Predictive Analytics Lecture 1

Adam Kapelner

Stat 422/722
at The Wharton School of the University of Pennsylvania

January 17 & 18, 2017

Define: Prediction and Forecast

“statement about an uncertain event”,

Define: Prediction and Forecast

“statement about an uncertain event”, “informed guess or opinion”

predict (v.) 1620s (implied in predicted), "*foretell, prophesy,*" a back formation from prediction or else from Latin praedicatus, past participle of praedicere "foretell, advise, give notice,"

forecast (n.) early 15c., "*forethought, prudence,*" probably from forecast (v.). Meaning "conjectured estimate of a future course" is from 1670s.

I will be using predict and forecast interchangeably.

Examples

We make predictions all the time, saying things like:

- "Apple stock will go up tomorrow",
- "This condo will sell for $500K"

and sometimes unknowingly

- "Going skiing this weekend will make me happy",

How do we make predictions?

Examples

We make predictions all the time, saying things like:

- “Apple stock will go up tomorrow”,
- “This condo will sell for $500K”

and sometimes unknowingly

- “Going skiing this weekend will make me happy”,

How do we make predictions? We use a *model*.

Define: model

Model: a functional decription of a system

Define: model

Model: a functional decription of a system

An example model is:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

aphorism: (2) a concise statement of a scientific principle (and scientific principles are *models* of the observable universe)

All models have **input(s)** and **output(s)**. In the model above, what are the...

Inputs?

Define: model

Model: a functional decription of a system

An example model is:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

aphorism: (2) a concise statement of a scientific principle (and scientific principles are *models* of the observable universe)

All models have **input(s)** and **output(s)**. In the model above, what are the...

Inputs? bedtime schedule, waking schedule, ...
Outputs?

Define: model

Model: a functional decription of a system

An example model is:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

aphorism: (2) a concise statement of a scientific principle (and scientific principles are *models* of the observable universe)

All models have **input(s)** and **output(s)**. In the model above, what are the...

Inputs? bedtime schedule, waking schedule, ...
Outputs? health, wealth and wisdom

Synonyms for Inputs and Outputs

Here, the inputs and outputs are

- *features*
- *attributes*
- *characteristics*
- *variables* / *variates*

of a person. A person features health, a person has the characteristic of going to bed early.

What are “observations”?

Here, we have features of a person. Generally, inputs and outputs are features of the

- *observation* or
- *unit* or
- *record* or
- *subject*.

Thus the model relates some *feature(s) of the observation* to other *feature(s) of the observation*. Here, we are relating specific people’s bedtime schedule and waking schedule to their health, wealth and wisdom.

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

- What does "early to bed" mean?
- What does "early to rise" mean?
- What does "healthy" mean?
- What does "wealthy" mean?
- What does "wise" mean?

Without resolving these ambiguities,

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

- What does “early to bed” mean?
- What does “early to rise” mean?
- What does “healthy” mean?
- What does “wealthy” mean?
- What does “wise” mean?

Without resolving these ambiguities, the model is unusable and of course, untestable.

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

- What does "early to bed" mean?
- What does "early to rise" mean?
- What does "healthy" mean?
- What does "wealthy" mean?
- What does "wise" mean?

Without resolving these ambiguities, the model is unusable and of course, untestable.

In order to make this precise and defined, there is a necessity to use numbers.

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

- What does "early to bed" mean?
- What does "early to rise" mean?
- What does "healthy" mean?
- What does "wealthy" mean?
- What does "wise" mean?

Without resolving these ambiguities, the model is unusable and of course, untestable.

In order to make this precise and defined, there is a necessity to use numbers. Thus, features of the observation must be *measured* (to be defined later).

Ambiguity of Models Defined by Words

Models phrased in language such as:

> *Early to bed and early to rise makes a man healthy, wealthy and wise.*

are usually ambiguous, imprecise, vague and ill-defined. Why?

- What does "early to bed" mean?
- What does "early to rise" mean?
- What does "healthy" mean?
- What does "wealthy" mean?
- What does "wise" mean?

