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,"

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.

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",

Examples

We make predictions all the time, saying things like:

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

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",

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.*

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)

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.

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”?

Generally, inputs and outputs are features of the

- *observation* or
- *unit* or
- *record* or
- *object* or
- *subject*

i.e. the “person” here.
Thus the model relates some *feature(s) of the observation*

What are “observations”?

Generally, inputs and outputs are features of the

- *observation* or
- *unit* or
- *record* or
- *object* or
- *subject*

i.e. the “person” here.
Thus the model relates some *feature(s) of the observation* to other *feature(s) of the observation*.

What are “observations”?

Generally, inputs and outputs are features of the

- *observation* or
- *unit* or
- *record* or
- *object* or
- *subject*

i.e. the “person” here.

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?

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 must be *measured* or *assessed*.

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 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,

$$\begin{array}{c}\text{the measured}\\ \text{outputs of an}\\ \text{observation}\end{array} = \text{model}\left(\begin{array}{c}\text{the measured}\\ \text{inputs of an}\\ \text{observation}\end{array}\right)$$

It is traditional to put the outputs on the left hand side.

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,

$$\begin{array}{c}\text{the measured}\\\text{outputs of an}\\\text{observation}\end{array} = \text{model}\left(\begin{array}{c}\text{the measured}\\\text{inputs of an}\\\text{observation}\end{array}\right)$$

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,

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,

$$\underbrace{\begin{matrix}\text{the } \textbf{guessed} \\ \text{output} \\ \text{measurements}\end{matrix}}_{\text{prediction}} = \text{model}\begin{pmatrix}\text{the measured} \\ \text{inputs of an} \\ \text{observation}\end{pmatrix}$$

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,

$$\underbrace{\begin{matrix}\text{the \textbf{guessed}}\\ \text{output}\\ \text{measurements}\end{matrix}}_{\text{prediction}} = \text{model}\begin{pmatrix}\text{the measured}\\ \text{inputs of an}\\ \text{observation}\end{pmatrix}$$

As an example new person Bob,

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

Measurements as Variables

Instead of "a measured quantity ..." we can use algebraic *variables* to denote the numerical quantities.

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.

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:

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 the true 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.

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.

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,

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,

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

and for a new observation,

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

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 into simple models with few inputs:

$$\begin{aligned} F &= G\frac{m_1 m_2}{d^2} \\ V &= IR \\ K &= \frac{1}{2}mv^2 \\ PV &= nRT \end{aligned}$$

It took us thousands of years to figure this out.

Focus: models with univariate responses.

Although general models have any number of outputs, this semester we will only consider models with one output.

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.

(I just took the first one arbitrarily).

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.

(I just took the first one arbitrarily). So, for an old observation,

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

and a new observation,

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.

(I just took the first one arbitrarily). So, for an old observation,

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

and a new observation,

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

Idea $\rightarrow$ English $\rightarrow$ Math

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

What is the response metric?

Idea → English → Math

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?

Idea → English → Math

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...

Idea → English → Math

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)?

ALSO: one also gets a feeling from the wording, there is either "healthy" or "not healthy".

Idea → English → Math

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)?

ALSO: 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.

Idea → English → Math

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)?

ALSO: 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}.

Idea → English → Math

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)?

ALSO: 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.

Idea → English → Math

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)?

ALSO: 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

Define the response clearly

Response: Healthy for ages 25–65

We still need a clear definition. Ideas?

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") can look 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

Dataframes

The historical *data frame* or *dataset* (or even more colloquially, the "data") can look 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

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

Dataframes

The historical *data frame* or *dataset* (or even more colloquially, the "data") can look 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

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 considered a predictor.

Dataframes

The historical *data frame* or *dataset* (or even more colloquially, the "data") can look 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

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 considered 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?

Dataframes

The historical *data frame* or *dataset* (or even more colloquially, the "data") can look 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

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 considered 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.

Mathematizing, this becomes $y = f(x_1, x_2)$.

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

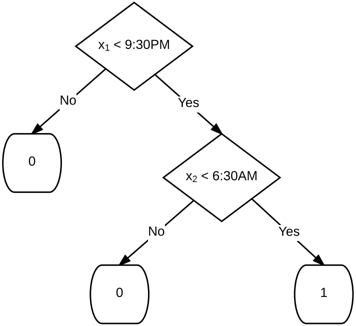
Mathematical Models are Deterministic

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

Mathematizing, this becomes $y = f(x_1, x_2)$.

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



Are Models *Really* Deterministic?

This is an ancient question... it touches on the free will vs. determinism debate.

Are Models *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 our model deterministic?

Are Models *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 our model deterministic? NO.

Thus, this model is **wrong**. Why?

Are Models *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 our 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

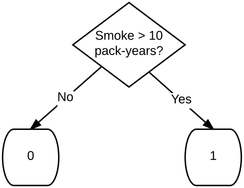
Smoking and Lung Cancer

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?



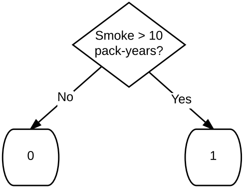
Smoking and Lung Cancer

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

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

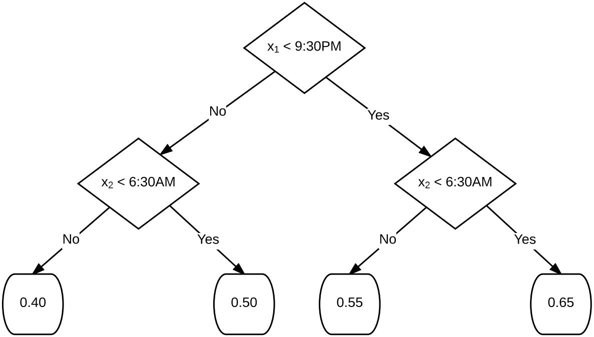
can become forgiving to **errors** in f by allowing for y to be modeled non-deterministically as a random variable (r.v.), uppercase Y. For our case of binary classification, this r.v. is the Bernoulli:

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

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

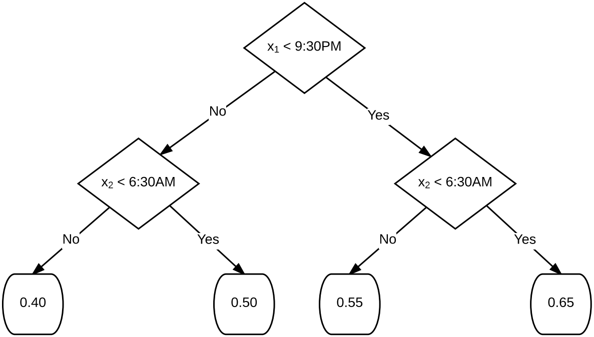
A Statistical Model

A more conceivable model f is:



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

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*

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 (and we will fix it soon):

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

we need some way to measure a quantity of healthiness on a continuous scale.

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 (and we will fix it soon):

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

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?

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

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

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

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

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).

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

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,

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 are deterministic.

Regression Models

When the response is continuous, the statistical model is called a *regression model*. What does regression mean?

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" — why?)

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" — why?) This r.v. necessarily must have no mean, $\mathbb{E}[\mathcal{E}] = 0$. Can you explain why?

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" — why?) 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:

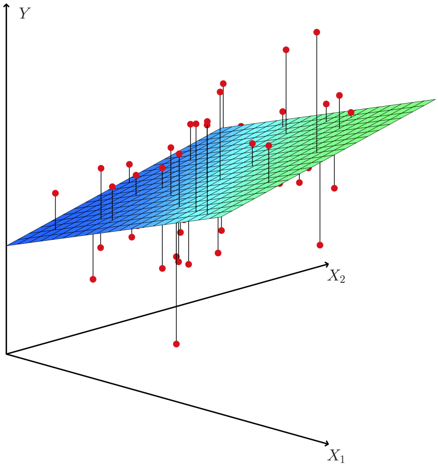
$$Y = \mathbb{E}\left[Y \mid x_1, x_2\right] + \mathcal{E}$$

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,

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

Specifying the model for f is sometimes called "conditional mean modeling".
How does one think of $\mathbb{E}\left[Y \mid x_1, x_2\right]$?

