Week 6: Statistical inferential goals

ANTH 674: Research Design & Analysis in Anthropology

Professor Andrew Du

Andrew.Du2@colostate.edu

Statistical vignette

- The parable of the Baltimore stockbroker
- Sends out email each week, predicting an increase/decrease for a given stock

Week 1 Email

Week 2 Email

- Gets 10 weeks in a row right
- Asks you to invest money w/ him (w/ commission)

What's going on behind the scenes

Week 1: Emails 10,240 people

Week 2: Emails 5,120 people

Week 3: Emails 2,560 people

⋮

Week 10: Ten people received ten straight weeks of correct picks

Should you invest or not?

Lecture outline

- Different types of statistical inferential goals:

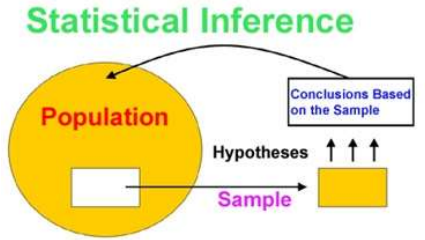
1. Exploratory
2. Confirmatory AKA hypothesis testing
   - Problem of multiple comparisons
3. Prediction
   - Cross-validation w/ *independent* data

Quick review of statistical inference

5

What is statistical inference?

- To understand properties of some larger statistical population by analyzing a smaller sample from said population



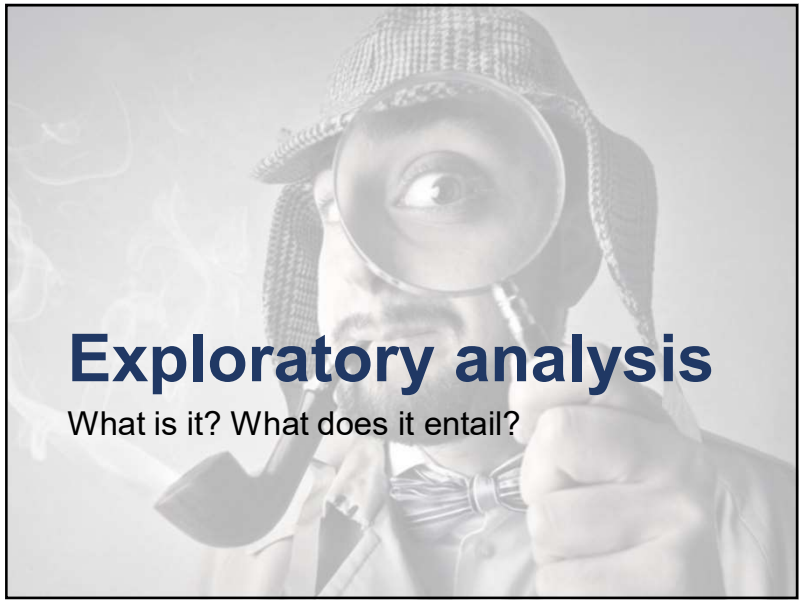
6

Three modes of inference

1. Exploratory
2. Confirmatory or hypothesis testing
3. Prediction

- Each has different, mutually exclusive goals
- Knowing which one is right for your question makes data analysis more straightforward!
- Part of translating research question into statistical question

7



Exploratory analysis

What is it? What does it entail?

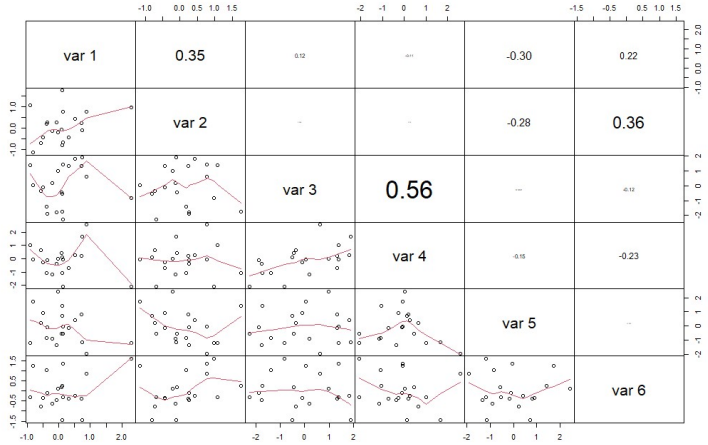
8

What is exploratory analysis?

