Samuel Brown, Patrick Dunnington, Will Sivolella

Getting Big (and Fast) Final Report

Problem Description

Athletes of all skill and experience levels seek to optimize their exercise routines to reach personal fitness goals. However, there are often too many proposed training methods to clearly understand what actions actually harm or improve performance. This leads to a lack of clarity in fitness circles. Furthermore, there is a significant difference in fitness types. Maximizing anaerobic performance requires different training methods than maximizing aerobic performance.

As a result, both the amount of proposed training methods and variety in fitness goals makes it challenging for amateur athletes to optimize their workouts and lifestyles. This confusion leads to suboptimal workout and lifestyle routines which may even prevent athletes from achieving their fitness goals.

Data Description

In order to better understand effective training practices we explored a Crossfit Athletes kaggle dataset (see link in references). This dataset is drawn from a survey given to crossfit athletes of all experience levels, and can be applied to the fitness journeys of regular people who aren't necessarily spending hours in the gym every day. It contains hundreds of thousands of responses, although only about 19,000 were used in our modeling process after cleaning. Most importantly, the survey contains athlete information regarding rest days, eating habits, weekly workout quantity and years of crossfit experience, all changeable factors individuals can utilize to improve their fitness.

The data also provides a variety of performance metrics. Some are specific to crossfit such as the filthy fifty (a combination of anaerobic and aerobic exercises in crossfit combinations), while others such as weight for deadlift and squat or times for 400 meter and 5K are more universal measures of strength and speed.

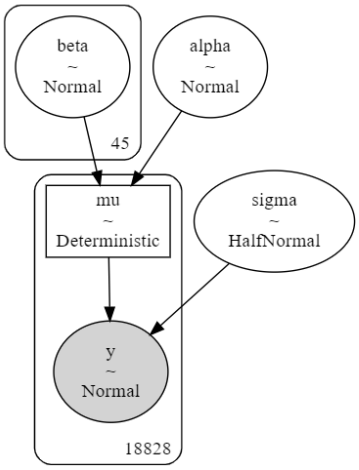
In order to prepare the data for modeling, we first cleaned the dataset, removing outliers and NA values from our relevant numeric variables. We then used one-hot encoding for every categorical predictor variable we intended to use in our model, extracting some of these answers from the complicated multiple-choice answers that the dataset gave to us. To add to our analysis, we added in two new variables: age squared and BMI, as we thought that age may have a nonlinear relationship with performance and that BMI would be a better measure of body size than raw height or weight for running.

Probability Model

In order to determine which predictors had the greatest impact on these crossfit athletes' performance, we built two Bayesian Linear Regression models: one for max deadlift weight in pounds, and one for fastest 5k time in seconds. The predictor variables remained almost the same, with height and weight in the deadlift model being replaced by BMI for running.

The first step was to place priors on the intercept and predictor coefficients. Since we needed to capture positive and negative values for predictors we did not fully understand, uninformed gaussian priors seemed best. Then, the standard deviation needed to be bounded to the positive domain, so that prior was a Half Normal distribution.

A Gaussian distribution was used for the likelihood function in the deadlift model, since the observed deadlift results loosely followed the normal curve. However, the likelihood distribution for the 5k model was adjusted to a right skewed normal distribution to better fit observed values after testing different models.



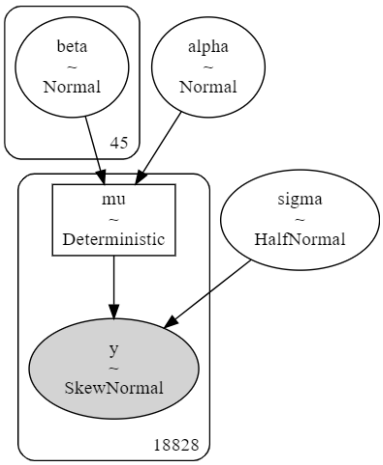


Fig. 1: Deadlift Model Graph Fig. 2: 5K Model Graph

These probability models make sense in context because deadlift weight likely follows a wider more evenly distributed pattern based on athletes' varying strength. However, fit individuals will probably run a 5k within 3-5 minutes of a total 20 minute time. This shifts the mean left and increases the length of the right tail where some individuals run significantly slower times. From these probability distributions the posterior predictive plots are built below.