Without resolving these ambiguities, the model is unusable and of course, untestable.

In order to make this precise and defined, there is a necessity to use numbers. Thus, features of the observation must be *measured* (to be defined later).

The Model as a Functional Relationship

Thus the model relates some *measured feature(s) of the observation* to other *measured feature(s) of the observation*. The relationship is a function taking in inputs (within the parentheses) and "returning" the outputs (the equal sign). For any observation,

$$\begin{matrix}\text{the measured}\\ \text{outputs of an}\\ \text{observation}\end{matrix} = \text{model}\begin{pmatrix}\text{the measured}\\ \text{inputs of an}\\ \text{observation}\end{pmatrix}$$

It is traditional to put the outputs on the left hand side. This is assumed that the outputs were measured. This type of observation is called

- old or
- historical or
- known

and predictions here are not needed (obviously). In our aphorism model, for the observation being a known person named Joe:

$$\begin{bmatrix}\text{a measured quantity of Joe's health}\\ \text{a measured quantity of Joe's wealth}\\ \text{a measured quantity of Joe's wisdom}\end{bmatrix} = \text{model}\left(\begin{bmatrix}\text{a measured quantity of Joe's bedtime}\\ \text{a measured quantity of Joe's waketime}\\ \vdots\end{bmatrix}\right)$$

Updated Definition of Prediction

Now we can hone our definition of prediction. For a

- new or
- heretofore unseen or
- future

observation, where the inputs have been measured / assessed but the output has not been measured / assessed,

$$\underbrace{\begin{array}{c}\text{the guessed}\\ \text{output}\\ \text{measurements}\end{array}}_{\text{prediction}} = \text{model}\left(\begin{array}{c}\text{the measured}\\ \text{inputs of an}\\ \text{observation}\end{array}\right)$$

$$\begin{bmatrix} \text{a guessed quantity of Bob's health}\\ \text{a guessed quantity of Bob's wealth}\\ \text{a guessed quantity Bob's wisdom} \end{bmatrix} = \text{model}\left(\begin{bmatrix} \text{a measured quantity of Bob's bedtime}\\ \text{a measured quantity of Bob's waketime}\\ \vdots \end{bmatrix}\right)$$

Measurements as Variables

Instead of "a measured quantity ..." we can use algebraic *variables* to denote the numerical quantities. It is traditional to use x's to represent inputs and y's to represent outputs. Here would be the relationship for Joe:

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = \text{model}\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}\right)$$

and for Bob:

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = \text{model}\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \end{bmatrix}\right)$$

We will use the "hat" symbol (ˆ) to indicate a prediction of the output $\hat{y}$ to distinguish it from a known value of the output y.

More Vocabulary

Even though measured inputs and outputs are features of an observation, they each go by special names that emphasize their roles.

Each output y is called a

- *response* (the model "responds" to inputs)
- *outcome* / *outcome metric* (the result of inputs)
- ~~*endpoint*~~ (only used in clinical trial context)

and they are the target of prediction — what we want to ultimately predict.

Inputs x's then can go by the following terms of art:

- *covariates* (because the vary with the response, co-vary)
- *predictors* (since they will be the inputs used to make predictions)

and they are what we make use of to predict. I will try to use "response" and "predictors" in this course.

Mathematical Model

Now that we have predictors and responses measured an numeric and an equal sign relating them. We have officially created a *mathematical model*. The word "model" now will be represented as a function, f. So for an old observation,

$$\begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} = f\left(x_1, x_2, \ldots\right)$$

and for a new observation,

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ \hat{y}_3 \end{bmatrix} = f\left(x_1, x_2, \ldots\right)$$

It is said that the "model explains the response". What does this mean?

Science is based on Mathematical Models

We have become quite successful at shrink-wrapping interesting variables in the world around us

$$F$$

Focus: models with univariate responses.

Although general models have any number of outputs, this semester we will only consider models with one output. Thus, we will be looking at models such as

Early to bed and early to rise makes a man healthy.

