Predicting HIV Infection Time for Infants

Maggie Russell, Carolyn Fish, Dara Lehman, Erick Matsen

May 21, 2020

My background:

- Graduated from Montana State University in Neuroscience

My background:

- Graduated from Montana State University in Neuroscience
- Spent a year (also at Montana State) in a Master's program for Math

My background:

- Graduated from Montana State University in Neuroscience
- ~~Spent a year (also at Montana State) in a Master's program for Math~~
- Here I am!

Predicting HIV Infection Time for Infants

Maggie Russell, Carolyn Fish, Dara Lehman, Erick Matsen

Why predict HIV infection time?

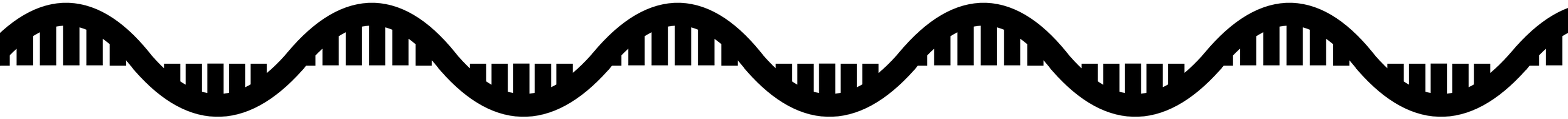
- HIV-1 establishes an infection which may last many years before diagnosis
- Three biomarkers have been identified to contain information about time since infection, but only distinguish between recent and long term infections or are imprecise
- Knowledge of true infection times would be helpful for learning about the true incidence of HIV-1 and viral pathogenesis
- Also, tracking time since HIV infection could be utilized for evaluating the efficacy of infection prevention methods such as HIV-1 vaccines and/or monoclonal antibodies

Rossenkhan, et. al. *Viruses*. 2019.

Puller, et. al. *PLOS Computational Biology*. 2017.

Carlisle, et. al. *The Journal of Infectious Diseases*. 2019.

HIV sequence



Average Pairwise Distance

Calculated from the frequencies $x_{i\alpha}$ of each nucleotide
$\alpha \in \{A, C, G, T\}$ at site $i = 1 \ldots L$

Average pairwise distance is the probability that **two randomly drawn sequences** have **different nucleotides** at a specific position, averaged over all positions

Average Pairwise Distance

Calculated from the frequencies $x_{i\alpha}$ of each nucleotide $\alpha \in \{A, C, G, T\}$ at site $i = 1 \ldots L$

Average pairwise distance is the <u>probability</u> that **two randomly drawn sequences** have **different nucleotides** at a specific position, <u>averaged over all positions</u>

$$\frac{1}{L} \sum_{i=1}^{L} \left[\sum_{\alpha} x_{i\alpha}(1 - x_{i\alpha}) \right]$$

Frequency Cutoff

x_i^m is the frequency of the dominant nucleotide at position i

The sum of all minor variants, $(1 - x_i^m)$, must be greater than a cutoff, x_c, to contribute to the diversity measure

Frequency Cutoff

x_i^m is the frequency of the dominant nucleotide at position i

The sum of all minor variants, $(1 - x_i^m)$, must be greater than a cutoff, x_c, to contribute to the diversity measure

$$\Theta(1 - x_i^m - x_c) = \begin{cases} 1 & \text{if } 1 - x_i^m > x_c \\ 0 & \text{if } 1 - x_i^m \leq x_c \end{cases}$$

Average Pairwise Diversity (APD)

Frequency Cutoff

$$\Theta(1-x_i^m-x_c)$$

Average Pairwise Distance

$$\frac{1}{L}\sum_{i=1}^{L}\Big[\sum_{\alpha}x_{i\alpha}(1-x_{i\alpha})\Big]$$

Average Pairwise Diversity (APD)

Frequency Cutoff

$$\Theta(1-x_i^m-x_c)$$

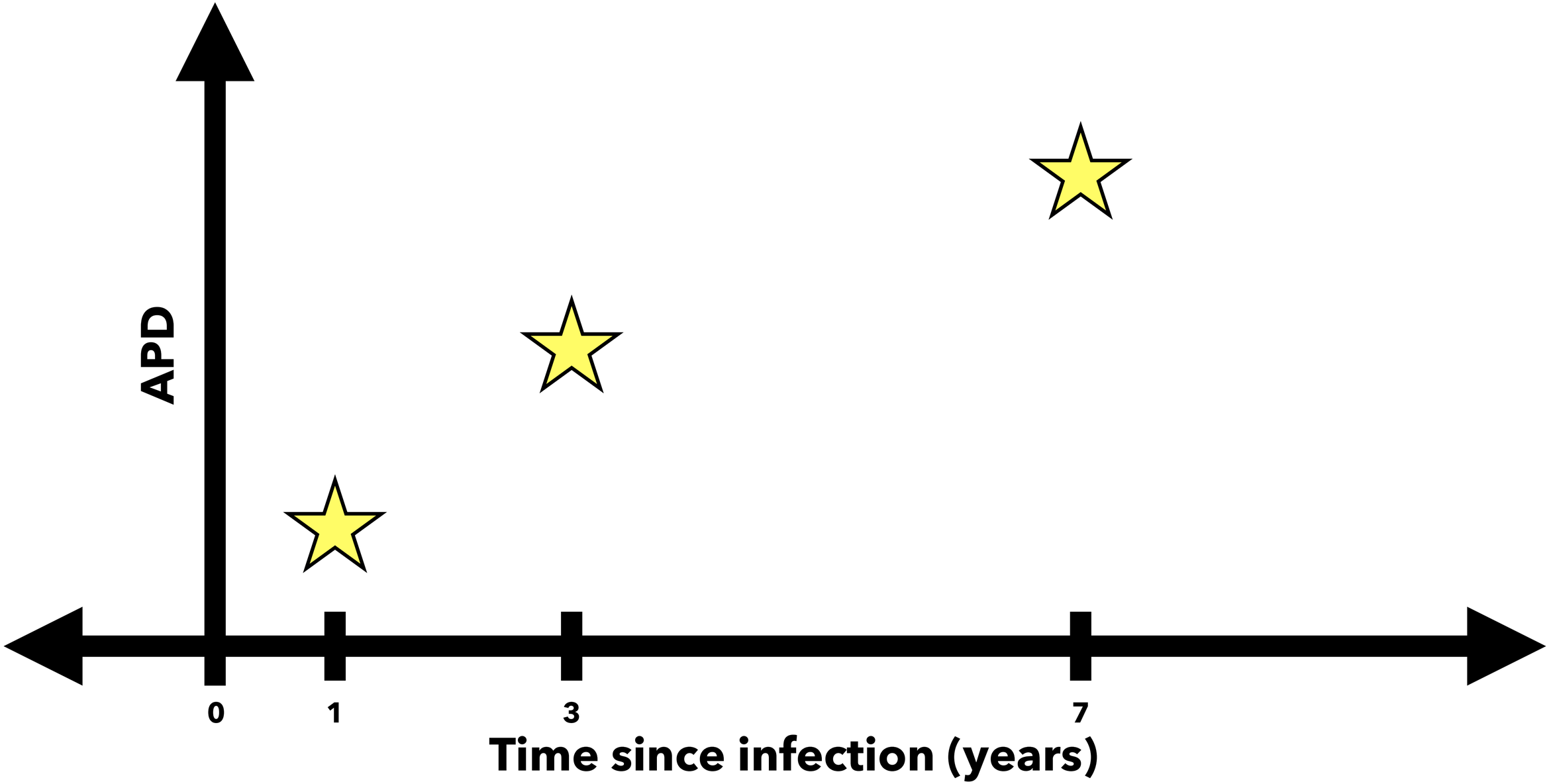
Average Pairwise Distance

$$\frac{1}{L}\sum_{i=1}^{L}\Big[\sum_{\alpha}x_{i\alpha}(1-x_{i\alpha})\Big]$$

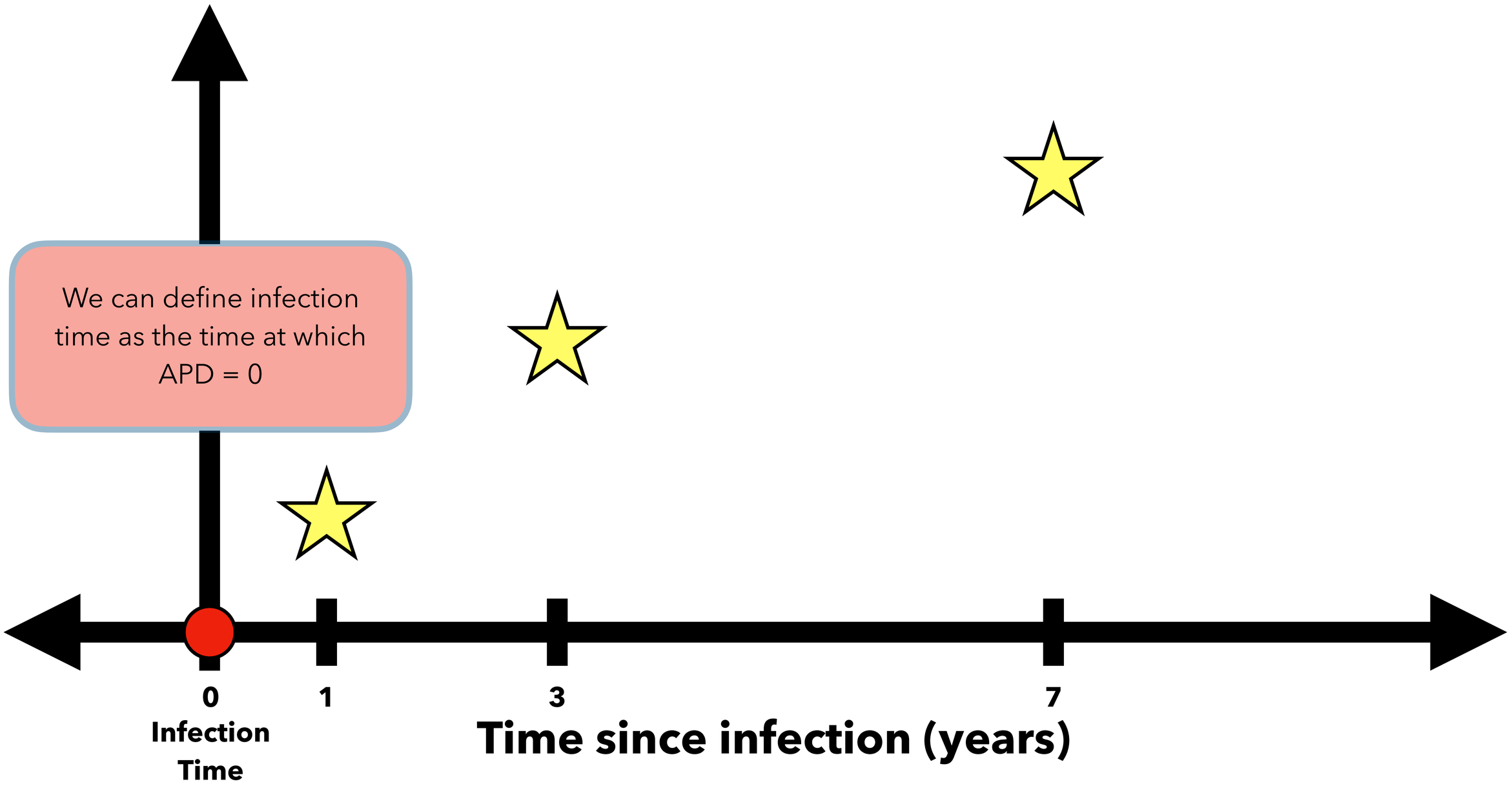
$$APD=\frac{1}{L}\sum_{i=1}^{L}\Theta(1-x_i^m-x_c)\Big[\sum_{\alpha}x_{i\alpha}(1-x_{i\alpha})\Big]$$

