回帰分析

予測と発展的なモデル

村田 昇

講義の内容

- 第 1 回: 回帰モデルの考え方と推定
- 第 2 回: モデルの評価
- **第 3 回: モデルによる予測と発展的なモデル**

回帰分析の復習

線形回帰モデル

- **目的変数** を **説明変数** で説明する関係式を構成
  - 説明変数: $x_1, \ldots, x_p$ (p 次元)
  - 目的変数: y (1 次元)
- **回帰係数** $\beta_0, \beta_1, \ldots, \beta_p$ を用いた一次式

$$y = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

- **誤差項** を含む確率モデルで観測データを表現

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i \quad (i = 1, \ldots, n)$$

問題設定

- 確率モデル

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- 式の評価 : **残差平方和** の最小化による推定

$$S(\boldsymbol{\beta}) = (\boldsymbol{y} - X\boldsymbol{\beta})^\mathsf{T}(\boldsymbol{y} - X\boldsymbol{\beta})$$

解

- 解の条件 : **正規方程式**

$$X^\mathsf{T} X\boldsymbol{\beta} = X^\mathsf{T}\boldsymbol{y}$$

- 解の一意性 : **Gram 行列** $X^\mathsf{T}X$ が正則

$$\hat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y}$$

実データによる例

- 東京の 8 月の気候 (気温, 降雨, 日射, 降雪, 風速, 気圧, 湿度, 雲量) に関するデータ (の一部)

	month	day	day_of_week	temp	rain	solar	snow	wdir	wind	press	humid	cloud
213	8	1	Sun	28.7	0.0	26.58	0	SSE	3.2	1000.2	76	2.3
214	8	2	Mon	28.6	0.5	19.95	0	SE	3.4	1006.1	80	7.0
215	8	3	Tue	29.0	3.0	19.89	0	S	4.0	1009.9	80	6.3
216	8	4	Wed	29.5	0.0	26.52	0	S	3.0	1008.2	76	2.8
217	8	5	Thu	29.1	0.0	26.17	0	SSE	2.8	1005.1	74	5.8
218	8	6	Fri	29.1	0.0	24.82	0	SSE	2.9	1004.2	75	4.0
219	8	7	Sat	27.9	2.0	11.43	0	NE	2.5	1003.1	85	9.0
220	8	8	Sun	25.9	90.5	3.43	0	N	3.0	998.0	97	10.0
221	8	9	Mon	28.1	2.0	13.34	0	S	6.1	995.4	84	6.0
222	8	10	Tue	31.0	0.0	22.45	0	SSW	4.7	996.3	58	4.8
223	8	11	Wed	29.2	0.0	21.12	0	SE	2.9	1008.0	61	9.3
224	8	12	Thu	26.0	0.5	8.34	0	SSE	2.4	1008.8	84	9.5
225	8	13	Fri	22.5	20.5	4.36	0	NE	2.7	1008.0	97	10.0
226	8	14	Sat	22.3	77.0	2.76	0	N	2.7	1003.6	100	10.0

- 作成した線形回帰モデルを検討する
  - モデル 1 : 気温 = F(気圧)
  - モデル 2 : 気温 = F(日射)
  - モデル 3 : 気温 = F(気圧, 日射)
  - モデル 4 : 気温 = F(気圧, 日射, 湿度)
  - モデル 5 : 気温 = F(気圧, 日射, 雲量)
- 説明変数と目的変数の関係

