

統計モデリング入門 2016 (c)
 Poisson regression, a generalized linear model (GLM)
 一般化線形モデル: ポアソン回帰

久保拓弥 kubo@ees.hokudai.ac.jp
 北大環境科学院の講義 <http://goo.gl/76c4i>
 2016-07-13

ファイル更新時刻: 2016-07-12 14:07

kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 1 / 47

agenda
今日のハナシ I

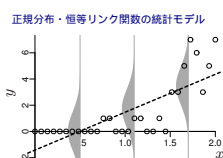
Poisson regression
 response variable explanatory variable
 応答変数 y と 説明変数 x

- ① **ポアソン回帰の統計モデル**
 植物個体の属性, あるいは実験処理が種子数に影響?
 how to specify GLM
- ③ **GLM の詳細を指定する**
 probability distribution, linear predictor and link function
 確率分布・線形予測子・リンク関数
- ④ **R で GLM のパラメーターを推定**
 あてはまりの良さは対数尤度関数で評価
- ⑤ **処理をした・しなかった 効果も統計モデルに入れる**
 factor type
 GLM の 因子型説明変数

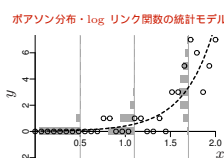
kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 2 / 47

agenda
今日のハナシ II

正規分布・恒等リンク関数の統計モデル



ポアソン分布・log リンク関数の統計モデル



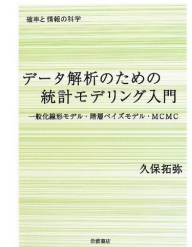
kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 3 / 47

今日の内容と「統計モデリング入門」との対応

<http://goo.gl/Ufq2>

今日はおもに「**第3章 一般化線形モデル (GLM)**」の内容を説明します。

- 著者: 久保拓弥
- 出版社: 岩波書店
- 2012-05-18 刊行



kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 4 / 47

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

- **ポアソン回帰** (Poisson regression)
- **ロジスティック回帰** (logistic regression)
- **直線回帰** (linear regression)
-

kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 5 / 47

Poisson regression
1. ポアソン回帰の統計モデル

response variable explanatory variable
 応答変数 y と 説明変数 x

一般化線形モデルにとりくんでみる

kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 6 / 47

ポアソン回帰の統計モデル 応答変数 y と 説明変数 x

statistical models appeared in the class
この授業であつかう統計モデルたち

The development of linear models

parameter estimation MCMC

Be more flexible

Generalized Linear Mixed Model (GLMM)

Incorporating random effects such as individuality

Generalized Linear Model (GLM)

MLE

Always normal distribution? That's non-sense!

Linear model

MSE

Kubo Doctrine: "Learn the evolution of linear-model family, firstly!"

kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 7 / 47

ポアソン回帰の統計モデル 応答変数 y と 説明変数 x

suppose that you have a "count data" set ...
0 個, 1 個, 2 個と数えられるデータ

カウントデータ ($y \in \{0, 1, 2, 3, \dots\}$ なデータ)

response variable y
応答変数

explanatory variable x
説明変数

- たとえば x は植物個体の大きさ, y はその個体の花数
- 体サイズが大きくなると花数が増えるように見えるが.....
- この現象を表現する統計モデルは?

kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 8 / 47

ポアソン回帰の統計モデル 応答変数 y と 説明変数 x

the normal distribution sucks!
正規分布を使った統計モデル ムリがある?

正規分布・恒等リンク関数の統計モデル

response variable y
応答変数

explanatory variable x
説明変数

NO!

とにかくセンシキがいいんでしょ
傾き「ゆーい」ならいいんでしょ
...という安易な発想のデータ解析

- タテ軸のばらつきは「正規分布」なのか?
- y の値は 0 以上なのに
- 平均値がマイナス?

kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 9 / 47

ポアソン回帰の統計モデル 応答変数 y と 説明変数 x

the Poisson distribution approximates data
ポアソン分布を使った統計モデルなら良さそう?!

ポアソン分布・対数リンク関数の統計モデル

response variable y
応答変数

explanatory variable x
説明変数

YES!

