Concepts in modeling and machine learning (ML)

Data Science for Biologists, Fall 2021
Dr. Spielman

First, a joke I shamelessly found online

- The optimist says, The glass is half full. The pessimist says, The glass is half empty.

- Microsoft Excel says, The glass is January 2nd.

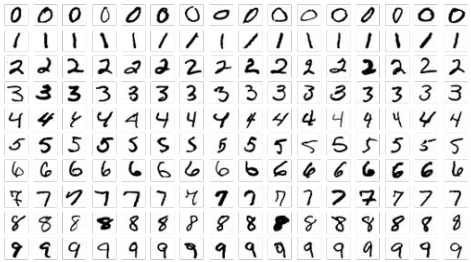
Let's just bash on Excel a little more

- **"How Excel may have caused the loss of 16,000 COVID test results in England"**
  - https://www.theguardian.com/politics/2020/oct/05/how-excel-may-have-caused-loss-of-16000-covid-tests-in-england
- **"Gene name errors are widespread in the scientific literature"**
  - https://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1044-7
- **"Scientists rename human genes to stop Microsoft Excel from misreading them as dates"**
  - https://www.theverge.com/2020/8/6/21355674/human-genes-rename-microsoft-excel-misreading-dates
- There is a WILD "undo" bug I have just learned about.

Caveats: This is not a statistics class

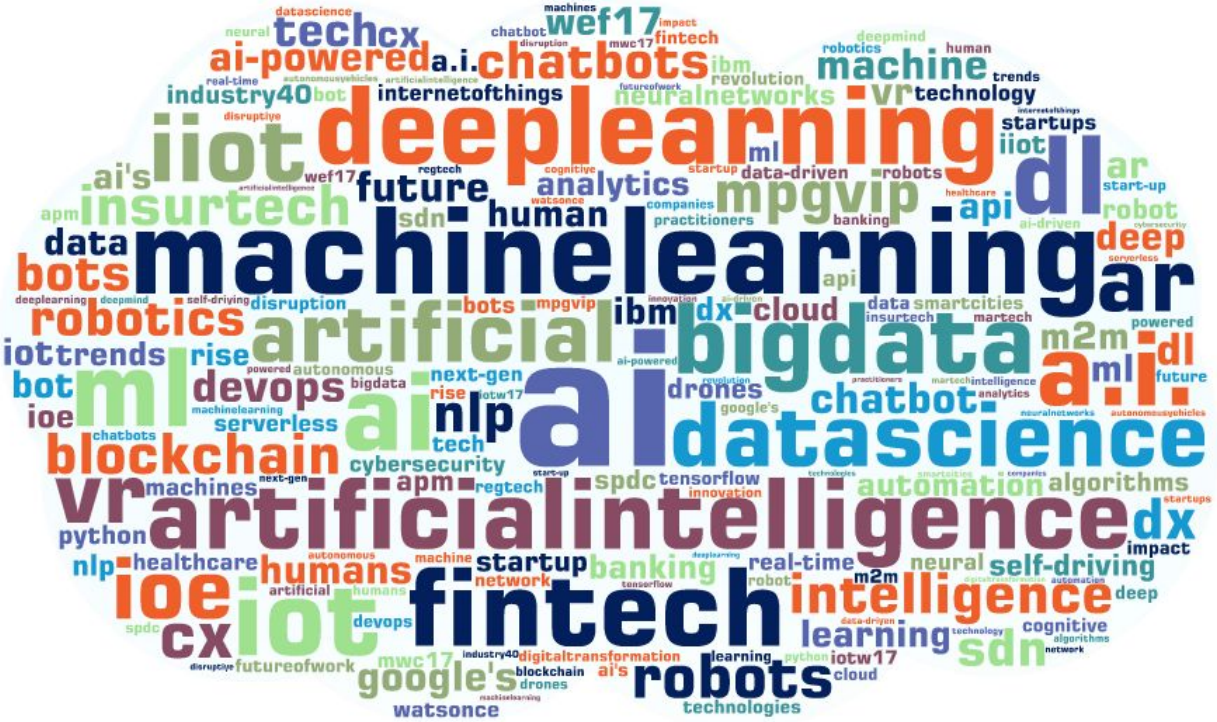
According to wikipedia...

