Lab 3: Logistic Regression Models

Sarah Wright (worked with Tyusha on

January 25 2022

Introduction

In the previous lab, we used a linear model to predict the value of a numerical response variable. However, often, we will want to predict the value of a *categorical* (or *qualitative*) response variable Y. Predicting the category of a response variable is a process known as *classification*.

For our first classification method, we will use a type of *generalized linear model* called a *logistic regression model*. If we have a *binomial random variable*, a random variable with just two possible outcomes (0 or 1), logistic regression gives us the probability that each outcome occurs based on some predictor variables X. Specifically, the form of the simple logistic regression equation with only one predictor variable is

$$P(Y=1|X) = \frac{e^{\beta_0+\beta_1 X}}{1+e^{\beta_0+\beta_1 X}}$$

In other words, this function gives us the probability that the outcome variable Y belongs to category 1 given a particular value for the predictor variable X. Notice that the function above will always be between 0 and 1 for any values of β_0, β_1, and X, which is what allows us to interpret this as a probability. Of course, the probability that the outcome variable is equal to 0 is just $1 - P(Y=1|X)$. Rearranging the formula above, we have

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \beta_0 + \beta_1 X.$$

and we see why logistic regression is considered a type of generalized *linear* regression. The quantity on the left is called the *log-odds* or *logit*, and so logistic regression models the log-odds as a linear function of the predictor variable. The coefficients are chosen via the *maximum likelihood criterion*, which you can read more about in the book in Section 4.3.2 if you would like.

Logistic Regression Lab

In this lab, we will practice applying logistic regression by working again with the weather data. As in the prior lab, we will build models on the data up through 2013, and then will evaluate the performance of those models on newer (2014-2016) data.

In contrast to our previous linear regression models, which predicted temperature (a continuous variable), we will now attempt to predict whether or not a (month's observed average high) temperature will be above normal.

