Module 2 - Data exploration

STAT 401

Section 1: Thinking about data

What are data? Every detail in the universe that is measured (or observed) and recorded is data. For example:

- The length, in meters, of every public swimming pool in Alaska
- The length, in meters, of a single swimming pool
- The length, in bologna sandwiches, of a single swimming pool
- The fact that there is a certain swimming pool in a certain location

- The type of bologna sandwich preferred by every public swimming pool custodian in Alaska
- What happens to a bologna sandwich when dropped in every public swimming pool in Alaska
- The full transcript of a 180-minute interview with each public swimming pool custodian in Alaska involving their ruminations on soggy bologna sandwiches

Note that data, to be data, do *not* need to satisfy any of the following:

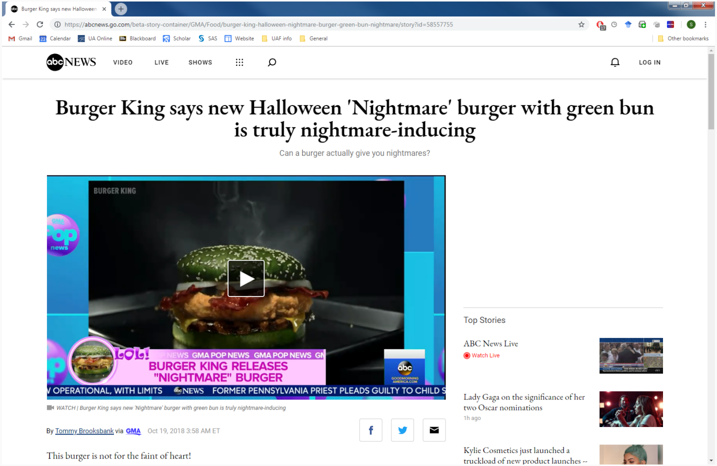
- Be measured or observed accurately
- Be measured or observed precisely
- Be recorded correctly
- Be recorded precisely
- Be measured, observed, or recorded consistently
- Be generalizable

- Be understood, in context, by anybody involved
- Be thoughtfully collected
- Include all the important variables
- Be ethical
- Be useful
- Stay any of these things permanently

Three good guidelines for working with data:

- Don’t worship Data
- Distrust all data (at least a little)
- Remember that not all data are created equal

A case study


Burger King says new Halloween
https://abcnews.go.com/beta-story-container/GMA/Food/burger-king-halloween-nightmare-burger-green-bun-nightmare/story?id=58557755
Gmail
Calendar
UA Online
Blackboard
Scholar
SAS
Website
UAF info
General
Other bookmarks
abcNEWS
VIDEO
LIVE
SHOWS
LOG IN
Burger King says new Halloween 'Nightmare' burger with green bun is truly nightmare-inducing
Can a burger actually give you nightmares?
BURGER KING
LOL!
GMA POP NEWS
BURGER KING RELEASES "NIGHTMARE" BURGER
OPERATIONAL, WITH LIMITS
FORMER PENNSYLVANIA PRIEST PLEADS GUILTY TO CHILD S
WATCH | Burger King says new 'Nightmare' burger with green bun is truly nightmare-inducing
By Tommy Brooksbank via GMA
Oct 19, 2018 3:58 AM ET
This burger is not for the faint of heart!
Top Stories
ABC News Live
Watch Live
Lady Gaga on the significance of her two Oscar nominations
1h ago
Kylie Cosmetics just launched a truckload of new product launches --

Burger King, which partnered with Paramount Trials and Florida Sleep & Neuro Diagnostic Services, Inc. on a study, claims that its new burger is nightmare-inducing. One hundred people ate the burger over the 10-night study. Burger King declined to provide a copy of the study to "GMA" but claims the data shows that participants reported that their nightmares increased 3.5 times. (ABC News)

What could possibly go wrong?

- Nocebo effect
- Experimenter's bias
- Measurement error
- Non-representative population
- Conflicts of interest

One additional guideline:

- Data without context is just noise (i.e. worthless)
- In short, respect the context

As an example, consider measurements of product quality in an industrial process

Dye pressure	13.1	12.1	11.0	14.3	15.8	14.7
Dye rotation	4	17	103	189	250	289
Edge deficiency	0	0	1	1	1	0

How does context help in interpreting the values of dye rotation? (Hint: which are further apart: 17 and 103, or 4 and 289?)