- **Machine learning**
  - "Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data."
  - **Deep learning** is a specific set of "black box" methods in machine learning that use something called *artificial neural networks*

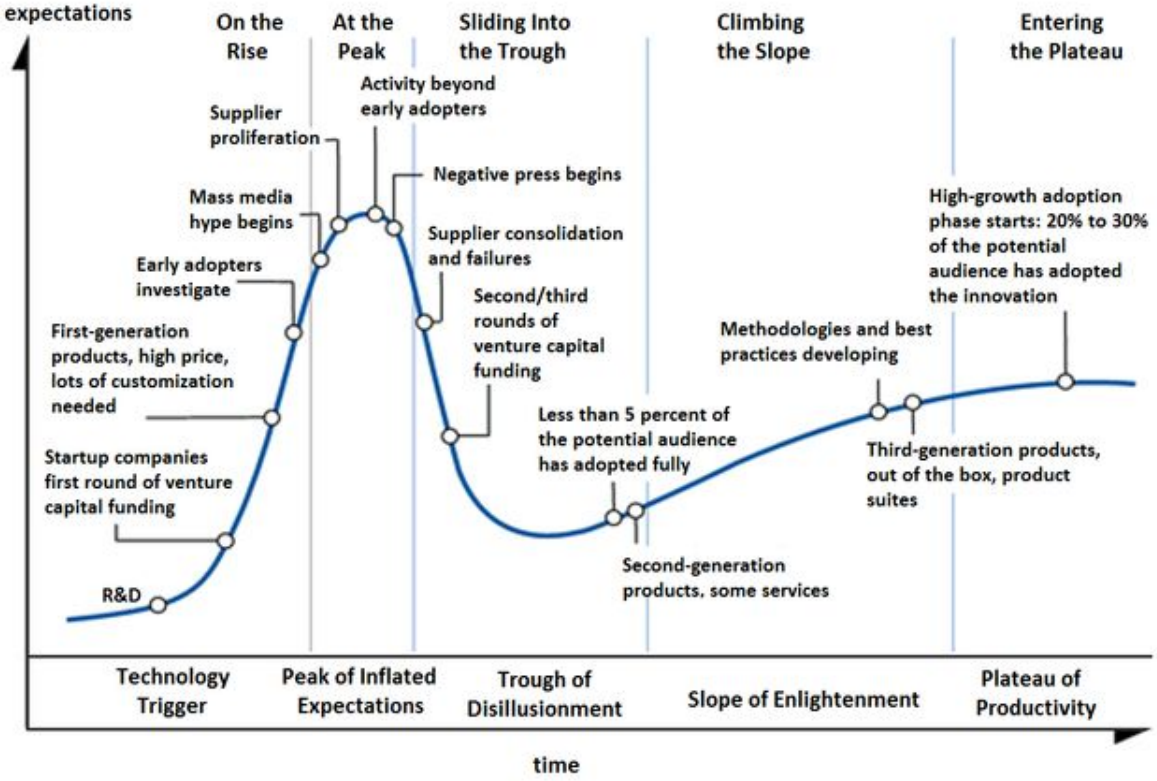


- **Artificial intelligence**
  - "Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to natural intelligence displayed by animals including humans."

Attack of the buzzwords



Some bullshit: The Gartner Hype Cycle



According to Spielman...

- **Machine learning**
  - ~~"Machine learning (ML) is the study of computer algorithms that can improve automatically through experience and by the use of data."~~
  - Machine learning occurs when you give a machine (computer) some data, and it uses an algorithm (code) to learn something from that data (build a model aka fancy mathematical formula)

- **Artificial intelligence**
  - ~~"Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to natural intelligence displayed by animals including humans."~~
  - AI is the WILD WEST OF BUZZWORDS!!!!!!!!!!!!!!!!!! People just want to use this word, and it never means *The Matrix* (IMO, AI = deep learning = neutral nets)


YOU KEEP USING THAT WORD
I DO NOT THINK IT MEANS WHAT YOU THINK IT MEANS


Adam Kucharski
@AdamJKucharski
Got a new fridge. It's amazing what AI can do these days.
6.Artificial intelligence
When the function is activated,the temperature of the refrigerator is automatically switched to
5°C ,the temperature of freezer is automatically switched to -18°C
2:30 PM · Sep 15, 2021 · Twitter Web App

AI is not NOT the hype it's made out to be

- **"Stop Calling Everything AI, Machine-Learning Pioneer Says"**
  - https://spectrum.ieee.org/stop-calling-everything-ai-machinelearning-pioneer-says
  - “People are getting confused about the meaning of AI in discussions of technology trends—that there is some kind of intelligent thought in computers that is responsible for the progress and which is competing with humans," he says. **“We don't have that, but people are talking as if we do."**

