \setminus \mathcal{D}_k = \bigcup_{i \neq k} \mathcal{D}_i.$$

2. $\ell = 1, \dots, K$ に対して以下を繰り返す.
 - (a) $\mathcal{D}^{(\ell)}$ を用いて学習: $\hat{f}_\ell(\mathbf{x})$
 - (b) \mathcal{D}_ℓ に対する $\hat{f}_\ell(\mathbf{x})$ の誤差を error_ℓ とする.

$$\text{error}_\ell := \frac{1}{|\mathcal{D}_\ell|} \sum_{(x,y) \in \mathcal{D}_\ell} (y - \hat{f}_\ell(\mathbf{x}))^2 \quad (\text{推定に使ってないデータの予測})$$

3. 出力: 予測誤差の推定値 $\hat{e}_{\text{cv}} = \frac{1}{K} \sum_{\ell=1}^K \text{error}_\ell.$

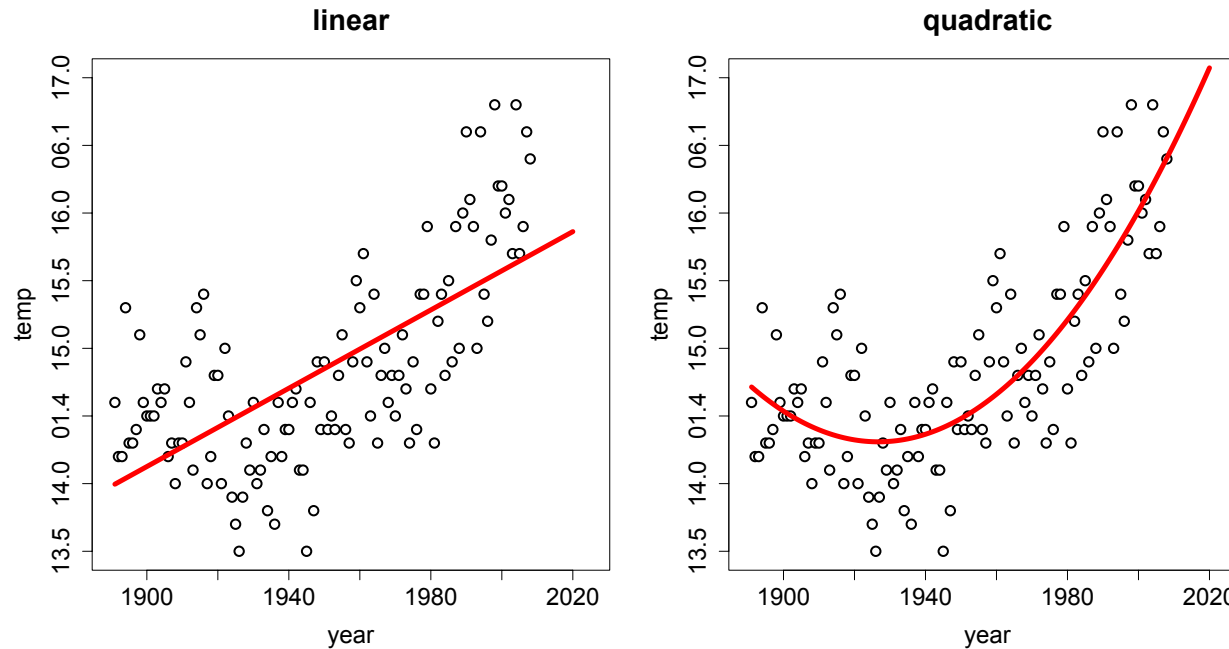
cross validation によるモデル選択

1. 統計モデル: M_1, \dots, M_ℓ
2. M_j を用いるときの予測誤差を K -**cv** で推定 $\rightarrow \hat{e}_{\text{cv},j}$
3. $\min_{j=1, \dots, \ell} \hat{e}_{\text{cv},j} \rightarrow \hat{j}$. 統計モデル $M_{\hat{j}}$ を使う.

天気データの例

5-CV ($K = 5$) : データをランダムにほぼ均等に5分割. \hat{e}_{CV} を計算.