Understanding a data set's context necessarily includes answering the following:

- Is there measurement error in the observations?
- How were observations recruited into the sample?
- What population is the sample representative of?
- What is desired to be learned from the data?

Section 2: Observation and experimentation

Observational data	Experimental data
Researchers record or measure variables without interfering	Researchers manipulate variables and measure a response
Can be used to demonstrate association	Can be used to demonstrate causation
Predictors are usually regarded as random	Predictors are usually regarded as fixed

Observational data:

Pros	Cons
Data may be convenient or cheap to collect	Causation cannot be established due to lurking variables
May be more ethical to collect	Data sets might not correspond directly to the research questions
	Data might not have been fully or accurately recorded

Observational data:

Pros	Cons
	Predictors might not vary much, obscuring their effect
	Predictors might be correlated
	Generalization to larger population may be limited

Experimental data:

Pros	Cons
Causation can be established via control	Inconvenient or costly or time consuming
Predictors can be varied artificially and independently	May be impossible or unethical
Generalization may be more feasible	

Research problem: A study is conducted on feedlots' effects on home values in two MN counties in 1993 and 94

Observational study: Researchers pull data on property sales price, property characteristics, and feedlot characteristics from public real estate records

Experimental study: Researchers purchase a set of identical mobile homes and randomly locate them either near to or far from feedlots. Homes are then market listed and sales prices are recorded.

Section 3: Understanding data

Some terminology:

Categorical variable: a variable that takes a values from a set of categories

Quantitative variable: a variable that takes numerical values

Explanatory variables: (aka predictors, regressors, independent variables) variables which are used to explain variability in the response variable

Response variable: (aka dependent variable) a variable of interest which we would like to model

Conventionally, data sets are organized in a rectangular array. Rows represent distinct observations. Columns represent distinct variables.

```r
head(Wool)

##   len amp load cycles
## 1 250   8   40    674
## 2 250   8   45    370
## 3 250   8   50    292
## 4 250   9   40    338
## 5 250   9   45    266
## 6 250   9   50    210
```

Statistical analyses of data are often intended to produce claims about population parameters. These analyses are called *inferential*. Statistical inference is a formal framework for extracting scientific insights from data.

Analyses may also be conducted in order to simply understand a data set's structure and content better. These analyses are called *exploratory*. Exploratory data analysis (EDA) often produces information that is useful for conducting an inferential data analysis (IDA). Hence, it is almost always prudent to perform EDA prior to IDA.

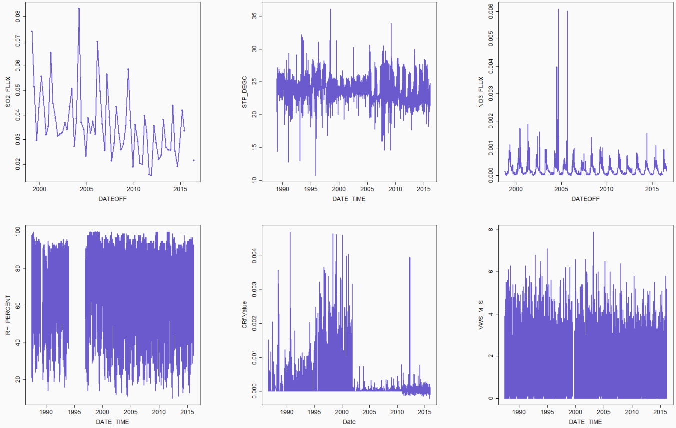
EDA is an iterative cycle for understanding your data. It involves:

1. Generating questions about your data
2. Searching for answers by visualizing, transforming, and modeling your data
3. Using what you learn to refine your questions

Some useful questions to ask during EDA include:

- Are there missing values? Is there any censoring? (e.g. minimum detectable limits)
- Is there measurement error?
- Are there outliers? Are any data values implausible?
- What degree of variation occurs within the variables? What degree of covariation occurs between the variables?
- How are variables distributed across their possible values?

For example, a scientist collected 30 years of atmospheric chemistry data at a single observation station in ≈ 50 variables.

