Multiple Logistic Regression

Section 4.3

Dr. Cathy Poliak, cpoliak@uh.edu

University of Houston

Recall Classification Problem

- The response variable, Y, is **qualitative** or **categorical**.
- Predicting a qualitative response for an observations can be referred to as **classifying** that observation.
- These methods predict the probability of each of the categories of a qualitative variables, as the basis for making the classification.

Logistic Regression

- Logistic regression can be used to model and solve problems when the Y (response) variable is a categorical variable with 2 classes.
- Also called binary classification problems.
- This models the **probability** that Y belongs to one of the two categories.

Example - Breast Cancer Database

- Using `R` in the `mlbench` package.
- The objective is to identify each cell benign or malignant classes based on some predictors.

$$Y = \begin{cases} 0 & \text{if benign} \\ 1 & \text{if malignant} \end{cases}$$

$$P(Y = 1 \mid \vec{X})$$

Predictor - Cell.size

```
fit.bc = glm(Class ~ Cell.size, family = "binomial", data = bc)
summary(fit.bc)
```

```
Call:
glm(formula = Class ~ Cell.size, family = "binomial", data = bc)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.1745     0.3879  -13.34   <2e-16 ***
Cell.size     1.5980     0.1335   11.97   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 254.76  on 681  degrees of freedom
AIC: 258.76

Number of Fisher Scoring iterations: 7
```

$$P(Y=1 \mid X) = \frac{\exp(-5.1745 + 1.5980 * \text{Cell.size})}{1 + \exp(-5.1745 + 1.5980 * \text{Cell.size})}$$

The Logistic Model

- Given $Y = 0$ or 1, let $p(X) = P(Y = 1|X)$. We want a model that shows the relationship between $p(X)$ and X.
- We use a model that gives outputs between 0 and 1 for all values of X. This is called the **logistic function**

$$p(X) = \frac{\exp^{\beta_0+\beta_1 X}}{1 + \exp^{\beta_0+\beta_1 X}}$$

- After some manipulation we get

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X$$

The left-hand side is called the *log-odds* or *logit*.

Confusion matrix

predict.bc	0	1
benign	433	37
malignant	11	202

$$\text{Accuracy} = \frac{433+202}{433+37+11+202} = 0.929$$

Multiple Logistic Regression

Predicting a binary response using multiple predictors.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X + \cdots + \beta_p X_p$$

Where $X = (X_1, \ldots, X_p)$ are p predictors. This can be rewritten as

$$p(X) = \frac{\exp^{\beta_0+\beta_1 X_1+\cdots\beta_p X_p}}{1+\exp^{\beta_0+\beta_1 X_1+\cdots\beta_p X_p}}$$

We again use the maximum likelihood method to estimate $\beta_0, \beta_1, \ldots, \beta_p$.

Three Predictors - Cl.thickness, Cell.shape and Cell.size

```
fit.bc3 = glm(Class ~ Cl.thickness + Cell.shape + Cell.size,
              family = "binomial", data = bc)
summary(fit.bc3)
```

```
Call:
glm(formula = Class ~ Cl.thickness + Cell.shape + Cell.size,
    family = "binomial", data = bc)

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)   -7.7210 = β0 0.6969 -11.079  < 2e-16 ***
Cl.thickness   0.5918 = β1 0.1030   5.746 9.14e-09 ***
Cell.shape     0.7240 = β2 0.1661   4.358 1.31e-05 ***
Cell.size      0.6390 = β3 0.1704   3.751 0.000176 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 176.50  on 679  degrees of freedom
AIC: 184.5
```

$H_0: \beta_3 = 0 \quad H_A: \beta_3 \neq 0$

$n - p - 1$

Comments

$$p(X) = \frac{\exp^{-7.721+0.592\times \text{Cl.thickness}+0.724\times \text{Cell.shape}+0.639\times \text{Cell.size}}}{1+\exp^{-7.721+0.592\times \text{Cl.thickness}+0.724\times \text{Cell.shape}+0.639\times \text{Cell.size}}}$$

- Interpret the coefficient for Cl.thickness.

As the Cl.thickness increases, the probability of being malignant increases.

- Predict the probability of malignant if Cl.thickness = 5, Cell.shape = 5, and Cell.size = 5.

$$P(Y=1 \mid \text{Cl.thickness}=5, \text{Cell.Shape}=5, \text{Cell size}=5)$$

$$= \frac{\exp[-7.721+0.592(5)+0.724(5)+0.639(5)]}{1+\exp[-7.721+0.592(5)+0.724(5)+0.639(5)]} = 0.9863$$

How Well Are We Predicting: Confusion Matrix

- Set up as follows:

		True Response	
		0	1
Predicted Response	0	true negatives	false positives
	1	false negatives	true positives

- Accuracy: Overall, how often is the classifier correct?

$$\frac{\text{true positives} + \text{true negatives}}{\text{total}}$$

- Miss-classification Rate: Overall, how often is the classifier wrong?

$$\frac{\text{false positives} + \text{false negatives}}{\text{total}}$$

- Sensitivity: When its actually positive, how often does it predict positive? Also called the true positive rate.