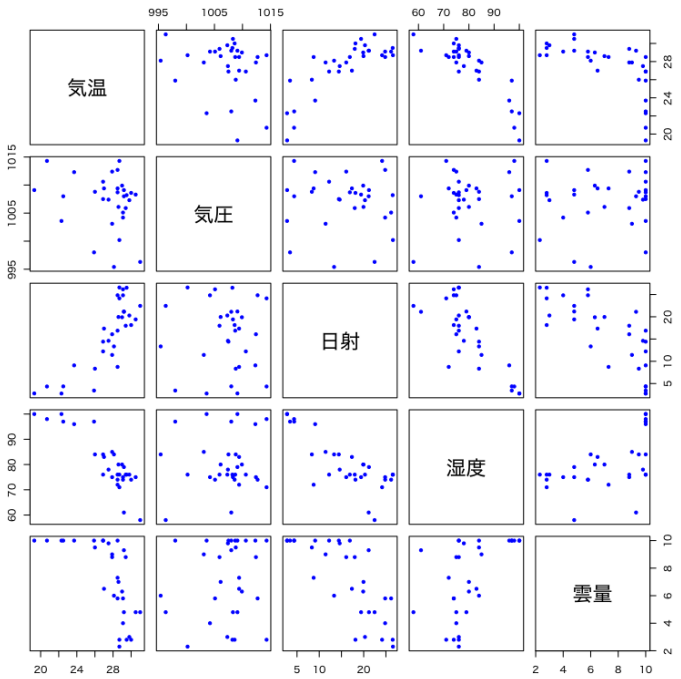

図 1: 説明変数と目的変数の散布図

- 観測値とあてはめ値の比較

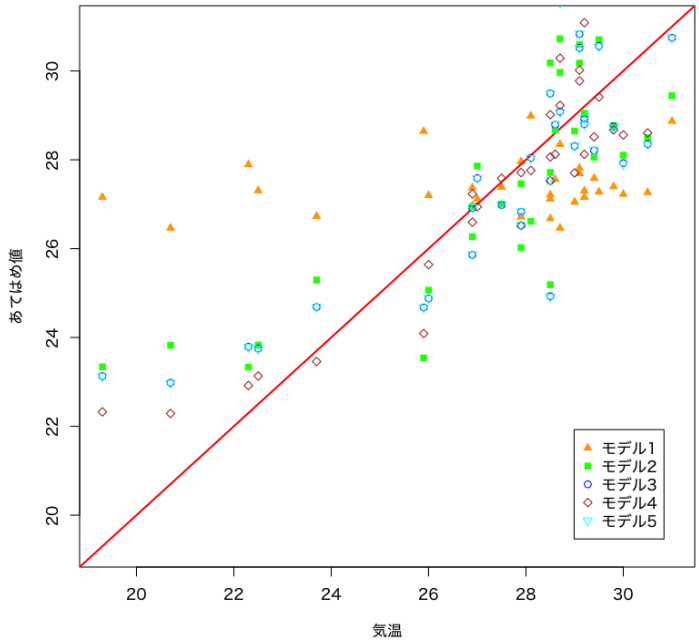

図 2: モデルの比較

寄与率

- **決定係数** (R-squared)

$$R^2 = 1 - \frac{\sum_{i=1}^{n} \hat{\epsilon}_i^2}{\sum_{i=1}^{n} (y_i - \bar{y})^2}$$

- **自由度調整済み決定係数** (adjusted R-squared)

$$\bar{R}^2 = 1 - \frac{\frac{1}{n-p-1} \sum_{i=1}^{n} \hat{\epsilon}_i^2}{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2}$$

  - 不偏分散で補正

モデルの評価

- 決定係数
  - モデル 1 : 気温 = F(気圧)

    `[1] "R2: 0.0483 ; adj. R2: 0.0155"`
  - モデル 2 : 気温 = F(日射)

    `[1] "R2: 0.663 ; adj. R2: 0.651"`
  - モデル 3 : 気温 = F(気圧, 日射)

    `[1] "R2: 0.703 ; adj. R2: 0.681"`
  - モデル 4 : 気温 = F(気圧, 日射, 湿度) (3 より改善している)

    `[1] "R2: 0.83 ; adj. R2: 0.811"`
  - モデル 5 : 気温 = F(気圧, 日射, 雲量) (3 より改善していない)

    `[1] "R2: 0.703 ; adj. R2: 0.67"`

F 統計量による検定

- 説明変数のうち 1 つでも役に立つか否かを検定する
  - 帰無仮説 H_0: $\beta_1 = \cdots = \beta_p = 0$
  - 対立仮説 H_1: $\exists j\ \beta_j \neq 0$ (少なくとも 1 つは役に立つ)
- F 統計量: 決定係数 (または残差) を用いて計算

$$F = \frac{n-p-1}{p} \frac{R^2}{1-R^2}$$

- p 値: 自由度 $p, n-p-1$ の F 分布で計算

モデルの評価

- 決定係数と F 統計量
  - モデル 1 : 気温 = F(気圧)

    ```
    [1] "R2: 0.0483 ; adj. R2: 0.0155 ; F-stat: 1.47 ; p-val: 0.235"
    ```
  - モデル 2 : 気温 = F(日射)

    ```
    [1] "R2: 0.663 ; adj. R2: 0.651 ; F-stat: 57 ; p-val: 2.52e-08"
    ```
  - モデル 3 : 気温 = F(気圧, 日射)

