C3M1_peer_reviewed

March 13, 2022

1 C3M1: Peer Reviewed Assignment

1.0.1 Outline:

The objectives for this assignment:

1. Apply Binomial regression methods to real data.
2. Understand how to analyze and interpret binomial regression models.
3. Flex our math skills by determining whether certain distributions are members of the exponential family.

General tips:

1. Read the questions carefully to understand what is being asked.
2. This work will be reviewed by another human, so make sure that you are clear and concise in what your explanations and answers.

[17]:

```
# Load required libraries
library(tidyverse)
library(dplyr)
```

1.1 Problem 1: Binomial (Logistic) Regression

The National Institute of Diabetes and Digestive and Kidney Diseases conducted a study of 768 adult female Pima Indians living near Phoenix, AZ. The purpose of the study was to investigate the factors related to diabetes.

Before we analyze these data, we should note that some have raised ethical issues with its collection and popularity in the statistics and data science community. We should think seriously about these concerns. For example, Maya Iskandarani wrote a brief piece on consent and privacy concerns raised by this dataset. After you familarize yourself with the data, we'll then turn to these ethical concerns.

First, we'll use these data to get some practice with GLM and Logistic regression.

[18]:

```
# Load the data
library(caret)
pima = read.csv("pima.txt", sep="\t")
# Here's a description of the data: https://rdrr.io/cran/faraway/man/pima.html
```

```
head(pima)
```

A data.frame: 6 × 9

	pregnant	glucose	diastolic	triceps	insulin	bmi	diabetes	age	test
	<int>	<int>	<int>	<int>	<int>	<dbl>	<dbl>	<int>	<int>
1	6	148	72	35	0	33.6	0.627	50	1
2	1	85	66	29	0	26.6	0.351	31	0
3	8	183	64	0	0	23.3	0.672	32	1
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
6	5	116	74	0	0	25.6	0.201	30	0

1.1.1 1. (a) Data Cleaning? What about Data Scrubbing? Data Sterilizing?

This is a real data set, which means that there's likely going to be gaps and missing values in the data. Before doing any modeling, we should inspect the data and clean it if necesary.

Perform simple graphical and numerical summaries of the data. Pay attention for missing or nonsensical values. Can you find any obvious irregularities? If so, take appropriate steps to correct these problems. In the markdown cell, specify what cleaning you did and why you did it.

Finally, split your data into training and test sets. Let the training set contain 80% of the rows and the test set contain the remaining 20%.

[19]:
```
# Your Code Here
summary(pima)
```

```
    pregnant         glucose        diastolic         triceps     
 Min.   : 0.000   Min.   :  0.0   Min.   :  0.00   Min.   : 0.00  
 1st Qu.: 1.000   1st Qu.: 99.0   1st Qu.: 62.00   1st Qu.: 0.00  
 Median : 3.000   Median :117.0   Median : 72.00   Median :23.00  
 Mean   : 3.845   Mean   :120.9   Mean   : 69.11   Mean   :20.54  
 3rd Qu.: 6.000   3rd Qu.:140.2   3rd Qu.: 80.00   3rd Qu.:32.00  
 Max.   :17.000   Max.   :199.0   Max.   :122.00   Max.   :99.00  
    insulin           bmi           diabetes           age       
 Min.   :  0.0   Min.   : 0.00   Min.   :0.0780   Min.   :21.00  
 1st Qu.:  0.0   1st Qu.:27.30   1st Qu.:0.2437   1st Qu.:24.00  
 Median : 30.5   Median :32.00   Median :0.3725   Median :29.00  
 Mean   : 79.8   Mean   :31.99   Mean   :0.4719   Mean   :33.24  
 3rd Qu.:127.2   3rd Qu.:36.60   3rd Qu.:0.6262   3rd Qu.:41.00  
 Max.   :846.0   Max.   :67.10   Max.   :2.4200   Max.   :81.00  
      test      
 Min.   :0.000  
 1st Qu.:0.000  
 Median :0.000  
 Mean   :0.349  
 3rd Qu.:1.000  
 Max.   :1.000  
```