- タテ軸に対応する「ばらつき」
- 負の値にならない「平均値」
- 正規分布を使ってるモデルよりましだね

kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 10 / 47

ポアソン回帰の例題: 架空植物の種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

2. ポアソン回帰の例題: 架空植物の種子数データ

植物個体の属性, あるいは実験処理が種子数に影響?

まずはデータの概要を調べる

kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 11 / 47

ポアソン回帰の例題: 架空植物の種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

body size x and fertilization f change seed number y ?
個体サイズと実験処理の効果を調べる例題

response variable seed number $\{y_i\}$

explanatory variable body size $\{x_i\}$

explanatory variable fertilization $\{f_i\}$

個体 i

せい
施肥処理 f_i
C: 肥料なし
T: 施肥処理

種子数 y_i

体サイズ x_i
施肥処理する前に測定したもの

sample size 標本数

control

- 無処理 ($f_i = C$): 50 sample ($i \in \{1, 2, \dots, 50\}$)

treated

- 施肥処理 ($f_i = T$): 50 sample ($i \in \{51, 52, \dots, 100\}$)

kubostat2016c (<http://goo.gl/76c4i>) 統計モデリング入門 2016 (c) 2016-07-13 12 / 47

ポアソン回帰の例題: 架空植物の種子数データ 植物個体の属性, あるいは実験処理が種子数に影響?

施肥処理 f を横軸とした図

```
> plot(d$f, d$y)
```

箱ひげ図

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 19 / 47

GLM の詳細を指定する 確率分布・線形予測子・リンク関数

how to specify GLM

3. GLM の詳細を指定する

probability distribution, linear predictor and link function
確率分布・線形予測子・リンク関数

ポアソン回帰では log link 関数を使うのが便利

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 20 / 47

GLM の詳細を指定する 確率分布・線形予測子・リンク関数

how to specify GLM

一般化線形モデルを作る

Generalized Linear Model

一般化線形モデル (GLM)

probability distribution

- 確率分布は?

linear predictor

- 線形予測子は?

link function

- リンク関数は?

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 21 / 47

GLM の詳細を指定する 確率分布・線形予測子・リンク関数

how to specify linear regression model, a GLM

GLM のひとつである直線回帰モデルを指定する

直線回帰のモデル

probability distribution Gaussian distribution

- 確率分布: 正規分布

linear predictor

- 線形予測子: e.g., $\beta_1 + \beta_2 x_i$

link function identity link function

- リンク関数: 恒等リンク関数

直線の式: (切片) + (傾き) $\times x_i$

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 22 / 47

GLM の詳細を指定する 確率分布・線形予測子・リンク関数

結果 ← 原因 (かも?) を表現する線形モデル

- 結果: 応答変数
- 原因: 説明変数
- 線形予測子 (linear predictor):

$$(\text{応答変数の平均}) = \text{定数 (切片)} + (\text{係数 } 1) \times (\text{説明変数 } 1) + (\text{係数 } 2) \times (\text{説明変数 } 2) + (\text{係数 } 3) \times (\text{説明変数 } 3) + \dots$$

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 23 / 47

GLM の詳細を指定する 確率分布・線形予測子・リンク関数

how to specify Poisson regression model, a GLM

GLM のひとつであるポアソン回帰モデルを指定する

ポアソン回帰のモデル

probability distribution Poisson distribution

- 確率分布: ポアソン分布

linear predictor

- 線形予測子: e.g., $\beta_1 + \beta_2 x_i$

link function log link function

- リンク関数: 対数リンク関数

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 24 / 47

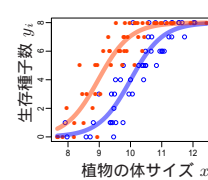
GLM の詳細を指定する 確率分布・線形予測子・リンク関数

how to specify logistic regression model, a GLM
GLM のひとつである **logistic 回帰モデル** を指定する

ロジスティック回帰のモデル

probability distribution binomial distribution
 linear predictor
 link function

- 確率分布: **二項分布**
- 線形予測子: e.g., $\beta_1 + \beta_2 x_i$
- リンク関数: **logit リンク関数**



生存種子数 y_i
植物の体サイズ x_i

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 25 / 47

GLM の詳細を指定する 確率分布・線形予測子・リンク関数

R で一般化線形モデル (GLM) の推定を.....