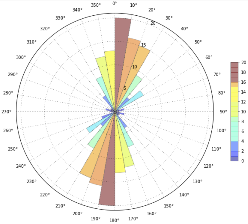


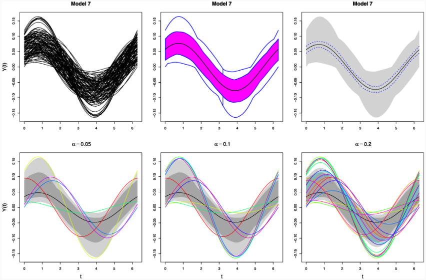
A geologist uses uranium-lead dating to estimate the ages of grains of sand from a certain layer of rock.

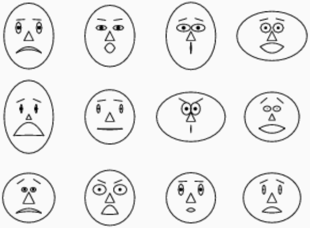
```
data <- read.csv("~\\Consulting\\Elisabeth Nadin\\DZMix B.csv")
summary(data$Mean.age)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##    1.68   97.13  366.70  779.50 1432.00 2905.00       5
```

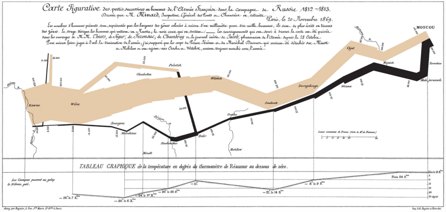
Section 4: Visualizing data

EDA can include strategies for visualizing data. Many ways have been developed.

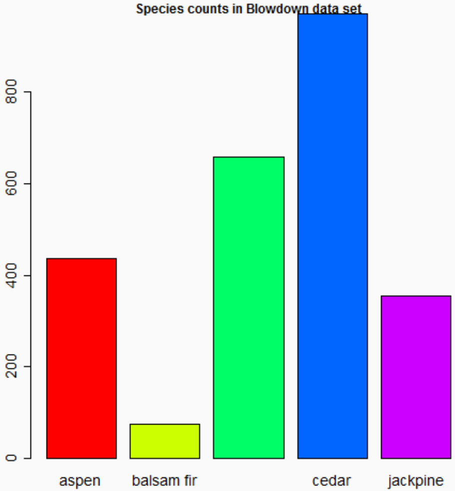




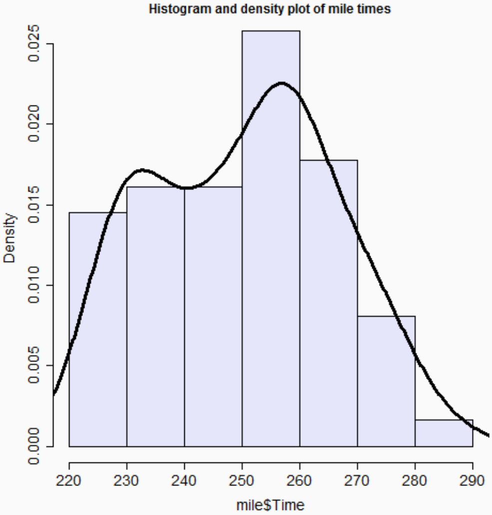




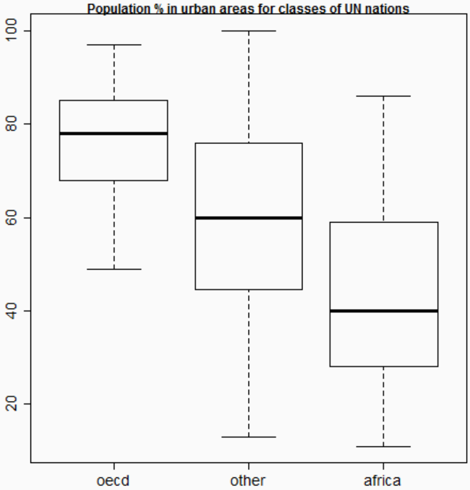
A categorical variable in isolation is easily visualized with a barplot



