データ解析
第四回「線形回帰分析の拡張：一般化線形モデル」

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

休講情報

5/20, 6/24 は休講

今日の講義内容

- 一般化線形モデル (glm)

ガウス・マルコフモデルの限界

ガウス・マルコフモデル:

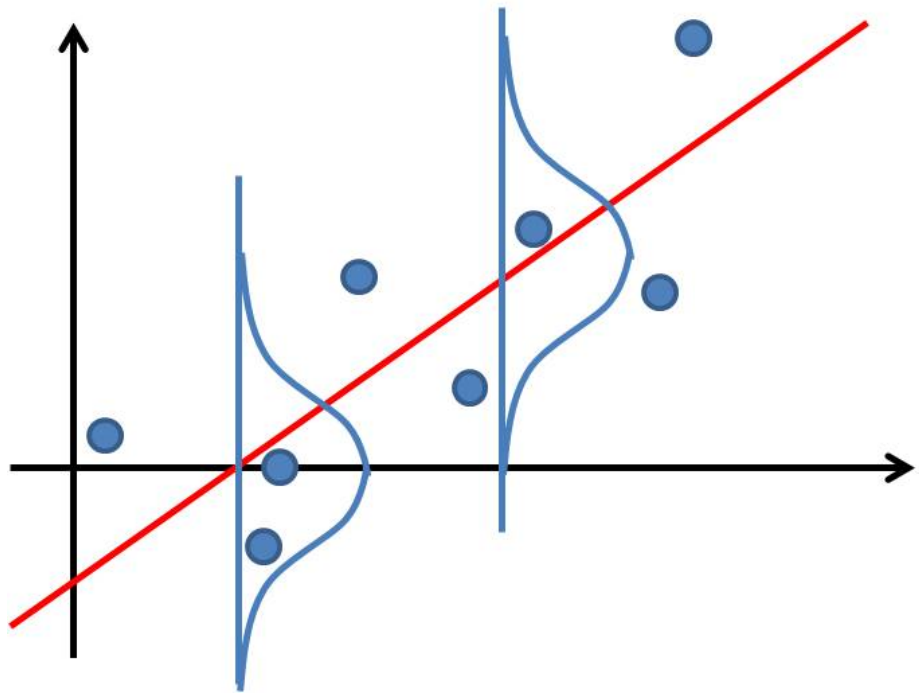
$$y = \beta^\top x + \epsilon \quad (\epsilon \sim N(0, \sigma^2))$$

- y は各 x でガウス分布.
- y の期待値は説明変数 x に対して線形.

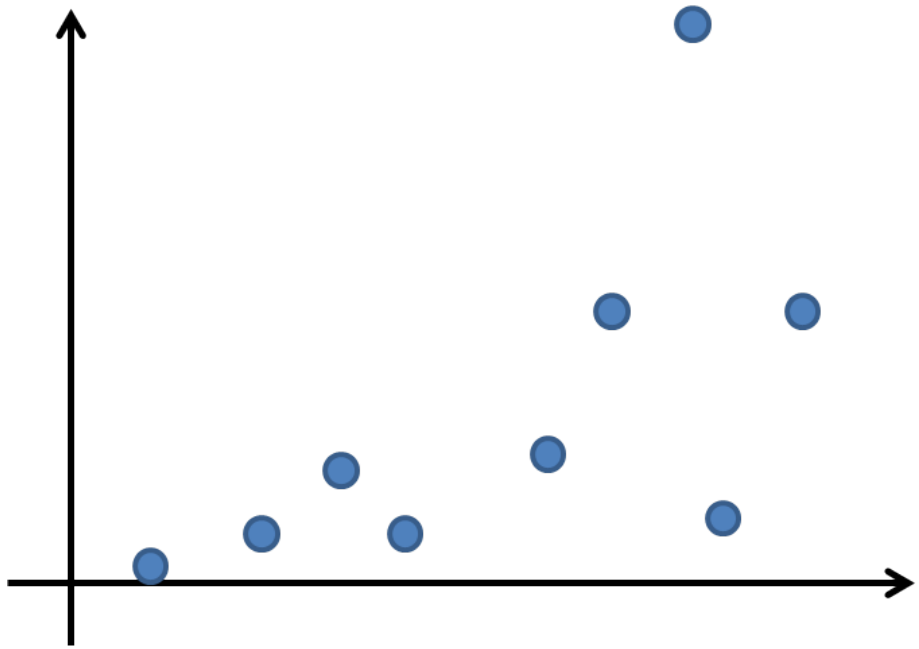
しかし，従属変数は離散値 (整数) を取ったり，非負値の制約があったり，説明変数と非線形な関係であったりする.
→ 「線形・正規分布」以上のことをしたい.