APD can be calculated at different time points

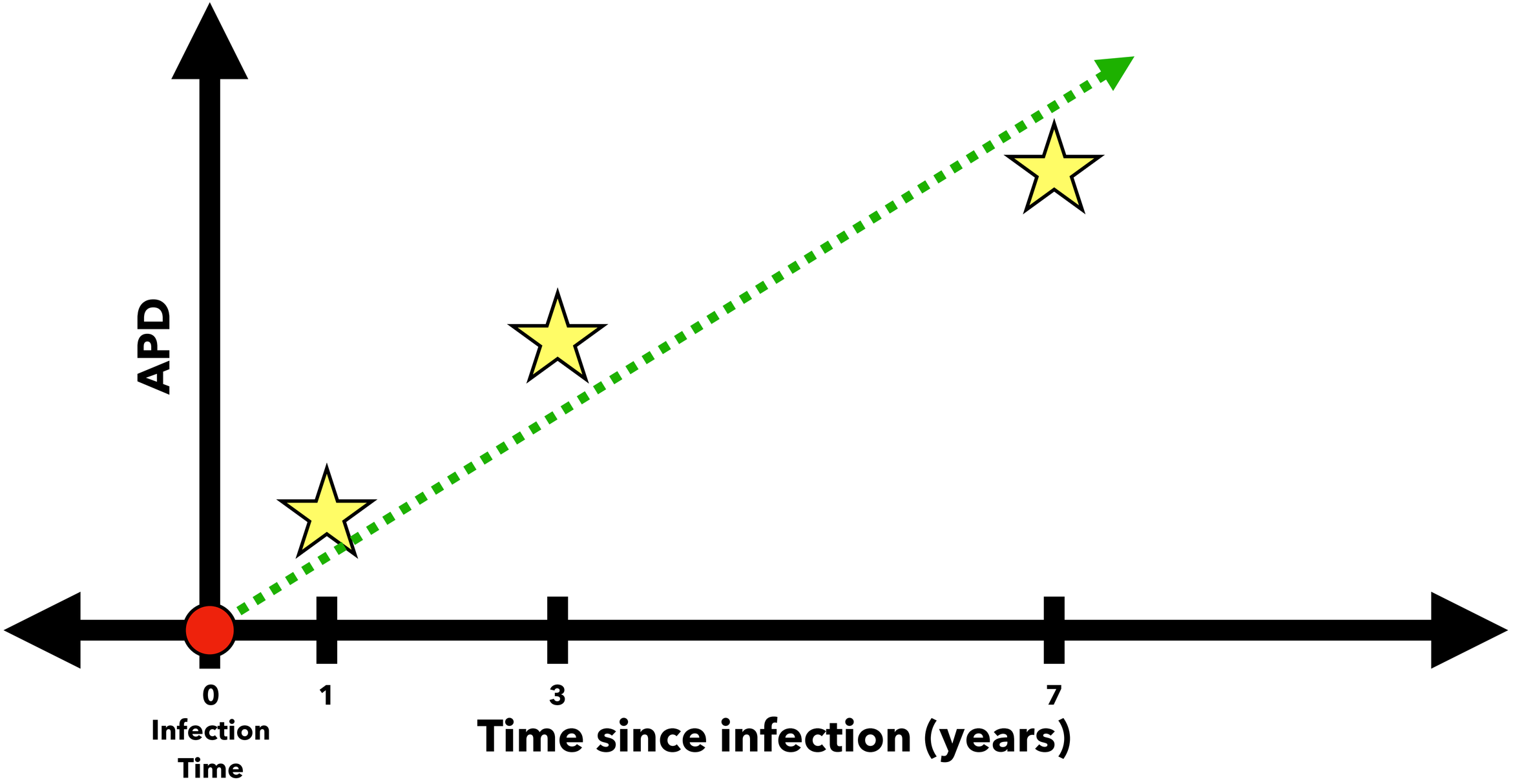
APD increases with Time



APD increases with Time



APD increases with Time

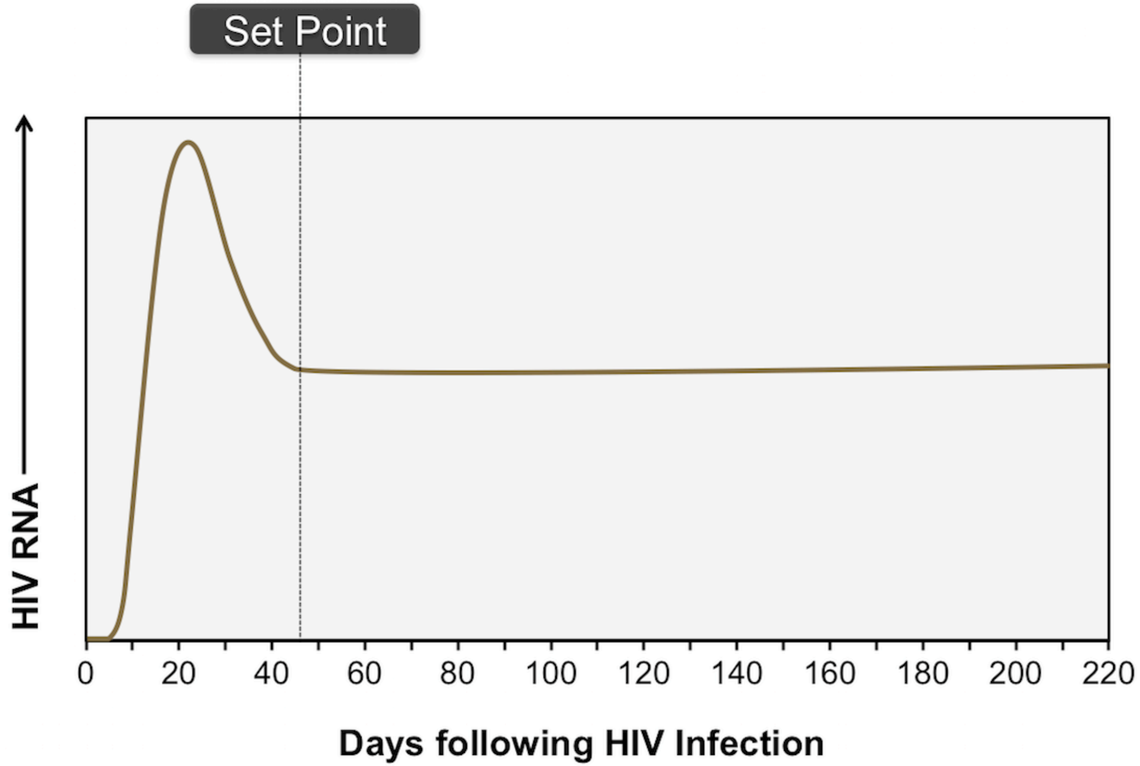


Why a linear model?

1. HIV sequences recombine extremely frequently

2. HIV evolutionary rates are not constant through time

3. Data is often sparse

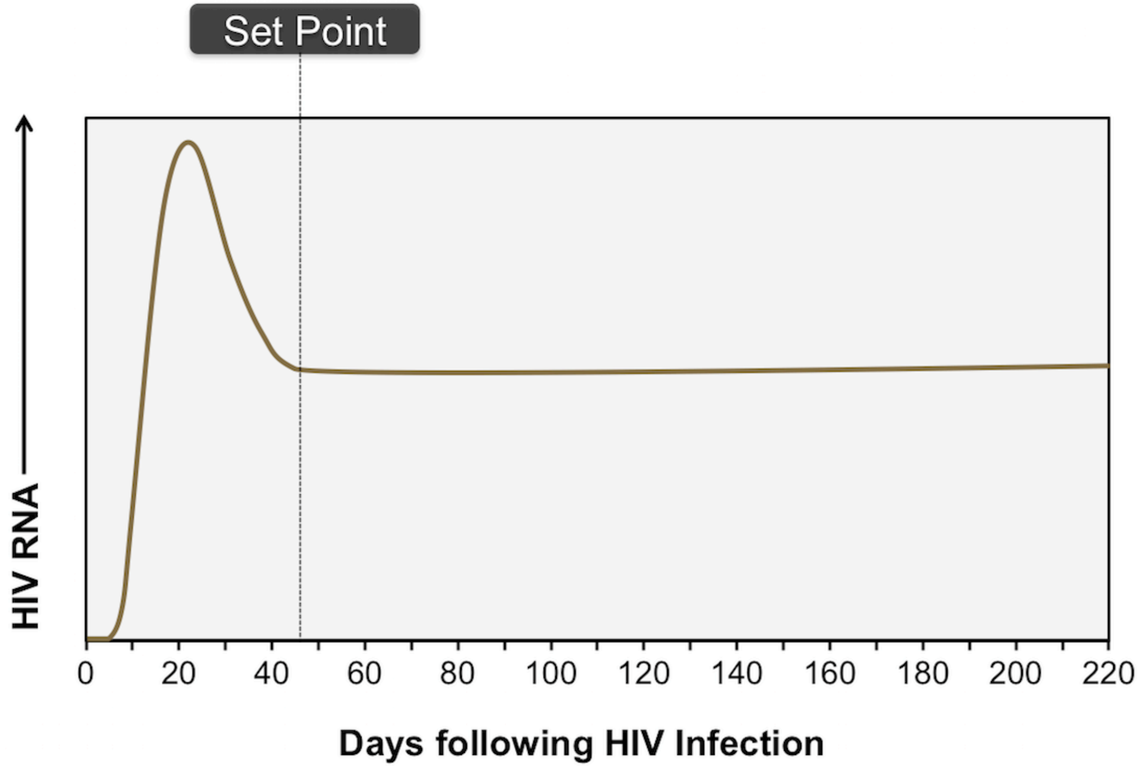


Why a linear model?

1. HIV sequences recombine extremely frequently

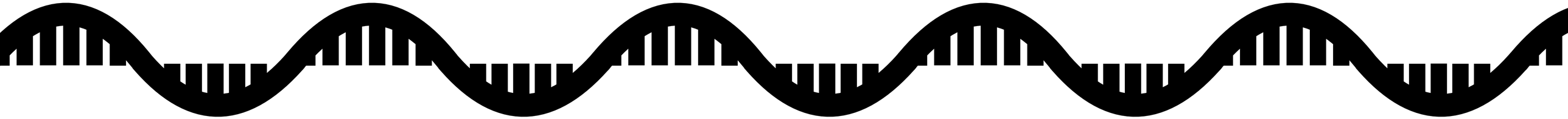
2. HIV evolutionary rates are not constant through time

3. Data is often sparse

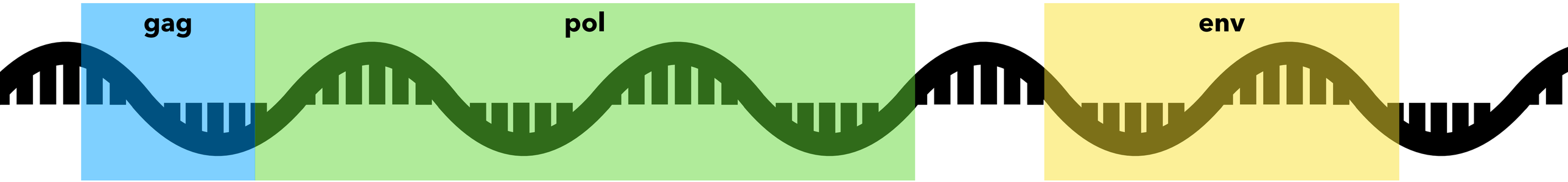


Want to be able to set priors and take a regression approach

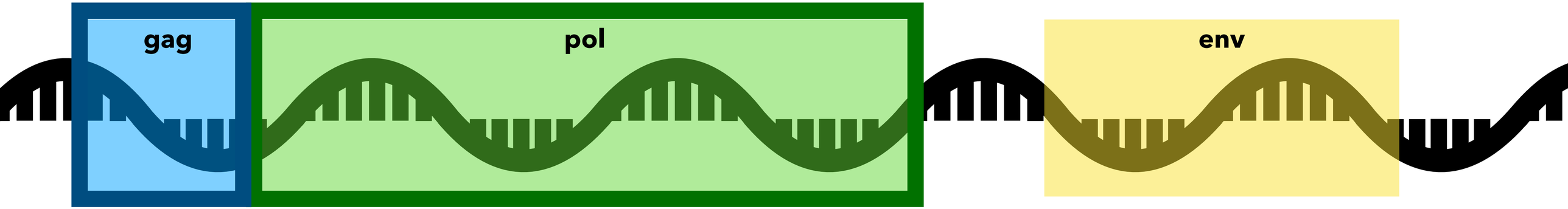
HIV sequence



HIV sequence



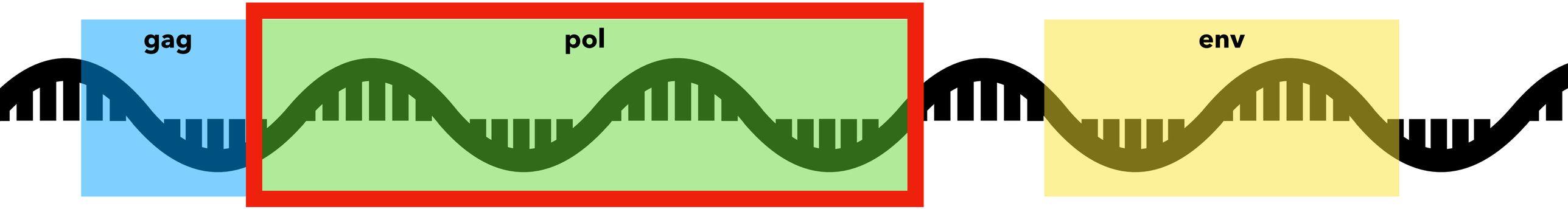
HIV sequence



Puller, et. al. *PLOS Computational Biology*. 2017.

Carlisle, et. al. *The Journal of Infectious Diseases*. 2019.

