Model Fitting: The Basic Concept
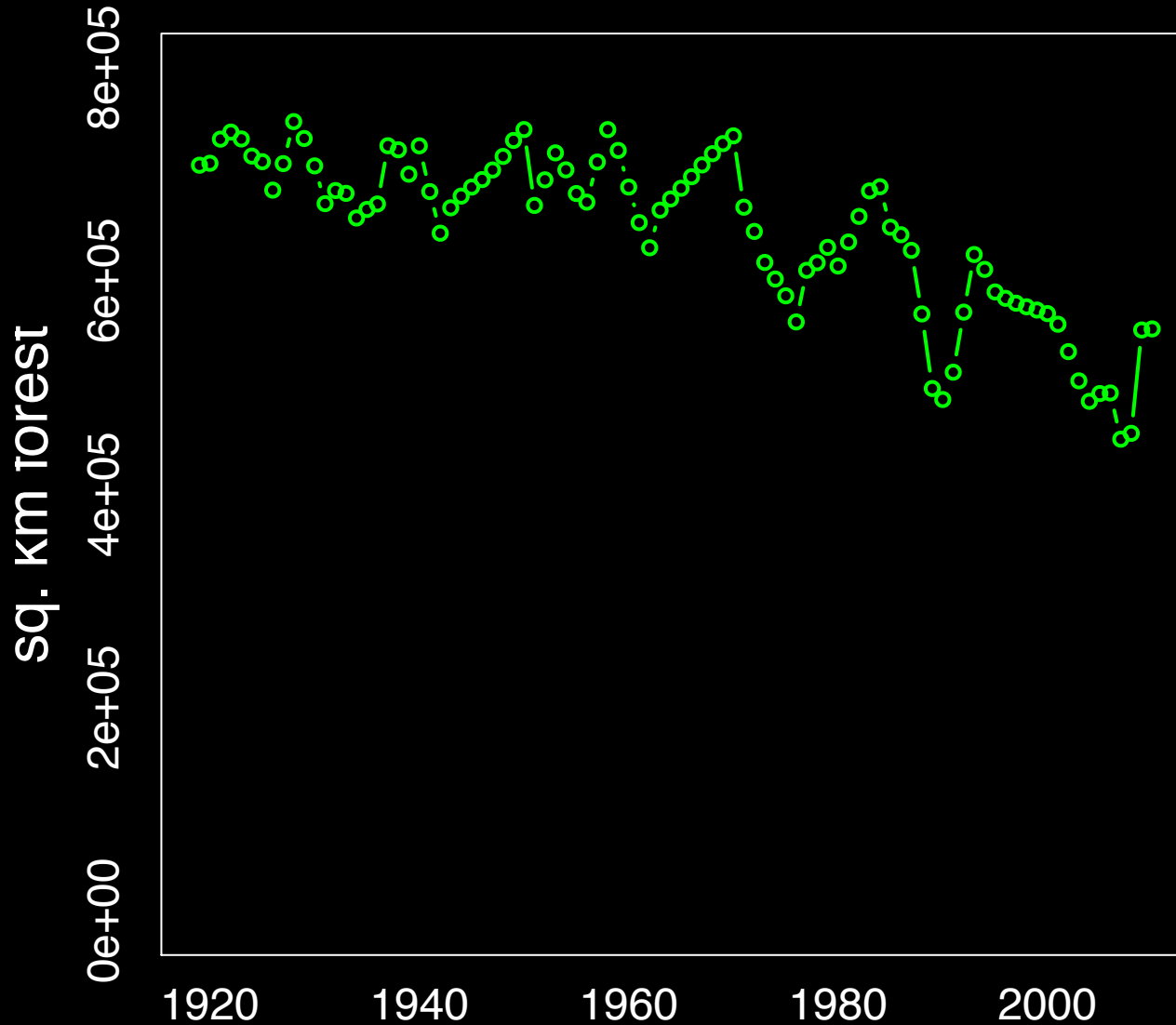

Cara Brook

E^2M^2, Centre ValBio

Ranomafana National Park, Madagascar

What happened to Madagascar's forest?

Model fitting in science

Model fitting in science

1. **Define your research question**

Model fitting in science

1. **Define your research question**
2. **Formulate a hypothesis**

Model fitting in science

1. **Define your research question**

2. **Formulate a hypothesis**

3. **Collect data**

Model fitting in science

1. **Define your research question**
2. **Formulate a hypothesis**
3. **Collect data**
4. **Construct a model** that demonstrates your hypothesis

Model fitting in science

1. **Define your research question**
2. **Formulate a hypothesis**
3. **Collect data**
4. **Construct a model** that demonstrates your hypothesis
5. **Assess model fit**: assuming our model is true, how likely are we to recover the observed data?

Model fitting in science

1. **Define your research question**
2. **Formulate a hypothesis**
3. **Collect data**
4. **Construct a model** that demonstrates your hypothesis
5. **Assess model fit**: assuming our model is true, how likely are we to recover the observed data?
6. **Optimize the parameters** behind the model to make it most likely to recover the data.

Model fitting in science

1. **Define your research question**
2. **Formulate a hypothesis**
3. **Collect data**
4. **Construct a model** that demonstrates your hypothesis
5. **Assess model fit**: assuming our model is true, how likely are we to recover the observed data?
6. **Optimize the parameters** behind the model to make it most likely to recover the data.
7. If need be, **restructure** your model to better match your data.

Statistical models are data-driven...

