Chapter 13.6 Career Trajectories

Jim Albert and Monika Hu

Chapter 13 Case Studies

Introduction

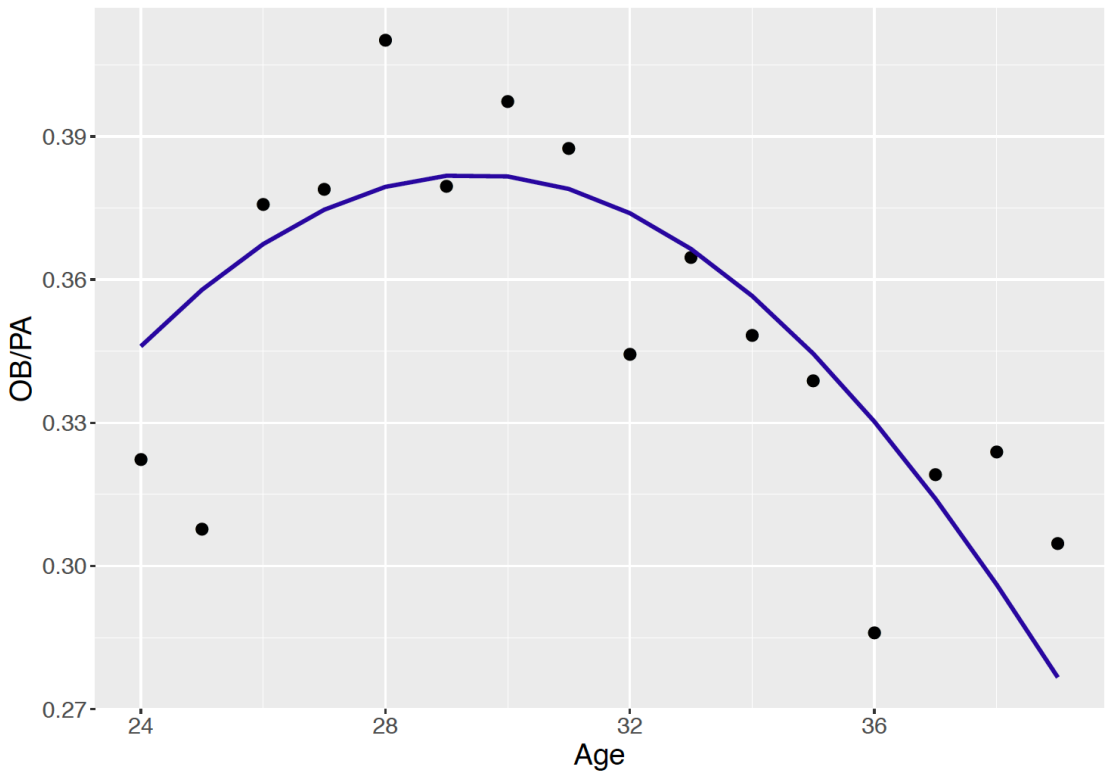
- The performance of a professional athlete typically begins at a small level, increases to a level where the player has peak performance, and then decreases until the player's retirement.
- This pattern of performance over a player's career is called the **career trajectory.**
- A general problem in sports is to predict future performance of a player and one relevant variable in this prediction is the player's age.
- We study career trajectories for baseball players, although the methodology will apply to athletes in other sports.

Measuring hitting performance in baseball

- Baseball players are measured by their ability to hit, pitch, and field.
- One of the more popular measures of batting performance is the on-base percentage or OBP.
- The OBP is defined to be the fraction of plate appearances where the batter reaches a base.
- Chase Utley in 2003 had 49 on-base events in 152 plate appearances and his OBP was $49/152 = 0.322$.