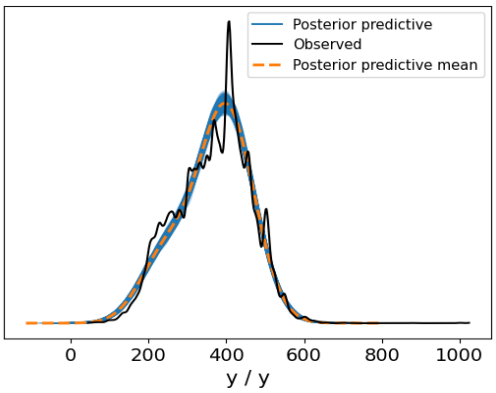


Fig. 3: Deadlift Posterior Predictions

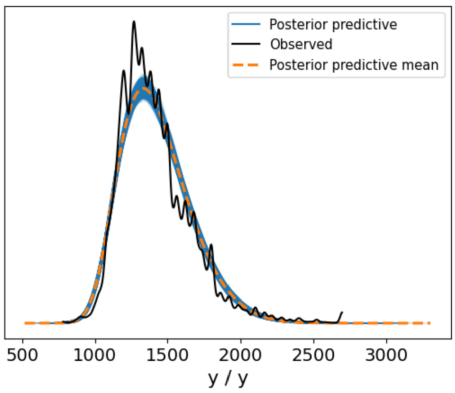


Fig. 4: 5K Posterior Predictions

Approach

In order to draw information about types of athletic activity performance, we isolated two predictor variables (5k time, deadlift weight). While this is an oversimplification of anaerobic and aerobic performance, it provided data on an athlete's strength and endurance that we could use. We then decided on a linear regression, as linear regressions are very easily interpretable, and ours was an inferential question.

Once we started making our linear regressions, we tested several model parameters and distributions before settling on our final model. We ended up giving our priors huge variance and making them very uninformed. This seemed to work well. We also tested different distributions for the likelihood prior before settling on our normal and skewed normal fits.

Results

Once the modeling was completed, we performed Markov chain Monte Carlo sampling, which converged nicely after drawing 2000 samples for the model trace. With the model trace we were able to analyze the forest, counterfactual and posterior predictive plots to determine which variables had the highest effect on outcomes. The forest plot beta values that were centered far from zero with tails that did not overlap zero indicated significant predictor variables. Comparing this to the variable list shows using a coach, less than 4 rest days a month, multiple workouts a day more than three days a week and training for over 4 years had the largest impact on deadlift weight. On the other hand, height, weight and diet seemed to have an insignificant effect (see posterior plots in Appendix A).

We analyzed the same plots and metrics for the 5k time model and found slightly different predictors were significant. Individuals with less than 4 rest days a month and multiple workouts a day more than three days a week still performed better. However, in the 5k time model eating 1-3 cheat meals per week and an athlete's BMI have significant impacts on results. Furthermore, using a coach and training for over 4 years had little to no impact on 5k time unlike with deadlift (plots in Appendix A).

The forest plots provided 94% high density interval lines through each predictor's mean to show variables' effect on the response for a range of possible values. As a result, the effect of the predictors discussed above may vary in strength. However, we only drew conclusions from predictors that were significantly far from zero so the probability they did not impact results was approximately zero.

Conclusions

Through the Crossfit Athletes kaggle dataset, we hoped to discover a variety of controllable variables related to aerobic and anaerobic fitness. Based on the analysis and modeling completed above, we achieved our goal. Bayesian linear modeling allowed us to uncover multiple significant predictor distributions which could help or harm athletes' performance. As a result, general training recommendations can be drawn from this analysis.

Unfortunately, this modeling process doesn't unlock the key to fitness for everyone. A one-size fits all program won't give everyone the results they are looking for, and customized training programs can be much more effective(perhaps this is why a coach is so helpful). Furthermore, the models were built on data which only included individuals involved in crossfit. Although this does include many "normal people", it is not a representative sample for every single athlete, and may not be applicable to them. Nevertheless, the modeling and analysis provided interesting insights.

References

“Untitled.” https://www.kaggle.com/sitemaps/5.xml (accessed: Dec. 11, 2023).

Appendix A

```
{0: 'age',
 1: 'age_squared',
 2: 'height',
 3: 'weight',
 4: 'Female',
 5: 'Male',
 6: 'eat_Decline to answer',
 7: 'eat_I eat 1-3 full cheat meals per week',
 8: "eat_I eat quality foods but don't measure the amount",
 9: 'eat_I eat strict Paleo',
 10: 'eat_I eat whatever is convenient',
 11: 'eat_I weigh and measure my food',
 12: 'train_Decline to answer',
 13: 'train_I have a coach who determines my programming',
 14: 'train_I incorporate CrossFit.com workouts',
 15: 'train_I record my workouts',
 16: 'train_I workout mostly at a CrossFit Affiliate',
 17: 'train_I workout mostly at home, work, or a traditional gym',
 18: 'train_I write my own programming',
 19: 'background_Decline to answer',
 20: 'background_I have no athletic background besides CrossFit',
 21: 'background_I played college sports',
 22: 'background_I played professional sports',
 23: 'background_I played youth or high school level sports',
 24: 'background_I regularly play recreational sports',
 25: 'experience_I began CrossFit by trying it alone (without a coach)',
 26: 'experience_I began CrossFit with a coach (e.g. at an affiliate)',
 27: 'experience_I have attended one or more specialty courses',
 28: 'experience_I have completed the CrossFit Level 1 certificate course',
 29: 'experience_I have had a life changing experience due to CrossFit',
 30: 'experience_I train other people',
 31: 'schedule_Decline to answer',
 32: 'schedule_I do multiple workouts in a day 1x a week',
 33: 'schedule_I do multiple workouts in a day 2x a week',
 34: 'schedule_I do multiple workouts in a day 3+ times a week',
 35: 'schedule_I strictly schedule my rest days',
 36: 'schedule_I typically rest 4 or more days per month',
 37: 'schedule_I typically rest fewer than 4 days per month',
 38: 'schedule_I usually only do 1 workout a day',
 39: 'howlong_1-2 years',
 40: 'howlong_2-4 years',
 41: 'howlong_4+ years',
 42: 'howlong_6-12 months',
 43: 'howlong_Decline to answer',
 44: 'howlong_Less than 6 months'}
```