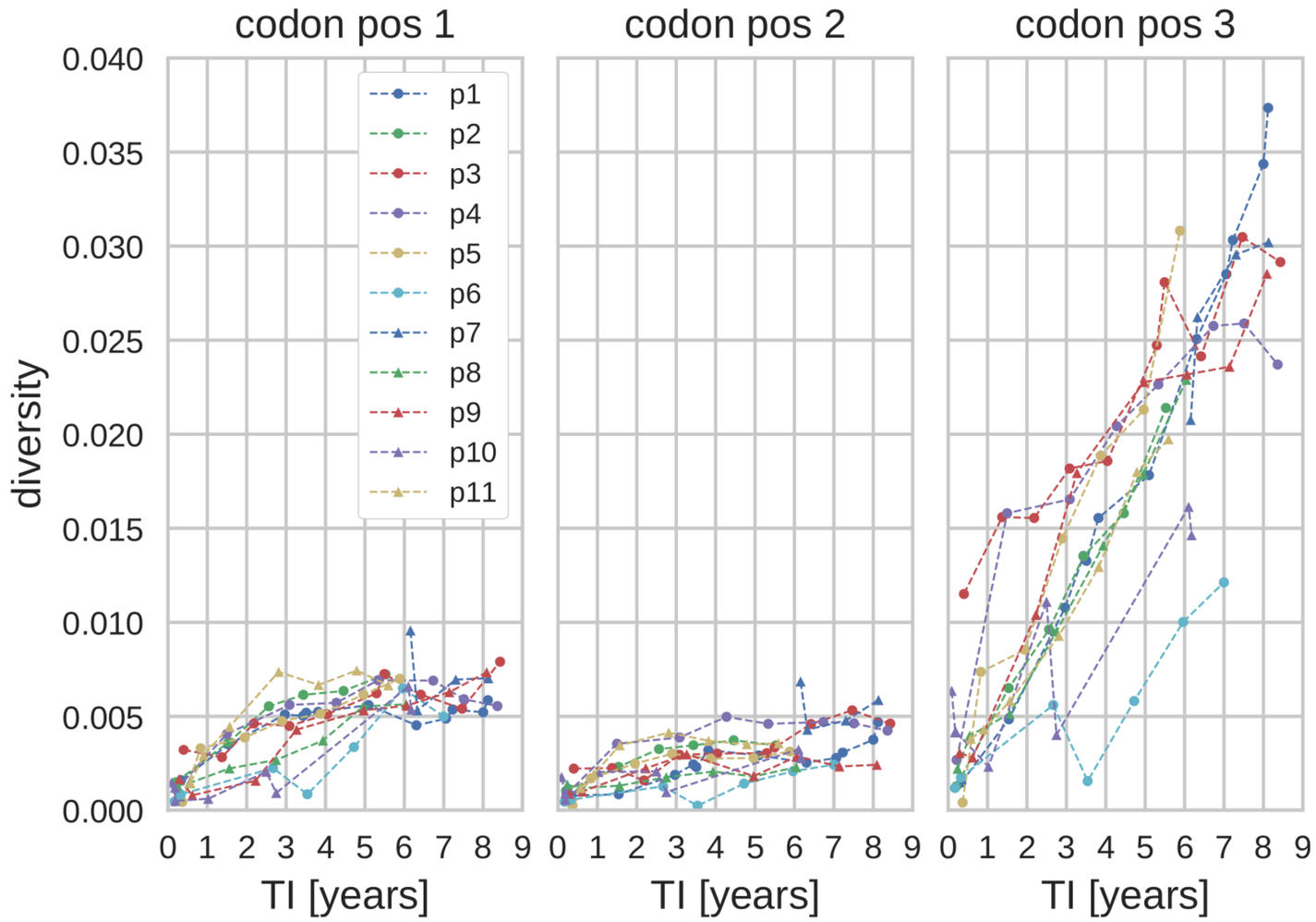
HIV sequence



Puller, et. al. *PLOS Computational Biology*. 2017.

APD increases with Time in Adults

Particularly at 3rd codon position



Puller, et. al. *PLOS Computational Biology*. 2017.

Puller et. al. models time since infection using APD

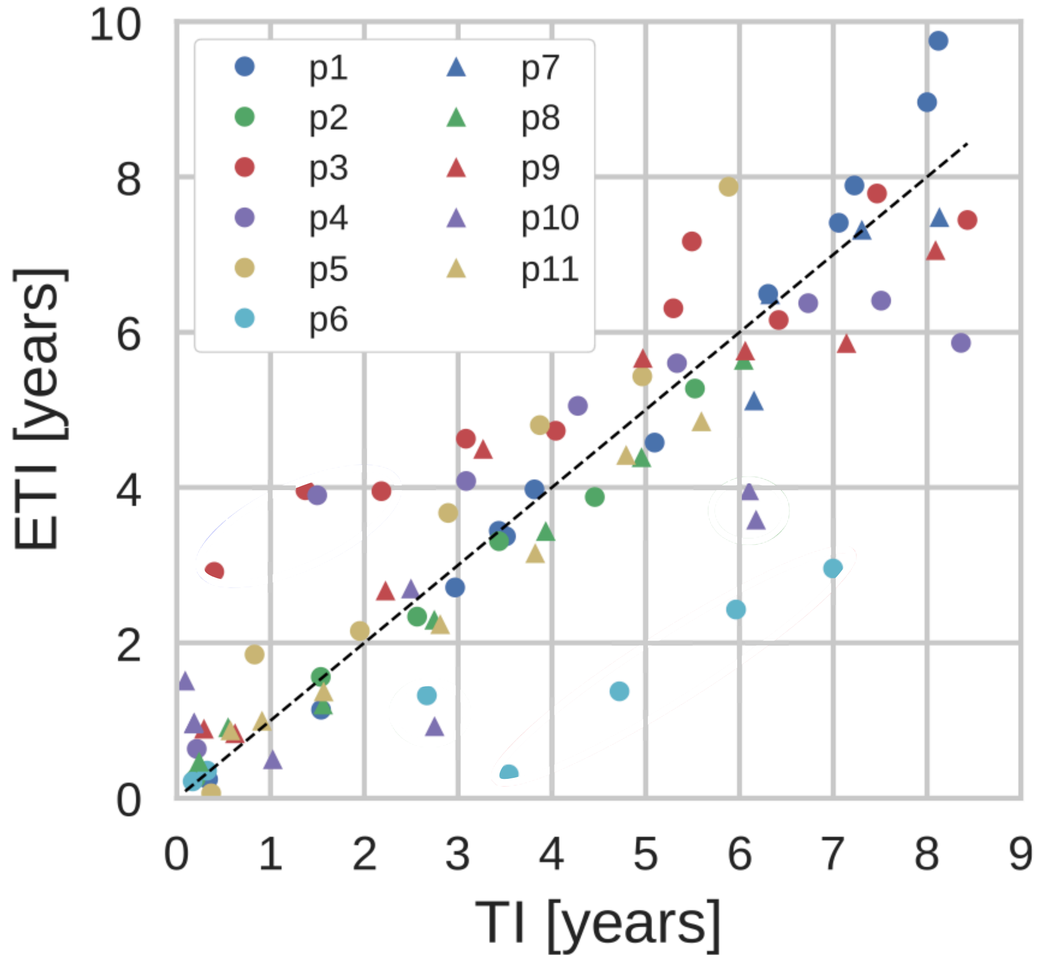
Given an APD measure, they model time since infection (TI) by

$$TI = m * APD + t_0$$

Slope, m , and intercept, t_0 , were estimated by **least absolute deviation (LAD) regression**

Puller, et. al. *PLOS Computational Biology*. 2017.

Puller et. al. models time since infection using APD



Puller, et. al. *PLOS Computational Biology*. 2017.

Why predict HIV infection time for infants?

- When a mother is HIV+, infants can be infected in utero, during birth, or through breastfeeding post birth
- HIV-infected infants typically progress to disease faster than adults, perhaps in part due to an immature immune system
- Infants can produce broadly neutralizing antibodies capable of neutralizing a range of HIV isolates earlier on in the infection compared to adults
- The ability to track time since infection in infants (as well as adults) could shed light on differences in infection response and progression between infants and adults

Goo, et. al. *Nature Medicine*. 2014.
Simonich, et. al. *Nature Communications*. 2019.

Infant dataset

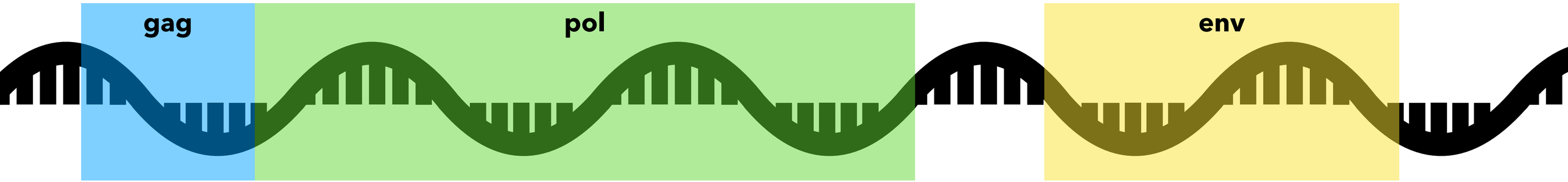
- 11 HIV+ infants
- These individuals were infected in utero (HIV+ at birth)
- Next generation sequencing data at 3-4 timepoints post birth

Infant dataset

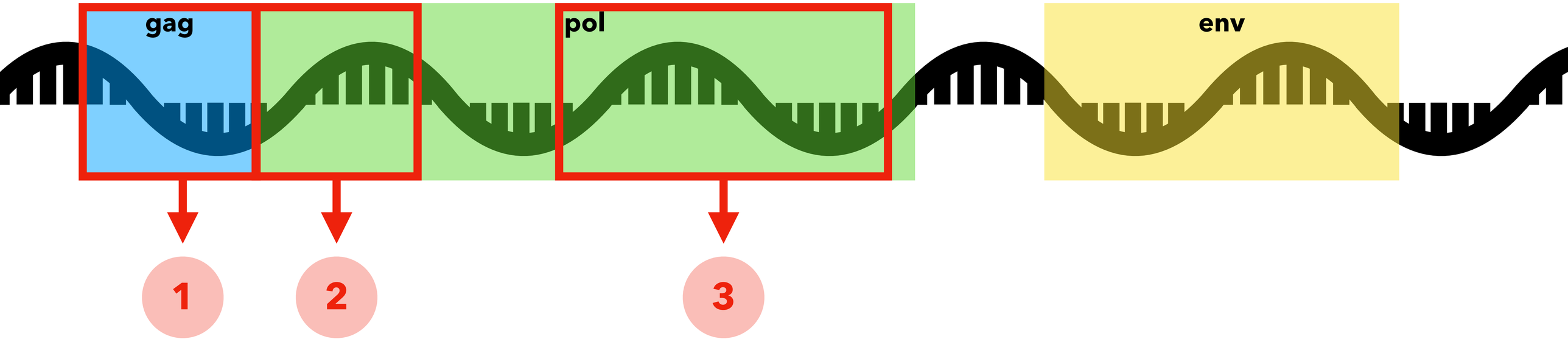
- 11 HIV+ infants
- These individuals were infected in utero (HIV+ at birth)
- Next generation sequencing data at 3-4 timepoints post birth

Exact infection time is not known, so for simplicity we will define **time of infection birth**

HIV sequence



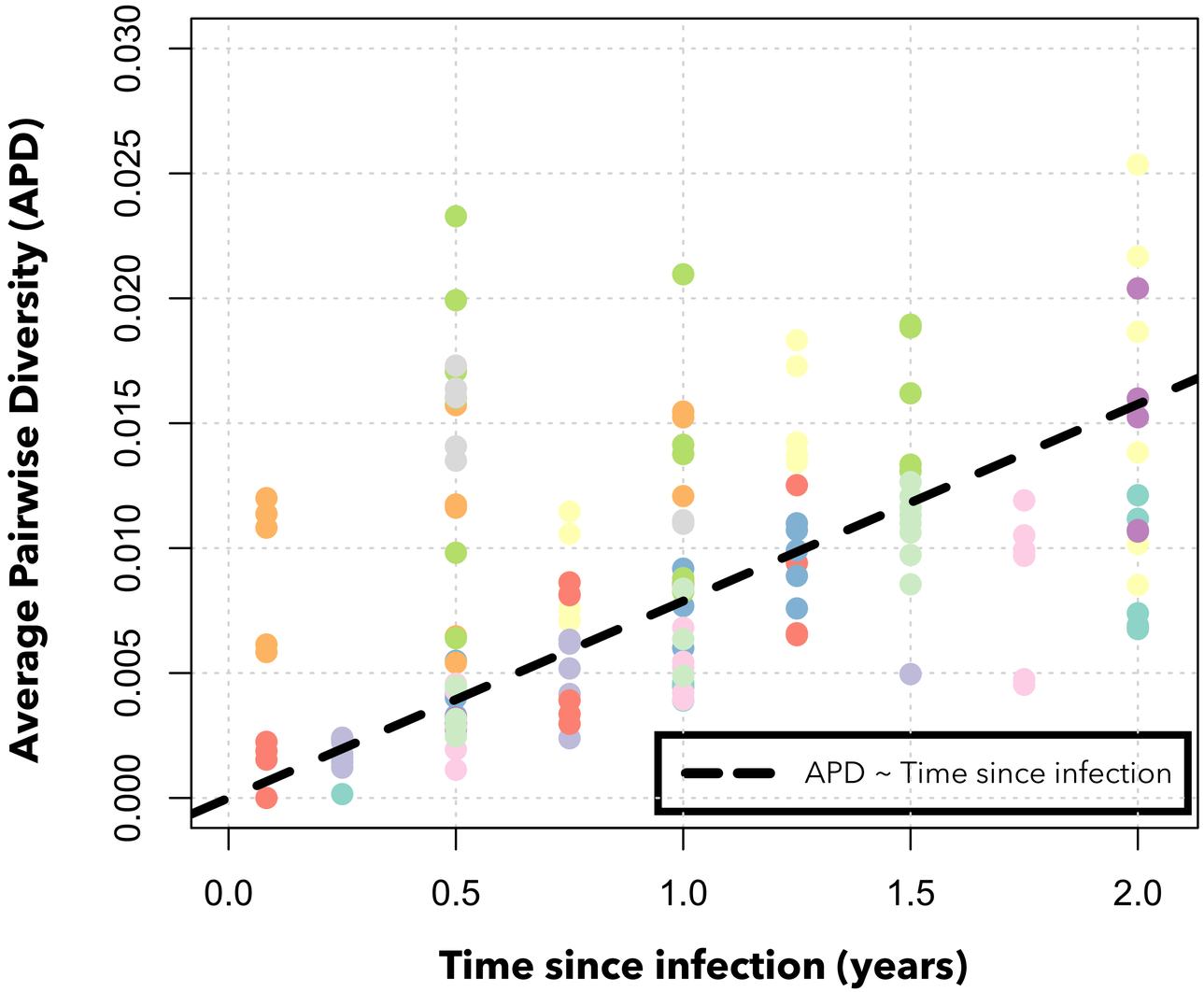
HIV sequence



Puller, et. al. *PLOS Computational Biology*. 2017.