1) Just like in the Multiple Linear Regression lab, create a data frame in R called `WeatherData` consisting of the data from `MonthlyWeatherData`. Then, using the `ifelse()` function, create a vector that is 1 when the observed average temperature (`WeatherData$Observed`) for the month is above normal (`WeatherData$Normal`) and 0 when it is below normal (HINT: Use `?ifelse` in the Console if you need to see how `ifelse()` works.) Add the vector you created to the `WeatherData` data frame as a column called `Binomial`.

```
WeatherData <- read.csv("/Users/sewii/Documents/CLASSES_Spring2022/Data325_AppliedDataScience/MonthlyWea
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 4.0.2

##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
Binomial <- ifelse(WeatherData$Normal <= WeatherData$Observed, 1, 0 )
WeatherData <- cbind(WeatherData, Binomial)#above normal = 1
WeatherData
```

```
##     X Month Normal First7D Observed Year Binomial
## 1   1   Jan   32.6    35.1     27.0 2009        0
## 2   2   Feb   35.7    32.9     39.1 2009        1
## 3   3   Mar   49.9    44.3     53.2 2009        1
## 4   4   Apr   61.5    54.9     61.2 2009        0
## 5   5   May   73.5    66.6     72.1 2009        0
## 6   6  June   79.8    73.0     78.6 2009        0
## 7   7  July   83.3    74.7     78.1 2009        0
## 8   8   Aug   82.0    78.6     80.1 2009        0
## 9   9   Sep   75.1    77.2     73.3 2009        0
## 10 10   Oct   62.1    61.3     58.9 2009        0
## 11 11   Nov   51.0    52.4     54.5 2009        1
## 12 12   Dec   40.4    40.6     36.7 2009        0
## 13 13   Jan   32.6    23.2     30.3 2010        0
## 14 14   Feb   35.7    32.2     31.8 2010        0
## 15 15   Mar   49.9    38.4     51.8 2010        1
## 16 16   Apr   61.5    78.5     66.7 2010        1
## 17 17   May   73.5    74.1     73.6 2010        1
## 18 18  June   79.8    77.0     80.2 2010        1
## 19 19  July   83.3    85.1     84.9 2010        1
## 20 20   Aug   82.0    83.6     84.2 2010        1
## 21 21   Sep   75.1    81.3     75.3 2010        1
## 22 22   Oct   62.1    59.5     63.8 2010        1
## 23 23   Nov   51.0    47.6     53.0 2010        1
## 24 24   Dec   40.4    29.7     30.2 2010        0
```

```
## 25 25   Jan   32.6    34.9     29.2 2011         0
## 26 26   Feb   35.7    33.0     37.6 2011         1
## 27 27   Mar   49.9    40.8     46.1 2011         0
## 28 28   Apr   61.5    52.4     60.7 2011         0
## 29 29   May   73.5    59.8     71.6 2011         0
## 30 30  June   79.8    81.5     79.3 2011         0
## 31 31  July   83.3    86.6     87.3 2011         1
## 32 32   Aug   82.0    86.4     81.5 2011         0
## 33 33   Sep   75.1    79.1     73.0 2011         0
## 34 34   Oct   62.1    65.4     61.6 2011         0
## 35 35   Nov   51.0    60.2     56.7 2011         1
## 36 36   Dec   40.4    45.9     43.6 2011         1
## 37 37   Jan   32.6    40.6     39.9 2012         1
## 38 38   Feb   35.7    44.1     42.4 2012         1
## 39 39   Mar   49.9    50.8     62.6 2012         1
## 40 40   Apr   61.5    59.0     60.2 2012         0
## 41 41   May   73.5    80.2     78.1 2012         1
## 42 42  June   79.8    71.4     81.6 2012         1
## 43 43  July   83.3    92.3     88.2 2012         1
## 44 44   Aug   82.0    88.4     83.5 2012         1
## 45 45   Sep   75.1    84.7     73.4 2012         0
## 46 46   Oct   62.1    65.4     61.5 2012         0
## 47 47   Nov   51.0    42.2     48.6 2012         0
## 48 48   Dec   40.4    52.6     43.6 2012         1
## 49 49   Jan   32.6    32.4     37.7 2013         1
## 50 50   Feb   35.7    27.7     34.7 2013         0
## 51 51   Mar   49.9    35.0     41.9 2013         0
## 52 52   Apr   61.5    51.4     61.6 2013         1
## 53 53   May   73.5    74.5     73.9 2013         1
## 54 54  June   79.8    71.6     78.5 2013         0
## 55 55  July   83.3    81.1     81.5 2013         0
## 56 56   Aug   82.0    76.8     80.2 2013         0
## 57 57   Sep   75.1    77.2     74.4 2013         0
## 58 58   Oct   62.1    76.2     63.8 2013         1
## 59 59   Nov   51.0    56.1     46.4 2013         0
## 60 60   Dec   40.4    45.3     38.8 2013         0
## 61 61   Jan   32.6    28.3     29.8 2014         0
## 62 62   Feb   35.7    29.6     31.9 2014         0
## 63 63   Mar   49.9    35.1     43.4 2014         0
## 64 64   Apr   61.5    57.9     62.2 2014         1
## 65 65   May   73.5    62.9     72.4 2014         0
## 66 66  June   79.8    78.3     80.7 2014         1
## 67 67  July   83.3    79.7     79.2 2014         0
## 68 68   Aug   82.0    79.8     80.7 2014         0
## 69 69   Sep   75.1    82.4     74.9 2014         0
## 70 70   Oct   62.1    62.9     61.6 2014         0
## 71 71   Nov   51.0    51.7     44.4 2014         0
## 72 72   Dec   40.4    41.2     40.8 2014         1
## 73 73   Jan   32.6    34.2     31.6 2015         0
## 74 74   Feb   35.7    33.7     25.9 2015         0
## 75 75   Mar   49.9    33.5     45.0 2015         0
## 76 76   Apr   61.5    59.2     60.4 2015         0
## 77 77   May   73.5    77.1     76.1 2015         1
## 78 78  June   79.8    73.3     77.9 2015         0
```

```
## 79 79  July   83.3    79.5     81.7 2015        0
## 80 80   Aug   82.0    81.1     81.0 2015        0
## 81 81   Sep   75.1    87.5     78.7 2015        1
## 82 82   Oct   62.1    65.7     63.4 2015        1
## 83 83   Nov   51.0    69.3     57.0 2015        1
## 84 84   Dec   40.4    45.8     50.8 2015        1
## 85 85   Jan   32.6    35.6     35.5 2016        1
## 86 86   Feb   35.7    49.1     42.5 2016        1
## 87 87   Mar   49.9    42.6     55.6 2016        1
## 88 88   Apr   61.5    52.4     59.4 2016        0
## 89 89   May   73.5    64.2     70.1 2016        0
## 90 90  June   79.8    80.0     81.5 2016        1
## 91 91  July   83.3    80.9     85.5 2016        1
## 92 92   Aug   82.0    88.1     86.4 2016        1
```