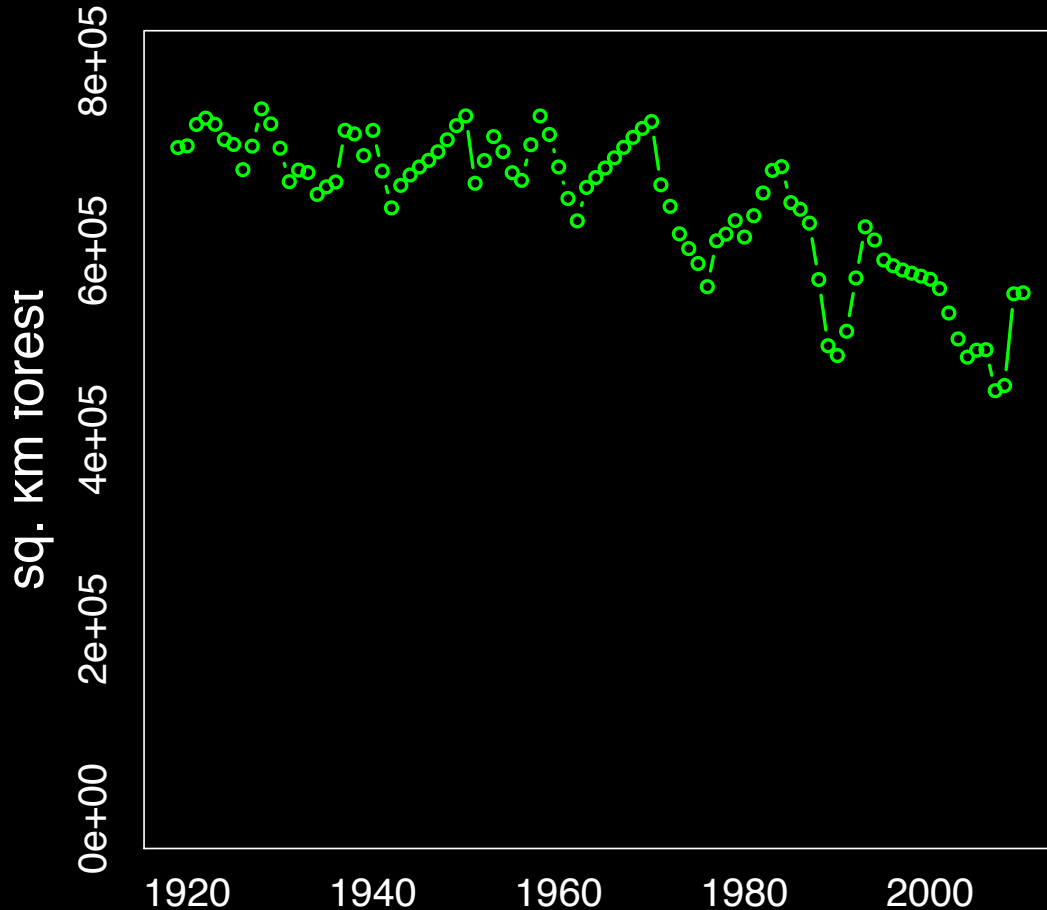- Goal: find patterns and correlations
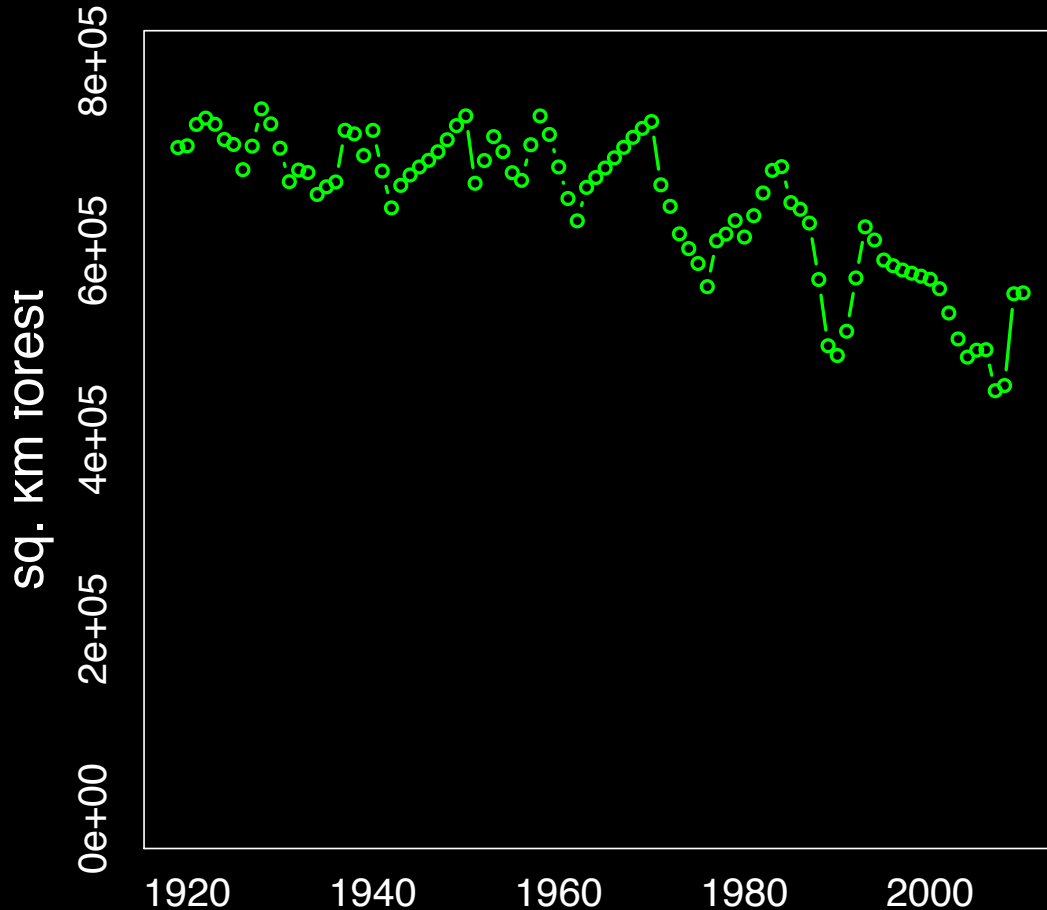
Statistical models
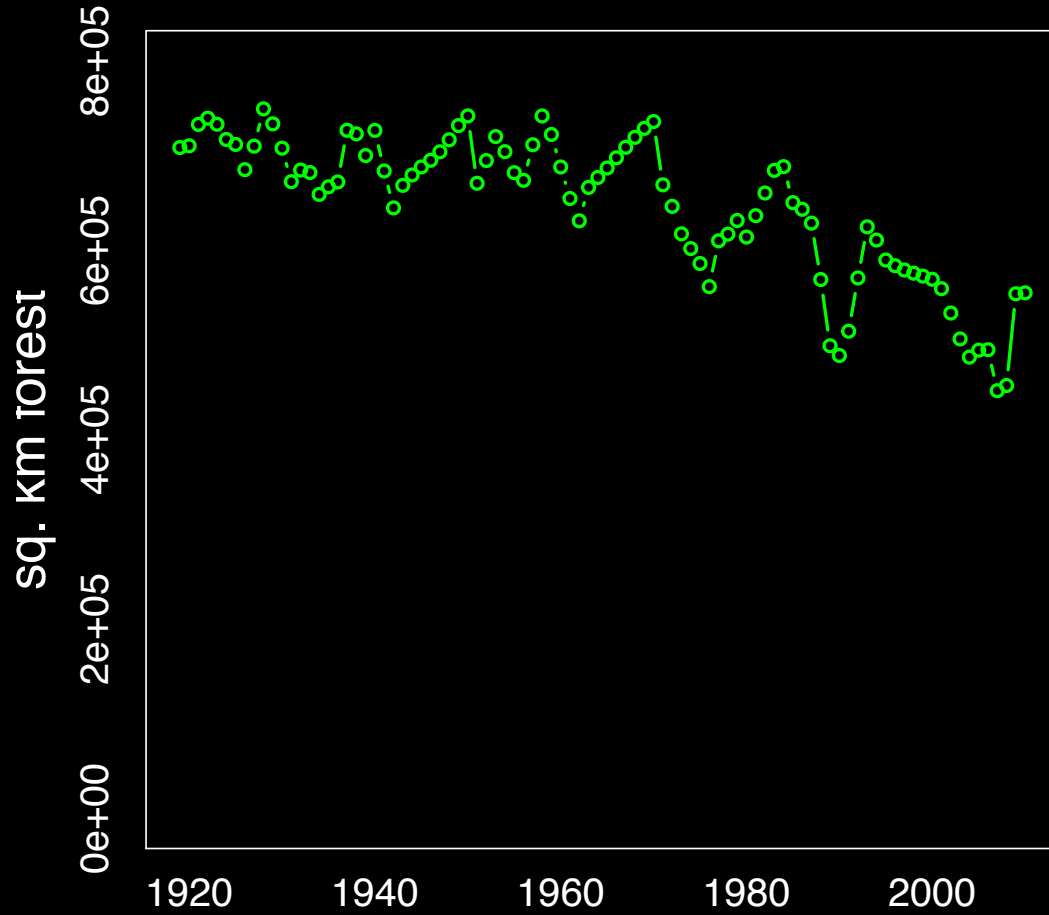
Statistical models

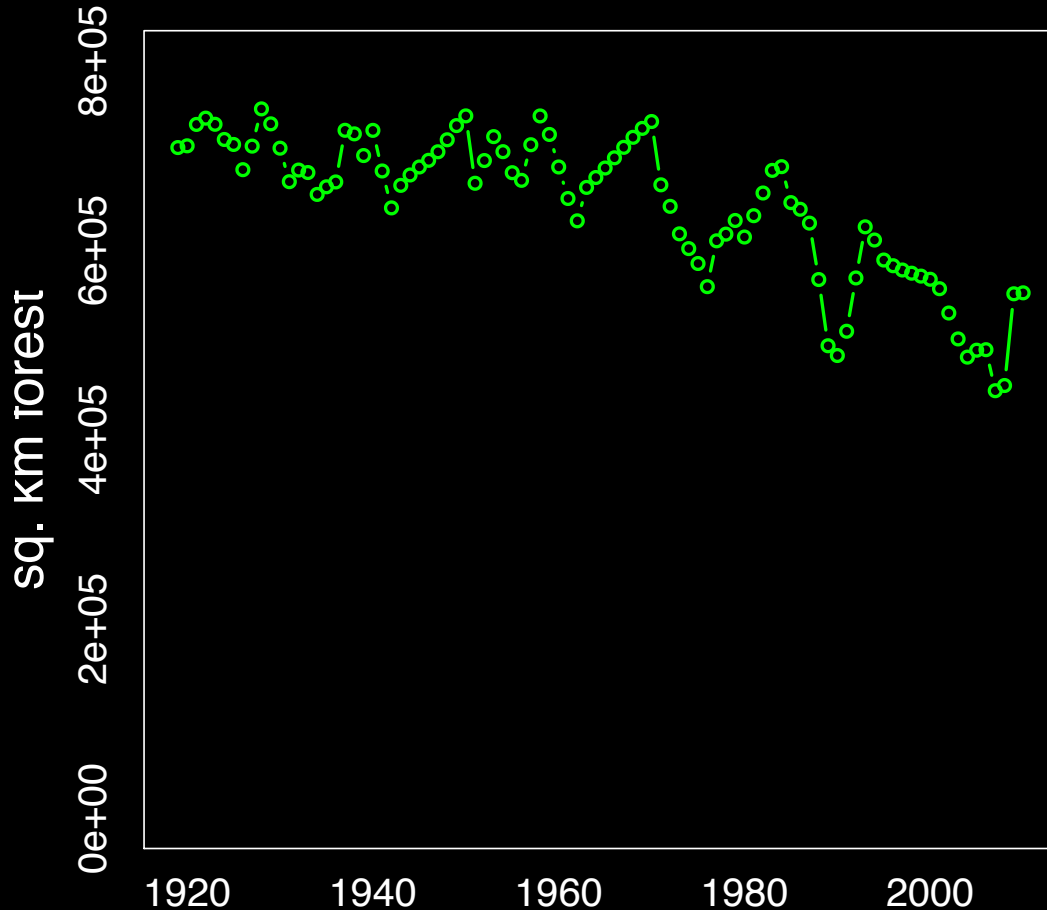- Goal: find patterns and correlations
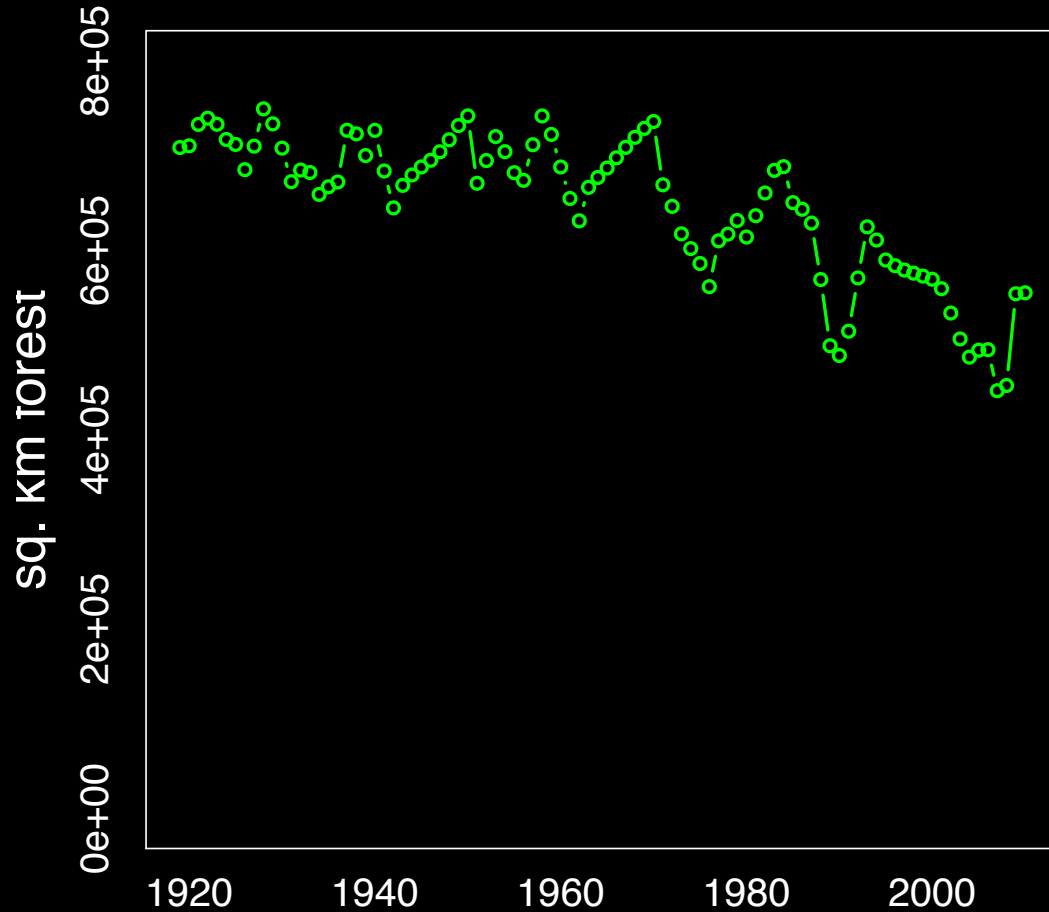
Statistical models are data-driven...