Carlisle, et. al. *The Journal of Infectious Diseases*. 2019.

APD increases with Time in Infants (mostly)



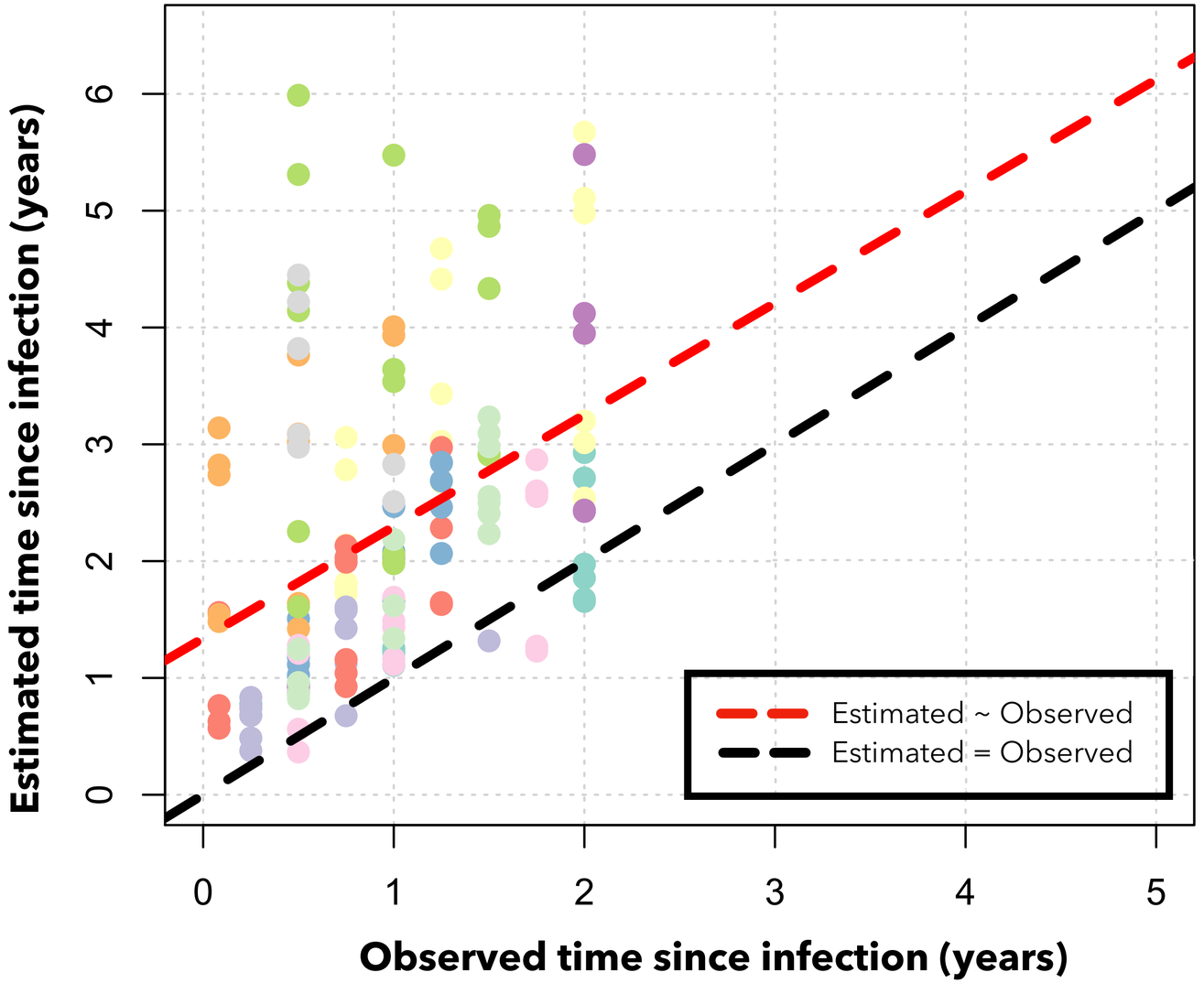
Approach:

1. Does the Puller et. al. adult model effectively predict time since infection for our **infant cohort**?

2. Does an infant specific linear regression model effectively predict time since infection for our **infant cohort**?

3. **What about a multilevel model?**

Model **overestimates** time since infection for infants



Approach:

1. Does the Puller et. al. adult model effectively predict time since infection for our **infant cohort**? **Not really**

2. Does an infant specific linear regression model effectively predict time since infection for our **infant cohort**?

3. **What about a multilevel model?**

Approach:

1. Does the Puller et. al. adult model effectively predict time since infection for our **infant cohort**? **Not really**

2. Does an infant specific linear regression model effectively predict time since infection for our **infant cohort**?

3. **What about a multilevel model?**

Here is what we can assume:

- “Training data” individuals were HIV+ at birth
- Time is time since infection
- Time zero is birth (for now)
- Infection time is defined as the time at which APD is zero

Infant linear regression model framework:

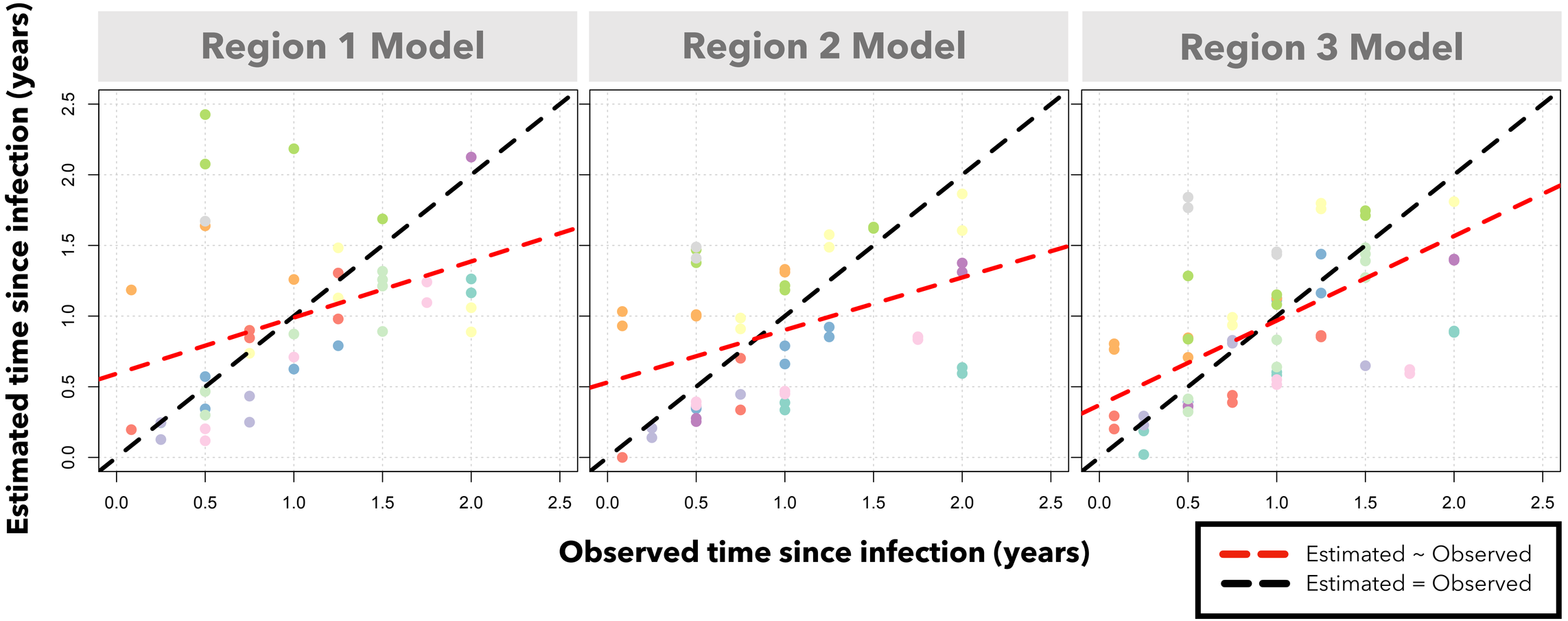
- "Training data" individuals were HIV+ at birth
- Time is time since infection
- Time zero is birth (for now)
- Infection time is defined as the time at which APD is zero

Given an APD measure, we will model time since infection (TI) by

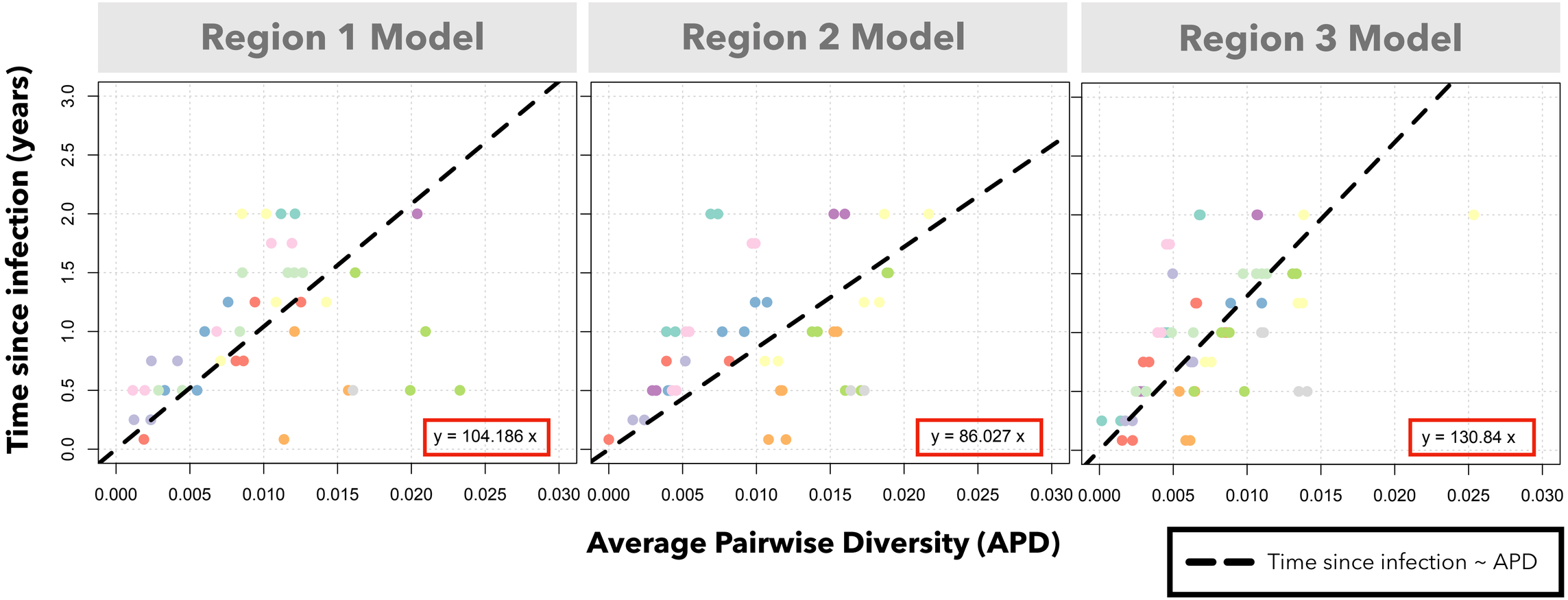
$$TI = m * APD + t_0$$

Slope, m, and intercept, t_0, will be estimated by **least absolute deviation (LAD) regression** for each sequence region