- Analyzing data where patterns and relationships are ***unknown***
- Thus, there is no *a priori* hypothesis to test!
- E.g., why do people pick their nose?
  - Might collect data on a bunch of IVs & DVs and see if there are any relationships

9

E.g., a scatter plot matrix



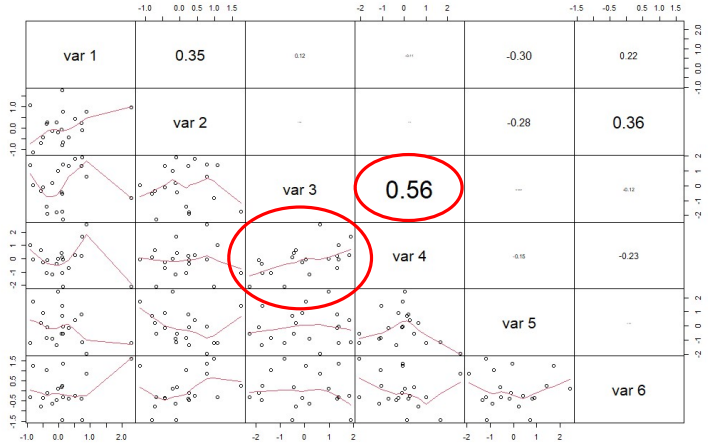
10

Necessary part of research!

- Important for a new field, or where we don't know a lot about the variables
- Can lead to the generation of hypotheses after the fact (*a posteriori*)

11

E.g., a scatter plot matrix



12

Necessary part of research!

- Important for a new field, or where we don't know a lot about the variables
- Can lead to the generation of hypotheses after the fact (*a posteriori*)
- "Finding the question is often more important than finding the answer."

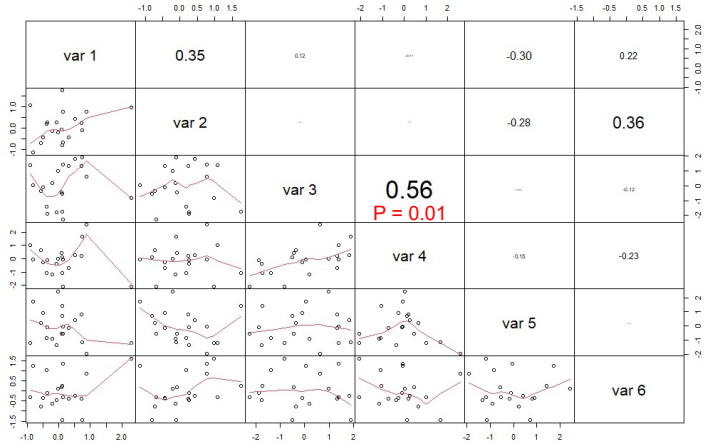
CANNOT calculate P-values!

John W. Tukey

13

Why no P-values?

All rnorm(20)



14

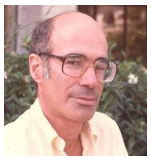
Why no P-values?

- Type I error: rejecting H_0 when H_0 is true
- Type I error rate = significance level (α) = 0.05
- On average, 5% of tests will be P < 0.05 just by chance when H_0 is true
- On average, 5% of tests will also have large statistics just by chance when H_0 is true

	H_0 True	H_0 False
Reject H_0	Type I Error	Correct Rejection
Fail to Reject H_0	Correct Decision	Type II Error

15

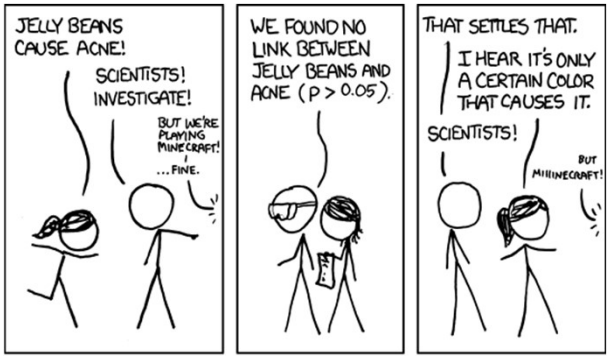
cf. Freedman's paradox



- Even if no relationships exist between variables, can look at many relationships until you get P < 0.05 (or a large statistic)
- But these significant relationships are ***not*** real (false positive)! Just got lucky (or in actuality, unlucky)
- Known pejoratively as "P-hacking", "P-fishing", "data dredging", and more