AI is not NOT the hype it's made out to be

- **"AI has a long way to go before doctors can trust it with your life"**
  - https://qz.com/2016153/ai-promised-to-revolutionize-radiology-but-so-far-its-failing/
  - "...when we collect data from Stanford Hospital, then we train and test on data from the same hospital, indeed, we can publish papers showing [the algorithms] are comparable to human radiologists in spotting certain conditions. **It turns out [that when] you take that same model, that same AI system, to an older hospital down the street, with an older machine, and the technician uses a slightly different imaging protocol, that data drifts to cause the performance of AI system to degrade significantly.** In contrast, any human radiologist can walk down the street to the older hospital and do just fine. So even though at a moment in time, on a specific data set, we can show this works, the clinical reality is that these models still need a lot of work to reach production.…All of AI, not just healthcare, has a proof-of-concept-to-production gap."

AI is not NOT the hype it's made out to be

- **"Deep neural networks for genomic prediction do not estimate marker effects"**
  - https://pubmed.ncbi.nlm.nih.gov/34596363/
  - "It has been suggested that the ability of powerful nonlinear models, such as deep neural networks, to capture complex epistatic effects between markers offers advantages for genomic prediction. **However, these methods tend not to outperform classical linear methods,** leaving it an open question why this capacity to model nonlinear effects does not seem to result in better predictive capability."

But there are some success stories!

- **"iRobot’s newest Roomba uses AI to avoid dog poop"**
  - https://www.theverge.com/2021/9/9/22660467/irobot-roomba-ai-dog-poop-avoidance-j7-specs-price
  - *He says iRobot has been working on the problem for years, even creating a huge database of fake pet mess to train their AI vision systems. “Robotics is supposed to be glamorous, but I don’t know how many Play-Doh models of poo we created,” says Angle. “Many, many thousands.”*

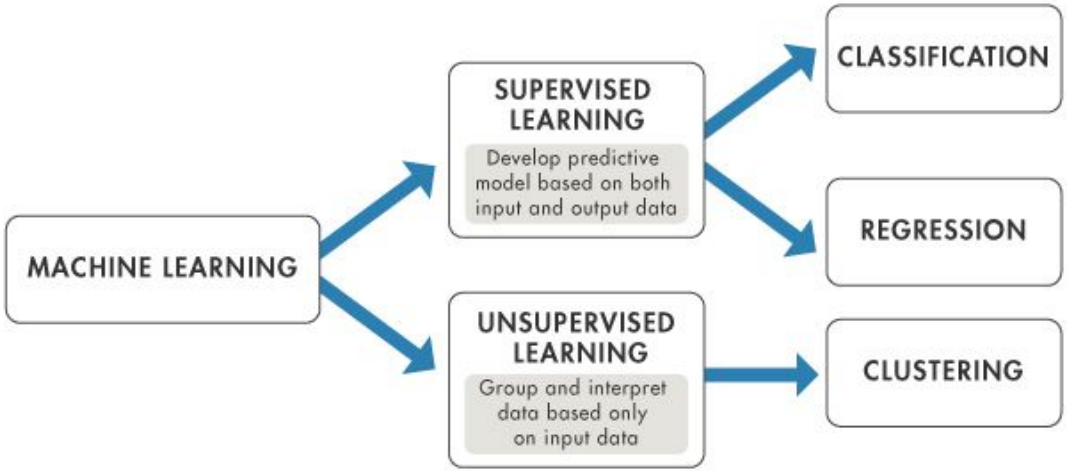
  - *The company began, he said, by buying "all the realistic gag poop you can buy on the internet," then branched out into making hundreds of Play-doh poop models, which it painted brown and photographed in different lighting and from different angles. He thinks every iRobot employee with a pet has had that animal's waste photographed from multiple angles.* https://www.cnn.com/2021/09/09/tech/roomba-ai-avoids-dog-poop/index.html

Methods in machine learning

Supervised learning: modeling data whose structure is "known"

Unsupervised learning: discovery novel patterns or structures to the data

There are *tons* of different methods/algorithms for each end-point shown here.



Supervised:
I know already which ones are dogs and which are muffins. Can the computer learn about the differences?

Unsupervised:
Are there distinct groups of things here?
Can I figure out those groups and predict something in the future based on which group is which?

(Bear in mind, computers have NO IDEA that "dogs" and "muffins" belong in different categories.)

