# Lab 4 Solutions

*IMC 490: Machine Learning for IMC*

*4/19/2017*

In this lab, we'll be going over:
- logistic regression
- making predictions
- logistic regression which requires data transformation (bottle return problem)

## Logistic Regression

**In this exercise, we will be using the Titanic data - a classic dataset with information on each passenger aboard the Titanic at the time of its sinking. We will build a model to predict the likelihood of survival for a passenger based on features like sex, age, class, and so on.**

*The dataset we are using is a cleaned version of the dataset used in the popular Kaggle competition, Titanic: Machine Learning from Disaster.* https://www.kaggle.com/c/titanic

```
titanic = read.csv("titanic_cleaned.csv")
str(titanic)
```

```
## 'data.frame':    1045 obs. of  6 variables:
##  $ pclass  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ survived: int  1 1 0 0 0 1 1 0 1 0 ...
##  $ name    : Factor w/ 1043 levels "Abbing, Mr. Anthony",..: 22 24 25 26 27 31 46 47 51 55 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
##  $ age     : num  29 0.92 2 30 25 48 63 39 53 71 ...
##  $ fare    : num  211 152 152 152 152 ...
```

In preparation for the Kaggle competition, we're going to get more hands on with handling the data in this lab. Download the dataset "titanic_cleaned.csv" from the Lab4 folder on Canvas, and do the following:

1. Read the data into R
2. Inspect the structure of the data
3. Get rid of the `name` column - there are too many levels to regress on, and including the names of the victims makes this exercise way too real.
4. The passenger class, `pclass`, should be a categorical feature. But since it is coded with the numbers 1-4, it is read as an integer field. Convert it to categorical.

```
# after cleaning
str(titanic)
```

```
## 'data.frame':    1045 obs. of  5 variables:
##  $ pclass  : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
##  $ survived: int  1 1 0 0 0 1 1 0 1 0 ...
##  $ sex     : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
##  $ age     : num  29 0.92 2 30 25 48 63 39 53 71 ...
##  $ fare    : num  211 152 152 152 152 ...
```

**Exercises:**

0. Generate a contingency table to analyze the effect of a passenger's gender on probability of survival. Are males or females more likely to survive? (hint: use `table()`)
1. Fit a logistic regression model to predict the probability of survival using only `sex` as the predictor and print the summary.
2. Write the regression equation.
3. Obtain the odds ratio for $\beta_{sex}$. What is the interpretation of this value?
4. Fit a logistic regression model to predict the probability of survival using all of the available predictors. How much better did the model get with the addition of predictors?
5. Predict the probability of survival for Eric, a 21-year old male riding economy class (3rd) with a $20 ticket.
6. Find the log odds value for the previous prediction. Use this to manually verify the predicted probability.

### 0. Generate a contingency table to analyze the effect of a passenger's gender on probability of survival. (hint: use `table()`)

It appears that many more females survived the titanic than males - likely due to the "women and children first" evacuation.

```r
table(titanic$sex, titanic$survived)
```

```
##         
##            0   1
##   female  96 292
##   male   522 135
```

### 1. Fit a logistic regression model to predict the probability of survival using only `sex` as the predictor and print the summary.

```r
fit1 = glm(survived ~ sex, family = "binomial", data = titanic)
summary(fit1)
```

```
## 
## Call:
## glm(formula = survived ~ sex, family = "binomial", data = titanic)
## 
## Deviance Residuals: 
##     Min       1Q   Median       3Q      Max  
## -1.6713  -0.6783  -0.6783   0.7540   1.7790  
## 
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)    
## (Intercept)   1.1124     0.1176   9.455   <2e-16 ***
## sexmale      -2.4648     0.1522 -16.195   <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: 1413.6  on 1044  degrees of freedom
## Residual deviance: 1101.5  on 1043  degrees of freedom
## AIC: 1105.5
## 
```

```
## Number of Fisher Scoring iterations: 4
```

### 2. Write out the regression equation.

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 * sex_{male}$$

### 3. Obtain the odds ratio for $\beta_{sex}$. What is the interpretation of this value?

Remember that the odds ratio is $e^{\beta}$. This odds ratio means that if a passenger is male, his odds of survival are multiplied by 8% compared to the female baseline.

```
2.71 ^ -2.46
```

```
## [1] 0.08607867
```

