The Importance of Statistics in Volleyball

Parker Booth

Introduction

- University of Utah's volleyball team finished the 2023 season with a record of 11 - 19
  - 15 - 16 in 2022 & 22 - 9 in 2021

- What volleyball game statistics are essential to winning?
- What movements are vital for success in both practice and games?

- The importance of Kills in Volleyball

- Four different relationships modeled using Stan
  - 2 Logistic Regressions & 2 Robust Linear Regressions

Data

Volleyball game statistics are aggregated as all statistics recorded by the team each match

Opponent game statistics were recorded as well

The practice and movement data highlights the eight critical players on the team throughout all practices and game days

All data is collected in an applied setting

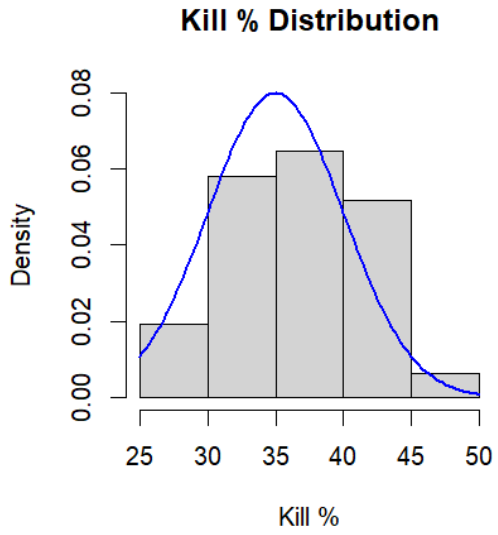
Practice and movement data is collected using Catapult IMUs

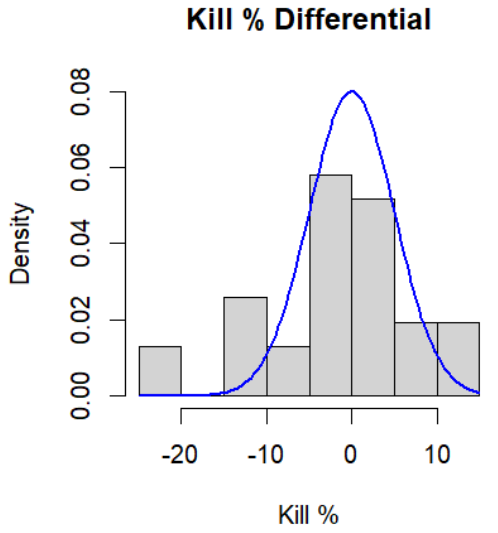
Metrics such as training load and jump statistics are relevant

Data Description

- **Kill %:** Percentage of attacks that directly lead to a point
- **Kill % Differential:** Difference between Utah’s kill percentage and the opponent's kill percentage
- **High Jump %:** Calculated by Catapult IMUs. Percentage of jumps classified as high
  - Setters’ medium jumps are relevant for this metric. Liberos removed from this metric
  - All other positions calculated with high jumps
- **Player Load per Min:** Calculated by Catapult IMUs. Player load is a metric that calculates the effort a given player exerts

- **Other metrics used:**
  - Dig %, Attacking Error %, Serving Error %, Ace %





Prior Specifications

- Sports statistics are commonly distributed around a mean with a low amount of outliers
  - Theoretically matches a normal distribution
  - Uniform priors do not fit well (Ace %)
- Normal Distributions for all priors but Catapult metrics
  - Normal priors gave unrealistic results for movement metrics
- Different priors were tried that gave lesser results
  - Exponential & Truncated Normal

Model 1: Logistic Regression

- Logistic Model of Important Volleyball Statistics on a Dichotomous Win Variable

- Win Dummy ~ Bernoulli Logit(Kill% + Attack Error% + Serve Error% + Ace% + Dig%)
  - Observations = 31; Iterations = 10,000; Chains = 3;
  - Priors:
    - Kill % ~N(35, 5); Attack Error % ~ N(8, 2); Dig % ~ N(65, 5); Serve Error % ~ N(12, 4); Ace % ~ N(8, 4)

- Shows which metrics are the most impactful on winning volleyball games

Stan Output

- Rhats all equal 1
- n_eff large for most parameters
- Coefficients defined in log odds

- Kill % has the largest impact on winning
  - Both in the point estimate in the 95% probability interval

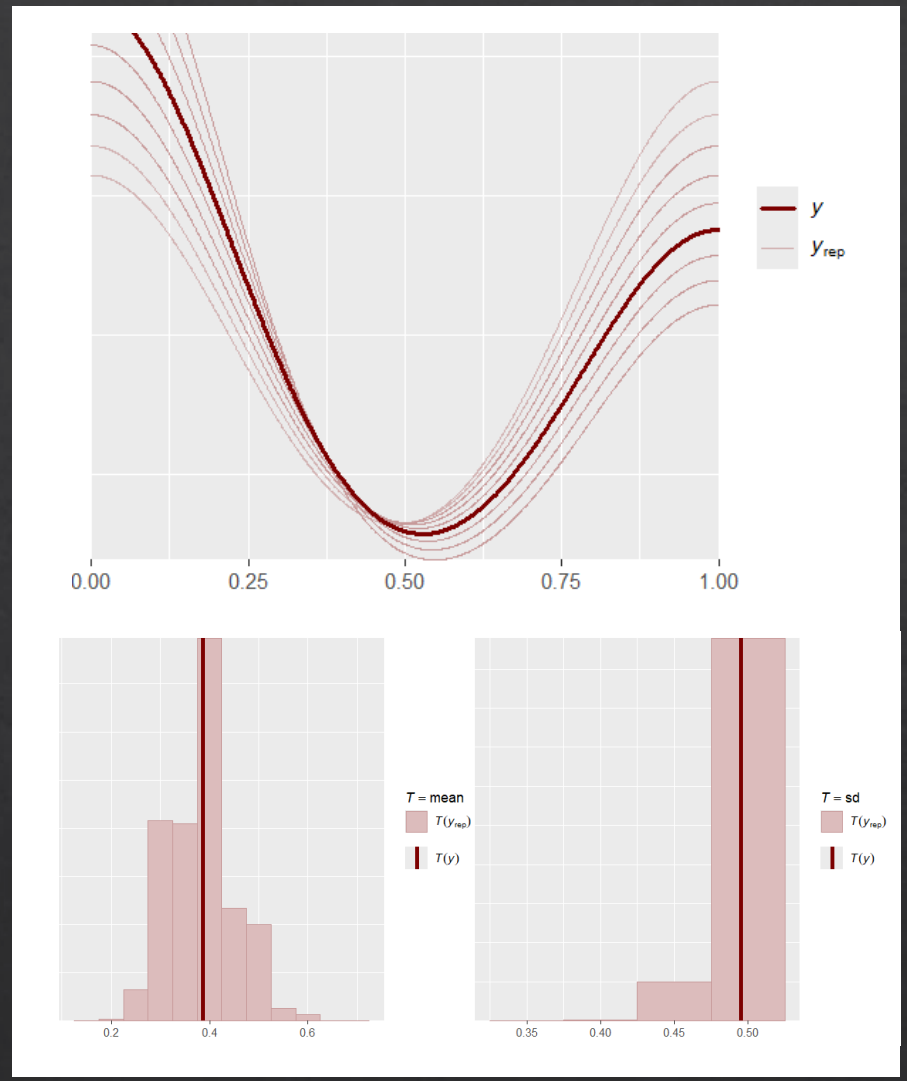
- Both error metrics are negative
  - Almost fully negative

