Linear and nonlinear regression

HUST Bioinformatics course series

Wei-Hua Chen (CC BY-NC 4.0)

09 August, 2021

section 1: TOC

前情提要

- R basics
- R data wrangler
- R plot
- R string, regular expression
- R parallel computing

本次提要

- linear regression
- nonlinear regression
- modeling and prediction
- **K**-**fold** & **X times** cross-validation
- external validation

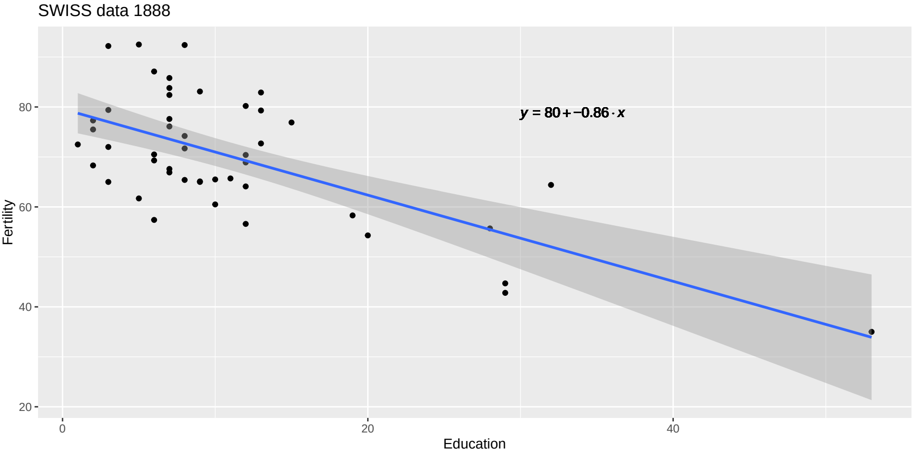
section 2: Linear regression

what is linear regression?

线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法

- Y 可以被一个变量 X 解释：一元线性回归
- Y 可以被 X, Z 等多个变量解释：multivariate linear regression

举例


SWISS data 1888
$y = 80 + -0.86 \cdot x$
Fertility
80
60
40
20
0
20
40
Education

解释

```
m <- lm(Fertility ~ Education, data = swiss);
summary(m);
```

```
##
## Call:
## lm(formula = Fertility ~ Education, data = swiss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.036  -6.711  -1.011   9.526  19.689
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  79.6101     2.1041  37.836  < 2e-16 ***
## Education    -0.8624     0.1448  -5.954 3.66e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.446 on 45 degrees of freedom
## Multiple R-squared:  0.4406, Adjusted R-squared:  0.4282
## F-statistic: 35.45 on 1 and 45 DF,  p-value: 3.659e-07
```

lm([target variable] ~ [predictor variables], data = [data source])

得到 Coefficients

```
coef( m );
```

```
## (Intercept)   Education
##  79.6100585  -0.8623503
```

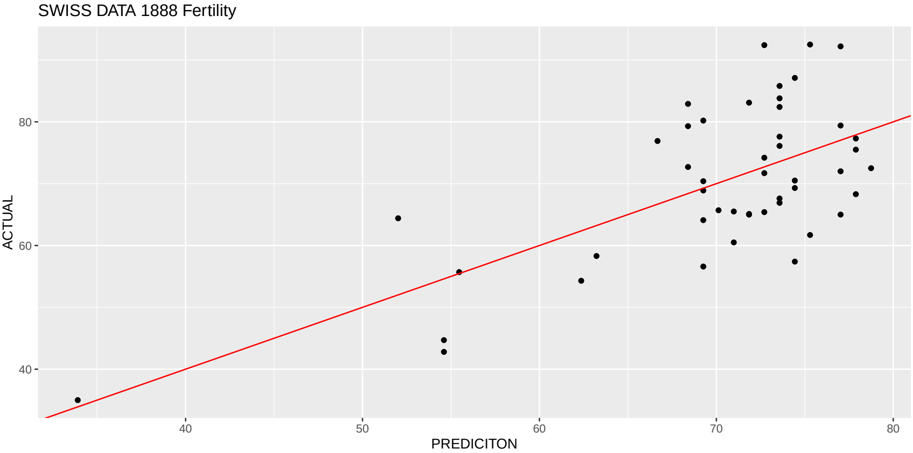
R-squred a.k.a R2 是怎么来的？

```r
library(magrittr);
library(caret);
predictions <- m %>% predict( swiss );

# Model performance
data.frame(
  RMSE = RMSE(predictions, swiss$Fertility),
  R2 = R2(predictions, swiss$Fertility)
)
```

```
##       RMSE        R2
## 1 9.242865 0.4406156
```