[20]:

```
nrow(pima)
```

```
768
```

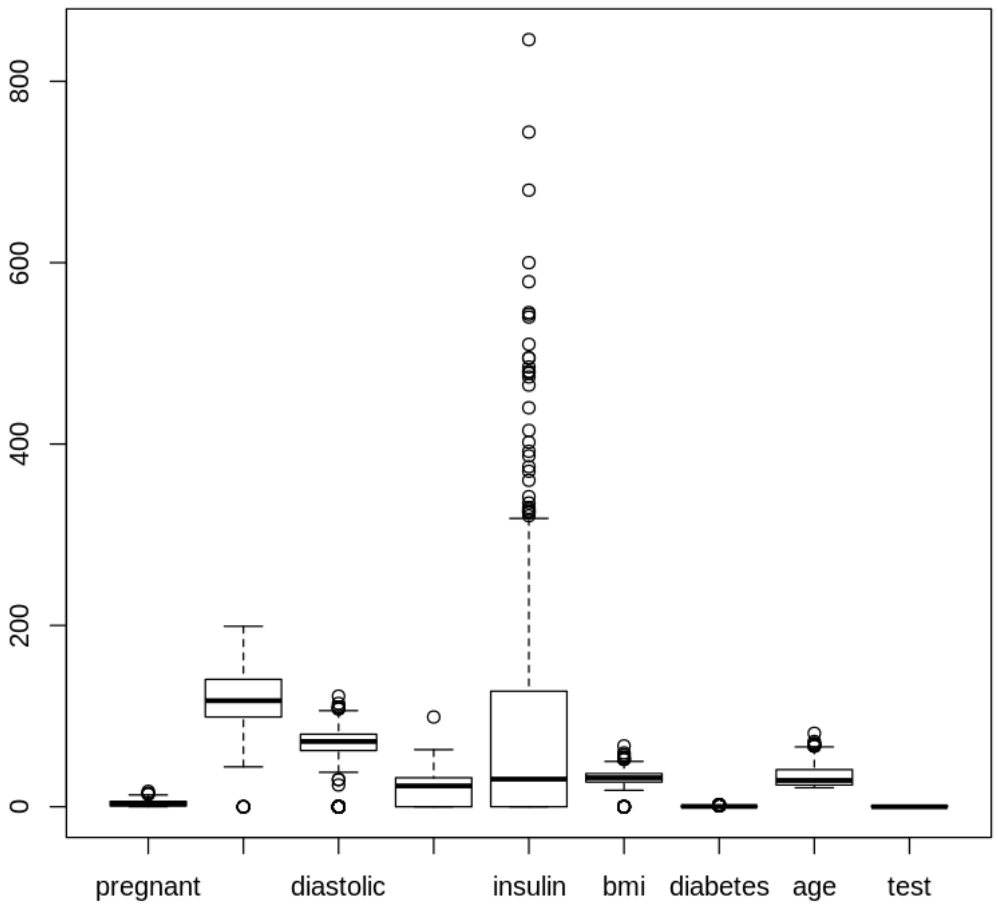
[21]:

```
colSums(is.na(pima))
```

```
pregnant  0 glucose  0 diastolic  0 triceps  0 insulin  0 bmi  0 diabetes  0 age  0 test  0
```