- Goal: find patterns and correlations

What is the trend in Madagascar's forest cover through time?

Statistical models are data-driven...
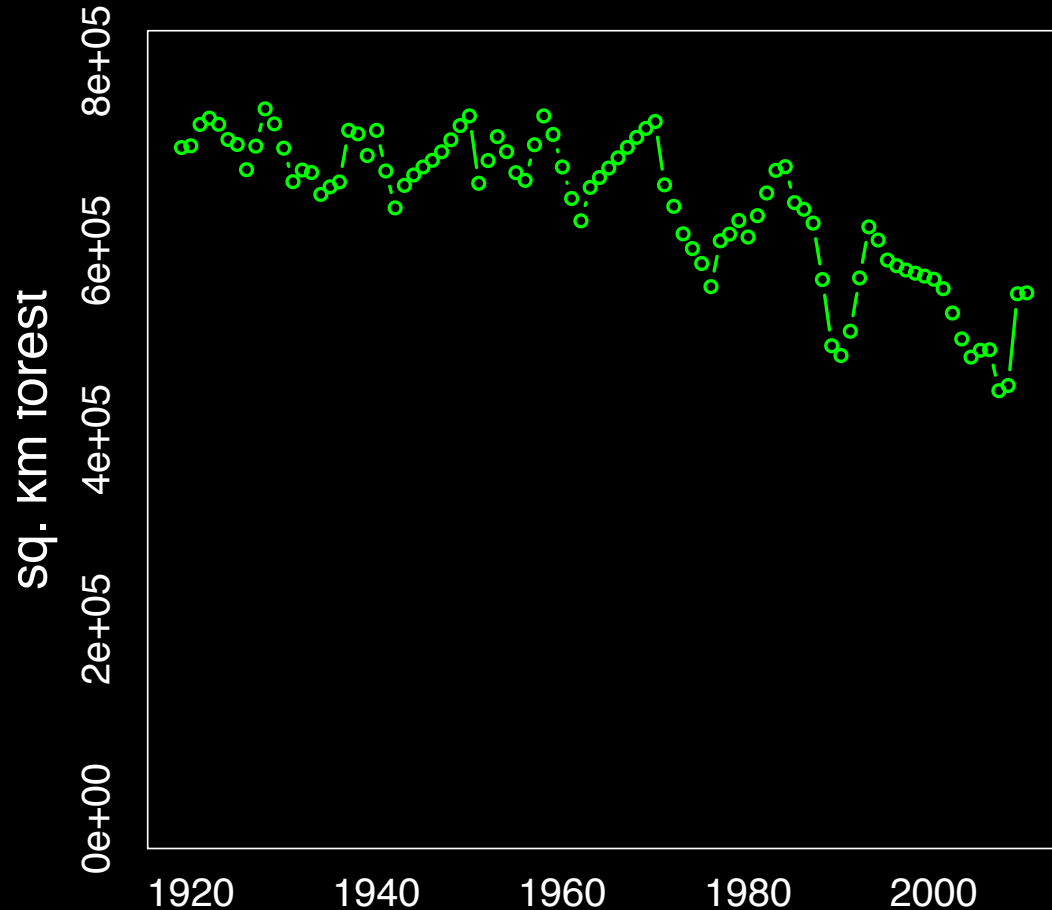
- Goal: find patterns and correlations

What is the trend in Madagascar's forest cover through time?

Can fit a linear regression.

Statistical models are data-driven...

- Goal: find patterns and correlations

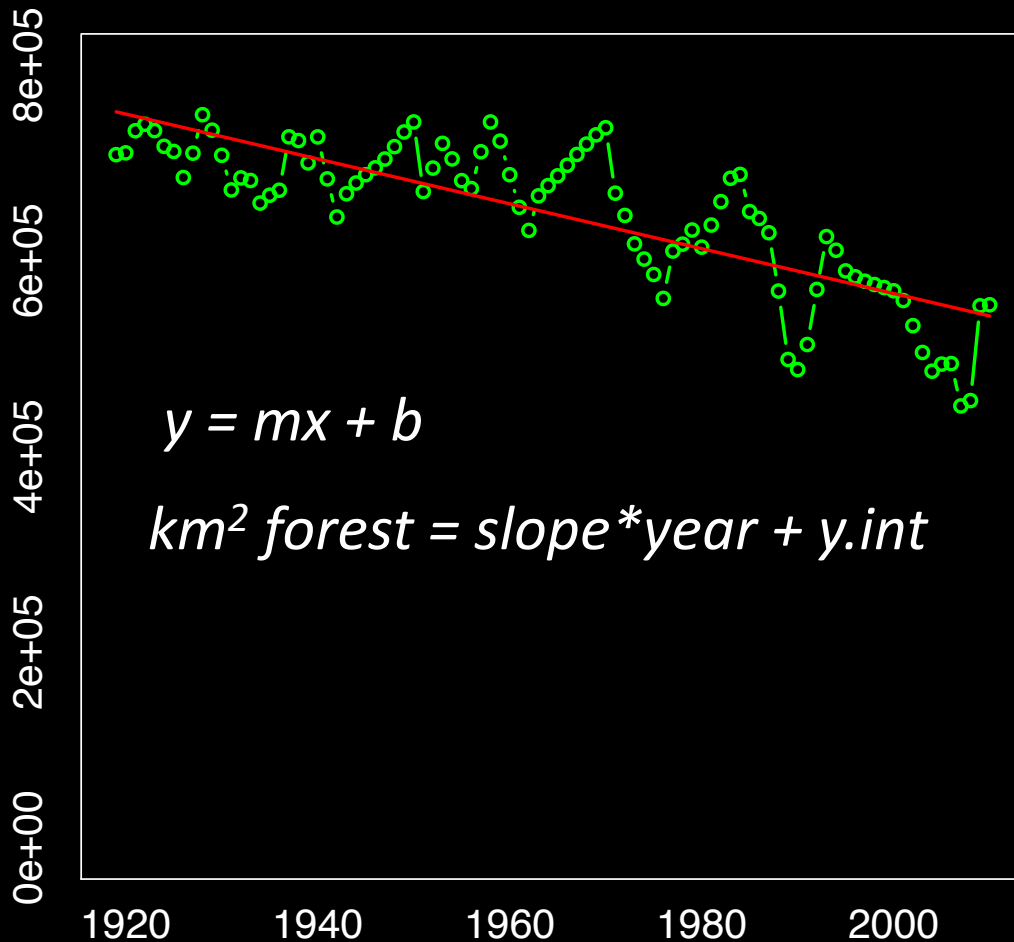1. **Construct a model** that demonstrates your hypothesis.