    ```
    [1] "R2: 0.703 ; adj. R2: 0.681 ; F-stat: 33.1 ; p-val: 4.23e-08"
    ```
  - モデル 4 : 気温 = F(気圧, 日射, 湿度)

    ```
    [1] "R2: 0.83 ; adj. R2: 0.811 ; F-stat: 43.8 ; p-val: 1.65e-10"
    ```
  - モデル 5 : 気温 = F(気圧, 日射, 雲量)

    ```
    [1] "R2: 0.703 ; adj. R2: 0.67 ; F-stat: 21.3 ; p-val: 2.81e-07"
    ```

t 統計量による検定

- 回帰係数 β_j が回帰式に寄与するか否かを検定する
  - 帰無仮説 H_0: $\beta_j = 0$
  - 対立仮説 H_1: $\beta_j \neq 0$ (β_j は役に立つ)
- t 統計量: 各係数ごと，ζ は $(X^\mathsf{T}X)^{-1}$ の対角成分

$$t = \frac{\hat{\beta}_j}{\hat{\sigma}\zeta_j}$$

- p 値: 自由度 $n-p-1$ の t 分布を用いて計算

モデルの評価

- 回帰係数の推定値と t 統計量
  - モデル 1 : 気温 = F(気圧)

    ```
                 Estimate Std. Error t value Pr(>|t|)
    (Intercept)  162.000     111.00    1.46    0.155
    press         -0.134       0.11   -1.21    0.235
    ```
    - 気圧単体では回帰係数は有意ではない
  - モデル 2 : 気温 = F(日射)

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)    22.50     0.7210   31.20 7.59e-24
solar           0.31     0.0411    7.55 2.52e-08
```

  - 日射単体の回帰係数は有意となる

- モデル 3 : 気温 = F(気圧, 日射)

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  145.000    63.1000    2.29 2.98e-02
press         -0.121     0.0627   -1.93 6.34e-02
solar          0.308     0.0393    7.85 1.50e-08
```

  - 気圧は日射と組み合わせることで有意となる

- 回帰係数の推定値と t 統計量 (つづき)
  - モデル 4: 気温 = F(気圧, 日射, 湿度)

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  147.000    48.7000    3.02 0.005470
press         -0.108     0.0484   -2.24 0.033800
solar          0.134     0.0492    2.73 0.011100
humid         -0.158     0.0353   -4.49 0.000121
```

  - モデル 5: 気温 = F(気圧, 日射, 雲量)

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 145.00000    65.0000    2.23 3.42e-02
press        -0.12200     0.0648   -1.88 7.15e-02
solar         0.31000     0.0624    4.97 3.31e-05
cloud         0.00686     0.1710    0.04 9.68e-01
```

  - このモデルでは雲量は有用でないことが示唆される

モデルの診断 (参考)

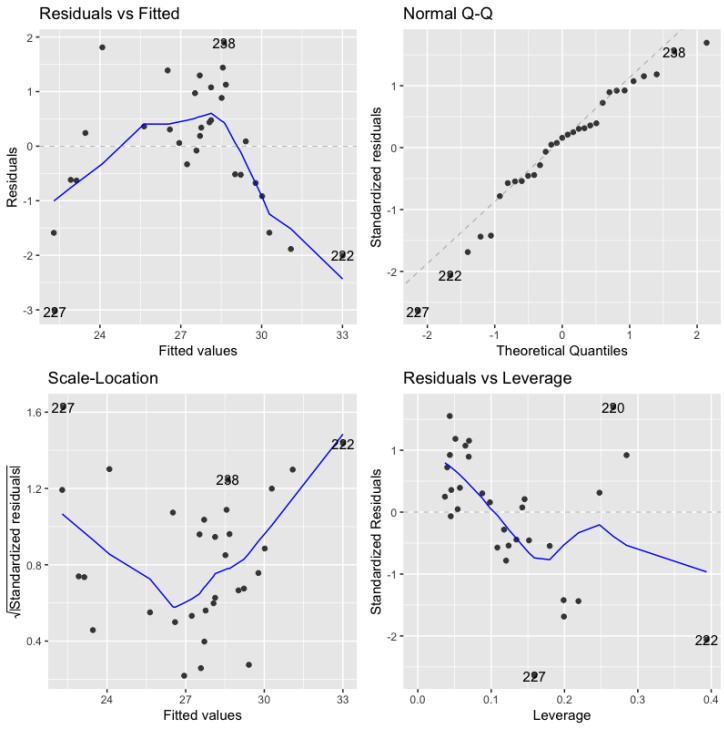

図 3: モデル 4 の診断プロット

回帰モデルによる予測

予測

- 新しいデータ (説明変数) $\boldsymbol{x}$ に対する **予測値**

$$\hat{y} = (1, \boldsymbol{x}^\mathsf{T})\hat{\boldsymbol{\beta}}, \qquad \hat{\boldsymbol{\beta}} = (X^\mathsf{T}X)^{-1}X^\mathsf{T}\boldsymbol{y}$$