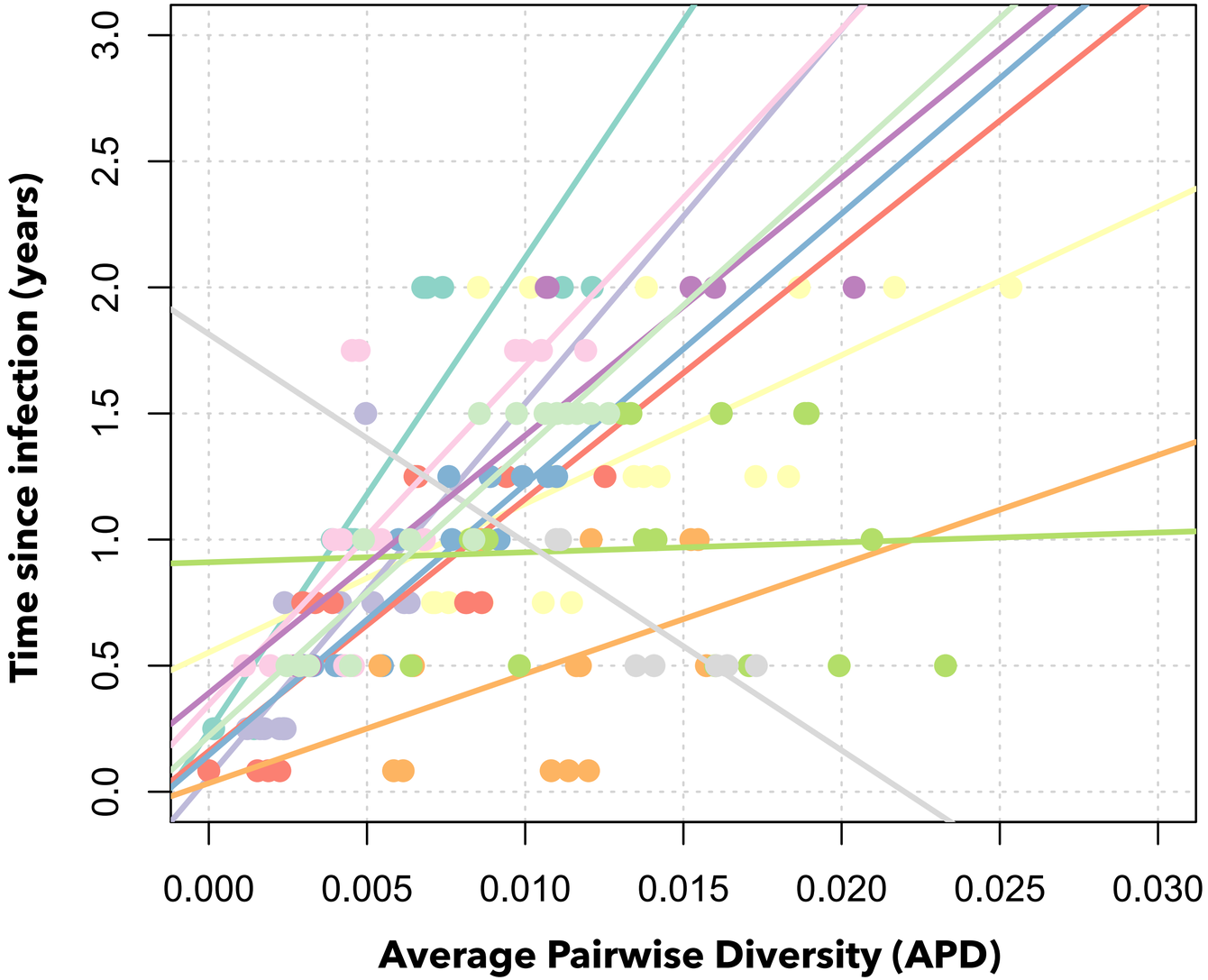
Model **overestimates** small time and **underestimates** large time for all sequence regions



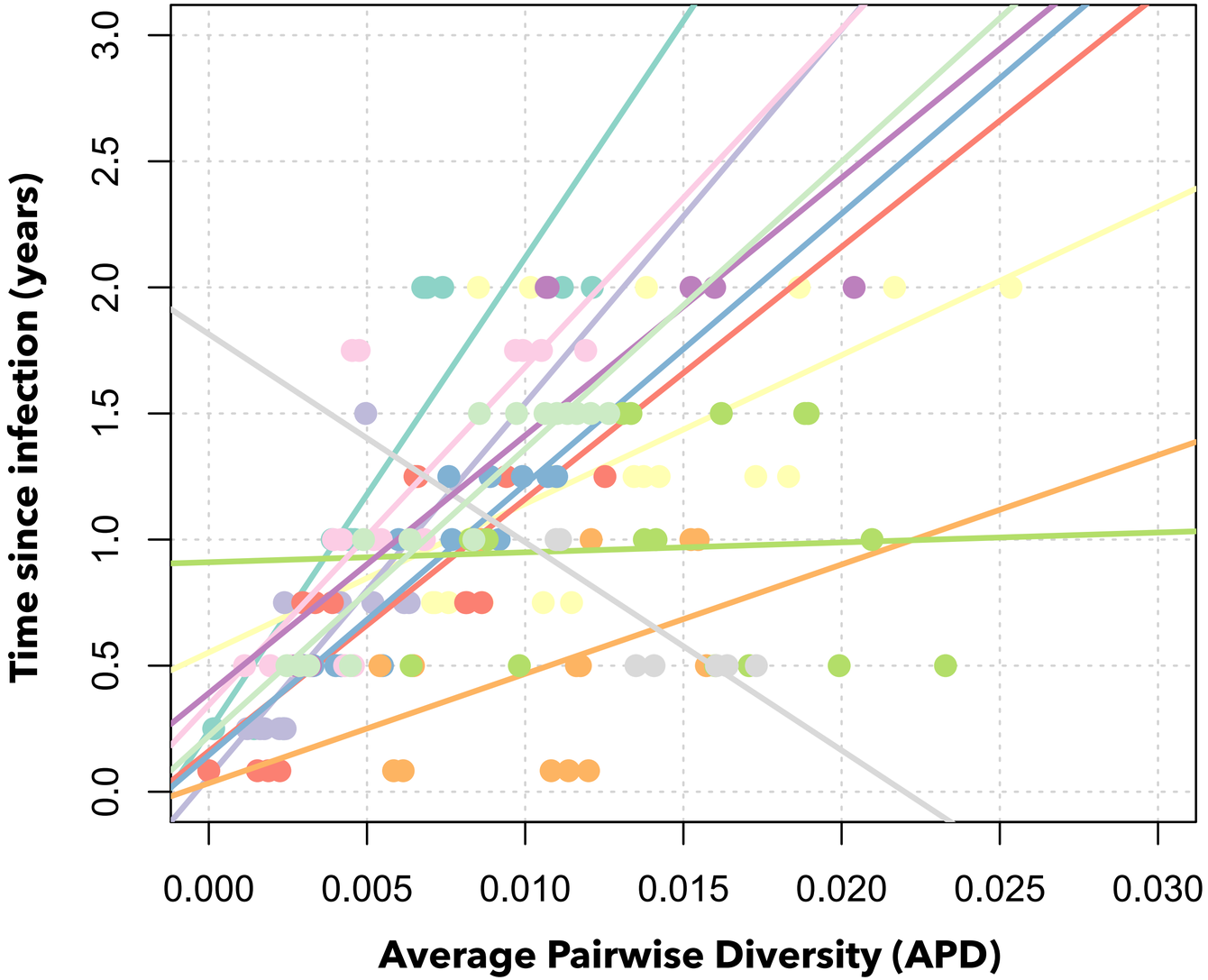
Each sequence region yields a **different regression slope**

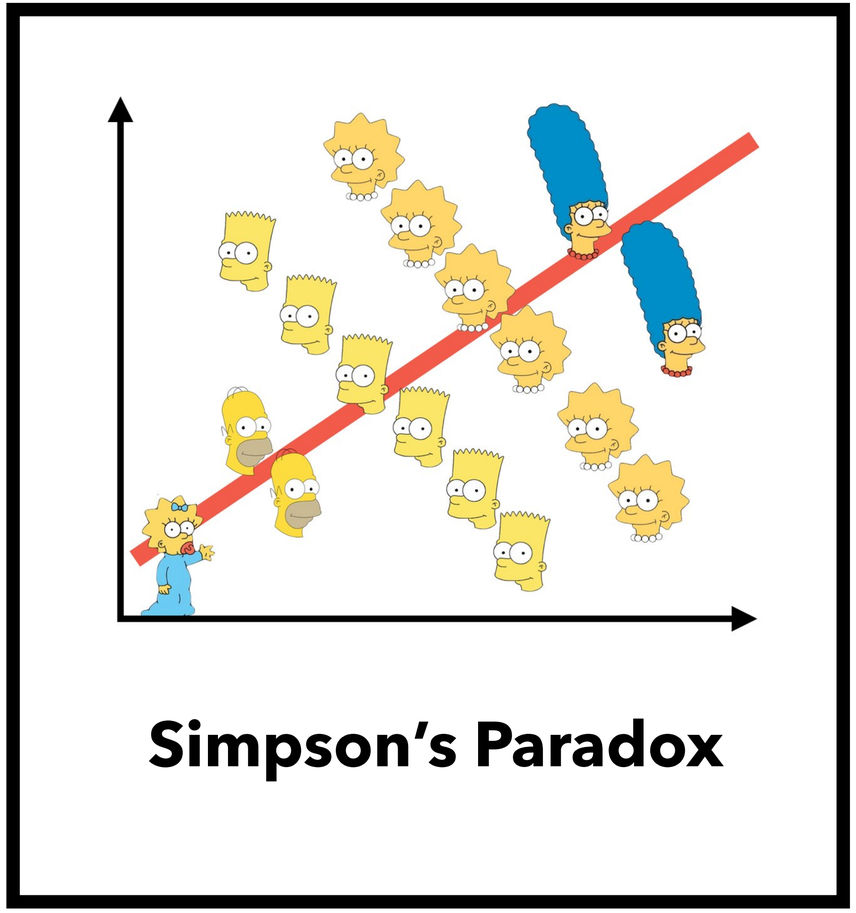


Each patient yields a **different regression slope**



Each patient yields a **different regression slope**





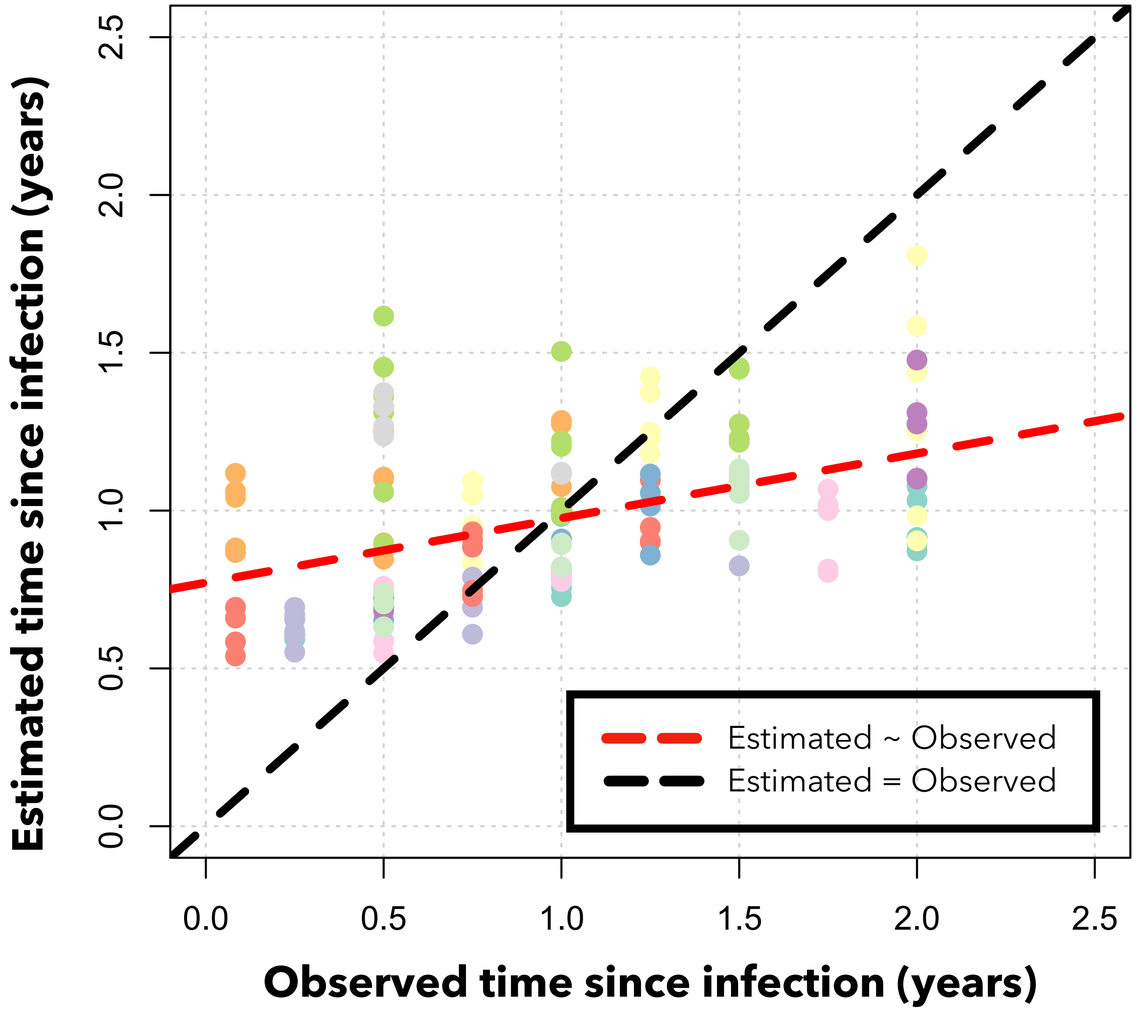
Generalized Estimating Equation approach?