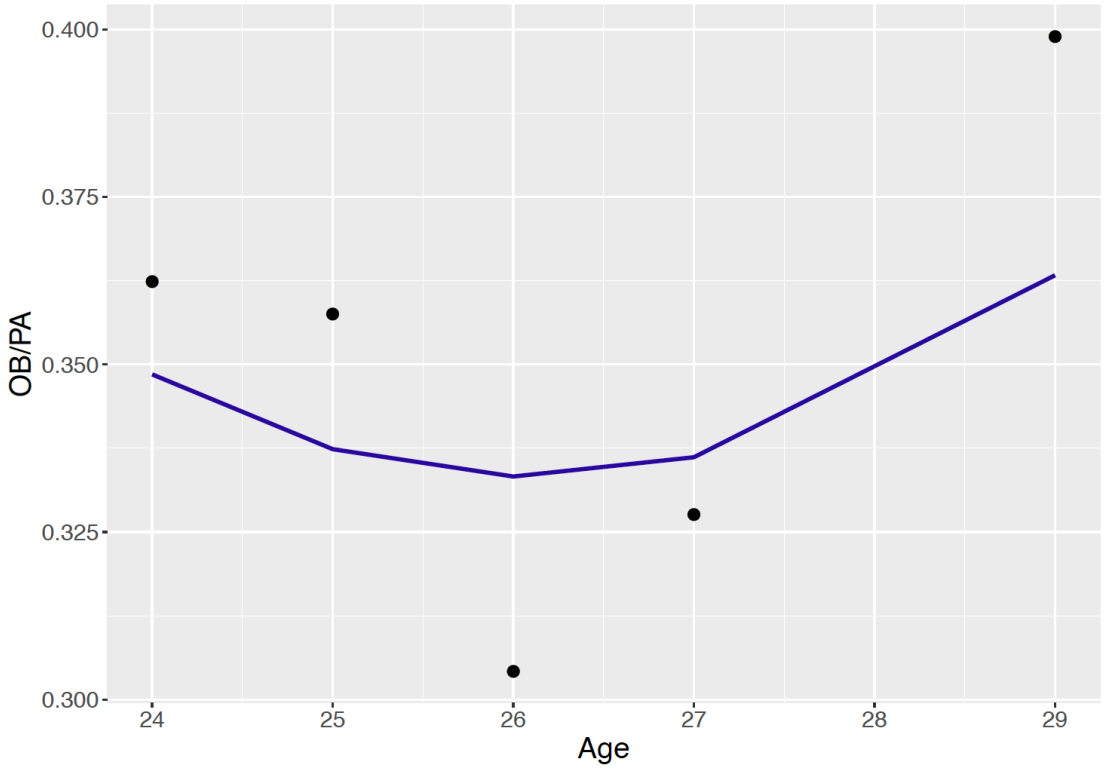
Chase Ultley's career trajectory

- We explore career trajectories of the OBP measure of a baseball player as a function of his age.
- Figure displays Chase Utley's OBP as a function of his age. A quadratic smoothing curve is added. Utley's OBP measure increases until about age 30 and then steadily decreases towards the end of his career.



Josh Phelps' Career Trajectory

- Figure displays the career trajectory of OBP for Josh Phelps who had a relatively short baseball career.
- Phelps does not have a clearly defined career trajectory.



Purpose of Case Study

- The purpose of this case study is to see if one can improve the career trajectory smooth of this player by a hierarchical Bayesian model that combines data from a number of baseball players.
- Recall in Chapter 10, we have seen how hierarchical Bayesian models have the pooling effect that could borrow information from other groups to improve the estimation of one group, especially for groups with small sample size.

Estimating a single trajectory

- Let y_j denote the number of on-base events in n_j plate appearances during a single hitter's j-th season.
- Assume that y_j has a Binomial distribution with parameters n_j and probability of success p_j.
- One represents the logit of the success probability p_j as a quadratic function of the player's age x_j:

$$\log\left(\frac{p_j}{1-p_j}\right) = \beta_0 + \beta_1(x_j - 30) + \beta_2(x_j - 30)^2$$

Comments

- The intercept β_0 is an estimate of the player's OBP performance at age 30.
- The quadratic function reaches its largest value at

$$h_1(\beta) = 30 - \frac{\beta_1}{2\beta_2}.$$

  This is the age where the player is estimated to have his peak on-base performance during his career.
- The maximum value of the curve, on the logistic scale, is

$$h_2(\beta) = \beta_0 - \frac{\beta_1^2}{4\beta_2}.$$

Comments (continued)

- The maximum value of the curve on the probability scale is

$$p_{max} = \exp(h_2(\beta))/(1 + \exp(h_2(\beta))).$$

  The parameter p_{max} is the estimated largest OBP of the player over his career.