2) Add another column to the data set called `Deg_From_Norm` that is the number of degrees the temperature in the first seven days of the month is *above* the normal temperature for that month.

```
Deg_From_Norm <-(WeatherData$First7D - WeatherData$Normal )
WeatherData <- cbind( WeatherData, Deg_From_Norm)
WeatherData
```

```
##      X Month Normal First7D Observed Year Binomial Deg_From_Norm
## 1    1   Jan   32.6    35.1     27.0 2009        0           2.5
## 2    2   Feb   35.7    32.9     39.1 2009        1          -2.8
## 3    3   Mar   49.9    44.3     53.2 2009        1          -5.6
## 4    4   Apr   61.5    54.9     61.2 2009        0          -6.6
## 5    5   May   73.5    66.6     72.1 2009        0          -6.9
## 6    6  June   79.8    73.0     78.6 2009        0          -6.8
## 7    7  July   83.3    74.7     78.1 2009        0          -8.6
## 8    8   Aug   82.0    78.6     80.1 2009        0          -3.4
## 9    9   Sep   75.1    77.2     73.3 2009        0           2.1
## 10 10   Oct   62.1    61.3     58.9 2009        0          -0.8
## 11 11   Nov   51.0    52.4     54.5 2009        1           1.4
## 12 12   Dec   40.4    40.6     36.7 2009        0           0.2
## 13 13   Jan   32.6    23.2     30.3 2010        0          -9.4
## 14 14   Feb   35.7    32.2     31.8 2010        0          -3.5
## 15 15   Mar   49.9    38.4     51.8 2010        1         -11.5
## 16 16   Apr   61.5    78.5     66.7 2010        1          17.0
## 17 17   May   73.5    74.1     73.6 2010        1           0.6
## 18 18  June   79.8    77.0     80.2 2010        1          -2.8
## 19 19  July   83.3    85.1     84.9 2010        1           1.8
## 20 20   Aug   82.0    83.6     84.2 2010        1           1.6
## 21 21   Sep   75.1    81.3     75.3 2010        1           6.2
## 22 22   Oct   62.1    59.5     63.8 2010        1          -2.6
## 23 23   Nov   51.0    47.6     53.0 2010        1          -3.4
## 24 24   Dec   40.4    29.7     30.2 2010        0         -10.7
## 25 25   Jan   32.6    34.9     29.2 2011        0           2.3
## 26 26   Feb   35.7    33.0     37.6 2011        1          -2.7
## 27 27   Mar   49.9    40.8     46.1 2011        0          -9.1
## 28 28   Apr   61.5    52.4     60.7 2011        0          -9.1
## 29 29   May   73.5    59.8     71.6 2011        0         -13.7
## 30 30  June   79.8    81.5     79.3 2011        0           1.7
```

```
## 31 31  July  83.3    86.6     87.3 2011         1             3.3
## 32 32   Aug  82.0    86.4     81.5 2011         0             4.4
## 33 33   Sep  75.1    79.1     73.0 2011         0             4.0
## 34 34   Oct  62.1    65.4     61.6 2011         0             3.3
## 35 35   Nov  51.0    60.2     56.7 2011         1             9.2
## 36 36   Dec  40.4    45.9     43.6 2011         1             5.5
## 37 37   Jan  32.6    40.6     39.9 2012         1             8.0
## 38 38   Feb  35.7    44.1     42.4 2012         1             8.4
## 39 39   Mar  49.9    50.8     62.6 2012         1             0.9
## 40 40   Apr  61.5    59.0     60.2 2012         0            -2.5
## 41 41   May  73.5    80.2     78.1 2012         1             6.7
## 42 42  June  79.8    71.4     81.6 2012         1            -8.4
## 43 43  July  83.3    92.3     88.2 2012         1             9.0
## 44 44   Aug  82.0    