

A Bayesian Mixture Distribution Model for Olympic Weightlifting Scores

Julia Platt

Department of Mathematical Sciences
Montana State University

May 4, 2018

A writing project submitted in partial fulfillment
of the requirements for the degree

Master of Science in Statistics

APPROVAL

of a writing project submitted by

Julia Platt

This writing project has been read by the writing project advisor and has been found to be satisfactory regarding content, English usage, format, citations, bibliographic style, and consistency, and is ready for submission to the Statistics Faculty.

Date

Andrew Hoegh
Writing Project Advisor

Date

Jennifer Green
Writing Project Advisor

Date

Mark C. Greenwood
Writing Projects Coordinator

Abstract

While performance enhancing drugs are an issue in nearly all professional sports, they are especially problematic in strength-based sports such as Olympic style weight lifting. In order to determine how much of an effect performance enhancing drug usage has on an athlete's strength, we chose to investigate how the amount of weight lifted by Olympic weight lifters differed between athletes that had tested positive for performance enhancing drugs and those who had not, after accounting for other factors such as sex, body weight, and age. An Olympic weightlifting athlete is scored based on two different events, where the score is equal to the sum of weight lifted for the two events. However if an athlete fails to successfully lift the weight with proper form in either event, he or she receives a score of zero. We analyzed data from twenty case-control pairs of athletes using a mixture model. One piece of the mixture distribution modeled the weight lifted for athletes who did not scratch using a Bayesian regression model with a random intercept for each athlete. The other portion of the distribution models the probability of scratching using a Bayesian probit regression model. Athletes who were caught using performance enhancing drugs lifted a total of 18.831 kg (\approx 41 lbs) more on average than athletes who were not caught cheating, when age, body weight, and sex were kept constant. However, there did not appear to be a difference in the probability of scratching between athletes who tested positive for performance enhancing drugs and those who did not. Because the cases (athletes caught cheating) were randomly selected, we should theoretically be able to infer any of our results back to the population of International Weightlifting Federation athletes who use performance enhancing drugs. However, since the controls (athletes who had not been caught cheating) were selected to closely match the cases and were not randomly selected, these athletes may not be entirely representative of the population of athletes who have not been caught cheating. Since this model was based on observational data, statistically we cannot conclude that performance enhancing drug usage caused the increase in the total amount of weight lifted between non-cheaters and cheaters.

A Bayesian Mixture Distribution Model for Olympic Weightlifting Scores

Julia Platt

May 2, 2018

1 Background

In the 1980s, researcher Robert Goldman surveyed elite and professional athletes about their use of performance enhancing drugs. Specifically, he asked whether each athlete would take an undetectable drug which would guarantee them a first place win if they would accept that it would kill them within five years. Over half of the athletes responded with a yes (3). Though the question Goldman posed may seem hyperbolic, performance enhancing drugs come with a whole host of side effects including depression, cancer, stroke, and heart failure (5) when used in large quantities. As the use of performance enhancing drugs has become more widespread— to the extent that countries such as Russia have instituted state-sponsored doping programs for their Olympic athletes, Goldman’s Dilemma is more relevant than ever in the world of professional sports.

While performance enhancing drugs are an issue in nearly all professional sports, they are especially problematic in strength-based sports such as powerlifting and Olympic style weight lifting. These sports are particularly prone to high rates of drug usage because strength is highly correlated with lean muscle mass (7), and performance

enhancing drugs are most commonly used to increase lean muscle and improve recovery. However, many side effects of these drugs such as hypertension and cardiomyopathy (5, 6) can negatively affect endurance, which makes them less popular for athletes that require both strength and endurance in their sport of choice. Unlike many other sports, weight lifting relies almost entirely on strength. Thus the rate of performance enhancing drug usage among weight lifters may be even higher than the rate among other elite athletes. In fact, Bob Gaynor, a powerlifting historian and a competitor himself, has estimated that up to 85% of competitive *amateur* powerlifters at the national level were using some sort of performance enhancing drug in mass quantities to improve their performance (2). This percentage could possibly be even greater for elite and professional athletes due to the additional incentives to win.

In order to determine how much of an effect performance enhancing drug usage has on an athlete's strength, we chose to investigate how the strength of Olympic weight lifters differed between athletes that had tested positive for performance enhancing drugs and those who had not (after accounting for other factors such as sex). As opposed to powerlifting, Olympic weight lifting presents itself as a better candidate to study performance enhancing drug usage, due to the rigorous drug testing procedures at the Olympic games. Though not every sample is tested for every drug, anti-doping experts from the World Anti-Doping Agency (WADA) try to predict which athletes are more likely to use certain drugs, and those athletes are tested with more scrutiny. Samples for all medal winners and nearly all weight lifting athletes are screened with more scrutiny, due to the rampant drug usage among weight lifters. Additionally, samples are stored for up to 10 years and are periodically retested for both drugs and masking agents as the accuracy of tests improves (1).

2 The Data

We investigated forty case-control pairs of Olympic weight lifters, paired based on age, sex, and weight-class at their most recent weight lifting meet. Case-control studies are a class of observational study in which two groups differing in one respect and are otherwise fairly similar to one another. Of each case-control pair, one athlete had been disqualified for performance enhancing drug usage and one had never been caught cheating. The twenty “cheaters” were selected by randomly sampling ten men and ten women from a list of Olympic athletes on the International Weightlifting Federation (IWF) website (14) who had tested positive for performance enhancing drugs and been disqualified at least once. Then each case was paired with a control of the same sex and weight class, and of similar age. These case-control pairs are listed in an appendix.

The data set was scraped from the federation’s website using the `rvest` package, pulling the html tables recorded by athlete one by one. However, the tables pulled from the IWF website recorded zero values for two different cases: when an athlete scratched at a meet and thus received a zero score and also when an athlete did not scratch but had been disqualified for performance enhancing drug usage. Though it makes sense for the official record books to record a disqualified athlete’s score as a zero, for our purposes the disqualifying score was an important observation in the data set. In most cases, we were able to pull these observations from the tables recorded by event on the IWF website. In a few of the older cases of disqualification, we had to find the numbers via other sources. This information was relatively easy to find since disqualifications, especially when an Olympic medal was recinded, are well-covered by media sources.

It is important to note that there are hundreds of drugs that can result in an athlete’s disqualification, not all of which are performance-enhancing (e.g. marijuana). The IWF

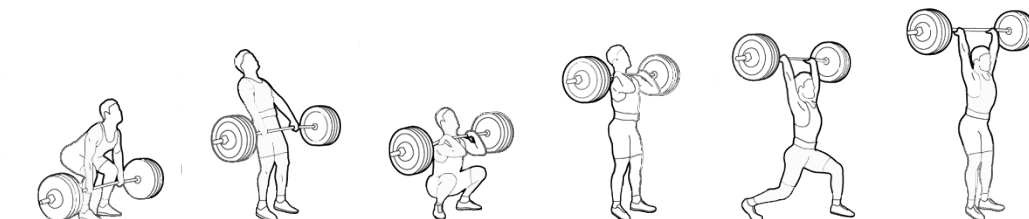
website does not always list which substance resulted in an athlete's disqualification. However, they do list which category of drug an athlete was banned for using. For our purposes, we will define a performance enhancing drug as any drug on WADA's list of banned substances which are prohibited at all times (15). This list includes anabolic androgenic agents (steroids), anti-catabolic agents, growth hormones, insulin, estrogen blockers, weight-cutting agents, diuretics, and masking agents. In order to compete, an athlete must not have used any banned substances within the last five years.

The data structure is rather unusual because of how Olympic weight lifting is scored. Scores are based on two different lifts, the snatch and the clean-and-jerk (figure 1, 2). An athlete is given three attempts at each style of lift, where he or she must lift a barbell loaded with weight plates overhead with proper form.

Figure 1: The snatch (8)



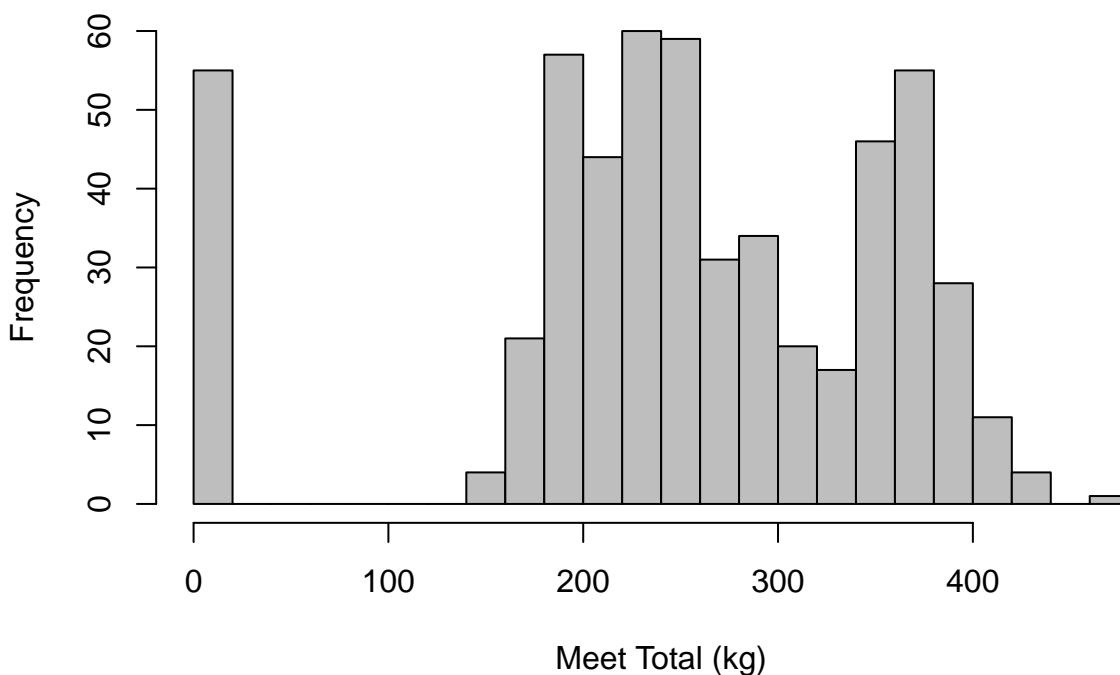
Figure 2: The clean-and-jerk (8)



The best (heaviest) successful attempt for each of the two styles of lift is recorded, and the two individual scores are added together for a total weight lifted score, recorded in kilograms. Placings are then awarded based on this meet total. However, if an athlete fails on all three attempts of one or both styles of lift, then he or she scratches

and receives a meet total of 0 (13). Hence the interesting data structure, where there are two separate pieces that make up the histogram of meet totals: the scratches and the total weight lifted when an athlete did not scratch (figure 3).