Fig. 1: Deadlift Beta Dictionary

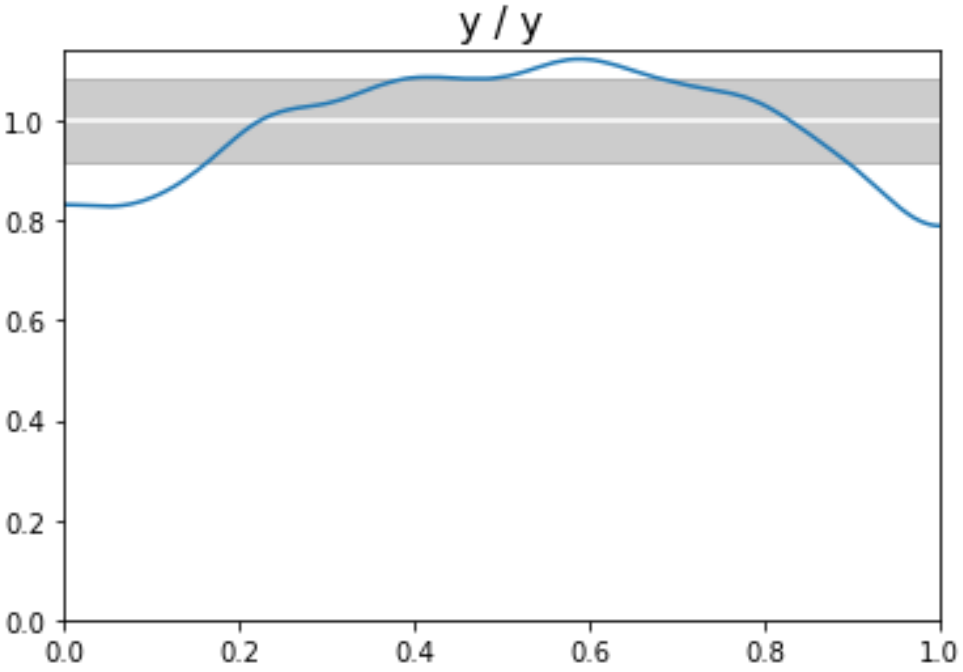


Fig. 2: Deadlift Bayesian P-value Plot

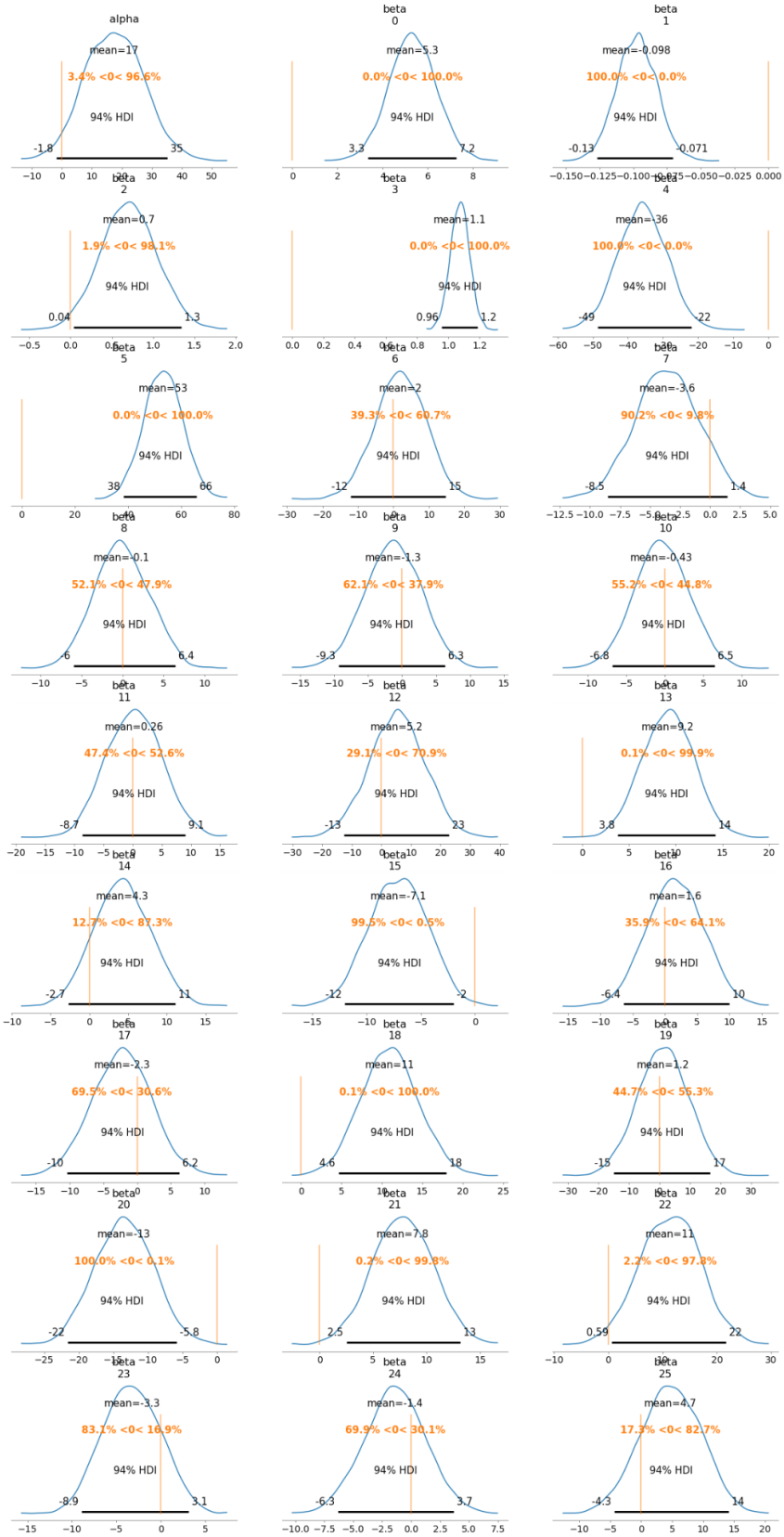