88.4     83.5 2012         1             6.4
## 45 45   Sep  75.1    84.7     73.4 2012         0             9.6
## 46 46   Oct  62.1    65.4     61.5 2012         0             3.3
## 47 47   Nov  51.0    42.2     48.6 2012         0            -8.8
## 48 48   Dec  40.4    52.6     43.6 2012         1            12.2
## 49 49   Jan  32.6    32.4     37.7 2013         1            -0.2
## 50 50   Feb  35.7    27.7     34.7 2013         0            -8.0
## 51 51   Mar  49.9    35.0     41.9 2013         0           -14.9
## 52 52   Apr  61.5    51.4     61.6 2013         1           -10.1
## 53 53   May  73.5    74.5     73.9 2013         1             1.0
## 54 54  June  79.8    71.6     78.5 2013         0            -8.2
## 55 55  July  83.3    81.1     81.5 2013         0            -2.2
## 56 56   Aug  82.0    76.8     80.2 2013         0            -5.2
## 57 57   Sep  75.1    77.2     74.4 2013         0             2.1
## 58 58   Oct  62.1    76.2     63.8 2013         1            14.1
## 59 59   Nov  51.0    56.1     46.4 2013         0             5.1
## 60 60   Dec  40.4    45.3     38.8 2013         0             4.9
## 61 61   Jan  32.6    28.3     29.8 2014         0            -4.3
## 62 62   Feb  35.7    29.6     31.9 2014         0            -6.1
## 63 63   Mar  49.9    35.1     43.4 2014         0           -14.8
## 64 64   Apr  61.5    57.9     62.2 2014         1            -3.6
## 65 65   May  73.5    62.9     72.4 2014         0           -10.6
## 66 66  June  79.8    78.3     80.7 2014         1            -1.5
## 67 67  July  83.3    79.7     79.2 2014         0            -3.6
## 68 68   Aug  82.0    79.8     80.7 2014         0            -2.2
## 69 69   Sep  75.1    82.4     74.9 2014         0             7.3
## 70 70   Oct  62.1    62.9     61.6 2014         0             0.8
## 71 71   Nov  51.0    51.7     44.4 2014         0             0.7
## 72 72   Dec  40.4    41.2     40.8 2014         1             0.8
## 73 73   Jan  32.6    34.2     31.6 2015         0             1.6
## 74 74   Feb  35.7    33.7     25.9 2015         0            -2.0
## 75 75   Mar  49.9    33.5     45.0 2015         0           -16.4
## 76 76   Apr  61.5    59.2     60.4 2015         0            -2.3
## 77 77   May  73.5    77.1     76.1 2015         1             3.6
## 78 78  June  79.8    73.3     77.9 2015         0            -6.5
## 79 79  July  83.3    79.5     81.7 2015         0            -3.8
## 80 80   Aug  82.0    81.1     81.0 2015         0            -0.9
## 81 81   Sep  75.1    87.5     78.7 2015         1            12.4
## 82 82   Oct  62.1    65.7     63.4 2015         1             3.6
## 83 83   Nov  51.0    69.3     57.0 2015         1            18.3
## 84 84   Dec  40.4    45.8     50.8 2015         1             5.4
```

```
## 85 85   Jan   32.6    35.6     35.5 2016        1           3.0
## 86 86   Feb   35.7    49.1     42.5 2016        1          13.4
## 87 87   Mar   49.9    42.6     55.6 2016        1          -7.3
## 88 88   Apr   61.5    52.4     59.4 2016        0          -9.1
## 89 89   May   73.5    64.2     70.1 2016        0          -9.3
## 90 90  June   79.8    80.0     81.5 2016        1           0.2
## 91 91  July   83.3    80.9     85.5 2016        1          -2.4
## 92 92   Aug   82.0    88.1     86.4 2016        1           6.1
```