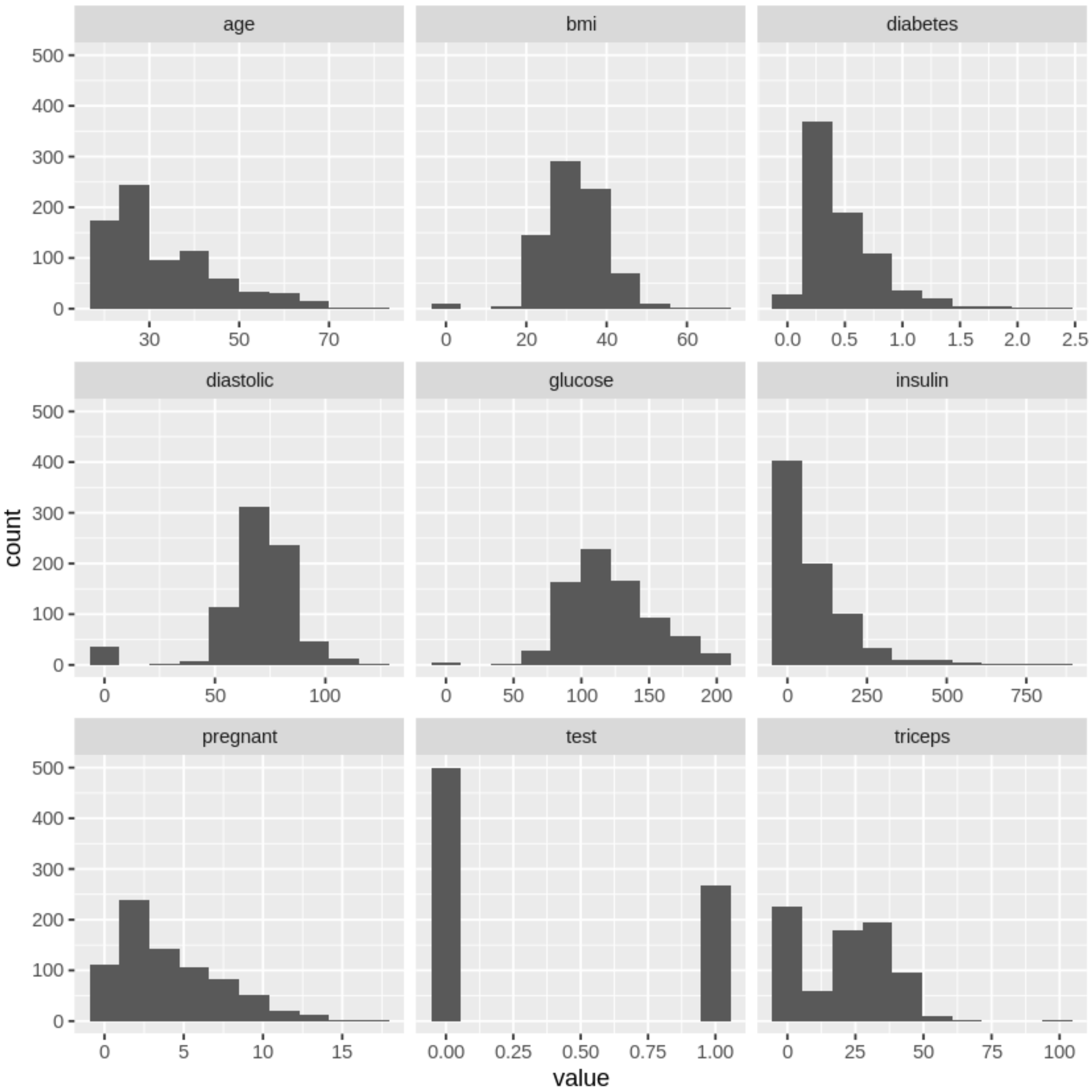
[22]:

```
boxplot(pima,use.cols=TRUE)
```



[23]:

```
ggplot(gather(pima), aes(value)) +
    geom_histogram(bins=10) +
    facet_wrap(~key, scales='free_x')
```



2 Explanation of Data Cleaning

Notice that many predictors have many values of 0. It's impossible to have a BMI of 0. People with diabetes are known to have a low diastolic measurement; however, having it measured at 0 is very rare. A glucose measure of 0 will cause a diabetic to go to the hospital. Also, an insulin level of 0 will cause major problems. Last, it is impossible for someone to have a tricep measurement of 0. Because a measurement of '0' seems to represent a missing a value for these predictors, we will remove them from the dataset. We will not remove the 0s of the categorical variables because the value represents a class.