We picked the most interesting output. So, for an old observation,

$$y = f\left(x_1, x_2, \ldots\right)$$

and a new observation,

$$\hat{y} = f\left(x_1, x_2, \ldots\right)$$

Many possible models

Early to bed and early to rise makes a man healthy.

What is the response metric?

Many possible models

Early to bed and early to rise makes a man healthy.

What is the response metric? What does "healthy" mean?

- Healthy for his whole life? Unlikely the model means this...
- Healthy for ages 25-65? Since we can expect health in infanthood and adolescence but not in elderly years

One also gets a feeling from the wording, there is either "healthy" or "not healthy". Thus the response metric will be the *categorical* data type and the model would be called a *classification* model.

Categorical measurements consist of discrete, mutually exclusive *levels*. Here, {healthy, not healthy}. Generally, {a, b, c, . . .}. Metrics with a large number of levels are difficult to model — keep it low.

I there are two levels, it is called *binary* or *dichotomous* and the model would be called a *binary response model* (or a classification) with elements 0 and 1.

Define the response clearly

Response: Healthy for ages 25–65

We still need a clear definition. Ideas? How about: healthy means no incidence of a "major" disease between the ages of 25–65? This can be assessed with medical records.

Define the predictors clearly

x_1: bedtime schedule

Definition?

Define the predictors clearly

x_1: bedtime schedule

Definition? Average bedtime.

Define the predictors clearly

x_1: bedtime schedule

Definition? Average bedtime. How to measure / assess? Survey?

x_2: waketime schedule

Definition?

Define the predictors clearly

x_1: bedtime schedule

Definition? Average bedtime. How to measure / assess? Survey?

x_2: waketime schedule

Definition? Average time to rise in the morning assessed via survey

Thus, x_1 and x_2 are a variant of a *timestamp* data type.

Dataframes

The historical *data frame* or *dataset* or even more colloquially, the "data" looks like:

Healthy? 1 = yes (y)	Average Bedtime (x_1)	Average Waketime (x_2)
1	9:32PM	6:42AM
0	11:55PM	7:53AM
0	10:33PM	7:02AM
$\vdots$		

Dataframes have n observations and p predictors. Here $n =$

Dataframes

The historical *data frame* or *dataset* or even more colloquially, the "data" looks like:

Healthy? 1 = yes (y)	Average Bedtime (x_1)	Average Waketime (x_2)
1	9:32PM	6:42AM
0	11:55PM	7:53AM
0	10:33PM	7:02AM
$\vdots$		

Dataframes have n observations and p predictors. Here $n = 3$ (only those viewable above) and $p = 2$. Thus, it is a matrix with n rows and $p + 1$ columns. The "+1" is for the response which is not a predictor.

What would a new observation look like? Tony went to bed on average 9:53PM and awoke on average at 6:13AM. Did he have a healthy life or not?

We don't know the model yet...

Mathematical Models are Deterministic

Early to bed and early to rise makes a man healthy.

which, after measuring inputs and outputs and mathematizing, becomes $y = f(x_1, x_2)$.

From the wording, it seems the model is unequivocal and deterministic. This means

Mathematical Models are Deterministic

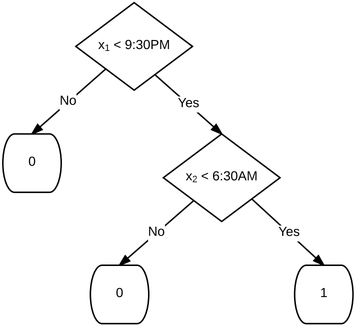
Early to bed and early to rise makes a man healthy.

which, after measuring inputs and outputs and mathematizing, becomes $y = f(x_1, x_2)$.

From the wording, it seems the model is unequivocal and deterministic. This means that for any input values (the measured values of x_1 and x_2), the output (the response) will have only one unique value.

Thus the functional form likely

looks like a *decision tree model.*



Is the Model *Really* Deterministic?

This is an ancient question... it touches on the free will vs. determinism debate. We will punt on the philosophy and ask: is this model deterministic?