RMSE mean squared error, the smaller the better **R2** R, higher the better * F-statistic: Higher the better

R-squred a.k.a R2 是怎么来的？cont.



Multivariate linear modeling

datarium package

```
install.packages("datarium");
```

```
library(datarium);
head(marketing);
```

```
##   youtube facebook newspaper sales
## 1  276.12    45.36     83.04 26.52
## 2   53.40    47.16     54.12 12.48
## 3   20.64    55.08     83.16 11.16
## 4  181.80    49.56     70.20 22.20
## 5  216.96    12.96     70.08 15.48
## 6   10.44    58.68     90.00  8.64
```

问题：广告投放在哪里对销售有帮助??

multivariate linear modeling , cont.

```
m1 <- lm( sales ~ youtube + facebook + newspaper, data = marketing );
m2 <- lm( sales ~ youtube, data = marketing);
m3 <- lm( sales ~ facebook, data = marketing);

summary(m1);
```

```
##
## Call:
## lm(formula = sales ~ youtube + facebook + newspaper, data = marketing)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10.5932  -1.0690   0.2902   1.4272   3.3951
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  3.526667   0.374290   9.422   <2e-16 ***
## youtube      0.045765   0.001395  32.809   <2e-16 ***
## facebook     0.188530   0.008611  21.893   <2e-16 ***
## newspaper   -0.001037   0.005871  -0.177     0.86
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.023 on 196 degrees of freedom
## Multiple R-squared:  0.8972, Adjusted R-squared:  0.8956
## F-statistic: 570.3 on 3 and 196 DF,  p-value: < 2.2e-16
```

facebook vs. youtube

```
coef( m1 );
```

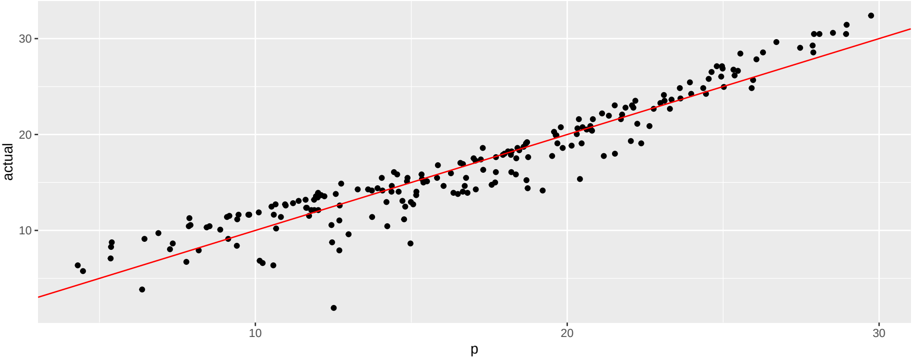
```
##  (Intercept)      youtube     facebook    newspaper
##  3.526667243  0.045764645  0.188530017 -0.001037493
```

```
data.frame( YOUTUBE = summary(m2)$r.squared, FACEBOOK = summary(m3)$r.squared);
```

```
##     YOUTUBE  FACEBOOK
## 1 0.6118751 0.3320325
```

predicted vs. actual

```
predicted.m1 <- m1 %>% predict( marketing );
data.frame(p = predicted.m1, actual = marketing$sales ) %>%
  ggplot( aes(x = p, y = actual) ) + geom_point() +
  geom_abline(intercept = 0, slope = 1, colour = "red");
```



performance evaluation

```
# Model performance
data.frame(
  RMSE = RMSE(predicted.m1, marketing$sales ),
  R2 = R2(predicted.m1, marketing$sales )
)
```

```
##       RMSE        R2
## 1 2.002284 0.8972106
```