- 予測値は元データの目的変数の重み付け線形和

$$\hat{y} = \boldsymbol{w}(\boldsymbol{x})^\mathsf{T}\boldsymbol{y}, \qquad \boldsymbol{w}(\boldsymbol{x})^\mathsf{T} = (1, \boldsymbol{x}^\mathsf{T})(X^\mathsf{T}X)^{-1}X^\mathsf{T}$$

  - 重みは元データと新規データの説明変数で決定

予測値の性質

- 推定量は以下の性質をもつ多変量正規分布

$$\begin{aligned}\mathbb{E}[\hat{\boldsymbol{\beta}}] &= \boldsymbol{\beta} \\ \mathrm{Cov}(\hat{\boldsymbol{\beta}}) &= \sigma^2(X^\mathsf{T}X)^{-1}\end{aligned}$$

- この性質を利用して以下の 3 つの値の違いを評価

$$\begin{aligned}\hat{y} &= (1, \boldsymbol{x}^\mathsf{T})\hat{\boldsymbol{\beta}} && \text{(回帰式による予測値)} \\ \tilde{y} &= (1, \boldsymbol{x}^\mathsf{T})\boldsymbol{\beta} && \text{(最適な予測値)} \\ y &= (1, \boldsymbol{x}^\mathsf{T})\boldsymbol{\beta} + \epsilon && \text{(観測値)}\end{aligned}$$

  - $\hat{y}$ と y は独立な正規分布に従うことに注意

演習

問題

- 誤差が平均 0 分散 σ^2 の正規分布に従うとき，以下の問に答えなさい
  - 予測値 $\hat{y}$ の平均を求めよ
  - 予測値 $\hat{y}$ の分散を求めよ

信頼区間

最適な予測値との差

- 差の分布は以下の平均・分散をもつ正規分布に従う

$$\begin{aligned}\mathbb{E}[\tilde{y} - \hat{y}] &= (1, \boldsymbol{x}^\mathsf{T})\boldsymbol{\beta} - (1, \boldsymbol{x}^\mathsf{T})\mathbb{E}[\hat{\boldsymbol{\beta}}] = 0 \\ \mathrm{Var}(\tilde{y} - \hat{y}) &= \underbrace{\sigma^2(1, \boldsymbol{x}^\mathsf{T})(X^\mathsf{T}X)^{-1}(1, \boldsymbol{x}^\mathsf{T})^\mathsf{T}}_{\hat{\boldsymbol{\beta}}\text{ の推定誤差による分散}} = \sigma^2\gamma_c(\boldsymbol{x})^2\end{aligned}$$

- 正規化による表現

$$\frac{\tilde{y} - \hat{y}}{\sigma\gamma_c(\boldsymbol{x})} \sim \mathcal{N}(0, 1)$$

信頼区間

- 未知の分散を不偏分散で推定

$$Z = \frac{\tilde{y} - \hat{y}}{\hat{\sigma}\gamma_c(\boldsymbol{x})} \sim \mathcal{T}(n-p-1) \quad (t \text{ 分布})$$

- 確率 α の信頼区間

$$\mathcal{I}_\alpha^c = (\hat{y} - C_\alpha \hat{\sigma}\gamma_c(\boldsymbol{x}),\ \hat{y} + C_\alpha \hat{\sigma}\gamma_c(\boldsymbol{x}))$$