16

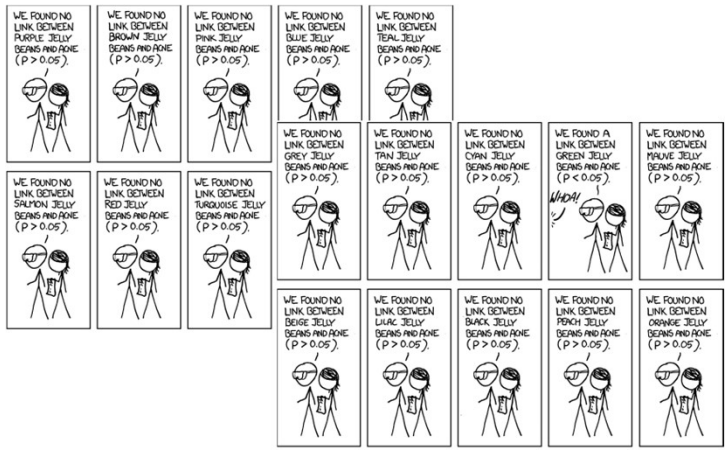
P-hacking



https://www.explainxkcd.com/wiki/index.php/882:_Significant

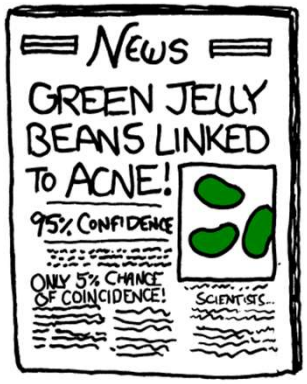
17

P-hacking



18

P-hacking



19

P-hacking: another way

- Different subsets of a variable can count as different samples from population; same with collecting more data for a variable
- E.g., “looked at body mass of all primates, but let’s now look at great apes only”
- Part of “researcher degrees of freedom” or “garden of forking paths”

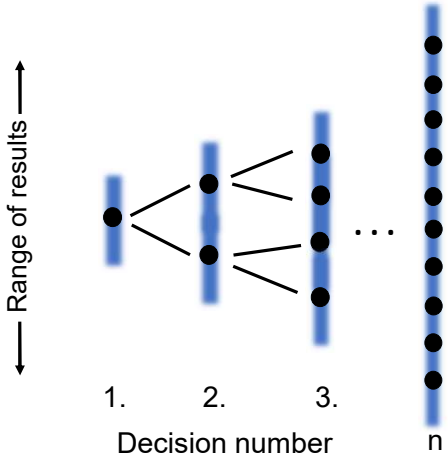
"If you don't reveal some insights soon, I'm going to be forced to slice, dice, and drill!"

20

Garden of forking paths

- Each (subconscious) decision in data analysis represents a "fork in the road"
  1. Choosing among IVs and DVs
  2. Collect more data or exclude data
  3. Running different tests
  4. Not reporting certain tests (file-drawer effect)
  5. Reviewers saying you should run a test a different way or more tests
  6. And many more

Garden of forking paths



- On average, 5% will lead to false positives if H_0 is true
- More forks you explore, more likely to get false positives if H_0 is true

An example

False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant

2011

Joseph P. Simmons[1], Leif D. Nelson[2], and Uri Simonsohn[1]
[1]The Wharton School, University of Pennsylvania, and [2]Haas School of Business, University of California, Berkeley

- "To help illustrate the problem, we conducted two experiments designed to demonstrate something false: that certain songs can change listeners' age. *Everything reported here actually happened*." (italics mine)