[30]:

```
concerning_cols = c('bmi', 'diastolic', 'glucose', 'insulin', 'triceps')
pima[concerning_cols][pima[concerning_cols] == 0] = NA
pima = na.omit(pima)
pima$test = as.factor(pima$test)
head(pima)
```

A data.frame: 6 × 9

	pregnant	glucose	diastolic	triceps	insulin	bmi	diabetes	age	test
	<int>	<int>	<int>	<int>	<int>	<dbl>	<dbl>	<int>	<fct>
4	1	89	66	23	94	28.1	0.167	21	0
5	0	137	40	35	168	43.1	2.288	33	1
7	3	78	50	32	88	31.0	0.248	26	1
9	2	197	70	45	543	30.5	0.158	53	1
14	1	189	60	23	846	30.1	0.398	59	1
15	5	166	72	19	175	25.8	0.587	51	1

[31]:

```
sample_size = floor(0.8 * nrow(pima))

set.seed(21412)
train_indices = sample(seq_len(nrow(pima)), size = sample_size)

train_df = pima[train_indices,]
test_df = pima[-train_indices,]
```

[32]:

```
print(nrow(train_df))
print(nrow(test_df))
```

```
[1] 313
[1] 79
```

2.0.1 1. (b) Initial GLM modelling

Our data is clean and we're ready to fit! What kind of model should we use to fit these data? Notice that the test variable is either 0 or 1, for whether the individual tested positive for diabetes. Because test is binary, we should use logistic regression (which is a kind of binomial regression).

Fit a model with test as the response and all the other variables as predictors. Can you tell whether this model fits the data?

