431 Class 26

Thomas E. Love

2018-12-06

Today's R Setup

```
Loading required package: ggplot2
```

```
library(knitr); library(broom); library(tidyverse)
```

and a couple of secrets, hidden for now.

13 Data Sets (summarized) in the d_long tibble:

```
# A tibble: 13 x 7
     set     n mean_x  sd_x mean_y  sd_y `cor(x, y)`
   <dbl> <dbl>  <dbl> <dbl>  <dbl> <dbl>        <dbl>
 1     1   142   54.3  16.8   47.8  26.9        -0.06
 2     2   142   54.3  16.8   47.8  26.9        -0.07
 3     3   142   54.3  16.8   47.8  26.9        -0.07
 4     4   142   54.3  16.8   47.8  26.9        -0.06
 5     5   142   54.3  16.8   47.8  26.9        -0.06
 6     6   142   54.3  16.8   47.8  26.9        -0.06
 7     7   142   54.3  16.8   47.8  26.9        -0.07
 8     8   142   54.3  16.8   47.8  26.9        -0.07
 9     9   142   54.3  16.8   47.8  26.9        -0.07
10    10   142   54.3  16.8   47.8  26.9        -0.06
11    11   142   54.3  16.8   47.8  26.9        -0.07
12    12   142   54.3  16.8   47.8  26.9        -0.07
13    13   142   54.3  16.8   47.8  26.9        -0.07
```

New Data: Model for Set 1

```
set_1 <- lm(y ~ x, data = d_long %>% filter(set == 1))

tidy(set_1) %>% kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	53.43	7.69	6.94	0.00
x	-0.10	0.14	-0.76	0.45

```
glance(set_1) %>%
  select(r.squared, adj.r.squared, sigma, p.value) %>%
  kable(digits = 3)
```

	r.squared	adj.r.squared	sigma	p.value
value	0.004	-0.003	26.98	0.448

New Data: Model for Set 2

```
set_2 <- lm(y ~ x, data = d_long %>% filter(set == 2))

tidy(set_2) %>% kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	53.81	7.69	7.00	0.00
x	-0.11	0.14	-0.81	0.42

```
glance(set_2) %>%
  select(r.squared, adj.r.squared, sigma, p.value) %>%
  kable(digits = 3)
```

	r.squared	adj.r.squared	sigma	p.value
value	0.005	-0.002	26.968	0.417

New Data: Model for Set 3

```
set_3 <- lm(y ~ x, data = d_long %>% filter(set == 3))

tidy(set_3) %>% kable(digits = 2)
```

term	estimate	std.error	statistic	p.value
(Intercept)	53.80	7.69	6.99	0.00
x	-0.11	0.14	-0.81	0.42

```
glance(set_3) %>%
  select(r.squared, adj.r.squared, sigma, p.value) %>%
  kable(digits = 3)
```