Is the Model *Really* Deterministic?

This is an ancient question... it touches on the free will vs. determinism debate. We will punt on the philosophy and ask: is this model deterministic? NO.

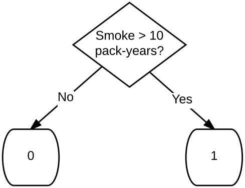
Thus, this model is wrong. Why? We can find at least one person who does not have a matching response when inputs are evaluated in f. Seems obvious but...

Smoking and Lung Cancer

Consider the model with the binary input

- y: contract lung cancer at some point (1) or not (0)
- x_1: smoke 10 pack years or more at some point in a lifetime (1) or not (0) and the response

Do you think the model should look like the below?



No... in fact "only" 16% of smokers get lung cancer compared to about 0.4% of non-smokers. Thus, the simpel model above is wrong because some responses (that is features of certain individuals) will not "fit" the model. Thus, should we throw out the whole enterprise of modeling?

Statistical Models

Mathematical models such as

$$y = f(x_1, x_2, \ldots)$$

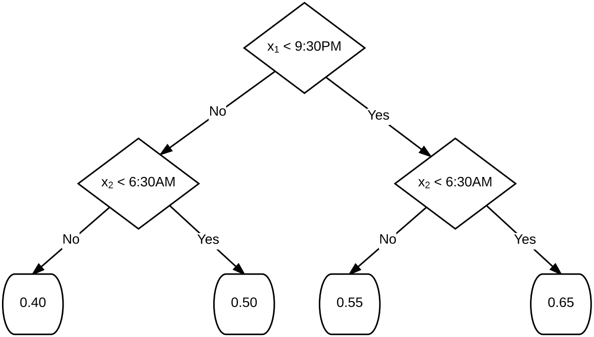
can become more forgiving to errors in f by allowing for Y to be modeled non-deterministically as a random variable (r.v.), uppercase Y. For our case of binary classification, this r.v. is the Bernoulli:

$$Y \sim \mathrm{Bernoulli}\left(f(x_1, x_2, \ldots)\right) := \begin{cases} 1 & \text{with probability } f(x_1, x_2, \ldots) \\ 0 & \text{otherwise} \end{cases}$$

Since the response is now a r.v., we call this a **statistical model**.

A Statistical Model

A more conceivable model f is:



Are there still reasons for x_1 and x_2 to be rigid binary values e.g. 1 if $x_2 <$ 6:30AM? No... but we haven't spoke about model fits nor parameters.... wait...

Is Health Dichotomous?

So, we should really update the text of the aphorism to reflect the introduction of the random variable response. It should read:

> *Early to bed and early to rise makes a man* ***more likely to be*** *healthy.*

However this seems to still suggest someone is either healthy or not healthy. Didn't the author of the aphorism, to be more accurate, say...

> *Early to bed and early to rise makes a man* ***healthier****.*

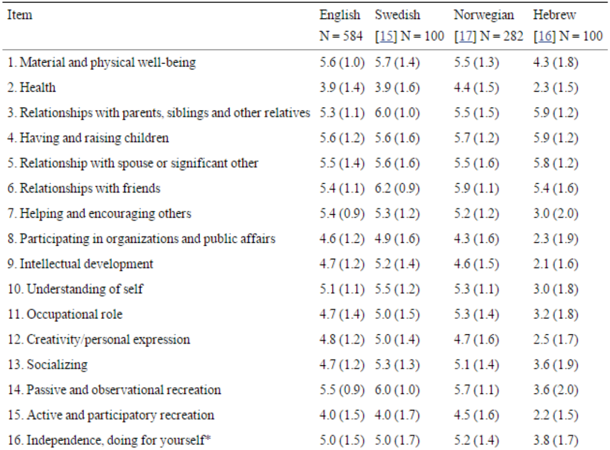
which is deterministic:

$$y = f(x_1, x_2)$$

we need some way to measure a quantity of healthiness on a continuous scale. Open problem. How can you shrink-wrap health into a single number? Diversion...