- The coefficient β_2, typically a negative value, tells us about the degree of curvature in the quadratic function.
- One simple interpretation is that β_2 represents the change in OBP from his peak age to one year later.

Fitting the Model

- Fit this Bayesian logistic model using the JAGS software.
- Assume the regression coefficients are independent with each coefficient assigned a Normal prior with mean 0 and precision 0.0001.
- The posterior density of β is given, up to an unknown proportionality constant, by

$$\pi(\beta \mid \{y_j\}) \propto \prod_j \left(p_j^{y_j} (1 - p_j)^{n_j - y_j} \right) \pi(\beta),$$

where p_j is defined by the logistic model and $\pi(\beta)$ is the prior density.

R Work

The JAGS model script is shown below.

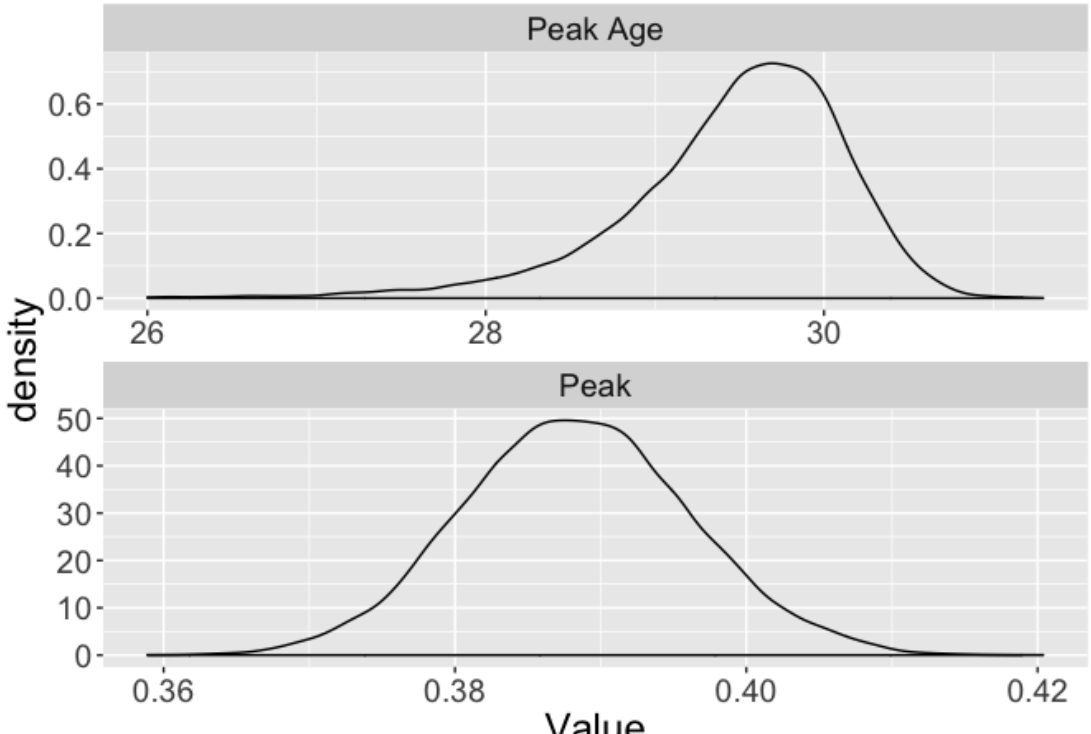
```
modelString = "
model {
## sampling
for (j in 1:N){
   y[j] ~ dbin(p[j], n[j])
   logit(p[j]) <- beta0 + beta1 * (x[j] - 30) +
            beta2 * (x[j] - 30) * (x[j] - 30)
}
## priors
beta0 ~ dnorm(0, 0.0001)
beta1 ~ dnorm(0, 0.0001)
beta2 ~ dnorm(0, 0.0001)
}
"
```