	r.squared	adj.r.squared	sigma	p.value
value	0.005	-0.002	26.963	0.419

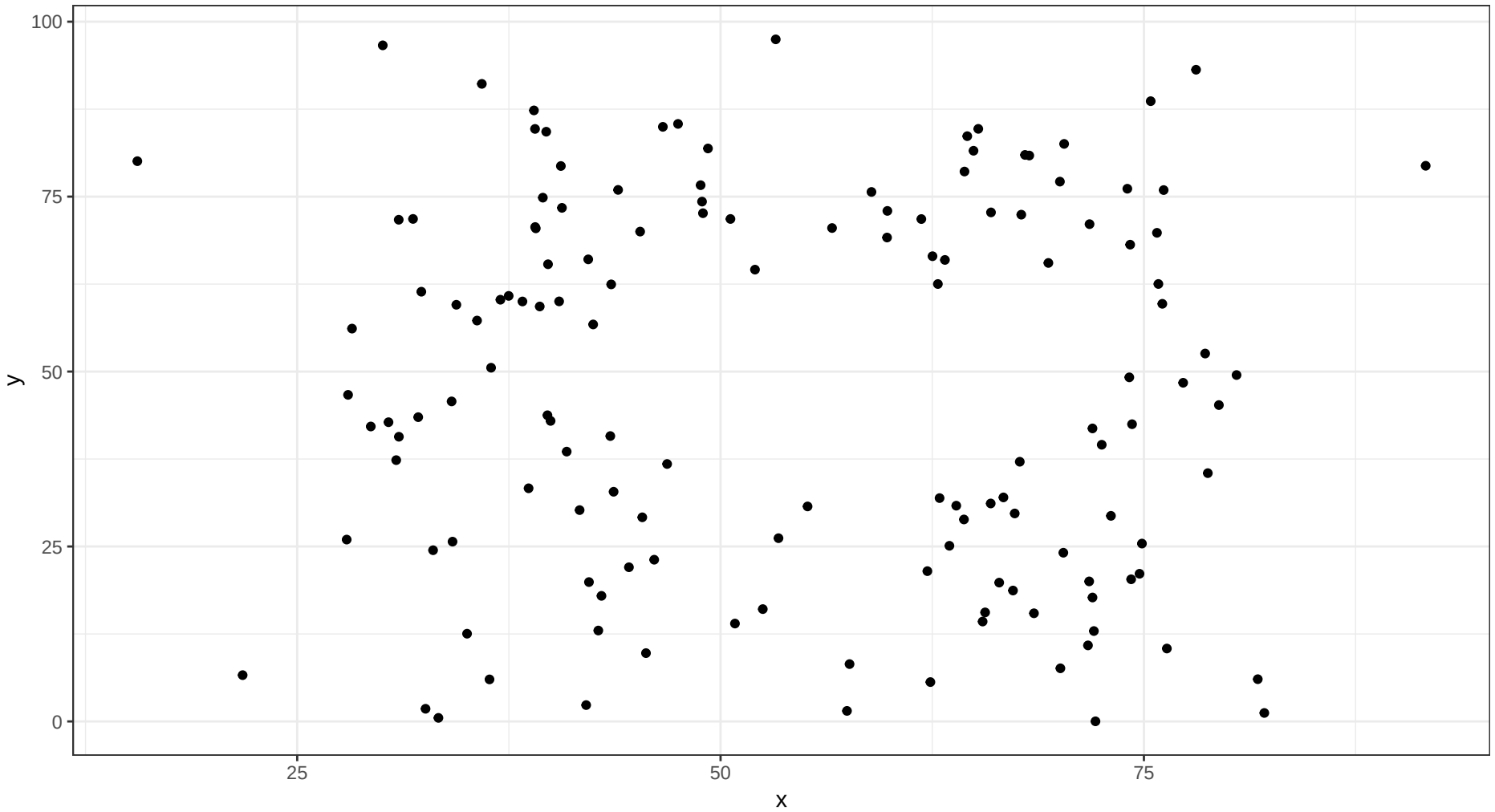
All 13 Models, at a glance

dataset	r.squared	adj.r.squared	sigma	p.value	AIC	BIC
1	0.004	-0.003	26.98	0.45	1343	1352
2	0.005	-0.002	26.97	0.42	1343	1352
3	0.005	-0.002	26.96	0.42	1343	1351
4	0.004	-0.003	26.98	0.45	1343	1352
5	0.004	-0.003	26.98	0.48	1343	1352
6	0.004	-0.003	26.98	0.47	1343	1352
7	0.005	-0.002	26.97	0.42	1343	1352
8	0.005	-0.002	26.97	0.41	1343	1352
9	0.005	-0.002	26.97	0.42	1343	1352
10	0.004	-0.003	26.97	0.46	1343	1352
11	0.005	-0.002	26.97	0.41	1343	1352
12	0.004	-0.003	26.97	0.43	1343	1352
13	0.004	-0.003	26.97	0.44	1343	1352