	probability distribution 確率分布	random number generation 乱数発生	GLM fitting GLM あてはめ
(離散)	ベルヌーイ分布	rbinom()	glm(family = binomial)
	二項分布	rbinom()	glm(family = binomial)
	ポアソン分布	rpois()	glm(family = poisson)
	負の二項分布	rnbinom()	glm.nb() in library(MASS)
(連続)	ガンマ分布	rgamma()	glm(family = gamma)
	正規分布	rnorm()	glm(family = gaussian)

- glm() で使える確率分布は上記以外にもある
- GLM は直線回帰・重回帰・分散分析・ポアソン回帰・ロジスティック回帰その他の「よせあつめ」と考えてもよいかも

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 26 / 47

GLM の詳細を指定する 確率分布・線形予測子・リンク関数

さてさて、種子数の例題にもどって

seed number y_i follows the Poisson distribution
種子数 y_i は平均 λ_i のポアソン分布にしたがうと
しましょう

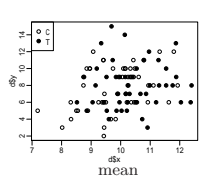
$$p(y_i | \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

個体 i の平均 λ_i を以下のようにおいてみたらどうだろう.....?

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$$

coefficient parameter
 係数 (パラメーター)

- β_1 と β_2 は 係数 (パラメーター)
- x_i は個体 i の体サイズ, f_i はとりあえず無視

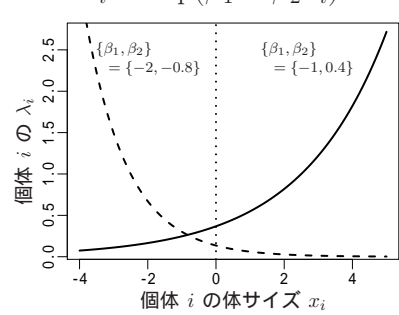


mean
body size no f_i , for simplicity

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 27 / 47

GLM の詳細を指定する 確率分布・線形予測子・リンク関数

exponential function 指数関数ってなんだっけ?

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$$


個体 i の λ_i
個体 i の体サイズ x_i

$\{\beta_1, \beta_2\} = \{-2, -0.8\}$
 $\{\beta_1, \beta_2\} = \{-1, 0.4\}$

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 28 / 47

GLM の詳細を指定する 確率分布・線形予測子・リンク関数

GLM のリンク関数と線形予測子 ← (直線の式)

mean
個体 i の平均 λ_i

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i)$$

⇕

log link function linear predictor
 $\log(\lambda_i) = \beta_1 + \beta_2 x_i$

log link function linear predictor
 $\log(\text{平均}) = \text{線形予測子}$

log リンク関数とよばれる理由は、上のようにになっているから

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 29 / 47

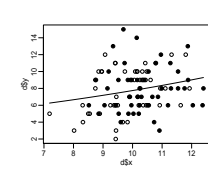
GLM の詳細を指定する 確率分布・線形予測子・リンク関数

a statistical model for this example この例題のための統計モデル

ポアソン回帰のモデル

probability distribution Poisson distribution
 linear predictor
 link function log link function

- 確率分布: **ポアソン分布**
- 線形予測子: $\beta_1 + \beta_2 x_i$
- リンク関数: **対数リンク関数**



kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 30 / 47

R で GLM のパラメーターを指定 あてはまりの良さは対数尤度関数で評価

4. R で GLM のパラメーターを推定

あてはまりの良さは対数尤度関数で評価

推定計算はコンピューターにおまかせ

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 31 / 47

R で GLM のパラメーターを指定 あてはまりの良さは対数尤度関数で評価

function
glm() 関数 の指定

```
> d
  y   x f
1  6 8.31 C
2  6 9.44 C
3  6 9.50 C
... (中略) ...
99 7 10.86 T
100 9 9.97 T
```

Is that all?
これだけ!

```
> fit <- glm(y ~ x, data = d, family = poisson)
```

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 32 / 47

R で GLM のパラメーターを指定 あてはまりの良さは対数尤度関数で評価

glm() 関数の指定の意味

```
fit <- glm(
  y ~ x,
  family = poisson(link = "log"),
  data = d
)
```

結果を格納するオブジェクト: fit
関数名: glm
モデル式: y ~ x
確率分布の指定: poisson
リンク関数の指定 (省略可): link = "log"
data.frame の指定: data = d

- モデル式 (線形予測子 z): どの説明変数を使うか?
- link 関数: z と応答変数 (y) 平均値 の関係は?
- family: どの確率分布を使うか?