Simulating from Posterior

- JAGS software is used to simulate a sample from the posterior distribution of β.
- One performs inference about the peak age function $h_1(\beta)$ by computing this function on the simulated β draws – the output is a posterior sample from the peak age function.
- In a similar fashion, one obtains a sample from the posterior of the maximum value function p_{max}

Some Posterior Distributions

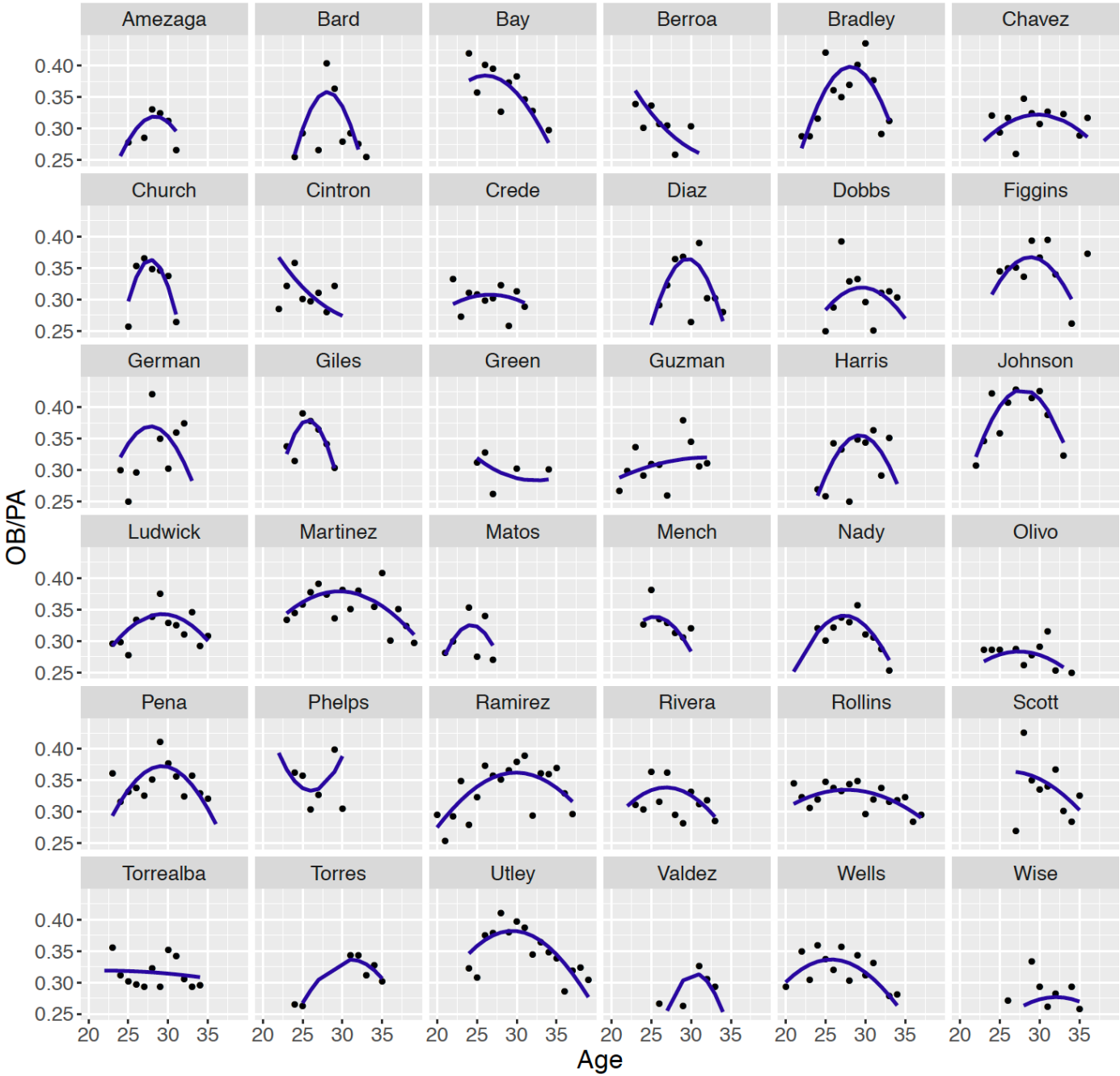
- Figure displays density estimates of the simulated values of $h_1(\beta)$ and p_{max}.
- Utley's peak performance was most likely achieved at age 29
- The posterior of the peak value p_{max} indicates that Utley's peak on-base probability ranged from 0.38 and 0.40.