Fig. 3: Deadlift Posterior Predictive Plots (alpha, beta 0-25)

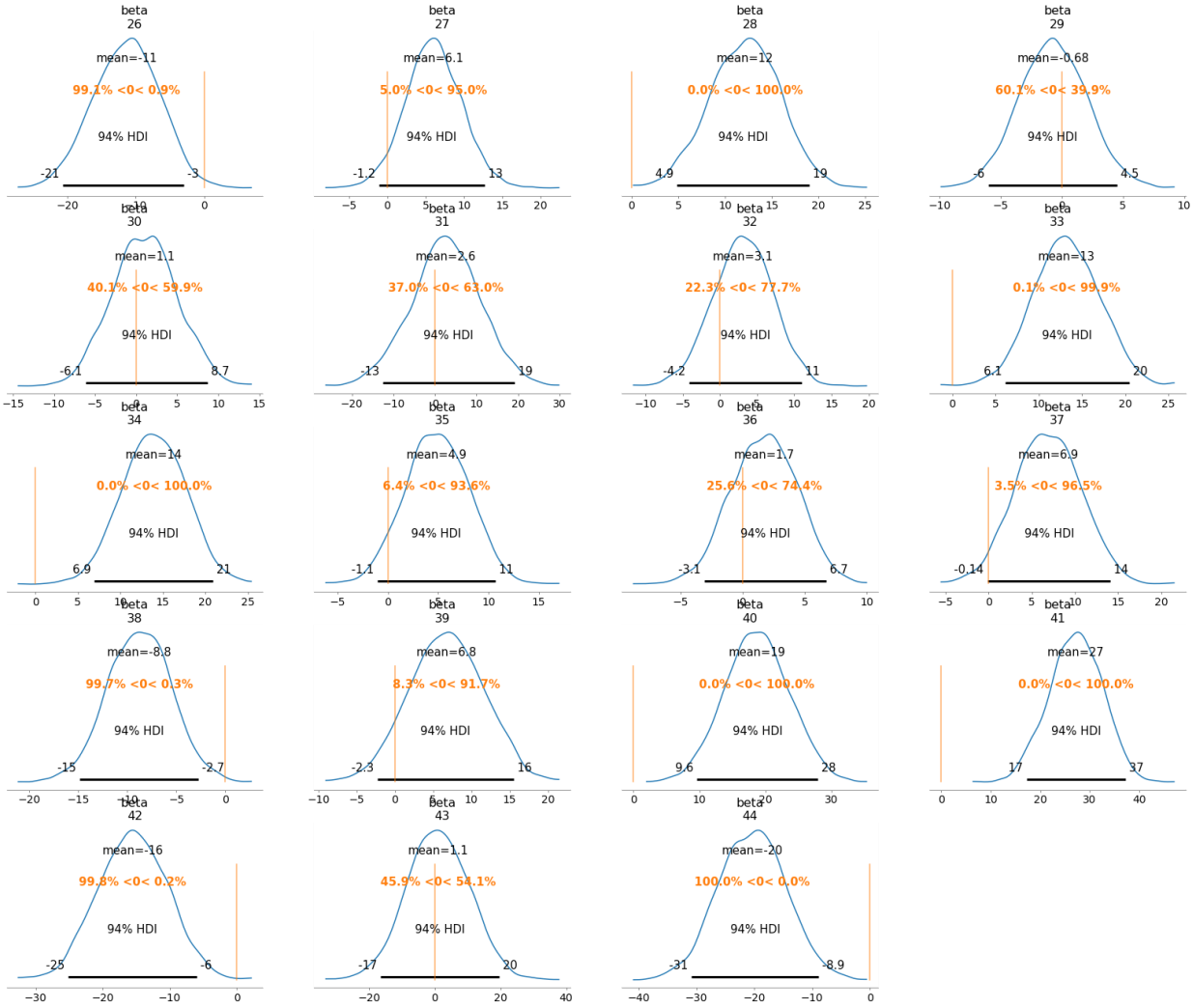


Fig. 4: Deadlift Posterior Predictive Plots (beta 26-44)