Figure 3: Histogram of Meet Totals (all athletes)



For these data, a single observation corresponds to a single meet for a particular athlete. The predictors available in this data set were the ranking at each meet, body weight as of the time of the weigh-in for that meet, the corresponding weight class the athlete competed in, the year of the meet, the athlete’s age, whether the athlete was caught cheating, and sex. If an athlete had been caught cheating, all the values of the “cheater” indicator variable are categorized as such (including the indicator for meets prior to when the athlete had been caught cheating).

Since the variables for body weight and weight class included very similar information, we chose to work with body weight (a quantitative predictor) rather than use the weight class (a categorical variable) because it contained more specific information. Additionally,

we wanted to account for how an athlete's performance changed over time by including age as a predictor. With a brief preliminary data exploration, we can see that there is visual evidence to suggest a quadratic trend in these data for both the age and body weight predictors (figure 4, 5).

Figure 4: Women's 75+ kg Weight Class

Cheater

Non-cheater

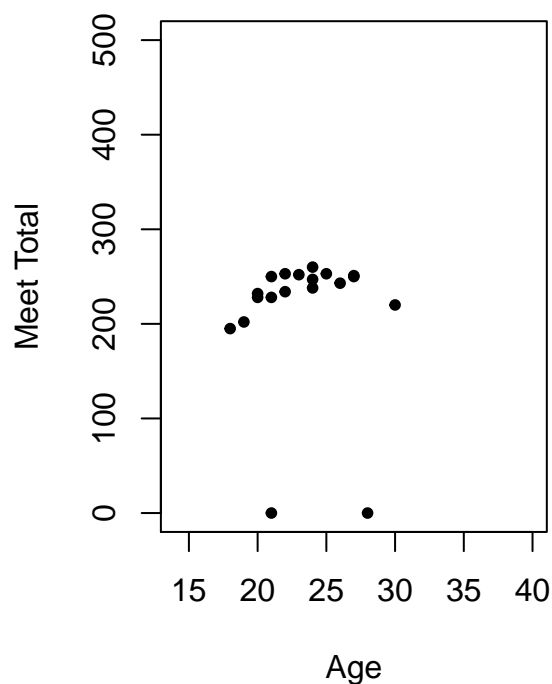
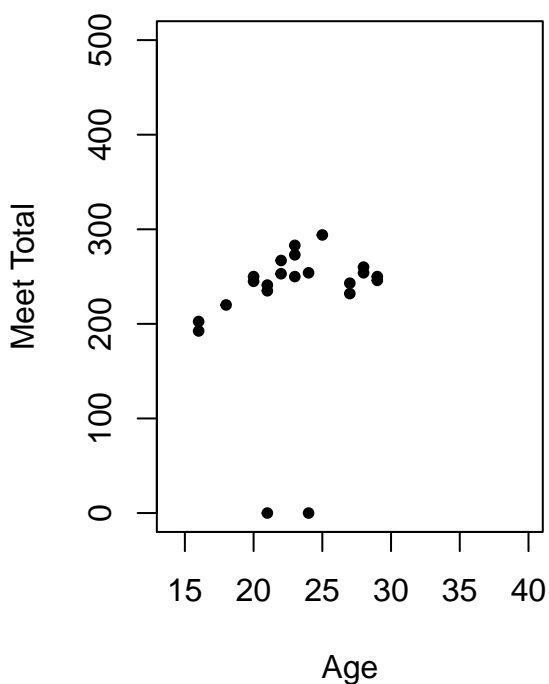
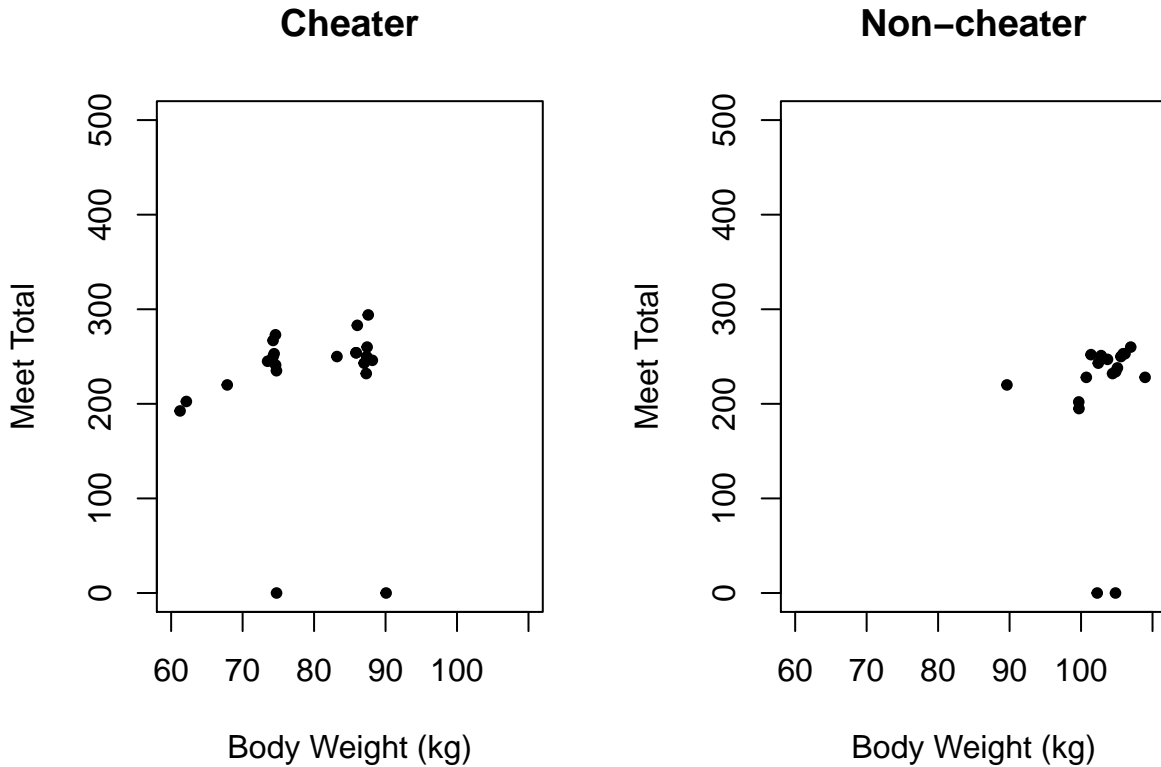


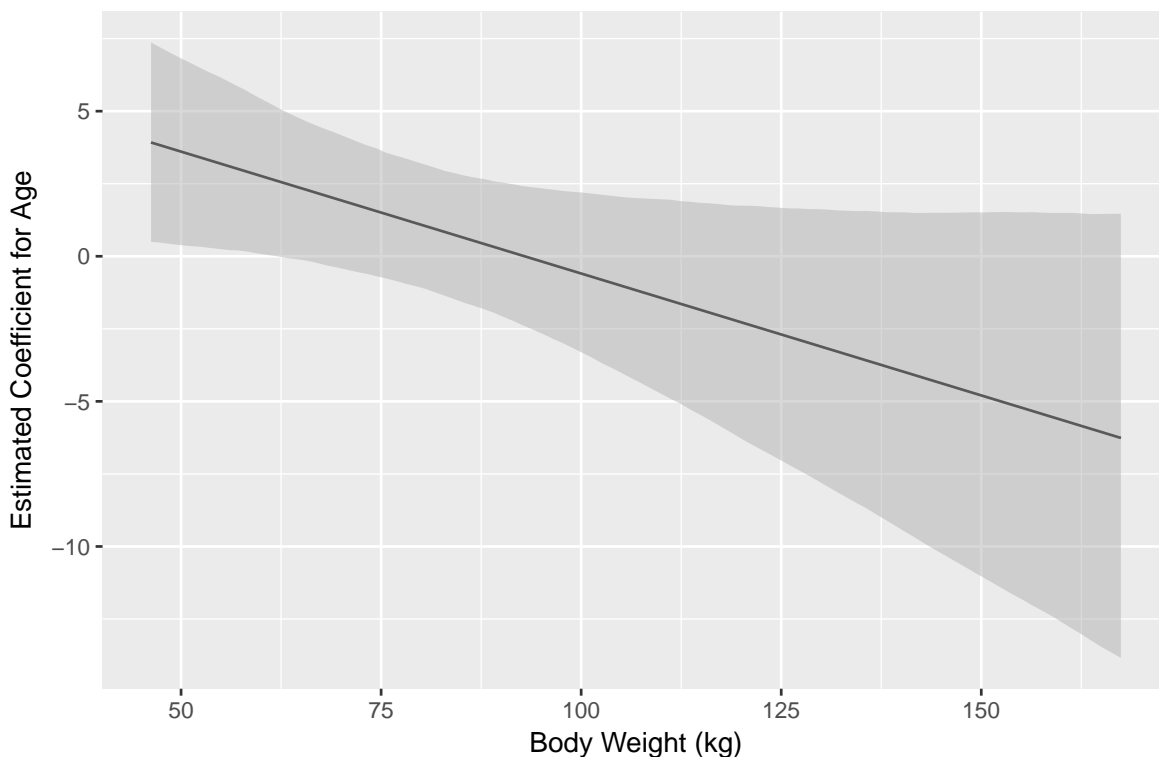
Figure 5: Women's 75+ kg Weight Class



Thus we chose to include age, a quadratic age term, bodyweight, and a quadratic bodyweight term in our initial model in addition to sex and whether or not the athlete had been caught cheating. The last two predictors are included in the model due to the large differences in natural testosterone levels between men and women and because the effect of the “cheater” variable is what we are interested in investigating.

Additionally, we can explore the interaction effect between age and body weight. In figure 6, the magnitude of the coefficient for age on the meet total decreases (along the y-axis) as body weight increases along the (x-axis), and the confidence band contains zero across most of the plot. Thus, there may be some weak evidence of an interaction effect, but we chose not to investigate it.

