Multiple Regression and Logistic Regression II

Dajiang Liu

@PHS 525

Apr-19-2016

Materials from Last Time

- Multiple regression model:
  - Include multiple predictors in the model
$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \epsilon_i$$
  - How to interpret the parameter estimate:
    - β_j represent the change in Y_i per unit of change in X_{ij} given $X_{i,1}, \ldots, X_{i,j-1}, X_{i,j+1}$ unchanged.
- Measures for model fitting
  - R^2
  - R^2_{adj}

Two Types of P-values

- P-values for the assessment of model fitting
  - $H_0: \beta_1 = \cdots = \beta_J = 0$
  - $H_A: \beta_1 \neq 0$ or $\beta_2 \neq 0$ or … $\beta_J \neq 0$

- P-values for testing the statistical significance for each predictor
  - $H_0: \beta_j = 0$
  - $H_A: \beta_j \neq 0$

Questions of Interest

- Not all predictors are useful
- Including “not useful” predictors in the model will reduce the accuracy of predictors
- **Full model** is the model that contains all predictors
- Question: Determine useful predictors from the full model

Approach I

- Fit the full model that contains the full set of predictors

- Determine which predictors are important by looking at
  - P-values for testing $H_0: \beta_j = 0$
  - Predictor j is important if p-values are significant for testing H_0

Mario_Kart Example Revisited

- Fit the full model including all predictors
  - Cond
  - Wheels
  - Duration
  - Stock_photo

- Which variables are important? Why?

Approach II

- Use of goodness of fit R^2
  - Larger values of R^2 (or R^2_{adj}) indicate the model is better

- Usually more preferred than the approach for examining each p-values for each predictor

Two Model Selection Strategies I – Backward Elimination Using R^2_{adj} as a Criterion

- Backward Elimination
  - Step 1: Fit the full model
  - Step 2: Remove the predictor with the least significant p-values
  - Step 3: Compare new model and old model based upon R^2_{adj}
  - Step 4: Repeat step 2 and 3 until the values for R^2_{adj} do not change “much”

Two Model Selection Strategies II – Forward Selection

- Forward selection
  - Step 1: Fit the null model with no predictors
  - Step 2: Examine each predictor, and add the predictor with the most significant p-values
  - Step 3: Compare new model and old model based upon R^2_{adj}
  - Step 4: Add the predictor if there R^2_{adj} change significantly. If the values for R^2_{adj} do not change much with all predictors, stop

Model Selection Using Akaike Information Criterion

- With more predictors, the fitting will always be better
  - Even when the predictors are not good

- You need to penalize the number of parameter models

- Instead of directing using R^2_{adj}

- AIC is sometimes used, which equals to

$$AIC = 2k - 2\log(L)$$

Logistic regression – Motivation

- The response variable may not be normally distributed
  - E.g. the response is a categorical variable
- When response variables are binary, a new method “generalized linear model” is used
- Two step modeling:
  - Step 1: model the response as a random variable, following a distribution (say binomial or Poisson)
  - Step 2: model the parameters of the distribution as function of the predictors

Email Data Revisited

variable	description
spam	Specifies whether the message was spam.
to_multiple	An indicator variable for if more than one person was listed in the *To* field of the email.
cc	An indicator for if someone was CCed on the email.
attach	An indicator for if there was an attachment, such as a document or image.
dollar	An indicator for if the word "dollar" or dollar symbol ($) appeared in the email.
winner	An indicator for if the word "winner" appeared in the email message.
inherit	An indicator for if the word "inherit" (or a variation, like "inheritance") appeared in the email.
password	An indicator for if the word "password" was present in the email.
format	Indicates if the email contained special formatting, such as bolding, tables, or links
re_subj	Indicates whether "Re:" was included at the start of the email subject.
exclaim_subj	Indicates whether any exclamation point was included in the email subject.

Table 8.13: Descriptions for 11 variables in the `email` data set. Notice that all of the variables are indicator variables, which take the value 1 if the specified characteristic is present and 0 otherwise.

Modeling the Probability for the Response

- When the response is two-level categorical variable (e.g. Yes or No), logistic regression model can be used to model the response

- We denote Y_i as the response variable. Y_i takes two values 0 and 1.

- We denote the probability of Y_i having value of 1 as
$$p_i = \Pr(Y_i = 1).$$

- The probability for $\Pr(Y_i = 0) = 1 - p_i$.