$y = mx + b$

$km^2 \text{ forest} = slope*year + y.int$

Statistical models are data-driven…

- Goal: find patterns and correlations

$y = mx + b$

$km^2 \ forest = slope*year + y.int$

1. **Construct a model** that demonstrates your hypothesis.

2. **Assess model fit**: use statistical tool (e.g., least squares) to ask, assuming our model is true, how likely are we to recover the observed data?
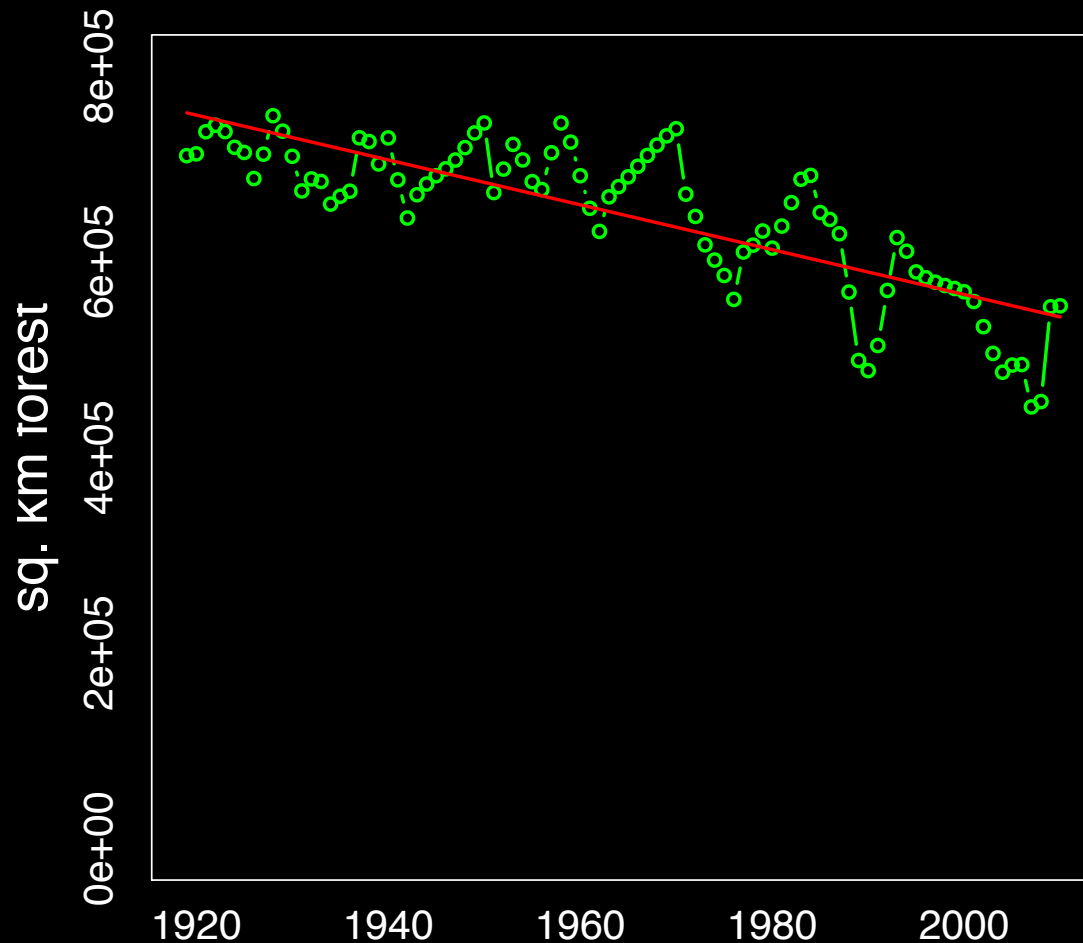
Statistical models are data-driven…

- Goal: find patterns and correlations

$y = mx + b$

m = -2293

b = 5152515.5

$km^2 \ forest = slope*year + y.int$
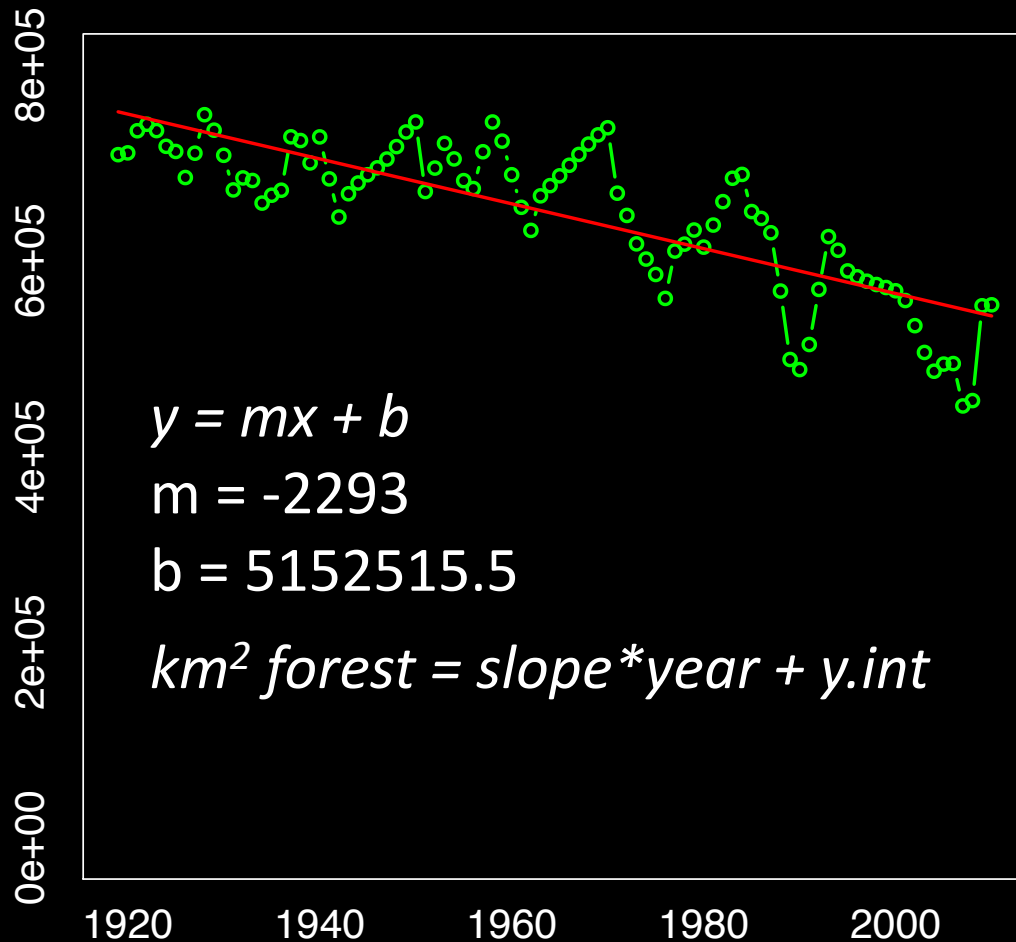
1. **Construct a model** that demonstrates your hypothesis.

2. **Assess model fit**: use least squares to ask, assuming our model is true, how likely are we to recover the observed data?

3. **Optimize the parameters** to make the model most likely to recover the data.