Figure 6: Interaction Plot (Body Weight and Age)



3 Models

These data can be modeled using mixture model, which is a probabilistic model which represents an overall population by combining probability distributions of random variables that represent different subpopulations. There are several common examples of mixture models, such as zero-inflated Poisson models. Zero-inflated Poisson models account for data that are consistent with a Poisson distribution but have an excess of zero-counts. There are two components to this zero-inflating processes, one in which a binary distribution is used to generate zeros and one in which the Poisson model is used to generate other counts, some of which are also zeros. While there is some overlap in the two pieces of the zero-inflated Poisson model (i.e. a zero count can be generated by either piece of the distribution), our data can be modeled by a mixture distribution in

two completely separate pieces because an athlete will never attempt a lift with zero pounds:

$$\text{total}_{ij} = \begin{cases} \beta_{0i} + \beta_1 \text{age}_{ij} + \beta_{11} \text{age}_{ij}^2 + \beta_2 \text{weight}_{ij} + \beta_{22} \text{weight}_{ij}^2 + \beta_3 \text{cheater}_{ij} + \beta_4 \text{sex}_{ij} + \epsilon_{ij} \\ 0 \end{cases} \quad \text{with probability } \theta$$

for the i^{th} athlete at his or her j^{th} competition and where β_{0i} represents a random intercept for the i^{th} athlete.

The probability of scratching at a meet θ can be modeled using probit regression:

$$\text{probit}(\theta = Pr(Z_{ij} = 1|X)) = \beta_{5i} + \beta_6 \text{age}_{ij} + \beta_{66} \text{age}_{ij}^2 + \beta_7 \text{weight}_{ij} + \beta_{77} \text{weight}_{ij}^2 + \beta_8 \text{cheater}_{ij} + \beta_9 \text{sex}_{ij}$$

where we define the indicator variable

$$Z_{ij} = \begin{cases} 1 & \text{scratch} \\ 0 & \text{doesn't scratch} \end{cases}$$

4 Methods

4.1 Meet Total: Random Intercept Model

The first piece of the model

$$y_{ij} = \beta_{0i} + \beta_1 \text{age}_{ij} + \beta_{11} \text{age}_{ij}^2 + \beta_2 \text{weight}_{ij} + \beta_{22} \text{weight}_{ij}^2 + \beta_3 \text{cheater}_{ij} + \beta_4 \text{sex}_{ij} + \epsilon_{ij}$$

has the following form:

$$Y_{ij} \sim MVN(X\tilde{\beta}, \sigma^2 I)$$

$$\text{where } X = \begin{bmatrix} 1 & 0 & \dots & 0 & x_{1,11} & x_{1,11}^2 & x_{2,11} & x_{2,11}^2 & x_{3,11} & x_{4,11} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & \dots & 0 & x_{1,1n_1} & x_{1,1n_1}^2 & x_{2,1n_1} & x_{2,1n_1}^2 & x_{3,1n_1} & x_{4,1n_1} \\ 0 & 1 & \dots & 0 & x_{1,21} & x_{1,21}^2 & x_{2,21} & x_{2,21}^2 & x_{3,21} & x_{4,21} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & \dots & 0 & x_{1,2n_2} & x_{1,2n_2}^2 & x_{2,2n_2} & x_{2,2n_2}^2 & x_{3,2n_2} & x_{4,2n_2} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & x_{1,401} & x_{1,401}^2 & x_{2,401} & x_{2,401}^2 & x_{3,401} & x_{4,401} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & \dots & 1 & x_{1,40n_{40}} & x_{1,40n_{40}}^2 & x_{2,40n_{40}} & x_{2,40n_{40}}^2 & x_{3,40n_{40}} & x_{4,40n_{40}} \end{bmatrix}$$

and n_1, \dots, n_{40} are the respective sample sizes for the 40 athletes, y represents the meet

total, x_1 represents age, x_2 represents body weight, $x_3 = \begin{cases} 1 & \text{cheater} \\ 0 & \text{not caught cheating} \end{cases}$, and

$$x_4 = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}.$$

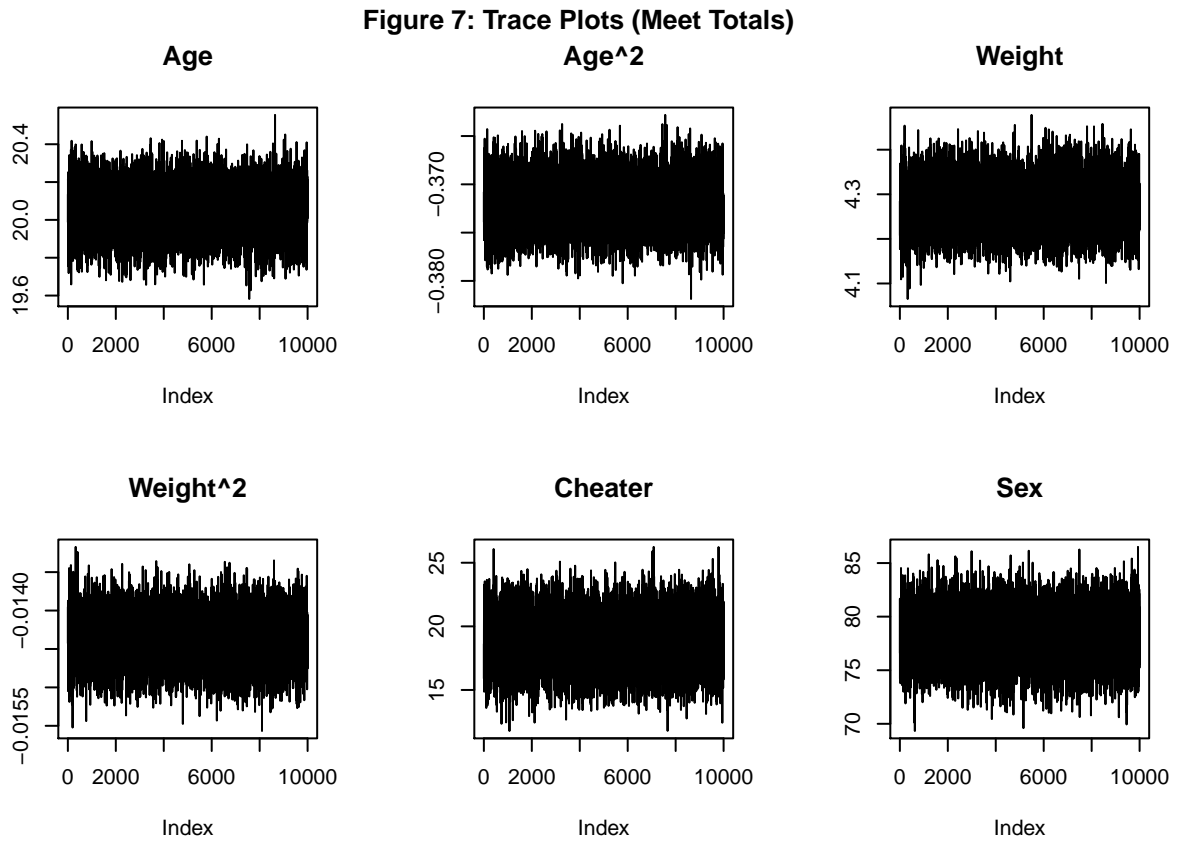
The priors for this model are as follows:

$$\beta_{0i} \sim N(\alpha, \delta^2)$$

$$\tilde{\beta} \sim MVN(\tilde{\beta}_0, \Sigma_0)$$

$$\sigma^2 \sim IG\left(\frac{\nu_0}{2}, \frac{\nu_0 \sigma_0^2}{2}\right)$$

The trace plots for each of the coefficients can be seen in the figures below (figure 7).



The coefficient values and their 95% credible intervals (table 1) indicate that all

of the predictors in this portion of the model are useful in determining an athlete’s meet total, due to the fact that none of the credible intervals contain zero. We are most interested in the “cheater” effect on the meet total, which has a coefficient of 18.831. Athletes who were caught using performance enhancing drugs lifted a total of 18.831 kg (\approx 41 lbs) more on average than athletes who were not caught cheating, when age, body weight, and sex were kept constant. This difference in meet total is a substantial amount, given that weight lifting meets are often won with scores less than five kilograms apart.

Table 1: Random Intercept Model

Parameter	Estimate	2.5%	97.5%
Age	20.045	19.8158	20.2803
Weight	4.278	4.1772	4.3805
Age ²	-0.372	-0.3765	-0.3673
Weight ²	-0.014	-0.015	-0.0138
Cheater	18.831	15.0188	22.6112
Sex	78.166	73.7823	82.5236

4.2 Scratches: Random Intercept Probit Regression Model

For the second piece of the model, we begin by defining $z_{ij} = \begin{cases} 1 & \text{scratch} \\ 0 & \text{doesn't scratch} \end{cases}$.

Then we define the probability of scratching as $\theta = Pr(Z_{ij} = 1|X)$

where X is the same as above and n_1, \dots, n_{40} are the respective sample sizes for the 40

athletes, x_1 represents age, x_2 represents body weight, $x_3 = \begin{cases} 1 & \text{cheater} \\ 0 & \text{not caught cheating} \end{cases}$,

and $x_4 = \begin{cases} 1 & \text{male} \\ 0 & \text{female} \end{cases}$.

The model has the following form:

$$\phi^{-1}(\theta) = \beta_{5i} + \beta_6 \text{age}_{ij} + \beta_{66} \text{age}_{ij}^2 + \beta_7 \text{weight}_{ij} + \beta_{77} \text{weight}_{ij}^2 + \beta_8 \text{cheater}_{ij} + \beta_9 \text{sex}_{ij}$$