- 1次式モデル M_1 : $\hat{e}_{CV,1} = 0.366$
- 2次式モデル M_2 : $\hat{e}_{CV,2} = 0.311$



2次式モデルのほうが予測精度が高いことが期待される.

K -CV と予測誤差

- K -CV : 推定量 $D \mapsto \hat{f}_D$ (D は観測データ)

$$E_D[e(\hat{f}_D)] = E_D[E_{x,y}[(y - \hat{f}_D(x))^2]] \text{ を推定.}$$

データ数 $n(K - 1)/K$ のときの予測誤差.

K が小さいと推定バイアスが大きい.

- K の決めかた : K が大きいほど計算コストが大きい.
計算リソースから決める. $K = 5 \sim 10$ 程度.

§ 統計解析特論：高次元モデルと正則化

予測誤差が小さな推定量が望ましい。

—— 2つのアプローチ ——

- モデル選択：いくつかのモデルから選択
- 正則化：大きなモデルを適当に制約

正則化が最近の主流：**Ridge**正則化，**Lasso**正則化。

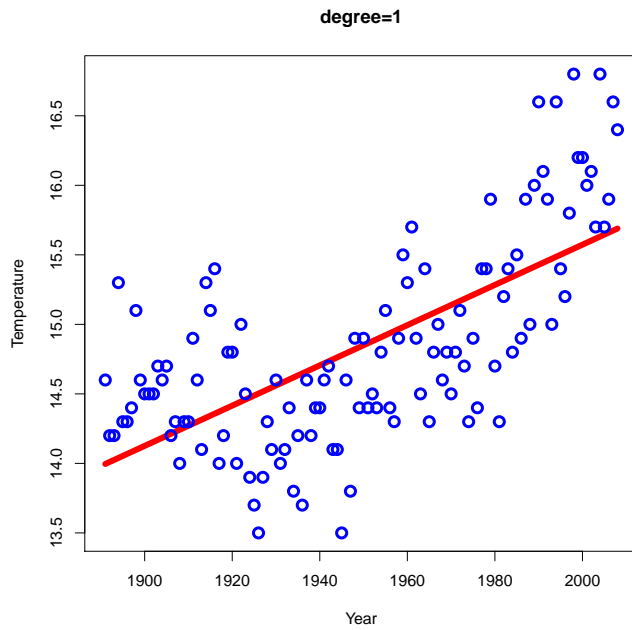
モデル選択より扱いやすい：

大きなモデルを1つ設定。階層的なモデルを設定しなくてよい。

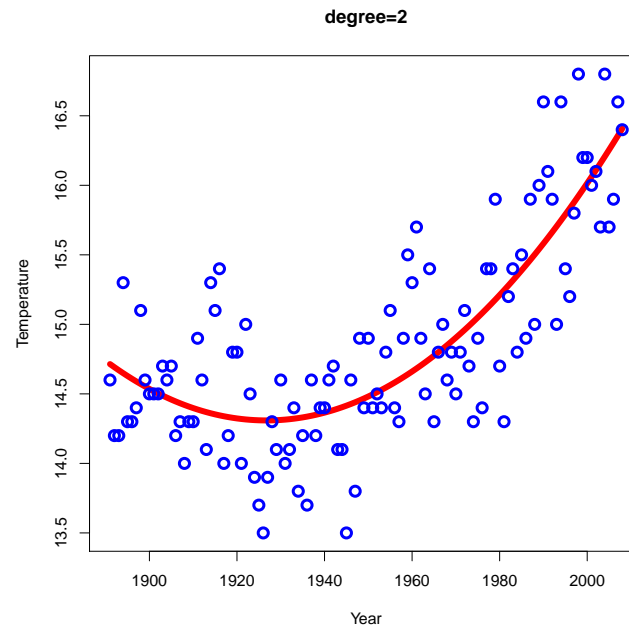
- 複雑なデータの解析
 - 自由度の大きなモデルを使う。
- モデルの自由度が大きすぎてもうまくいかない：
 - データに過剰に適合して、将来の予測がうまくいかない

