Lecture 7: Logistic Regression and Model Comparison

Will Welch (adapted from materials by Matias Salibian-Barrera)

Department of Statistics, University of British Columbia

STAT 447B: Methods for Statistical Learning

September–December 2014

Today's Learning Goals

1. Review of Logistic Regression
2. Comparing Models Via Parameter Inference
3. Comparing Models Via Analysis of Deviance
4. Comparing Models Via Misclassification
5. Cross Validation

Vaso Constriction Data

```
> help(vaso, package='robustbase')

(...)

 A data frame with 39 observations on the following 3 variables.
 'Volume' Inhaled volume of air
 'Rate' Rate of inhalation
 'Y' vector of 0 or 1 values.
```

v is Volume in the following.

Vaso Constriction: Logistic Regression

- The p_i are determined by only 2 parameters, β_0 and β_1

$$p_i = \Pr(Y_i = 1 | v = v_i) = \frac{\exp(\beta_0 + \beta_1 v_i)}{1 + \exp(\beta_0 + \beta_1 v_i)}$$

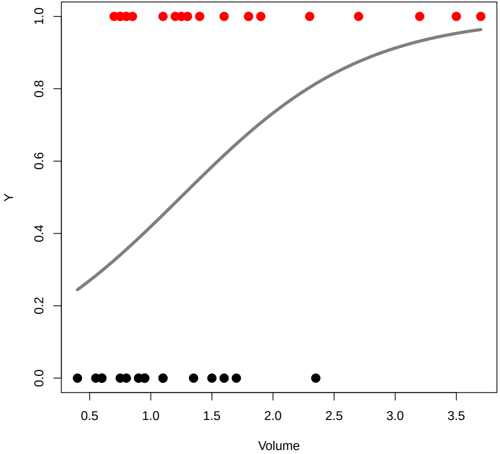
- We maximize the likelihood over β_0 and β_1

$$L(y_1, \ldots, y_n; v_1, \ldots, v_n; \beta_0, \beta_1)$$

- Given a value of Volume, the predicted probability of $Y = 1$ is

$$\frac{\exp\left(\hat{\beta}_0 + \hat{\beta}_1 v\right)}{1 + \exp\left(\hat{\beta}_0 + \hat{\beta}_1 v\right)}$$

Vaso Constriction: Logistic Regression



The curve is the prediction of $\Pr(Y = 1)$. Note the logistic shape.

Vaso Constriction: Use Volume and Rate?

- $\mathbf{x}_i = (v_i, r_i) = (\text{Volume}_i, \text{Rate}_i)$
- Linear predictor $\eta(\mathbf{x}_i) = \beta_0 + \beta_1 v_i + \beta_2 r_i$
- $p_i = \exp\left(\eta(\mathbf{x}_i)\right) / \left(1 + \exp\left(\eta(\mathbf{x}_i)\right)\right)$

```
> summary(vaso.glm)

Call:
glm(formula = Y ~ Volume + Rate, family = "binomial", data = vaso)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -9.5296     3.2332  -2.947  0.00320 **
Volume        3.8822     1.4286   2.717  0.00658 **
Rate          2.6491     0.9142   2.898  0.00376 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Should Rate be in the model? **Clicker question 1.**

Model Comparison Via Comparing Likelihoods

- Recall (STAT 305) Wilk's approximate likelihood ratio test

$$W = 2 \ln \frac{\text{likelihood of unrestricted model}}{\text{likelihood of model under } H_0} \sim \chi^2 \quad \text{under } H_0,$$

  with degrees of freedom the change in the number of parameters
- 2 (ln ratio of likelihoods) = difference in 2 (ln likelihood)
- Instead of 2 (ln likelihood), we use the deviance
- Deviance of a model:

$$D = 2 \left[\ln \left(\text{likelihood of saturated model} \right) - \ln \left(\text{likelihood of model} \right) \right]$$

- The saturated model has the best possible likelihood: for Bernoulli data the best possible estimated p_i is $\hat{p}_i = Y_i$ (perfect fit)

Analysis of Deviance

- For Bernoulli data, it can be shown (just plug into likelihood)

$$D = 2\left[\sum_{i=1} y_i \ln\left(\frac{y_i}{\hat{p}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{p}_i}\right)\right]$$

- Higher likelihood corresponds to smaller deviance
- Smaller deviance is good, like residual sum of squares

Vaso Constriction: Analysis of Deviance

```
> summary(vaso.glm)

Call:
glm(formula = Y ~ Volume, family = "binomial", data = vaso)

(...)