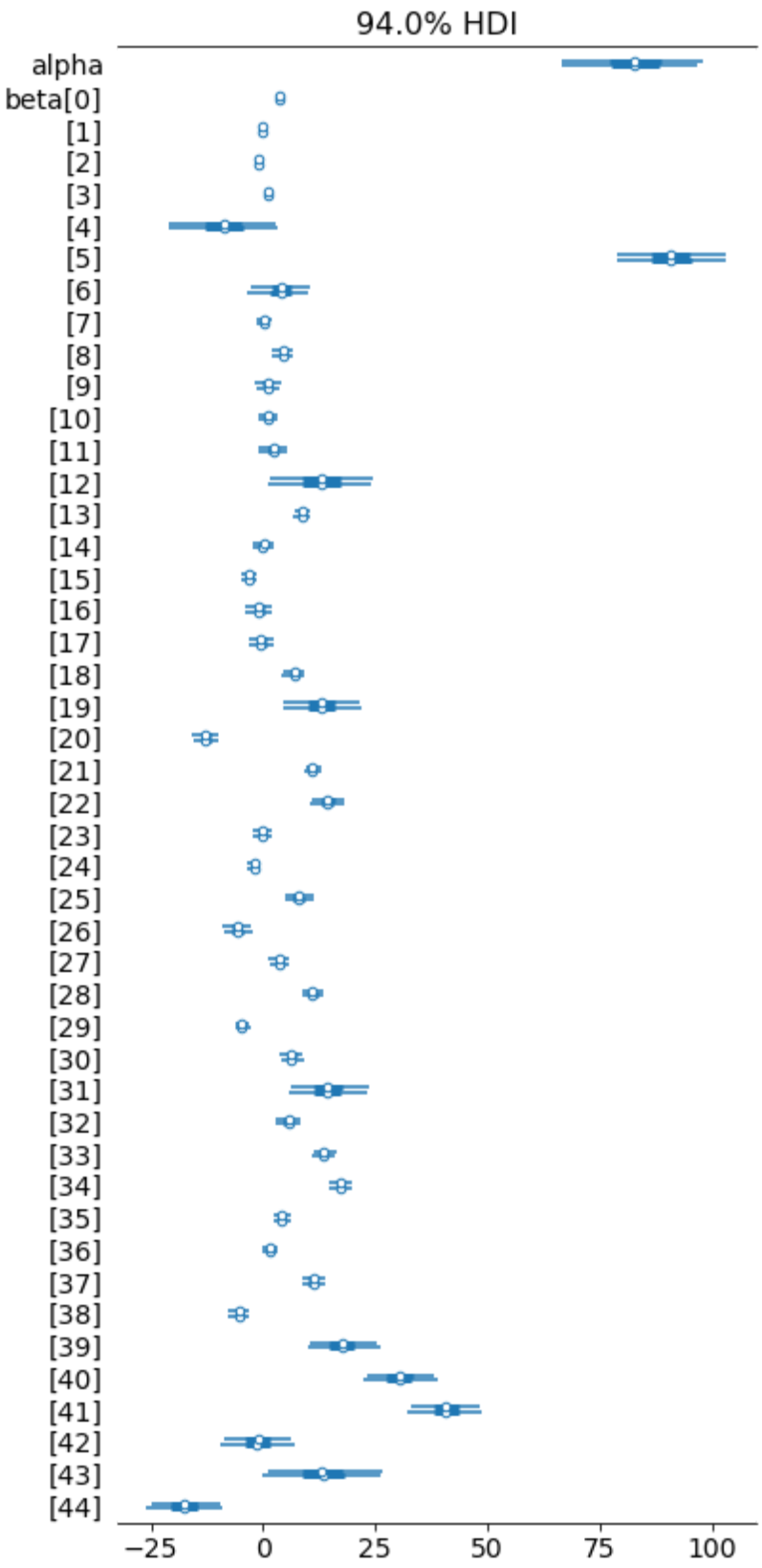


Fig. 5: Deadlift Posterior Forest Plot

```
{0: 'age',
 1: 'age_squared',
 2: 'BMI',
 3: 'Female',
 4: 'Male',
 5: 'eat_Decline to answer',
 6: 'eat_I eat 1-3 full cheat meals per week',
 7: "eat_I eat quality foods but don't measure the amount",
 8: 'eat_I eat strict Paleo',
 9: 'eat_I eat whatever is convenient',
 10: 'eat_I weigh and measure my food',
 11: 'train_Decline to answer',
 12: 'train_I have a coach who determines my programming',
 13: 'train_I incorporate CrossFit.com workouts',
 14: 'train_I record my workouts',
 15: 'train_I workout mostly at a CrossFit Affiliate',
 16: 'train_I workout mostly at home, work, or a traditional gym',
 17: 'train_I write my own programming',
 18: 'background_Decline to answer',
 19: 'background_I have no athletic background besides CrossFit',
 20: 'background_I played college sports',
 21: 'background_I played professional sports',
 22: 'background_I played youth or high school level sports',
 23: 'background_I regularly play recreational sports',
 24: 'experience_I began CrossFit by trying it alone (without a coach)',
 25: 'experience_I began CrossFit with a coach (e.g. at an affiliate)',
 26: 'experience_I have attended one or more specialty courses',
 27: 'experience_I have completed the CrossFit Level 1 certificate course',
 28: 'experience_I have had a life changing experience due to CrossFit',
 29: 'experience_I train other people',
 30: 'schedule_Decline to answer',
 31: 'schedule_I do multiple workouts in a day 1x a week',
 32: 'schedule_I do multiple workouts in a day 2x a week',
 33: 'schedule_I do multiple workouts in a day 3+ times a week',
 34: 'schedule_I strictly schedule my rest days',
 35: 'schedule_I typically rest 4 or more days per month',
 36: 'schedule_I typically rest fewer than 4 days per month',
 37: 'schedule_I usually only do 1 workout a day',
 38: 'howlong_1-2 years',
 39: 'howlong_2-4 years',
 40: 'howlong_4+ years',
 41: 'howlong_6-12 months',
 42: 'howlong_Decline to answer',
 43: 'howlong_Less than 6 months'}
```