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.

The wording "early to bed" and "early to rise" seem to smack of

Generalizing the Inputs

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

The wording "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.

Generalizing the Inputs

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

The wording "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:

Generalizing the Inputs

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

The wording "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.

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?

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".

Bedtime and Waketime Again

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

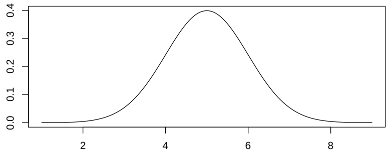
We now need to use continuous measures for x_1 and x_2. How can we do this? Should we use 9:42PM, 10:14PM, etc. as before? What is later 11:55PM or 1:02AM?

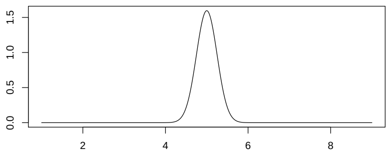
We should not use timestamps as they fail the monotonicity property that we desire to capture "lateness".

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?

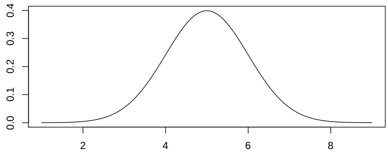


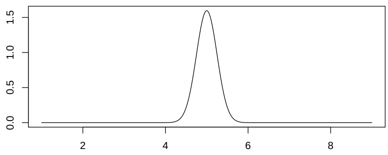


These are two bedtime distributions over many, many years. They both have the same average: 10PM. Who do you think is healthier on average?

The Average Is Misleading

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





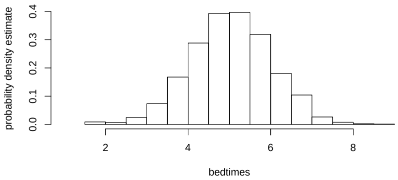
These are two bedtime distributions over many, many years. They both have the same average: 10PM. Who do you think is healthier on average? The person on the right. Why?

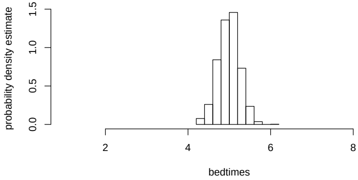
Designing Better Inputs

How can we get more "information" out of a person's bedtime and waketime that is relevant to predicting health outcomes?

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:





All bin values can be used as inputs. So in this model, $p > 2$. It could be almost 100. This is called **featurization** — designing features and we will talk more about it later.

Flexible Inputs $p = 2 \to p \approx 100$

- Advantage: We can fit more exact rules like if the proportion of times went to bed past 1AM is 10% ... then health drops considerably and then more so for 2AM

Flexible Inputs $p = 2 \rightarrow p \approx 100$

- Advantage: We can fit more exact rules like if the proportion of times went to bed past 1AM is 10% ... then health drops considerably and then more so for 2AM
- Disadvantage: It makes the model hard to fit and interpret. There are a lot of "degrees of freedom" now (a term you've heard before). A lot more in this later as this is the most important topic in this course.

Summary

Models relate inputs to outputs. Inputs and outputs are measurements or assessments on objects / observations. Here, we consider only one output and name it the response.

Summary

Models relate inputs to outputs. Inputs and outputs are measurements or assessments on objects / observations. Here, we consider only one output and name it the response. All the inputs we then consider to be predictors that explain the response via the model f.

Summary

Models relate inputs to outputs. Inputs and outputs are measurements or assessments on objects / observations. Here, we consider only one output and name it the response. All the inputs we then consider to be predictors that explain the response via the model f.

Data frames have rows that are observations and columns that are predictors (with one column that is a response).

Summary

Models relate inputs to outputs. Inputs and outputs are measurements or assessments on objects / observations. Here, we consider only one output and name it the response. All the inputs we then consider to be predictors that explain the response via the model f.

Data frames have rows that are observations and columns that are predictors (with one column that is a response).

Each predictor and the response have a certain data type.

Summary

Models relate inputs to outputs. Inputs and outputs are measurements or assessments on objects / observations. Here, we consider only one output and name it the response. All the inputs we then consider to be predictors that explain the response via the model f.

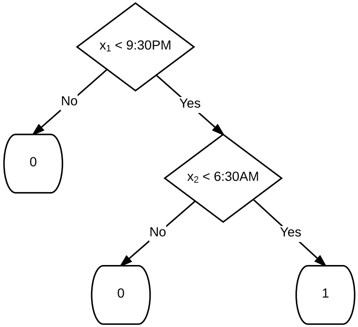
Data frames have rows that are observations and columns that are predictors (with one column that is a response).

Each predictor and the response have a certain data type. If the data type of the response is categorical (or binary as in two categories), we have a classification model. If the data type of the response is a continuous metric, we have a regression model.

The Aphorism Revisited

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

which may imply binary x_1, x_2 and y and thus f likely looks like:



New question: where did this model come from??

History

As the olde englysshe prouerbe sayth in this wyse. Who soo woll ryse erly shall be holy helthy & zely.
- The Book of St. Albans, 1486

Earely to bed and earely to rise, makes a man healthy, wealthy, and wise.
- John Clarke's Paroemiologia Anglo-Latina, 1639 (collection of proverbs)
-Benjamin Franklin, 1735 (popularized it in American English)

We, as a people, built this model and it's been validated over centuries. How did we build it?

Humans as Model Builders

1. We made it up. Don't laugh... you will see on the homework why this may be subtle.

Humans as Model Builders

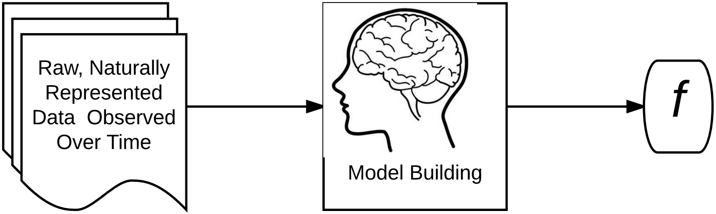
1. We made it up. Don't laugh... you will see on the homework why this may be subtle.
2. We used a "data-driven approach".

Humans as Model Builders

1. We made it up. Don't laugh... you will see on the homework why this may be subtle.
2. We used a "data-driven approach".
   Likely, many individuals independently *observed* people (the unit of observation), assessed both y, their healthiness level over their lives and **assessed / measured all features and noted that** their x_1, x_2 bedtime and waketime habits were related via a simple pattern relating these inputs to the response.

Humans as Model Builders

1. We made it up. Don't laugh... you will see on the homework why this may be subtle.
2. We used a "data-driven approach".
   Likely, many individuals independently *observed* people (the unit of observation), assessed both y, their healthiness level over their lives and **assessed / measured all features and noted that** their x_1, x_2 bedtime and waketime habits were related via a simple pattern relating these inputs to the response.



How does this work?

We notice phenomena around ourselves we wish to explain. For example, a person...

- ... gets a job at Citadel Capital or not

How does this work?

We notice phenomena around ourselves we wish to explain. For example, a person...

- ... gets a job at Citadel Capital or not
- ... gets a mortgage approved or not

How does this work?

We notice phenomena around ourselves we wish to explain. For example, a person...

- ... gets a job at Citadel Capital or not
- ... gets a mortgage approved or not
- ... runs a mile in t minutes

How does this work?

We notice phenomena around ourselves we wish to explain. For example, a person...