- Dig % & Ace % are positive
  - Dig % is fully positive
  - Ace % not determined

Parameter	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
Intercept	49.47	0.26	20.52	-95.28	-16.36	6201	1
Kill %	0.79	0.00	0.32	0.29	1.53	6878	1
Attack Error %	-0.52	0.00	0.33	-1.23	0.07	13651	1
Dig %	0.40	0.00	0.19	0.07	0.83	7264	1
Serve Error %	-0.40	0.00	0.24	-0.93	0.01	9955	1
Ace %	0.31	0.00	0.28	-0.17	0.95	13657	1

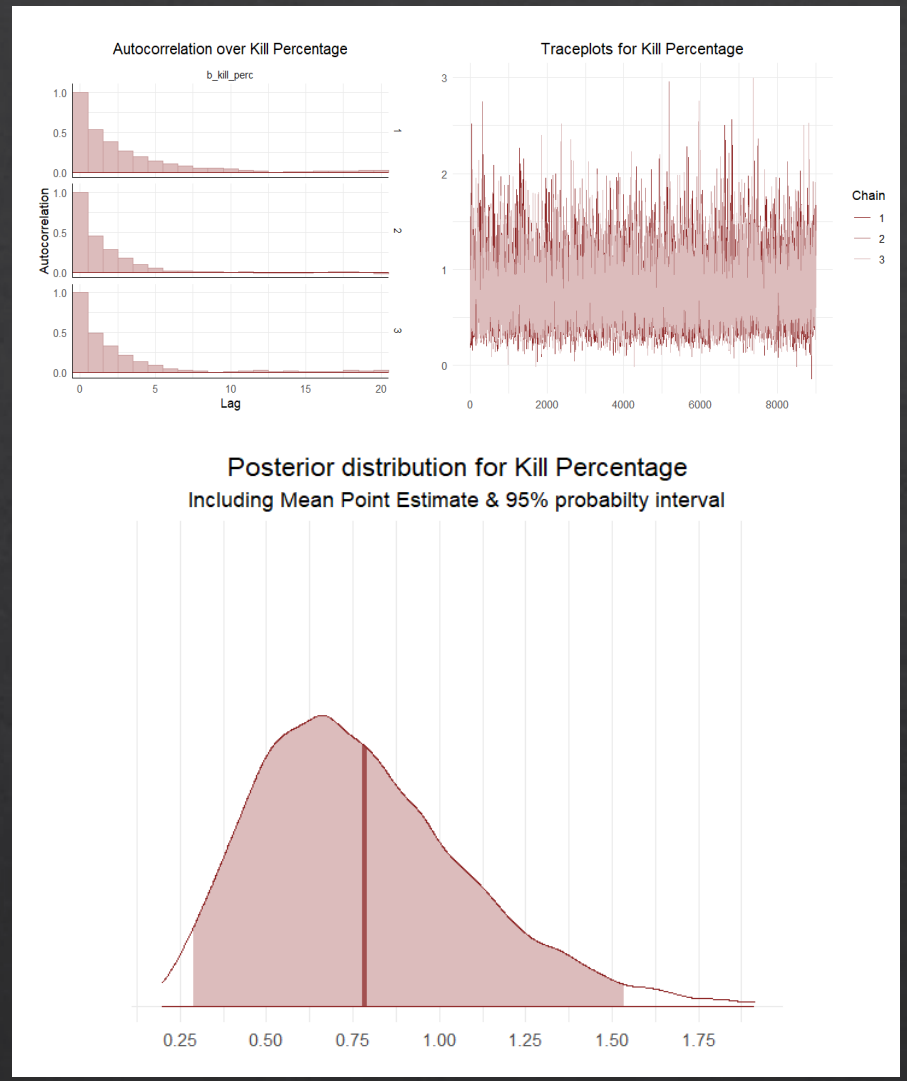
Posterior Predictive Checks

- Generative data follows the distribution of the volleyball data
  - Also puts all weight on ones and zeros
- Mean and standard deviation recovered well
  - Max and Min also recovered
    - Shown by generative data putting weights on ones and zeros
- Mean distributed normally around large center



Posterior Predictive Checks

- Autocorrelation dies out after approx. 5 periods
- Trace plots show convergence and efficient sampling
  - Samples negative values only a few times
- Coefficient on Kill % ranges from around 0.3 to 1.5
  - 99% interval plotted; 95% interval highlighted
  - Purely positive



Model 2: Logistic Regression

- Logistic Model of Kill % Differential on a Dichotomous Win Variable

- Win Dummy ~ Bernoulli Logit(Kill_Diff)
  - Observations = 31; Iterations = 10,000; Chains = 3;
  - Priors: Kill Diff. ~ N(0,5);

- Isolates the impact of kill percentage on winning
- Kill differential considers both offensive and defensive aspects of the game
  - Is the team attacking efficiently and defending opposing attacks efficiently?
  - More telling metric than Kill %