Fig. 6: 5k Beta dictionary

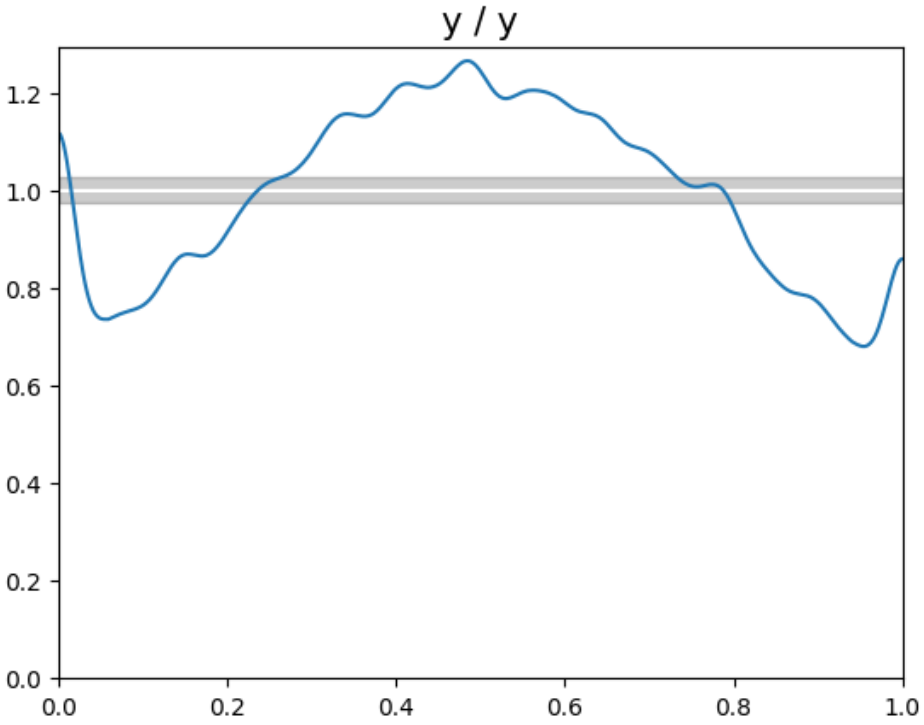


Fig. 7: 5k Bayesian P-Value Plot

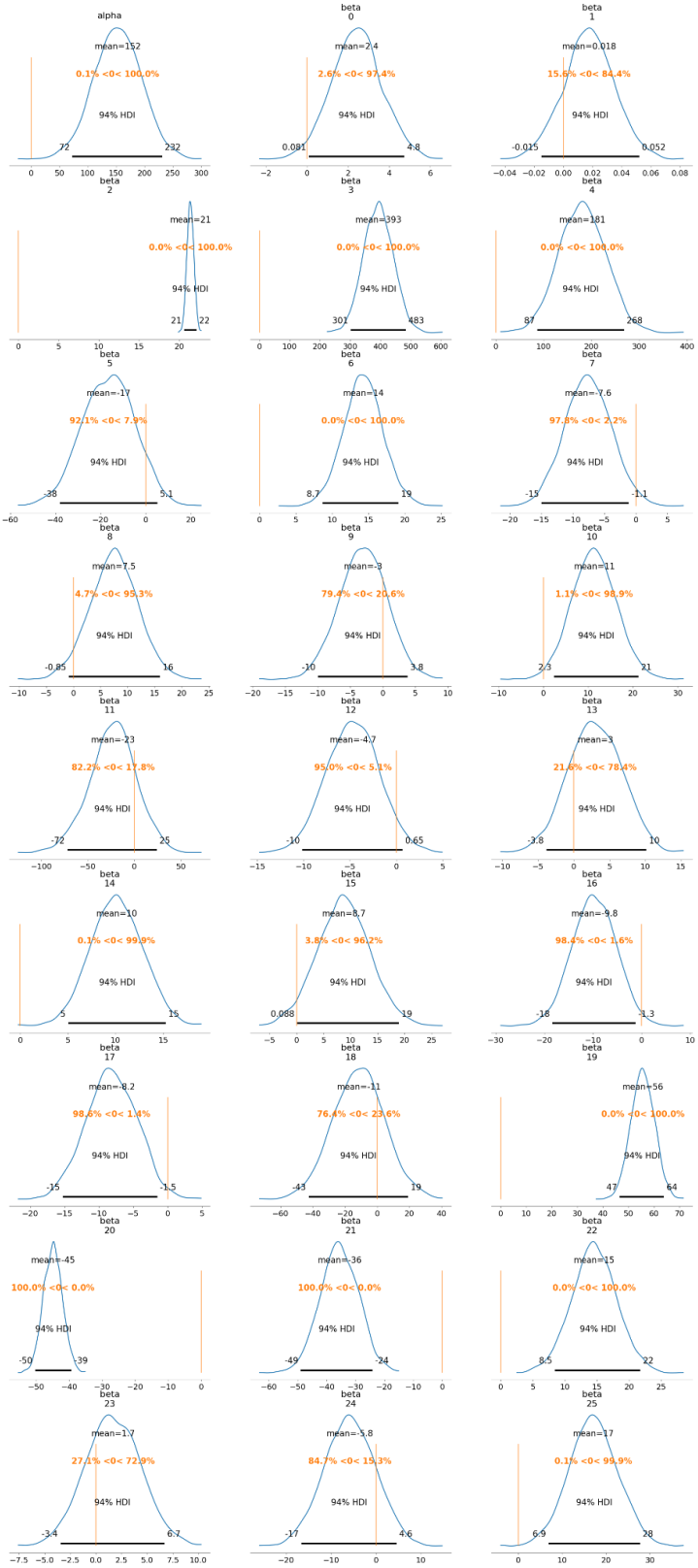


Fig. 8: 5k Posterior Predictive Plots (alpha, beta 0-25)

