broom パッケージの紹介

- 回帰結果の診断、利用に向けて -

@nonsabotage

NagoyaStat#3　LT
2016.11.26　ヤフー株式会社 名古屋オフィス

動機：モデルの比較表をつくるには？

表 6.2　種子の生存確率のモデルの AIC など．各列については表 4.2 の説明を参照．f, x, x + f モデルはそれぞれ，施肥処理 f_i だけ，体サイズ x_i だけ，そして両者に依存するモデル．

モデル	k	$\log L^*$	deviance $-2\log L^*$	residual deviance	AIC
一定	1	−321.2	642.4	499.2	644.4
f	2	−316.9	633.8	490.6	637.8
x	2	−180.2	360.3	217.2	364.3
x+f	3	−133.1	266.2	123.0	272.2
フル	100	−71.6	143.2	0.0	343.2

モデルの比較表をつくる方法

- summary 関数の出力を整形
- broom パッケージを利用

broomパッケージの
利用をオススメします

summaryの出力

```
> str (summary (fit), 1)
List of 17
 $ call          : language glm(formula = cbind(y, N - y) ~ x + f, family = binomial, data = obs)
 $ terms         :Classes 'terms', 'formula'  language cbind(y, N - y) ~ x + f
  .. ..- attr(*, "variables")= language list(cbind(y, N - y), x, f)
  .. ..- attr(*, "factors")= int [1:3, 1:2] 0 1 0 0 0 1
  .. .. ..- attr(*, "dimnames")=List of 2
  .. ..- attr(*, "term.labels")= chr [1:2] "x" "f"
  .. ..- attr(*, "order")= int [1:2] 1 1
  .. ..- attr(*, "intercept")= int 1
  .. ..- attr(*, "response")= int 1
  .. ..- attr(*, ".Environment")=<environment: R_GlobalEnv>
  .. ..- attr(*, "predvars")= language list(cbind(y, N - y), x, f)
  .. ..- attr(*, "dataClasses")= Named chr [1:3] "nmatrix.2" "numeric" "
  .. .. ..- attr(*, "names")= chr [1:3] "cbind(y, N - y)" "x" "f"
 $ family        :List of 12
  ..- attr(*, "class")= chr "family"
 $ deviance      : num 123
 $ aic           : num 272
 $ contrasts     :List of 1
 $ df.residual   : int 97
 $ null.deviance : num 499
 $ df.null       : int 99
 $ iter          : int 5
 $ deviance.resid: Named num [1:100] -1.624 0.215 -1.409 -0.831 -0.718 ...
  ..- attr(*, "names")= chr [1:100] "1" "2" "3" "4" ...
 $ coefficients  : num [1:3, 1:4] -19.536 1.952 2.022 1.414 0.139 ...
  ..- attr(*, "dimnames")=List of 2
 $ aliased       : Named logi [1:3] FALSE FALSE FALSE
  ..- attr(*, "names")= chr [1:3] "(Intercept)" "x" "fT"
 $ dispersion    : num 1
 $ df            : int [1:3] 3 97 3
 $ cov.unscaled  : num [1:3, 1:3] 1.9987 -0.1955 -0.1975 -0.1955 0.0193 ...
  ..- attr(*, "dimnames")=List of 2
 $ cov.scaled    : num [1:3, 1:3] 1.9987 -0.1955 -0.1975 -0.1955 0.0193 ...
  ..- attr(*, "dimnames")=List of 2
 - attr(*, "class")= chr "summary.glm"
```

回帰結果に関する
全ての情報がでてくる

broom::glanceの出力

```
> glance(fit)
  null.deviance df.null    logLik      AIC      BIC deviance df.residual
1      499.2321      99 -133.1056 272.2111 280.0266 123.0339          97
```

モデルに関する
情報だけがでてくる

broom::tidyの出力

```
> tidy(fit)
          term   estimate std.error  statistic      p.value
1 (Intercept) -19.536066 1.4137671 -13.818447 1.972811e-43
2           x   1.952406 0.1388724  14.058998 6.783963e-45
3          fT   2.021506 0.2312828   8.740406 2.322734e-18
```

パラメータに関する
情報だけがでてくる

broom::augmentの出力

```
> augment(fit) %>% head()
  cbind.y..N...y..y cbind.y..N...y..V2     x f    .fitted   .se.fit     .resid
1                 1                   7  9.76 C -0.4805796 0.1416902 -1.6236041
2                 6                   2 10.48 C  0.9251529 0.1401079  0.2154538
3                 5                   3 10.83 C  1.6084952 0.1632051 -1.4091200
4                 6                   2 10.94 C  1.8232599 0.1726660 -0.8312953
5                 1                   7  9.37 C -1.2420181 0.1693114 -0.7177105
6                 1                   7  8.81 C -2.3353657 0.2264088  0.3469797
        .hat   .sigma     .cooksd .std.resid
1 0.03792029 1.119402 0.030589483 -1.6552920
2 0.03192734 1.131858 0.000513808  0.2189779
3 0.02961405 1.122626 0.026158235 -1.4304600
4 0.02855187 1.128802 0.008297381 -0.8434232
5 0.03987396 1.129608 0.006513129 -0.7324621
6 0.03299174 1.131506 0.001580629  0.3528491
```

データ・診断に関する情報だけがでてくる

モデルの比較表を作るまでの例

```
> models <-
+     tibble(
+         name    = c ("fixed", "x", "f", "x+f", "x*f"),
+         formula = c (
+             "cbind(y, N-y) ~ 1",
+             "cbind(y, N-y) ~ x",
+             "cbind(y, N-y) ~ f",
+             "cbind(y, N-y) ~ x+f",
+             "cbind(y, N-y) ~ x+f+x:f"
+         )
+     ) %>%
+     group_by (name, formula) %>%
+     do (
+         fit = glm(.$formula, data=obs, family=binomial)
+     )
> models %>%
+     glance (fit)
```

```
Source: local data frame [5 x 9]
Groups: name, formula [5]

   name                 formula null.deviance df.null    logLik      AIC      BIC deviance df.residual
  <chr>                   <chr>         <dbl>   <int>     <dbl>    <dbl>    <dbl>    <dbl>       <int>
1     f       cbind(y, N-y) ~ f      499.2321      99 -316.8799 637.7598 642.9701 490.5825          98
2 fixed       cbind(y, N-y) ~ 1      499.2321      99 -321.2047 644.4093 647.0145 499.2321          99
3     x       cbind(y, N-y) ~ x      499.2321      99 -180.1727 364.3454 369.5558 217.1682          98
4   x*f cbind(y, N-y) ~ x+f+x:f      499.2321      99 -132.8053 273.6106 284.0313 122.4334          96
5   x+f     cbind(y, N-y) ~ x+f      499.2321      99 -133.1056 272.2111 280.0266 123.0339          97
```

まとめ

broomパッケージでglmの結果から
表形式で情報を抽出できる

- glance　：モデルレベルの情報
- tidy　：パラメータレベルの情報
- augment　：データレベルの情報

参考文献

- Hadley Wickham : ggplot2 Elegant Graphics for Data Analysis, Second Edition, Springer, 2016.