GEEs are used to estimate the parameters of a generalized linear model with a **possible unknown correlation between outcomes**

Here, correlations within **patients** and **sequence regions**

Generalized Estimating Equation approach?

Similar overestimation/underestimation pattern



Approach:

1. Does the Puller et. al. adult model effectively predict time since infection for our **infant cohort**? **Not really**

2. Does an infant specific linear regression model effectively predict time since infection for our **infant cohort**? **Also, not really**

3. **What about a hierarchical model?**

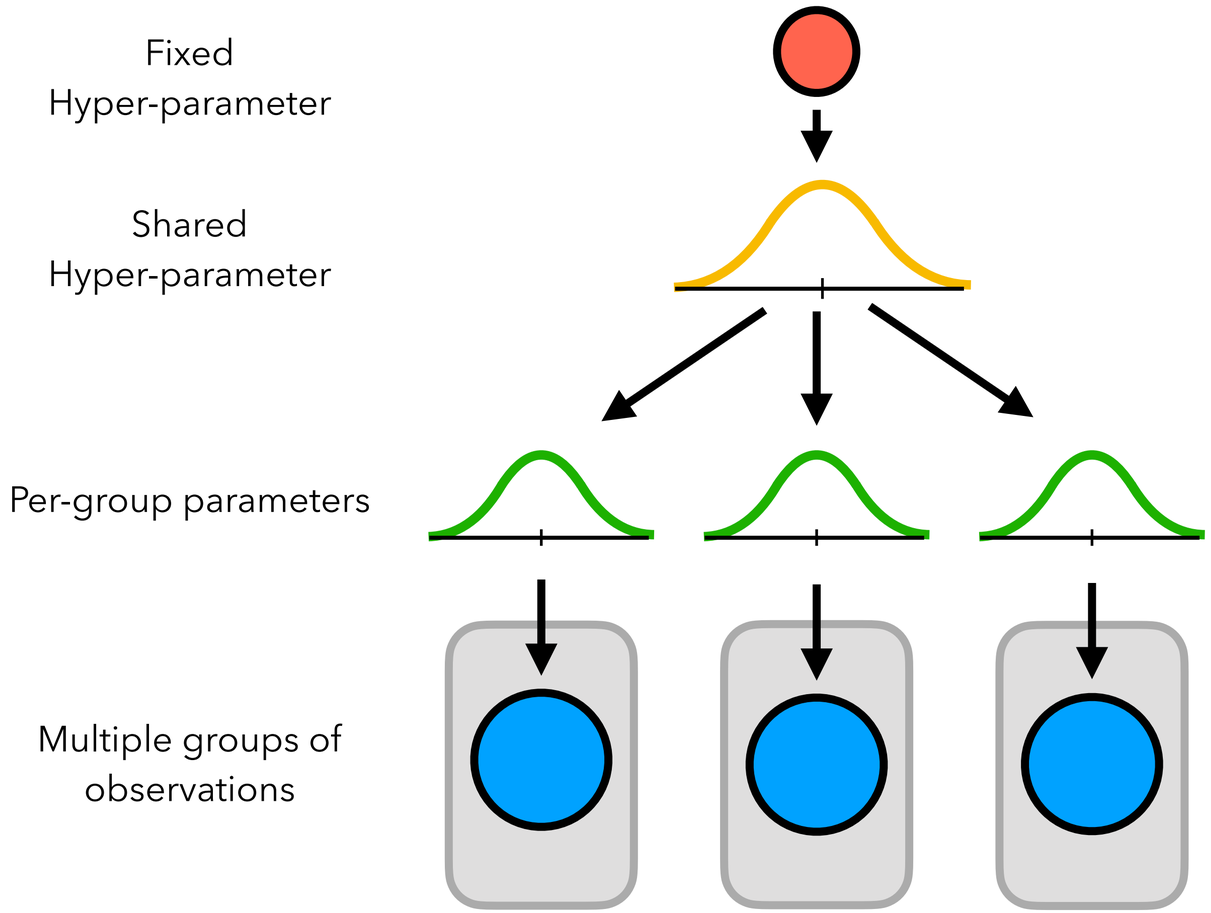
Approach:

1. Does the Puller et. al. adult model effectively predict time since infection for our **infant cohort**? **Not really**

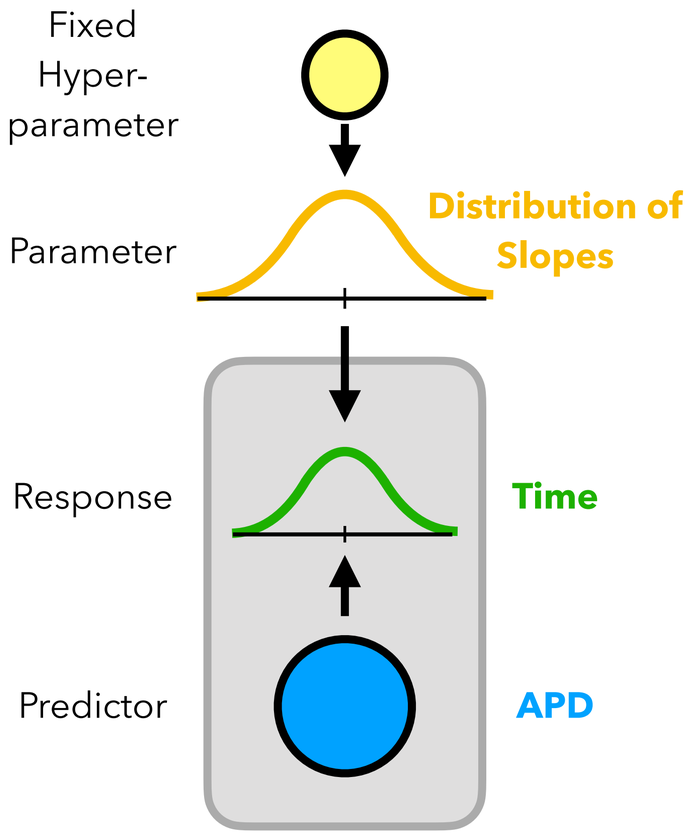
2. Does an infant specific linear regression model effectively predict time since infection for our **infant cohort**? **Also, not really**

3. **What about a hierarchical model?**

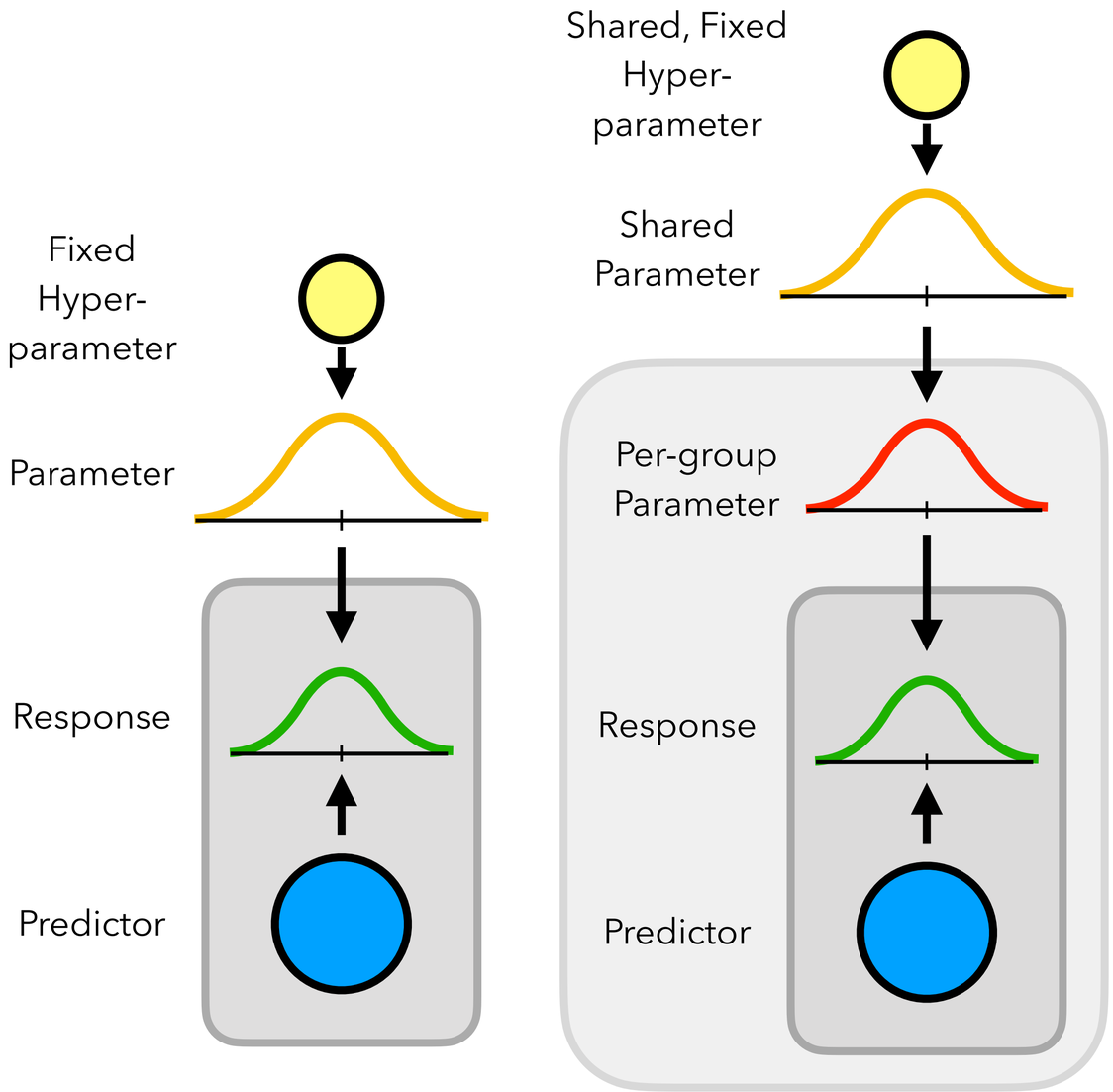
What is a hierarchical model?



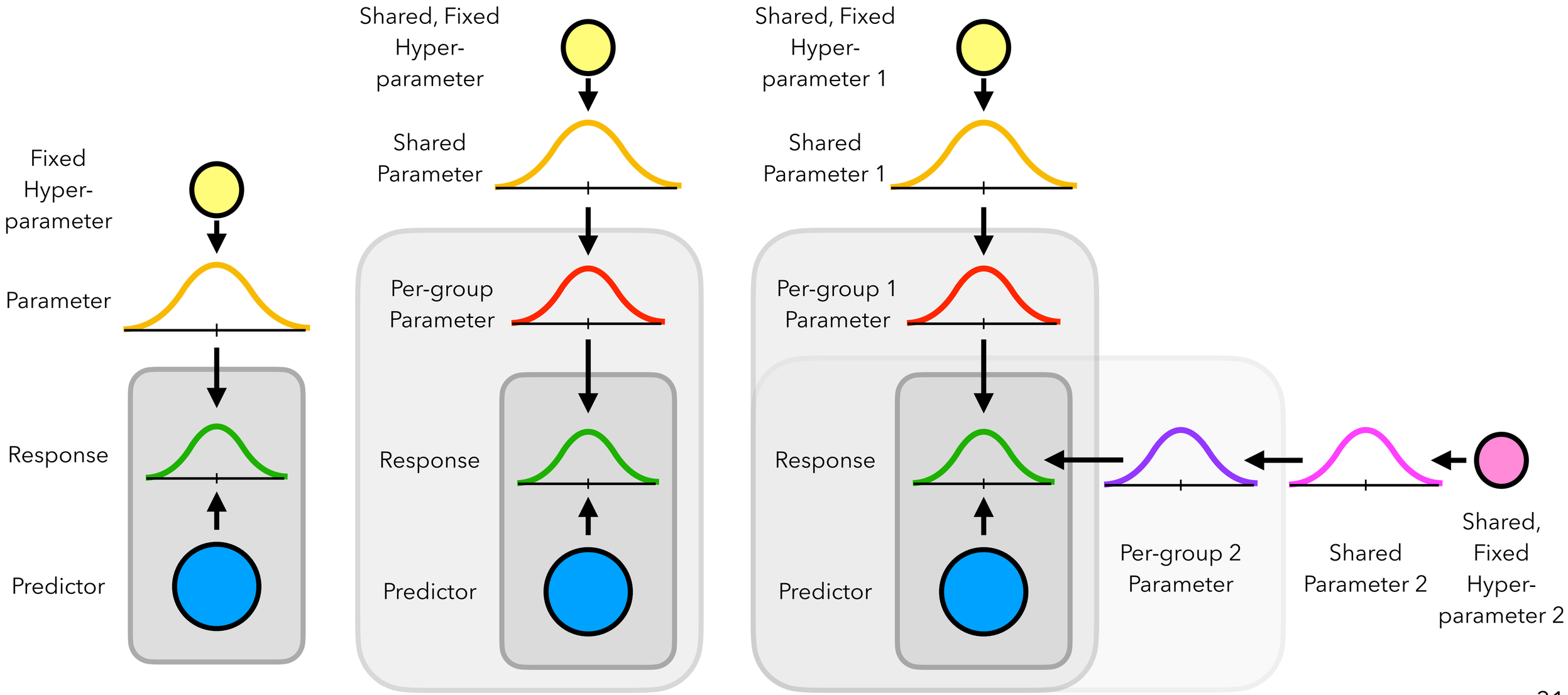
What is a hierarchical regression model?