● 最小2乗法で推定した結果：

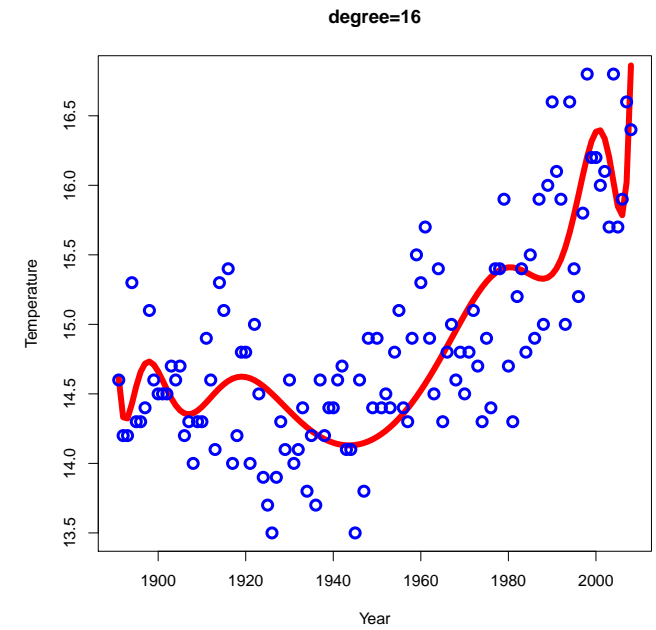
自由度 (次数): 小



自由度 (次数): 中



自由度 (次数): 大



正則化：モデルの自由度を調整

- 大きな次元の統計モデル：さまざまなデータに対応

$$Y = \sum_{k=1}^d \theta_k \phi_k(\mathbf{x}) + b + \varepsilon, \quad d \gg 1$$

定数項 b と基底関数 $\phi_k(\mathbf{x})$ を分けることが多い。

例 1. $\phi(x) = (x, x^2, x^3, \dots, x^{100})$.

データ : $\{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$.

正則化項付き最小2乗法

$$\min_{\theta, b} \sum_{i=1}^n (y_i - (\phi(\mathbf{x}_i)^T \theta + b))^2 + \frac{\lambda R(\theta)}{\text{正則化項}}$$

\implies 最適解 $\hat{\theta}, \hat{b}$. $\hat{f}(\mathbf{x}) = \phi(\mathbf{x})^T \theta + \hat{b}$

正則化パラメータ $\lambda > 0$.

- $\lambda \rightarrow$ 大 : 正則化強い (モデルの自由度小さい)
- $\lambda \rightarrow$ 小 : 正則化弱い (モデルの自由度大きい)

正則化項 $R(\theta)$: データへの **overfit** を防ぐ.

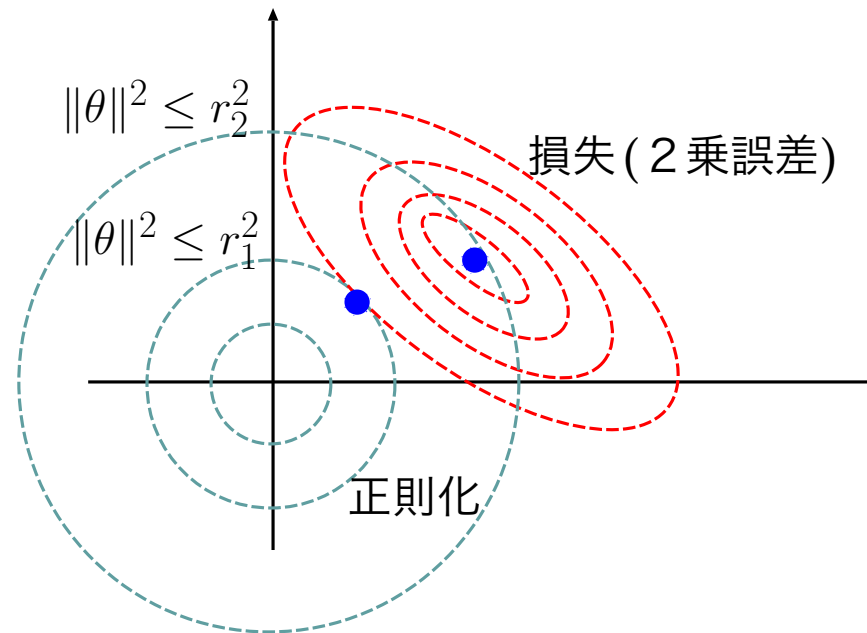
- L_0 正則化 : $R(\theta) = \|\theta\|_0$: 「 $(\theta_1, \dots, \theta_d)$ 中の非ゼロ要素の数」
 - 基底数を制約. 組合せ最適化. 高次元だと実用的でない.
- L_1 正則化 (**Lasso**) : $R(\theta) = \|\theta\|_1 = \sum_{i=1}^d |\theta_i|$
 - L_0 正則化の近似. 凸2次計画問題.
スパース解 (θ の要素はほとんど0) を得る.
- L_2 正則化 (**Ridge**) : $R(\theta) = \|\theta\|_2^2 = \sum_{i=1}^d \theta_i^2$
 - 計算しやすい.