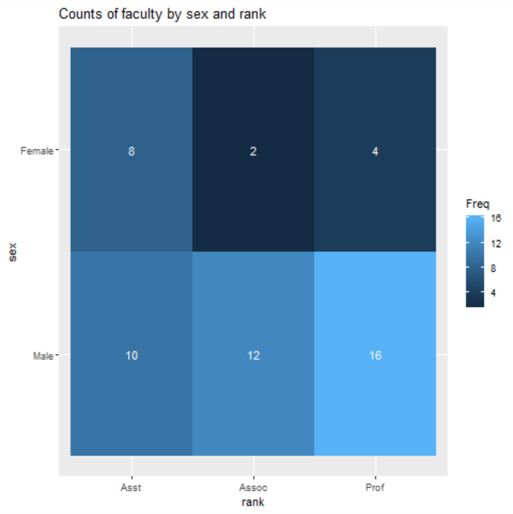
A quantitative variable in isolation is often visualized with a histogram or a density plot



A quantitative variable cross-classified by a categorical variable is often visualized with a boxplot



A categorical variable cross-classified by a categorical variable is often visualized with a heatmap



Section 5: Scatter plots

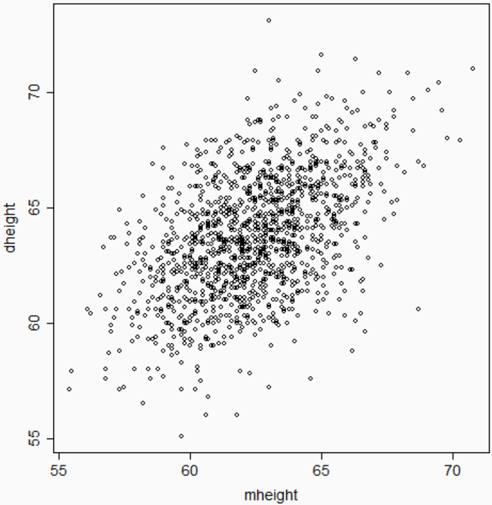
Scatter plots: plots of points on a two-dimensional plane where each dimension represents some variable. They are used to illustrate association.

To make a scatter plot, plot (x_i, y_i) for $i = 1, \ldots, n$

For example, consider the `Heights` data set in the `alr4` library:

```
head(Heights)
```

```
##   mheight dheight
## 1    59.7    55.1
## 2    58.2    56.5
## 3    60.6    56.0
## 4    60.7    56.8
## 5    61.8    56.0
## 6    55.5    57.9
```

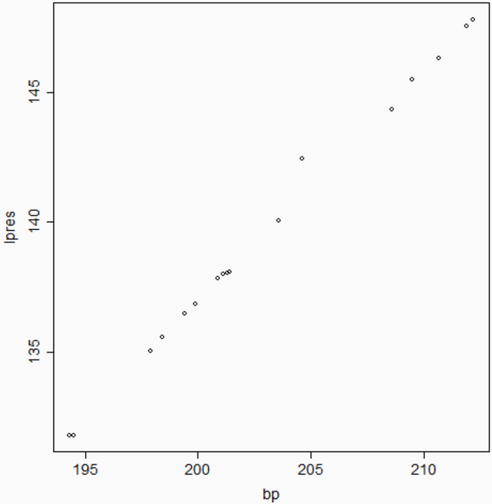

dheight
70
65
60
55
55
60
65
70
mheight

Another example: the `Forbes` data set:

```
head(Forbes)
```

```
##       bp  pres  lpres
## 1 194.5 20.79 131.79
## 2 194.3 20.79 131.79
## 3 197.9 22.40 135.02
## 4 198.4 22.67 135.55
## 5 199.4 23.15 136.46
## 6 199.9 23.35 136.83
```

Scatter plots are built with the “response vs. explanatory variable”



A scatter plot helps us see features such as

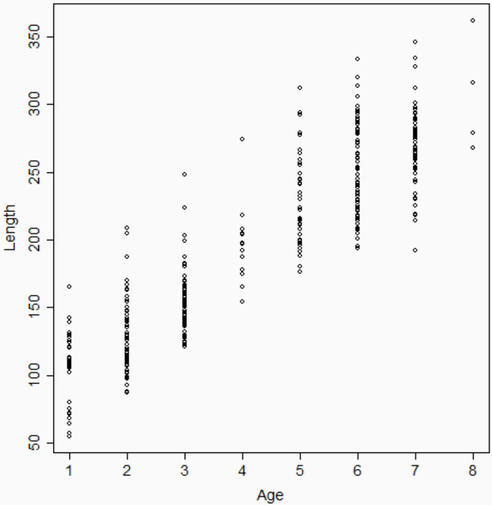
- functional relationship
- direction of association
- strength of association
- outliers
- leverage points

Strength and direction of association were also measured by r_{xy}.