Someone really did this

https://blog.cloudsight.ai/chihuahua-or-muffin-1bdf02ec1680


chocolate cookie


fawn smooth Chihuahua


baked blueberry muffin


white chihuahua


fawn smooth Chihuahua


blueberry muffin


fawn smooth Chihuahua


muffin


[unknown]


brown coated Chihuahua


baked muffin


beige short coated puppy


tan smooth Chihuahua puppy

blueberry cupcakes


three smooth Chihuahua puppies


white and black muffin

Here's some more for fun

https://www.boredpanda.com/dog-food-comparison-bagel-muffin-lookalike-teenybiscuit-karen-zack/

Another example...

Imagine I know the weight and time slept for 83 mammals!!

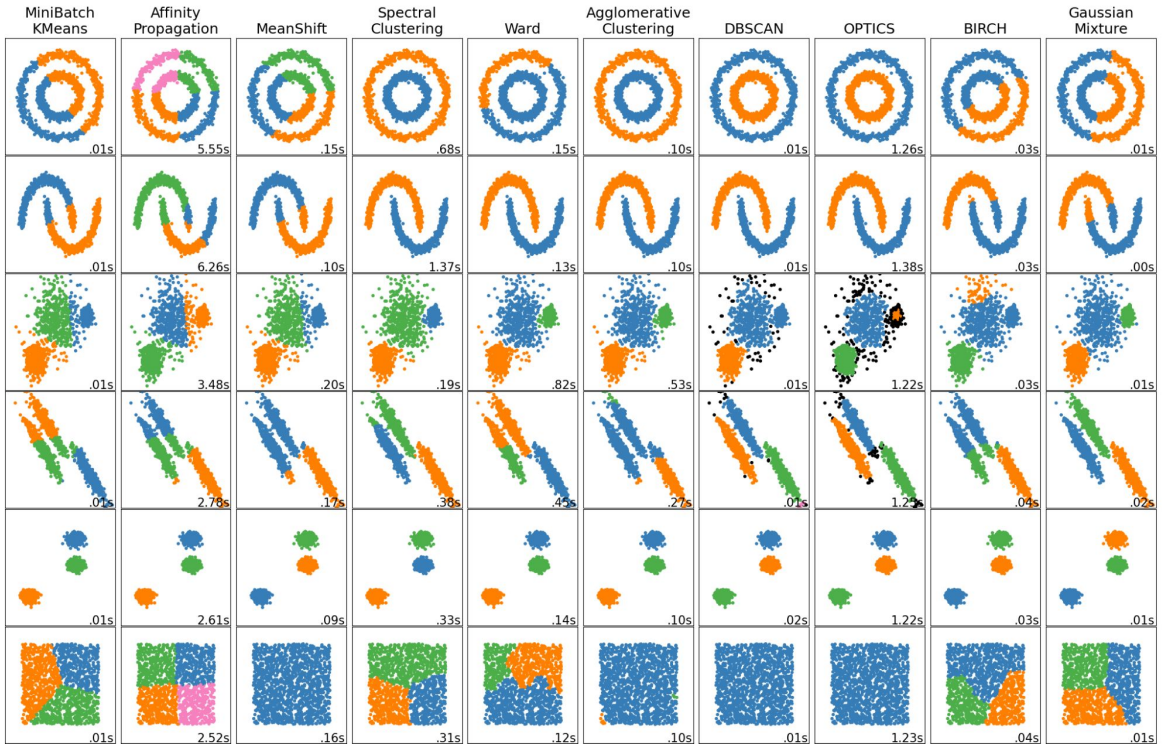
- **Supervised scenario**:
  - *I also know what these mammals eat.* I can use information about weight/sleep to explore relationship with vore and build a model!
  - If that model is good, maybe later I can use that model to predict what other uncharacterized mammals eat, if I know their weight/sleep.

- **Unsupervised scenario:**
  - *I have no idea what they eat.* Can I use the information I have to get the computer to build a model that will figure out natural groupings in the data that maybe could be informative about what they eat?

Unsupervised learning is really hard.