正則化項付き最小 2 乗法は以下と等価：

$$\min_{\theta, b} \sum_{i=1}^n (y_i - (\phi(\mathbf{x}_i)^T \theta + b))^2, \quad \text{s.t. } R(\theta) \leq r$$

λ と r の対応：大きい $\lambda \Leftrightarrow$ 小さい r .

Ridge の場合：



Ridge 推定量の計算：

$$\min_{\theta, b} \|Y - \Phi\theta - b\mathbf{1}\|^2 + \lambda\|\theta\|^2, \quad (\Phi_{ij} = \phi_j(\mathbf{x}_i))$$

関数 $L(\theta, b) = \|Y - \Phi\theta - b\mathbf{1}\|^2 + \lambda\|\theta\|^2$ として

$$\frac{\partial L}{\partial \theta} = -2\Phi^T(Y - \Phi\theta - b\mathbf{1}) + 2\lambda\theta = 0$$

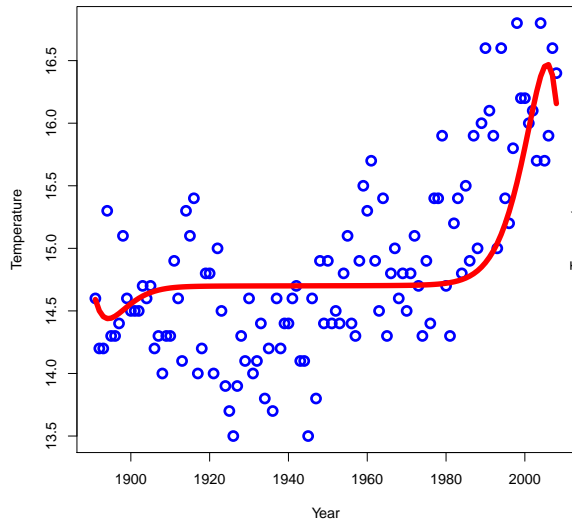
$$\frac{\partial L}{\partial b} = -2\mathbf{1}^T(Y - \Phi\theta - b\mathbf{1}) = 0$$

$$\begin{pmatrix} \Phi^T\Phi + \lambda I & \Phi^T\mathbf{1} \\ \mathbf{1}^T\Phi & n \end{pmatrix} \begin{pmatrix} \theta \\ b \end{pmatrix} = \begin{pmatrix} \Phi^TY \\ \mathbf{1}^TY \end{pmatrix}$$

- 15次多項式モデル：
$$y = \sum_{k=1}^{15} x^k \theta_k + b + \varepsilon$$
- Ridge の正則化パラメータ $\lambda > 0$