For example: the `wblake` data set:

```
head(wblake)
```

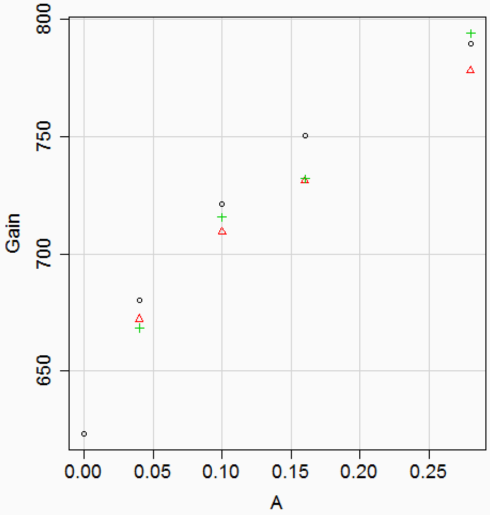
```
##   Age Length   Scale
## 1   1     71 1.90606
## 2   1     64 1.87707
## 3   1     57 1.09736
## 4   1     68 1.33108
## 5   1     72 1.59283
## 6   1     80 1.91602
```



Length
Age
50
100
150
200
250
300
350
1
2
3
4
5
6
7
8

Another example: the `turkey` data set:

```
head(turkey)

##         A  Gain S  m        SD
## 1 0.00 623.0 1 10 19.459359
## 2 0.04 680.2 1  5  7.190271
## 3 0.10 721.4 1  5 21.454603
## 4 0.16 750.4 1  5 17.487138
## 5 0.28 789.4 1  5 14.673105
## 6 0.04 672.2 2  5 26.508489
```



800
750
700
650
Gain
0.00
0.05
0.10
0.15
0.20
0.25
A

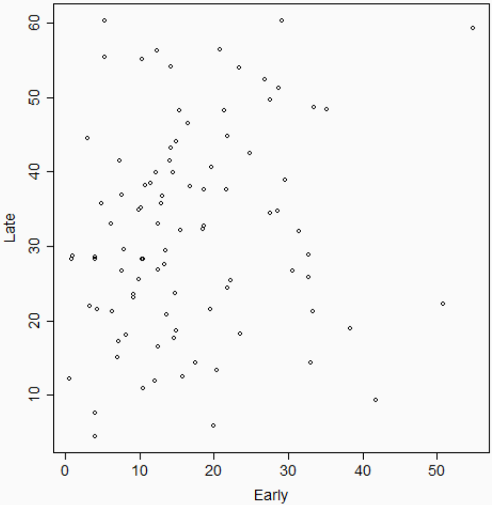
It is important to remember that just because we see a linear relationship between a response and some predictor(s), there's no implication that a change in the predictor causes a change in the response. This is the classical distinction between correlation and causation.

For example:

X	Y
umbrellas in use on campus	precipitation fallen that day
winter outdoor temperature	airborne smoke particulates
antibiotic use	missing days at school/work

Another example: the `ftcollinssnow` data set:

```
head(ftcollinssnow)

##    YR1 Early Late
## 1 1900   3.0 44.5
## 2 1901  15.8 12.5
## 3 1902  13.1 36.8
## 4 1903   4.0  7.6
## 5 1904   1.0 28.7
## 6 1905  10.5 28.3
```