Stan Output

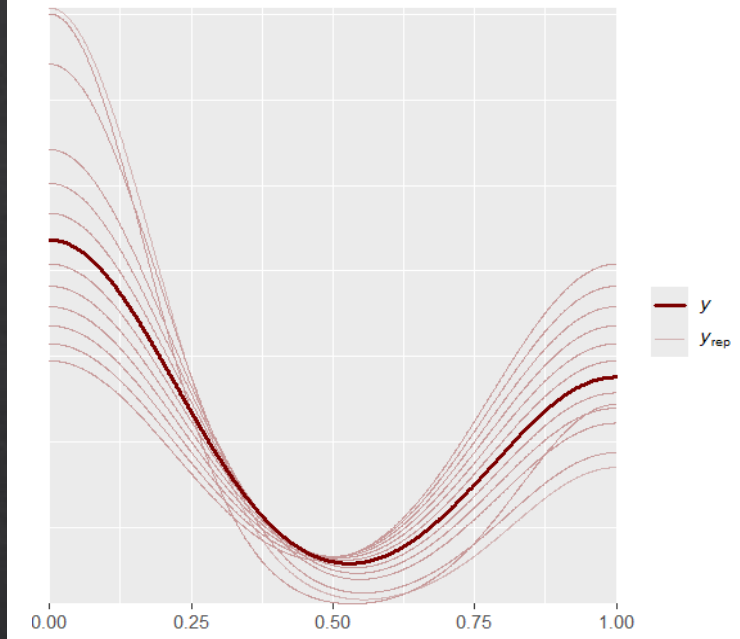
Parameter	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
Intercept	-0.84	0.00	0.57	-2.03	0.21	18392	1
Kill Differential	0.38	0.00	0.14	0.16	0.70	15966	1

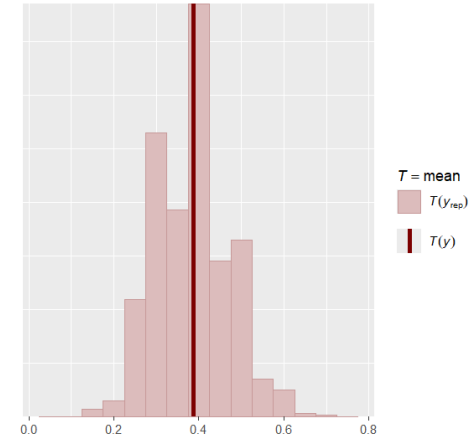
- Same positive relationship exists between kill differential and winning
  - 95% interval is fully positive

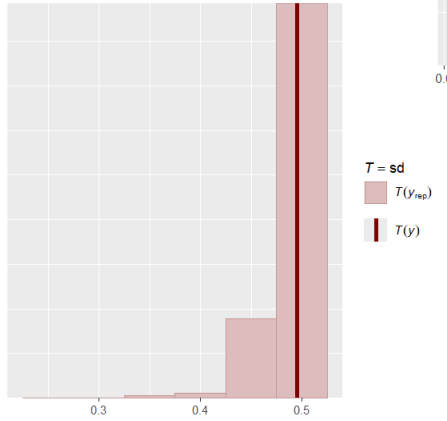
- Rhats equal to 1 for both parameters
- n_eff is high for both parameters

Posterior Predictive Checks

- Generative data follows the distribution of the volleyball data
  - Again, puts all weight on ones and zeros
- Mean and standard deviation recovered well
  - Max and Min also recovered
    - Shown by generative data putting weights on ones and zeros
- Mean distributed somewhat normally