QOL: a Response Metric?

One such scale is found in Flanagan (1978) invented the precursor to the modern “Quality of Life Scale” (QOLS) metric based on assessing 7-point Likert scales. It takes 5 minutes and scores range from 16–112. Here are the categories:



Item	English N = 584	Swedish [15] N = 100	Norwegian [17] N = 282	Hebrew [16] N = 100
1. Material and physical well-being	5.6 (1.0)	5.7 (1.4)	5.5 (1.3)	4.3 (1.8)
2. Health	3.9 (1.4)	3.9 (1.6)	4.4 (1.5)	2.3 (1.5)
3. Relationships with parents, siblings and other relatives	5.3 (1.1)	6.0 (1.0)	5.5 (1.5)	5.9 (1.2)
4. Having and raising children	5.6 (1.2)	5.6 (1.6)	5.7 (1.2)	5.9 (1.2)
5. Relationship with spouse or significant other	5.5 (1.4)	5.6 (1.6)	5.5 (1.6)	5.8 (1.2)
6. Relationships with friends	5.4 (1.1)	6.2 (0.9)	5.9 (1.1)	5.4 (1.6)
7. Helping and encouraging others	5.4 (0.9)	5.3 (1.2)	5.2 (1.2)	3.0 (2.0)
8. Participating in organizations and public affairs	4.6 (1.2)	4.9 (1.6)	4.3 (1.6)	2.3 (1.9)
9. Intellectual development	4.7 (1.2)	5.2 (1.4)	4.6 (1.5)	2.1 (1.6)
10. Understanding of self	5.1 (1.1)	5.5 (1.2)	5.3 (1.1)	3.0 (1.8)
11. Occupational role	4.7 (1.4)	5.0 (1.5)	5.3 (1.4)	3.2 (1.8)
12. Creativity/personal expression	4.8 (1.2)	5.0 (1.4)	4.7 (1.6)	2.5 (1.7)
13. Socializing	4.7 (1.2)	5.3 (1.3)	5.1 (1.4)	3.6 (1.9)
14. Passive and observational recreation	5.5 (0.9)	6.0 (1.0)	5.7 (1.1)	3.6 (2.0)
15. Active and participatory recreation	4.0 (1.5)	4.0 (1.7)	4.5 (1.6)	2.2 (1.5)
16. Independence, doing for yourself*	5.0 (1.5)	5.0 (1.7)	5.2 (1.4)	3.8 (1.7)

Making Up Metrics

Yes, metrics are essentially “made up”. Good ones are engineered to carefully capture the information sought. Examples:

- The Human Freedom Index
- Democracy-Dictatorship Index
- S&P 500
- Visual Acuity 20/20, 20/40, etc

It is most important for these metrics to be monotonic (i.e. higher always means better or worse).
We also would appreciate these metrics being approximately linear. So an increase of 1 “point” on the scale means the same increase/decrease in quality. But that is usually too much to ask.

Back to Modeling

We now are considering health as a continuous number (the data type is called "continuous") but the model is still deterministic. How to we reengineer the aphorism to allow for stochasticity (randomness)?

> *Early to bed and early to rise makes a man* ***healthier on average****.*

We can then build a statistical model:

$$Y \sim g\left(f(x_1, x_2), \sigma^2, \ldots\right)$$

where $f(x_1, x_2)$ now represents the mean health for these inputs, σ^2 is now variance around that mean, and the ellipses is a technicality dealing with higher moments such as skew, etc that we will ignore for the purposes of this class. Thus, health scores are realized randomly but the mean health scores do not.

Regression Models

When the response is continuous, the statistical model is called a *regression model*. What does regression mean? Loosely, when you hear regression, you know you're modeling some continuous response (e.g. price, blood pressure, lens power). The typical way these models are written are:

$$Y = f(x_1, x_2) + \mathcal{E}$$

The equals sign makes us feel like we're back in a deterministic model. But we're not; the $\mathcal{E}$ is a r.v. known as the "noise". (The British call it the "errors" but that is confusing!) This r.v. necessarily must have no mean, $\mathbb{E}[\mathcal{E}] = 0$. Can you explain why?