それを可能にするのが**一般化線形モデル**.
線形判別分析 ともつながる.

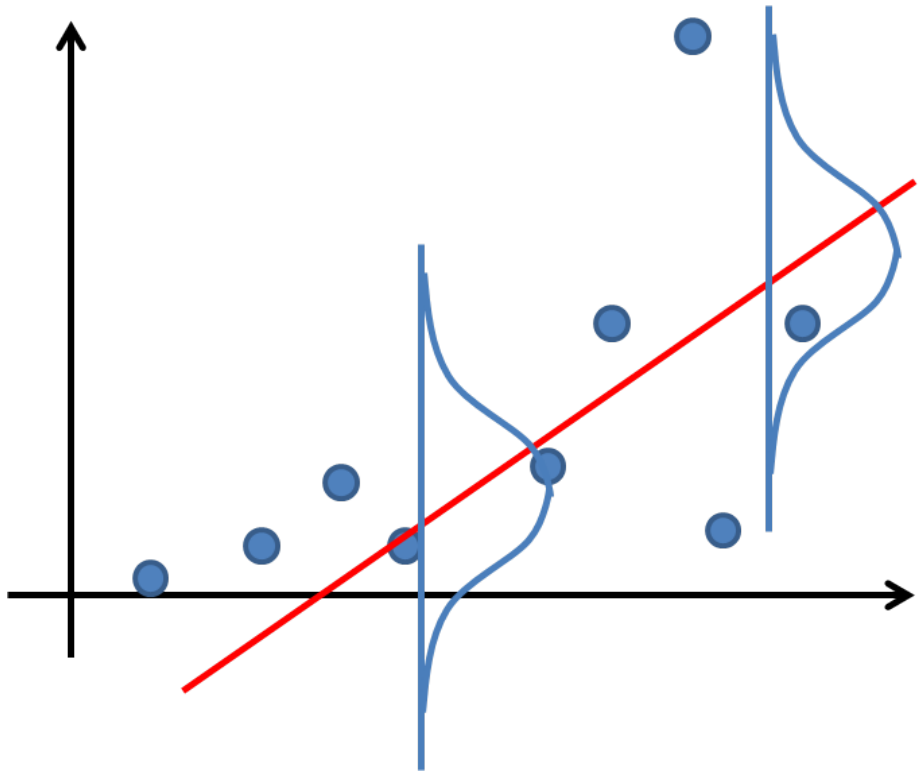
ガウスマルコフモデルのイメージ



このようなデータでは..？

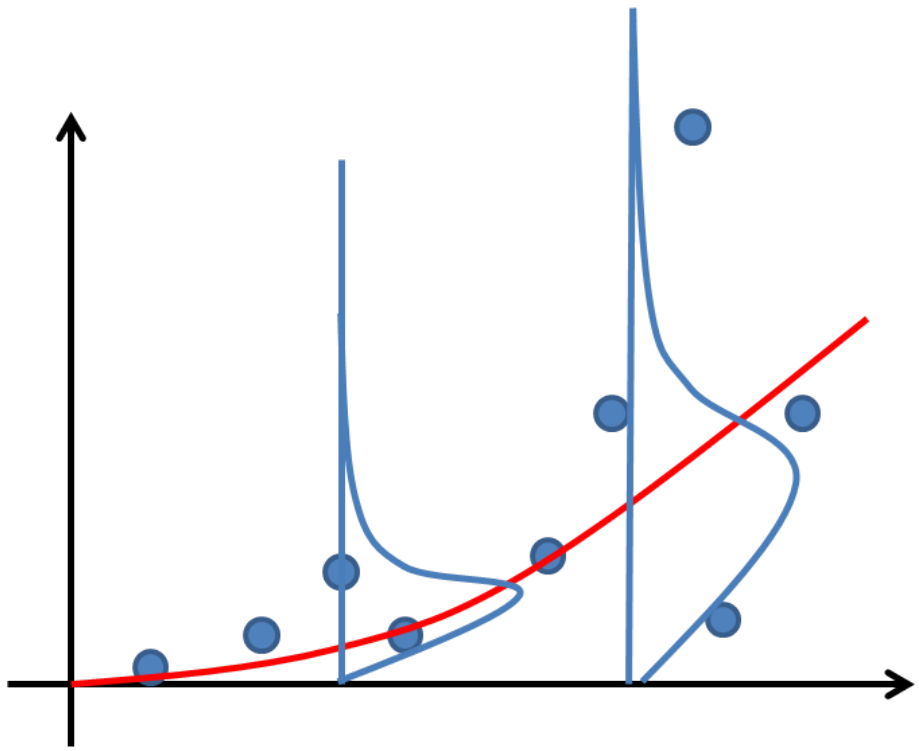


線形回帰してみる



あまり良くなさそう.

一般化線形回帰



こちらの方が良さそう.

→ **一般化線形モデル**

一般化線形モデルの基本形 (重要)

二つの構成要素

- あるパラメータ θ によって特徴づけられる分布 $P_\theta(Y)$.
- $u = \beta^\top x$ に対して Y の分布 P_θ のパラメータ θ を定める関数 $g(u)$: g^{-1} のことを**リンク関数**と呼ぶ.

一般化線形モデル：

$$y_i \sim P_{\theta=g(\beta^\top x_i)} \quad (i = 1, \ldots, n).$$

例：ガウスマルコフモデルでは $P_\theta = N(\theta, \sigma^2)$, $g(u) = u$.

