Final Report

1. Introduction
   - Dataset

     The dataset used in this project is obtained from https://hbiostat.org/data/. This data is originally collected to develop a screening program for female relatives of boys with DMD. The initial program's goal was to inform a woman of her chances of being a carrier based on serum markers as well as her family pedigree. This data consists of mostly real and integer variables with some boolean variables.The variable included in this data set are as such :
     - Hospital ID (hospid)
     - Age in years (age)
     - Date of study (sdate)
     - Creatine Kinase (ck), real values which represent the level of creatine kinase, an enzyme located in skeletal and heart muscles which plays a role in energy production.
     - Hemopexin (h), real values representing the level of hemopexin a patient which is a plasma glycoprotein which plays a role in transport of heme in blood.
     - Pyruvate Kinase (pk), real values representing level of pyruvate kinase, an enzyme which plays a crucial role in glycolysis, transforming glucose into ATP.
     - Lactate Dehydrogenase (ld), real values which represent the level of lactate dehydrogenase, an enzyme which plays a role in energy production specifically anaerobic metabolism.
     - Carrier of DMD (carrier), boolean variable indicating whether the the patient is a carry of DMD or not.
     - Observation number within patient (obsno), integer indicating the number of times the patient has been observed in the institute.

   - Literature Review and Problem Formulation

     Duchenne muscular dystrophy is a form of inherited muscular dystrophy which does not exhibit prediction for any race or ethnic group. This disease affects primarily male as the disease occurred due to mutations in the X genes. According to the John Hopkins Medicine website, this disease primarily occurs on young boys withing the age of 3 - 6 years old (Medicine, 2025). The symptom of this disease includes mobility issues, facial weakness, and in some cases heart problems. The website also suggested some ways of diagnosing DMD such as, blood tests, muscle biopsy, EMG, or EKG. It is stated that the first line of treatment for this disease is corticosteroids, which have been proven to decrease the rate of muscle deterioration of DMD patients.

     Research done in 2016 by van Westering, Betts, and Wood further explores the development of therapeutic strategies for Duchenne muscular dystrophy, emphasizing the role of corticosteroids in delaying disease progression and the potential of gene-targeted treatments in clinical (van Westering, 2015). This research indicates the existence levels of this disease which are BMD (Becker muscular dystrophy) which is a milder form of DMD. This research then dives deeper into the topic of ways of diagnosing this disease through gene analysis, specifically on the analysis of the DMD genes which are responsible for the creation of dystrophin protein. In this project I would do a diagnosis approach not using genes analysis but instead by determining carriers of this disease by using the provided stats such as levels of the provided enzyme, and age.

     The main problem of this project would be creating a logistic model to detect female carriers of this disease using the provided data using a Bayesian approach. Based on this Bayesian regression I would then compare this model to another model with a frequentist approach which is made to be as simple as possible and verify it with LASSO method to make sure that all the variables selected are relevant.

2. Methodology

- Data preparation

  As the data used has been previously used to create a model with similar purposes it is expected that all the variables will be significant enough for the model. My only concern about this data would be the correlation between the variables which may be a problem when creating a good model. For the splits I have chosen to do a 0.8 as in my opinion it would be a fair split considering the size of the data. I have also omitted those with NA values as there would still be sufficient data even though those have been omitted. After analyzing the correlations between variables, I have decided to remove Pyruvate Kinase, as it is highly correlated with Creatine Kinase. I have also decided to remove unnecessary variables such as Hospital ID, Study date, and Number of observations as these variables seem to be useless for the purposes of this study.

- Frequentist Approach

  The first approach I decided to use the glm() method with carrier as the response variable, and family to be binomial for logistic regression. From this simple regression I obtained a result as such:

```
Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -17.79216    3.23170  -5.506 3.68e-08 ***
age           0.14454    0.04935   2.929 0.003405 **
ck            0.04736    0.01406   3.369 0.000754 ***
h             0.08280    0.02749   3.012 0.002598 **
ld            0.01239    0.00556   2.228 0.025850 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 204.956  on 154  degrees of freedom
Residual deviance:  76.277  on 150  degrees of freedom
AIC: 86.277

Number of Fisher Scoring iterations: 8
```

  And when doing the lasso method to verify the variable selection by cross checking the Lasso coefficient path plot to the cross-validation plot, I confirmed that all 4 variables are not redundant. Using these variables, I obtained a model with an accuracy of 94.8% when using a threshold of 0.5 for diagnosing the carriers.

- Bayesian Approach with Naïve Priors

  From the previous approach I have decided to select a prior distribution as such:

  - Intercept (alpha) ~ Norm(-5, 10) this is selected as the number of female carriers of DMD is roughly 0.6% which is closest to -5 when in logit terms.
  - Age (beta[1]) ~ Norm(0.1, 0,1) this is selected as from my EDA I have noticed that most of the patience are in the 30s – 50s and there might sort of 10 years to a 10% increase in the probability.
  - Ck, H, Ld (beta[2], beta[3], beta[4]) ~ Norm(0.01, 0.1) this is selected as I observed that most of the variable are different in mean at around 100 for all the variables.