where the function ϕ^{-1} is the inverse CDF of a standard normal distribution

$$Z_{ij}^* \sim MVN(X\tilde{\beta}, I)$$

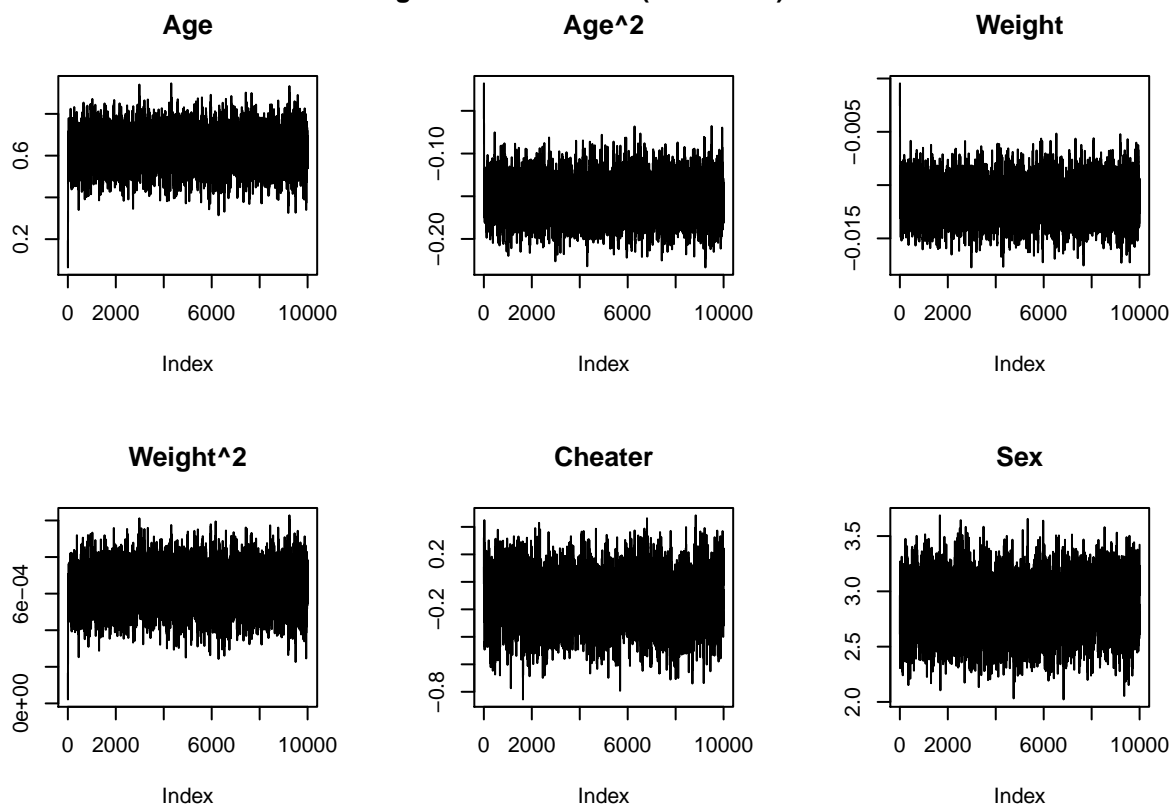
for the i^{th} athlete at his or her j^{th} competition and where Z_{ij}^* is the latent trait where $Z_{ij} = 1$ if the underlying trait $Z_{ij}^* > 0$.

The model has the following prior:

$$\tilde{\beta} \sim MVN(\tilde{\beta}_0, \sigma_0^2 I)$$

The trace plots for the regression coefficients are shown below (figure 8).

Figure 8: Trace Plots (Scratches)



For this piece of the model, there did not appear to be a “cheater” effect on the probability of an athlete scratching at a meet. However, all the other predictors appeared to be useful in predicting a scratch since those credible intervals did not contain zero (table 2).

Table 2: Probit Regression Model

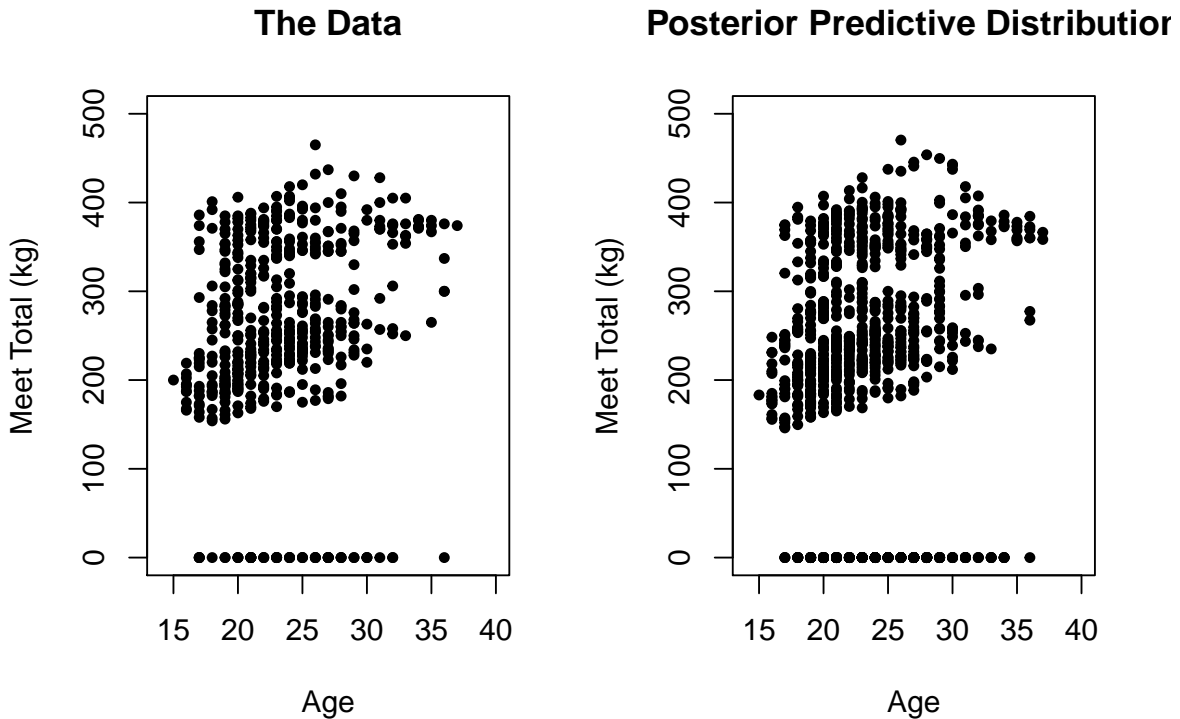
Parameter	Estimate	2.5%	97.5%
Age	0.631612	0.464836	0.791319
Weight	-0.15253	-0.195164	-0.108869
Age ²	-0.011417	-0.014722	-0.007975
Weight ²	0.00063	0.000416	0.000838
Cheater	-0.135898	-0.493308	0.206814

Parameter	Estimate	2.5%	97.5%
Sex	2.860543	2.422033	3.294198

4.3 Posterior Predictive Distribution

When comparing the original data and the posterior predictive distribution, we can see that the two plots look fairly similar and thus the model is doing a reasonably good job of describing these data (figure 9).

Figure 9: The Data versus the Posterior Predictive Distribution

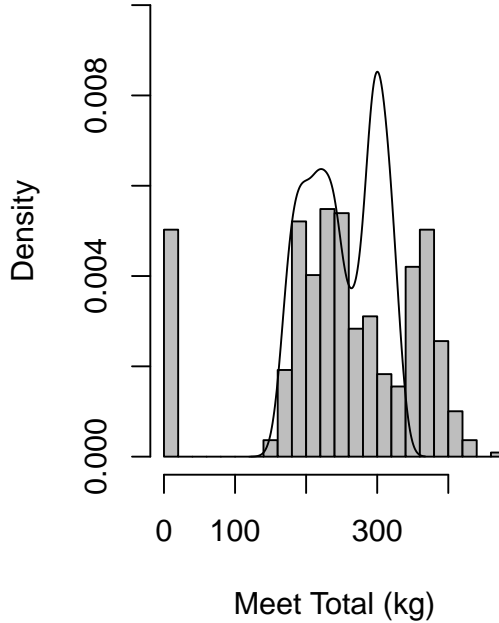
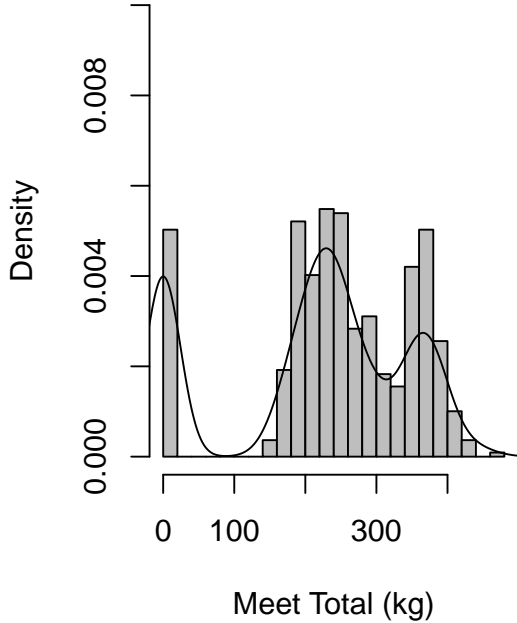


We can see the importance of accounting for the scratches with a mixture distribution since the model that fails to account for scratching does not fit the data nearly as well as the mixture distribution does. This result can be seen below (figure 10), where the posterior predictive distribution overlays the histogram of the data.

Figure 10: Why use a mixture distribution?

Mixture Distribution

Not Accounting for Scratches



5 Implications and Conclusions

Because the cases (athletes caught cheating) were randomly selected, we should theoretically be able to infer any of our results back to the population of IWF athletes who use performance enhancing drugs. However, as noted in the introduction, not every sample is tested for every drug, but certain athletes are tested with more scrutiny. According to WADA, athletes who perform well are listed as high-risk for doping and are tested at least three times in 18 months (16). This method of selection creates an interesting phenomenon where athletes who use performance enhancing drugs and perform well are more likely to get caught than athletes who use and yet do not perform well. Due to this potential bias we may be overestimating the “cheater” effect in our model.

It is also important to note that the controls (athletes who had not been caught cheating) could potentially have used performance enhancing drugs in the past or could be using them currently as well. The IWF allows athletes who have not tested positive within the last five years to compete in their organization (15). Additionally, drug tests for performance enhancing drugs are notorious for their high type II error rate (16), due to new drugs that do not yet have tests, artificial urine, cleanse drinks, bribes, and sample tampering (as was the case in Russia's state-sponsored doping scandal). Even though IWF registered athletes can still be subject to random drug tests outside of the competition season, the rate of drug tests and the scrutiny of the testing procedures are not as intense as they are during competition. As such, athletes tend to use performance enhancing drugs during their training season and cycle off of them just in time to test negative for competition, while still experiencing the performance-enhancing effects of their training season cycle. These controls also were not randomly sampled from the population of IWF athletes because they were selected to closely match the cases. Thus, these athletes may not be entirely representative of the population of athletes who have not been caught cheating.

Since this model was based on observational data and there was no option to randomly assign treatments, statistically we cannot conclude that performance enhancing drug usage caused the increase in the meet total between non-cheaters and cheaters.

There are many routes we could take this project from here. In the future, we could use Bayesian model comparison techniques to assess which predictors should be included in each model. Additionally, we could include a random slope for each athlete, in addition to the random intercept. Given the large the cheater effect and the large number of WADA doping violations that occur each year, there is clearly an issue with performance enhancing drug usage among elite athletes and a need to study these

effects.