$P_{\theta=g(x)}$ の密度関数を $p(y|g(\beta^\top x))$ と書くと，対数尤度は

$$\sum_{i=1}^{n} \log(p(y_i|g(\beta^\top x_i))),$$

となる.

一般化線形モデルの例 (離散)

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

$$y_i \sim \mathrm{Po}(\theta = \exp(\beta^\top x))$$

$g(\beta^\top x) = \exp(\beta^\top x)$, $g^{-1}(\theta) = \log(\theta) = \beta^\top x$: 対数リンク関数.
$\beta^\top x$ が負の値をとっても大丈夫！

一般化線形モデルの例 (離散)

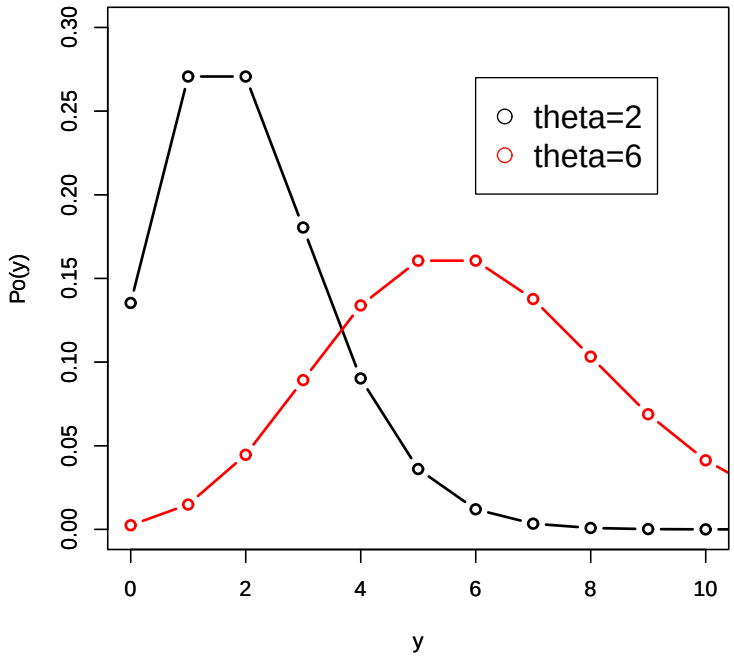
- ポアソン分布と対数リンク関数:
  $\mathrm{Po}_\theta(Y) = \frac{\theta^Y e^{-\theta}}{Y!}$ $(\theta > 0, Y = 0, 1, 2, \dots)$.