  After preparing the stan object I prepared the data set that is going to be used which would consist of 4 items, the number of rows in the training dataset (N), Number of coefficients needed (k), the explanatory variables (X), and the response variable (y). After setting up the Bayesian approach I would then compute the posterior using 5 independent Markov chains and 2000 iteration. I have selected 5 independent Markov chains as I believe they would be sufficient and not too expensive computationally. This would also be the same justification as why I selected 2000 iterations. From these

computations I would obtain 5000 posteriors which I would then take the mean of to create the final posterior coefficients. The final posterior coefficients are as such :

	mean	se_mean	sd	2.5%	50%	97.5%	n_eff	Rhat
alpha	-17.66	0.06	2.92	-23.82	-17.56	-12.31	2430	1
beta[1]	0.14	0.00	0.04	0.06	0.14	0.23	3305	1
beta[2]	0.05	0.00	0.01	0.02	0.05	0.08	3114	1
beta[3]	0.08	0.00	0.03	0.03	0.08	0.14	2638	1
beta[4]	0.01	0.00	0.01	0.00	0.01	0.02	3723	1

After comparing the values to the frequentist approach and then verifying the trace plot, I would confirm that this model is well mixed. I am also convinced that the model is well made as the values obtained are quite like those obtained with the frequentist approach.

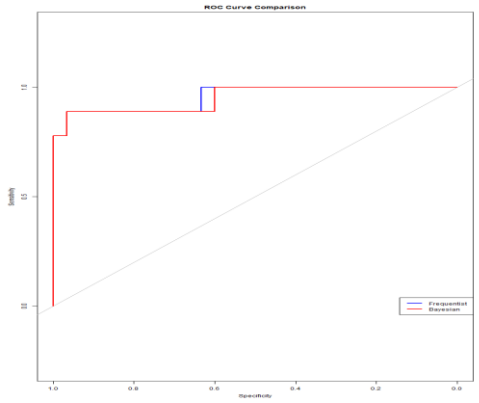
3. Discussion

- Results
- From the 2 approaches done above I have obtained a confusion matrix as such:

```
> print(conf_matrix_freq)
          Actual
Predicted  0  1
        0 30  2
        1  0  7
```

```
> print(conf_matrix_bayes)
          Actual
Predicted  0  1
        0 30  2
        1  0  7
```

The confusion matrix showed that the frequentist approach has achieved the same results as the Bayesian approach. The fact that the two results are similar is expected as both coefficients seem to be similar, making the prediction to be similar. The only error made in this would be a false positive which would be unfavorable in the medical field, but considering that this prediction is made to diagnose carriers and not the actual patients, using 0.5 as the threshold is still justifiable as in this case knowing that someone is a carrier of a disease is not as urgent and making a miss diagnosis would cause panic to the patients when there is no urgency. The second plot I have generated to compare the performance of the two model is the ROC plot where blue is the frequentist approach and red is the Bayesian approach.



The plot also indicates that the frequentist approach performs slightly better than the Bayesian approach with AUC of 0.955 compared to 0951. As this result seems weird, I have done multiple iterations of these modeling with different seeds and the results seem to be random where sometimes the Bayesian approach outperforms the frequentist approach by a slight margin and the other way around. From this result I have the suspicion that the two model performs similarly since the sample size is quite large whereas in the assignments, we have always used smaller data set making the prior selection the main advantage of the Bayesian approach. The fact that most of the carriers have very high Creatine Kinase or Lactate Dehydrogenase would also mean that the Bayesian approach would not be as good as we want it to be, because this means that any prior belief would just be overwritten by this fact. Which would be the edge that the Bayesian approach would have if we knew of the estimated prior.

- Limitations

  Some main limitations of this model are the fact the variable has been preselected as the selected data are those which have been previously used by doctors and previous researchers to diagnose this disease. Which means that the variables are expected to be good indicators of the disease. Other limitations of this model would be the fact that we assumed that all the predictors are linearly correlated to the logit of the probability. We did not explore any other relations such as quadratic, logarithmic, or other form of multiplicative variable's combination. Another problem that would be a limitation is the fact that the data is obtained from 1960s research which means that it may very well been outdated. The source of the data set is also not clear form which regions, meaning that there may be some sort of latent variable connected to the area the data is obtained from. We also did not use cross validation to verify the model therefore, there might be an issue with overfitting when deployed into the real world.

  As for the performance of the Bayesian model, we could not create a better model as we have no access to previous models used to predict the carriers of DMD therefore, we could not construct a model with better prior. Another limitation to the comparison is the fact that the frequentist approach has a method of verifying that all 4 variables are necessary whereas the Bayesian model is done as simple as possible with no method of verifying the importance of all variables. This implies that there may be a simpler model for the Bayesian approach which we did not explore.

Reference

Medicine, J. H. (2025, 04 18). Retrieved from John Hopkins Medicine: https://www.hopkinsmedicine.org/health/conditions-and-diseases/duchenne-muscular-dystrophy

van Westering, T. L. (2015). The importance of genetic diagnosis for Duchenne muscular dystrophy. *Disease Models & Mechanisms*, 195 - 213.