$$\frac{\text{true positives}}{\text{total postitives}}$$

- Specificity: When it is actually negative, how often does it predict negative? Also called true negative rate.

$$\frac{\text{true negatives}}{\text{total negatives}}$$

Getting the Predicted Responses in R

```
percent.bc = predict.glm(fit.bc, type = "response")
predict.bc = ifelse(percent.bc < 0.5, "benign","malignant")
(conf.bc = table(predict.bc,bc$Class))
```

	Actual	
predict.bc	0	1
benign	433	37
malignant	11	202

Pred

444 239 | 683

$$\text{Error rate} = \frac{37+11}{683} = 0.0703$$

$$\text{Sensitivity} = \frac{202}{239} = 0.9484$$

$$\text{specificity} = \frac{433}{444} = 0.9213$$

Confusion Matrix from Example

- Confusion Matrix from model: $p(X) = \frac{\exp^{-5.1745+1.598\times \text{Cell.size}}}{1+\exp^{-5.1745+1.598\times \text{Cell.size}}}$

		True Response	
		Benign	Malignant
Predicted Response	Benign	433	37
	Malignant	11	202

Accuracy = 0.93

- From model:

$$p(X) = \frac{\exp^{-7.721+0.592\times \text{Cl.thickness}+0.724\times \text{Cell.shape}+0.639\times \text{Cell.size}}}{1+\exp^{-7.721+0.592\times \text{Cl.thickness}+0.724\times \text{Cell.shape}+0.639\times \text{Cell.size}}}$$

		True Response	
		Benign	Malignant
Predicted Response	Benign	430	20
	Malignant	14	219
		444	239

Accuracy = 0.95

Lab Questions

1. What is the accuracy rate for the model with three predictors?

a) 0.95
c) 0.92
b) 0.05
d) 0.97

$\frac{430+219}{683} = 0.95$

2. What is the specificity rate for the model with three predictors?

a) 0.95
c) 0.92
b) 0.05
d) 0.97

$\frac{430}{444}$

Testing and Training Sets

- It is important to recall that the confusion matrix will be always biased towards unrealistic good classification rates if it is computed in the same sample used for fitting the logistic model.
- A familiar analogy is asking to your mother (data) whether you (model) are a good-looking human being (good predictive accuracy) – the answer will be highly positively biased.
- To get a fair confusion matrix, the right approach is to split randomly the sample into two: a training data set, used for fitting the model, and a test data set, used for evaluating the predictive accuracy.
- From https://bookdown.org/egarpor/SSS2-UC3M/logreg-deviance.html

Split into Test and Training

```
set.seed(100)
sample = sample.int(n = nrow(bc),
                    size = floor(.75*nrow(bc)),
                    replace = FALSE)
train.data.bc = bc[sample,]
test.data.bc = bc[-sample,]
train.bc = glm(Class ~ Cl.thickness + Cell.shape + Cell.size,
               data = train.data.bc,
               family = "binomial")
#Using the test data to determine the confusion matrix
glm.pred = predict.glm(train.bc,newdata = test.data.bc,
                       type = "response")
yHat = ifelse(glm.pred < 0.5, "benign","malignant")
table(yHat,test.data.bc$Class)
```

```
yHat           0   1
  benign     104   5
  malignant    4  58
```

108 63 171

Accuracy $\frac{104+58}{171} = 0.947$

Sensitivity $\frac{58}{63} = 0.9206$

Specificity $\frac{104}{108} = 0.963$

Deviance

- Recall coefficients in logistic regression are determined by maximizing the log-likelihood function.
- Likelihood is joint probability, between 0 and 1.
- Natural log, $ln(x)$, is negative if x is between 0 and 1.
- In in classified problems we have a quantity called *deviance*, this is defined to be -2 times the log-likelihood.

$$\text{Deviance} = -2ln(\text{likelihood})$$

- Maximizing log-likelihood is the same as minimizing the deviance.

Deviances in Logistic Regression

$L(x) = \prod_{i=1}^{n} p \prod_{i=1}^{n} (1-p)$

- **Residual** deviance is the value of deviance for the fitted model. This shows how well the response variable is predicted by a model that includes the independent variables. The residual deviance is calculated by:

LR: $RSS = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$

$$D_{resid} = -2\sum_{i=1}^{n} [y_i ln\hat{y}_i + (1-y_i) ln(1-\hat{y}_i)]$$

- **Null** deviance the value of the deviance for the model with only the intercept. This shows how well the response variable is predicted by a model that includes only the intercept. The null deviance is given by the formula:

LR: $TSS = \sum_{i=1}^{n} (y_i - \bar{y})^2$ $D_{null} = -2\sum_{i=1}^{n} [y_i ln\bar{y} + (1-y_i) ln(1-\bar{y})]$ → proportion of malignant.