[35]:

```
logit1 = glm(test~., family=binomial, data=train_df)
summary(logit1)
par(mfrow = c(2,2))
plot(logit1)
```

```
Call:
glm(formula = test ~ ., family = binomial, data = train_df)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

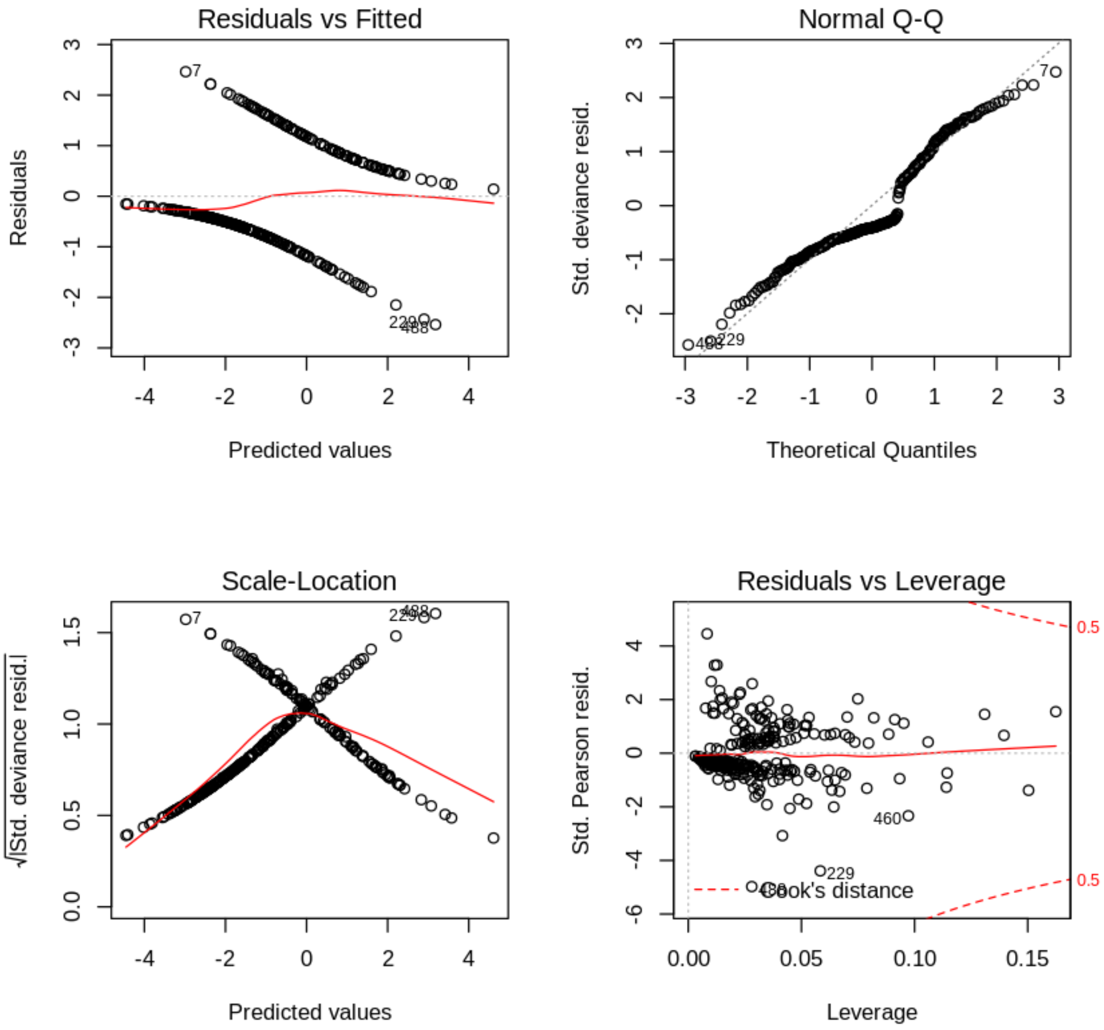
```
-2.5383  -0.6614  -0.4020   0.6681   2.4619

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -9.380267   1.292969  -7.255 4.02e-13 ***
pregnant     0.062436   0.061323   1.018   0.3086
glucose      0.038305   0.006342   6.040 1.54e-09 ***
diastolic   -0.006308   0.013000  -0.485   0.6275
triceps      0.008483   0.018438   0.460   0.6454
insulin     -0.001511   0.001520  -0.994   0.3202
bmi          0.067658   0.029475   2.295   0.0217 *
diabetes     0.826397   0.451822   1.829   0.0674 .
age          0.042264   0.020424   2.069   0.0385 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 400.73  on 312  degrees of freedom
Residual deviance: 284.21  on 304  degrees of freedom
AIC: 302.21

Number of Fisher Scoring iterations: 5
```



3 Answer

We cannot tell if this model is a good fit. Even using the deviance test wouldn't work because our response is a 0-1 Bernoulli distribution. Thus, the goodness of fit metrics do not work.

3.0.1 1. (c) Remember Bayes

A quick analytical interlude.

Is diastolic blood pressure significant in the regression model? Do women who test positive have higher diastolic blood pressures? Explain the distinction between the two questions and discuss why the answers are only apparently contradictory.

[38]:

```
diastolic_reg = lm(diastolic ~ test, data = train_df)
summary(diastolic_reg)
```

```
Call:
lm(formula = diastolic ~ test, data = train_df)