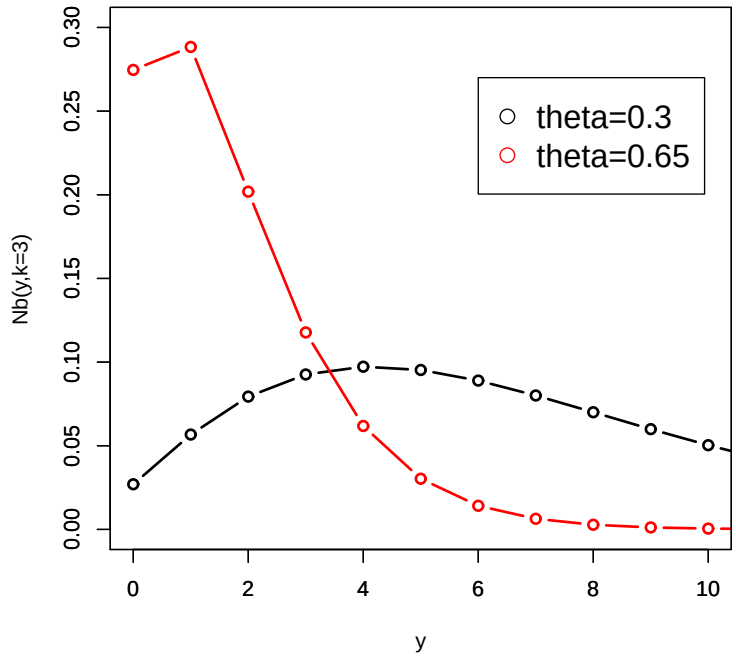
$$y_i \sim \mathrm{Po}(\theta = \exp(\beta^\top x))$$

  $g(\beta^\top x) = \exp(\beta^\top x)$, $g^{-1}(\theta) = \log(\theta) = \beta^\top x$: 対数リンク関数.
  $\beta^\top x$ が負の値をとっても大丈夫!
- 負の二項分布とロジットリンク関数:
  $\mathrm{Nb}_{\theta|k}(Y) = \binom{Y+k-1}{k-1}\theta^k(1-\theta)^Y$ $(\theta \in (0,1),\ Y = 0, 1, 2, \dots)$.

$$y_i \sim \mathrm{Nb}\left(\theta = \frac{1}{1+\exp(-\beta^\top x)} | k\right)$$

  $g(\beta^\top x) = \frac{1}{1+\exp(-\beta^\top x)}$, $\beta^\top x = g^{-1}(\theta) = \log\left(\frac{\theta}{1-\theta}\right)$: ロジット関数.
  $\beta^\top x$ が $-\infty$ から ∞ までの値をとることで, θ は $(0,1)$ 区間を動く.





ポアソン分布と負の二項分布

一般化線形モデルの例 (離散つづき)

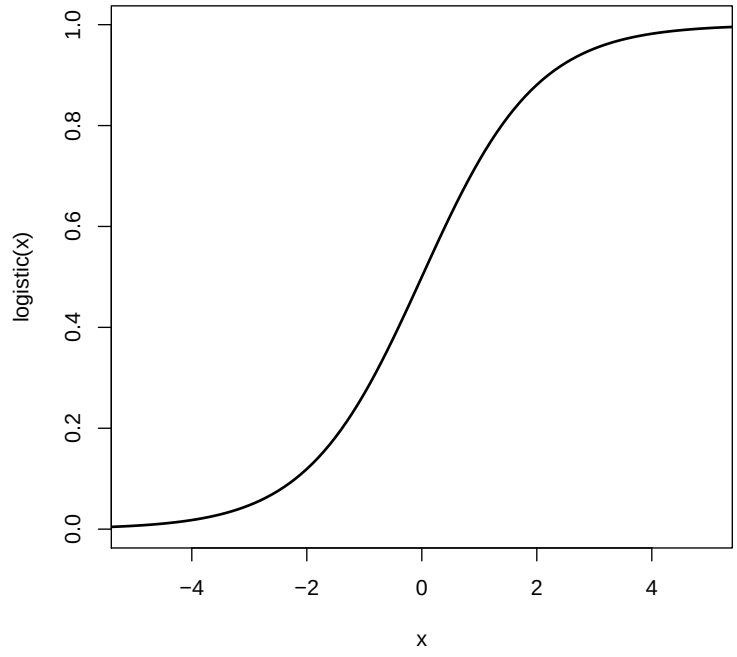
- 二項分布とロジットリンク関数:
  $\mathrm{Bin}_{\theta|N}(Y) = \binom{N}{Y}\theta^{Y}(1-\theta)^{N-Y}$ $(\theta \in (0,1),\ Y = 0,1,\ldots,N)$.