Each column is a different unsupervised method trying to discover unknown groupings

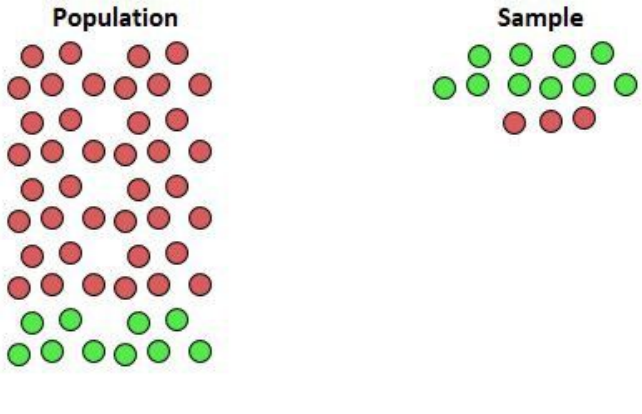
Each row is a dataset plotted on X/Y axes

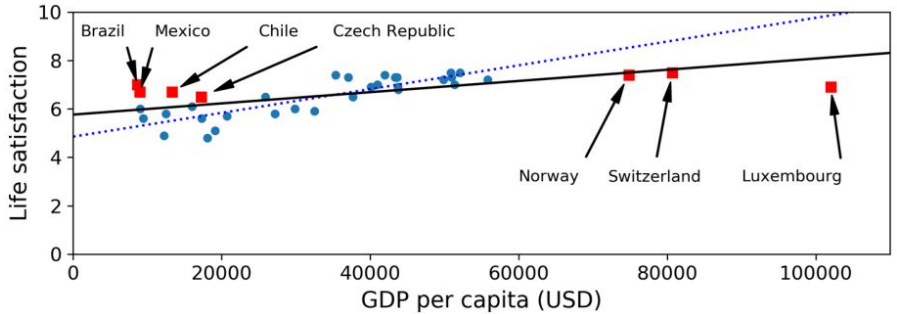


A model is only as good as the data used to build (train) it

Good data to [build, train, optimize, fit] your model

- Dataset should be HUGE
- Dataset should be DIVERSE and REPRESENTATIVE of real-world application





Machine learning "industry" approach

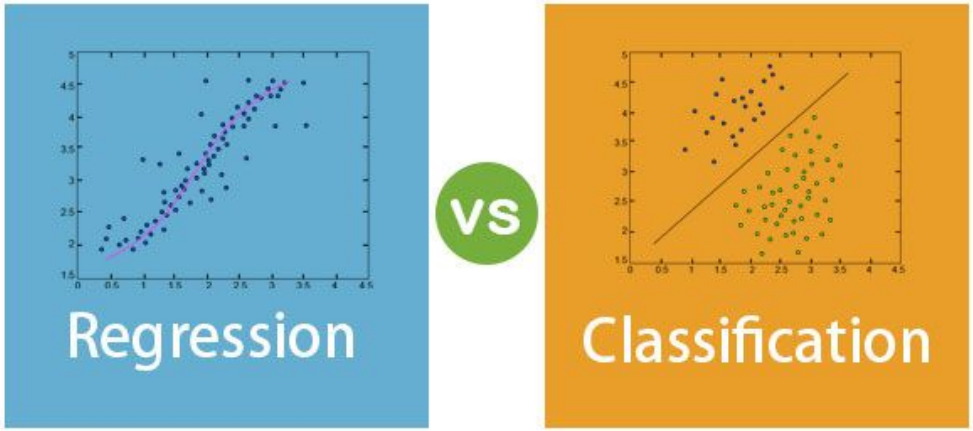
- **Step 1: Train (build/fit/optimize) your model on a huge, diverse, representative dataset.**
  - Resulting model is basically a mathematical formula.

- **Step 2: Test the model on different data to *validate* its performance.**
  - Does it work well on this new *testing data?*
  - Or, does it only work well on the *training data*?

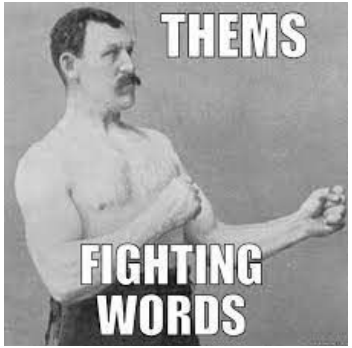
We'll learn some "supervised learning" approaches

Generalized linear models (GLMs): A more generalized framework for performing a bunch of different types of regressions