get rid of newspaper

```
m4 <- lm( sales ~ youtube + facebook, data = marketing );
anova(m1, m4);
```

```
## Analysis of Variance Table
##
## Model 1: sales ~ youtube + facebook + newspaper
## Model 2: sales ~ youtube + facebook
##   Res.Df    RSS Df Sum of Sq      F Pr(>F)
## 1    196 801.83
## 2    197 801.96 -1  -0.12775 0.0312 0.8599
```

```
data.frame(
  with_newspapers = summary(m1)$r.squared,
  without_newspapers = summary(m4)$r.squared
)
```

```
##   with_newspapers without_newspapers
## 1       0.8972106          0.8971943
```

relative importance analysis

```
library(relaimpo);
calc.relimp( sales ~ youtube + facebook + newspaper, data = marketing );
```

```
## Response variable: sales
## Total response variance: 39.19947
## Analysis based on 200 observations
##
## 3 Regressors:
## youtube facebook newspaper
## Proportion of variance explained by model: 89.72%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##                  lmg
## youtube   0.58527298
## facebook  0.28878652
## newspaper 0.02315114
##
## Average coefficients for different model sizes:
##
##                   1X        2Xs          3Xs
## youtube   0.04753664 0.04632801  0.045764645
## facebook  0.20249578 0.19351941  0.188530017
## newspaper 0.05469310 0.02543180 -0.001037493
```

interactions

interactions 考虑因素之间的依赖关系或互作关系，比如，在一平台上投放广告会促进另一个平台上广告的效果，因为两个平台的用户可能是重叠的。他们在两个平台都看到广告时，更可能购买产品。

```
m5 <- lm( sales ~ youtube + facebook + youtube:facebook, data = marketing );
anova(m4, m5);
```

```
## Analysis of Variance Table
##
## Model 1: sales ~ youtube + facebook
## Model 2: sales ~ youtube + facebook + youtube:facebook
##   Res.Df    RSS Df Sum of Sq      F    Pr(>F)
## 1    197 801.96
## 2    196 251.26  1     550.7 429.59 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
data.frame(
  no_interactions = summary(m4)$r.squared,
  with_interactions = summary(m5)$r.squared
)
```

```
##   no_interactions with_interactions
## 1       0.8971943         0.9677905
```

interactions, cont.

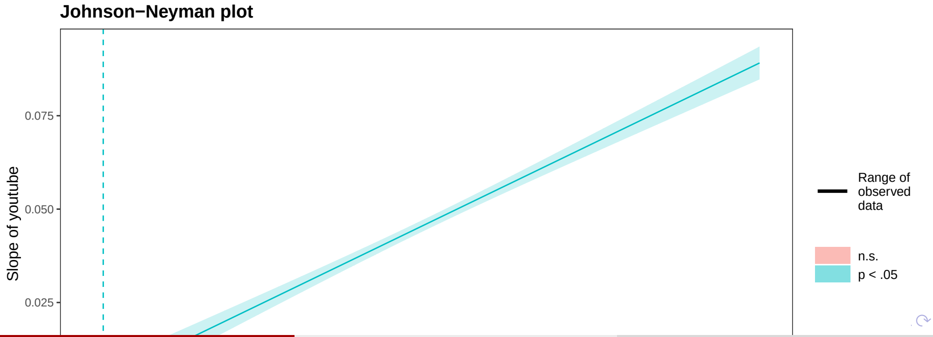
```
## m5 <- lm( sales ~ youtube + facebook + youtube:facebook, data = marketing );

## 上面的 m5 可以直接写为:
m6 <- lm( sales ~ youtube*facebook, data = marketing );
summary(m6);
```

```
##
## Call:
## lm(formula = sales ~ youtube * facebook, data = marketing)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.6039 -0.4833  0.2197  0.7137  1.8295
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)      8.100e+00  2.974e-01  27.233   <2e-16 ***
## youtube          1.910e-02  1.504e-03  12.699   <2e-16 ***
## facebook         2.886e-02  8.905e-03   3.241   0.0014 **
## youtube:facebook 9.054e-04  4.368e-05  20.727   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.132 on 196 degrees of freedom
## Multiple R-squared:  0.9678, Adjusted R-squared:  0.9673
## F-statistic:  1963 on 3 and 196 DF,  p-value: < 2.2e-16
```