6 Sources

- (1) Associated Press. “How It Works: A Look at the Retests of Olympic Drug Samples.” NBC Olympics, www.nbcolympics.com/news/how-it-works-look-retests-olympic-drug-samples.
- (2) Easter, Michael. “What Powerlifting Tells Us About The Effects Of PEDs.” FiveThirtyEight, FiveThirtyEight, 9 Dec. 2014, fivethirtyeight.com/features/what-powerlifting-tells-us-about-the-effects-of-peds/.
- (3) Reynolds, Gretchen. “Phys Ed: Will Olympic Athletes Dope If They Know It Might Kill Them?” The New York Times, The New York Times, 20 Jan. 2010, well.blogs.nytimes.com/2010/01/20/phys-ed-will-olympic-athletes-dope-if-they-know-it-might-kill-them/.
- (4) “9 Olympians, Including 6 Medallists, Caught for Beijing Doping | CBC Sports.” CBCnews, CBC/Radio Canada, www.cbc.ca/sports/olympics/ioc-doping-olympians-beijing-1.3822355.
- (5) Wedro, Benjamin. “Steroids: Abuse & Side Effects for Performance Enhancing Drugs.” EMedicineHealth, www.emedicinehealth.com/steroids/article_em.htm#what_are_the_side_effects_of_prolonged_steroid_abuse.

- (6) “Doping: Do You Know the Facts about Performance-Enhancing Drugs?” Mayo Clinic, Mayo Foundation for Medical Education and Research, 15 Oct. 2015, www.mayoclinic.org/healthy-lifestyle/fitness/in-depth/performance-enhancing-drugs/art-20046134.
- (7) Melekoglu Tuba, Ocal Defne, Badil Bilgehan, Sonmez M.. “Muscle Strength in Relation to Body Composition in the Turkish Male National Judo Team.” [Http://Www.analefefs.ro](http://www.analefefs.ro), 2012, www.analefefs.ro/anale-fefs/2012/issue-2/pe-autori/11.pdf.
- (8) “Olympic Lifts.” Canadian Masters Weightlifting Federation, cdnmastersweightlifting.org/olympic-lifts/.
- (9) Albert, James H, and Chibb Siddhartha. Bayesian Analysis of Binary and Polychotomous Response Data . American Statistical Association, www.stat.cmu.edu/~brian/905-2009/all-papers/albert-chib-1993.pdf.
- (10) Erwin, Mark E. Normal Random Effects Model. www.markirwin.net/stat220/Lecture/Lecture12.pdf.
- (11) Bayes Shortcourse. Virginia Tech, www.lisa.stat.vt.edu/sites/default/files/Bayes%20Shortcourse%202011.pdf.
- (12) “Bayesian Simple Linear Regression with Gibbs Sampling in R.” R-Bloggers, 7 Aug. 2017, www.r-bloggers.com/bayesian-simple-linear-regression-with-gibbs-sampling-in-r/.

- (13) “2018 Rules and Regulations.” International Weightlifting Federation, http://www.iwf.net/wp-content/uploads/downloads/2018/01/IWF-TCRR_01012018.pdf.
- (14) International Weightlifting Federation, www.iwf.net/.
- (15) “What Is Prohibited.” World Anti-Doping Agency, www.wada-ama.org/en/content/what-is-prohibited/prohibited-at-all-times/.
- (16) Romano, John. “Beating the CrossFit Drug Test.” T NATION, www.t-nation.com/pharma/beating-the-crossfit-drug-test.

7 Appendix of R Code

```
library(rvest)
library(stringr)
require(xml2)
require(purrr)
require(LearnBayes)
require(mvtnorm)
require(MCMCpack)
require(knitr)
require(nlme)
require(utils)
library(interplot)

# Men
artykov <-
  read_html("http://www.iwf.net/results/athletes/?athlete=artykov-izzat-
            1993-09-08&id=2770")
A1 <- artykov %>% html_nodes('table') %>% html_table()
a1 <- data.frame(A1)
```

```

a1$Year <- a1$Var.1
a1$Age <- a1$Year - 1993
a1 <- a1[, 2:10]
a1$Total[1] <- 339
a1$Snatch[1] <- 151
a1$Cl.Jerk[1] <- 188
a1$ID <- 1
a1$PED <- 1
a1$pairID <- "a"

sanchez.lopez <-
  read_html("http://www.iwf.net/results/athletes/?athlete=sanchez-lopez-david
            -1994-07-20&id=11088")
A2 <- sanchez.lopez %>% html_nodes('table') %>% html_table()
a2 <- data.frame(A2)
a2$Year <- a2$Var.1
a2$Age <- a2$Year - 1994
a2 <- a2[, 2:10]
a2$ID <- 2
a2$PED <- 0
a2$pairID <- "a"

martirosyan <-
  read_html("http://www.iwf.net/results/athletes/?athlete=martirosyan-tigran
            --1988-06-09&id=2583")
B1 <- martirosyan %>% html_nodes('table') %>% html_table()
b1 <- data.frame(B1)
b1$Year <- b1$Var.1
b1$Age <- b1$Year - 1988
b1 <- b1[, 2:10]
b1$Total[10] <- 338
b1$Snatch[10] <- 153
b1$Cl.Jerk[10] <- 185
b1$ID <- 3
b1$PED <- 1
b1$pairID <- "b"

minasidis <-
  read_html("http://www.iwf.net/results/athletes/?athlete=minasidis-dimitris
            -1989-04-29&id=2168")
B2 <- minasidis %>% html_nodes('table') %>% html_table()
b2 <- data.frame(B2)

```



```

b2$Year <- b2$Var.1
b2$Age <- b2$Year - 1989
b2 <- b2[, 2:10]
b2$ID <- 4
b2$PED <- 0
b2$pairID <- "b"

perepechenov <-
  read_html("http://www.iwf.net/results/athletes/?athlete=perepechenov-oleg
            -1975-09-06&id=1196")
C1 <- perepechenov %>% html_nodes('table') %>% html_table()
c1 <- data.frame(C1)
c1$Year <- c1$Var.1
c1$Age <- c1$Year - 1975
c1 <- c1[, 2:10]
c1$Total[6] <- 365
c1$Snatch[6] <- 170
c1$Cl.Jerk[6] <- 195
c1$ID <- 5
c1$PED <- 1
c1$pairID <- "c"

maposua <-
  read_html("http://www.iwf.net/results/athletes/?athlete=maposua-uati-1976
            -07-26&id=274")
C2 <- maposua %>% html_nodes('table') %>% html_table()
c2 <- data.frame(C2)
c2$Year <- c2$Var.1
c2$Age <- c2$Year - 1976
c2 <- c2[, 2:10]
c2$ID <- 6
c2$PED <- 0
c2$pairID <- "c"

sinraian <-
  read_html("http://www.iwf.net/results/athletes/?athlete=sinraian-gabriel-1988
            -12-21&id=1454")
D1 <- sinraian %>% html_nodes('table') %>% html_table()
d1 <- data.frame(D1)
d1$Year <- d1$Var.1
d1$Age <- d1$Year - 1988
d1 <- d1[, 2:10]

```

```

d1$Total[1] <- 390
d1$Snatch[1] <- 173
d1$Cl.Jerk[1] <- 217
d1$ID <- 7
d1$PED <- 1
d1$pairID <- "d"

bardis <-
  read_html("http://www.iwf.net/results/athletes/?athlete=bardis-giovanni-battista-
            1987-05-21&id=7889")
D2 <- bardis %>% html_nodes('table') %>% html_table()
d2 <- data.frame(D2)
d2$Year <- d2$Var.1
d2$Age <- d2$Year - 1987
d2 <- d2[, 2:10]
d2$ID <- 8
d2$PED <- 0
d2$pairID <- "d"

aukhadov <-
  read_html("http://www.iwf.net/results/athletes/?athlete=aukhadov-apti-1992-11-
            18&id=7218")
E1 <- aukhadov %>% html_nodes('table') %>% html_table()
e1 <- data.frame(E1)
e1$Year <- e1$Var.1
e1$Age <- e1$Year - 1992
e1 <- e1[, 2:10]
e1$Total[6] <- 385
e1$Snatch[6] <- 210
e1$Cl.Jerk[6] <- 175
e1$ID <- 9
e1$PED <- 1
e1$pairID <- "e"

rostami <-
  read_html("http://www.iwf.net/results/athletes/?athlete=rostami-kianoush-1991-
            07-23&id=2249")
E2 <- rostami %>% html_nodes('table') %>% html_table()
e2 <- data.frame(E2)
e2$Year <- e2$Var.1
e2$Age <- e2$Year - 1991
e2 <- e2[, 2:10]

```

```

e2$ID <- 10
e2$PED <- 0
e2$pairID <- "e"

rybakou <-
  read_html("http://www.iwf.net/results/athletes/?athlete=rybakou-andrei-1982-
            03-04&id=264")
F1 <- rybakou %>% html_nodes('table') %>% html_table()
f1 <- data.frame(F1)
f1$Year <- f1$Var.1
f1$Age <- f1$Year - 1982
f1 <- f1[, 2:10]
f1$Total[5] <- 394
f1$Snatch[5] <- 185
f1$Cl.Jerk[5] <- 209
f1$ID <- 11
f1$PED <- 1
f1$pairID <- "f"

martirosyan2 <-
  read_html("http://www.iwf.net/results/athletes/?athlete=martirosyan-tigran-1983-
            03-03&id=1246")
F2 <- martirosyan2 %>% html_nodes('table') %>% html_table()
f2 <- data.frame(F2)
f2$Year <- f2$Var.1
f2$Age <- f2$Year - 1983
f2 <- f2[, 2:10]
f2$ID <- 12
f2$PED <- 0
f2$pairID <- "f"

ilyin <-
  read_html("http://www.iwf.net/results/athletes/?athlete=ilyin-ilya-1988-05-
            24&id=7895")
G1 <- ilyin %>% html_nodes('table') %>% html_table()
g1 <- data.frame(G1)
g1$Year <- g1$Var.1
g1$Age <- g1$Year - 1988
g1 <- g1[, 2:10]
g1$Total[3] <- 418
g1$Snatch[3] <- 185
g1$Cl.Jerk[3] <- 233

```

```

g1$Total[6] <- 406
g1$Snatch[6] <- 180
g1$Cl.Jerk[6] <- 226
g1$ID <- 13
g1$PED <- 1
g1$pairID <- "g"

anttiroiko <-
  read_html("http://www.iwf.net/results/athletes/?athlete=anttiroiko-miika-matias-
            1988-11-20&id=1492")
G2 <- anttiroiko %>% html_nodes('table') %>% html_table()
g2 <- data.frame(G2)
g2$Year <- g2$Var.1
g2$Age <- g2$Year - 1988
g2 <- g2[, 2:10]
g2$ID <- 14
g2$PED <- 0
g2$pairID <- "g"

ciricu <-
  read_html("http://www.iwf.net/results/athletes/?athlete=ciricu-anatoli-1988-
            09-14&id=3456")
H1 <- ciricu %>% html_nodes('table') %>% html_table()
h1 <- data.frame(H1)
h1$Year <- h1$Var.1
h1$Age <- h1$Year - 1988
h1 <- h1[, 2:10]
h1$Cl.Jerk[1] <- 210
h1$Total[3] <- 407
h1$Snatch[3] <- 181
h1$Cl.Jerk[3] <- 226
h1$ID <- 15
h1$PED <- 1
h1$pairID <- "h"

karina <-
  read_html("http://www.iwf.net/results/athletes/?athlete=karina-endri-1989-03-
            02&id=1507")
H2 <- karina %>% html_nodes('table') %>% html_table()
h2 <- data.frame(H2)
h2$Year <- h2$Var.1
h2$Age <- h2$Year - 1989