Estimating many trajectories by a hierarchical model

- Want to simultaneously estimate the career trajectories for a group of players.
- Focus on the Major League players who were born in the year 1978 and had at least 1000 career at-bats.
- Figure on next slide displays scatterplots of age and OBP with quadratic smoothing curves for the 36 players in this group.
- Many of the curves follow a familiar concave down shape with the player achieving peak performance near an age of 30.
- But for some players, especially for those players who played a small number of seasons, note that the trajectories have different shapes.

Many Career Trajectories



Partially Pooling

- May be desirable to partially pool the data from the 36 players using a hierarchical model to obtain improved trajectory estimates for all players.
- For the i-th player, one observes the on-base events $\{y_{ij}\}$ which are Binomial with sample sizes $\{n_{ij}\}$ and probabilities of on-base success $\{p_{ij}\}$. Have logistic model

$$\log\left(\frac{p_{ij}}{1-p_{ij}}\right) = \beta_{i0} + \beta_{i1}(x_{ij} - 30) + \beta_{i2}(x_{ij} - 30)^2,$$

  where x_{ij} is the age of the i-th player during the j-th season.
- Want to estimating the regression vectors $(\beta_1, ..., \beta_N)$ for the N players in the study.

Hierarchical Prior

- Construct a two-stage prior on these regression vectors.
- At Stage 1, one assumes that $\beta_1, ..., \beta_N$ are independent distributed from a common multivariate Normal distribution with mean vector μ_β and precision matrix τ_β.
- At Stage 2, vague prior distributions are assigned to the unknown values of μ_β and τ_β.

R Work

- The JAGS script defining this model is a straightforward extension of the JAGS script for a logistic regression model for a single career trajectory.
- Variables are the player ids `player`, and the logistic parameters for the ith player are `beta0[i]`, `beta1[i]`, and `beta2[i]` represent the logistic regression parameters for
- The vector `B[j, 1:3]` represents a vector of parameters for one player and `mu.beta` and `Tau.B` represent respectively the second-stage prior mean vector and precision matrix values.
- Variables `mean`, `prec`, `Omega` are specified parameters reflectin weak information about the parameters at the second stage.