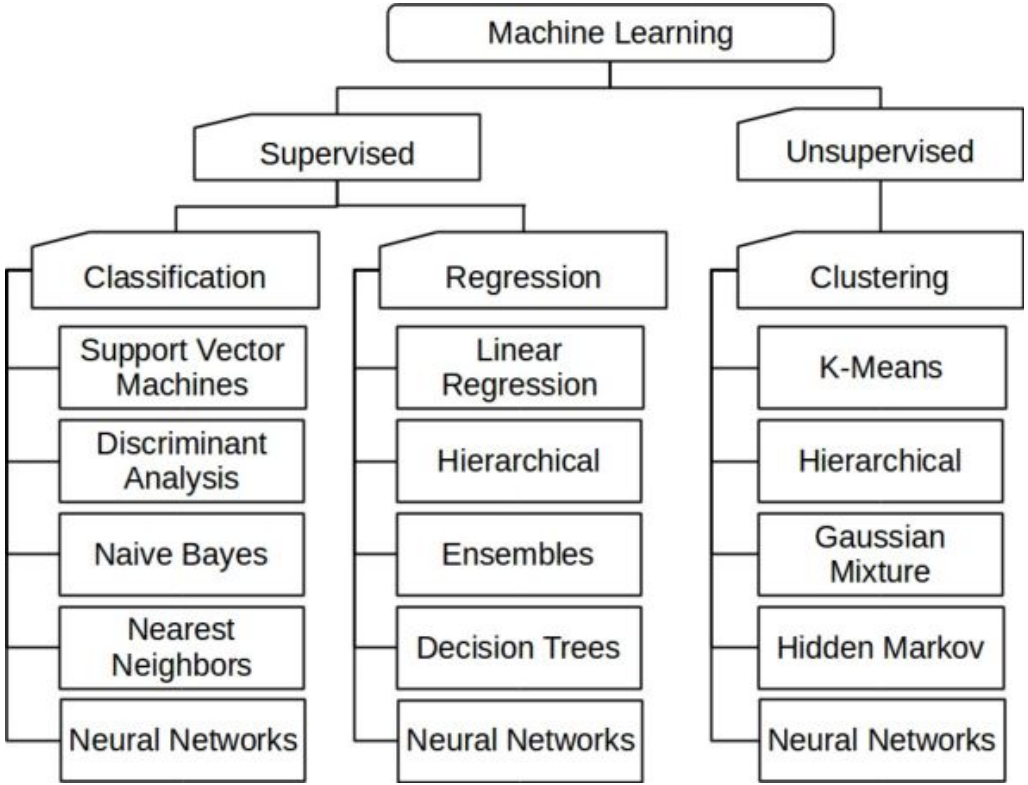
General linear models (*regressions*) model a *numeric response variable*



Logistic regressions model a *binary response variable*



We are only just barely grazing the surface, as is this flowchart!

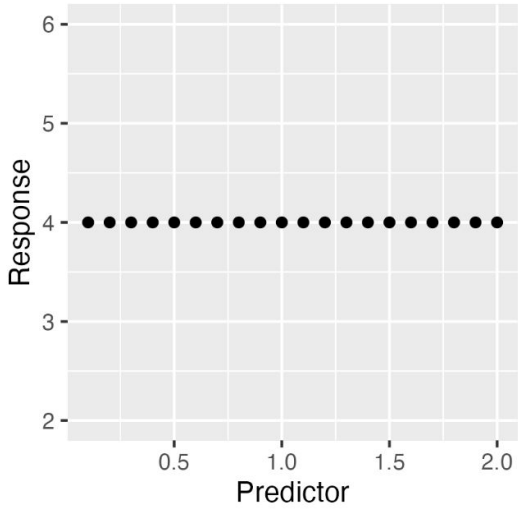


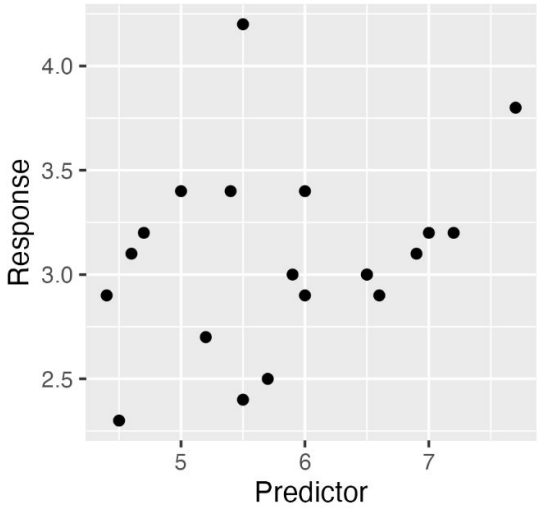
Regression and friends model *numeric responses*