3) Split the data into `WeatherTrain` (the first 60 observations) and `WeatherTest` (the last 32 observations).

```
WeatherTrain <- slice(WeatherData, n = 1:60)
WeatherTest <- slice_tail(WeatherData, n = 32 )
WeatherTrain
```

```
##      X Month Normal First7D Observed Year Binomial Deg_From_Norm
## 1    1   Jan   32.6    35.1     27.0 2009        0           2.5
## 2    2   Feb   35.7    32.9     39.1 2009        1          -2.8
## 3    3   Mar   49.9    44.3     53.2 2009        1          -5.6
## 4    4   Apr   61.5    54.9     61.2 2009        0          -6.6
## 5    5   May   73.5    66.6     72.1 2009        0          -6.9
## 6    6  June   79.8    73.0     78.6 2009        0          -6.8
## 7    7  July   83.3    74.7     78.1 2009        0          -8.6
## 8    8   Aug   82.0    78.6     80.1 2009        0          -3.4
## 9    9   Sep   75.1    77.2     73.3 2009        0           2.1
## 10 10   Oct   62.1    61.3     58.9 2009        0          -0.8
## 11 11   Nov   51.0    52.4     54.5 2009        1           1.4
## 12 12   Dec   40.4    40.6     36.7 2009        0           0.2
## 13 13   Jan   32.6    23.2     30.3 2010        0          -9.4
## 14 14   Feb   35.7    32.2     31.8 2010        0          -3.5
## 15 15   Mar   49.9    38.4     51.8 2010        1         -11.5
## 16 16   Apr   61.5    78.5     66.7 2010        1          17.0
## 17 17   May   73.5    74.1     73.6 2010        1           0.6
## 18 18  June   79.8    77.0     80.2 2010        1          -2.8
## 19 19  July   83.3    85.1     84.9 2010        1           1.8
## 20 20   Aug   82.0    83.6     84.2 2010        1           1.6
## 21 21   Sep   75.1    81.3     75.3 2010        1           6.2
## 22 22   Oct   62.1    59.5     63.8 2010        1          -2.6
## 23 23   Nov   51.0    47.6     53.0 2010        1          -3.4
## 24 24   Dec   40.4    29.7     30.2 2010        0         -10.7
## 25 25   Jan   32.6    34.9     29.2 2011        0           2.3
## 26 26   Feb   35.7    33.0     37.6 2011        1          -2.7
## 27 27   Mar   49.9    40.8     46.1 2011        0          -9.1
## 28 28   Apr   61.5    52.4     60.7 2011        0          -9.1
## 29 29   May   73.5    59.8     71.6 2011        0         -13.7
## 30 30  June   79.8    81.5     79.3 2011        0           1.7
## 31 31  July   83.3    86.6     87.3 2011        1           3.3
## 32 32   Aug   82.0    86.4     81.5 2011        0           4.4
## 33 33   Sep   75.1    79.1     73.0 2011        0           4.0
## 34 34   Oct   62.1    65.4     61.6 2011        0           3.3
## 35 35   Nov   51.0    60.2     56.7 2011        1           9.2
## 36 36   Dec   40.4    45.9     43.6 2011        1           5.5
```

```
## 37 37   Jan   32.6    40.6     39.9 2012        1            8.0
## 38 38   Feb   35.7    44.1     42.4 2012        1            8.4
## 39 39   Mar   49.9    50.8     62.6 2012        1            0.9
## 40 40   Apr   61.5    59.0     60.2 2012        0           -2.5
## 41 41   May   73.5    80.2     78.1 2012        1            6.7
## 42 42  June   79.8    71.4     81.6 2012        1           -8.4
## 43 43  July   83.3    92.3     88.2 2012        1            9.0
## 44 44   Aug   82.0    88.4     83.5 2012        1            6.4
## 45 45   Sep   75.1    84.7     73.4 2012        