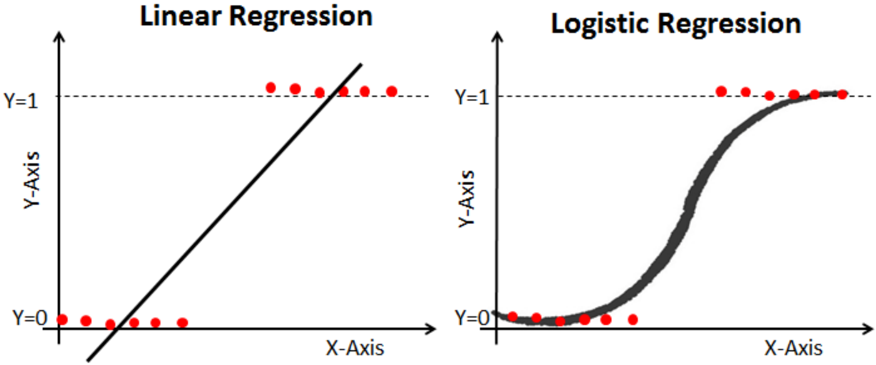
Logistic regression

Yang Wang

College of Charleston

Logistic Regression



Logistic Regression

- Response variable is categorical
- Types of logistic regression:
  - Binary logistic regression (to be studied in this course): response variable has two possible outcomes
  - Nominal logistic regression: response variable is described using 3 or more categories with no natural ordering
  - Ordinal logistic regression: response variable is described using 3 or more categories and there exists a natural ordering associated to the categories
- Challenges related to modeling categorical response:
  - nonnormal error terms
  - nonconstant error variance
  - constraints on the response function (e.g. the response variable can take values 0 and 1 only)

Logistic Regression

- Suppose π is the probability that the response variable *y* takes the value 1. π is also known as "success probability".
- Note: π ranges from 0 to 1. How to model?
- The ratio $\frac{\pi}{1-\pi}$ is called odds. This is the ratio of "success probability" and "failure probability".
- Note: if odds > 1 implies "success probability" is greater than "failure probability".
- It is easier to model $\log(\frac{\pi}{1-\pi})$ because the $\log$ function takes values from negative infinity to infinity.
- $\log(\frac{\pi}{1-\pi})$ is known as the logit transformation of the probability π

Logistic Regression

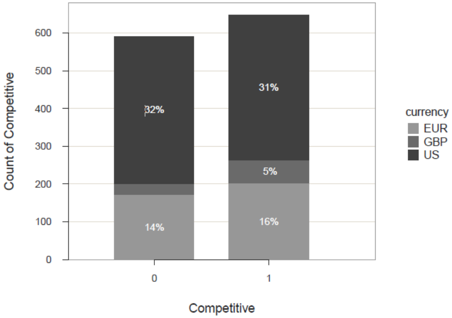
- Logistic regression model:
  $\log(\frac{\pi(y)}{1-\pi(y)}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$
- Equivalent algebraic forms:

$$\log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$
$$\frac{\pi}{1-\pi} = e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}$$
$$\pi = \frac{e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}}{1 + e^{(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}} = \frac{1}{1 + e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p)}}$$

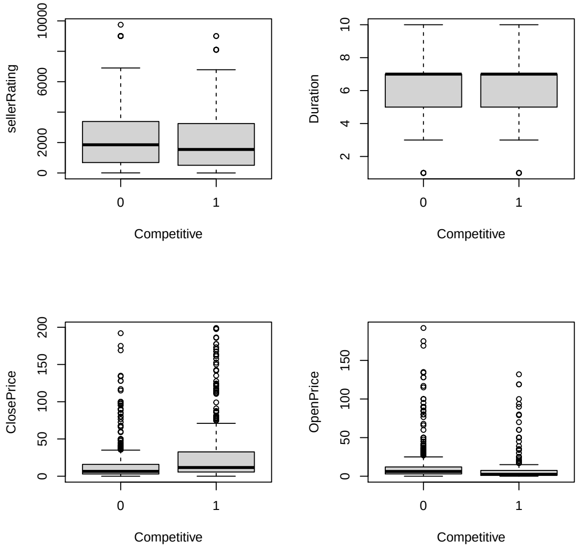
- In logistic regression the logit is a linear function of the regressor variable(s).
- If $\hat{\pi} > 0.5$ assign $\hat{y} = 1$. Otherwise $\hat{y} = 0$.