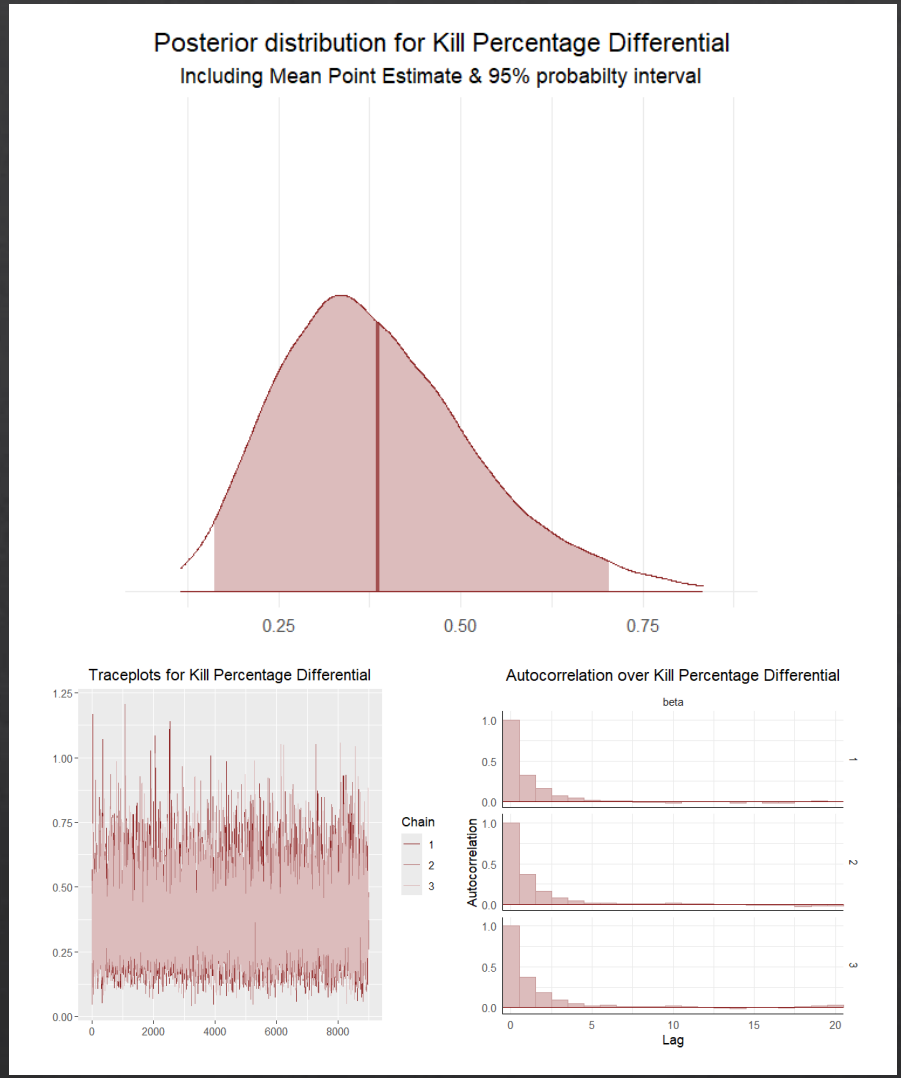






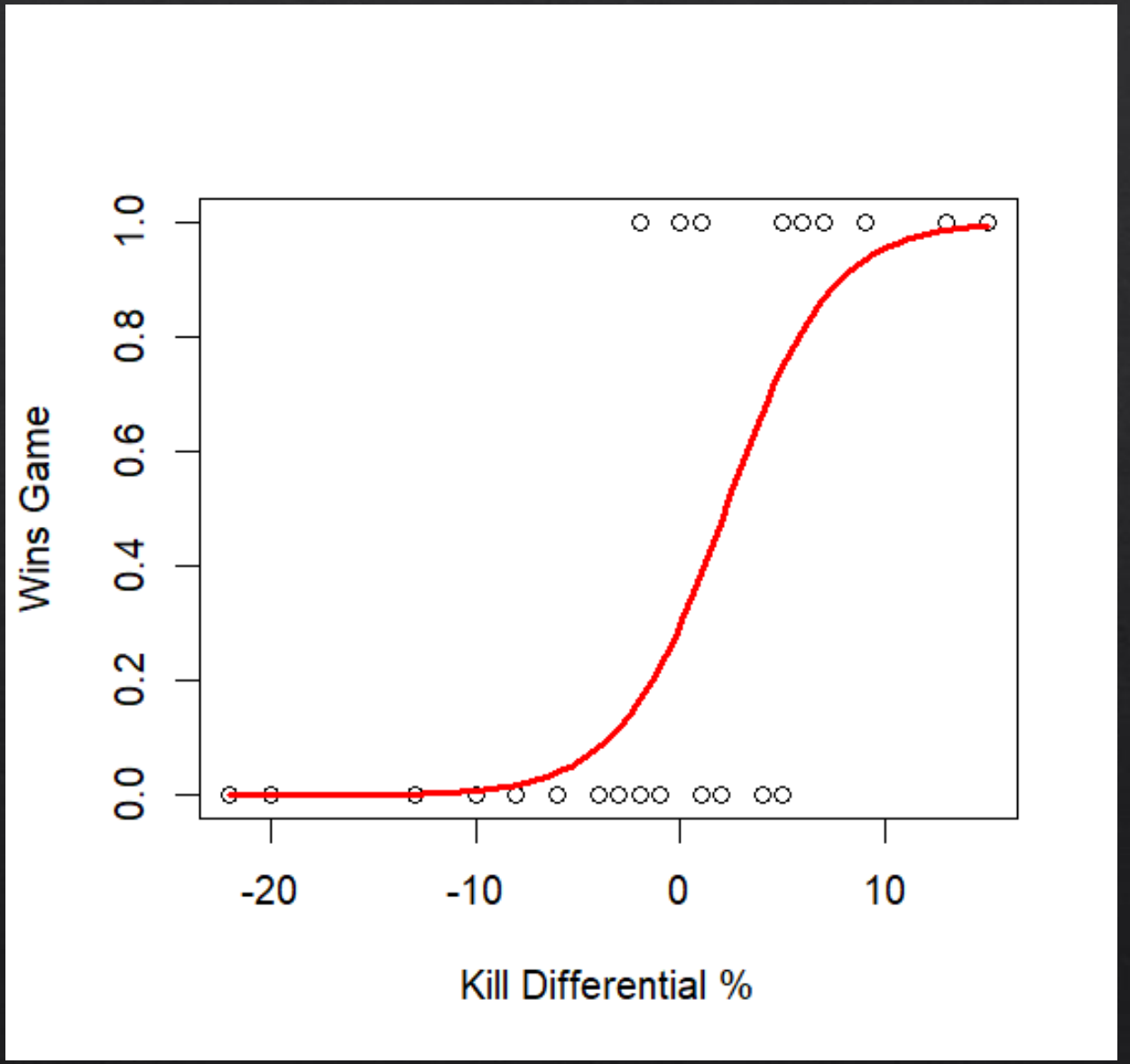
Posterior Predictive Checks

- Autocorrelation dies out after approx. 5 periods
- Trace plots show convergence and efficient sampling
  - Negative values are not sampled
- Coefficient on Kill % ranges from around 0.1 to 0.7
  - 99% interval plotted; 95% interval highlighted
  - Once again purely positive



Logistic Sigmoid Curve

- Plotting the stan output gives a logistic curve over the wins and losses
  - Wins = 1; Losses = 0
- Kill diff. of 0 ≈ 40% chance to win
- Changes in win likelihood happen between -10 and 10
- Lowest win was a kill diff. of ≈ -2
  - Only wins with negative diff.



Model 3: Robust Normal Regression

- Robust Linear Model of Good Game Jumps on Kill % Differential

- Kill Differential ~ student_t(nu, X * beta, sigma)
  - Observations = 29; Iterations = 10,000; Chains = 3;
  - Priors: nu ~ Gamma(2,.1)

- Ignorance prior used for beta
- Jumps collected by Catapult IMUs within games
  - High Jumps for most positions; Medium for Setters; Liberos excluded
- Two games did not have catapult data (observations: 31 ➔ 29)

Stan Output

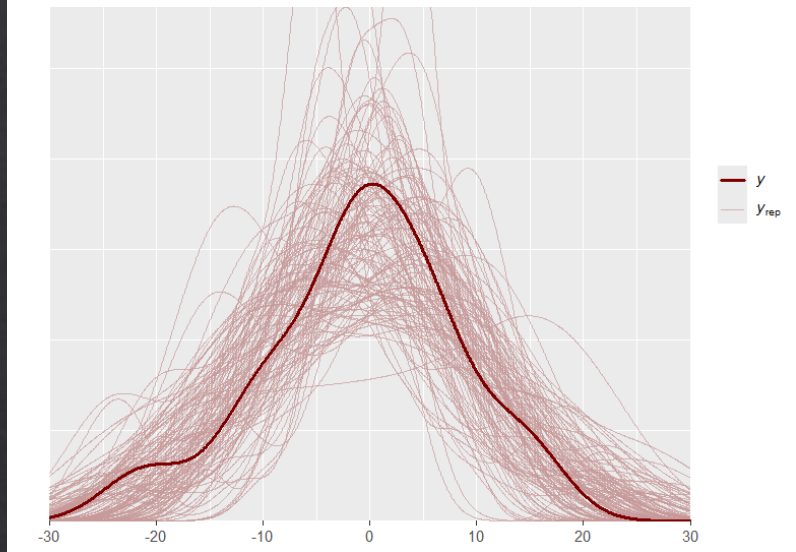
- Good Jumps has an almost fully positive relationship with Kill diff.
  - 95% interval crosses 0
- Rhats equal to 1 for all parameters
- n_eff is high for all parameters