### 4. Fit a logistic regression model to predict the probability of survival using all of the available predictors.

```
fit2 = glm(survived ~ ., family = "binomial", data = titanic)
summary(fit2)
```

```
##
## Call:
## glm(formula = survived ~ ., family = "binomial", data = titanic)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -2.6449  -0.6976  -0.4356   0.6695   2.3948
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)  3.4839588  0.3759535   9.267  < 2e-16 ***
## pclass2     -1.2570304  0.2551755  -4.926 8.39e-07 ***
## pclass3     -2.2614127  0.2654700  -8.519  < 2e-16 ***
## sexmale     -2.4938106  0.1669963 -14.933  < 2e-16 ***
## age         -0.0342311  0.0063613  -5.381 7.40e-08 ***
## fare         0.0003398  0.0017811   0.191    0.849
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 1413.57  on 1044  degrees of freedom
## Residual deviance:  982.35  on 1039  degrees of freedom
## AIC: 994.35
##
## Number of Fisher Scoring iterations: 4
```

**5. Predict the probability of survival for Eric, a 21-year old male riding economy (3rd) class with a $20 ticket.**

```
new = data.frame(pclass = as.factor(3), sex = "male", age = 21, fare = 20)
predict(fit2, new, type = "response")
```

```
##         1
## 0.1209653
```

**6. Find the log odds value for the previous prediction. Use this to manually verify the predicted probability.**

Remove the `type = "response"` parameter from `predict()` to get the log odds. To manually verify the predicted probability, solve the log odds equation for $\pi$.

$$log(\frac{\pi}{1-\pi}) = \beta_0 + \beta_1 x_1 \ + \ ... \ = -1.983$$

```
predict(fit2, new)
```

```
##         1
## -1.983321
```

```
# solving for the regression equation
e = exp(1)
e ^ -1.9833 / (1 + e ^ -1.9833)
```

```
## [1] 0.1209675
```

## Logistic Regression (requiring data transformation)

A zoologist is researching squirrels. In a study on squirrels' favorite foods, the zoologist chooses three types of food - almonds, acorns, and cashews - and attempts to hand feed the squirrels around campus. The zoologist records data on the number of squirrels he approaches (num_approached), and the number of squirrels that take the food from him (num_fed). **Let's build a simple logistic regression to predict the probability of a squirrel taking food from the zoologist, depending on the type of the food.**

```
squirrels = data.frame(num_approached = c(12, 15, 13),
                       food = as.factor(c("almonds", "acorns", "cashews")),
                       num_fed = c(2, 13, 6))
```

**Exercises:**

1. Transform the data into a form suitable for logistic regression.
2. Fit the regression and print the summary.
3. Looking at the regression coefficients, which food appears to be the squirrels' favorite? Is this result consistent with your intuitive conclusion from the collected data?

### 1. Transform the data into a form suitable for logistic regression.

```
squirrels_transformed =
  data.frame(food = rep(squirrels$food, 2),
             count = c(squirrels$num_fed, squirrels$num_approached - squirrels$num_fed),
             y = c(rep(1, 3), rep(0, 3))
             )

squirrels
```

```
##   num_approached    food num_fed
## 1             12 almonds       2
## 2             15  acorns      13
## 3             13 cashews       6
```

```
squirrels_transformed
```

```
##      food count y
## 1 almonds     2 1
## 2  acorns    13 1
## 3 cashews     6 1
## 4 almonds    10 0
## 5  acorns     2 0
## 6 cashews     7 0
```

**2. Fit the regression and print the summary.**

```
fit = glm(y ~ food, family = "binomial", weights = count, data = squirrels_transformed)
summary(fit)
```

```
##
## Call:
## glm(formula = y ~ food, family = "binomial", data = squirrels_transformed,
##     weights = count)
##
## Deviance Residuals:
##      1       2       3       4       5       6
##  2.677   1.929   3.046  -1.910  -2.839  -2.944
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)   1.8718     0.7595   2.465  0.01372 *
## foodalmonds  -3.4812     1.0848  -3.209  0.00133 **
## foodcashews  -2.0260     0.9415  -2.152  0.03140 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 55.352  on 5  degrees of freedom
## Residual deviance: 40.539  on 3  degrees of freedom
## AIC: 46.539
##
## Number of Fisher Scoring iterations: 4
```

**3. Looking at the regression coefficients, which food appears to be the squirrels' favorite? Is this result consistent with your intuitive conclusion from the collected data?**

Since the coefficients for almonds and cashews are both negative, the squirrels' favorite food appears to be the baseline response for the dummy variables - acorns. This is consistent because in the original data, the most squirrels accepted acorns (13 out of 15) compared to almonds (2 out of 12) and cashews (6 out of 13).