```

```

h2 <- h2[, 2:10]
h2$ID <- 16
h2$PED <- 0
h2$pairID <- "h"

gyurkovics <-
  read_html("http://www.iwf.net/results/athletes/?athlete=gyurkovics-ferenc-1979
            -09-03&id=555")
I1 <- gyurkovics %>% html_nodes('table') %>% html_table()
i1 <- data.frame(I1)
i1$Year <- i1$Var.1
i1$Age <- i1$Year - 1979
i1 <- i1[, 2:10]
i1$Rank[16] <- "DSQ"
i1$Total[16] <- 420
i1$Snatch[16] <- 195
i1$Cl.Jerk[16] <- 225
i1$ID <- 17
i1$PED <- 1
i1$pairID <- "i"

vysniauskas <-
  read_html("http://www.iwf.net/results/athletes/?athlete=vysniauskas-ramunas-1976
            -09-23&id=309")
I2 <- vysniauskas %>% html_nodes('table') %>% html_table()
i2 <- data.frame(I2)
i2$Year <- i2$Var.1
i2$Age <- i2$Year - 1976
i2 <- i2[, 2:10]
i2$ID <- 18
i2$PED <- 0
i2$pairID <- "i"

danielyan <-
  read_html("http://www.iwf.net/results/athletes/?athlete=danielyan-ashot-1974-04
            -11&id=332")
J1 <- danielyan %>% html_nodes('table') %>% html_table()
j1 <- data.frame(J1)
j1$Year <- j1$Var.1
j1$Age <- j1$Year - 1974
j1 <- j1[, 2:10]
j1$Total[5] <- 465

```

```

j1$Snatch[5] <- 207.5
j1$Cl.Jerk[5] <- 257.5
j1$ID <- 19
j1$PED <- 1
j1$pairID <- "j"

sobotka <-
  read_html("http://www.iwf.net/results/athletes/?athlete=sobotka-petr-1975-04
            -30&id=338")
J2 <- sobotka %>% html_nodes('table') %>% html_table()
j2 <- data.frame(J2)
j2$Year <- j2$Var.1
j2$Age <- j2$Year - 1975
j2 <- j2[, 2:10]
j2$ID <- 20
j2$PED <- 0
j2$pairID <- "j"

# Women

ozkan.konak <-
  read_html("http://www.iwf.net/results/athletes/?athlete=ozkan-konak-sibel-
            1988-03-03&id=3370")
K1 <- ozkan.konak %>% html_nodes('table') %>% html_table()
k1 <- data.frame(K1)
k1$Year <- k1$Var.1
k1$Age <- k1$Year - 1988
k1 <- k1[, 2:10]
k1$Total[11] <- 199
k1$Snatch[11] <- 88
k1$Cl.Jerk[11] <- 111
k1$ID <- 21
k1$PED <- 1
k1$pairID <- "k"

karpinska <-
  read_html("http://www.iwf.net/results/athletes/?athlete=karpinska-marzena-1988-
            02-19&id=1350")
K2 <- karpinska %>% html_nodes('table') %>% html_table()
k2 <- data.frame(K2)
k2$Year <- k2$Var.1
k2$Age <- k2$Year - 1988

```

```

k2 <- k2[, 2:10]
k2$ID <- 22
k2$PED <- 0
k2$pairID <- "k"

iovu <-
  read_html("http://www.iwf.net/results/athletes/?athlete=iovu-cristina-1992-11-
            08&id=13007")
L1 <- iovu %>% html_nodes('table') %>% html_table()
l1 <- data.frame(L1[[3]])
l1$Year <- l1$Var.1
l1$Age <- l1$Year - 1992
l1 <- l1[, 2:10]
l1$Total[3] <- 219
l1$Snatch[3] <- 99
l1$Cl.Jerk[3] <- 120
l1$ID <- 23
l1$PED <- 1
l1$pairID <- "1"

hsu <-
  read_html("http://www.iwf.net/results/athletes/?athlete=hsu-shu-ching-1991-
            05-09&id=2220")
L2 <- hsu %>% html_nodes('table') %>% html_table()
l2 <- data.frame(L2)
l2$Year <- l2$Var.1
l2$Age <- l2$Year - 1991
l2 <- l2[, 2:10]
l2$ID <- 24
l2$PED <- 0
l2$pairID <- "1"

chinshanlo <-
  read_html("http://www.iwf.net/results/athletes/?athlete=chinshanlo-zulfiya-
            1993-07-25&id=2862")
M1 <- chinshanlo %>% html_nodes('table') %>% html_table()
m1 <- data.frame(M1)
m1$Year <- m1$Var.1
m1$Age <- m1$Year - 1993
m1 <- m1[, 2:10]
m1$Total[3] <- 226
m1$Snatch[3] <- 95

```

```

m1$Cl.Jerk[3] <- 131
m1$ID <- 25
m1$PED <- 1
m1$pairID <- "m"

yagi <-
  read_html("http://www.iwf.net/results/athletes/?athlete=yagi-kan-ae-1992-07-16&id=2838")
M2 <- yagi %>% html_nodes('table') %>% html_table()
m2 <- data.frame(M2)
m2$Year <- m2$Var.1
m2$Age <- m2$Year - 1992
m2 <- m2[, 2:10]
m2$ID <- 26
m2$PED <- 0
m2$pairID <- "m"

tsarukaeva <-
  read_html("http://www.iwf.net/results/athletes/?athlete=tsarukaeva-svetlana-1987-12-25&id=889")
N1 <- tsarukaeva %>% html_nodes('table') %>% html_table()
n1 <- data.frame(N1)
n1$Year <- n1$Var.1
n1$Age <- n1$Year - 1987
n1 <- n1[, 2:10]
n1$Total[1] <- 237
n1$Snatch[1] <- 112
n1$Cl.Jerk[1] <- 125
n1$ID <- 27
n1$PED <- 1
n1$pairID <- "n"

maneva <-
  read_html("http://www.iwf.net/results/athletes/?athlete=maneva-milka-mikova-1985-06-07&id=394")
N2 <- maneva %>% html_nodes('table') %>% html_table()
n2 <- data.frame(N2)
n2$Year <- n2$Var.1
n2$Age <- n2$Year - 1985
n2 <- n2[, 2:10]
n2$ID <- 28
n2$PED <- 0

```



```

n2$pairID <- "n"

maneza <-
  read_html("http://www.iwf.net/results/athletes/?athlete=maneza-maiya-1985-
            11-01&id=3431")
O1 <- maneza %>% html_nodes('table') %>% html_table()
o1 <- data.frame(O1)
o1$Year <- o1$Var.1
o1$Age <- o1$Year - 1985
o1 <- o1[, 2:10]
o1$Total[3] <- 245
o1$Snatch[3] <- 110
o1$Cl.Jerk[3] <- 135
o1$ID <- 29
o1$PED <- 1
o1$pairID <- "o"

girard <-
  read_html("http://www.iwf.net/results/athletes/?athlete=girard-christine-
            1985-01-03&id=809")
O2 <- girard %>% html_nodes('table') %>% html_table()
o2 <- data.frame(O2)
o2$Year <- o2$Var.1
o2$Age <- o2$Year - 1985
o2 <- o2[, 2:10]
o2$ID <- 30
o2$PED <- 0
o2$pairID <- "o"

shkermankova <-
  read_html("http://www.iwf.net/results/athletes/?athlete=shkermankova-maryna-
            1990-04-09&id=3693")
P1 <- shkermankova %>% html_nodes('table') %>% html_table()
p1 <- data.frame(P1)
p1$Year <- p1$Var.1
p1$Age <- p1$Year - 1990
p1 <- p1[, 2:10]
p1$Total[6] <- 256
p1$Snatch[6] <- 113
p1$Cl.Jerk[6] <- 143
p1$ID <- 31
p1$PED <- 1

```

```

p1$pairID <- "p"

cocos <-
  read_html("http://www.iwf.net/results/athletes/?athlete=cocos-roxana-daniela-
            1989-06-05&id=1406")
P2 <- cocos %>% html_nodes('table') %>% html_table()
p2 <- data.frame(P2)
p2$Year <- p2$Var.1
p2$Age <- p2$Year - 1989
p2 <- p2[, 2:10]
p2$ID <- 32
p2$PED <- 0
p2$pairID <- "p"

zabolotnaya <-
  read_html("http://www.iwf.net/results/athletes/?athlete=zabolotnaya-natalya-
            1985-08-15&id=6021")
Q1 <- zabolotnaya %>% html_nodes('table') %>% html_table()
q1 <- data.frame(Q1)
q1$Year <- q1$Var.1
q1$Age <- q1$Year - 1985
q1 <- q1[, 2:10]
q1$Total[1] <- 291
q1$Snatch[1] <- 131
q1$Cl.Jerk[1] <- 160
q1$ID <- 33
q1$PED <- 1
q1$pairID <- "q"

valentin.perez <-
  read_html("http://www.iwf.net/results/athletes/?athlete=valentin-perez-lidia-
            1985-02-10&id=829")
Q2 <- valentin.perez %>% html_nodes('table') %>% html_table()
q2 <- data.frame(Q2)
q2$Year <- q2$Var.1
q2$Age <- q2$Year - 1985
q2 <- q2[, 2:10]
q2$ID <- 34
q2$PED <- 0
q2$pairID <- "q"

kulesha <-