- ... gets a job at Citadel Capital or not
- ... gets a mortgage approved or not
- ... runs a mile in t minutes
- ... has a degree of healthiness (in our aphorism example)

We then pore through the *raw, naturally represented data*. What kind of information is this??

How does this work?

We notice phenomena around ourselves we wish to explain. For example, a person...

- ... gets a job at Citadel Capital or not
- ... gets a mortgage approved or not
- ... runs a mile in t minutes
- ... has a degree of healthiness (in our aphorism example)

We then pore through the *raw, naturally represented data*. What kind of information is this??
Everything we notice about the people we observe! Imagine every encounter with the person for years and years (10 minutes here, 5 minutes there), full video clips, audio recordings, smells, touches, tastes, everything.

How does this work?

We notice phenomena around ourselves we wish to explain. For example, a person...

- ... gets a job at Citadel Capital or not
- ... gets a mortgage approved or not
- ... runs a mile in t minutes
- ... has a degree of healthiness (in our aphorism example)

We then pore through the *raw, naturally represented data*. What kind of information is this??
Everything we notice about the people we observe! Imagine every encounter with the person for years and years (10 minutes here, 5 minutes there), full video clips, audio recordings, smells, touches, tastes, everything.
How many measurements is that? (What's the dimensionality of the input space?) Immeasurable and cannot be defined. But we know it's HUGE! Note: this is technically called **deep learning**.

A Deep Learning Model You've Built

Is this a cat?

Response?

A Deep Learning Model You've Built

Is this a cat?

Response? Cat or not (0 or 1 numerically). Predictors in the model?

A Deep Learning Model You've Built

Is this a cat?

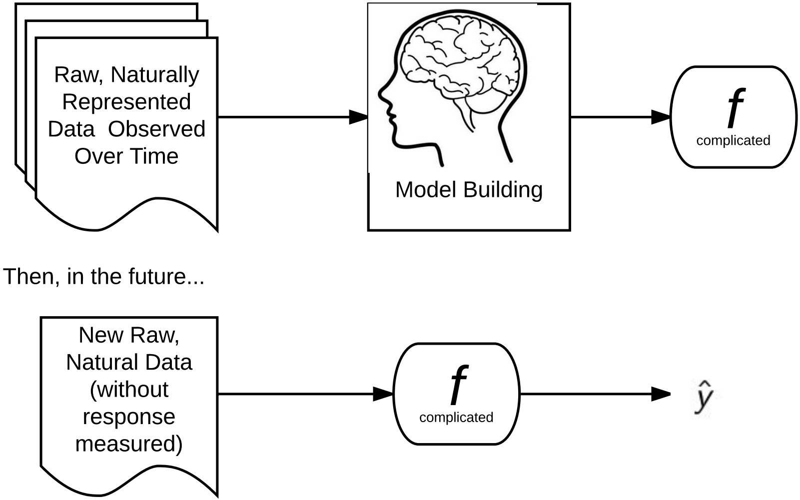
Response? Cat or not (0 or 1 numerically). Predictors in the model? Entire image... raw natural data representation.

A Deep Learning Model You've Built

Is this a cat?

Response? Cat or not (0 or 1 numerically). Predictors in the model? Entire image... raw natural data representation. The brain shrinks the space down to a small number of predictors. You've already built this model (but only in your head... and you don't even know how it works).

Prediction in a Deep-Learned Human Model



Weakness in a Deep-Learned Human Model

Weaknesses

- The model cannot be communicated precisely.

Weakness in a Deep-Learned Human Model

Weaknesses

- The model cannot be communicated precisely.
- The inputs are too complicated.

Weakness in a Deep-Learned Human Model

Weaknesses

- The model cannot be communicated precisely.
- The inputs are too complicated.
- The functional form cannot be shared.

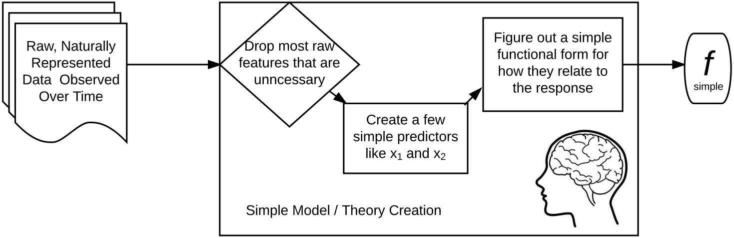
Weakness in a Deep-Learned Human Model

Weaknesses

- The model cannot be communicated precisely.
- The inputs are too complicated.
- The functional form cannot be shared.
- And thus no prediction can be made with it (unless it is the same brain that makes the prediction that built the model).

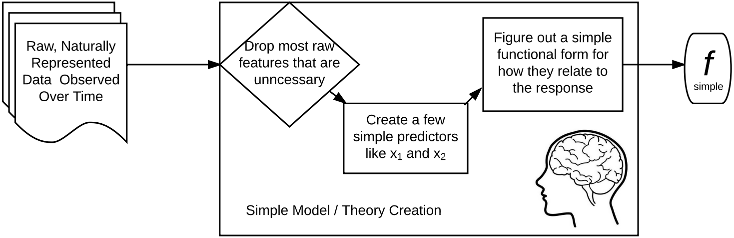
But we Create Simple Models. How?

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



But we Create Simple Models. How?

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



This is the process by which we put forth the model "early to bed and early to rise makes a man healthy".

Models we are Good and Bad At

We are ...

Good at Building Statistical Models when...

Models we are Good and Bad At

We are ...

Good at Building Statistical Models when...

the number of variables that truly matter is small and there's low noise.

Bad at Building Statistical Models when...

Models we are Good and Bad At

We are ...

Good at Building Statistical Models when...

the number of variables that truly matter is small and there's low noise.

Bad at Building Statistical Models when...

inputs are already derivative features of the raw data representation, are numeric and there a lot of them (p large), and noise is large ($\mathbb{V}\text{ar}\left[\mathcal{E}\right] >> 0$).

We are Frequently Bad...

Famous finding: Paul Meehl in 1954 found when comparing predictions from a panel of clinicians with PhD's and predictions with a linear model (a simple statistical model),

We are Frequently Bad...

Famous finding: Paul Meehl in 1954 found when comparing predictions from a panel of clinicians with PhD's and predictions with a linear model (a simple statistical model), he found the linear model beat the PhD's.

> *When one is dealing with human lives and life opportunities, it is immoral to adopt a mode of decision-making which has been demonstrated repeatedly to be either inferior in success rate...*

We are Frequently Bad...

Famous finding: Paul Meehl in 1954 found when comparing predictions from a panel of clinicians with PhD's and predictions with a linear model (a simple statistical model), he found the linear model beat the PhD's.

> *When one is dealing with human lives and life opportunities, it is immoral to adopt a mode of decision-making which has been demonstrated repeatedly to be either inferior in success rate...*

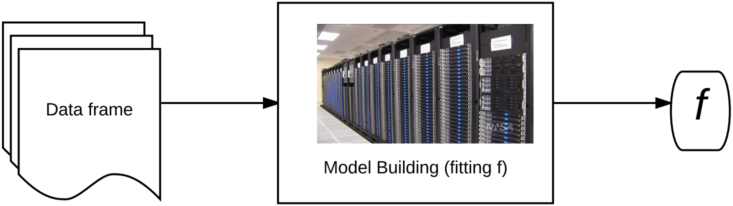
The clinicians make a diagnosis on the basis of a quick meeting and a whole bunch of numeric variables: age, serum glucose, blood pressure, symptom measurements... difficult models for us to build.

Can we Use Artificial Intelligence (AI)?

Can we use computers to build models, especially the models we're bad at?