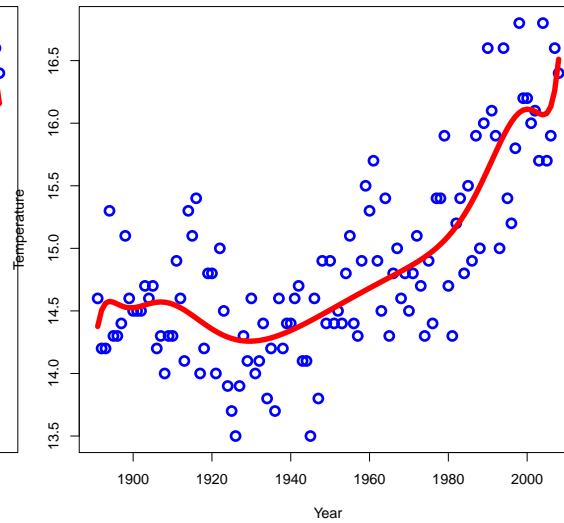
λ : 大

lambda=10000



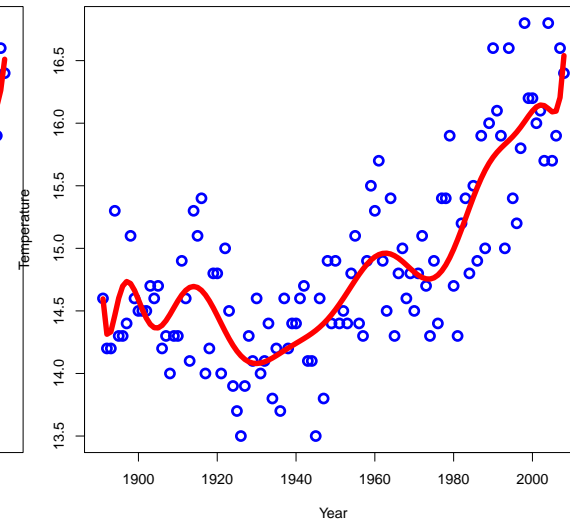
λ : 中

lambda=1



λ : 小

lambda=1e-05



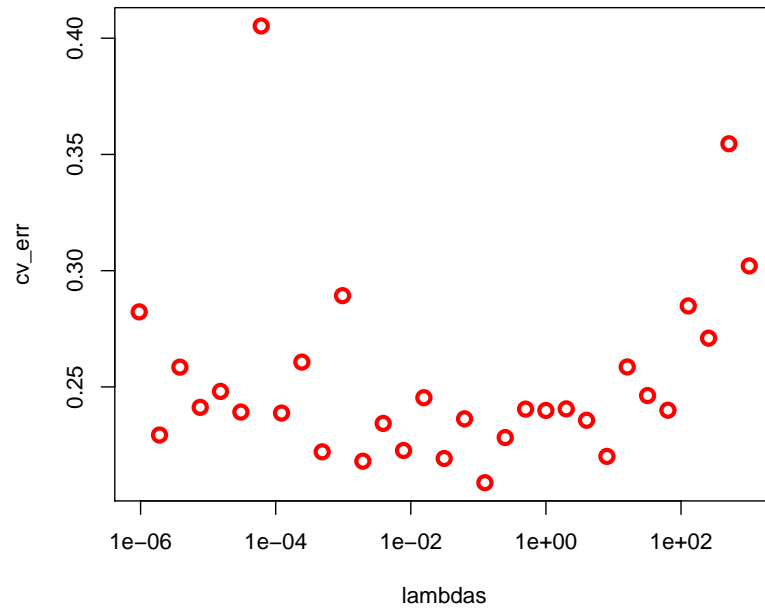
小さい ← モデル自由度 (表現力) → 大きい

正則化パラメータの選択

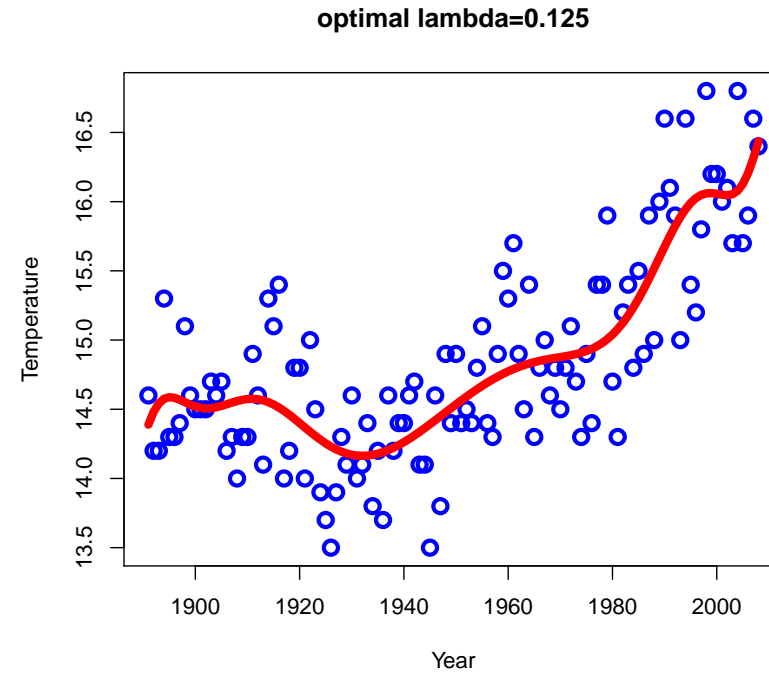
- 正則化パラメータをうまく選べば, 予測誤差が小さくなる.
- **cross validation** で正則化パラメータを選ぶ.
 1. λ の候補 $\{\lambda_1, \lambda_2, \dots, \lambda_m\}$ を設定.
 2. 各 $\lambda_k, k = 1, \dots, m$ に対して, **cv** で予測誤差を推定.
 $\hat{e}_{cv}(\lambda_k), k = 1, \dots, m$ を得る.
 3. $\hat{e}_{cv}(\lambda_k)$ を最小にする正則化パラメータ λ_k を用いる.