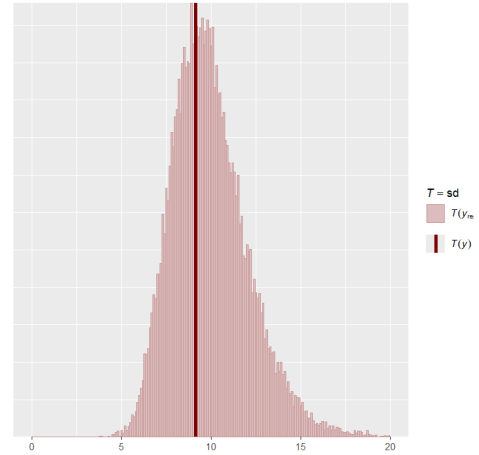
Parameter	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
Intercept	-43.34`	0.25	23.83	-89.96	4.08	8763	1
Good Jumps	0.62	0.00	0.14	-0.06	1.31	8765	1
Sigma	8.66	0.01	1.38	6.34	11.74	12087	1
Nu	22.16	0.11	14.04	4.43	57.05	15823	1

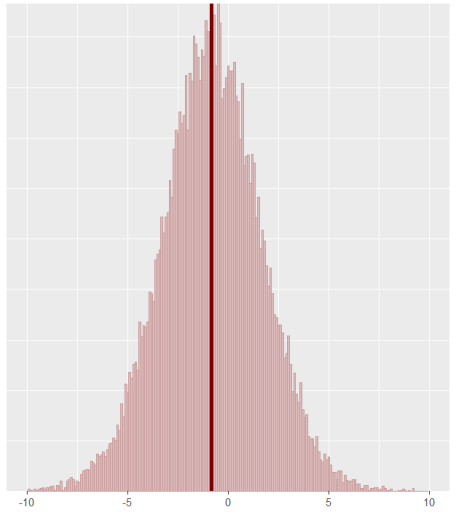
Posterior Predictive Checks

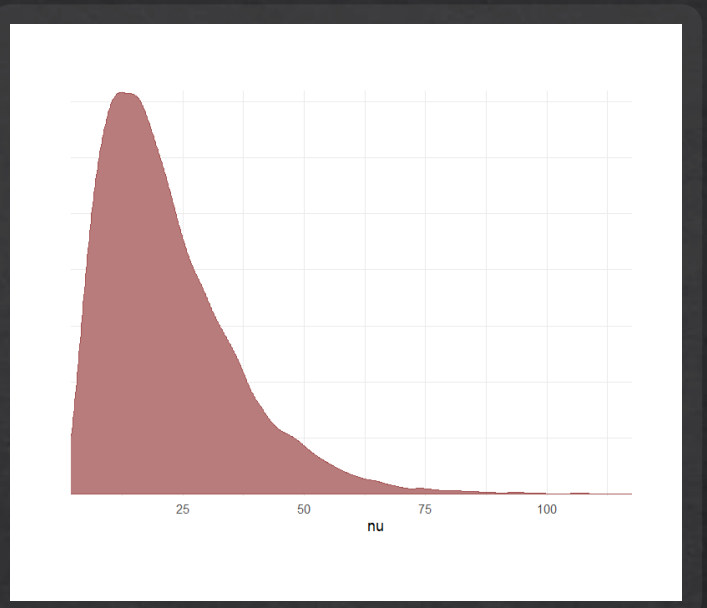
- Generative data follows the normal distribution of the volleyball data
  - Spread of generative data explains high variance

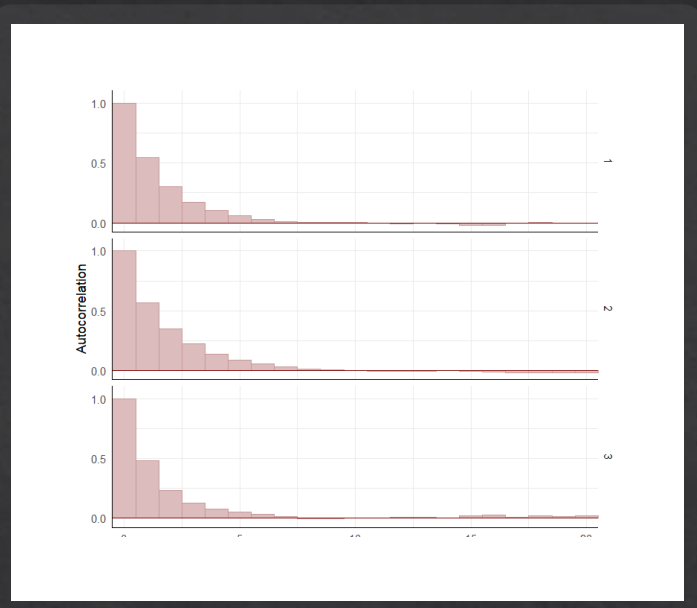
- Mean and standard deviation recovered well
  - Large spread for both of these variables
  - Max and Min also recovered