0            9.6
## 46 46   Oct   62.1    65.4     61.5 2012        0            3.3
## 47 47   Nov   51.0    42.2     48.6 2012        0           -8.8
## 48 48   Dec   40.4    52.6     43.6 2012        1           12.2
## 49 49   Jan   32.6    32.4     37.7 2013        1           -0.2
## 50 50   Feb   35.7    27.7     34.7 2013        0           -8.0
## 51 51   Mar   49.9    35.0     41.9 2013        0          -14.9
## 52 52   Apr   61.5    51.4     61.6 2013        1          -10.1
## 53 53   May   73.5    74.5     73.9 2013        1            1.0
## 54 54  June   79.8    71.6     78.5 2013        0           -8.2
## 55 55  July   83.3    81.1     81.5 2013        0           -2.2
## 56 56   Aug   82.0    76.8     80.2 2013        0           -5.2
## 57 57   Sep   75.1    77.2     74.4 2013        0            2.1
## 58 58   Oct   62.1    76.2     63.8 2013        1           14.1
## 59 59   Nov   51.0    56.1     46.4 2013        0            5.1
## 60 60   Dec   40.4    45.3     38.8 2013        0            4.9
```

```
WeatherTest
```

```
##      X Month Normal First7D Observed Year Binomial Deg_From_Norm
## 1  61   Jan   32.6    28.3     29.8 2014        0           -4.3
## 2  62   Feb   35.7    29.6     31.9 2014        0           -6.1
## 3  63   Mar   49.9    35.1     43.4 2014        0          -14.8
## 4  64   Apr   61.5    57.9     62.2 2014        1           -3.6
## 5  65   May   73.5    62.9     72.4 2014        0          -10.6
## 6  66  June   79.8    78.3     80.7 2014        1           -1.5
## 7  67  July   83.3    79.7     79.2 2014        0           -3.6
## 8  68   Aug   82.0    79.8     80.7 2014        0           -2.2
## 9  69   Sep   75.1    82.4     74.9 2014        0            7.3
## 10 70   Oct   62.1    62.9     61.6 2014        0            0.8
## 11 71   Nov   51.0    51.7     44.4 2014        0            0.7
## 12 72   Dec   40.4    41.2     40.8 2014        1            0.8
## 13 73   Jan   32.6    34.2     31.6 2015        0            1.6
## 14 74   Feb   35.7    33.7     25.9 2015        0           -2.0
## 15 75   Mar   49.9    33.5     45.0 2015        0          -16.4
## 16 76   Apr   61.5    59.2     60.4 2015        0           -2.3
## 17 77   May   73.5    77.1     76.1 2015        1            3.6
## 18 78  June   79.8    73.3     77.9 2015        0           -6.5
## 19 79  July   83.3    79.5     81.7 2015        0           -3.8
## 20 80   Aug   82.0    81.1     81.0 2015        0           -0.9
## 21 81   Sep   75.1    87.5     78.7 2015        1           12.4
## 22 82   Oct   62.1    65.7     63.4 2015        1            3.6
## 23 83   Nov   51.0    69.3     57.0 2015        1           18.3
## 24 84   Dec   40.4    45.8     50.8 2015        1            5.4
## 25 85   Jan   32.6    35.6     35.5 2016        1            3.0
## 26 86   Feb   35.7    49.1     42.5 2016        1           13.4
```

```
## 27 87   Mar   49.9    42.6     55.6 2016        1           -7.3
## 28 88   Apr   61.5    52.4     59.4 2016        0           -9.1
## 29 89   May   73.5    64.2     70.1 2016        0           -9.3
## 30 90  June   79.8    80.0     81.5 2016        1            0.2
## 31 91  July   83.3    80.9     85.5 2016        1           -2.4
## 32 92   Aug   82.0    88.1     86.4 2016        1            6.1
```