- 15次多項式モデル： $y = \sum_{k=1}^{15} x^k \theta_k + b + \varepsilon$
- 正則化パラメータ $\lambda = 2^{-20} \sim 2^{10}$

λに対する 5-CV のプロット



CVで選ばれたλによる回帰

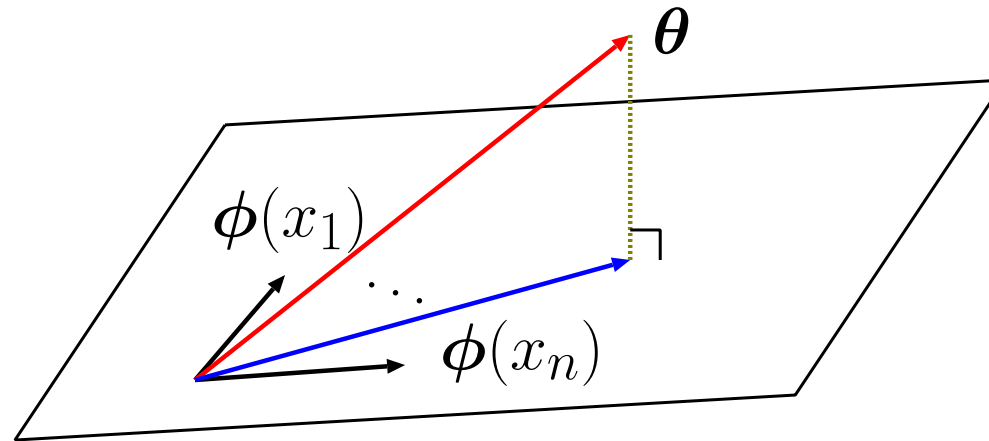


§ カーネル回帰分析

「大きなモデル+正則化」 \implies カーネル法

$$\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \sum_{i=1}^n (y_i - \boldsymbol{\theta}^T \boldsymbol{\phi}(\mathbf{x}_i))^2 + \lambda \|\boldsymbol{\theta}\|^2$$

- $\boldsymbol{\theta} \in \text{span}\{\boldsymbol{\phi}(\mathbf{x}_1), \dots, \boldsymbol{\phi}(\mathbf{x}_n)\}$ の範囲で考えれば十分.