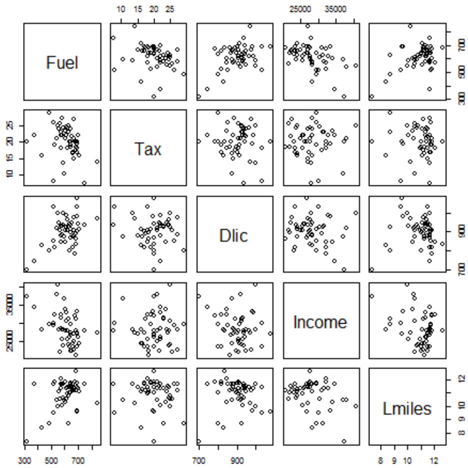


Late
Early
60
50
40
30
20
10
0
10
20
30
40
50

Scatter plot matrices can be used to graph two-way scatter plots between every pair of variables in a data frame

```r
head(fuel2001)
```

```
##      Drivers    FuelC Income  Miles      MPC      Pop  Tax     Fuel
## AL   3559897  2382507  23471  94440 12737.00  3451586 18.0 690.2644
## AK    472211   235400  30064  13628  7639.16   457728  8.0 514.2792
## AZ   3550367  2428430  25578  55245  9411.55  3907526 18.0 621.4751
## AR   1961883  1358174  22257  98132 11268.40  2072622 21.7 655.2927
## CA  21623793 14691753  32275 168771  8923.89 25599275 18.0 573.9129
## CO   3287922  2048664  32949  85854  9722.73  3322455 22.0 616.6115
##          Dlic    Lmiles
## AL 1031.3801 11.455720
## AK 1031.6411  9.519882
## AZ  908.5972 10.919533
## AR  946.5706 11.494069
## CA  844.7033 12.036298
## CO  989.6062 11.360403
```


Fuel
Tax
Dlic
Income
Lmiles
10 15 20 25
25000 35000
300 500 700
700 900
8 9 10 12

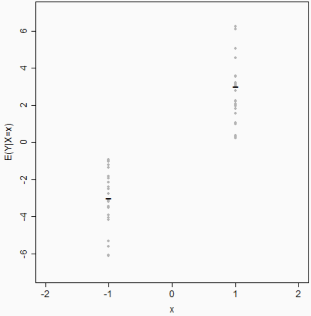
Section 6: Linear models

The expectation of a random variable Y was notated as $E(Y)$. It could be that, for some random variable Y, $E(Y) = -3$ or $E(Y) = 3$ or $E(Y) = a$ for some constant a.

It's also possible that the expected value of some random variable Y depends on the value of some other random variable X. For instance, if X can only be -1 or 1, then there could be a R.V. Y such that

$E(Y|X = -1) = -3$ and

$E(Y|X = 1) = 3$



This relationship between X and Y can be rewritten as

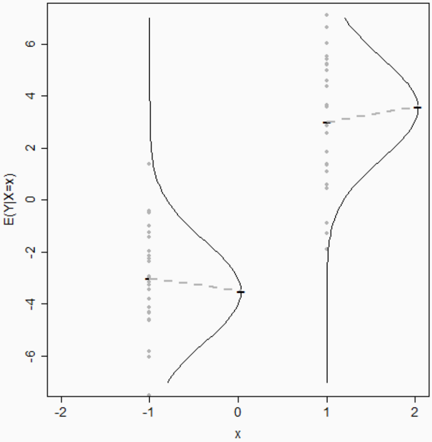
$$E(Y|X = x) = 3x \text{ for } x = -1 \text{ or } x = 1$$

Conditional expectation is a powerful idea. It allows the mean of the distribution of Y (given X) to shift based on the value of x, while leaving the distribution's shape and variance alone.

For instance, if

$$(Y|X = x) \sim N(3x,\ 2)$$

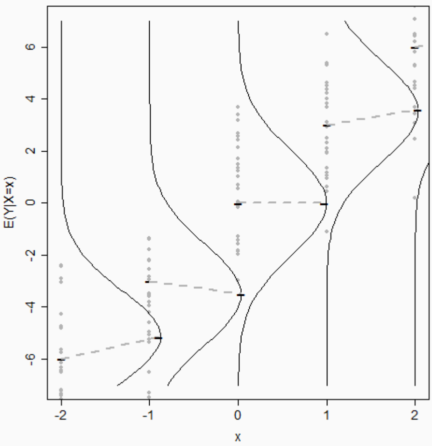
for $x = -1$ or 1, Y has one of two normal distributions, depending on whether $X = -1$ or $X = 1$.