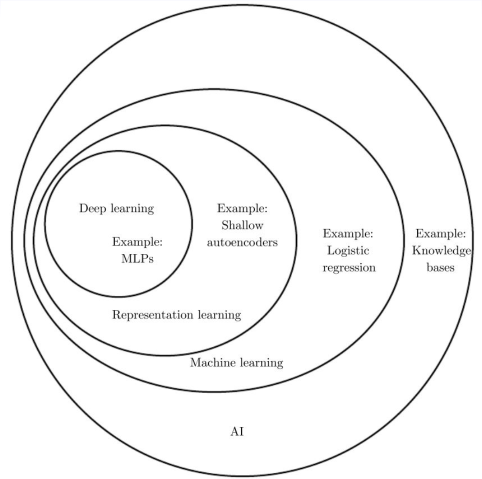
Can we Use Artificial Intelligence (AI)?

Can we use computers to build models, especially the models we're bad at?



Luckily, yes. And this is a main advantage of artificial intelligence.

Types of AI?



(Fig 1.4 in Goodfellow et al., 2017)

What Does Input to AI Look Like?

Let's return to our model:

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

We observe every encounter with the person for years and years (10 minutes here, 5 minutes there), full video clips, audio recordings, smells, touches, tastes, everything.

What Does Input to AI Look Like?

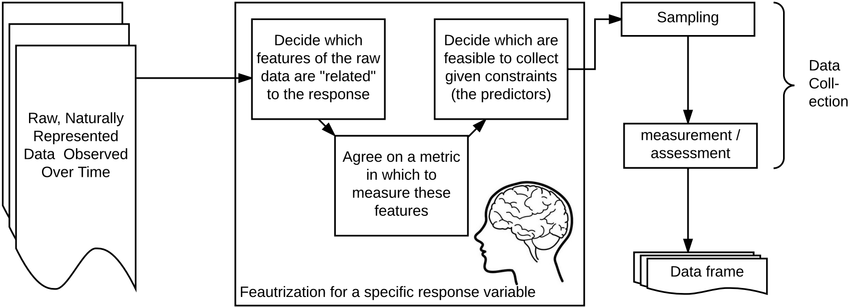
Let's return to our model:

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

We observe every encounter with the person for years and years (10 minutes here, 5 minutes there), full video clips, audio recordings, smells, touches, tastes, everything.

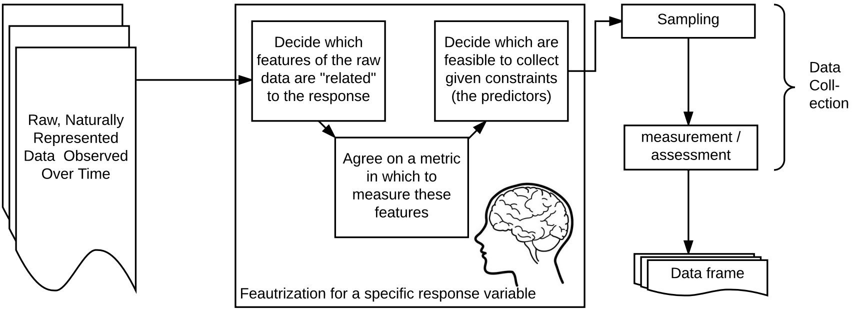
Can we enter all this into a computer? No, not now, possibly not ever. Also, willl be using statistical modeling so it needs well-defined measurements. So we need to "featurize" and then "collect data".

What is Featurization & Data Collection?



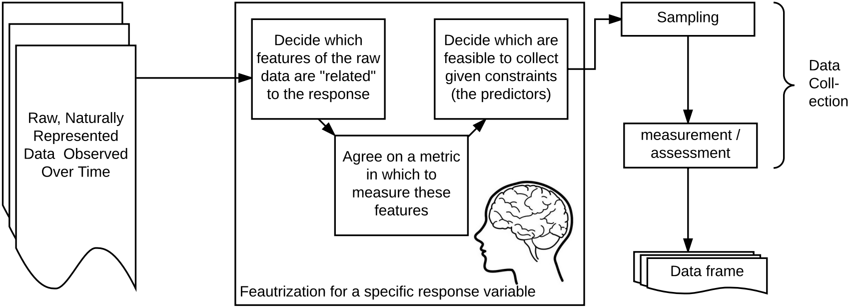
What is featurization?

What is Featurization & Data Collection?



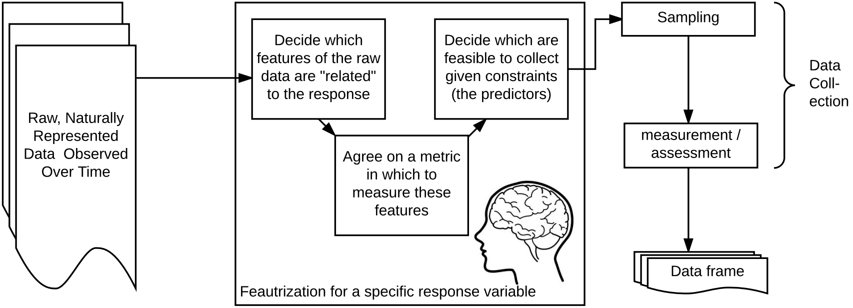
What is featurization? Deciding predictors. What is data collection?

What is Featurization & Data Collection?



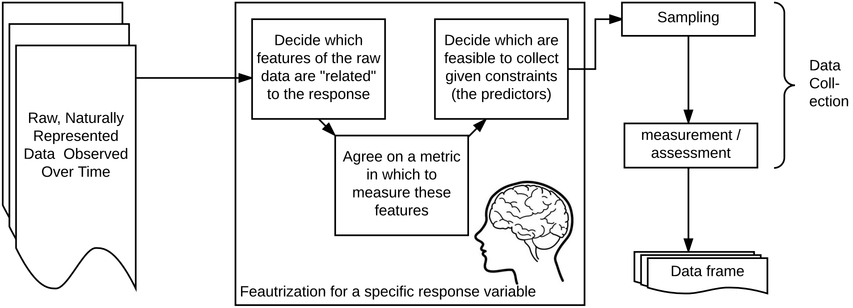
What is featurization? Deciding predictors. What is data collection? (a) sampling units then

What is Featurization & Data Collection?



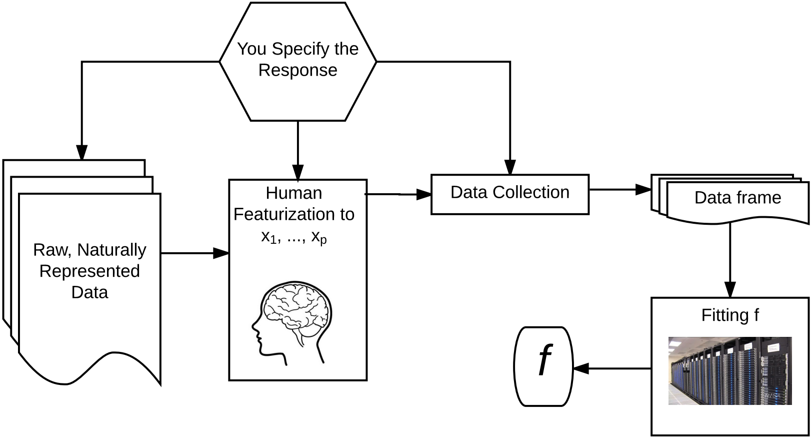
What is featurization? Deciding predictors. What is data collection? (a) sampling units then (b) measuring the numeric values of the predictors

What is Featurization & Data Collection?