Deviance Calculations From example

```
y.class = BreastCancer[complete.cases(BreastCancer), ]$Class
y.i = ifelse( y.class == "malignant", 1, 0)
## Null Deviance
-2*sum(y.i*log(mean(y.i))+(1 - y.i)*log(1-mean(y.i)))
```

```
[1] 884.3502
```

```
## Residual Deviance for Cell.size as only predictor
-2*sum(y.i*log(fit.bc$fitted.values)+(1 - y.i)*log(1-fit.bc$fitted.values))
```

```
[1] 254.7596
```

```
## Residual Deviance for Cl.thickness, Cell.shape, and Cell.size as predictors
-2*sum(y.i*log(fit.bc3$fitted.values)+(1 - y.i)*log(1-fit.bc3$fitted.values))
```

```
[1] 176.4952
```

These values are in the *summary* output for the fitted models.

Psuedo R^2

- Recall in linear regression: $R^2 = 1 - \frac{RSS}{TSS}$.
- RSS is similar to the residual deviance and TSS is similar to null deviance. Thus for logistic regression we can do:

$$R^2 = 1 - \frac{D_{resid}}{D_{null}}$$

- The larger the R^2 the better the fit.
- This can be used as an indicator for the "goodness of fit" of a model. [1]

cell size only

$$R^2 = 1 - \frac{254.7596}{894.3502} = 0.7119$$

[1] http://courses.atlas.illinois.edu/fall2016/STAT/STAT200/RProgramming/LogisticRegression.html

AIC

- Recall from linear regression: $AIC = 2(p+1) + n * ln\left(\frac{RSS}{n}\right)$
- For logistic regression: $AIC = 2(p+1) + D_{resid}$

For cell.size only

$AIC = 2(1+1) + 254.7596 = 258.7596$

Lab Questions

Predictors	Null	Residual	R^2	AIC
Cell.size	884.35	254.76	$1 - \frac{254.76}{884.35} = 0.71$	$2 * 2 + 254.76 = 258.76$
Cl.thickness+Cell.shape+Cell.size	884.35	176.50	0.8004 / 0.8832	184.5 / 119.27

Final

3. What is the R^2 for the model the three predictors?

a) 0.7119
c) 1.000
b) 0.8004
d) 0.2881

$R^2 = 1 - \frac{176.5}{884.35}$

4. What is the AIC for the model the three predictors?

a) 182.5
c) 176.5
b) 184.5
d) 892.5

$AIC = 2(3+1) + 176.5$

Using all predictors which is best?

```
fit.bc.all = glm(Class ~ ., family = "binomial", data = bc)
summary(fit.bc.all)
```

```
Call:
glm(formula = Class ~ ., family = "binomial", data = bc)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)    -10.10394    1.17488  -8.600  < 2e-16 ***
Cl.thickness     0.53501    0.14202   3.767 0.000165 ***
Cell.size       -0.00628    0.20908  -0.030 0.976039
Cell.shape       0.32271    0.23060   1.399 0.161688
Marg.adhesion    0.33064    0.12345   2.678 0.007400 **
Epith.c.size     0.09663    0.15659   0.617 0.537159
Bare.nuclei      0.38303    0.09384   4.082 4.47e-05 ***
Bl.cromatin      0.44719    0.17138   2.609 0.009073 **
Normal.nucleoli  0.21303    0.11287   1.887 0.059115 .
Mitoses          0.53484    0.32877   1.627 0.103788
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 102.89  on 673  degrees of freedom
AIC: 122.89

Number of Fisher Scoring iterations: 8
```

$AIC = 2(10) + 102.89$

$R^2 = 1 - \frac{102.89}{884.35} = 0.8837$

Best Fit?

```
fit.bc.final = glm(Class ~ Cl.thickness+Cell.shape+Marg.adhesion+Bare.nuclei+Bl.cromatin+Normal.nucleoli+Mitose
summary(fit.bc.final)
```

```
Call:
glm(formula = Class ~ Cl.thickness + Cell.shape + Marg.adhesion +
    Bare.nuclei + Bl.cromatin + Normal.nucleoli + Mitoses, family = "binomial",
    data = bc)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -9.98278    1.12610  -8.865  < 2e-16 ***
Cl.thickness     0.53400    0.14079   3.793 0.000149 ***
Cell.shape       0.34529    0.17164   2.012 0.044255 *
Marg.adhesion    0.34249    0.11922   2.873 0.004068 **
Bare.nuclei      0.38830    0.09356   4.150 3.32e-05 ***
Bl.cromatin      0.46194    0.16820   2.746 0.006025 **
Normal.nucleoli  0.22606    0.11097   2.037 0.041644 *
Mitoses          0.53119    0.32446   1.637 0.101598
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 884.35  on 682  degrees of freedom
Residual deviance: 103.27  on 675  degrees of freedom
AIC: 119.27

Number of Fisher Scoring iterations: 8
```

$= 2(8) + 103.27$

$R^2 = 1 - \frac{103.27}{884.35} = 0.8832$

Lab Questions

5. What is the R^2 for this final model?

   a) 0.7119
   c) 0.8832
   b) 0.8004
   d) 0.8837

6. Below is the confusion matrix from this model. What is the accuracy rate for this final model?

   a) 0.97
   c) 0.95
   b) 0.03
   d) 0.98

yHat	0	1
benign	434	11
malignant	10	228

$\frac{434+228}{683} = 0.969$