Residuals:
    Min      1Q  Median      3Q     Max 
-45.295  -9.038   0.705   7.962  36.705 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  69.2947     0.8712  79.543  < 2e-16 ***
test1         4.7430     1.4970   3.168  0.00169 ** 
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 12.53 on 311 degrees of freedom
Multiple R-squared:  0.03127,Adjusted R-squared:  0.02815 
F-statistic: 10.04 on 1 and 311 DF,  p-value: 0.001685
```

3.1 Answer:

Diastolic blood pressure is not signficant in the logistic regression model at an .05 significance level. Looking at the regression modeling comparing the test result (as the predictor) and the women's diastolic measurement, women who test positive do have higher diastolic pressure. The first question asks if the `diastolic` variable is significant in the model; in other words, we are measuring the liklihood of having a positive test, given the measurement of the diastolicy test (conditional). On the other hand, the second question asks if we are given a positive test, will it be a high diastolic test (testing diastolic based on if they test positive or not). These two tests are not the same.

3.1.1 1. (d) GLM Interpretation

We've seen so many regression summaries up to this point, how is this one different from all the others? Well, to really understand any model, it can be helpful to loop back and plug the fitted results back into the model's mathematical form.

Explicity write out the equation for the binomial regression model that you fit in (b). Then, in words, explain how a 1 unit change of `glucose` affects `test`, assuming all other predictors are held constant.

3.1.2 Model:

$\log\left(\frac{p}{1-p}\right)$ = -9.380 + 0.062* Pregnant + 0.038 * Glucose + -0.006 * Diastolic + -0.008 * Triceps + -0.002 * Insulin + 0.067 * BMI + 0.826 * diabetes + 0.042 * Age

3.1.3 Answer:

Assuming all other predictors are held constant, a 1 unit change in glucose will increase the log odds of success of test by 0.038.

3.1.4 1. (e) GLM Prediction

One of the downsides of Logistic Regression is that there isn't an easy way of evaulating the goodness of fit of the model without predicting on new data. But, if we have more data to test with, then there are many methods of evaluation to use. One of the best tools are confusion matrices, which (despite the name) are actually not that hard to understand.

A confusion matrix compares the predicted outcomes of a Logistic Regression Model (or any classification model) with the actual classifications. For binary classification, it is a 2×2 matrix where the rows are the models' predicted outcome and the columns are the actual classifications. An example is displayed below.

	True	False
1	103	37
0	55	64

In the example, we know the following information: * The [1,1] cell is the number of datapoints that were correctly predicted to be 1. The value (103) is the number of True Positives (TP). * The [2,2] cell is the number of datapoints that were correctly predicted to be 0. The value is the number of True Negatives (TN). * The [1, 2] cell is the number of datapoints that were predicted to be 1 but where actually 0. This is the number of False Positives (FP), also called Type I error. In the context of our diabetes dataset, this would mean our model predicted that the person would have diabetes, but they actually did not. * The [2, 1] cell is the number of datapoints that were predicted to be 0 but where actually 1. This is the number of False Negatives (FN), also called Type 2 error. In the context of our diabetes dataset, this would mean our model predicted that the person would not have diabetes, but they actually did have diabetes.

Use your model to predict the outcomes of the test set. Then construct a confusion matrix for these predictions and display the results.

[40]:

```r
# Your Code Here
# Predictions on the row names and Actual Results as the column names
predictions = predict(logit1, newdata = test_df, type="response")
predictions = factor(ifelse(predictions > 0.5, 1, 0))
table(predictions, test_df$test)
```

```
predictions  0  1
          0 50  8
          1  5 16
```

This is a similar to the confusion matrix of:

	True	False
1	16	5
0	50	8

3.1.5 1. (f) Evaluation Statistics

Using the four values from the confusion matrix, we can construct evaulation statistics to get a numerical approximation for our model's performance. Spend some time researching accuracy, precision, recall and F score.

Calculate these values for your model's predictions on the test set. Clearly display your results. How well do you think your model fits the data?