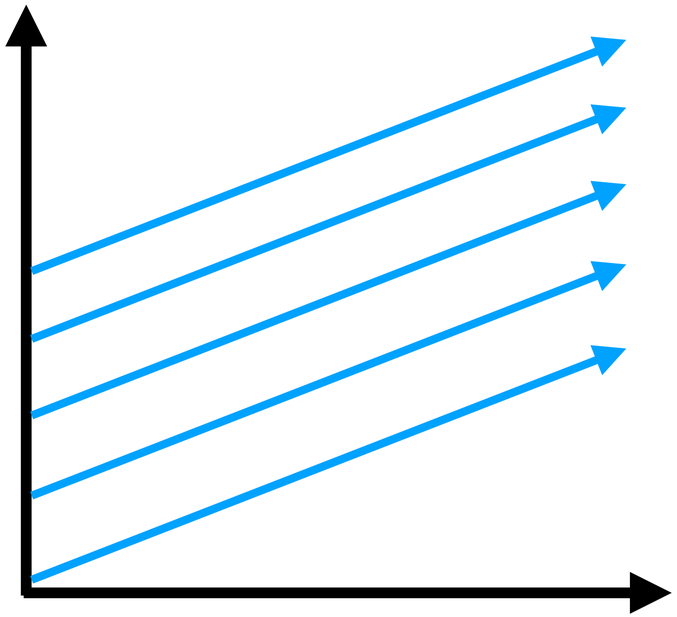
What is a hierarchical regression model?



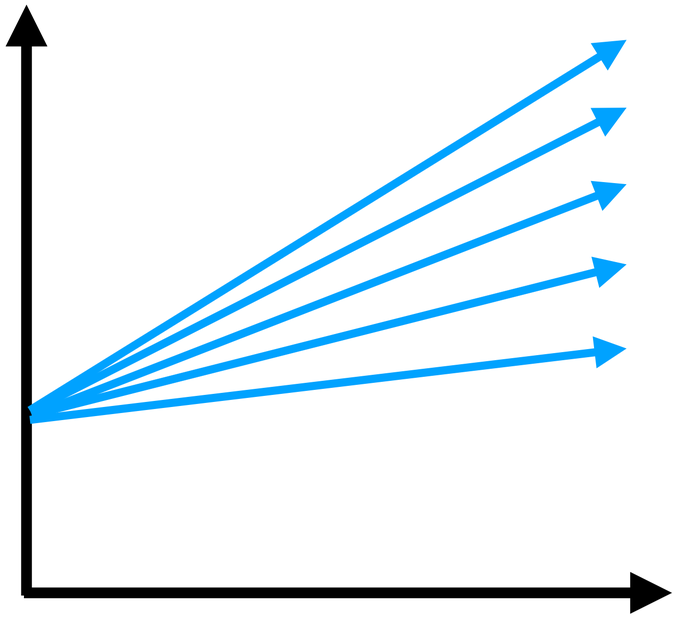
What is a hierarchical regression model?



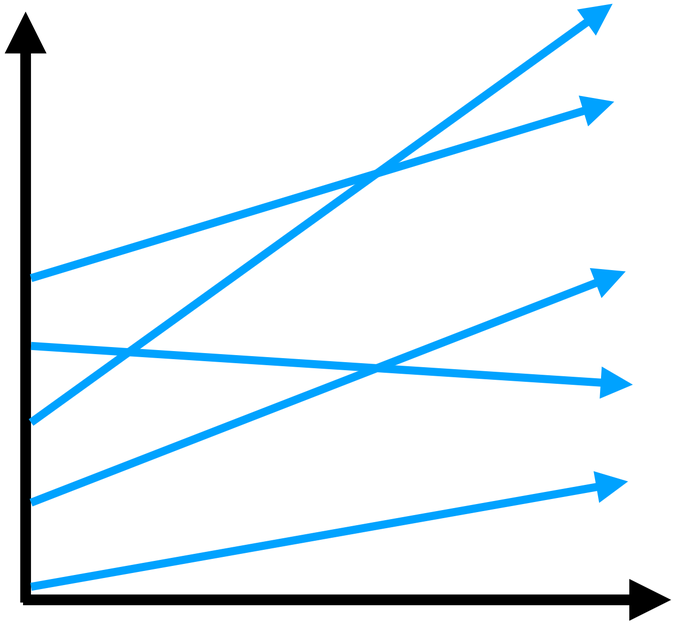
What is a hierarchical regression model?



If we vary intercepts by groups



If we vary slopes by groups



If we vary slopes **and** intercepts by groups

Here is what we can assume:

- “Training data” individuals were HIV+ at birth
- Time is time since infection
- Time zero is birth
- Infection time is defined as the time at which APD is zero
- The rate of APD change over time may be different for each individual
- The rate of APD change over time may be different for each sequence region

Here is what we can assume:

- "Training data" individuals were HIV+ at birth
- Time is time since infection
- Time zero is birth
- Infection time is defined as the time at which APD is zero

For now, we will restrict the intercept to be zero

- The rate of APD change over time may be different for each individual
- The rate of APD change over time may be different for each sequence region

Here is what we can assume:

- "Training data" individuals were HIV+ at birth
- Time is time since infection
- Time zero is birth
- Infection time is defined as the time at which APD is zero

For now, we will restrict the intercept to be zero

- The rate of APD change over time may be different for each individual

Varying slopes by subject

- The rate of APD change over time may be different for each sequence region

Varying slopes by sequence region

Hierarchical model framework:

We model the function (APD) -> (time) as a linear function.

$$\text{time} \sim Normal(\text{mean}, \text{sd})$$

$$\text{mean} \leftarrow \text{total\_intercept} + \text{total\_slope} * APD$$

Hierarchical model framework:

We model the function (APD) -> (time) as a linear function.

$$\text{time} \sim Normal(\text{mean}, \text{sd})$$

$$\text{mean} \leftarrow \text{total\_intercept} + \text{total\_slope} * APD$$

We model the total slope of this function to be average slope + individual subject variable slope + sequence region variable slope

$$\text{total\_slope} \leftarrow \text{avg\_slope} + \text{subject\_slope}\ [\text{subject index}] + \text{region\_slope}\ [\text{region index}]$$

Hierarchical model framework:

We model the function (APD) -> (time) as a linear function.

We model the total slope of this function to be average slope + individual subject variable slope + sequence region variable slope

$$\text{time} \sim Normal(\text{mean}, \text{sd})$$

$$\text{mean} \leftarrow \text{total\_intercept} + \text{total\_slope} * APD$$

$$\text{total\_slope} \leftarrow \text{avg\_slope} + \text{subject\_slope}\ [\text{subject index}] + \text{region\_slope}\ [\text{region index}]$$

$$\text{subject\_slope}\ [\text{subject index}] \sim Normal(\text{subject\_mean}, \text{subject\_sd})$$

$$\text{region\_slope}\ [\text{region index}] \sim Normal(\text{region\_mean}, \text{region\_sd})$$

Hierarchical model framework:

We model the function (APD) -> (time) as a linear function.

We model the total slope of this function to be average slope + individual subject variable slope + sequence region variable slope

Priors:

$$\text{time} \sim Normal(\text{mean}, \text{sd})$$

$$\text{mean} \leftarrow \text{total\_intercept} + \text{total\_slope} * APD$$

$$\text{total\_slope} \leftarrow \text{avg\_slope} + \text{subject\_slope}\ [\text{subject index}] + \text{region\_slope}\ [\text{region index}]$$

$$\text{subject\_slope}\ [\text{subject index}] \sim Normal(\text{subject\_mean}, \text{subject\_sd})$$

$$\text{region\_slope}\ [\text{region index}] \sim Normal(\text{region\_mean}, \text{region\_sd})$$

Hierarchical model framework:

We model the function (APD) -> (time) as a linear function.

We model the total slope of this function to be average slope + individual subject variable slope + sequence region variable slope

Priors:

$$\text{time} \sim Normal(\text{mean}, \text{sd})$$

$$\boxed{\text{sd} \sim Cauchy(0, 0.5)}$$

$$\text{mean} \leftarrow \text{total\_intercept} + \text{total\_slope} * APD$$

$$\text{total\_slope} \leftarrow \text{avg\_slope} + \text{subject\_slope}\ [\text{subject index}] + \text{region\_slope}\ [\text{region index}]$$

$$\text{subject\_slope}\ [\text{subject index}] \sim Normal(\text{subject\_mean}, \text{subject\_sd})$$

$$\text{region\_slope}\ [\text{region index}] \sim Normal(\text{region\_mean}, \text{region\_sd})$$

Hierarchical model framework:

We model the function (APD) -> (time) as a linear function.

We model the total slope of this function to be average slope + individual subject variable slope + sequence region variable slope

Priors:

$$\text{time} \sim Normal(\text{mean}, \text{sd})$$

$$\text{sd} \sim Cauchy(0, 0.5)$$

$$\text{mean} \leftarrow \text{total\_intercept} + \text{total\_slope} * APD$$

$$\text{total\_intercept}$$

For now, restrict to be 0

$$\text{total\_slope} \leftarrow \text{avg\_slope} + \text{subject\_slope}\ [\text{subject index}] + \text{region\_slope}\ [\text{region index}]$$

$$\text{subject\_slope}\ [\text{subject index}] \sim Normal(\text{subject\_mean}, \text{subject\_sd})$$

$$\text{region\_slope}\ [\text{region index}] \sim Normal(\text{region\_mean}, \text{region\_sd})$$

Multilevel model framework:

We model the function (APD) -> (time) as a linear function.

We model the total slope of this function to be average slope + individual subject variable slope + sequence region variable slope

Priors:

$$\text{time} \sim Normal(\text{mean}, \text{sd})$$

$$\boxed{\text{sd} \sim Cauchy(0, 0.5)}$$

$$\text{mean} \leftarrow \text{total\_intercept} + \text{total\_slope} * APD$$

$$\boxed{\text{total\_intercept}}$$ For now, restrict to be 0

$$\text{total\_slope} \leftarrow \text{avg\_slope} + \text{subject\_slope}\ [\text{subject index}] + \text{region\_slope}\ [\text{region index}]$$

$$\boxed{\text{avg\_slope} \sim Uniform(0, 150)}$$

$$\text{subject\_slope}\ [\text{subject index}] \sim Normal(\text{subject\_mean}, \text{subject\_sd})$$

$$\text{region\_slope}\ [\text{region index}] \sim Normal(\text{region\_mean}, \text{region\_sd})$$

Hierarchical model framework:

We model the function (APD) -> (time) as a linear function.

We model the total slope of this function to be average slope + individual subject variable slope + sequence region variable slope

Priors:

$$\text{time} \sim Normal(\text{mean}, \text{sd})$$

$$\text{sd} \sim Cauchy(0, 0.5)$$

$$\text{mean} \leftarrow \text{total\_intercept} + \text{total\_slope} * APD$$

$$\text{total\_intercept}$$

For now, restrict to be 0

$$\text{total\_slope} \leftarrow \text{avg\_slope} + \text{subject\_slope}\ [\text{subject index}] + \text{region\_slope}\ [\text{region index}]$$

$$\text{avg\_slope} \sim Uniform(0, 150)$$

$$\text{subject\_slope}\ [\text{subject index}] \sim Normal(\text{subject\_mean}, \text{subject\_sd})$$

$$\text{subject\_mean} \sim Normal(0, 1)$$

$$\text{subject\_sd} \sim Cauchy(0, 20)$$

$$\text{region\_slope}\ [\text{region index}] \sim Normal(\text{region\_mean}, \text{region\_sd})$$

Hierarchical model framework:

We model the function (APD) -> (time) as a linear function.