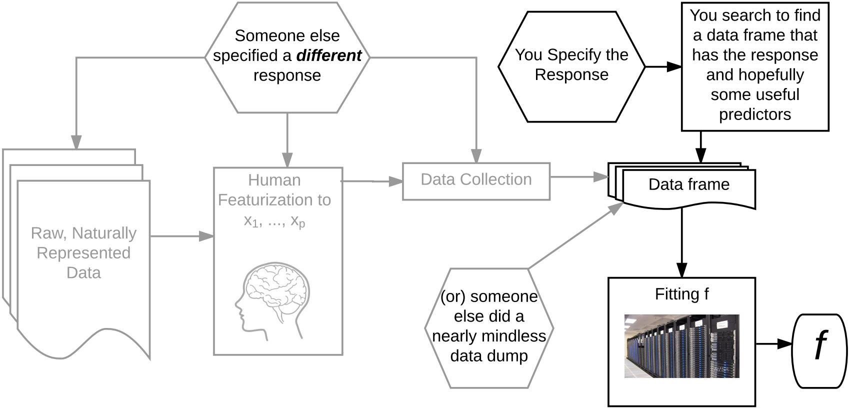


What is featurization? Deciding predictors. What is data collection? (a) sampling units then (b) measuring the numeric values of the predictors (note: some measurements may be missing).

What is Good Machine Learning?



What is Day-Day Machine Learning?

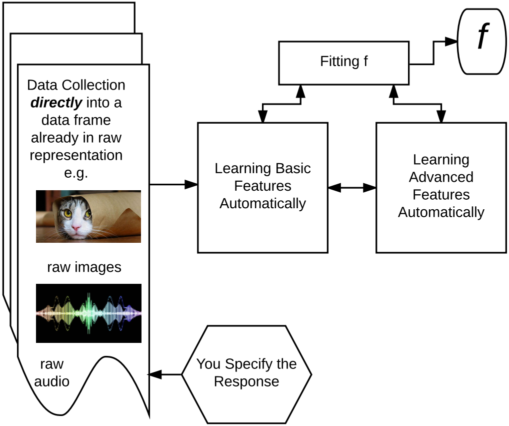


This is bad... but it's what we generally do all day...

Baseball Data Questions

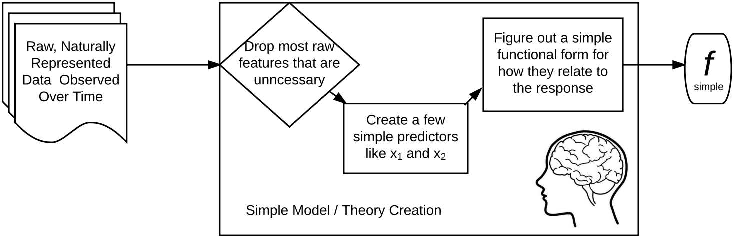
- Response?
- Predictors?
- $n = ?\ \ p = ?$
- How was this data frame likely collected?
- Do you think by studying this dataset you can predict salaries?

Aside: what is Deep Learning?



Learning complex features automatically (the cutting edge).
Software running self-driving cars use this.

What are we really bad at?



Figuring out those functional forms... and the computer excels at it.

The Fundamental Statistical Problem

We know our response variable, we've picked features, made measurements and now have an $n \times (p+1)$ data frame. We know the response looks like

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

where f represents the conditional mean $\mathbb{E}\left[Y \mid x_1, x_2, \ldots, x_p\right]$ and the $\mathcal{E}$ r.v. is random noise added atop the conditional mean but we don't know f!

The Fundamental Statistical Problem

We know our response variable, we've picked features, made measurements and now have an $n \times (p+1)$ data frame. We know the response looks like

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

where f represents the conditional mean $\mathbb{E}[Y \mid x_1, x_2, \ldots, x_p]$ and the $\mathcal{E}$ r.v. is random noise added atop the conditional mean but we don't know f!

So we have to learn / infer f the best we could from the historical dataset. We denote this fit $\hat{f}$.

The Fundamental Statistical Problem

We know our response variable, we've picked features, made measurements and now have an $n \times (p+1)$ data frame. We know the response looks like

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

where f represents the conditional mean $\mathbb{E}[Y \mid x_1, x_2, \ldots, x_p]$ and the $\mathcal{E}$ r.v. is random noise added atop the conditional mean but we don't know f!

So we have to learn / infer f the best we could from the historical dataset. We denote this fit $\hat{f}$. What does this mean in the baseball data?

Two Worldviews: (I) Parametric

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

Philosophy of creating the fit, $\hat{f}$

Two Worldviews: (I) Parametric

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

Philosophy of creating the fit, $\hat{f}$

- We believe that f has a very “nice” form.

Two Worldviews: (I) Parametric

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

Philosophy of creating the fit, $\hat{f}$

- We believe that f has a very "nice" form.
- We value "simplicity" because we wholeheartedly believe in Occam's Razor (simplest theory is the one that should be retained).

Two Worldviews: (I) Parametric

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

Philosophy of creating the fit, $\hat{f}$

- We believe that f has a very "nice" form.
- We value "simplicity" because we wholeheartedly believe in Occam's Razor (simplest theory is the one that should be retained).
- We love the simple mathematical models such as $V = IR$ and $F = G\frac{m_1 m_2}{d^2}$ and always try to live up to their elegance.

Two Worldviews: (I) Parametric

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

Philosophy of creating the fit, $\hat{f}$

- We believe that f has a very "nice" form.
- We value "simplicity" because we wholeheartedly believe in Occam's Razor (simplest theory is the one that should be retained).
- We love the simple mathematical models such as $V = IR$ and $F = G\frac{m_1 m_2}{d^2}$ and always try to live up to their elegance.
- We care very much about knowing how f works inside i.e. how each of the x_j predictors are affect the conditional mean. Knowing how this system works is our top priority.

Two Worldviews: (I) Parametric

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

Philosophy of creating the fit, $\hat{f}$

- We believe that f has a very "nice" form.
- We value "simplicity" because we wholeheartedly believe in Occam's Razor (simplest theory is the one that should be retained).
- We love the simple mathematical models such as $V = IR$ and $F = G\frac{m_1 m_2}{d^2}$ and always try to live up to their elegance.
- We care very much about knowing how f works inside i.e. how each of the x_j predictors are affect the conditional mean. Knowing how this system works is our top priority.
- Absolute prediction accuracy is not our #1 focus. $f \not\approx \hat{f}$

Two Worldviews: (II) Non-Parametric

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

Philosophy of creating the fit, $\hat{f}$

Two Worldviews: (II) Non-Parametric

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

Philosophy of creating the fit, $\hat{f}$

- We do not make any assumptions about the form of f "nice" or not.

Two Worldviews: (II) Non-Parametric

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

Philosophy of creating the fit, $\hat{f}$

- We do not make any assumptions about the form of f "nice" or not.
- We have no value for simplicity. We are okay with a world being messy and complex.

Two Worldviews: (II) Non-Parametric

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

Philosophy of creating the fit, $\hat{f}$

- We do not make any assumptions about the form of f “nice” or not.
- We have no value for simplicity. We are okay with a world being messy and complex.
- We believe simple mathematical models such as $V = IR$ and $F = G\frac{m_1 m_2}{d^2}$ work for idealized situations, but never real-life situations with aribitrarily selected predictors measured in arbitrary ways.

Two Worldviews: (II) Non-Parametric

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

Philosophy of creating the fit, $\hat{f}$

- We do not make any assumptions about the form of f "nice" or not.
- We have no value for simplicity. We are okay with a world being messy and complex.
- We believe simple mathematical models such as $V = IR$ and $F = G\frac{m_1 m_2}{d^2}$ work for idealized situations, but never real-life situations with aribitrarily selected predictors measured in arbitrary ways.
- We would like to know how f works, but it's likely very complex, so it's not our top priority.