An example

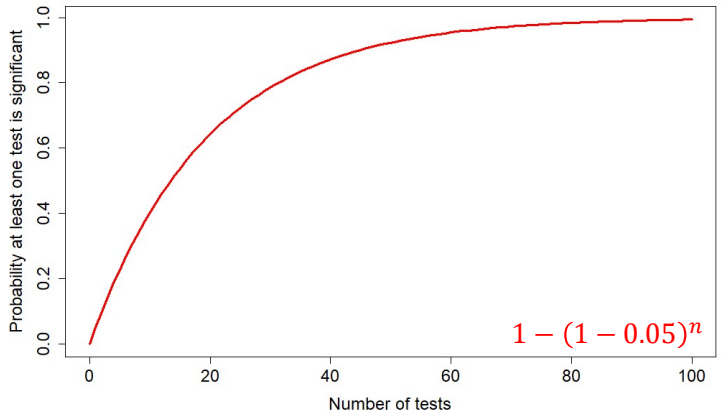
- Are subjects younger after listening to a song?

"Kalimba" by Mr. Scruff (control)

"When I'm 64" by The Beatles

- Simulated experiment, data collection, choosing different variables, collecting more data, doing different analyses, etc. (lots of forks!)
- Found subjects were 1.5 years younger after listening to "When I'm Sixty-Four" compared to the control, "Kalimba" ($P = 0.04$)

Assuming H_0 is true…



25

Summary: exploratory

- Used when nothing is known about data, and there are no *a priori* hypotheses to test
- Explore what data look like and relationships between variables (i.e., fish to your heart's content! It's okay & even necessary!)
- ***BUT DO NOT CALCULATE P-VALUES!***
- Can generate hypotheses, but emphasize they are *a posteriori*
- Need to be tested/confirmed with independent dataset

26

Questions?

27

Confirmatory analysis AKA hypothesis testing

What is it? What does it entail?

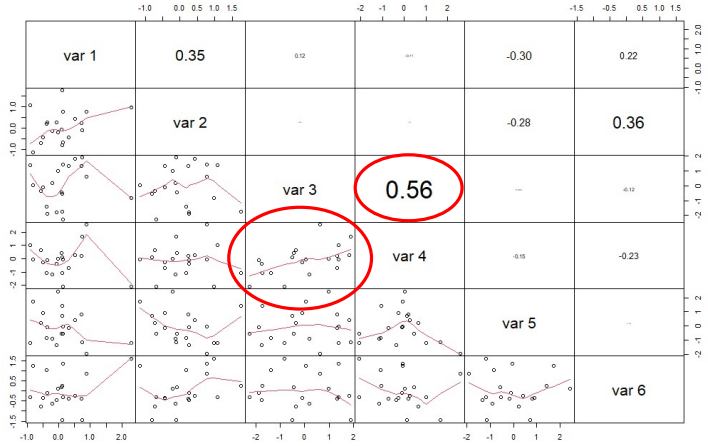
28

What is confirmatory analysis?

- Testing an *a priori* hypothesis with data using confidence intervals or P-values
- Only time you should calculate P-values!
- Hypotheses can come from intuition, theory, or previous exploratory analyses
- For the third, need an ***independent*** dataset

29

Our scatter plot matrix



30

Testing the hypothesis

- Collect more data for var3 and var4
- E.g., if var3 is time spent nose-picking & var4 is age, collect data from another group of people
- Conduct a test & calculate P-value
- If $P < 0.05$, H_0 is ***now*** falsified (though more replication/confirmation is always good!)
- If $P > 0.05$, perhaps original exploratory correlation was spurious, or need more testing (e.g., if sample size or effect size too small)

31

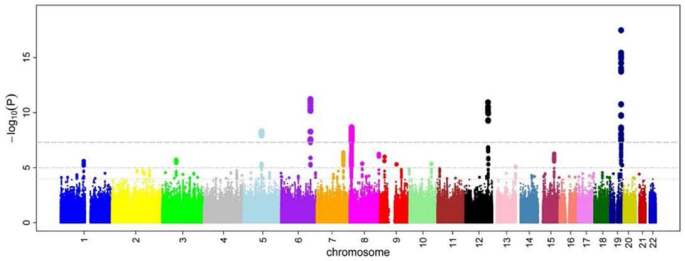
Separate exploratory and confirmatory!