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 33 / 47

R で GLM のパラメーターを指定 あてはまりの良さは対数尤度関数で評価

output
glm() 関数の 出力

```
> fit <- glm(y ~ x, data = d, family = poisson)

all: glm(formula = y ~ x, family = poisson, data = d)

Coefficients:
(Intercept)          x
      1.2917         0.0757

Degrees of Freedom: 99 Total (i.e. Null); 98 Residual
Null Deviance: 89.5
Residual Deviance: 85 AIC: 475
```

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 34 / 47

R で GLM のパラメーターを指定 あてはまりの良さは対数尤度関数で評価

glm() 関数のくわしい出力

```
> summary(fit)
Call:
glm(formula = y ~ x, family = poisson, data = d)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.368  -0.735  -0.177   0.699   2.376

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.2917    0.3637    3.55  0.00038
x              0.0757    0.0356    2.13  0.03358

..... (以下, 省略) .....
```

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 35 / 47

R で GLM のパラメーターを指定 あてはまりの良さは対数尤度関数で評価

推定値と標準誤差のいめーじ (かなりいいかげんな説明)

- 確率 p は ゼロからの距離 をあらわしている
- p がゼロに近いほど 推定値 $\hat{\beta}$ はゼロから離れている
- p が 0.5 に近いほど 推定値 $\hat{\beta}$ はゼロに近い

(注: 頻度主義的な信頼区間の正しい解釈はもっとめんどくさい)

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 36 / 47

R で GLM のパラメーターを推定 あてはまりの良さは対数尤度関数で評価

推定値と標準誤差のいめーじ (何がめんどくさいの?)

β_2 (Estimate 0.0757, SE 0.0356)

(Estimate 1.29, SE 0.364)

β_1

- 区間 95% 内に「ゼロ」があるとしよう → 「だから何？」
- 多数のパラメーターがある場合には?
- 授業の後半であつかうベイズ統計モデルでの解釈は **簡単**になるはず.....

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 37 / 47

R で GLM のパラメーターを推定 あてはまりの良さは対数尤度関数で評価

model prediction モデルの予測

```
> fit <- glm(y ~ x, data = d, family = poisson)
...
Coefficients:
(Intercept)      x
      1.2917      0.0757
```

```
> plot(d$x, d$y, pch = c(21, 19)[d$f]) # data
> xp <- seq(min(d$x), max(d$x), length = 100)
> lines(xp, exp(1.2917 + 0.0757 * xp))
```

the figure shows the relationship between model prediction and data
ここでは観測データと予測の関係を見ているだけ、なのだが

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 38 / 47

処理をした・しなかった 効果も統計モデルに入れる GLM の 因子型説明変数

5. 処理をした・しなかった 効果も統計モデルに入れる

factor type
GLM の 因子型説明変数

数量型 + 因子型 という組み合わせで

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 39 / 47

処理をした・しなかった 効果も統計モデルに入れる GLM の 因子型説明変数

Add fertilization effects 肥料の効果 f_i もいれましょう

seed number y_i follows the Poisson distribution
種子数 y_i は平均 λ_i のポアソン分布にしたがうと
しましょう

$$p(y_i | \lambda_i) = \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!}$$

個体 i の平均 λ_i を次のようにする

$$\lambda_i = \exp(\beta_1 + \beta_2 x_i + \beta_3 d_i)$$

fertilization effects coefficient
 β_3 は施肥処理の効果の係数
dummy variable
 f_i の ダミー変数

$$d_i = \begin{cases} 0 & (f_i = C \text{ の場合}) \\ 1 & (f_i = T \text{ の場合}) \end{cases}$$