$$P(|Z| < C_\alpha | Z \sim \mathcal{T}(n-p-1)) = \alpha$$

  - 最適な予測値 $\tilde{y}$ が入ることが期待される区間

演習

問題

- 以下の問に答えなさい
  - 信頼区間について以下の式が成り立つことを示せ

$$P(\tilde{y} \in \mathcal{I}_\alpha^c) = \alpha$$

  - 観測値と予測値の差 $y - \hat{y}$ の平均と分散を求めよ

予測区間

観測値との差

- 差の分布は以下の平均・分散をもつ正規分布に従う

$$\begin{aligned}
\mathbb{E}[y - \hat{y}] &= (1, \boldsymbol{x}^\mathsf{T})\boldsymbol{\beta} + \mathbb{E}[\boldsymbol{\epsilon}] - (1, \boldsymbol{x}^\mathsf{T})\mathbb{E}[\hat{\boldsymbol{\beta}}] = 0 \\
\mathrm{Var}(y - \hat{y}) &= \underbrace{\sigma^2 (1, \boldsymbol{x}^\mathsf{T})(X^\mathsf{T}X)^{-1}(1, \boldsymbol{x}^\mathsf{T})^\mathsf{T}}_{\hat{\boldsymbol{\beta}} \text{ の推定誤差による分散}} + \underbrace{\sigma^2}_{\text{誤差の分散}} = \sigma^2 \gamma_p(\boldsymbol{x})^2
\end{aligned}$$

- 正規化による表現

$$\frac{y - \hat{y}}{\sigma\gamma_p(\boldsymbol{x})} \sim \mathcal{N}(0, 1)$$

予測区間

- 未知の分散を不偏分散で推定

$$Z = \frac{y - \hat{y}}{\hat{\sigma}\gamma_p(\boldsymbol{x})} \sim \mathcal{T}(n-p-1) \quad (t \text{ 分布})$$

- 確率 α の予測区間

$$\mathcal{I}_\alpha^p = \left(\hat{y} - C_\alpha \hat{\sigma} \gamma_p(\boldsymbol{x}),\ \hat{y} + C_\alpha \hat{\sigma} \gamma_p(\boldsymbol{x})\right)$$