- It is ***unethical*** to explore data, calculate P-values, and then present the significant results as *a priori* hypothesis tests (i.e., P-hacking)
- Only get to test one hypothesis for one question for one dataset
- ***BIG*** reason why so many results in science are not replicable
- If you do wind up subsetting/collecting more data, doing more tests, etc., should be transparent about it

32

Multiple comparisons

- But what if research question demands multiple tests and computed P-values?
- Common in genetics (e.g., genome-wide association studies)

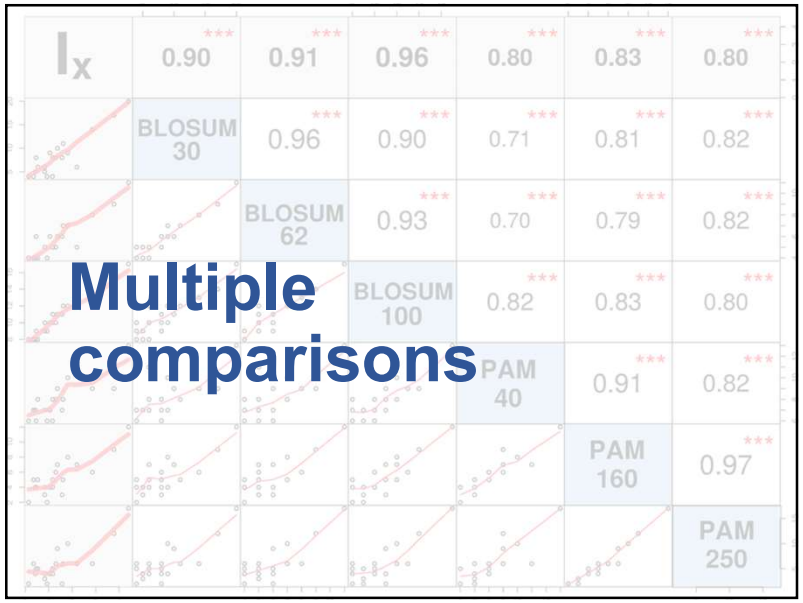


33

Questions?

34

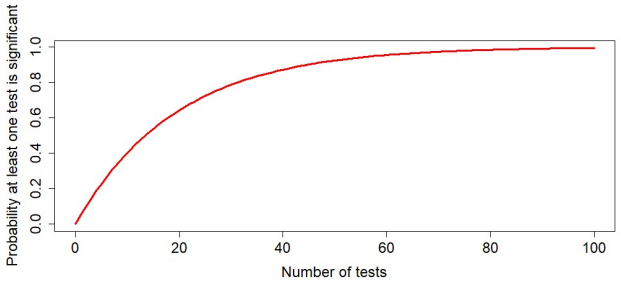
Multiple comparisons



35

Multiple comparisons

- If H_0 is true, more tests means more likely to get at least one false positive
- Thus, need to correct P-values if you calculate a lot of them



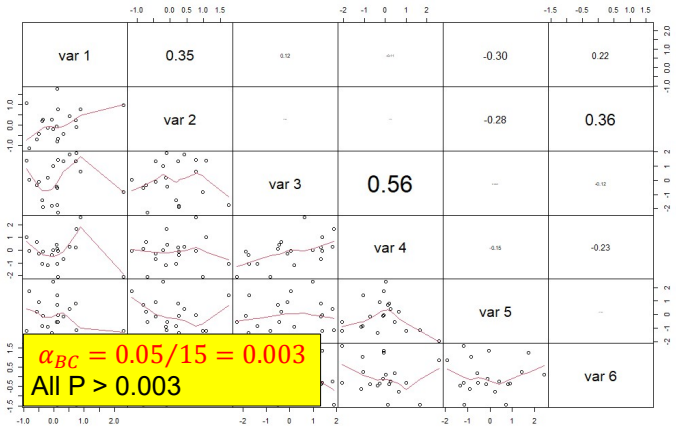
36

1. Bonferroni correction

- Adjusts the ***family-wise error rate*** (FWER): the probability of ≥ one Type I error in your tests
- Creates a new significance level, $\alpha_{BC} = \alpha/k$, where α is the original significance level (0.05) and k is the number of tests
- E.g., with 20 tests, the probability of at least one false positive is 0.64 at $\alpha = 0.05$
- With 20 tests and $\alpha_{BC} = 0.05/20 = 0.0025$, probability of at least one false positive is 0.05