```

```

    read_html("http://www.iwf.net/results/athletes/?athlete=kulesha-iryna-
              1986-06-26&id=958")
R1 <- kulesha %>% html_nodes('table') %>% html_table()
r1 <- data.frame(R1)
r1$Year <- r1$Var.1
r1$Age <- r1$Year - 1986
r1 <- r1[, 2:10]
r1$Total[1] <- 269
r1$Snatch[1] <- 121
r1$Cl.Jerk[1] <- 148
r1$Total[4] <- 255
r1$Snatch[4] <- 118
r1$Cl.Jerk[4] <- 137
r1$ID <- 35
r1$PED <- 1
r1$pairID <- "r"

lim <-
    read_html("http://www.iwf.net/results/athletes/?athlete=lim-ji-hye-1985-
              10-28&id=3463")
R2 <- lim %>% html_nodes('table') %>% html_table()
r2 <- data.frame(R2)
r2$Year <- r2$Var.1
r2$Age <- r2$Year - 1985
r2 <- r2[, 2:10]
r2$ID <- 36
r2$PED <- 0
r2$pairID <- "r"

podobedova <-
    read_html("http://www.iwf.net/results/athletes/?athlete=podobedova-svetlana-
              1986-05-25&id=827")
S1 <- podobedova %>% html_nodes('table') %>% html_table()
s1 <- rbind(data.frame(S1[[1]]), data.frame(S1[[2]]))
s1$Year <- s1$Var.1
s1$Age <- s1$Year - 1986
s1 <- s1[, 2:10]
s1$Total[3] <- 291
s1$Snatch[3] <- 130
s1$Cl.Jerk[3] <- 161
s1$ID <- 37
s1$PED <- 1

```

```

s1$pairID <- "s"

mizdal <-
  read_html("http://www.iwf.net/results/athletes/?athlete=mizdal-ewa-justyna-
            1987-07-18&id=9935")
S2 <- mizdal %>% html_nodes('table') %>% html_table()
s2 <- data.frame(S2)
s2$Year <- s2$Var.1
s2$Age <- s2$Year - 1987
s2 <- s2[, 2:10]
s2$ID <- 38
s2$PED <- 0
s2$pairID <- "s"

khurshudyan <-
  read_html("http://www.iwf.net/results/athletes/?athlete=khurshudyan-hripsime-
            1987-07-27&id=818")
T1 <- khurshudyan %>% html_nodes('table') %>% html_table()
t1 <- data.frame(T1)
t1$Year <- t1$Var.1
t1$Age <- t1$Year - 1987
t1 <- t1[, 2:10]
t1$Total[2] <- 246
t1$Snatch[2] <- 110
t1$Cl.Jerk[2] <- 136
t1$Total[7] <- 294
t1$Snatch[7] <- 128
t1$Cl.Jerk[7] <- 166
t1$Total[16] <- 235
t1$Snatch[16] <- 105
t1$Cl.Jerk[16] <- 130
t1$ID <- 39
t1$PED <- 1
t1$pairID <- "t"

shimamoto <-
  read_html("http://www.iwf.net/results/athletes/?athlete=shimamoto-mami-
            1987-09-24&id=445")
T2 <- shimamoto %>% html_nodes('table') %>% html_table()
t2 <- data.frame(T2)
t2$Year <- t2$Var.1
t2$Age <- t2$Year - 1987

```

```

t2 <- t2[, 2:10]
t2$ID <- 40
t2$PED <- 0
t2$pairID <- "t"

# All Athletes
men <- rbind(a1, a2, b1, b2, c1, c2, d1, d2, e1, e2, f1, f2, g1, g2,
            h1, h2, i1, i2, j1, j2)
men$Sex <- rep(1, length(men$Total))

women <- rbind(k1, k2, l1, l2, m1, m2, n1, n2, o1, o2, p1, p2, q1, q2, r1,
              r2, s1, s2, t1[,1:12], t2)
women$Sex <- rep(0, length(women$Total))

athletes <- rbind(men, women)

write.csv(athletes, file = "athletes.csv", sep = ",",
          col.names = TRUE, row.names = FALSE)
athletes <- read.csv("~/athletes.csv")

# Initial Plots
hist(athletes$Total, xlab = "Meet Total (kg)",
     main = "Figure 3: Histogram of Meet Totals (all athletes)",
     breaks = 17, col = "grey")

par(mfrow = c(1, 2))
data <- athletes[which(athletes$pairID == "t"), ]

plot(Total ~ Age, data = data[which(data$PED == 1), ], xlab = "Age",
     ylab = "Meet Total", main = "Cheater", xlim = c(14, 40), ylim = c(0, 500),
     pch = 20)
plot(Total ~ Age, data = data[which(data$PED == 0), ], xlab = "Age",
     ylab = "Meet Total", main = "Non-cheater", xlim = c(14, 40),
     ylim = c(0, 500), pch = 20)
title("Figure 4: Women's 75+ kg Weight Class", line = -1, outer = TRUE)

par(mfrow = c(1, 2))
data <- athletes[which(athletes$pairID == "t"), ]

plot(Total ~ BWT, data = data[which(data$PED == 1), ],
     xlab = "Body Weight (kg)",
     ylab = "Meet Total", main = "Cheater",
     xlim = c(60, 110), ylim = c(0, 500), pch = 20)

```

```

plot(Total ~ BWT, data = data[which(data$PED == 0), ],
     xlab = "Body Weight (kg)",
     ylab = "Meet Total", main = "Non-cheater",
     xlim = c(60, 110),
     ylim = c(0, 500), pch = 20)
title("Figure 5: Women's 75+ kg Weight Class",
      line = -1, outer = TRUE)

```

```

par(mfrow = c(1, 1))
model <- lm(Total ~ Age*BWT + as.factor(PED) + as.factor(Sex),
            data = athletes)
athletes$Age <- as.numeric(athletes$Age)
interplot(m = model, var1 = "Age", var2 = "BWT", plot = TRUE) +
  xlab('Body Weight (kg)') +
  ylab('Estimated Coefficient for Age') +
  ggtitle("Figure 6: Interaction Plot (Body Weight and Age)")

```

```

# Meet Total Model Portion
athletes$nonzero.data <- rep(0, length(athletes$Age))
for (i in 1:length(athletes$Age)){
  athletes$nonzero.data[i] <- ifelse(athletes$Total[i] == 0, NA,
                                    athletes$Total[i])
}

data <- na.omit(athletes)
N <- length(data)

ybar <- vector()
sampvar <- vector()
n <- vector()

for (j in 1:40){

  ybar[j] <- mean(data[data$ID == j, 7])
  sampvar[j] <- var(data[data$ID == j, 7])
  n[j] <- length(data[data$ID == j, 7])
}

b_0 <- ybar
delta2 <- mean(sampvar)

# Set up X-matrix

```

```

X1 <- rep(NA, 492)
for(i in 1:40){
  X <- rep(1, n[i])
  X1 <- c(X1, X, rep(0, 492))
}

X2 <- matrix(X1[493:(length(X1) - 492)], 492, 40)

X <- cbind(X2, data$Age, data$BWT, data$Age^2, data$BWT^2, data$PED, data$Sex)
y <- data$Total
m <- length(unique(data$ID))
# ybar <- n <- sv <- rep(0, length(unique(data$ID)))

model2 <- lm(Total ~ Age + BWT + I(Age^2) + I(BWT^2)
             + as.factor(PED) + as.factor(Sex),
             data = data)
model3 <- lme(Total ~ Age + BWT + I(Age^2) + I(BWT^2)
             + as.factor(PED) + as.factor(Sex),
             random = ~ 1|ID, data = data)

p <- length(X[1, ])

### starting values
beta.0 <- c(rep(model2$coefficients[1], 40), model2$coefficients[2:7])
Sigma.0 <- diag(c(rep(20, 40)^2, summary(model2)$coefficients[2:7 , 2]^2), 46)
nu.0 <- 0
sigma.0 <- 1
sims <- 10000

sigma <- 1
invSigma.0 <- solve(Sigma.0)
XtX <- t(X)%*%X
b_0i.post <- vector()
b_0i.post[1] <- model2$coefficients[1]
beta.post <- matrix(nrow = sims, ncol = p)
sigma.post <- vector()

for(sims in 1:sims) {

#update beta
  Sigma.n <- solve( invSigma.0 + XtX/sigma^2 )

```

```

beta.n <- Sigma.n%*( invSigma.0%*beta.0 + t(X)%*y/sigma^2 )
beta <- t(rmvnorm(1, beta.n, Sigma.n))
#update sigma
nu.n <- nu.0 + N
ssr <- t(y - X%*beta)%*(y - X%*beta)
sigma.sq <- rinvgamma(1, (nu.n)/2, (nu.0*sigma.0^2 + ssr)/2)

beta.post[sims, ] <- t(beta)
sigma.post[sims] <- sqrt(sigma.sq)
}

Parameter <- c("Age", "Weight", "Age^2", "Weight^2", "Cheater", "Sex")
Estimate <- round(c(colMeans(beta.post[, 41:46])), 3)

vec <- vector()
j <- 1
for (i in 1:40){
  n.temp <- n[i]
  end <- j + n.temp - 1
  vec[j:end] <- rep(mean(beta.post[, i]), n.temp)
  j <- end + 1
}
variance <- var(vec)

post <- cbind(beta.post, sigma.post)

par(mfrow = c(2, 3))
plot(beta.post[, 41], type = 'l', main = "Age", ylab = " ")
plot(beta.post[, 43], type = 'l', main = "Age^2", ylab = " ")
plot(beta.post[, 42], type = 'l', main = "Weight", ylab = " ")
plot(beta.post[, 44], type = 'l', main = "Weight^2", ylab = " ")
plot(beta.post[, 45], type = 'l', main = "Cheater", ylab = " ")
plot(beta.post[, 46], type = 'l', main = "Sex", ylab = " ")
title("Figure 7: Trace Plots (Meet Totals)", line = -1, outer = TRUE)

# par(mfrow = c(2, 3))
# hist(beta.post[, 41], main = "Age", xlab = " ")
# hist(beta.post[, 43], main = "Age^2", xlab = " ")
# hist(beta.post[, 42], main = "Weight", xlab = " ")
# hist(beta.post[, 44], main = "Weight^2", xlab = " ")
# hist(beta.post[, 45], main = "Cheater", xlab = " ")
# hist(beta.post[, 46], main = "Sex", xlab = " ")

```



```

# title("Figure 8: Posterior Distributions (Meet Totals)", line = -1,
#       outer = TRUE)

Credible_Interval <- rbind(quantile(post[, 41], p = c(0.025, 0.975)),
                           quantile(post[, 42], p = c(0.025, 0.975)),
                           quantile(post[, 43], p = c(0.025, 0.975)),
                           quantile(post[, 44], p = c(0.025, 0.975)),
                           quantile(post[, 45], p = c(0.025, 0.975)),
                           quantile(post[, 46], p = c(0.025, 0.975)))

kable(cbind(Parameter, Estimate,
             round(Credible_Interval, 4)),
      caption = "Random Intercept Model")

# Scratch Model Portion
athletes$Scratch <- ifelse(is.na(athletes$nonzero.data), 1, 0)

model4 <- glm(Scratch ~ Age + BWT + I(Age^2) +
              I(BWT^2) + as.factor(PED) +
              as.factor(Sex), data = athletes,
              family = binomial(link = "probit"))

require(LearnBayes)

# Set up X-matrix

X1 <- rep(NA, 547)
for(i in 1:40){
  X <- rep(1, n[i])
  X1 <- c(X1, X, rep(0, 547))
}

X2 <- matrix(X1[548:(length(X1) - 547)], 547, 40)

X <- cbind(X2, athletes$Age, athletes$BWT, athletes$Age^2, athletes$BWT^2,
           athletes$PED, athletes$Sex)

prior <- list(beta = c(rep(model4$coefficients[1], 40), model4$coefficients[2:7]),
              P = diag(c(rep(summary(model4)$coeff[1, 2]^2, 40),
                          summary(model4)$coeff[2:7, 2]^2)))

sims <- 10000