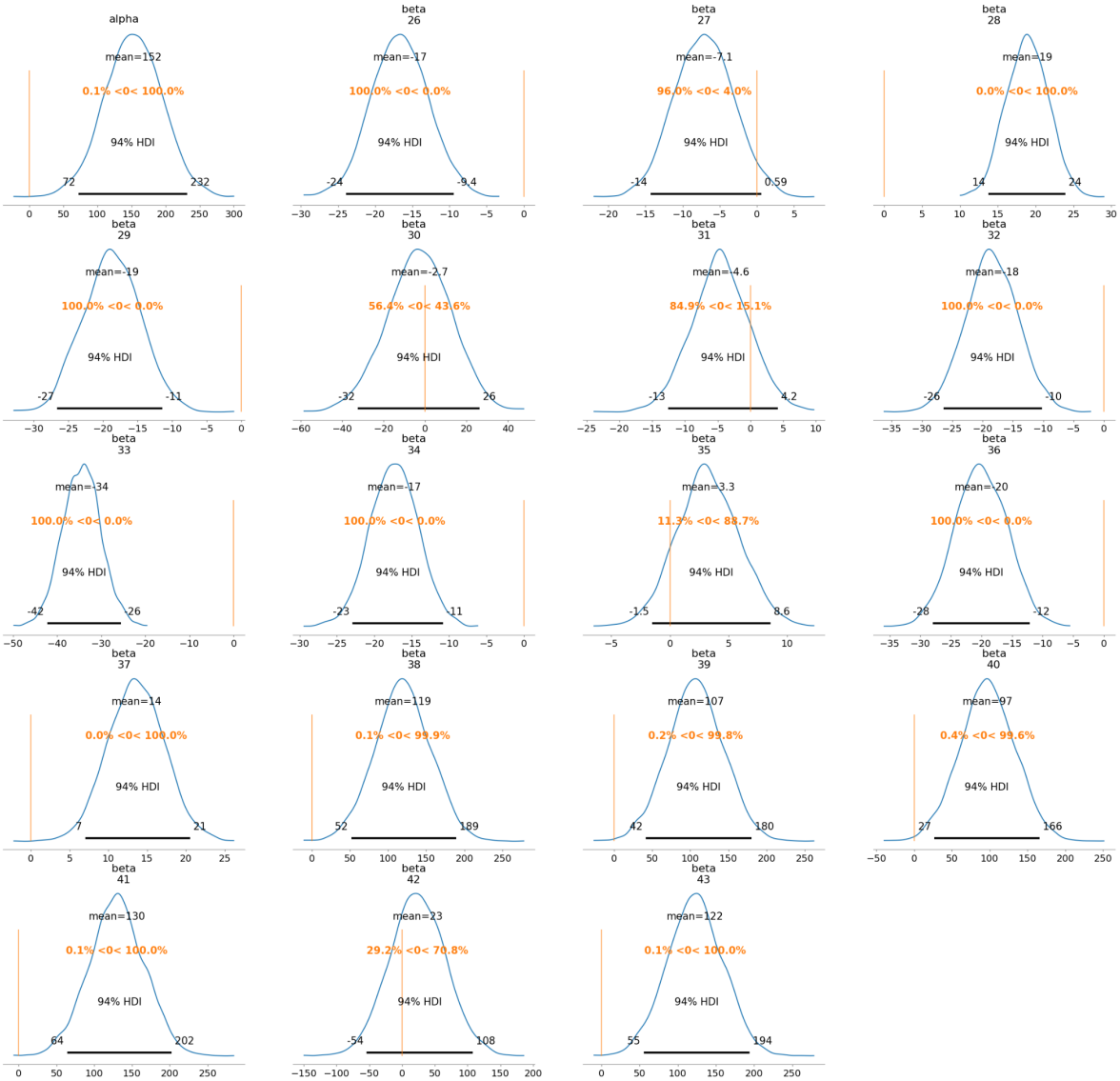


Fig. 9: 5k Posterior Predictive Plots (beta 26-44)