37

1. Bonferroni correction



38

1. Bonferroni: issues

- Should k be the # of tests you publish, the total # tests you ran, total # of tests in the journal?
- Assumes H_0 is true for **ALL** tests, so it is overly conservative (OK if this is your goal)
- That is, it decreases Type I error but at the expense of increasing Type II error
- Are you really that far off base that **NONE** of your H_0 are false in reality?!
- A better solution is the Benjamini-Hochberg procedure

39

2. Benjamini-Hochberg

- Instead of adjusting FWER, adjusts false discovery rate (FDR): proportion of false positives **in set of rejected H_0 (i.e., P < 0.05)**

	Null True	Alternative True	Total
Not Called Significant	*U*	*T*	*m - R*
Called Significant	*V*	*S*	*R*
	m_0	$m\text{-}m_0$	*m*

$$FDR = \frac{V}{R}$$

40

2. Benjamini-Hochberg

1. Order raw P-values in increasing order
2. Find test with highest rank, j, for which corresponding P-value is ≤ $(j/m) \times \delta$, where δ is FDR level (0.05) and m is number of tests
3. P-values of rank ≤ j are significant

2. Benjamini-Hochberg

Rank (j)	P-value	(j/m)× δ	Reject H_0 ?
1	0.0008	0.005	1
2	0.009	0.010	1
3	0.165	0.015	0
4	0.205	0.020	0
5	0.396	0.025	0
6	0.450	0.030	0
7	0.641	0.035	0
8	0.781	0.040	0
9	0.900	0.045	0
10	0.993	0.050	0

Or just use `p.adjust()` function in R

Comparing corrections

- I simulated data and computed 4950 pairwise correlations and P-values (H_0 is false for 1225)
- Using Bonferroni, 94 (8%) were significant
- Using Benjamini-Hochberg, 1165 (95%) were significant
- So BH is **MUCH** better, but still not perfect
- Best solution is to not calculate so many P-values in the first place!

My rules for hypothesis testing

- Distill research question down to as few hypotheses as possible → calculate as few P-values as possible
- Lots of thinking before you collect data and run tests (i.e., go down as few forks as possible)
- Simulate fake data to think and work through your analyses and code
- Communicate/write down your hypothesis & methods before data collection

Questions?

45

46

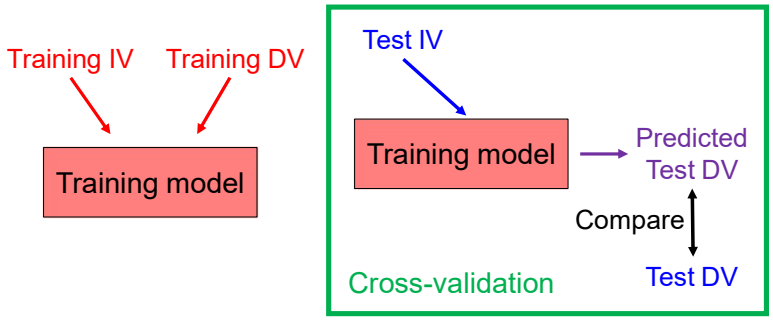
What is prediction?

- Fit a model to your data to predict unknown DV values, given **NEW** IV values
- E.g., use `lm(body.mass ~ femur.length)` to predict body mass using new femoral specimens
- Thus, need to assess how your model does on a **NEW** dataset where DV and IV values are known (i.e., **cross-validation**)

47

Cross-validation

- **Training data**: data used to fit model
- **Test data**: data used to test trained model



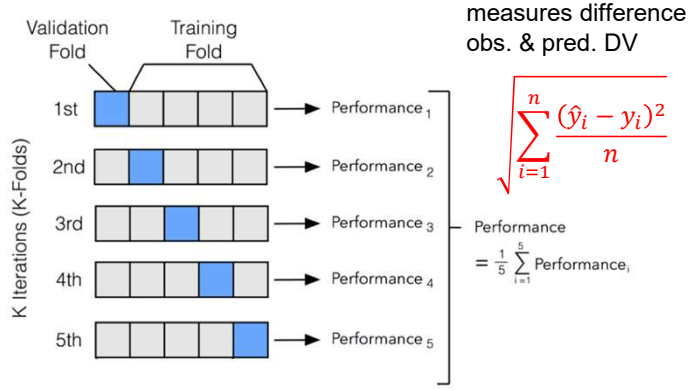
48

Cross-validation

1. Holdout method: 80% data reserved for training, 20% for testing (or 75-25, 70-30, etc.)
2. *k*-fold: E.g., 10-fold → use 1st 10% of data to test, other 90% to train; 2nd 10% to test, other 90% to train; repeat 10x & average model predictions