Statistical models are data-driven…

- Goal: find patterns and correlations

$y = mx + b$

m = -2,293
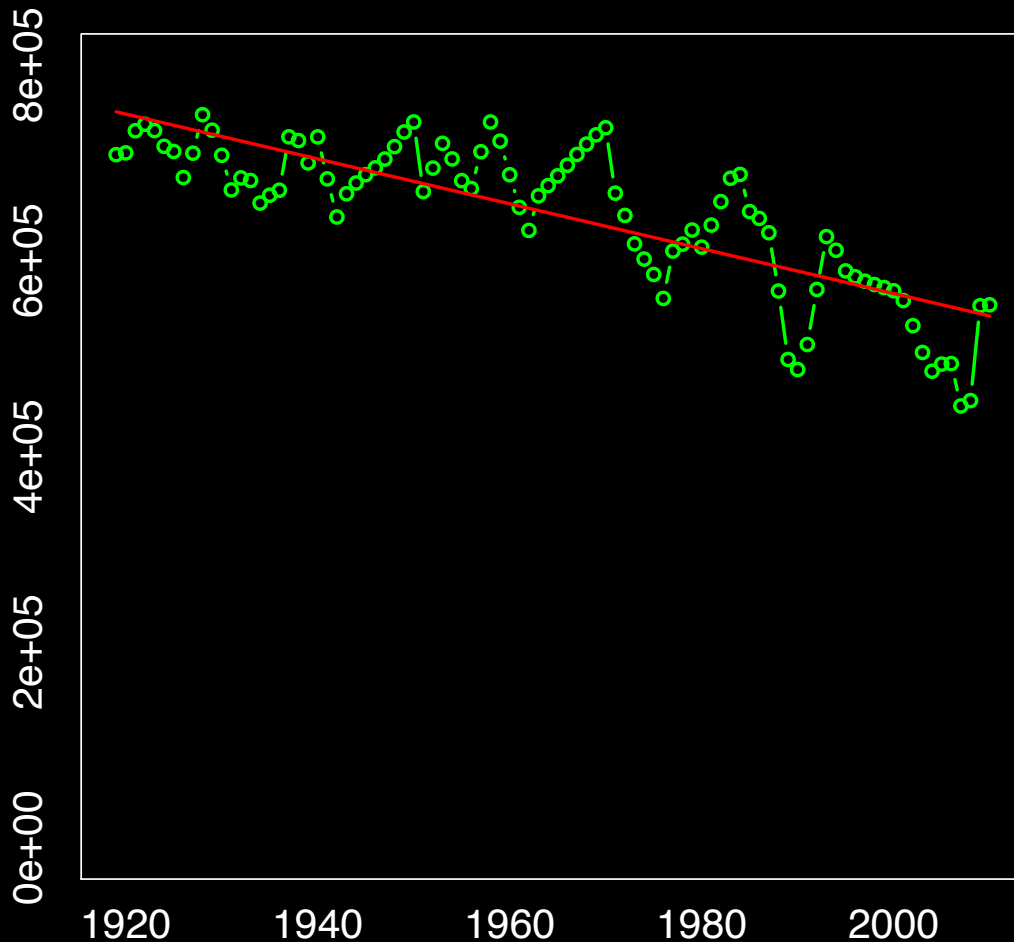
b = 5,152,515.5

- *This tells us that the time trend in forest cover is negative (declining) and that there was ~5*10^6 sq. km of forest in 1920.*

Statistical models are data-driven...

- Goal: find patterns and correlations

$y = mx + b$

m = -2293

b = 5152515.5

- *This tells us that the time trend in forest cover is negative (declining) and that there was ~5*10^6 sq. km of forest in 1920.*
- *But we know nothing about causation.*

Mechanistic modeling is process-driven…

- We want to understand *what* happened, *when* it happened, and *why* it happened
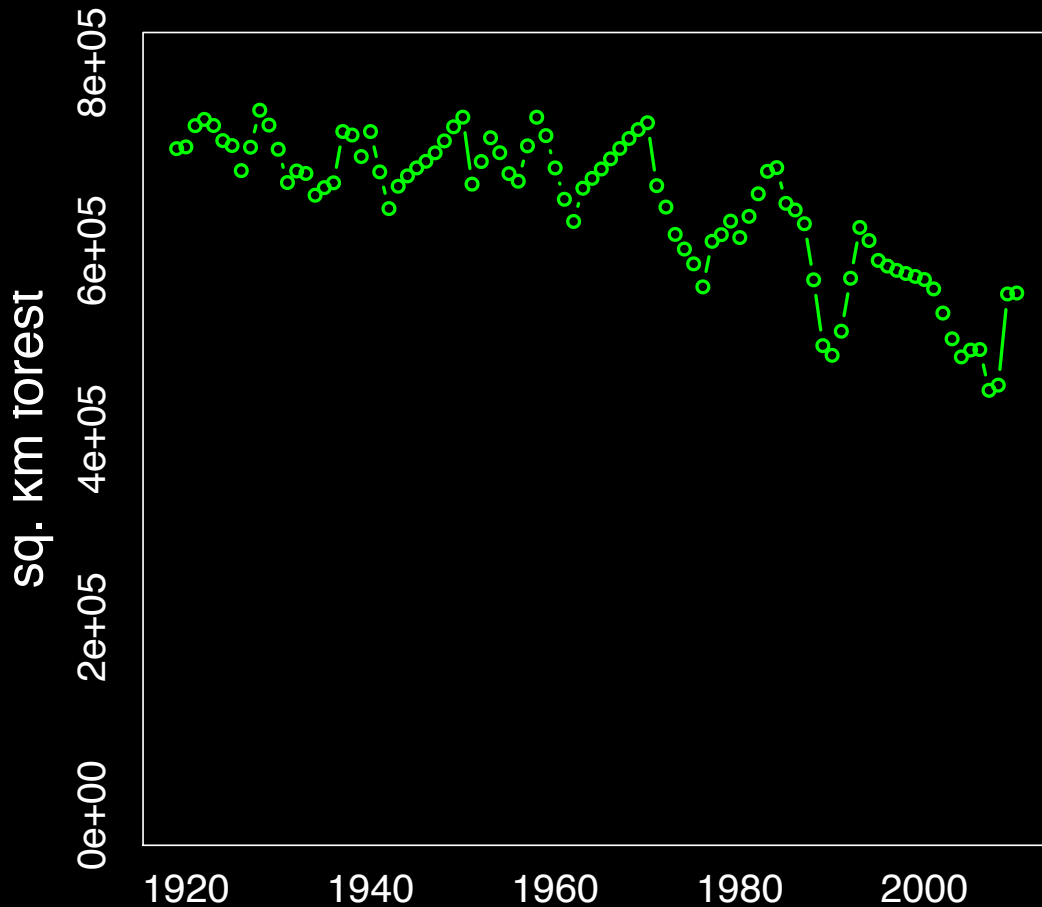
Mechanistic modeling is process-driven…

- We want to understand *what* happened, *when* it happened, and *why* it happened