Logistic Regression

- Data: eBayAuctions
- Response variable: competitive (1 if atleast two bids are placed on the item auctioned, and 0 otherwise)
- Regressors: *currency*, *sellerRating*, *Duration*, *ClosePrice*, and *OpenPrice*
- *currency* is categorical with values *EUR*, *GBP*, and *US*



Logistic Regression



Example

```
RNGkind (sample.kind = "Rounding")
set.seed(0) ## set seed so that you get same partition each time
p2 <- partition.2(ebay, 0.7) ## creating 70:30 partition
training.data <- p2$data.train
test.data <- p2$data.test

> logistic.model <- glm(Competitive ~ ., family=binomial(link='logit'),data=training.data)
Warning message:
glm.fit: fitted probabilities numerically 0 or 1 occurred
> summary(logistic.model)

Call:
glm(formula = Competitive ~ ., family = binomial(link = "logit"),
    data = training.data)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-4.6681  -0.9454   0.0010   0.9713   2.5820

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.154e-01  3.651e-01  -1.412  0.15800
currencyGBP   1.121e+00  2.808e-01   3.993 6.51e-05 ***
currencyUS    5.809e-01  1.903e-01   3.053  0.00226 **
sellerRating -5.430e-05  3.765e-05  -1.442  0.14931
Duration     -3.141e-02  3.958e-02  -0.794  0.42741
ClosePrice    1.389e-01  1.290e-02  10.773  < 2e-16 ***
OpenPrice    -1.653e-01  1.402e-02 -11.795  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)
```

Example: Fitted model

- Fitted model: $\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = -0.5154 + 1.1213 currencyGBP + 0.5809 currencyUS - 0.0001 sellerRating - 0.0314 Duration + 0.1389 ClosePrice - 0.1653 OpenPrice$
- Fitted value of $\hat{\pi} = 1/(1 + e^{-(-0.5154 + 1.1213 currencyGBP + 0.5809 currencyUS + \cdots - 0.1653 OpenPrice)})$
- Here $\hat{\pi}$ is the estimated probability that the response variable *Competitive* takes the value 1 for given values of the predictor variables
- If $\hat{\pi} > 0.5$ assign $\hat{y} = 1$. Otherwise $\hat{y} = 0$.

Point estimate for success probability

```
### Prediction on new data ###
x0 <- data.frame(currency="US", sellerRating = 2000, Duration = 7,
                 ClosePrice = 13.01, OpenPrice = 9.99)
> predict(logistic.model, newdata = x0,type = "response")
        1
0.4732401
```

- Since the estimated success probability is less than 0.5, the estimated response is 0.

Example: Interpretation of regression coefficients

- Interpretation of $\hat{\beta}_0$:
  - Recall: $\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j x_j + \cdots + \hat{\beta}_p x_p$
  - Recall: the ratio $\frac{\pi}{1-\pi}$ is called odds
  - Let $\eta = \log(\frac{\pi}{1-\pi})$.
  - $\eta = \hat{\beta}_0$ is the log of the odds when all regressors are 0.
  - May not be meaningful when 0 is not a possible value for the regressor variables.

Example: Interpretation of regression coefficients

- Interpretation of $\hat{\beta}_j$:
  - Recall: $\log(\frac{\hat{\pi}}{1-\hat{\pi}}) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j x_j + \cdots + \hat{\beta}_p x_p$
  - $\hat{\eta}|_{x_j=c} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j c + \cdots + \hat{\beta}_p x_p$
  - $\hat{\eta}|_{x_j=c+1} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \cdots + \hat{\beta}_j (c+1) + \cdots + \hat{\beta}_p x_p$
  - Thus, $\hat{\eta}|_{x_j=c+1} - \hat{\eta}|_{x_j=c} = \hat{\beta}_j$
  - Recall: $\eta = \log(\frac{\pi}{1-\pi})$ = log(odds)
  - $\log(odds|_{x_j=c+1}) - \log(odds|_{x_j=c}) = \log\left(\frac{odds|_{x_j=c+1}}{odds|_{x_j=c}}\right) = \hat{\beta}_j$
  - From above we obtain **odds ratio** $= \hat{O}_R = \frac{odds|_{x_j=c+1}}{odds|_{x_j=c}} = e^{\hat{\beta}_j}$
  - The odds ratio can be interpreted as the estimated increase in the odds of success associated to a one-unit change in the value of the predictor variable.