$$P(|Z| < C_\alpha | Z \sim \mathcal{T}(n-p-1)) = \alpha$$

  - 観測値 y が入ることが期待される区間
  - $\gamma_p > \gamma_c$ なので信頼区間より広くなる

解析の事例

信頼区間と予測区間

- 東京の気候データを用いて以下を試みる
  - 8 月のデータで回帰式を推定する

    気温 = F(気圧, 日射, 湿度)
  - 上記のモデルで 9 月のデータを予測する

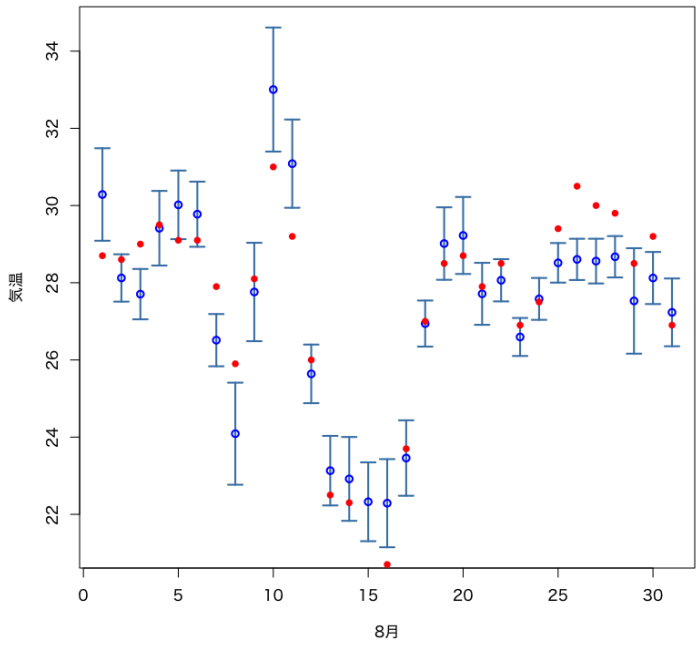

図 4: 8 月のあてはめ値の信頼区間

発展的なモデル

非線形性を含むモデル

- 目的変数 Y
- 説明変数 $X_1, \ldots, X_p$

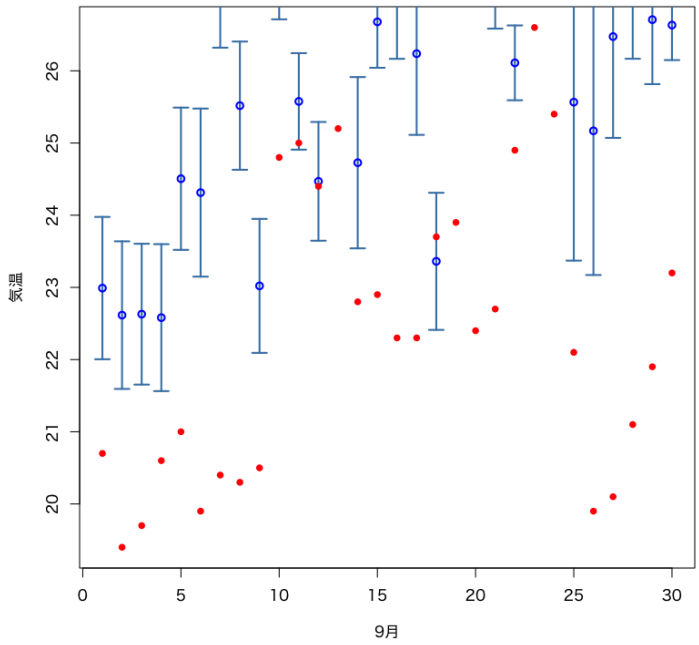

図 5: 8 月モデルによる 9 月の予測値の信頼区間

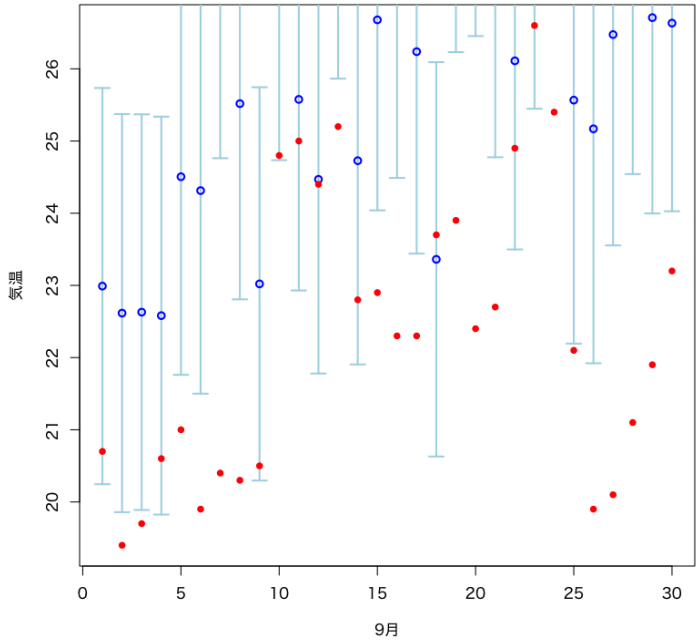

図 6: 8 月モデルによる 9 月の予測値の予測区間

- 説明変数の追加で対応可能
  - 交互作用 (交差項) : $X_i X_j$ のような説明変数の積
  - 非線形変換 : $\log(X_k)$ のような関数による変換

カテゴリカル変数を含むモデル

- 数値ではないデータ
  - 悪性良性
  - 血液型
- 適切な方法で数値に変換して対応:
  - 2 値の場合は 1,0 (真，偽) を割り当てる
    - 悪性 : 1
    - 良性 : 0
  - 3 値以上の場合は **ダミー変数** を利用する (カテゴリ数-1 個)
    - A 型 : (1,0,0)
    - B 型 : (0,1,0)
    - O 型 : (0,0,1)
    - AB 型 : (0,0,0)

解析の事例

非線形変換による線形化

- 体重と脳の重さの関係

	body	brain
Mountain beaver	1.350	8.1
Cow	465.000	423.0
Grey wolf	36.330	119.5
Goat	27.660	115.0
Guinea pig	1.040	5.5
Dipliodocus	11700.000	50.0
Asian elephant	2547.000	4603.0
Donkey	187.100	419.0
Horse	521.000	655.0
Potar monkey	10.000	115.0
Cat	3.300	25.6
Giraffe	529.000	680.0
Gorilla	207.000	406.0
Human	62.000	1320.0
African elephant	6654.000	5712.0
Triceratops	9400.000	70.0
Rhesus monkey	6.800	179.0
Kangaroo	35.000	56.0
Golden hamster	0.120	1.0
Mouse	0.023	0.4
Rabbit	2.500	12.1
Sheep	55.500	175.0
Jaguar	100.000	157.0
Chimpanzee	52.160	440.0
Rat	0.280	1.9
Brachiosaurus	87000.000	154.5

```
Mole                   0.122    3.0
Pig                  192.000  180.0
```