- We start by building a model that uses explicit processes to recover the same outcomes ("states") as our data
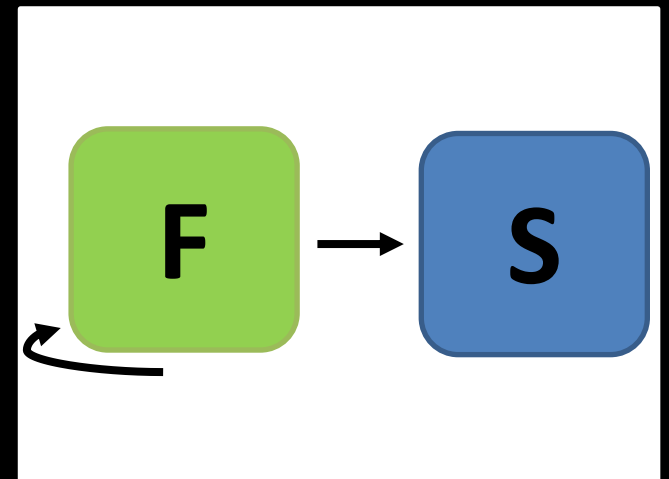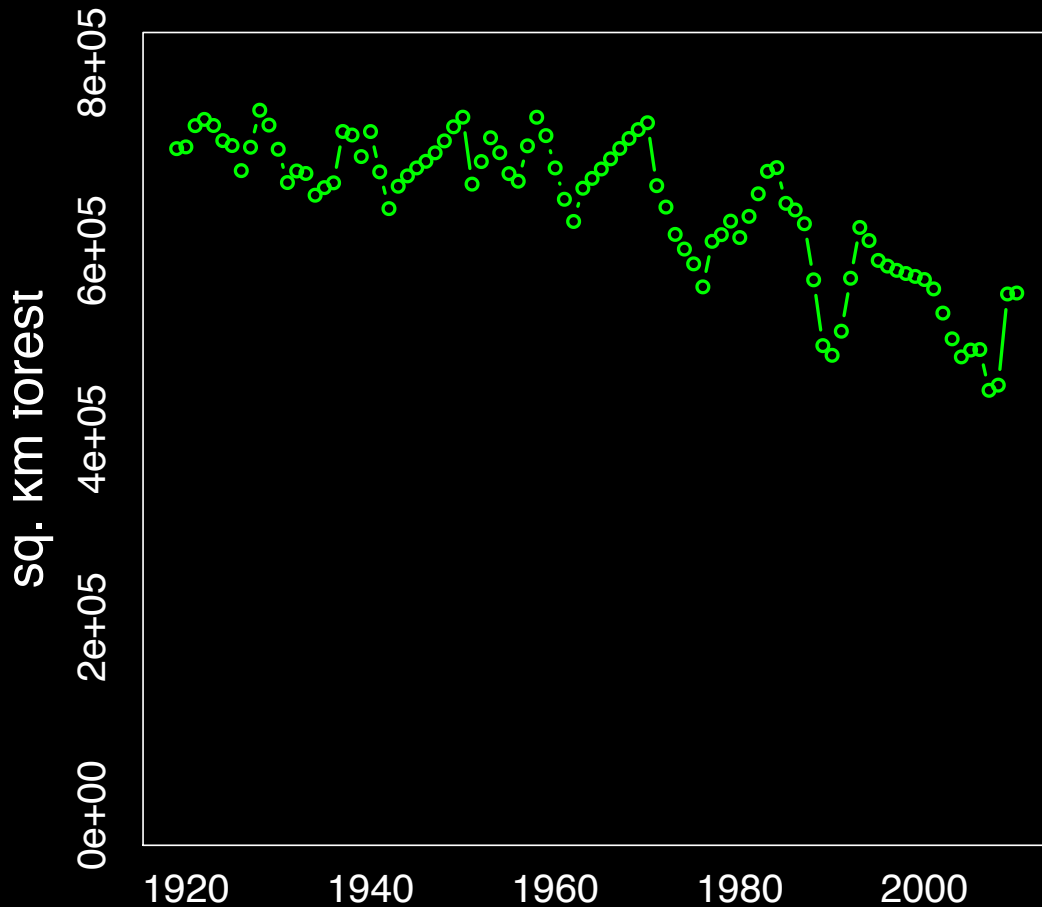
These data give us forest over time…
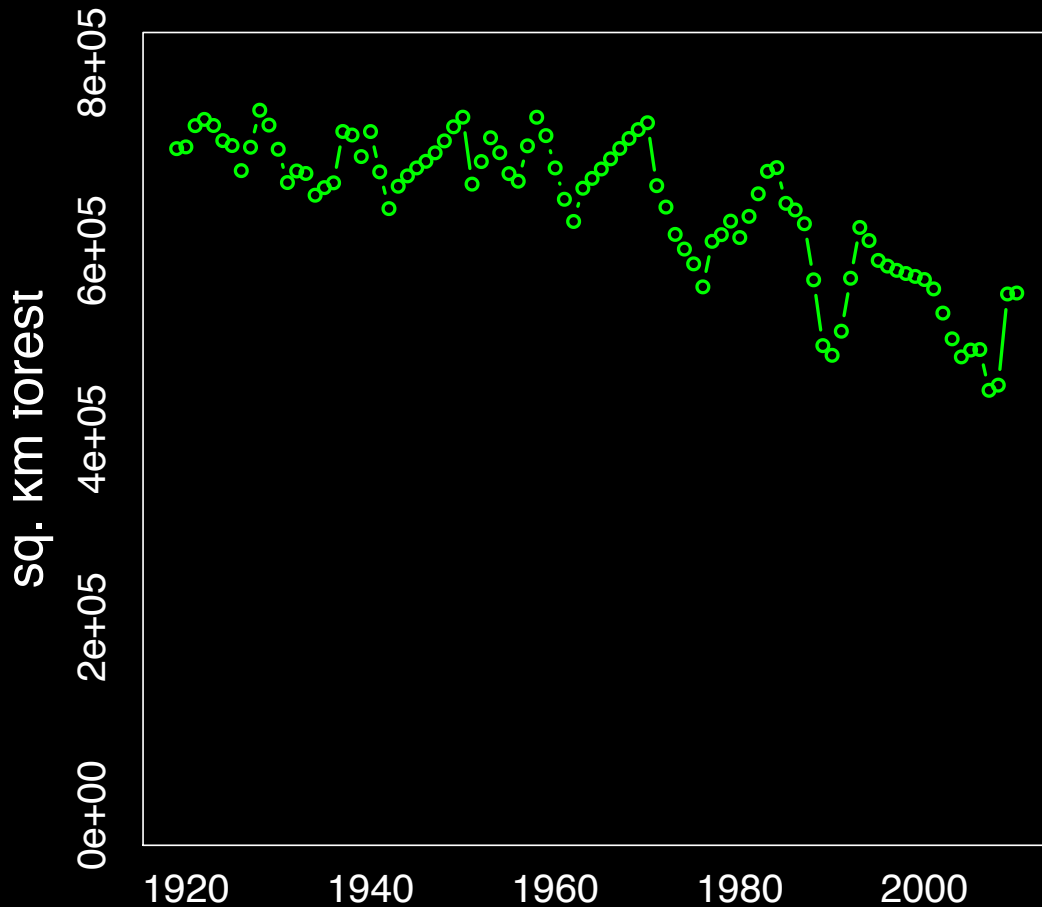
*What **states** and are captured in our data?*
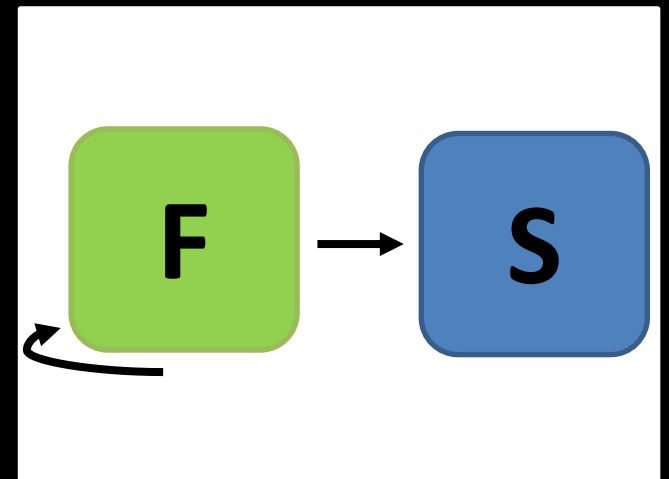
These data give us forest over time…
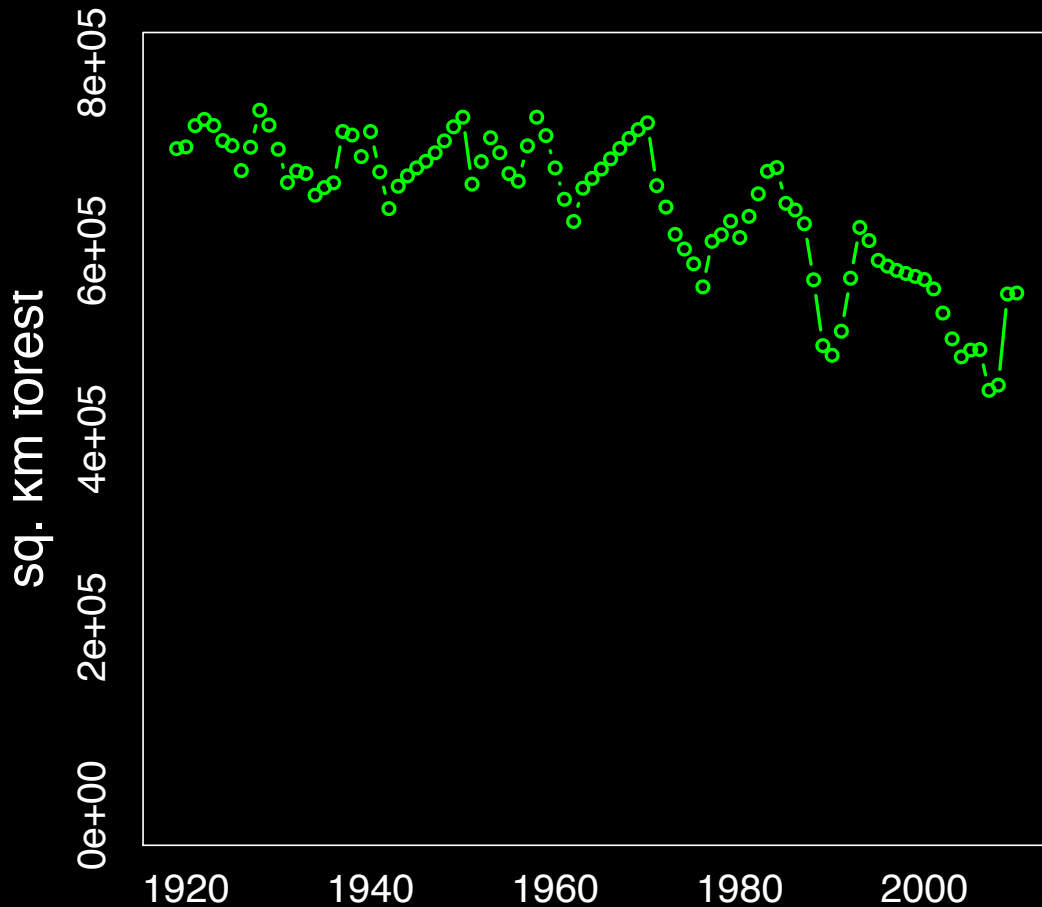

*What **states** and are captured in our data?*
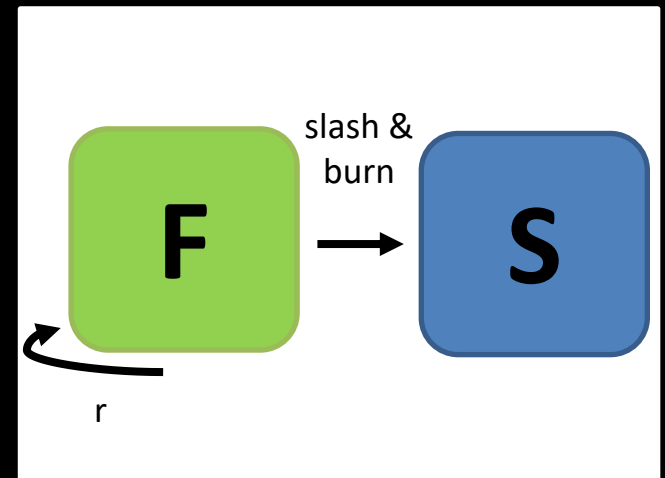
These data give us forest over time…