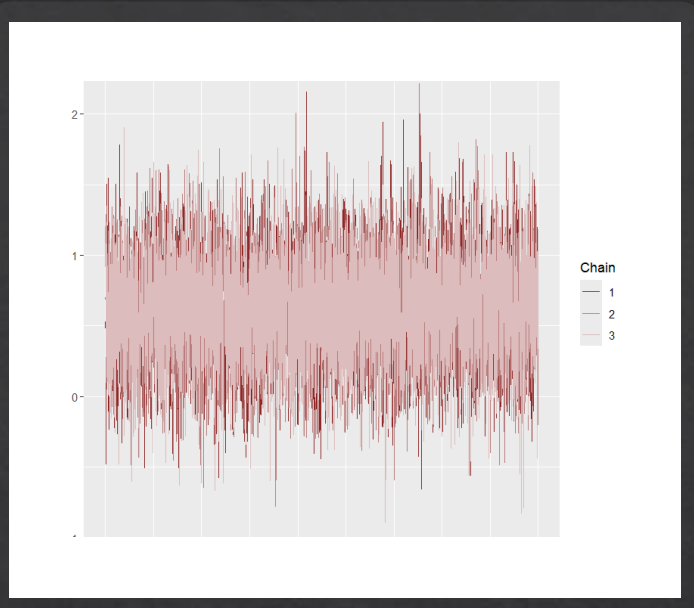












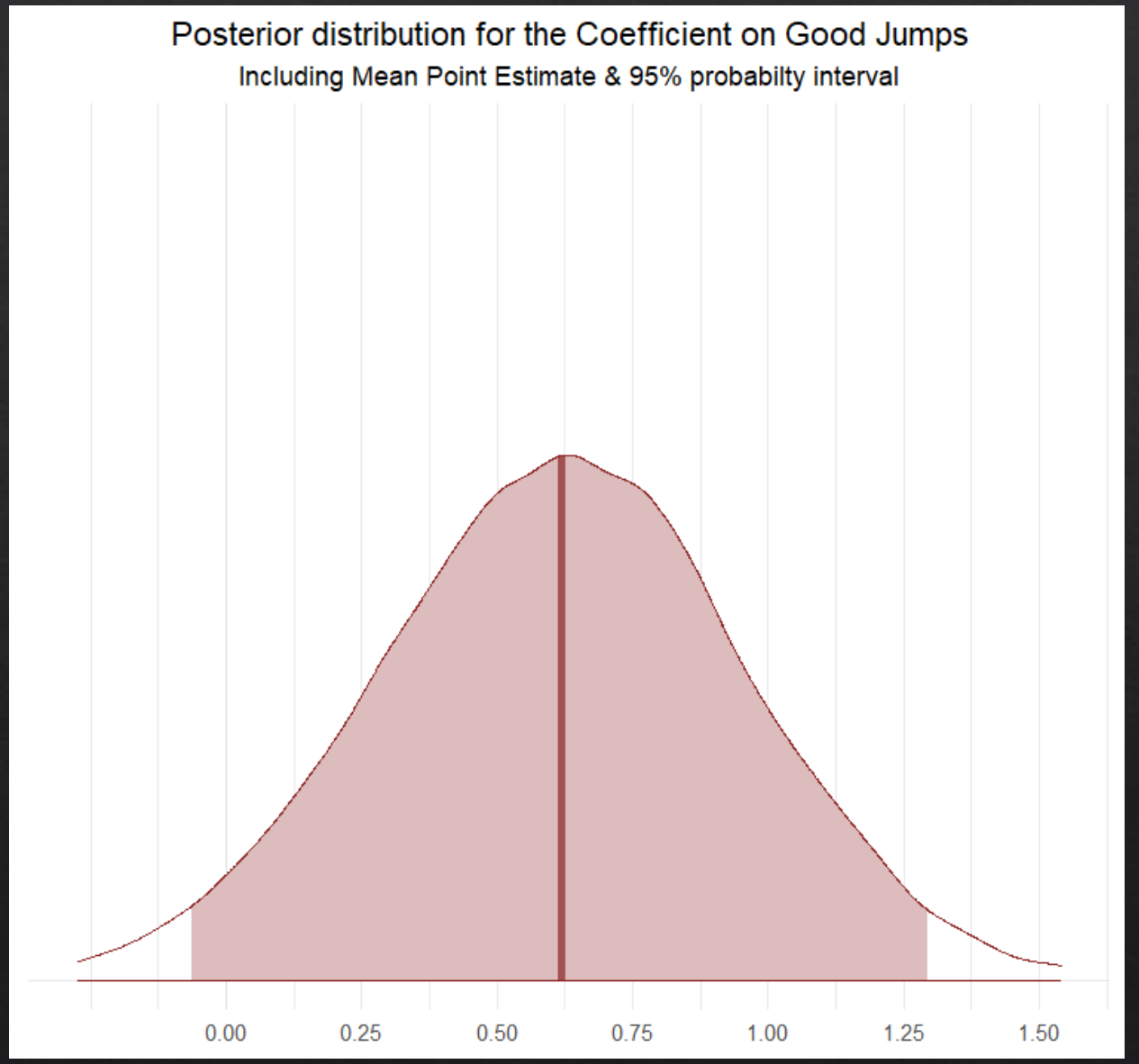
Posterior Predictive Checks

- Trace plot shows efficient sampling
  - Samples over 0 regularly
- Autocorrelation dies out after approx. 5 periods

- Nu contains large values
  - Indication of student-t converging to normal distribution

Good Jumps Density Curve

- Coefficient on Good Jumps ranges from -0.1 to 1.3
- 90% interval does not contain negative values
  - Setting is applied
  - Jumping is all but required to get a kill
- Kill Diff adds an interesting angle
  - Team could be jumping higher than ever, and still get outperformed



Model 4: Robust Normal Regression

- Robust Linear Model of Practice Good Jumps & Player Load on Kill % Differential:

- Kill Differential ~ student_t(nu, X * beta, sigma)
  - Observations = 29; Iterations = 10,000; Chains = 3;
  - Priors: nu ~ Gamma(2,.1);

- Ignorance prior used for beta
- Two practices did not have catapult data (observations: 31 → 29)
- Only practices analyzed were the practices the day before a game