Example: Interpretation of odds ratio

- Recall: $odds|_{x=x_j} = \frac{\pi}{1-\pi}|_{x=x_j}$
- This is the ratio of "success probability" and "failure probability" at a given value of the regressor variable $x = x_j$. Note that, if odds > 1 implies "success probability" is greater than "failure probability".
- Odds ratio = $\hat{O}_R = \frac{odds|_{x=x_j+1}}{odds|_{x=x_j}} = \frac{(\pi/(1-\pi))|_{x=c+1}}{(\pi/(1-\pi))|_{x=c}}$
- Odds ratio represents the change in odds for every unit increase in x_j when other predictors remain constant.

Example: Interpretation of odds ratio

- Odds ratio $= 1$ implies that a unit change in x_j does not have any effect on the odds. Thus, there is no association between y and x_j.
- Odds ratio > 1 implies that if x_j is increased by one unit then the odds of success becomes higher.
- Odds ratio < 1 implies that if x_j is increased by one unit then the odds of success gets lower.
- Odds ratio values farther from 1 represent stronger degrees of association.

Example: Interpretation of odds ratio

- Recall: odds ratio$= \hat{O}_R = \frac{odds|_{x_j=c+1}}{odds|_{x_j=c}} = e^{\hat{\beta}_j}$
- In this example, $\hat{\beta}_3 = -0.0001$
- $e^{\hat{\beta}_3} = e^{-0.0001} = 0.9999$
- Hence, for every unit increase in *sellerRating* the odds of competitiveness decreases multiplicatively by 0.9999 times when the other predictors remain constant.
- If we are interested in d unit increase in the regressor variable x, then the corresponding odds ratio can be expressed as $\hat{O}_R = \frac{odds|_{x_j=c+d}}{odds|_{x_j=c}} = e^{d\hat{\beta}_j}$
- In the context of current example, let $d = 1000$.
- $e^{d\hat{\beta}_3} = e^{1000\times-0.0001} = 0.9048$
- We can conclude, for every 1000 unit increase in *sellerRating* the odds of competitiveness decreases multiplicatively by 0.9048 times when other predictors remain constant.

Example: Interpretation of odds ratio

- For every unit increase in *Duration* the odds of competitiveness decreases multiplicatively by $e^{\hat{\beta}_4} = e^{-0.0314} = 0.9691$ times when the other predictors remain constant.
- For every unit increase in *ClosePrice* the odds of competitiveness increases multiplicatively by $e^{\hat{\beta}_5} = e^{0.1389} = 1.1491$ times when the other predictors remain constant.
- For every unit increase in *OpenPrice* the odds of competitiveness decreases multiplicatively by $e^{\hat{\beta}_6} = e^{-0.1653} = 0.8476$ times when the other predictors remain constant.

Example: Interpretation of odds ratio

- For categorical predictors with d levels, $d - 1$ dummy variables are created.
- *currency* $=$ *EUR* has been treated as baseline.
- The odds of competitiveness increases multiplicatively by $e^{\hat{\beta}_1} = e^{1.1213} = 3.0687$ times when the *currency* variable changes from *EUR* to *GBP* and the other predictors remain constant.
- The odds of competitiveness increases multiplicatively by $e^{\hat{\beta}_2} = e^{0.5809} = 1.7877$ times when the *currency* variable changes from *EUR* to *US* and the other predictors remain constant.

Testing the significance of regressors

- The parameters β_j's are estimated using maximum likelihood estimation (MLE) approach.
- According to property of MLE, β_j's are asymptotically normally distributed.
- We want to test $H_0 : \beta_j = 0$ vs. $H_1 : \beta_j \neq 0$
- Test statistic $z_0 = \frac{\hat{\beta}_j}{s.e.(\hat{\beta}_j)}$ follows normal distribution
- Reject H_0 if $|z_0| \geq z_{1-\alpha/2}$
- $100 \times (1-\alpha)\%$ confidence interval for β_1 is given by $\hat{\beta}_j \pm z_{1-\alpha/2} s.e.(\hat{\beta}_j)$