If we allow X to be *any* integer and maintain the requirement that

$$E(Y|X = x) = 3x$$

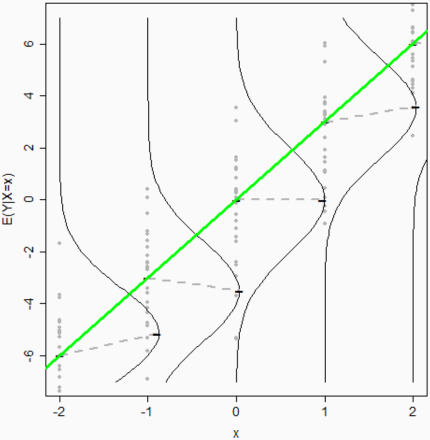
then there are infinite distributions for Y given X: all of them normal.



The means of these distributions are all given by

$$E(Y|X = x) = 3x$$

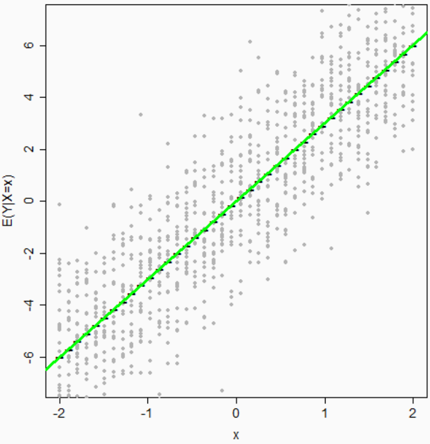
which is the equation of a line with slope 3 and intercept 0.



Now let X be allowed to take any value on the real line. If

$$E(Y|X = x) = 3x$$

then all means of Y fall on the green line.



We call the assumed relationship

$$E(Y|X = x) = 3x$$

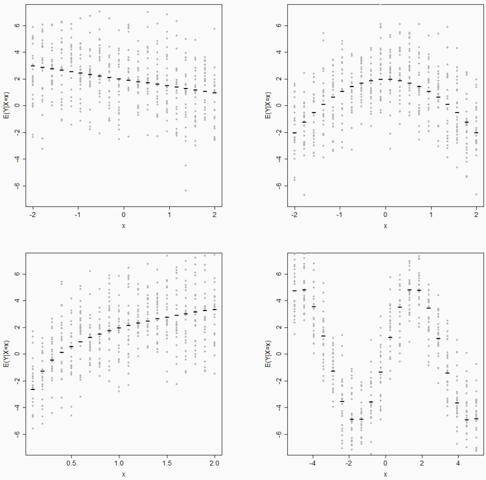
a *model for the conditional mean function of* Y, or simply the *mean function of* Y. In general, for random variables X and Y, the mean function of Y is commonly assumed to be $E(Y|X = x) = \beta_0 + \beta_1 x$ for some constants β_0 and β_1.

β_0 and β_1 are parameters because they are fixed, unknown quantities that could be exactly calculated if we had access to the entire population of X and Y. β_0 plays the role of an intercept and β_1 plays the role of a slope.

The model for the mean function of Y

$$E(Y|X=x) = \beta_0 + \beta_1 x$$

is a *linear model*. Other models are also possible via other mean functions.



Model: a mathematical function that approximates a relationship between two or more variables. We do not need to believe the model captures the right relationship; it only needs to mimic reality to a satisfying degree.

In addition to the assumption on the mean function:

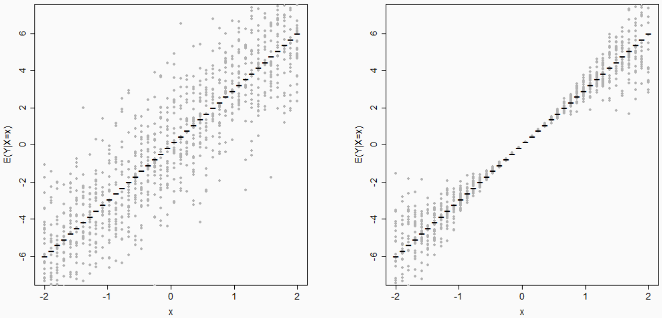
$$E(Y|X = x) = \beta_0 + \beta_1 x$$

many linear models make an additional assumption on the variance of Y given X. Although any mathematical relationship is possible, the simplest possible assumption is that

$$V(Y|X = x) = \sigma^2$$

and it is adequate for many data sets.

Compare data from the original scenario where the variance assumption is satisfied versus data where it is not.



In practice, the violation of the variance assumption can look like the trend seen in the BigMac data set.