*What **processes** and are captured in our data?*

These data give us forest over time…

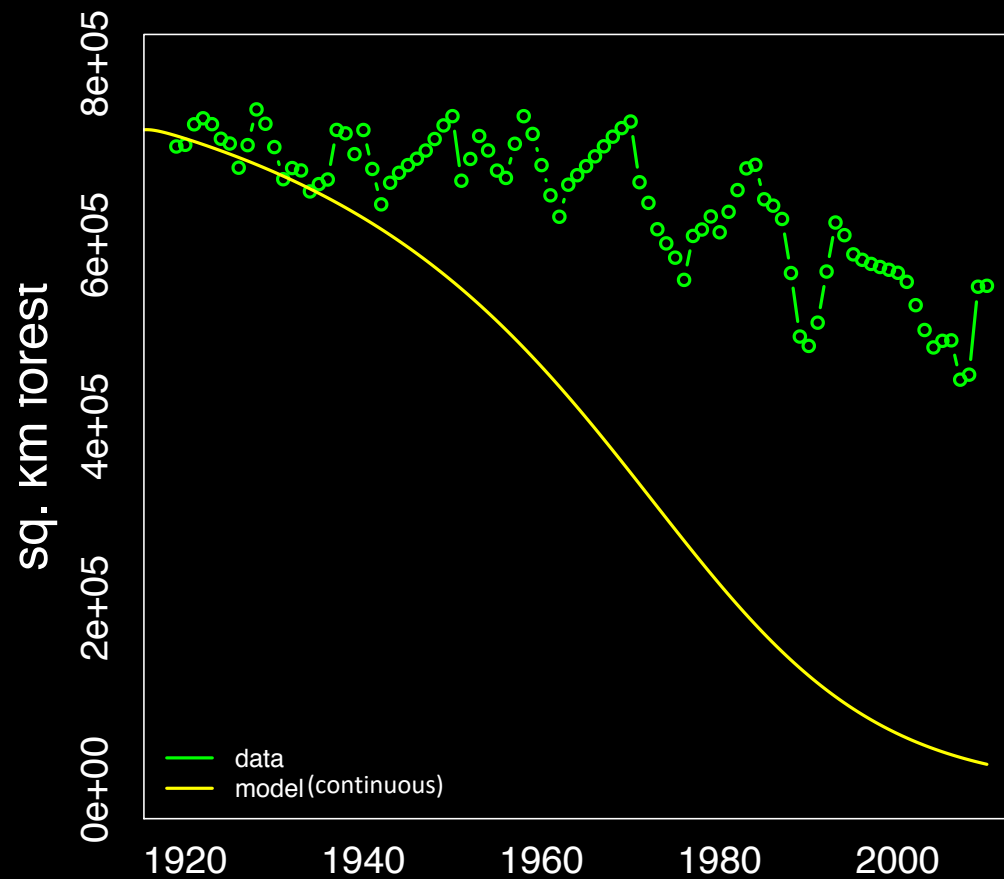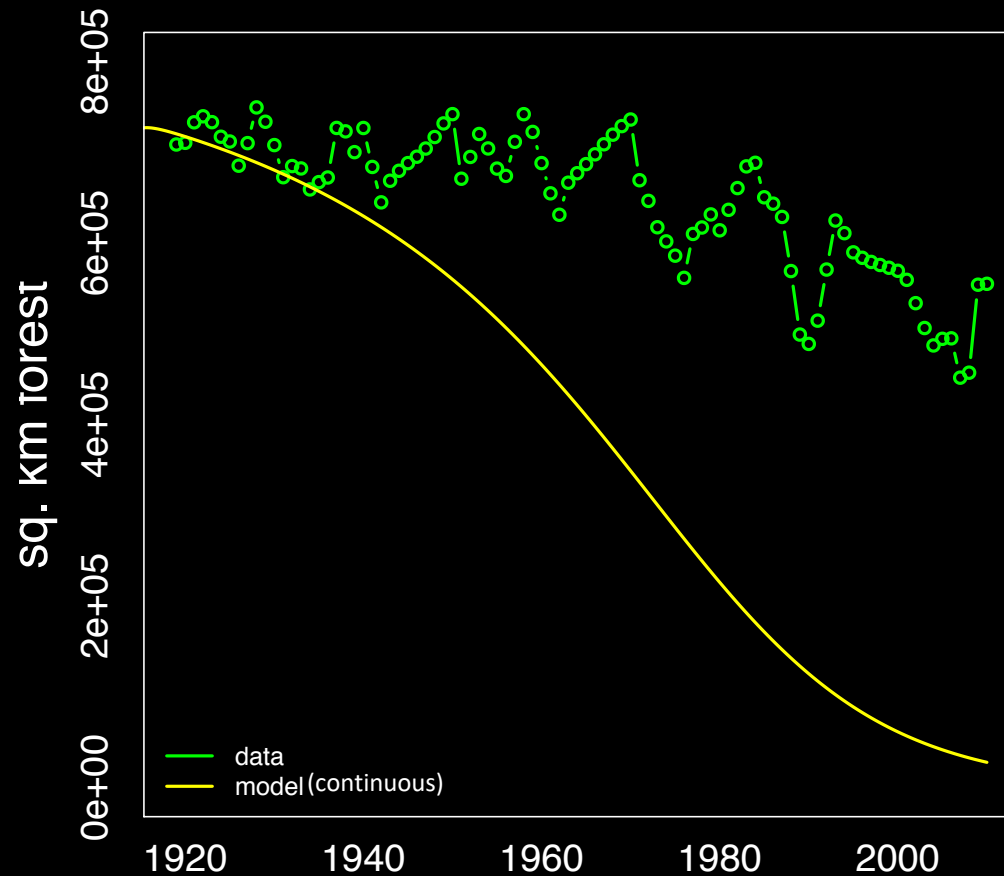*What **processes** and are captured in our data?*

How to fit a model to data

1. Build a mechanistic model that uses explicit processes to recover the same states as the data.

How to fit a model to data

2. Using least squares we ask, assuming our model is true, how likely are we to recover the observed data?
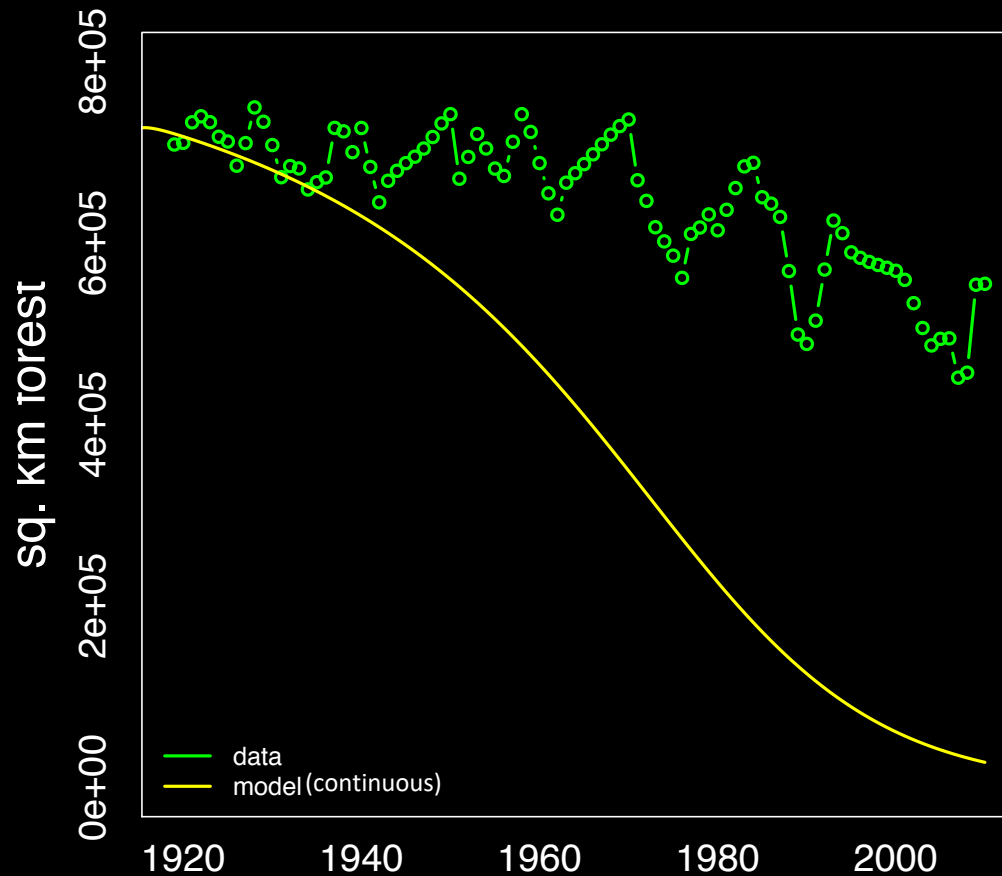

Model has the right trajectory but forest declines faster than in the data.

What does this suggest about our guess for the slash and burn rate?

How to fit a model to data

2. Using least squares we ask, assuming our model is true, how likely are we to recover the observed data?

Model has the right trajectory but forest declines faster than in the data.