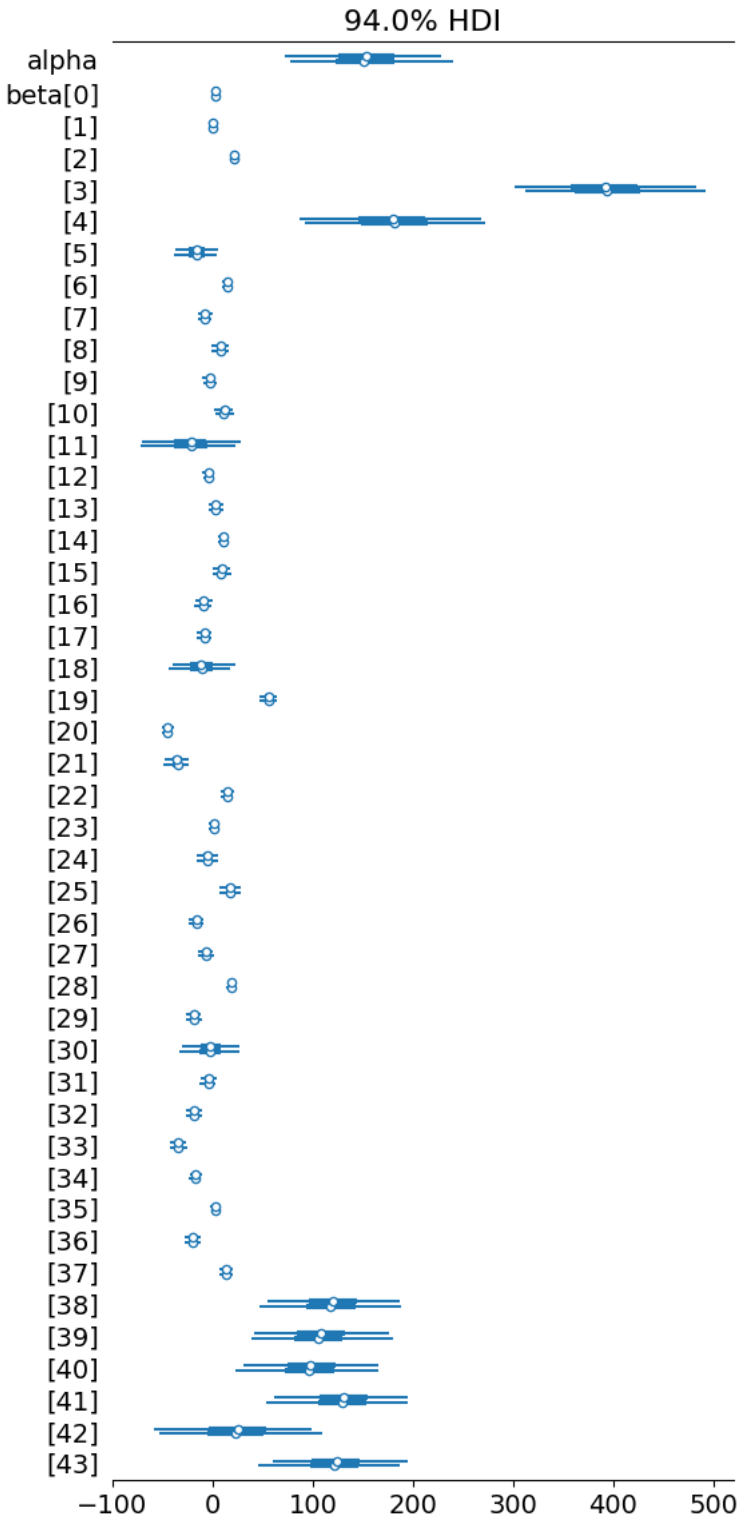


Fig. 10: 5k Posterior Forest Plot