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

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

二項分布は特に $N = 2$ の場合が重要.
その場合，**ロジスティック回帰**と呼ばれる.
→ **二値判別，判別分析**.

例えば，顔認識では x が画像で，$y = 1$ のときにその画像が顔画像，$y = 0$ のときに顔画像以外というように用いる.



一般化線形モデルの例 (連続)

- ガンマ分布と逆数リンク関数:
  $\Gamma_{\theta|\alpha}(Y) = \frac{Y^{\alpha-1}e^{-Y/\theta}}{\Gamma(\alpha)\theta^{\alpha}}$ $(\theta > 0,\ Y \geq 0)$.

$$y_i \sim \mathrm{Gamma}(\theta = \frac{1}{\beta^\top x}|\alpha)$$

  $g(\beta^\top x) = \frac{1}{\beta^\top x}$, $\beta^\top x = g^{-1}(\theta) = \frac{1}{\theta}$: 逆数.
  R では形状パラメータ α も同時に推定される.

- 正規分布と恒等リンク関数:
  $N(Y|\theta = \beta^\top x, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y-\theta)^2}{2\sigma^2}\right)$ $(\theta\ Y \in \mathbb{R})$.

$$y_i \sim N(\theta = \beta^\top x, \sigma^2)$$

  $g(\beta^\top x) = \beta^\top x$, $\beta^\top x = g^{-1}(\theta) = \theta$: 恒等写像.
  これはガウスマルコフモデル.

※ここで挙げた分布とリンク関数の組は，よく使われる例として挙げただけで，この組み合わせでなくてはいけないということはない.

一般化線形モデルの推定

前のスライドに書いたように β の対数尤度は

$$\sum_{i=1}^{n} \log(p(y_i|g(\beta^\top x_i))),$$

となる．ここで，$\ell(y_i, \beta^\top x_i) := \log(p(y_i|g(\beta^\top x_i)))$ と書くと，β の最尤推定量は

$$\hat{\beta} = \arg\max_{\beta} \sum_{i=1}^{n} \ell(y_i, \beta^\top x_i),$$

となる．ℓ をロス関数と言ったりもする．最適化は汎用最適化ソルバーなどを用いる (最急降下法，準ニュートン法など).

p_θ や g を介さずに直接ロス関数 ℓ を設計することもある．

漸近正規性

ここで，もしモデルが正しかったら，

$$\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow{d} N(0, \mathrm{E}_{\beta^*}[\nabla_\beta \ell(Y, \beta^\top X) \nabla_\beta^\top \ell(Y, \beta^\top X)|_{\beta=\beta^*}]^{-1}).$$

(漸近正規性)

なお，Fisher 情報行列は

$$\begin{aligned}
&\mathrm{E}_{\beta^*}[\nabla_\beta \ell(Y, \beta^\top X) \nabla_\beta^\top \ell(Y, \beta^\top X)|_{\beta=\beta^*}] \\
\simeq &\frac{1}{n} \sum_{i=1}^{n} \mathrm{E}_{\hat{\beta}|x_i}[\nabla_\beta \ell(Y, \beta^\top x_i) \nabla_\beta^\top \ell(Y, \beta^\top x_i)|_{\beta=\hat{\beta}}]
\end{aligned}$$

で近似可能.
→ 信頼区間の構築や検定が可能に.

今回のデモで必要なライブラリ

MASS
faraway

install.packages(..)
library(..)
でインストール可能.

一般化線形モデルを R で実際に使ってみる

```
data(gala,package="faraway")
```

ガラパゴス諸島の 30 の島と亀の種類との関連
7 変数 30 サンプル
Species：その島の亀の種類の数 (従属変数)
Endemics：亀固有種の数 (説明変数)
Area：島の面積 (km2) (説明変数)
Elevation：島の標高 (m) (説明変数)
Nearest：最近隣の島との距離（km）(説明変数)
Scruz：Santa Cruz 島との距離 (km) (説明変数)
Adjacent：近隣の島のエリア（km2）(説明変数)