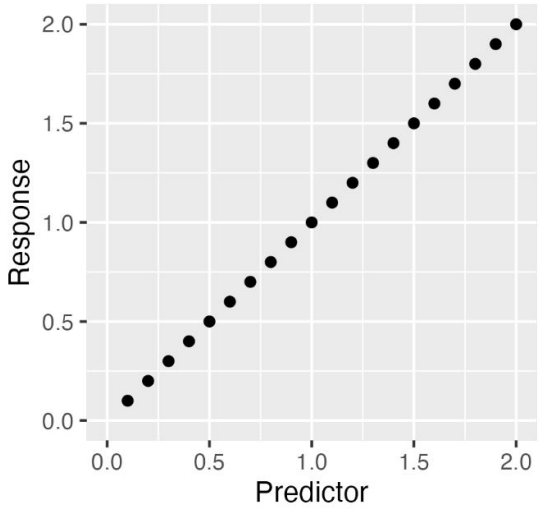
- Simple regression
  - Single numeric predictor
- ANOVA
  - Single categorical predictor
- ANCOVA
  - A numeric and a categorical predictor

General linear model: *WHO CARES* what the predictors are. The math is all the same! What's with these names?!!?

The goal of linear regression is to explain variation in a numeric response variable







Regressions calculate lines of best fit.

$$Y = \beta_1 X_1 + \beta_0 + \epsilon$$

$$Y = mX + b$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 \ldots + \beta_n X_n + \epsilon$$

"Experimental" approach to modeling/regression

- I designed an experiment to assess influence of oral vitamin C supplementation on tooth length (in mm) in guinea pigs
- I took 100 guinea pigs, half male and half female, and randomly placed them into two groups:
  - Half receive regular food
  - Half receive regular food with a vitamin C supplement
- I also measured *confounding influences* like their sex and age
  - Confounding influence = anything that might influence results but isn't our variable "of interest"
- **Analyze with a regression: Does Vit C supplementation affect tooth length, when accounting for confounding influences of age and sex?**
  - **I specifically want to know the specific relationship between Vit C and tooth growth.**

"Exploratory" approach to modeling/regression

- I wonder what kinds of things are related to guinea pig tooth growth?
- I'm gonna collect a BUNCH OF DATA on guinea pigs, including their tooth lengths and moooore!
- **Analyze with a regression: What is the best* way to model tooth length in guinea pigs so I can accurately predict guinea pig tooth lengths?**
    - ***What is "best"?**

"Experimental" vs "exploratory" approaches

- Both are building a model to explore tooth length, but...

- The "experimentalist" (scientist, researcher, academic, etc.) has specific hypotheses they want to test about the tooth length in the context of their setup.

- The "explorer" (industry data scientist or "big data" scientists [also in academic research and science!]) wants know as much as they can about tooth length in guinea pigs and maybe predict new tooth lengths from new guinea pigs

We will live in "exploratory" world in this class

- Want to go into data science? You need *all of it and then some.* Fun fact: data science is really just a rebrand of statistics. Sorry.
  - You also need to learn Python and pay *really close attention* after Thanksgiving.

- All models have certain *assumptions*, aka conditions that must be met in order for the fitted model to be *reliable*. We're going to assume assumptions are reasonably met all around.
  - **But I must say:** IT IS NOT TRUE THAT THE DATA HAS TO BE NORMALLY DISTRIBUTED FOR LINEAR REGRESSION. I REPEAT: IT IS NOT TRUE.

Let's head to R!

We'll go back and forth to some of the other slides for additional context.

P-values: The biggest hoax in science.

- P-values are one of the most notoriously misunderstood concepts. They tell you:
  - **Assuming the null hypothesis is true, what is the probability of observing my data?**
  -
- They do NOT tell you:
  - What is the probability that this result I observe is real?
  - Is the null hypothesis wrong?
  - Is the null hypothesis right?
  - Is the alternative hypothesis wrong?
  - Is the alternative hypothesis right?

Null hypotheses are set in stone

- Each statistical test you do relies on a highly specific null hypothesis that is *always associated with that statistical test.* There is 0 creativity or wiggle-room.

- In linear models, the null hypotheses are:
  - All coefficients (betas) = 0
  - $R^2 = 0$

- Each estimated parameter ("model quantity") has an associated P-value

Statistical significance is mental gymnastics

If a P-value is very very small (usually P<.05), we say....

Gee! That's a small probability! I don't think it's likely that things with low probabilities happen, so maybe actually something else besides the null is going on. **We call this significant.**

If a P-value is not very small (usually P>=.05), we say...