Two Worldviews: (II) Non-Parametric

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

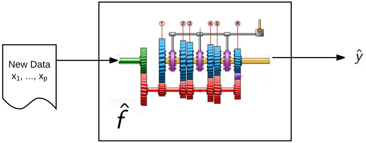
Philosophy of creating the fit, $\hat{f}$

- We do not make any assumptions about the form of f "nice" or not.
- We have no value for simplicity. We are okay with a world being messy and complex.
- We believe simple mathematical models such as $V = IR$ and $F = G\frac{m_1 m_2}{d^2}$ work for idealized situations, but never real-life situations with aribitrarily selected predictors measured in arbitrary ways.
- We would like to know how f works, but it's likely very complex, so it's not our top priority.
- Absolute prediction accuracy is indeed our #1 focus. It's the bottom line. We want $f \approx \hat{f}$ as close as possible.

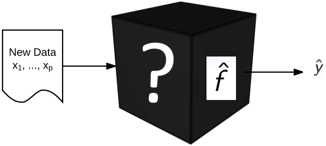
Two Worldviews

The process to find $\hat{f}$ is known by many names:

- parametric modeling
- model fitting
- statistical modeling
- white box modeling



- non-parametric modeling
- function fitting
- function approximation
- response surface methodology
- machine learning
- black box modeling



What is a Parametric Model?

If we see f is a parametric model, we mean that

$$f(x_1, x_2, \ldots, x_p) \approx s(x_1, x_2, \ldots, x_p; \theta_1, \theta_2, \ldots, \theta_\ell)$$

i.e. f has a assjumed simple form s that has various knobs that can be adjusted based on the data.

What is a Parametric Model?

If we see f is a parametric model, we mean that

$$f(x_1, x_2, \ldots, x_p) \approx s(x_1, x_2, \ldots, x_p; \theta_1, \theta_2, \ldots, \theta_\ell)$$

i.e. f has a assjumed simple form s that has various knobs that can be adjusted based on the data. These knobs are called the **unknown parameters**. Here there are ℓ of them $\theta_1, \theta_2, \ldots, \theta_\ell$. It is said the model has ℓ "degrees of freedom".

What is a Parametric Model?

If we see f is a parametric model, we mean that

$$f(x_1, x_2, \dots, x_p) \approx s(x_1, x_2, \dots, x_p; \theta_1, \theta_2, \dots, \theta_\ell)$$

i.e. f has a assjumed simple form s that has various knobs that can be adjusted based on the data. These knobs are called the **unknown parameters**. Here there are ℓ of them $\theta_1, \theta_2, \dots, \theta_\ell$. It is said the model has ℓ "degrees of freedom". An example of this is the linear model,

$$s(x_1, x_2, \dots, x_p) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

with p predictors, there are $p + 1$ degrees of freedom (the intercept is also a knob that can be twisted). It is traditional to call the θ's in the linear model β's due to historical reasons.

What is $\hat{f}$ in a linear model?

Then we need to create a fit $\hat{f}$ that means we need estimates of all the parameters:

$$\hat{f}(x_1, x_2, \ldots, x_p) = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \ldots + \hat{\beta}_p x_p$$

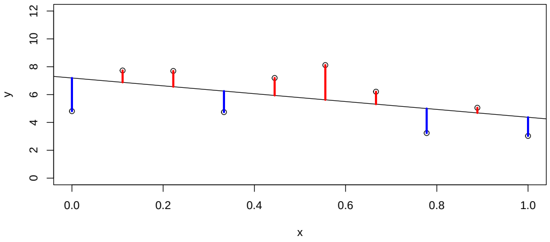
then we can use $\hat{f}$ on new data (where the response is not observed), say $x_1^*, x_2^*, \ldots, x_p^*$ to get a prediction:

$$\hat{y} = \hat{f}(x_1^*, x_2^*, \ldots, x_p^*)$$

Where do we get $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$ from?

What defines a good fit?

Consider the simple linear regression (one predictor) i.e. $s(x) = \beta_0 + \beta_1 x$ thus we need to figure out $\hat{\beta}_0$ and $\hat{\beta}_1$, constituting the model fitting. Let's say given data, we guess that $\hat{\beta}_0 = 7.2$ and $\hat{\beta}_1 = -2.8$. Would this be a good fit?



The line is our $\hat{y}(x)$ i.e. all possible predictions. Seems sometimes we undershot the response (the red) i.e. $y - \hat{y} > 0$ and sometimes overshot the response (the blue) i.e. $y - \hat{y} < 0$.

The Role of the Residuals

We call $e_i := y_i - \hat{y}_i$ the ith residual. Since we fit all of our historical data there are n residuals $e_1, \ldots, e_n$. Wouldn't it be nice to keep these small?

The Role of the Residuals

We call $e_i := y_i - \hat{y}_i$ the ith residual. Since we fit all of our historical data there are n residuals $e_1, \ldots, e_n$. Wouldn't it be nice to keep these small?

How to create $\hat{f}$

- Minimize $e_1, \ldots, e_n$ while
- staying true to our assumption $s(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

What does "minimize $e_1, \ldots, e_n$" mean?

The Role of the Residuals

We call $e_i := y_i - \hat{y}_i$ the ith residual. Since we fit all of our historical data there are n residuals $e_1, \ldots, e_n$. Wouldn't it be nice to keep these small?

How to create $\hat{f}$

- Minimize $e_1, \ldots, e_n$ while
- staying true to our assumption $s(x_1, x_2, \ldots, x_p) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$

What does "minimize $e_1, \ldots, e_n$" mean? We need to define an overall error metric. This is called the **loss function**.

Loss Functions

Let $L = L(e_1, \ldots, e_n)$ be a loss function. What's the best we can do?

Loss Functions

Let $L = L(e_1, \ldots, e_n)$ be a loss function. What's the best we can do? If we have $e_1 = 0, \ldots, e_n = 0$, then we have no error whatsoever. Thus $L \geq 0$.

Loss Functions

Let $L = L(e_1, \ldots, e_n)$ be a loss function. What's the best we can do? If we have $e_1 = 0, \ldots, e_n = 0$, then we have no error whatsover. Thus $L \geq 0$. It is common to sum up the individual losses for each residual. Here are some loss functions below:

- $L = |e_1| + |e_2| + \ldots + |e_n| = \sum_{i=1}^{n} |e_i|$

Loss Functions

Let $L = L(e_1, \ldots, e_n)$ be a loss function. What's the best we can do? If we have $e_1 = 0, \ldots, e_n = 0$, then we have no error whatsoever. Thus $L \geq 0$. It is common to sum up the individual losses for each residual. Here are some loss functions below:

- $L = |e_1| + |e_2| + \ldots + |e_n| = \sum_{i=1}^{n} |e_i|$
  This is known as L1 error or absolute error or sum of absolute error

Loss Functions

Let $L = L(e_1, \ldots, e_n)$ be a loss function. What's the best we can do? If we have $e_1 = 0, \ldots, e_n = 0$, then we have no error whatsoever. Thus $L \geq 0$. It is common to sum up the individual losses for each residual. Here are some loss functions below:

- $L = |e_1| + |e_2| + \ldots + |e_n| = \sum_{i=1}^{n} |e_i|$
  This is known as L1 error or absolute error or sum of absolute error
- $L = e_1^2 + e_2^2 + \ldots + e_n^2$

Loss Functions

Let $L = L(e_1, \ldots, e_n)$ be a loss function. What's the best we can do? If we have $e_1 = 0, \ldots, e_n = 0$, then we have no error whatsoever. Thus $L \geq 0$. It is common to sum up the individual losses for each residual. Here are some loss functions below:

- $L = |e_1| + |e_2| + \ldots + |e_n| = \sum_{i=1}^{n} |e_i|$
  This is known as L1 error or absolute error or sum of absolute error
- $L = e_1^2 + e_2^2 + \ldots + e_n^2$
  This is known as L2 error or sum of squared error (SSE)

Loss Functions

Let $L = L(e_1, \ldots, e_n)$ be a loss function. What's the best we can do? If we have $e_1 = 0, \ldots, e_n = 0$, then we have no error whatsover. Thus $L \geq 0$. It is common to sum up the individual losses for each residual. Here are some loss functions below:

- $L = |e_1| + |e_2| + \ldots + |e_n| = \sum_{i=1}^{n} |e_i|$
  This is known as L1 error or absolute error or sum of absolute error
- $L = e_1^2 + e_2^2 + \ldots + e_n^2$
  This is known as L2 error or sum of squared error (SSE)

Linear model fitting and machine learning algorithms mostly use SSE.