Proof. $\mathcal{S} = \text{span}\{\phi(\mathbf{x}_1), \dots, \phi(\mathbf{x}_n)\}$ とおく. $\boldsymbol{\theta} = \boldsymbol{\theta}_1 + \boldsymbol{\theta}_2$, $\boldsymbol{\theta}_1 \in \mathcal{S}$, $\boldsymbol{\theta}_2 \in \mathcal{S}^\perp$ とおくと, $\boldsymbol{\theta}^T \phi(\mathbf{x}_i) = \boldsymbol{\theta}_1^T \phi(\mathbf{x}_i)$, $\|\boldsymbol{\theta}\| \geq \|\boldsymbol{\theta}_1\|$. したがって

$$\sum_{i=1}^n (y_i - \boldsymbol{\theta}^T \phi(\mathbf{x}_i))^2 + \lambda \|\boldsymbol{\theta}\|^2 \geq \sum_{i=1}^n (y_i - \boldsymbol{\theta}_1^T \phi(\mathbf{x}_i))^2 + \lambda \|\boldsymbol{\theta}_1\|^2$$

よって \mathcal{S} の範囲で探索すれば十分. ■

$$\boldsymbol{\theta} = \sum_{j=1}^n \phi(\mathbf{x}_j) \beta_j = \Phi \boldsymbol{\beta} \text{ を 2 乗誤差に代入. } \boldsymbol{\beta} = (\beta_1, \dots, \beta_n)^T$$

$$\begin{aligned} & \sum_{i=1}^n (y_i - \boldsymbol{\theta}^T \phi(\mathbf{x}_i) - b)^2 + \lambda \|\boldsymbol{\theta}\|^2 \\ &= \|\mathbf{y} - \Phi^T \Phi \boldsymbol{\beta} - b \mathbf{1}\|^2 + \lambda \boldsymbol{\beta}^T \Phi^T \Phi \boldsymbol{\beta} \longrightarrow \min_{\boldsymbol{\beta}, b} \end{aligned}$$

カーネル法

- カーネル関数の定義： $k(\mathbf{x}, \mathbf{x}') := \phi(\mathbf{x})^T \phi(\mathbf{x}') \in \mathbb{R}$
 $n \times n$ 行列 $K_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) \implies K = \Phi^T \Phi$

正則化項付き 2乗誤差： $\|\mathbf{y} - K\boldsymbol{\beta} - b\mathbf{1}\|^2 + \lambda\boldsymbol{\beta}^T K\boldsymbol{\beta}$

→ 最適解 $\hat{\boldsymbol{\beta}}, \hat{b}$

- 回帰関数： $\hat{f}(\mathbf{x}) = \phi(\mathbf{x})^T \sum_{i=1}^n \phi(\mathbf{x}_i) \hat{\beta}_i + \hat{b} = \sum_{i=1}^n k(\mathbf{x}, \mathbf{x}_i) \hat{\beta}_i + \hat{b}$
- 関数 $k(\mathbf{x}, \mathbf{x}')$ だけから推定量 $\hat{f}(\mathbf{x})$ を計算可能.
→ 最初から $k(\mathbf{x}, \mathbf{x}')$ を与えて推定.

カーネル回帰分析の特徴

- $\phi(\mathbf{x}) = (\phi_1(\mathbf{x}), \phi_2(\mathbf{x}), \dots) \in \mathbb{R}^D$ とする.
 $D = \infty$ でも, $k(\mathbf{x}, \mathbf{x}')$ を簡単に計算できる場合がある.
例: ガウシアンカーネル (→次頁)
- 線形モデル $y = \mathbf{x}^T \boldsymbol{\theta} + \varepsilon$ とほぼ同じ手間で, 表現力が大幅アップ.

カーネル関数の例： $\mathbf{x} \in \mathbb{R}^d \mapsto \phi(\mathbf{x}) \in \mathbb{R}^D$.

- 線形カーネル： $k(\mathbf{x}, \mathbf{x}') = \mathbf{x}^T \mathbf{x}'$, $D = d$
- 多項式カーネル： $k(\mathbf{x}, \mathbf{x}') = (1 + \mathbf{x}^T \mathbf{x}')^k$, $k \in \mathbb{N}$, $D = \frac{(k + d)!}{k! d!}$.
 k 次以下の単項式全体の線形和.
- ガウシアンカーネル： $D = \infty$

$$k(\mathbf{x}, \mathbf{x}') = \exp\{-\gamma \cdot \|\mathbf{x} - \mathbf{x}'\|_2^2\}, \quad \gamma > 0$$