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 40 / 47

処理をした・しなかった 効果も統計モデルに入れる GLM の 因子型説明変数

output glm(y ~ x + f, ...) の出力

```
> summary(glm(y ~ x + f, data = d, family = poisson))
... (略) ...
```

Coefficients:

	Estimate	Std. Error	z value	Pr(> z)
(Intercept)	1.2631	0.3696	3.42	0.00063
x	0.0801	0.0370	2.16	0.03062
fT	-0.0320	0.0744	-0.43	0.66703

..... (以下, 省略)

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 41 / 47

処理をした・しなかった 効果も統計モデルに入れる GLM の 因子型説明変数

model prediction x + f モデルの予測

```
> plot(d$x, d$y, pch = c(21, 19)[d$f]) # data
> xp <- seq(min(d$x), max(d$x), length = 100)
> lines(xp, exp(1.2631 + 0.0801 * xp), col = "blue", lwd = 3) # C
> lines(xp, exp(1.2631 + 0.0801 * xp - 0.032), col = "red", lwd = 3) # T
```

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 42 / 47

処理をした・しなかった 効果も統計モデルに入れる GLM の 因子型説明変数

multiple explanatory variables
複数の説明変数をいれた場合の統計モデル

- $f_i = C: \lambda_i = \exp(1.26 + 0.0801x_i)$
- $f_i = T: \lambda_i = \exp(1.26 + 0.0801x_i - 0.032)$
 $= \exp(1.26 + 0.0801x_i) \times \exp(-0.032)$

平均種子数 λ_i

control 無処理
fertilization 施肥処理

体サイズ x_i

施肥効果である $\exp(-0.032)$ はかけ算できくことに注意!

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 43 / 47

処理をした・しなかった 効果も統計モデルに入れる GLM の 因子型説明変数

model interpretation depends on link function
リンク関数が違うとモデルの解釈が異なる

log link function (A) 対数リンク関数
 $\lambda = \exp(\beta_1 + \beta_2x + \dots)$

identity link function (B) 恒等リンク関数
 $\lambda = \beta_1 + \beta_2x + \dots$

multiplicative 相乗的
additive 相加的

平均種子数 λ_i

体サイズ x_i

無処理 施肥処理

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 44 / 47

処理をした・しなかった 効果も統計モデルに入れる GLM の 因子型説明変数

probability distribution link function
GLM: 適切な 確率分布 とリンク関数 を選ぶ

正規分布・恒等リンク関数の統計モデル

ポアソン分布・log リンク関数の統計モデル

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 45 / 47

処理をした・しなかった 効果も統計モデルに入れる GLM の 因子型説明変数

statistical models appeared in the class
この講義であつかう統計モデルたち

線形モデルの発展

階層ベイズモデル (HBM) 推定計算方法 MCMC

もっと自由な統計モデリングを!

一般化線形混合モデル (GLMM) 最尤推定法

個体差・場所差といった変量効果をあつかいたい

一般化線形モデル (GLM) 最小二乗法

正規分布以外の確率分布をあつかいたい

線形モデル

データの特徴にあわせて線形モデルを改良・発展させる

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 46 / 47

処理をした・しなかった 効果も統計モデルに入れる GLM の 因子型説明変数

statistical models appeared in the class
この授業であつかう統計モデルたち

The development of linear models

Hierarchical Bayesian Model parameter estimation MCMC

Be more flexible

Generalized Linear Mixed Model (GLMM) MLE

Incorporating random effects such as individuality

Generalized Linear Model (GLM) MSE

Always normal distribution? That's non-sense!

Linear model

Kubo Doctrine: "Learn the evolution of linear-model family, firstly!"

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 47 / 47

処理をした・しなかった 効果も統計モデルに入れる GLM の 因子型説明変数

次回予告
The next topic

(A) $k = 1$ Too simple?

(B) $k = 7$ Too complex?

種子数 y

体サイズ x

モデル選択と統計学的検定
Model selection and statistical test

kubostat2016c (http://goo.gl/76c4i) 統計モデリング入門 2016 (c) 2016-07-13 48 / 47