Testing the significance of regressors

```
> summary(logistic.model)
Coefficients:
               Estimate Std. Error z value Pr(>|z|)
(Intercept)  -5.154e-01  3.651e-01  -1.412  0.15800
currencyGBP   1.121e+00  2.808e-01   3.993 6.51e-05 ***
currencyUS    5.809e-01  1.903e-01   3.053  0.00226 **
sellerRating -5.430e-05  3.765e-05  -1.442  0.14931
Duration     -3.141e-02  3.958e-02  -0.794  0.42741
ClosePrice    1.389e-01  1.290e-02  10.773  < 2e-16 ***
OpenPrice    -1.653e-01  1.402e-02 -11.795  < 2e-16 ***
---
```

- We want to test $H_0 : \beta_3 = 0$ vs. $H_1 : \beta_3 \neq 0$
- Test statistic = -1.442, p-value = 0.14931
- Decision: Fail to reject H_0 at $\alpha = 0.05$. We conclude that *sellerRating* is not a significant contributor to the model.
- We want to test $H_0 : \beta_5 = 0$ vs. $H_1 : \beta_5 \neq 0$
- Test statistic = 10.773, p-value = $< 2e - 16$
- Decision: Reject H_0 at $\alpha = 0.05$. We conclude that *ClosePrice* is a significant contributor to the model.

Confidence interval

```
> confint.default(logistic.model) ## confidence interval for regression coefficients
                   2.5 %        97.5 %
(Intercept)  -1.2309017149  2.001009e-01
currencyGBP   0.5709338851  1.671574e+00
currencyUS    0.2080199583  9.537960e-01
sellerRating -0.0001280976  1.950434e-05
Duration     -0.1089752004  4.615840e-02
ClosePrice    0.1136588041  1.642152e-01
OpenPrice    -0.1928174764 -1.378667e-01
> exp(confint.default(logistic.model)) ## confidence interval for odds ratio
                 2.5 %    97.5 %
(Intercept)  0.2920291 1.2215260
currencyGBP  1.7699192 5.3205340
currencyUS   1.2312377 2.5955435
sellerRating 0.9998719 1.0000195
Duration     0.8967527 1.0472403
ClosePrice   1.1203698 1.1784679
OpenPrice    0.8246325 0.8712148
```

- Recall: odds ratio = e^{β_j}
- Confidence interval for $\beta_3 = (-0.0001280976, 1.950434e-05)$
- Confidence interval for odds ratio associated to *sellerRating* is given by $(e^{-0.0001280976}, e^{1.950434e-05}) = (0.9998719, 1.0000195)$
- If the confidence interval for odds ratio do not include 1 that implies that the associated variable is a significant contributor to the model.

Evaluation of logistic regression model

```
library(caret)

# Confusion matrix for training data
pred.prob.train <- logistic.model$fitted.values
pred.y.train <- ifelse(pred.prob.train > 0.5, 1, 0) # using cutoff = 0.5
confusionMatrix(as.factor(pred.y.train), as.factor(training.data$Competitive),
                positive = "1")

# Confusion matrix for test data
pred.prob.test <- predict(logistic.model, newdata = test.data,type = "response")
pred.y.test <- ifelse(pred.prob.test > 0.5, 1, 0) # using cutoff = 0.5
confusionMatrix(as.factor(pred.y.test), as.factor(test.data$Competitive),
                positive = "1")
```

Fitting logistic regression using k-fold cross validation approach

```
library(caret)

training.data$Competitive <- as.factor(training.data$Competitive)
levels(training.data$Competitive) <- c("no", "yes")

## K-fold Cross Validation
# value of K equal to 10
set.seed(0)
train_control <- trainControl(method = "cv", number = 10,
                              classProbs = TRUE, summaryFunction = twoClassSummary)

train_control <- trainControl(method = "cv", number = 10)

# Fit K-fold CV model
logistic_kcv <- train(Competitive ~ ., data = training.data,
                 method = "glm", family = "binomial",
                 metric = "Kappa", trControl = train_control)
print(logistic_kcv)
logistic_kcv$finalModel

# Confusion matrix for test data
pred.prob.test <- predict(logistic_kcv, newdata = test.data, type = "prob")
pred.y.test <- ifelse(pred.prob.test[,2] > 0.5, 1, 0) # using cutoff = 0.5
confusionMatrix(as.factor(pred.y.test), as.factor(test.data$Competitive),
                positive = "1")
```