```
[15]: length(predictions)
```

152

```
[41]: # Your Code Here
      #Accuracy
      print(paste0("The Accuracy of the Model is: ", ((16 + 50) /␣
       ↪length(predictions))))

      #Precision
      precision = 16 / (16 + 5)
      print(paste0("The Precision of the Model is: ", precision))

      #Recall
      recall = 16 / (16 + 8)
      print(paste0("The Recall of the Model is: ", recall))
      #F1-Score
      print(paste0("The F1 Score of the Model is: ", 2 * ((precision * recall) /␣
       ↪(precision + recall))))
```

```
[1] "The Accuracy of the Model is: 0.835443037974684"
[1] "The Precision of the Model is: 0.761904761904762"
[1] "The Recall of the Model is: 0.666666666666667"
[1] "The F1 Score of the Model is: 0.711111111111111"
```

Overall, this model has high accuracy and precision. However, the recall is fairly low; this causes the F1 score (the harmonic mean of the Precision and Recall) to drop. Thus, I would say the model fits our data okay.

3.1.6 1. (g) Understanding Evaluation Statistics

Answer the following questions in the markdown cell below.

1. Give an example scenario for when accuracy would be a misleading evaulation statistic.
2. Confusion matrices can also be used for non-binary classification problems. Describe what a confusion matrix would look like for a response with 3 levels.
3. You'll have to take our word on the fact (or spend some time researching) that Type I error and Type II error are inversely related. That is, if a model is very good at detecting false positives, then it will be bad at detecting false negatives. In the case of our diabetes dataset, would you prefer a model that overestimates the Type 1 error or overestimates the Type II error. Justify your answer.

1. Accuracy would be misleading in an algoirithm where we are attempting to identify spam email. If 99 out of every 100 emails are not spam, we could receive 99% accuracy but missclassify the spam emails often. In this case, recall may be the best measurement. In a case of email spam, it measures the rate at we identify actual postive cases as positive.
2. A confusion matrix with 3 levels would have each level label as a row for the predicted class and each level label as a column for the actual class. Thus, we would have a 3x3 matrix. Each column intersection would have a count for that predicted label/ actual label combination.
3. In classification recall is often used instead of Type I error. In our situation with diabetes, we don't want to miss a needed diagnosis. However, if we misdiagnose someone with diabetes, it won't be as harmful. Therefore, we would want to have a model that overestimates the Type I error.

3.1.7 1. (h) Ethical Issues in Data Collection

Read Maya Iskandarani's piece on consent and privacy concerns raised by this dataset. Summarize those concerns here.

First, Maya states that there is concern with how accessible the personal information of the Prima tribe is. It intervenes with numerous privacy concerns. Also, Maya states the Prima tribe had no knowledge their health data was being collected. Lastly, Iskandarani states that the dataset has been available fore many years; maintaining this data for even longer can be harmful.

3.2 Problem 2: Practicing those Math skills

One of the conditions of GLMs is that the "random component" of the data needs to come from the Exponential Family of Distributions. But how do we know if a distribution is in the Exponential Family? Well, we could look it up. Or we could be proper mathematicians and check the answer ourselves! Let's flex those math muscles.

3.2.1 2. (a) But it's in the name...

Show that $Y \sim exponential(\lambda)$, where λ is known, is a member of the exponential family.

The PDF can be written as the following:

$$f(y;\theta;\phi) = \exp\Big\{\frac{y^{\phi} - b(\phi)}{a(\phi)} + c(y,\phi)\Big\}$$

Therfore, **Y** is exponentially distributed if it is in the form of:

$$\exp\Big(\frac{\lambda y - \log(\lambda)}{-1} + 0\Big)$$

3.2.2 2. (b) Why can't plants do math? Because it gives them square roots!

Let $Y_i \sim \mathit{exponential}(\lambda)$ where $i \in \{1, \ldots, n\}$. Then $Z = \sum_{i=1}^{n} Y_i \sim \mathit{Gamma}(n, \lambda)$. Show that Z is also a member of the exponential family.

$$f(y;n,\lambda) = \frac{\lambda^n}{\Gamma(n)} y^{n-1} e^{-\lambda y}$$

```
[ ]:
```