Cross-validation

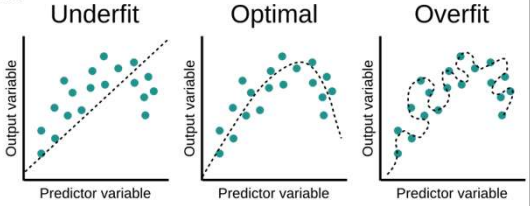
Performance metrics
1. R^2 btw obs. & pred. DV
2. Root-mean-square-error (RMSE): measures difference btw obs. & pred. DV



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

Overfitting

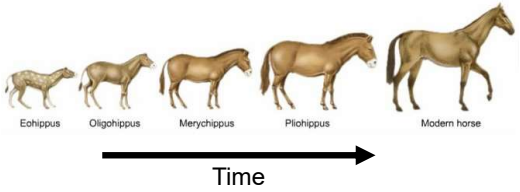
- Cross-validation is done to prevent overfitting (fitting the noise structure specific to one system instead of the signal)
- Therefore, test data (and its noise structure) must be **INDEPENDENT** from training data
- This is complicated by the presence of autocorrelation



Autocorrelation

- How one variable is correlated with itself → correlated errors (noise)
- Specifically, how closer values are more similar

1. Temporal autocorrelation: values closer in time are more similar
   - E.g., "tomorrow is likely to be sunny like today"



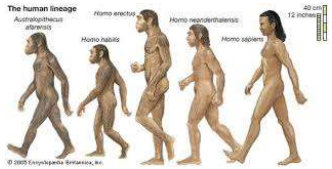
Also, evolution! (descent w/ modification)

Autocorrelation

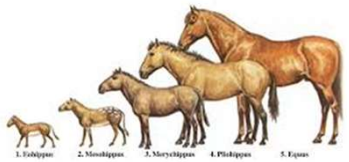
1. Temporal autocorrelation
2. Spatial autocorrelation: values closer in space are more similar
   - E.g., Tobler's first law of geography

Autocorrelation

1. Temporal autocorrelation
2. Spatial autocorrelation
3. Phylogenetic autocorrelation: values in more closely related taxa are more similar

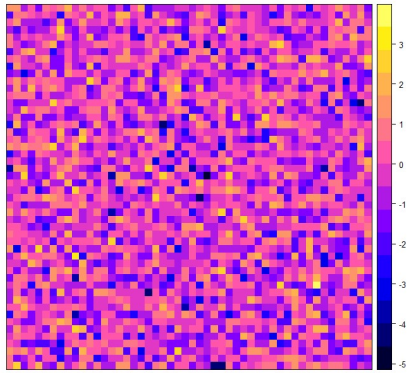


vs.



Cross-validation w/ no spatial autocorrelation

Dependent variable



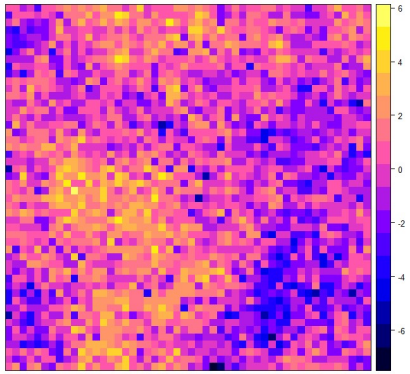
IV is white noise; fxn is noisy quadratic

- 60-40 holdout
- LOESS curve (span = 0.01)
- Training data
  1. R^2 = 0.43
  2. RMSE = 0.86
- Test data
  1. R^2 = 0.035
  2. RMSE = 1.09

Both metrics worsened!