4) Using the data from `WeatherTrain` and the `glm()` function, build a logistic regression model to predict whether or not a month will be above normal based only on the how many degrees the first seven days are above normal. NOTE: There are lots of models that are "generalized linear models." To use a logistic model, you must specify `family = binomial` in the `glm()` function.

```
Weather_glm  <- glm( Binomial ~ Deg_From_Norm, data = WeatherTrain,family = "binomial")
Weather_glm
```

```
##
## Call:  glm(formula = Binomial ~ Deg_From_Norm, family = "binomial",
##     data = WeatherTrain)
##
## Coefficients:
##   (Intercept)  Deg_From_Norm
##      -0.09254        0.11819
##
## Degrees of Freedom: 59 Total (i.e. Null);  58 Residual
## Null Deviance:       82.91
## Residual Deviance: 74.36     AIC: 78.36
```

5) Use the `predict()` function to evaluate your model with integers from -20 to 20. NOTE: To use `predict()` the `newdata` must be a data frame where the columns have the same names as the those in the data frame you used to train your model.

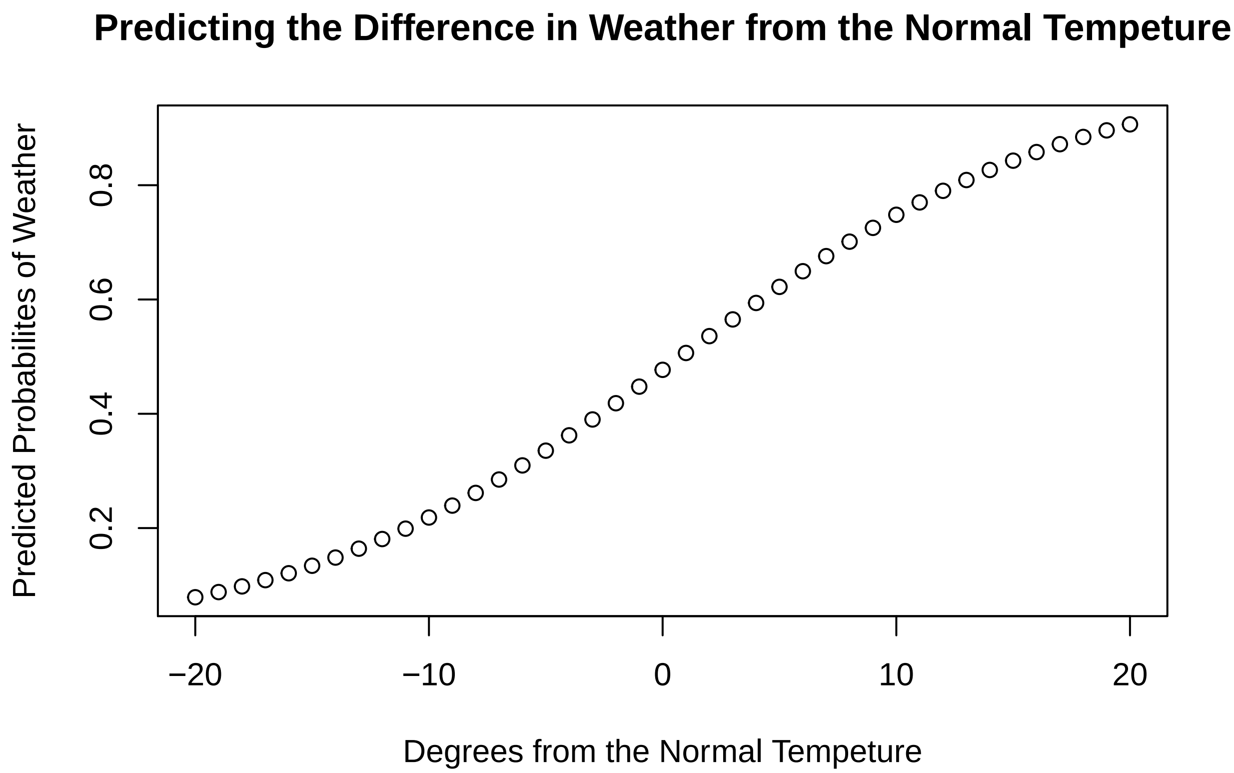
```
new_data <- data.frame(Deg_From_Norm = c(-20:20))  #why is is -20:20
new_data['predicted'] <- predict(Weather_glm, new_data, type = "response")
new_data
```

```
##    Deg_From_Norm  predicted
## 1            -20 0.07897658
## 2            -19 0.08801274
## 3            -18 0.09797280
## 4            -17 0.10892538
## 5            -16 0.12093822
## 6            -15 0.13407654
## 7            -14 0.14840122
## 8            -13 0.16396654
## 9            -12 0.18081781
## 10           -11 0.19898873
## 11           -10 0.21849863
## 12            -9 0.23934981
## 13            -8 0.26152493
## 14            -7 0.28498479
## 15            -6 0.30966662
## 16            -5 0.33548312
```

```
## 17              -4 0.36232228
## 18              -3 0.39004830
## 19              -2 0.41850355
## 20              -1 0.44751164
## 21               0 0.47688147
## 22               1 0.50641205
## 23               2 0.53589795
## 24               3 0.56513500
## 25               4 0.59392585
## 26               5 0.62208523
## 27               6 0.64944445
## 28               7 0.67585505
## 29               8 0.70119137
## 30               9 0.72535209
## 31              10 0.74826061
## 32              11 0.76986457
## 33              12 0.79013450
## 34              13 0.80906189
## 35              14 0.82665674
## 36              15 0.84294497
## 37              16 0.85796565
## 38              17 0.87176830
## 39              18 0.88441033
## 40              19 0.89595477
## 41              20 0.90646816
```