How to create $\hat{f}$

Loss Functions

Let $L = L(e_1, \ldots, e_n)$ be a loss function. What's the best we can do? If we have $e_1 = 0, \ldots, e_n = 0$, then we have no error whatsover. Thus $L \geq 0$. It is common to sum up the individual losses for each residual. Here are some loss functions below:

- $L = |e_1| + |e_2| + \ldots + |e_n| = \sum_{i=1}^{n} |e_i|$
  This is known as L1 error or absolute error or sum of absolute error
- $L = e_1^2 + e_2^2 + \ldots + e_n^2$
  This is known as L2 error or sum of squared error (SSE)

Linear model fitting and machine learning algorithms mostly use SSE.

How to create $\hat{f}$

- Minimize SSE via a

Loss Functions

Let $L = L(e_1, \ldots, e_n)$ be a loss function. What's the best we can do? If we have $e_1 = 0, \ldots, e_n = 0$, then we have no error whatsover. Thus $L \geq 0$. It is common to sum up the individual losses for each residual. Here are some loss functions below:

- $L = |e_1| + |e_2| + \ldots + |e_n| = \sum_{i=1}^{n} |e_i|$
  This is known as L1 error or absolute error or sum of absolute error
- $L = e_1^2 + e_2^2 + \ldots + e_n^2$
  This is known as L2 error or sum of squared error (SSE)

Linear model fitting and machine learning algorithms mostly use SSE.

How to create $\hat{f}$

- Minimize SSE via a
- search over all possible values of $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$

Loss Functions

Let $L = L(e_1, \ldots, e_n)$ be a loss function. What's the best we can do? If we have $e_1 = 0, \ldots, e_n = 0$, then we have no error whatsover. Thus $L \geq 0$. It is common to sum up the individual losses for each residual. Here are some loss functions below:

- $L = |e_1| + |e_2| + \ldots + |e_n| = \sum_{i=1}^{n} |e_i|$
  This is known as L1 error or absolute error or sum of absolute error
- $L = e_1^2 + e_2^2 + \ldots + e_n^2$
  This is known as L2 error or sum of squared error (SSE)

Linear model fitting and machine learning algorithms mostly use SSE.

How to create $\hat{f}$

- Minimize SSE via a
- search over all possible values of $\left\{\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p\right\}$

Luckily the computer does all of this for you and it just pops out its answer $\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p$.

Is Minimal SSE what you ACTUALLY want?

Imagine your response is how much IBM stock moves on a percentage basis for one trading day. Consider two scenarios:

Is Minimal SSE what you ACTUALLY want?

Imagine your response is how much IBM stock moves on a percentage basis for one trading day. Consider two scenarios:

- If $y = -0.25\%$ and $\hat{y} = 0.5\%$ then $e = -0.25\% - 0.5\% = -0.75\%$ and the squared loss for your prediction would be 0.00005625.

Is Minimal SSE what you ACTUALLY want?

Imagine your response is how much IBM stock moves on a percentage basis for one trading day. Consider two scenarios:

- If $y = -0.25\%$ and $\hat{y} = 0.5\%$ then $e = -0.25\% - 0.5\% = -0.75\%$ and the squared loss for your prediction would be 0.00005625.
- If $y = 1.25\%$ and $\hat{y} = 0.5\%$ then $e = 1.25\% - 0.5\% = 0.75\%$ and the squared loss for your prediction would be 0.00005625.

The loss that the "computer sees" for both scenarios is the same. But isn't the first prediction "worse for you" than the second prediction?

Is Minimal SSE what you ACTUALLY want?

Imagine your response is how much IBM stock moves on a percentage basis for one trading day. Consider two scenarios:

- If $y = -0.25\%$ and $\hat{y} = 0.5\%$ then $e = -0.25\% - 0.5\% = -0.75\%$ and the squared loss for your prediction would be 0.00005625.
- If $y = 1.25\%$ and $\hat{y} = 0.5\%$ then $e = 1.25\% - 0.5\% = 0.75\%$ and the squared loss for your prediction would be 0.00005625.

The loss that the "computer sees" for both scenarios is the same. But isn't the first prediction "worse for you" than the second prediction?

We won't have time to get into custom and asymmetric cost functions. But you need to keep this in mind when you consider the predictions you make with all of the $\hat{f}$'s we discuss in this class.

An Interpretable Measure of Fit

Let $L = L(e_1, \ldots, e_n) = SSE$ be our loss function. Okay, SSE = 654.4567. Did the model fit well?

An Interpretable Measure of Fit

Let $L = L(e_1, \ldots, e_n) = SSE$ be our loss function. Okay, SSE = 654.4567. Did the model fit well? What are the units of SSE?

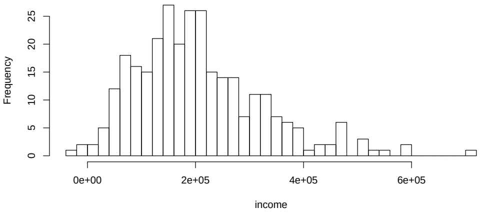
An Interpretable Measure of Fit

Let $L = L(e_1, \ldots, e_n) = SSE$ be our loss function. Okay, SSE = 654.4567. Did the model fit well? What are the units of SSE? The response units squared.

An Interpretable Measure of Fit

Let $L = L(e_1, \ldots, e_n) = SSE$ be our loss function. Okay, SSE = 654.4567. Did the model fit well? What are the units of SSE? The response units squared. It also grows linearly with n. We cannot compare models using SSE.

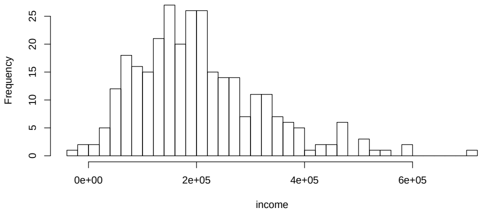
Consider a dataset for income among HS grads that didn't go to college. You have $n = 1,000$ historical responses (the incomes). Here they are:



An Interpretable Measure of Fit

Let $L = L(e_1, \ldots, e_n) = SSE$ be our loss function. Okay, SSE = 654.4567. Did the model fit well? What are the units of SSE? The response units squared. It also grows linearly with n. We cannot compare models using SSE.

Consider a dataset for income among HS grads that didn't go to college. You have $n = 1,000$ historical responses (the incomes). Here they are:



Imagine there was no predictors but you still would like to produce predictions. What would you do?

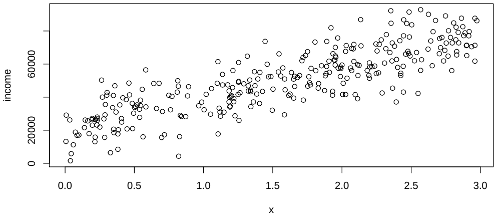
Shoot Blind

The best guess (i.e. the one with minimal SSE) is the average $\hat{y} = \bar{y}$ for any new observation. Using this prediction model, albeit very basic, the SSE is 115512239042 (usually called SST).

Shoot Blind

The best guess (i.e. the one with minimal SSE) is the average $\hat{y} = \bar{y}$ for any new observation. Using this prediction model, albeit very basic, the SSE is 115512239042 (usually called SST).