- 散布図 (変換なし)

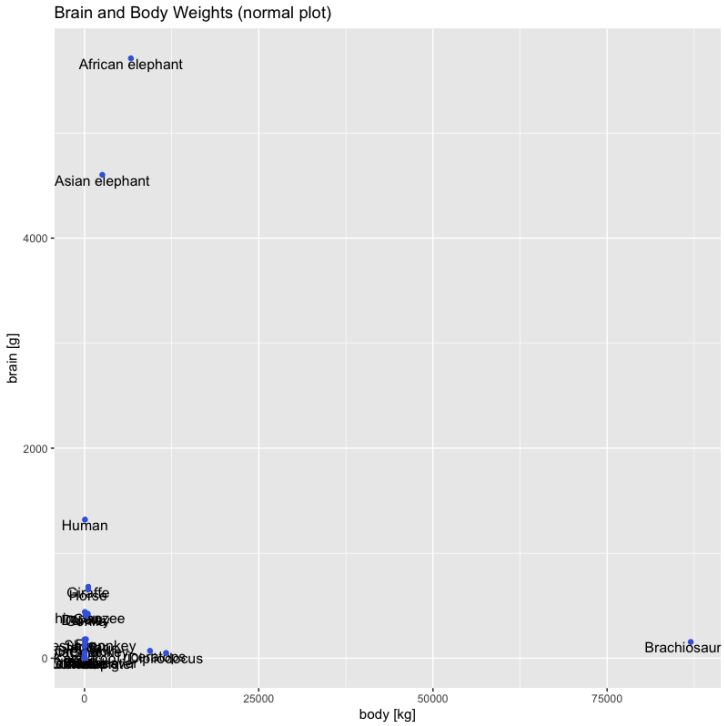

図 7: 散布図 (データの変換なし)

- 散布図 (x 軸を対数変換)

- 散布図 (xy 軸を対数変換)

非線形な関係の分析

- 東京の気候データを用いて気温に影響する変数の関係を検討する
  - 日射量と気圧の線形回帰モデル
    (日射量と気圧が気温にどのように影響するか検討する)
  - これらの交互作用を加えた線形回帰モデル
    (日射量と気圧の相互の関係の影響を検討する)

- 日射量，気圧の線形回帰モデル

```
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  145.000    63.1000    2.29 2.98e-02
solar          0.308     0.0393    7.85 1.50e-08
press         -0.121     0.0627   -1.93 6.34e-02
```

```
[1] "R2: 0.703 ; adj. R2: 0.681 ; F-stat: 33.1 ; p-val: 4.23e-08"
```

  - 係数の正負から
    * 日射が高くなるほど

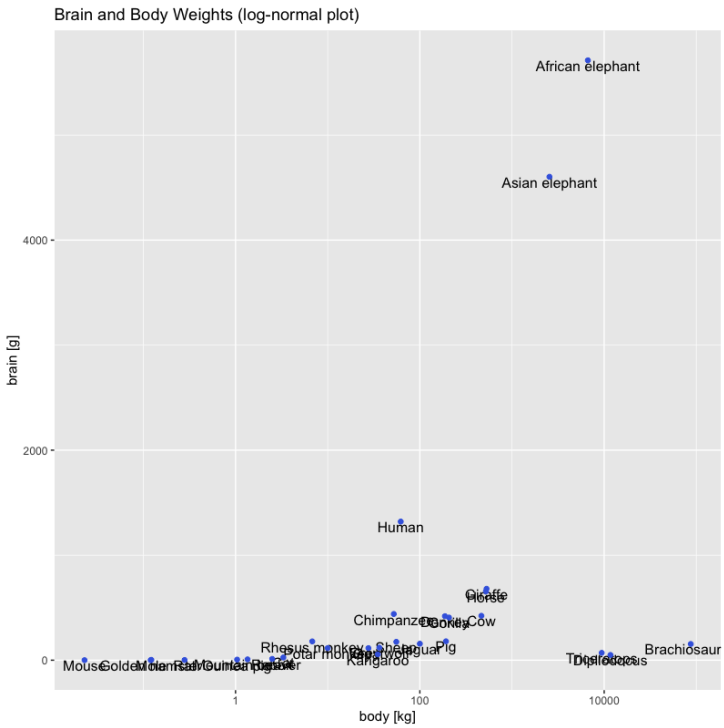

図 8: 散布図 (体重を対数変換)