   Null deviance: 54.040  on 38  degrees of freedom
Residual deviance: 46.989  on 37  degrees of freedom

> summary(vaso.glm)

Call:
glm(formula = Y ~ Volume + Rate, family = "binomial", data = vaso)

(...)

    Null deviance: 54.040  on 38  degrees of freedom
Residual deviance: 29.772  on 36  degrees of freedom
```

Mini Activity

- We want to compare two models by Analysis of Deviance. The `R` model formulas (specifiying their linear predictors) are
  - `Y ~ Volume`, i.e., $\eta(\mathbf{x}_i) = \beta_0 + \beta_1 v_i$
  - `Y ~ Volume + Rate`, i.e., $\eta(\mathbf{x}_i) = \beta_0 + \beta_1 v_i + \beta_2 r_i$
- Answer the following questions
  - What is the null hypothesis, H_0?
  - What is W, i.e., the change in deviance between the two models? (Be careful about how Wilk's W is defined and its relationship to deviance and the change in deviance.)
  - To test H_0, how many degrees of freedom are used in the χ^2 distribution?
  - Is H_0 rejected?
- You have 10 minutes.

Clicker questions 2, 3, 4, and 5.

Comparison Via Misclassification Rate

- For regression, error was measured by prediction root mean squared error

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}(\mathbf{x}_i))^2}$$

- For classification, the fitted values are $\hat{p}_i$
- We can turn the $\hat{p}_i$ into hard (0/1) predictions by thresholding

$$\hat{y}_i = \begin{cases} 0 & \text{if } \hat{p}_i \leq 0.5 \\ 1 & \text{if } \hat{p}_i > 0.5 \end{cases}$$

$y_i - \hat{y}_i$ is 0 (no error) or 1 (error)

- Often this is summarized by a misclassification matrix or misclassification table

Vaso Constriction: Comparison Via Misclassification Rate

True y	Y $\sim$ Volume		Y $\sim$ Volume + Rate	
	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$
0	14	5	18	1
1	8	12	3	17
Misclass. rate	$(5 + 8)/39 = 0.33$		$(1 + 3)/39 = 0.10$	

Which model predicts the training data better? **Clicker question 6.**

Vaso Constriction: 10-Fold Cross Validation

$\hat{y}$ here is from 10-fold cross-validation

	Y $\sim$ Volume		Y $\sim$ Volume + Rate	
True y	$\hat{y} = 0$	$\hat{y} = 1$	$\hat{y} = 0$	$\hat{y} = 1$
0	13	6	14	5
1	9	11	4	16
Misclass. rate	$(6 + 9)/39 = 0.38$		$(5 + 4)/39 = 0.23$	

What is the best explanation of why the Volume + Rate model seems to predict the (cross-validation) test data better? **Clicker question 7.**

Vaso Constriction: Summary

- We compared 2 models
  - `Y ~ Volume`, i.e., $\eta(\mathbf{x}_i) = \beta_0 + \beta_1 v_i$
  - `Y ~ Volume + Rate`, i.e., $\eta(\mathbf{x}_i) = \beta_0 + \beta_1 v_i + \beta_2 r_i$
- 3 comparisons say the second model is better
  - $H_0 : \beta_2 = 0$ is rejected using a test based on approximate normality
  - A likelihood ratio test or equivalent analysis of deviance rejects $H_0 : \beta_2 = 0$
  - The model with Volume and Rate predicts better under cross validation

Digit Recognition

- Again from the UCI Machine Learning Repository
  `https://archive.ics.uci.edu/ml/machine-learning-databases/mfeat/mfeat-pix`
- 10 classes, one for each of the digits $0, \ldots, 9$
- Can turn this into a 2-class problem by considering only two digits, e.g., “8” and ”9”
- 240 explanatory variables from 15×16 averages of pixels from a grey-scale image of a handwritten digit, taking values 0–7
- Database has 200 cases for each of the 10 digits (“0” data first, then “1” data, etc.)
- We will not compare models yet, just assess the model with linear predictor using all 240 explanatory variables

$$\eta(\mathbf{x}_i) = \beta_0 + \beta_1 x_{i,1} + \cdots + \beta_{240} x_{i,240}$$

Digit Recognition: Misclassification Rate on Training Data

True y	$\hat{y} = 0$	$\hat{y} = 1$
0 (“8”)	200	0
1 (“9”)	0	200
Misclass. rate	$(0 + 0)/400 = 0$	

Perfect prediction!

Digit Recognition: Cross-Validated Misclassification Rate

$\hat{y}$ here is from 10-fold cross-validation

True y	$\hat{y} = 0$	$\hat{y} = 1$
0 ("8")	196	4
1 ("9")	4	196
Misclass. rate	$(4+4)/400 = 0.02$	

2% error rate