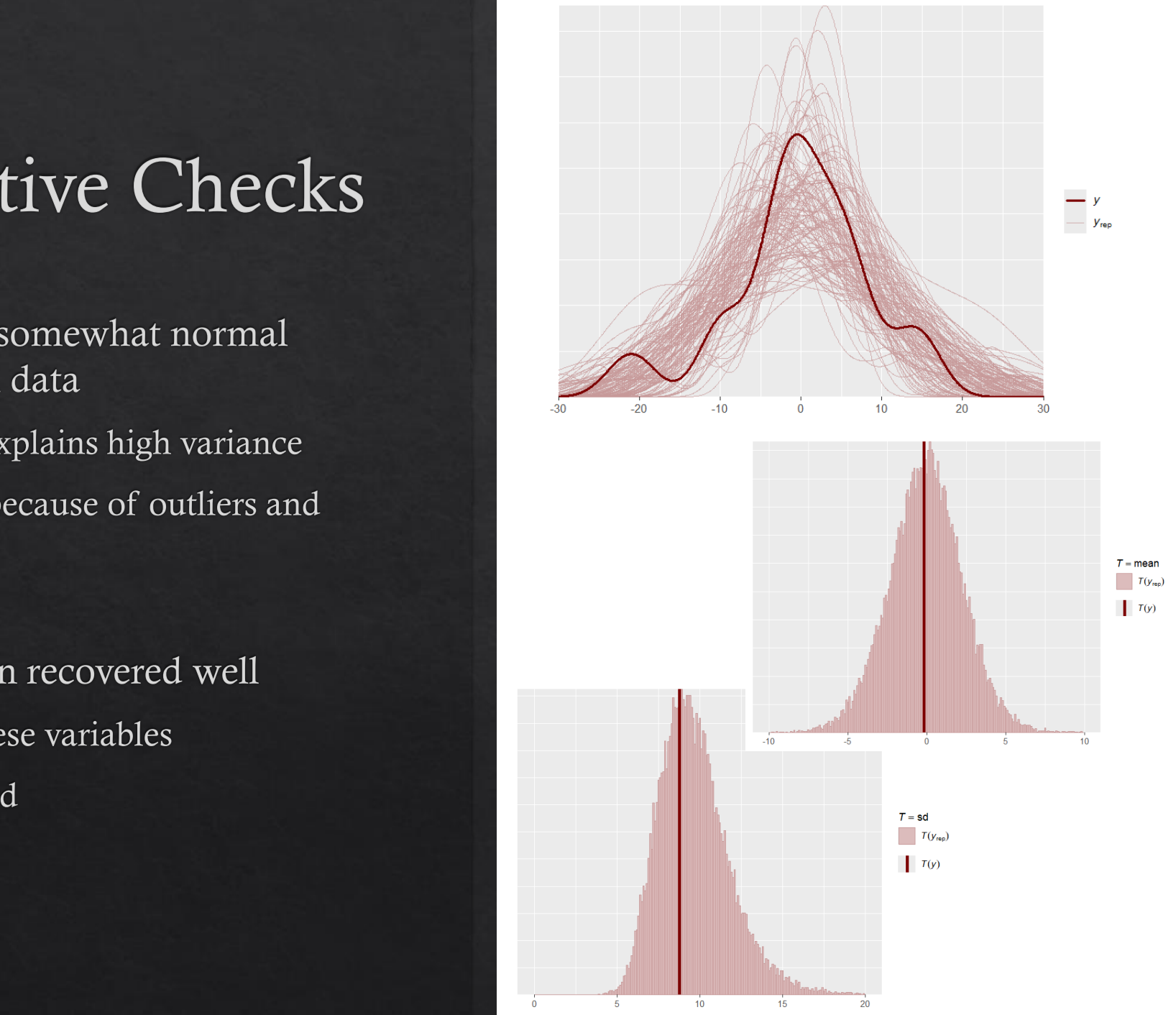
Stan Output

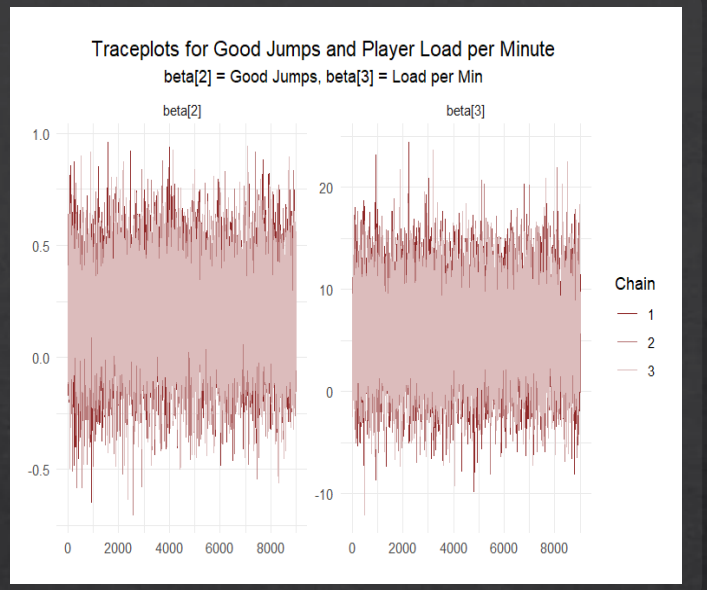
Parameter	mean	se_mean	sd	2.5%	97.5%	n_eff	Rhat
Intercept	-31.58	0.17	18.02	-66.57	4.17	11008	1
Good Jumps	0.18	0.00	0.20	-0.21	0.58	13859	1
Player Load	5.95	0.03	3.93	-1.70	13.80	13580	1
Sigma	8.13	0.01	1.47	5.56	11.33	12916	1
Nu	18.96	0.10	13.62	3.29	53.88	17761	1

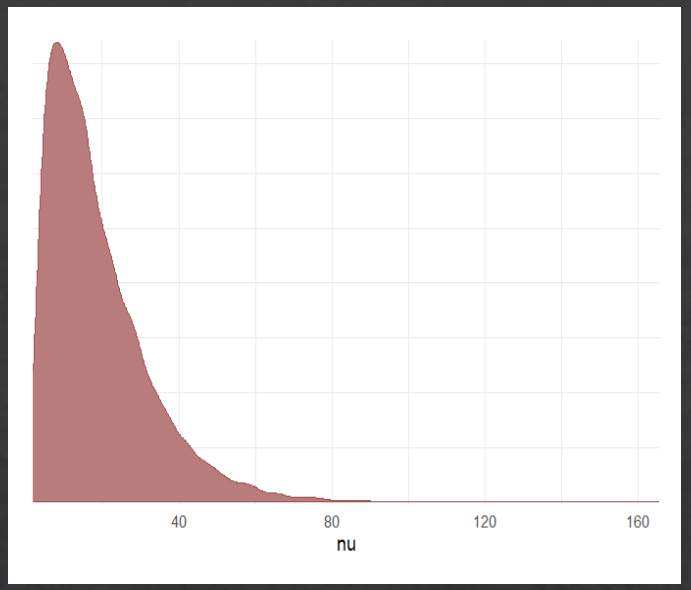
- Much more inconclusive results
- Both Good Jumps and Player Load cross into the negative values
- Rhats equal 1
- n_eff values are sufficiently large

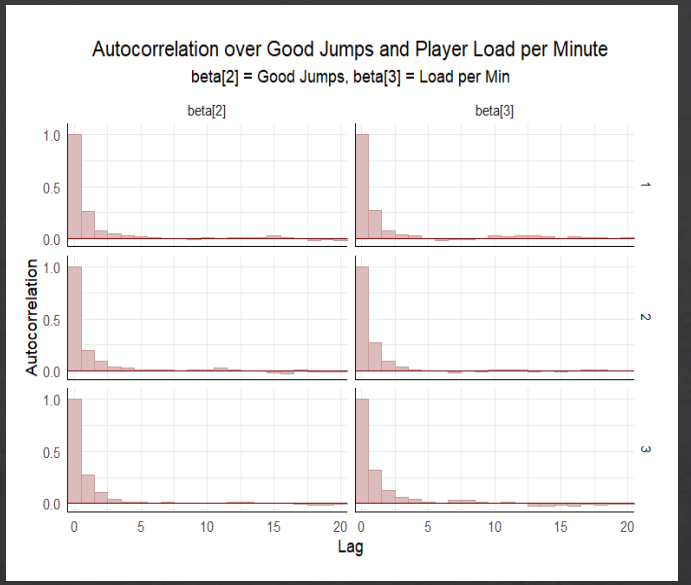
Posterior Predictive Checks

- Generative data follows the somewhat normal distribution of the volleyball data
  - Spread of generative data explains high variance
  - Weird shape of Utah data because of outliers and low amount of data
- Mean and standard deviation recovered well
  - Large spread for both of these variables
  - Max and Min also recovered