Cross-validation WITH spatial autocorrelation

Dependent variable



IV is white noise; fxn is noisy quadratic

- 60-40 holdout
- LOESS curve (span = 0.01)
- Training data
  1. R^2 = 0.45
  2. RMSE = 1.31
- Test data
  1. R^2 = 0.14
  2. RMSE = 1.67

Worsened, but less so!

What happened?

No spatial autocorrelation
- R^2: 0.43 → 0.035
- RMSE: 0.86 → 1.09

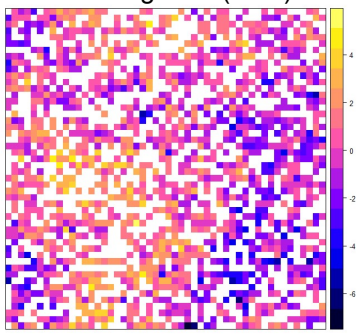
Spatial autocorrelation
- R^2: 0.45 → 0.14
- RMSE: 1.31 → 1.67

- Same spatial autocorrelation structure is present in both training and test data (not truly independent!)

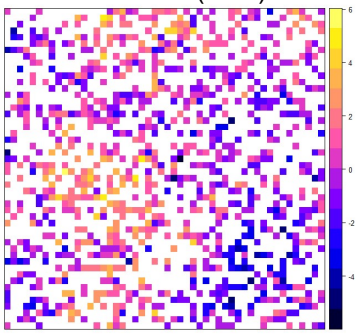
57

What happened?

Training data (60%)



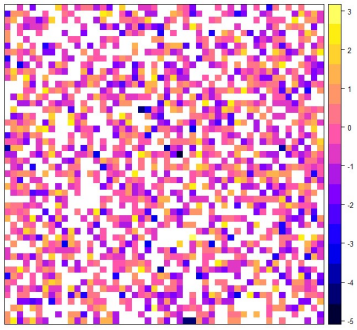
Test data (40%)



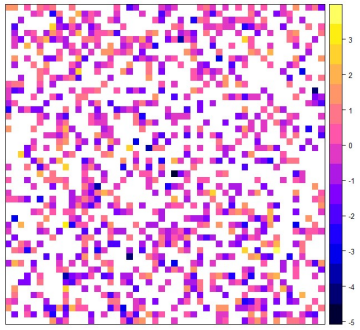
58

Compare with no spatial autocorrelation

Training data (60%)



Test data (40%)



59

What happened?

No spatial autocorrelation
- R^2: 0.43 → 0.035
- RMSE: 0.86 → 1.09

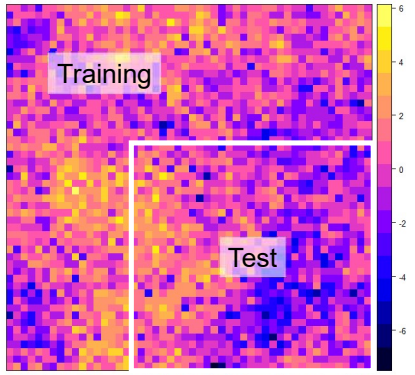
Spatial autocorrelation
- R^2: 0.45 → 0.14
- RMSE: 1.31 → 1.67

- Same spatial autocorrelation structure is present in both training and test data (not truly independent!)
- Training model is fitting some of the noise structure, which is present in test data
- Thus, overconfident in how model generalizes to new datasets (likely has diff. noise structure)

60

What should be done

Dependent variable



IV is white noise; fxn is noisy quadratic

- Training data
1. R^2 = 0.46
2. RMSE = 1.22
- Test data
1. R^2 = 0.00
2. RMSE = 1.88

Worsened way more! More honest assessment of model.

61

Questions?

62

Summary

- Where does your research fall?
- Which is best for your question?

	Exploratory	Confirmatory	Prediction
Frequentist/ Monte Carlo			
Likelihood			
Bayesian			

63

Summary

- Three main modes of statistical inference:

1. Exploratory data analysis
   - Explore patterns and relationships in your data
   - **DO NOT** calculate P-values
2. Confirmatory data analysis
   - Tests *a priori* hypotheses with CIs & P-values
   - Correct for multiple comparisons if necessary
3. Prediction
   - Using fitted model to predict DV in new dataset
   - **MUST** cross-validate with **INDEPENDENT** dataset

64