6) Create a plot with the integers from -20 to 20 on the x-axis and the predicted probabilities on the y-axis. Give the plot some descriptive labels.

```
plot(x = c(-20:20), y = new_data$predicted, xlab = "Degrees from the Normal Tempeture", ylab = "Predicte
```



7) Estimate the input that would be needed to give an output of 0.75. What does this mean in the context of the model?

   In order for the weather to have a 0.75 or a 75% probabiliry of the weather the predicted tempeture must be be around 10 degrees from the normal tempeture.

For a classification problem, we want a prediction of which class the outcome variable belongs to. In order to get a prediction from a binomial logistic regression model, we define a *threshold*. If the output of the model is above the threshold, then we predict class 1, and if it is below the threshold we predict class 0.

8) Using a threshold value of 0.5, obtain a vector of class predictions for the data set `WeatherTrain`. HINT: the `ifelse` function might be useful here.

```
probabilityWeather <- predict(Weather_glm, WeatherTrain, type = "response")
classWeatherTrain <- ifelse(probabilityWeather  > 0.5, 1, 0)
classWeatherTrain
```

```
##  1  2  3  4  5  6  7  8  9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26
##  1  0  0  0  0  0  0  0  1  0  1  0  0  0  0  1  0  0  1  1  1  0  0  0  1  0
## 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52
##  0  0  0  1  1  1  1  1  1  1  1  1  1  0  1  0  1  1  1  1  0  1  0  0  0  0
## 53 54 55 56 57 58 59 60
##  1  0  0  0  1  1  1  1
```

9) Use the `table()` function to construct a *confusion matrix* for your predictions. What is the accuracy of your predictions? How many false positives (months incorrectly classified as above average) are there?

```r
#binomial = actual temps
#classWeatherTrain = predicted temps
table(WeatherTrain$Binomial, classWeatherTrain)
```

```
##    classWeatherTrain
##      0  1
##   0 20 12
##   1 11 17
```

```r
37/60
```

```
## [1] 0.6166667
```

false positives : 12 false negatives: 11 mis-calssifacitions: 23 (bad)

```
There is a 61.6% accuracy in this prediction.
```

10) A threshold of 0.5 isn't necessarily the best choice for the threshold. Experiment with other values for the threshold to see if you can obtain any greater accuracy on the training set.

```r
classWeatherTrain2 <- ifelse(probabilityWeather  > 0.2, 1, 0)
table(WeatherTrain$Binomial, classWeatherTrain2)
```

```
##    classWeatherTrain2
##      0  1
##   0  2 30
##   1  1 27
```

```r
classWeatherTrain3 <- ifelse(probabilityWeather  > 0.8, 1, 0)
table(WeatherTrain$Binomial, classWeatherTrain3)
```

```
##    classWeatherTrain3
##      0  1
##   0 32  0
##   1 26  2
```

```r
classWeatherTrain4 <- ifelse(probabilityWeather  > 0.6, 1, 0)
table(WeatherTrain$Binomial, classWeatherTrain2)
```

```
##    classWeatherTrain2
##      0  1
##   0  2 30
##   1  1 27
```

```r
classWeatherTrain5 <- ifelse(probabilityWeather  > 0.7, 1, 0)
table(WeatherTrain$Binomial, classWeatherTrain2)
```

```
##    classWeatherTrain2
##      0  1
##   0  2 30
##   1  1 27
```

true positives : 11 true negatives: 28 miscalssifacitions: 11 (GOOD 0.6)

11) Regardless of the model used, there is at least one flawed assumption inherent in using the temperatures from a month's first seven days to predict the temperature for the rest of the month. What is the issue? (Hint: think about seasons)

```
Because the data is periodic, the tempetures are always going to alternate regularly in the summe
```