Model the Event Probability as Functions of the Predictors

- A GLM-based multiple regression model usually takes the form

$$transform(p_i) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$$

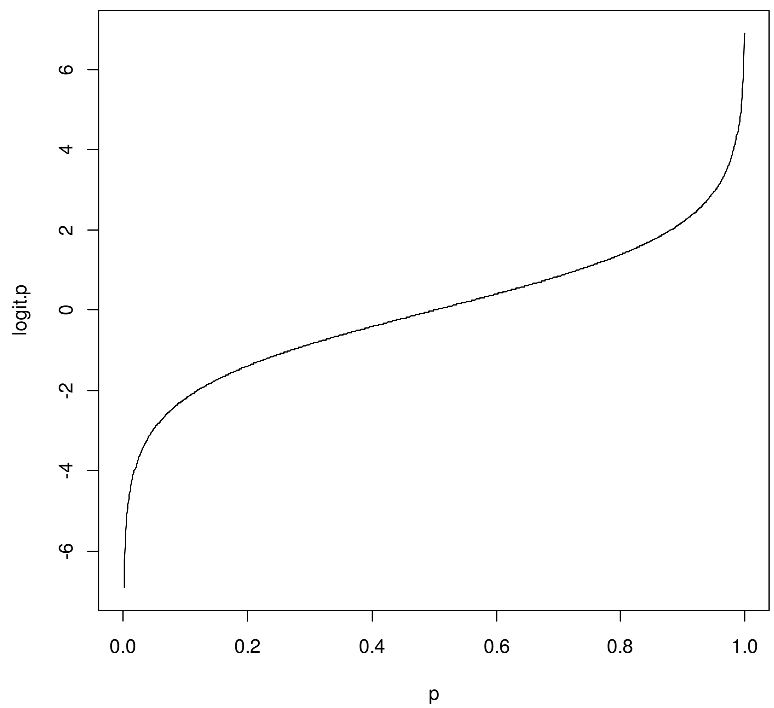
- The transformation can be the **logit** function

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1 - p_i}\right)$$

- GLMs using logit as link function is called logistic regression

$$\log\left(\frac{p_i}{1 - p_i}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_K X_K$$

What does Logistic Link Function Look Like?



The logit for a probability has range from (-Inf,Inf)

Interpret the Coefficients I

- The parameters estimated in logistic regression models can be used to estimate the probability of the response variables:

- Example: in the *Email* dataset, regressing variable $Spam$ on the variable $to\_multiple$, we obtain

$$\log\left(\frac{p_i}{1-p_i}\right) = -2.12 - 1.81 \times to\_multiple$$

- Question: What is the probability of a given email being a spam?

Interpreting the Coefficients II

- Using simple linear regression model, we have

$$\hat{p}_i = \frac{\exp(-2.12 - 1.81 \times to\_multiple)}{1 + \exp(-2.12 - 1.81 \times to\_multiple)}$$

- What is the predicted probability for an email being spam if it is sent to multiple users?

Interpreting the Coefficients III

- How to interpret the parameter estimates from logistic regression model:
- The coefficient estimates represent **log odds ratio**:

What is an odds:

$$O_1 = \Pr(Y_i = 1|X_i = 1) \,/\, \Pr(Y_i = 0|X_i = 1)$$
$$O_0 = \Pr(Y_i = 1|X_i = 0) \,/\, \Pr(Y_i = 0|X_i = 0)$$

What is an odds ratio:

$$OR = O_1/O_0$$

Odds ratio

- Using the simplest model $\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_1$
- $O_1 = \Pr(Y_i = 1|X_i = 1)/\Pr(Y_i = 0|X_i = 1) = \exp(\beta_0 + \beta_1)$
- $O_0 = \Pr(Y_i = 1|X_i = 0)/\Pr(Y_i = 0|X_i = 0) = \exp(\beta_0)$
- $OR = \frac{O_1}{O_0} = \exp(\beta_1)$
- $\log(OR) = \beta_1$

A Tabular View of Odds Ratio

- The odds ratio can be calculated by the quotient of the product of diagonal element over the product of the off-diagonal element:

	$\boldsymbol{Y = 0}$	$\boldsymbol{Y = 1}$
$X = 0$	$\Pr(Y = 0 \mid X = 0)$	$\Pr(Y = 1 \mid X = 0)$
$X = 1$	$\Pr(Y = 0 \mid X = 1)$	$\Pr(Y = 1 \mid X = 1)$

Practical Exercise:

- Email dataset revisited:
- Can you repeat the analyses regressing SPAM over to_multiple?

```
data=read.table('email.txt',header=T,sep='\t');
summary(data)
names(data)
summary(glm(spam~to_multiple,data=data,family='binomial'))
```

Any Other Variables Important to SPAM classification?

- Perform multiple logistic regression models
- Similar to multiple linear regression, multiple logistic regression models can be performed to incorporate multiple predictors

$$\log\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3}$$

- How to interpret the parameters?

Email Data: Multiple Predictors

- Include addition predictors into the model

summary(glm(spam ~ to_multiple + cc + image + attach + winner + dollar,family='binomial',data=data))

```
Call:
glm(formula = spam ~ to_multiple + cc + image + attach + winner +
    dollar, family = "binomial", data = data)

Deviance Residuals:
    Min      1Q   Median      3Q     Max
-2.4908  -0.4744  -0.4744  -0.2020   3.5959

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept) -2.12767    0.06176 -34.450  < 2e-16 ***
to_multiple -2.01934    0.30788  -6.559 5.42e-11 ***
cc           0.01770    0.02102   0.842 0.399659
image       -4.98117    2.11866  -2.351 0.018718 *
attach       0.72125    0.11335   6.363 1.98e-10 ***
winneryes    1.88412    0.29818   6.319 2.64e-10 ***
dollar      -0.07626    0.02018  -3.779 0.000157 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2437.2  on 3920  degrees of freedom
Residual deviance: 2271.5  on 3914  degrees of freedom
AIC: 2285.5

Number of Fisher Scoring iterations: 9
```