We model the total slope of this function to be average slope + individual subject variable slope + sequence region variable slope

Priors:

$$\text{time} \sim Normal(\text{mean}, \text{sd})$$

$$\text{sd} \sim Cauchy(0, 0.5)$$

$$\text{mean} \leftarrow \text{total\_intercept} + \text{total\_slope} * APD$$

$$\text{total\_intercept}$$

For now, restrict to be 0

$$\text{total\_slope} \leftarrow \text{avg\_slope} + \text{subject\_slope}\ [\text{subject index}] + \text{region\_slope}\ [\text{region index}]$$

$$\text{avg\_slope} \sim Uniform(0, 150)$$

$$\text{subject\_slope}\ [\text{subject index}] \sim Normal(\text{subject\_mean}, \text{subject\_sd})$$

$$\text{subject\_mean} \sim Normal(0, 1)$$
$$\text{subject\_sd} \sim Cauchy(0, 20)$$

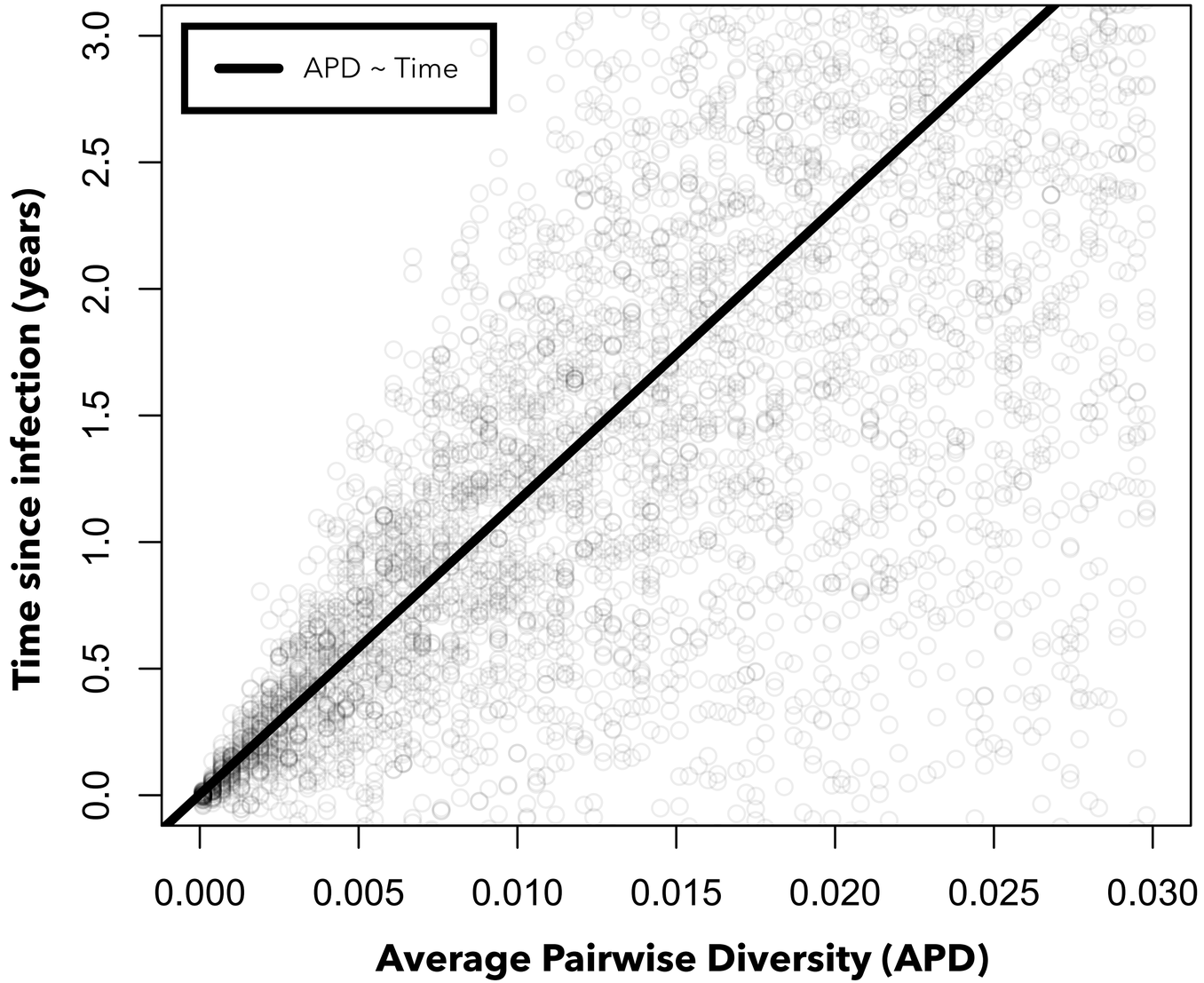
$$\text{region\_slope}\ [\text{region index}] \sim Normal(\text{region\_mean}, \text{region\_sd})$$

$$\text{region\_mean} \sim Normal(0, 1)$$
$$\text{region\_sd} \sim Cauchy(0, 5)$$

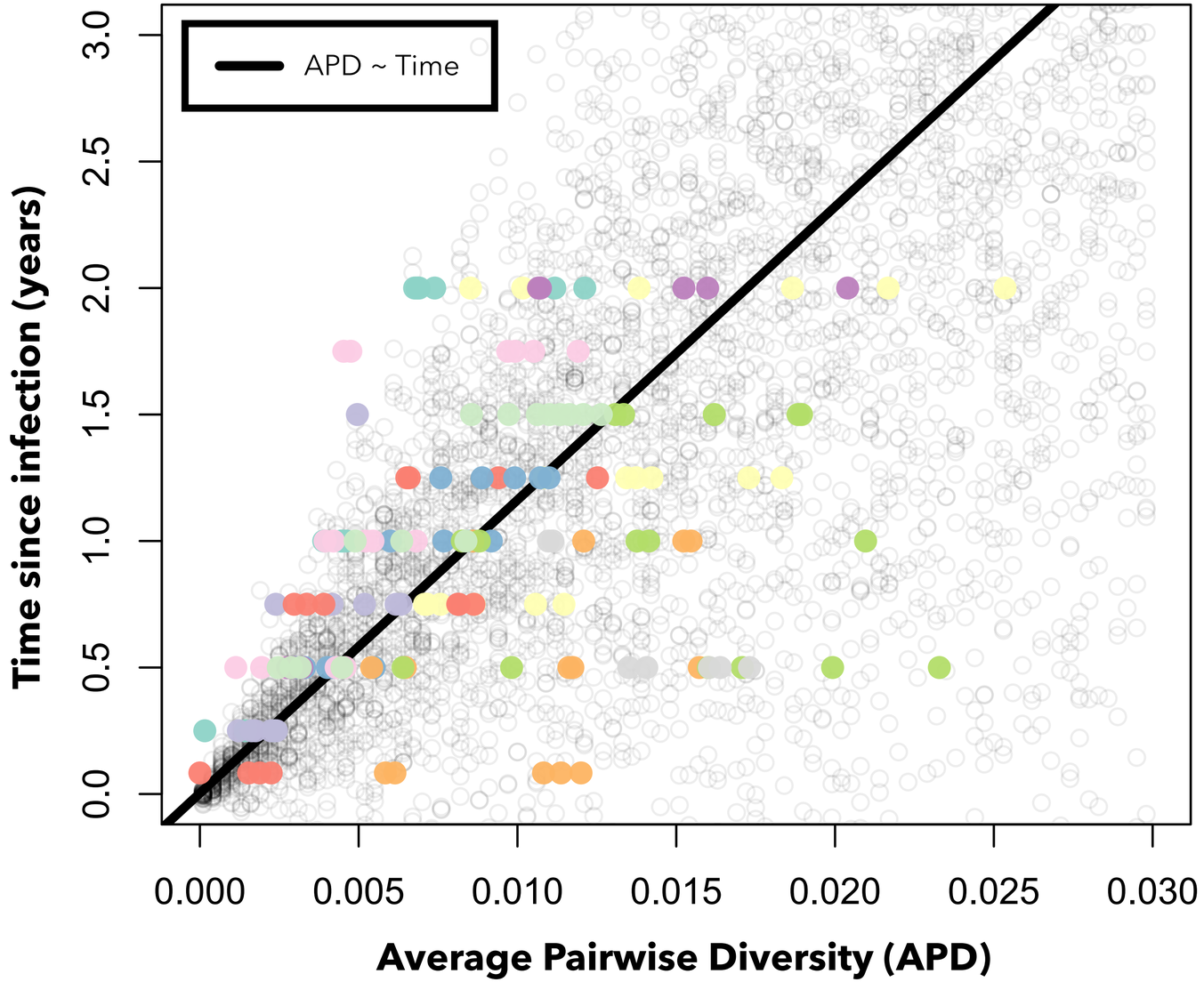
Hierarchical Model Performance

We can use this model to **simulate data** by sampling from the posterior distribution of time, given an APD value

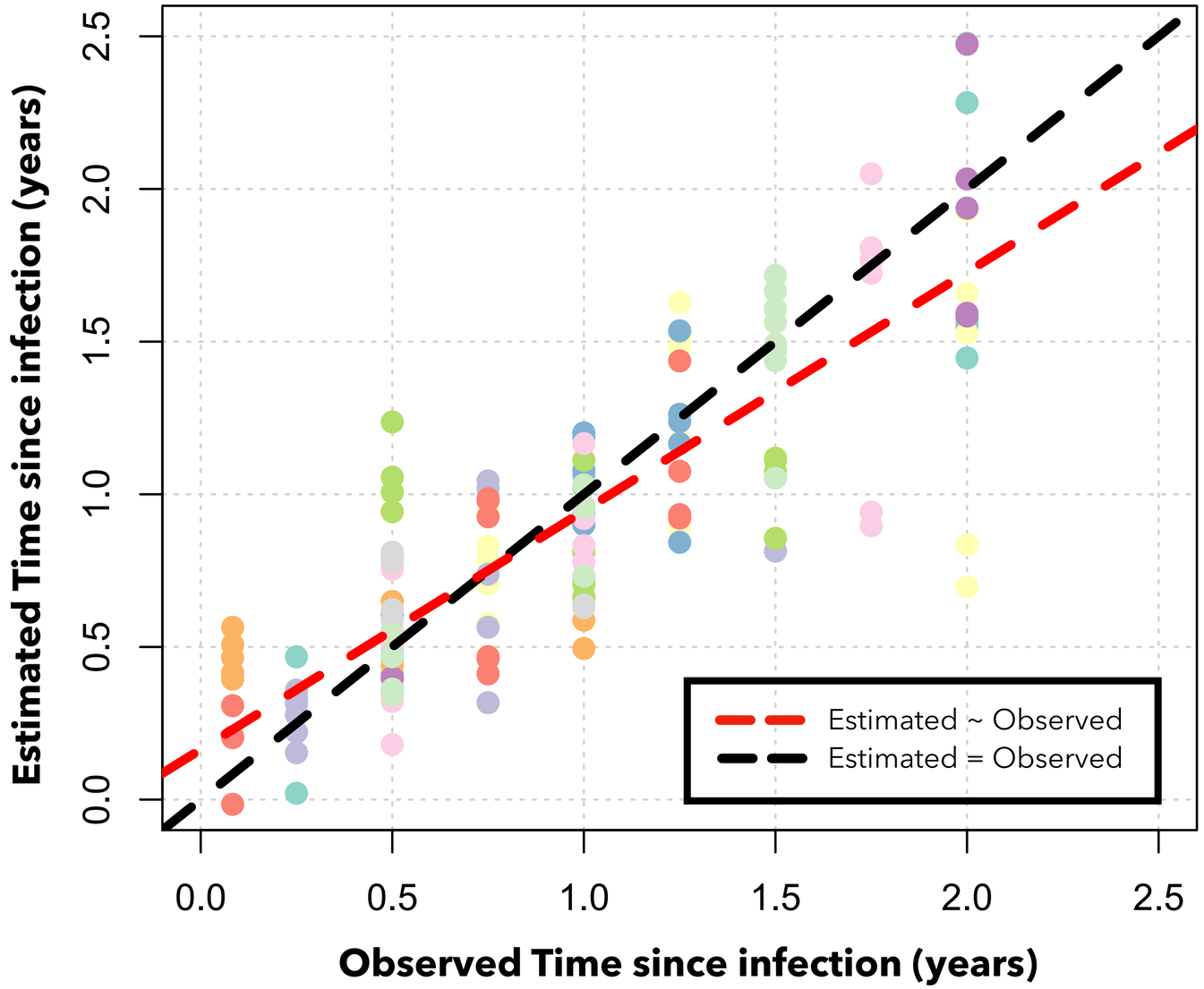
Hierarchical Model Performance



Hierarchical Model Performance



Hierarchical Model Performance



Approach:

1. Does the Puller et. al. adult model effectively predict time since infection for our **infant cohort**? **Not really**

2. Does an infant specific linear regression model effectively predict time since infection for our **infant cohort**? **Also, not really**

3. **What about a hierarchical model?** **Maybe!**

Next Steps/Questions

- Instead of defining time of infection to be birth, allow time of infection to be before birth
- Because each individual may have a different infection time, we need to incorporate a varying intercept level into the model
- Are we sampling all of the diversity of an individual?
- How can we explain individuals for which APD decreases with time?