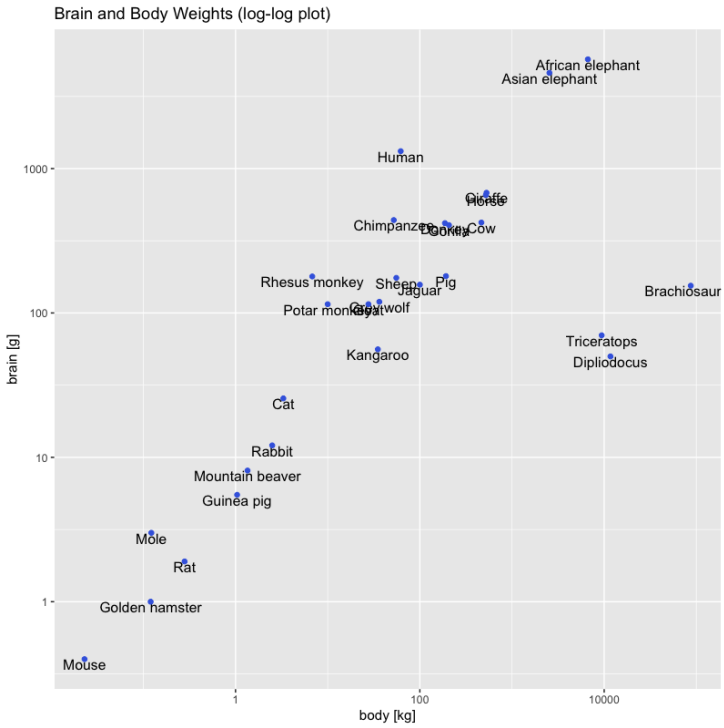

図 9: 散布図 (体重と脳の重さを対数変換)

    * 気圧が低くなるほど

  気温が高くなることが示唆される

- 交互作用を加えた線形回帰モデル

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 343.0000   1.31e+02    2.61   0.0146
solar       -12.6000   7.58e+00   -1.67   0.1070
press        -0.3180   1.30e-01   -2.44   0.0216
solar:press   0.0129   7.53e-03    1.71   0.0995

[1] "R2: 0.732 ; adj. R2: 0.702 ; F-stat: 24.5 ; p-val: 7.19e-08"
```

  - 2 次式を整理すると
    * ある気圧より低い場合には日射量が高くなるほど
    * ある日射量より低い場合には気圧が高くなるほど

    気温が高くなることが示唆される
  - 係数の有意性は低いのでより多くのデータでの分析が必要

カテゴリカル変数の利用

- 東京の気候データを用いて気温を回帰するモデルを検討する
  - 降水の有無を表すカテゴリカル変数を用いたモデル
    (雨が降ると気温が変化することを検証する)
  - 月をカテゴリカル変数として加えたモデル
    (月毎の気温の差を考慮する)

- 降水の有無を表すカテゴリカル変数を用いたモデル

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)     15.8      0.472   33.50 6.25e-113
rainTRUE         2.6      0.810    3.21  1.45e-03

[1] "R2: 0.0276 ; adj. R2: 0.0249 ; F-stat: 10.3 ; p-val: 0.00145"
```

  - 降水の有無は気温の予測に無関係ではないと考えられる
  - 決定係数から回帰式としての説明力は極めて低い

- 月をカテゴリカル変数として加えたモデル

```
    Estimate Std. Error t value  Pr(>|t|)
(Intercept)     5.57      0.463   12.00  3.59e-28
rainTRUE       -1.02      0.296   -3.44  6.45e-04
month2          2.95      0.669    4.41  1.40e-05
month3          7.55      0.654   11.50  2.28e-26
month4          9.84      0.658   14.90  1.87e-39
month5         14.60      0.662   22.00  2.80e-68
month6         17.60      0.663   26.50  7.04e-86
month7         20.80      0.656   31.70 4.89e-105
month8         22.30      0.657   33.90 5.11e-113
month9         17.20      0.662   26.00  6.79e-84
month10        13.00      0.656   19.90  2.20e-59
month11         8.32      0.657   12.70  1.62e-30
month12         2.55      0.652    3.92  1.08e-04

[1] "R2: 0.884 ; adj. R2: 0.88 ; F-stat: 224 ; p-val: 0"
```

  - 月毎に比較すると雨の日の方が気温が低いことが支持される

次回の予定

- **第 1 回: 主成分分析の考え方**
- 第 2 回: 分析の評価と視覚化