Gee! I think that probabilities that are not very small could totally come to pass. It's not unreasonable to maybe observe this data under the null. **We call this not significant.**

But 0.05 is not a special number

It's an historical accident based on a single thing RA Fisher wrote:
https://www.bmj.com/rapid-response/2011/11/03/origin-5-p-value-threshold

"...If one in twenty does not seem high enough odds, we may, if we prefer it, draw the line at one in fifty or one in a hundred. Personally, the writer prefers to set a low standard of significance at the 5 percent point, and ignore entirely all results which fails to reach this level. A scientific fact should be regarded as experimentally established only if a properly designed experiment rarely fails to give this level of significance..."

Metrics for evaluating performance of linear regressions

- **R^2 : The percent of variation in the response that is explained by the model.**
  - $R^2 = 0$ → The model explains *nothing* about the response.
  - $R^2 = 1$ → The model explains *everything* about the response.
  - $0 < R^2 < 1$ → The model explains *some amount of what there is to know* about the response

- **RMSE (Root-Mean Squared Error): The spread of error in the model**
  - The standard deviation of model residuals in units of response
  - Model residuals represent *error*, aka what does the model NOT explain?
  - **High RMSE = high error spread**. Points are generally more spread out from the line-of-best-fit
  - **Low RMSE = low error spread**. Points are generally closer to the line-of-best-fit

Model validation/evaluation

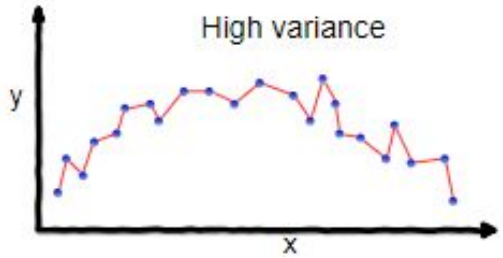
- How good is the model at explaining variation *in data it does NOT know about?*
- Should we even bother using our model to predict future outcomes?

- For example, if I built my model from only Maine Coon cats, it might perform really well on other Maine coons cats, but terribly on Siamese cats.

- If I built my model only with data from cats that weigh < 4 kg, it might perform really poorly on cats that weight more than 4 kg.

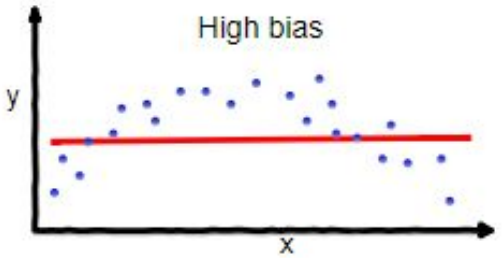
The "bias/variance" trade-off

Model is "overly tailored" to training data. It will probably not be good at predicting from new data.



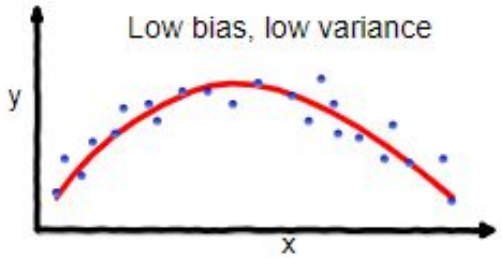
overfitting

Model doesn't even capture the complexity of the training data! It probably won't well in any circumstance.



underfitting

The Goldilocks Zone of modeling



Good balance

How do we avoid problems from overfitting/underfitting?

- Use validation/evaluation procedures to determine model performance on new data

- When building the model, choose our predictors so they only contribute information and not as much *noise* to the outcome
  - Not possibly in "experimental" scenarios, but possible in "exploratory" scenarios.

Model validation procedures

1. Train (build) model on some data
2. *Test* (evaluate! validate!) model on *other data that wasn't used during training* whose outcomes are KNOWN, and ask how good those predictions are.
    a. Good prediction = low error in predictions. The model systematically "gets it right"

- Problem: If we use all our data for training, we have none left for testing!! What do?? **Split the data up into testing and training groups!**

- There are *many ways* to do this. We'll see one: a simple testing/training split.
    - See here for more: https://towardsdatascience.com/validating-your-machine-learning-model-25b4c8643fb7

Training and testing splits

1. *Randomly* split your data into two groups:
    - **Training data**, usually 60-80% of the data is used to build aka train your model
    - **Testing data**, the rest, is used to evaluate the model trained on your training data
2. Build your model with the training data
3. Ask how well the model performs on testing data

Pop quiz: Do you think models generally perform better on training or testing data?