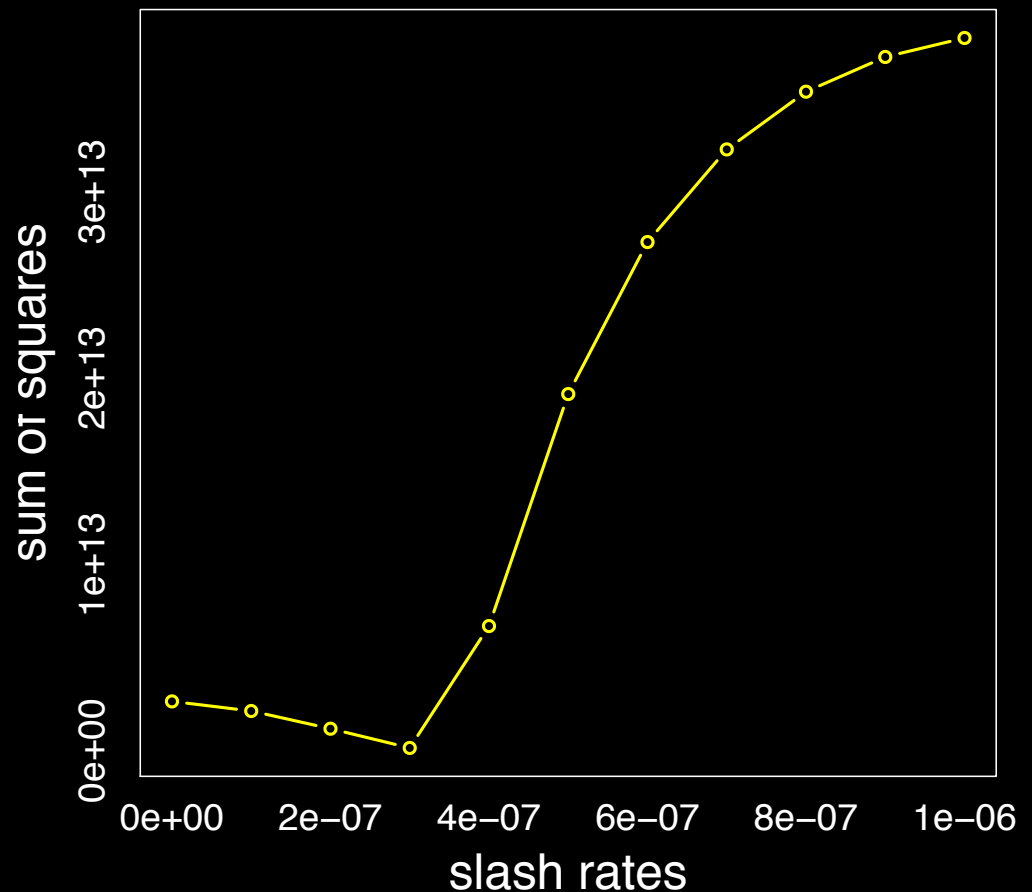What does this suggest about our guess for the slash and burn rate?

sum.sq = 3.914575e+13

Can we make that smaller?

How to fit a model to data

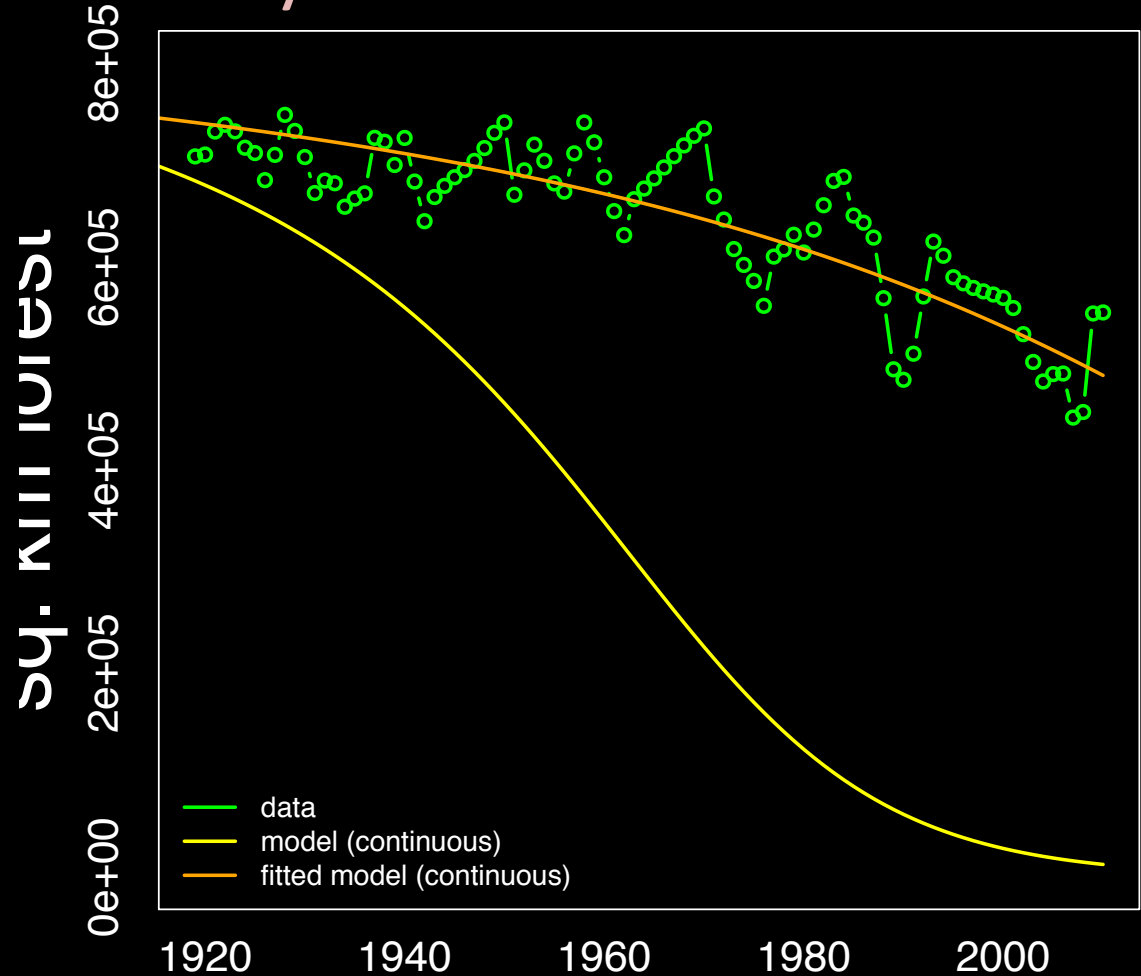3. Optimize the parameters behind the processes to make the model most likely to recover the data.

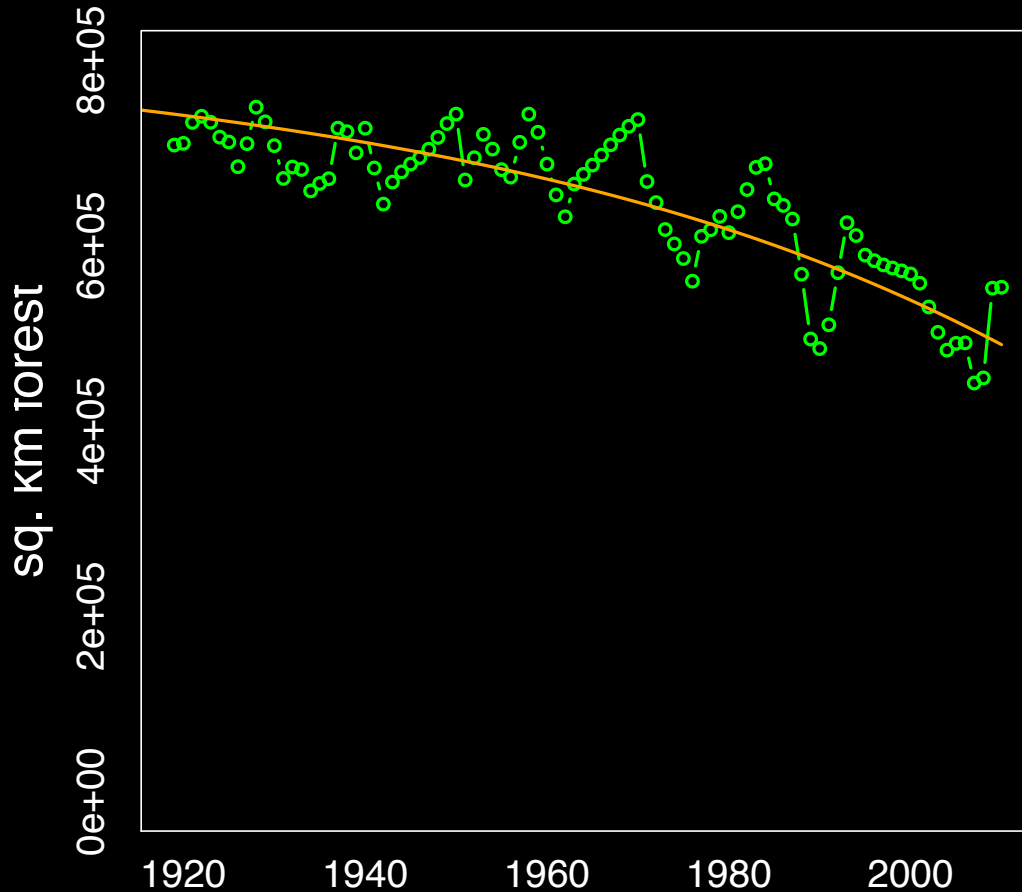slash.list[which.min(sm.sq)] = 3e-7

How to fit a model to data

3. Optimize the parameters behind the processes to make the model most likely to recover the data.

New slash fits better!

How to fit a model to data

4. If need be, restructure your model to better match your data.

We are good!