一般化線形モデル (glm) 基本形

一般化線形モデルの最尤推定:

```
gala.pm1<-glm(Species~ ., data = gala,
                  family = poisson(link="log"))
```

family で分布族を決定．link でリンク関数を決定.
ポアソン回帰の場合はリンク関数のデフォルトが log なので link="log" は省略可能.

```
gala.pm1<-glm(Species~ ., data = gala, family = poisson)
```

family と link 関数の例

family とデフォルトの link 関数の組:

```
binomial(link = "logit")
gaussian(link ="identity")
Gamma(link = "inverse")
inverse.gaussian(link = "1/mu^2")
poisson(link = "log")
quasi(link = "identity", variance = "constant")
quasibinomial(link = "logit")
quasipoisson(link = "log")
negative.binomial(link = "log")
```

negative.binomial は library(MASS) が必要.

結果の要約

要約の表示: summary(gala.pm1).

```
Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  2.828e+00  5.958e-02  47.471  < 2e-16 ***
Endemics     3.388e-02  1.741e-03  19.459  < 2e-16 ***
Area        -1.067e-04  3.741e-05  -2.853  0.00433 **
Elevation    2.638e-04  1.934e-04   1.364  0.17264
Nearest      1.048e-02  1.611e-03   6.502 7.91e-11 ***
Scruz       -6.835e-04  5.802e-04  -1.178  0.23877
Adjacent     4.539e-05  4.800e-05   0.946  0.34437
```

係数 β の推定値，標準偏差，Wald 統計量 (z), p-値.
ここで，係数の標準偏差の導出には最尤推定量の漸近正規性を用い，正規分布に近似して求めている (正規分布で近似してしまえば，最小二乗法と同じ理屈).
Wald 統計量は最尤推定量を (推定された) 標準偏差で割ったもの．$\beta_i^* = 0$ なる帰無仮説のもと，漸近的に正規分布に従う.
つまり，`Pr(>|z|)` が十分小さければ $\beta_i^* = 0$ なる仮説は棄却される.

結果の要約 2

```
    Null deviance: 3510.73  on 29  degrees of freedom
Residual deviance:  313.36  on 23  degrees of freedom
```

deviance (逸脱度) とは，あてはまりの「悪さ」の指標で，

$$D = -2\log(\hat{L})$$

で与えられる．すなわち，最尤推定量の対数尤度に -2 をかけたものである．

Residual deviance は D から「サンプル数分のパラメータを使ってあてはめたモデル」の逸脱度を引いたもの．
値が小さければ小さいほど手元にあるサンプルへの当てはまりが良い．モデルが正しければ漸近的に「degree of freedom」と同じ自由度の χ^2 分布に従う．
Null deviance は切片のみのモデルの Residual deviance である．

この例の場合，Residual deviance は大きく，ポアソンモデルがそこまで良いとは言えない．
→ 負の二項分布など別の分布を試してみるとどうなるか? (レポート課題)

モデル選択

Residual deviance が小さいモデルは手元にあるデータへの当てはまりは良いが, 必ずしも予測力が高いわけではない.

予測誤差が小さくなるようなモデルを選ぶ場合，AIC を用いてモデル選択を行えばよい.

glm オブジェクトに対しても step(.) や AIC(.) が使える.

第二回レポート

- gala データに，ガウスマルコフモデルおよび負の二項分布モデルを当てはめ，講義で行ったポアソンモデルとあてはまりの良さや予測誤差 (AIC) 等を比較せよ.
- solder データでも同様にガウスマルコフモデルおよびポアソンモデルと比較せよ.
- （optional）余力があれば他の分布やリンク関数を当てはめてみよ.

レポートの提出方法

- 私宛にメールにて提出.
- 件名に必ず「データ解析第 n 回レポート」と明記し，R のソースコードと結果をまとめたレポートを送付のこと.
- 氏名と学籍番号も忘れず明記すること.
- レポートは本文に載せても良いが，pdf などの電子ファイルにレポートを出力して添付ファイルとして送付することが望ましい (これを期に tex の使い方を覚えることを推奨します).
- 提出期限は講義最終回まで.

※相談はしても良いですが，コピペは厳禁です.

講義情報ページ

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

一般化線形モデル参考文献

[1] 久保拓弥: データ解析のための統計モデリング入門ー一般化線形モデル・階層ベイズモデル・MCMC. 岩波書店, 2012.

[2] J.F. Faraway: *Extending the linear model with R*. Chapman and HallCRC, 2005.