visualize interactions

```
## install.packages("interactions"); 如需要，请安装这个包
library(interactions); ## 装入
sim_slopes(m6, pred = youtube, modx = facebook, jnplot = TRUE)
```

```
## JOHNSON-NEYMAN INTERVAL
##
## When facebook is OUTSIDE the interval [-26.70, -16.41], the slope of
## youtube is p < .05.
##
## Note: The range of observed values of facebook is [0.00, 59.52]
```



relative importance analysis including interactions

```
library(relaimpo);
calc.relimp( sales ~ youtube*facebook, data = marketing );
```

```
## Response variable: sales
## Total response variance: 39.19947
## Analysis based on 200 observations
##
## 3 Regressors:
## youtube facebook youtube:facebook
## Proportion of variance explained by model: 96.78%
## Metrics are not normalized (rela=FALSE).
##
## Relative importance metrics:
##
##                         lmg
## youtube          0.58851843
## facebook         0.30867583
## youtube:facebook 0.07059629
##
## Average coefficients for different model sizes:
##
##                          1X        2Xs          3Xs
## youtube          0.04753664 0.04575482 0.0191010738
## facebook         0.20249578 0.18799423 0.0288603399
## youtube:facebook        NaN        NaN 0.0009054122
```

assumptions of linear regression

1. 任何检验都有基本的假设
2. 将检验应用于不符合假设的数据是统计学最大的滥用

assumptions

1. Linearity: The relationship between X and the mean of Y is linear.
2. Homoscedasticity: The variance of residual is the same for any value of X.
3. Independence: Observations are independent of each other.
4. Normality: For any fixed value of X, Y is normally distributed.

glm vs. lm

```
lm(formula, data, …)
glm(formula, family=gaussian, data, …)
```

glm:

1. 当 `family=gaussian` 时，二者是一样的。

```
library(texreg);
```

```
## Version:  1.37.5
## Date:     2020-06-17
## Author:   Philip Leifeld (University of Essex)
##
## Consider submitting praise using the praise or praise_interactive functions.
## Please cite the JSS article in your publications -- see citation("texreg").

##
## Attaching package: 'texreg'

## The following object is masked from 'package:magrittr':
##
##     extract

## The following object is masked from 'package:tidyr':
##
##     extract
```

glm 还可用于其它类型数据的分析

1. Logistic regression (family=binomial)

预测的结果（Y）是 binary 的分类，比如 Yes, No，且只能有两个值；

```
dat <- iris %>% filter( Species %in% c("setosa", "virginica") );
bm <- glm( Species ~ Sepal.Length + Sepal.Width + Petal.Length + Petal.Width,
           data = dat, family = binomial );
```

```
## Warning: glm.fit: algorithm did not converge

## Warning: glm.fit: fitted probabilities numerically 0 or 1 occurred
```

```
data.frame( predicted = bm %>% predict( dat, type = "response" ),
            original = dat$Species ) %>% sample_n(6) %>% arrange( original );
```

```
##         predicted  original
## 9   5.204109e-12     setosa
## 45  2.126479e-11     setosa
## 37  2.220446e-16     setosa
## 30  4.976501e-12     setosa
## 72  1.000000e+00  virginica
## 54  1.000000e+00  virginica
```

注意：

glm 的 Poisson regression (family=poisson)

Poisson regression is a special type of regression in which the response variable consists of **count data**.

Asumptions:

1. The response variable consists of count data.
2. Observations are independent.
3. The mean and variance of the model are equal.
4. The distribution of counts follows a Poisson distribution.

section 3: Non-linear regression (nls)

section 5: 小结及作业！

本次小结

XXX

相关包

下次预告

作业

- Exercises and homework 目录下 talkxx-homework.Rmd 文件;
- 完成时间: 见钉群的要求

important

- all codes are available at Github: https://github.com/evolgeniusteam/R-for-bioinformatics