Now imagine you had a predictor... let's say some measure of work ethic or intelligence, call it x. Here's a plot:



(see in R)

Shoot Less Blind

Now, the "new" SSE is 34198974577.

Shoot Less Blind

Now, the “new” SSE is 34198974577. This is a reduction in SSE by 81313264465 or as a percentage basis 81313264465 / 115512239042 = 70.39%. Seen this before?

Shoot Less Blind

Now, the “new” SSE is 34198974577. This is a reduction in SSE by 81313264465 or as a percentage basis 81313264465 / 115512239042 = 70.39%. Seen this before?

$$R^2 := \frac{SSE_0 - SSE}{SSE_0}$$

Why is this called “percentage of variance explained”?

Shoot Less Blind

Now, the "new" SSE is 34198974577. This is a reduction in SSE by 81313264465 or as a percentage basis 81313264465 / 115512239042 = 70.39%. Seen this before?

$$R^2 := \frac{SSE_0 - SSE}{SSE_0}$$

Why is this called "percentage of variance explained"?

$$\begin{aligned}
R^2 &:= \frac{SSE_0 - SSE}{SSE_0} \times \frac{n-1}{n-1} = \frac{\frac{1}{n-1} SSE_0 - \frac{1}{n-1} SSE}{\frac{1}{n-1} SSE_0} \\
&= \frac{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2 - \frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \bar{y})^2} \\
&= \frac{s_y^2 - s_e^2}{s_y^2} \approx \frac{\mathbb{V}\mathrm{ar}[Y] - \mathbb{V}\mathrm{ar}[E]}{\mathbb{V}\mathrm{ar}[Y]}
\end{aligned}$$

where $\mathbb{V}\mathrm{ar}[Y]$ is what was inexplicable before and $\mathbb{V}\mathrm{ar}[E]$ is what is inexplicable after. R^2 is really an *estimate* of the pctg var. explained.

Limitations of R^2

R^2 is the great equalizer among all models and all responses. They are immediately comparable. But R^2 doesn't mean much for my specific problem.

Limitations of R^2

R^2 is the great equalizer among all models and all responses. They are immediately comparable. But R^2 doesn't mean much for my specific problem. I'm now predicting income, $\hat{y} = \$70,000$ and $R^2 = 70\%$. How does knowing $R^2 = 70\%$ (i.e. the model fit is pretty good) tell me how good my $\hat{y}$ is?

Limitations of R^2

R^2 is the great equalizer among all models and all responses. They are immediately comparable. But R^2 doesn't mean much for my specific problem. I'm now predicting income, $\hat{y} = \$70,000$ and $R^2 = 70\%$. How does knowing $R^2 = 70\%$ (i.e. the model fit is pretty good) tell me how good my $\hat{y}$ is? How big is e in $y = \hat{y} + e$ where y is the true response for this prediciton. It doesn't...

How about the following metric?

Limitations of R^2

R^2 is the great equalizer among all models and all responses. They are immediately comparable. But R^2 doesn't mean much for my specific problem. I'm now predicting income, $\hat{y} = \$70,000$ and $R^2 = 70\%$. How does knowing $R^2 = 70\%$ (i.e. the model fit is pretty good) tell me how good my $\hat{y}$ is? How big is e in $y = \hat{y} + e$ where y is the true response for this prediciton. It doesn't...

How about the following metric?

$$s_e = \sqrt{s_e^2} = \sqrt{\frac{1}{n-1} SSE} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}$$

This is our best guess of the standard error of our estimate e, our residual (AKA “RMSE”). What is a standard error?

Recall the Empirical Rule

If $X \sim \mathcal{N}(\mu, \sigma^2)$, then

- $\mu \pm 1\sigma$ contains 68% of the realization values
- $\mu \pm 2\sigma$ contains 95% of the realization values
- $\mu \pm 3\sigma$ contains 99.7% of the realization values

Recall the Empirical Rule

If $X \sim \mathcal{N}\left(\mu,\, \sigma^2\right)$, then

- $\mu \pm 1\sigma$ contains 68% of the realization values
- $\mu \pm 2\sigma$ contains 95% of the realization values
- $\mu \pm 3\sigma$ contains 99.7% of the realization values

Thus for a realization x,

- $x \pm 1\sigma$ contains μ 68% of the time
- $x \pm 2\sigma$ contains μ 95% of the time
- $x \pm 3\sigma$ contains μ 99.7% of the time

Stretch the Empirical Rule

If σ is unknown then,

- $x \pm 1s$ contains μ about 68% of the time
- $x \pm 2s$ contains μ about 95% of the time
- $x \pm 3s$ contains μ about 99.7% of the time

Stretch the Empirical Rule

If σ is unknown then,

- $x \pm 1s$ contains μ about 68% of the time
- $x \pm 2s$ contains μ about 95% of the time
- $x \pm 3s$ contains μ about 99.7% of the time

and if the distribution of X is non-normal but not too funky,

- $x \pm 1s$ contains μ about about 68% of the time
- $x \pm 2s$ contains μ about about 95% of the time
- $x \pm 3s$ contains μ about about 99.7% of the time

Why is RMSE useful?

In our case, the r.v. is $Y \mid X_1, \ldots, X_p$ which is centered at $\mu = \mathbb{E}\left[Y \mid x_1, \ldots, x_p\right]$ and generally speaking, non-normal. $\hat{y} \approx \mu$ and $s_e \approx \mathbb{S}\mathrm{E}\left[Y \mid X\right]$. Thus,

Why is RMSE useful?

In our case, the r.v. is $Y \mid X_1, \ldots, X_p$ which is centered at $\mu = \mathbb{E}[Y \mid x_1, \ldots, x_p]$ and generally speaking, non-normal. $\hat{y} \approx \mu$ and $s_e \approx \mathbb{S}\mathrm{E}[Y \mid X]$. Thus,

- $\hat{y} \pm 1s_e$ contains 68% of the response values for a specific $x_1, \ldots, x_p$
- $\hat{y} \pm 2s_e$ contains 95% of the response values for a specific $x_1, \ldots, x_p$
- $\hat{y} \pm 3s_e$ contains 99.7% of the response values for a specific $x_1, \ldots, x_p$

Thus RMSE gives you an approximate means of assessing how variable the real response y could be give your predicted response $\hat{y}$.

All Three are Equivalent

Minimizing SSE, maximizing R^2 and minimizing s_e all give equivalent fits.

$$
\begin{aligned}
& SSE, \\
R^2 &:= \frac{SSE_0 - SSE}{SSE_0} \quad \text{and} \\
s_e &= \sqrt{s_e^2} = \sqrt{\frac{1}{n-1} SSE} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2}
\end{aligned}
$$

Inference

We haven't spoken about t tests. Why is that?

Inference

We haven't spoken about t tests. Why is that?

In order to have inference, we need to make explicit random variable model assumptions

$$Y \sim g(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p, \sigma^2, \ldots)$$

must be assumed to be something like

$$Y \sim \mathcal{N}\left(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,\ \sigma^2\right)$$

(we will explore next time)

R^2 vs. F test

In this case R^2 will be related to F, the omnibus test statistic for whether the model has any signal whatsoever.

$$R^2 = \frac{SSE_0 - SSE}{SSE_0} = \ldots = 1 - \left(1 + F\frac{p-1}{n-p}\right)^{-1}$$

$$\begin{aligned} F &= \frac{\frac{SSE_0 - SSE}{p-1}}{\frac{SSE}{n-p}} = \frac{SSE_0 - SSE}{SSE}\frac{n-p}{p-1} = \ldots \\ &= \underbrace{\frac{R^2}{1-R^2}}_{\substack{\text{ratio of variance} \\ \text{explained to} \\ \text{unexplained}}} \quad \underbrace{\frac{n-p}{p-1}}_{\substack{\text{penalty for} \\ \text{too many features}}} \end{aligned}$$