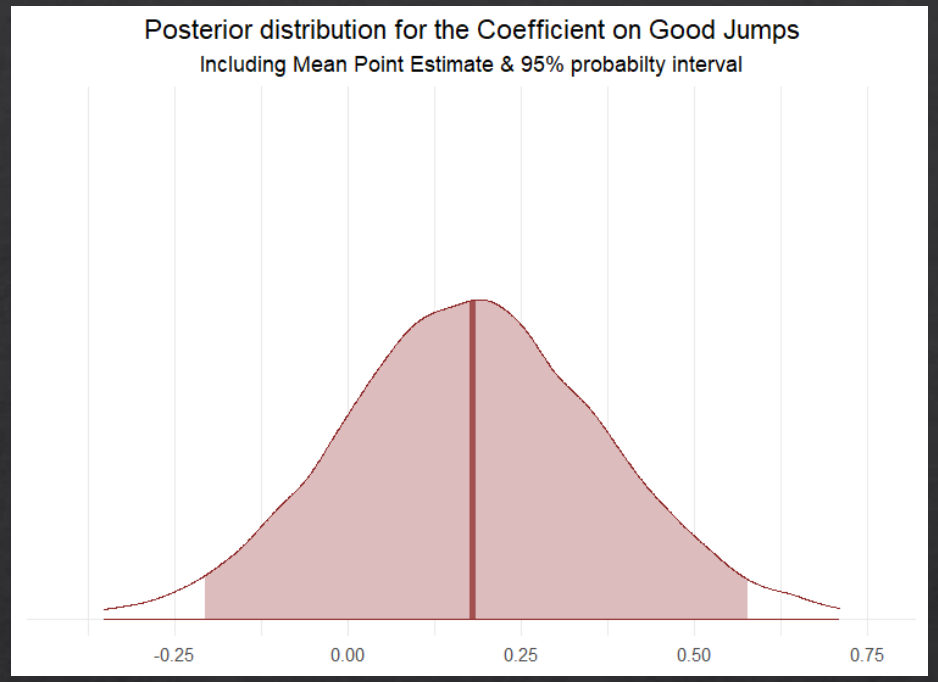
Posterior Predictive Checks

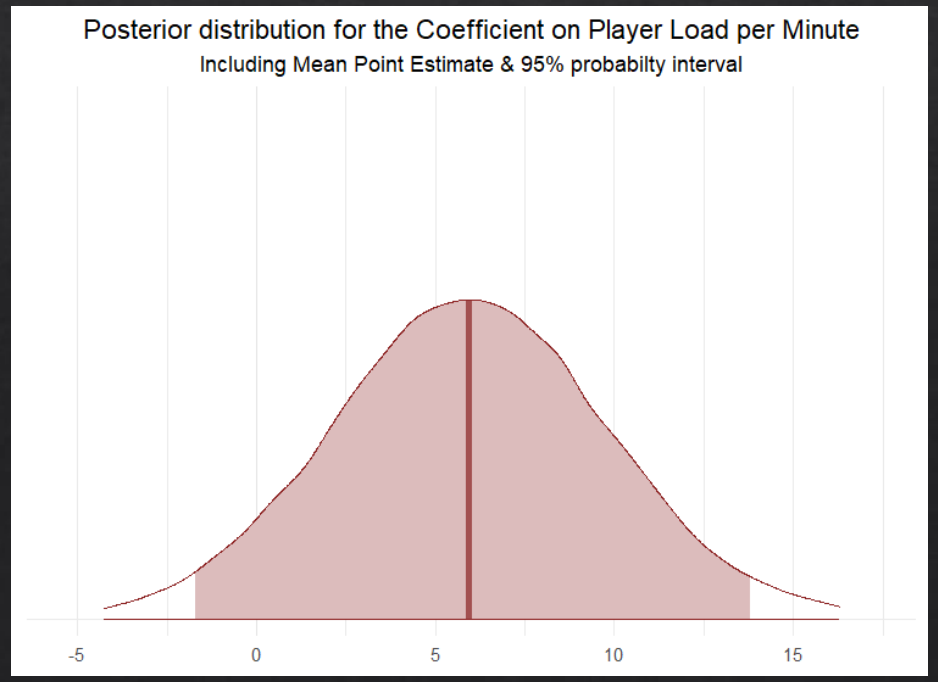
- Both trace plots show efficient sampling
  - Both sample negative values regularly
- Autocorrelation dies out almost immediately

- Nu contains somewhat large values
  - Indication of student-t converging to normal distribution
  - Not as large as the previous model

Density Curves

- Negative values indicate uncertainty
  - Coefficient on Good Jumps ranges from -0.2 to 0.6
  - Coefficient on Player Load ranges from -2 to 13
- Practice effort leads to improvements in kill differential, up to a point
- 90% intervals are mostly positive





Limitations

Data is collected in an applied settings

- Lots of confounding variables that cannot be accounted for

Small dataset size

- Analysis would be more concrete if multiple seasons were compared

Conclusion

- Kill percentage has the largest impact on winning volleyball games
- Kill % differential encapsulates offensive and defensive performance
  - Also has a large impact on winning
- The Bayesian framework and data collection in an applied setting allow for relaxed probability intervals
  - Good jumps in a game correlate with a higher kill differential
  - Effort in practice the day before a game has an undetermined impact on kill differential

Sources

- **Lots of advice from the University of Utah Applied Health and Performance Science Team as well as strength coaches, and information that I have heard indirectly from the coaching staff**
- "Women's Volleyball Statistics." NCAA.com, National Collegiate Athletic Association, https://www.ncaa.com/stats/volleyball-women/d1
- "Logistic/Probit Regression." Stan Users Guide, Stan Development Team, https://mc-stan.org/docs/stan-users-guide/regression.html#logistic-probit-regression.section.
- "Catapult Metric Descriptions." PlayerTekPlus, Catapult Sports, https://playertekplus.catapultsports.com/.
- "OpenAI Chat." OpenAI, OpenAI, https://chat.openai.com/?model=text-davinci-002-render-sha.
- PubMed Central, National Center for Biotechnology Information, U.S. National Library of Medicine, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7052708/#:~:text=Catapult%20Sports%20proposed%20that%20PL,et%20al.%2C%202011.
- Grammarly, Grammarly Inc., https://www.grammarly.com/.
- University of Utah Writing Center, University of Utah, https://writingcenter.utah.edu/.
- "Stan User Guide" Stan, Stan Development Team, https://mc-stan.org/.