```

```

model5 <- bayes.probit(athletes$Scratch, X, sims, prior)

Parameter <- c("Age", "Weight", "Age^2", "Weight^2", "Cheater", "Sex")
Estimate <- round(colMeans(model5$beta[, 41:46]), 6)

par(mfrow = c(2, 3))
plot(model5$beta[, 41], type = 'l', main = "Age", ylab = " ")
plot(model5$beta[, 42], type = 'l', main = "Age^2", ylab = " ")
plot(model5$beta[, 43], type = 'l', main = "Weight", ylab = " ")
plot(model5$beta[, 44], type = 'l', main = "Weight^2", ylab = " ")
plot(model5$beta[, 45], type = 'l', main = "Cheater", ylab = " ")
plot(model5$beta[, 46], type = 'l', main = "Sex", ylab = " ")
title("Figure 8: Trace Plots (Scratches)", line = -1, outer = TRUE)

#
# par(mfrow = c(2, 3))
# hist(model5$beta[, 2], main = "Age", xlab = " ")
# hist(model5$beta[, 3], main = "Age^2", xlab = " ")
# hist(model5$beta[, 4], main = "Weight", xlab = " ")
# hist(model5$beta[, 5], main = "Weight^2", xlab = " ")
# hist(model5$beta[, 6], main = "Cheater", xlab = " ")
# hist(model5$beta[, 7], main = "Sex", xlab = " ")
# title("Figure 10: Posterior Distributions (Scratches)", line = -1, outer = TRUE)

Credible_Interval <- rbind(quantile(model5$beta[, 41], p = c(0.025, 0.975)),
                          quantile(model5$beta[, 42], p = c(0.025, 0.975)),
                          quantile(model5$beta[, 43], p = c(0.025, 0.975)),
                          quantile(model5$beta[, 44], p = c(0.025, 0.975)),
                          quantile(model5$beta[, 45], p = c(0.025, 0.975)),
                          quantile(model5$beta[, 46], p = c(0.025, 0.975)))

kable(cbind(Parameter, Estimate, round(Credible_Interval, 6)),
      caption = "Probit Regression Model")

x <- vector()
Age <- vector()
BWT <- vector()
PED <- vector()
Sex <- vector()
ID <- vector()

```

```

n <- 1000

for (i in 1:n){
  b_1 <- sample(post[, 41], 1)
  b_2 <- sample(post[, 42], 1)
  b_11 <- sample(post[, 43], 1)
  b_22 <- sample(post[, 44], 1)
  b_3 <- sample(post[, 45], 1)
  b_4 <- sample(post[, 46], 1)
  e <- sample(post[, 47], 1)
  vars <- athletes[, c(4, 9, 10, 11, 13)]
  vars1 <- vars[sample(nrow(vars), 1),]
  age <- vars1$Age
  bwt <- vars1$BWT
  ped <- vars1$PED
  sex <- vars1$Sex
  id <- vars1$ID
  b_0 <- sample(post[, id], 1)
  Age[i] <- age
  BWT[i] <- bwt
  PED[i] <- ped
  Sex[i] <- sex
  ID[i] <- id
  theta <- pnorm(sample(model5$beta[, id], 1) + sample(model5$beta[, 41], 1)*age +
    sample(model5$beta[, 42], 1)*bwt + sample(model5$beta[, 43], 1)*age^2 +
    sample(model5$beta[, 44], 1)*bwt^2 + sample(model5$beta[, 45], 1)*ped +
    sample(model5$beta[, 46], 1)*sex, mean = 0, sd = 1)
  obs <- rbernoulli(1, theta)
  x[i] <- ifelse(obs == TRUE, 0, b_0 + b_1*age + b_11*age^2 + b_2*bwt +
    b_22*bwt^2 +
    b_3*ped + b_4*sex)
}

par(mfrow = c(1, 2))
plot(Total ~ Age, data = athletes, xlab = "Age",
  ylab = "Meet Total (kg)", main = "The Data",
  xlim = c(14, 40), ylim = c(0, 500), pch = 20)
plot(x ~ Age, xlab = "Age", ylab = "Meet Total (kg)",
  main = "Posterior Predictive Distribution",
  xlim = c(14, 40), ylim = c(0, 500), pch = 20)
title("Figure 9: The Data versus the Posterior Predictive Distribution",
  line = -1, outer = TRUE)

```

```

par(mfrow = c(1, 2))
hist(athletes$Total, xlab = "Meet Total (kg)", main = "Mixture Distribution",
     breaks = 17, col = "grey", freq = FALSE, ylim = c(0, 0.01))
lines(density(x))

data <- athletes
N <- length(data)
X <- cbind(rep(1, N), data$Age, data$BWT, data$Age^2, data$BWT^2, data$PED, data$Sex)
y <- data$Total

model2 <- lm(Total ~ Age + BWT + I(Age^2) + I(BWT^2) +
             as.factor(PED) +
             as.factor(Sex), data = data)

p <- length(X[, ])

### starting values
beta.0 <- model2$coefficients
Sigma.0 <- diag(summary(model2)$coefficients[, 2], 7)
nu.0 <- 0
sigma.0 <- 1
sims <- 10000

sigma <- 1
invSigma.0 <- solve(Sigma.0)
XtX <- t(X)%*%X
beta.post <- matrix(nrow = sims, ncol = p)
sigma.post <- vector()

for(sims in 1:sims) {
  #update beta
  Sigma.n <- solve( invSigma.0 + XtX/sigma^2 )
  beta.n <- Sigma.n%*( invSigma.0%*beta.0 + t(X)%*y/sigma^2 )
  beta <- t(rmvnorm(1, beta.n, Sigma.n))
  #update sigma
  nu.n <- nu.0 + N
  ssr <- t(y - X%*beta)%*(y - X%*beta)
  sigma.sq <- rinvgamma(1, (nu.n)/2, (nu.0*sigma.0^2 + ssr)/2)
  # sigma <- nu.0*sigma.0^2 + sum( (y - X%*beta)^2 )

  beta.post[sims, ] <- t(beta)
}

```

```

    sigma.post[sims] <- sqrt(sigma.sq)
  }

x <- vector()
Age <- vector()
BWT <- vector()
PED <- vector()
Sex <- vector()
n <- 1000

for (i in 1:n){
  b_0 <- sample(beta.post[, 1], 1)
  b_1 <- sample(beta.post[, 2], 1)
  b_2 <- sample(beta.post[, 3], 1)
  b_11 <- sample(beta.post[, 4], 1)
  b_22 <- sample(beta.post[, 5], 1)
  b_3 <- sample(beta.post[, 6], 1)
  b_4 <- sample(beta.post[, 7], 1)
  vars <- athletes[, c(4, 9, 11, 13)]
  vars1 <- vars[sample(nrow(vars), 1),]
  age <- vars1$Age
  bwt <- vars1$BWT
  ped <- vars1$PED
  sex <- vars1$Sex
  Age[i] <- age
  BWT[i] <- bwt
  PED[i] <- ped
  Sex[i] <- sex
  x[i] <- b_0 + b_1*age + b_11*age^2 + b_2*bwt + b_22*bwt^2 + b_3*ped + b_4*sex
}

hist(athletes$Total, xlab = "Meet Total (kg)",
     main = "Not Accounting for Scratches",
     breaks = 17, col = "grey", freq = FALSE, ylim = c(0, 0.01))
lines(density(x))
title("Figure 10: Why use a mixture distribution?", line = -1, outer = TRUE)

```

8 APPENDIX OF ATHLETES

Men 69 kg:

Cheater: ARTYKOV, Izzat

Non-cheater: SANCHEZ LOPEZ, David

Men 69 kg:

Cheater: MARTIROSYAN, Tigran

Non-cheater: MINASIDIS, Dimitris

Men 77 kg:

Cheater: PEREPECHENOV, Oleg

Non-cheater: MAPOSUA, Uati

Men 85 kg:

Cheater: SINCRAIAN, Gabriel

Non-cheater: BARDIS, Giovanni Battista

Men 85 kg:

Cheater: AUKHADOV, Apti

Non-cheater: ROSTAMI, Kianoush

Men 85 kg:

Cheater: RYBAKOU, Andrei

Non-cheater: MARTIROSYAN, Tigran

Men 94 kg:

Cheater: ILYIN, Ilya

Non-cheater: ANTTIROIKO, Miika Matias

Men 94 kg:

Cheater: CIRICU, Anatoli

Non-cheater: KARINA, Endri

Men 105 kg:

Cheater: GYURKOVICS, Ferenc

Non-cheater: VYSNIAUSKAS, Ramunas

Men 105+ kg:

Cheater: DANIELYAN, Ashot

Non-cheater: SOBOTKA, Petr

Women 48 kg:

Cheater: OZKAN KONAK, Sibel

Non-cheater: KARPINSKA, Marzena

Women 53 kg: Cheater: IOVU, Cristina
Non-cheater: HSU, Shu-Ching

Women 53 kg: Cheater: CHINSHANLO, Zulfiya
Non-cheater: YAGI, Kanae

Women 63 kg:
Cheater: TSARUKAEVA, Svetlana
Non-cheater: MANEVA Milka, Mikova

Women 63 kg:
Cheater: MANEZA, Maiya
Non-cheater: GIRARD, Christine

Women 69 kg: Cheater: SHKERMANKOVA, Maryna
Non-cheater: COCOS, Roxana Daniela

Women 75 kg:
Cheater: ZABOLOTNAYA, Natalya
Non-cheater: VALENTIN PEREZ, Lidia

Women 75 kg:
Cheater: KULESHA, Iryna
Non-cheater: LIM, Ji-Hye

Women 75 kg:
Cheater: PODOBEDOVA, Svetlana
Non-cheater: MIZDAL, Ewa Justyna

Women 75+ kg:
Cheater: KHURSHUDYAN, Hripsime
Non-cheater: SHIMAMOTO, Mami