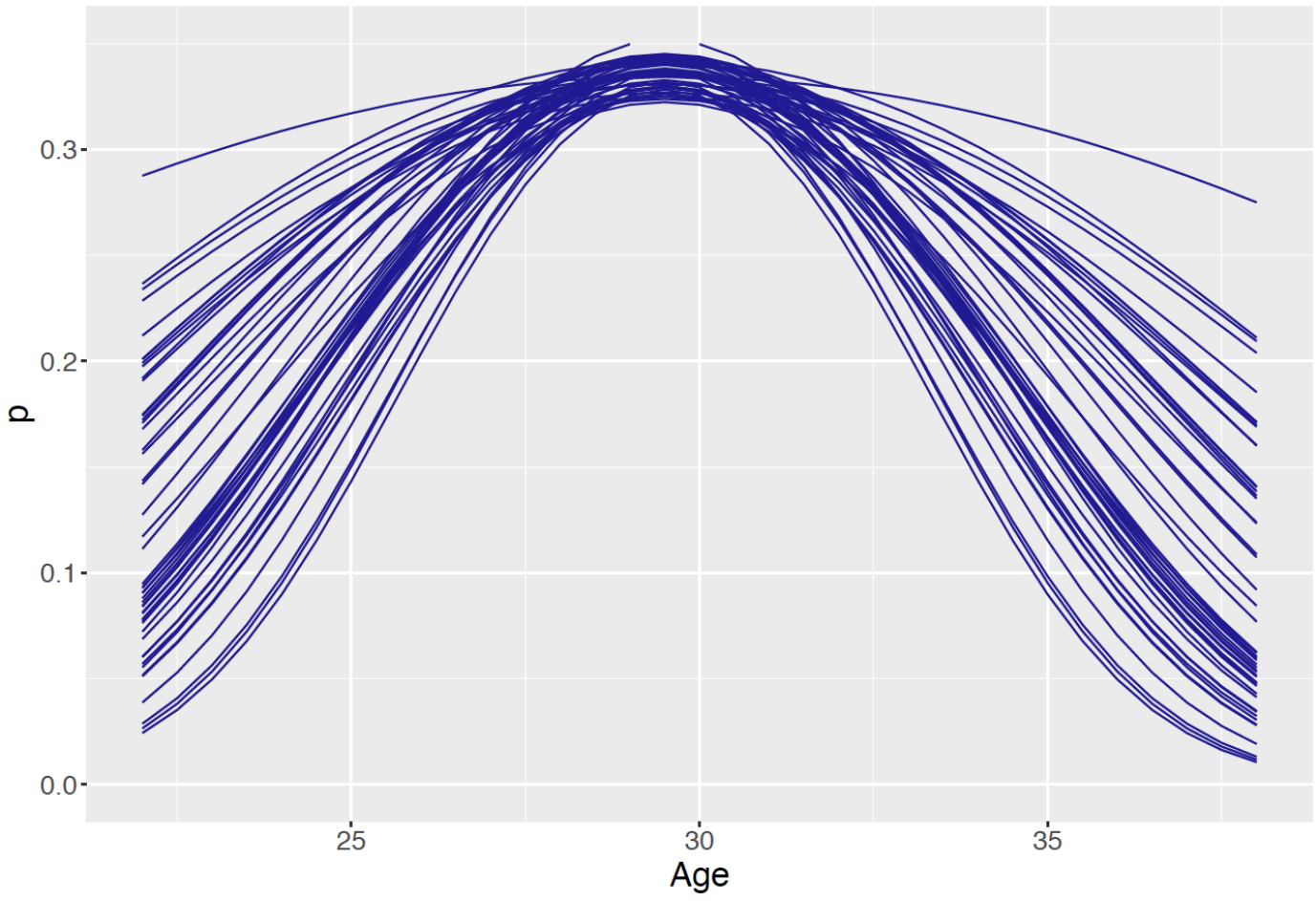
JAGS Script

```
modelString = "model {
for (i in 1:N){
   y[i] ~ dbin(p[i], n[i])
   logit(p[i]) <- beta0[player[i]] +
                  beta1[player[i]] * (x[i] - 30) +
                  beta2[player[i]] * (x[i] - 30) * (x[i]
}
for (j in 1:J){
   beta0[j] <- B[j,1]
   beta1[j] <- B[j,2]
   beta2[j] <- B[j,3]
   B[j,1:3] ~ dmnorm (mu.beta[], Tau.B[,])
}
mu.beta[1:3] ~ dmnorm(mean[1:3],prec[1:3 ,1:3 ])
Tau.B[1:3 , 1:3] ~ dwish(Omega[1:3 ,1:3 ], 3)
}
"
```

Posterior Inferences

- JAGS is used to simulate from the posterior distribution of this hierarchical model
- The player trajectories $\beta_1, ..., \beta_{36}$ are a sample from a Normal distribution with mean μ_β.
- Figure on next slide displays draws of the posterior of the mean peak age $h_1(\mu_\beta)$ expressed as probabilities over a grid of age values from 23 to 37.
- The takeaway if that a typical MLB player in this group peaks in on-base performance about age 29.5.

Posterior of mean peak age



Borrowing information

- Bayesian hierarchical model is helpful in borrowing information for estimating the career trajectories of players with limited career data.
- Figure on next page shows individual and hierarchical posterior mean fits of the career trajectories for two players.
- For Chase Utley, the two fits are very similar since Utley's career trajectory was well-estimated just using his data.
- In contrast, Phelps' career trajectory is corrected to be more similar to the concave down trajectory for most players.

Posterior distributions of 2 trajectories