Where does $\mathcal{E}$ come from?? Philosophical question... one we will return to soon.

Conditional Expectation

The model can be written even another way to belabor this point:

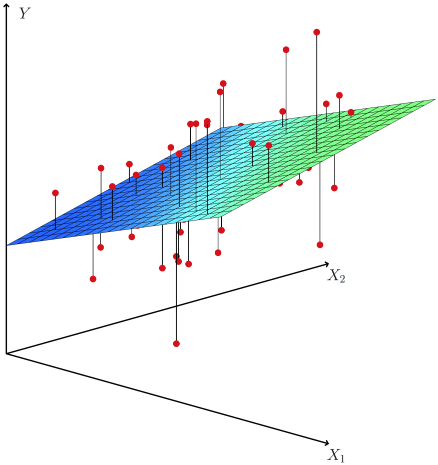
$$Y = \mathbb{E}\left[Y \mid x_1, x_2\right] + \mathcal{E}$$

where $\mathbb{E}\left[Y \mid x_1, x_2\right]$ is called the "conditional expectation function" or the "conditional mean function" and of course,

$$\mathbb{E}\left[Y \mid x_1, x_2\right] = f(x_1, x_2)$$

What does this look like?

A mock $\mathbb{E}[Y \mid x_1, x_2]$ Illustration



Generalizing the Inputs

> *Early to bed and early to rise makes a man healthier on average.*

"early to bed" and "early to rise" seem to smack of binary inputs. Either it's early or it's late... no in-between values. Again, it's probably not what the original author had in mind.

Let's reengineer the aphorism again to allow for grey area:

> ***The earlier*** *to bed and the* ***earlier to*** *rise makes a man healthier on average.*

Bedtime and Waketime Again

We began with the average bedtime and waketime and recorded it as a datetime.

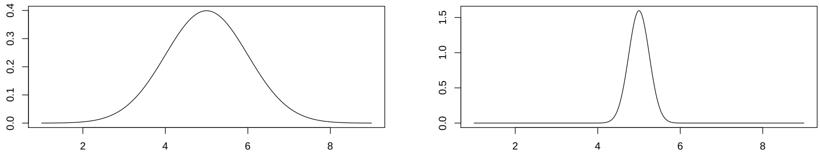
We now need to use continuous measures for x_1 and x_2. How can we do this? Should we use 9:42PM, 10:14PM, etc. as before? What is later 11:55PM or 1:02AM?

We should not use timestamps as they fail the monotonicity property that we desire to capture "lateness".

What should we do? Maybe just one number defined as the number of hours after an absurd average bedtime like 5PM? Thus, 9PM $\rightarrow x_1 = 4$ and 2AM $\rightarrow x_1 = 9$, etc. Ditto for waketime to avoid the problem of people on average waking up after 12:59PM.

The Average Is Misleading

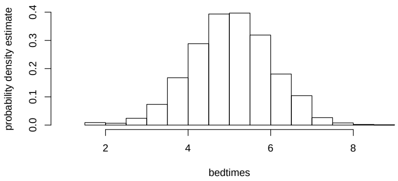
We are using average bedtime and waketime. What's wrong with an average?

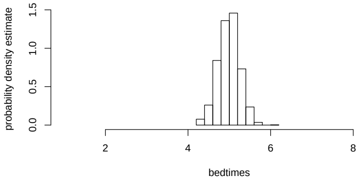


These are two bedtime distributions over many, many years. They both have the same average: 10PM. Who do you think is healthier on average? The person on the right. Why?

Designing Better Inputs

How can we get more “information” out of a person’s bedtime and waketime that is relevant to predicting health outcomes?
We likely don’t know which piece of the distribution will be helpful, so let’s just add all the information. Let’s bin by maybe 20min and record the probabilities over many years of being in that bin. For instance, 5 year bins for these two people may look like:


probability density estimate
bedtimes


probability density estimate
bedtimes