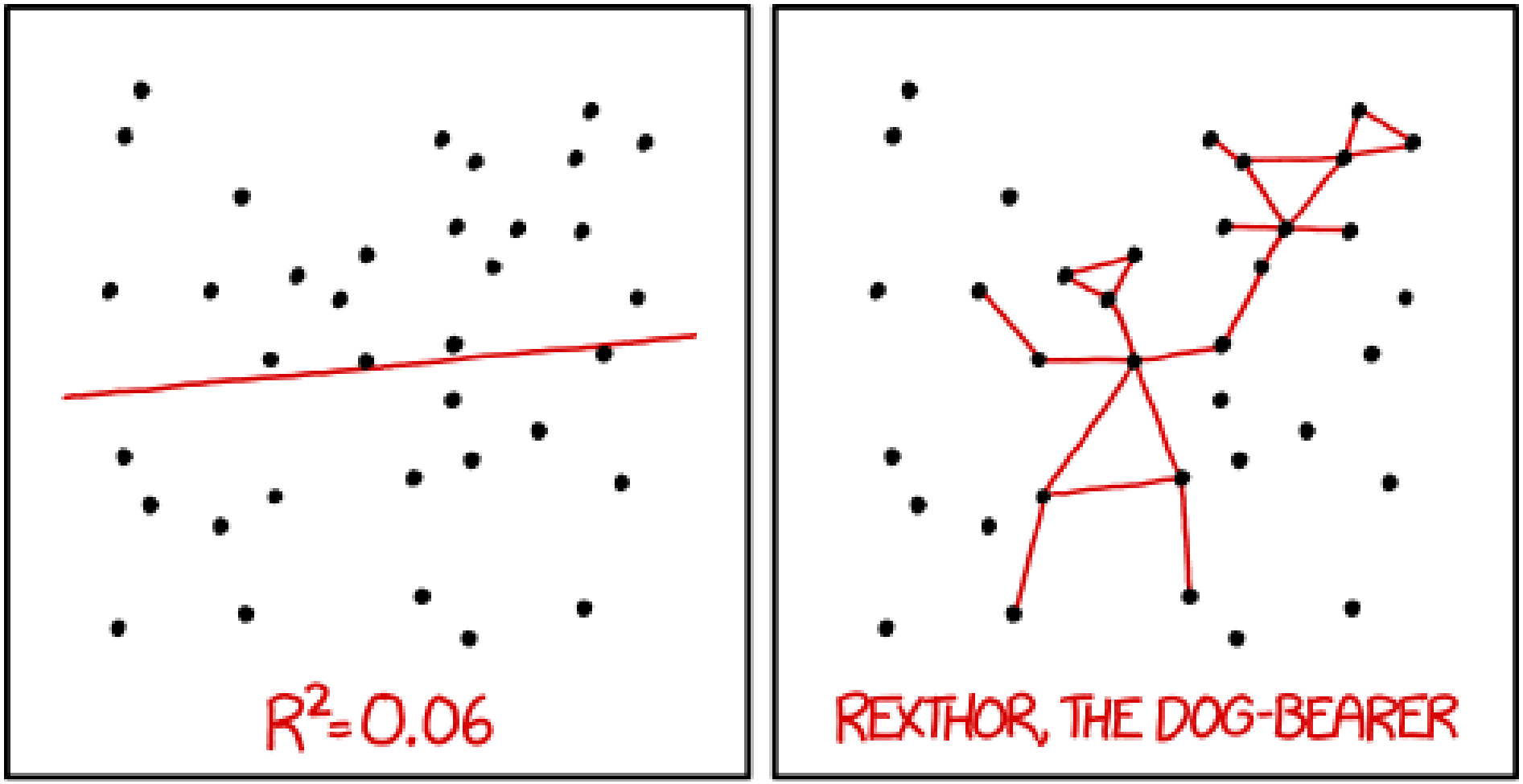
Plot for Set 1 (code)

```
d_long %>%
  filter(set == 1) %>%
  ggplot(., aes(x = x, y = y)) +
  geom_point() +
  theme_bw()
```

Plot for Set 1 (result)

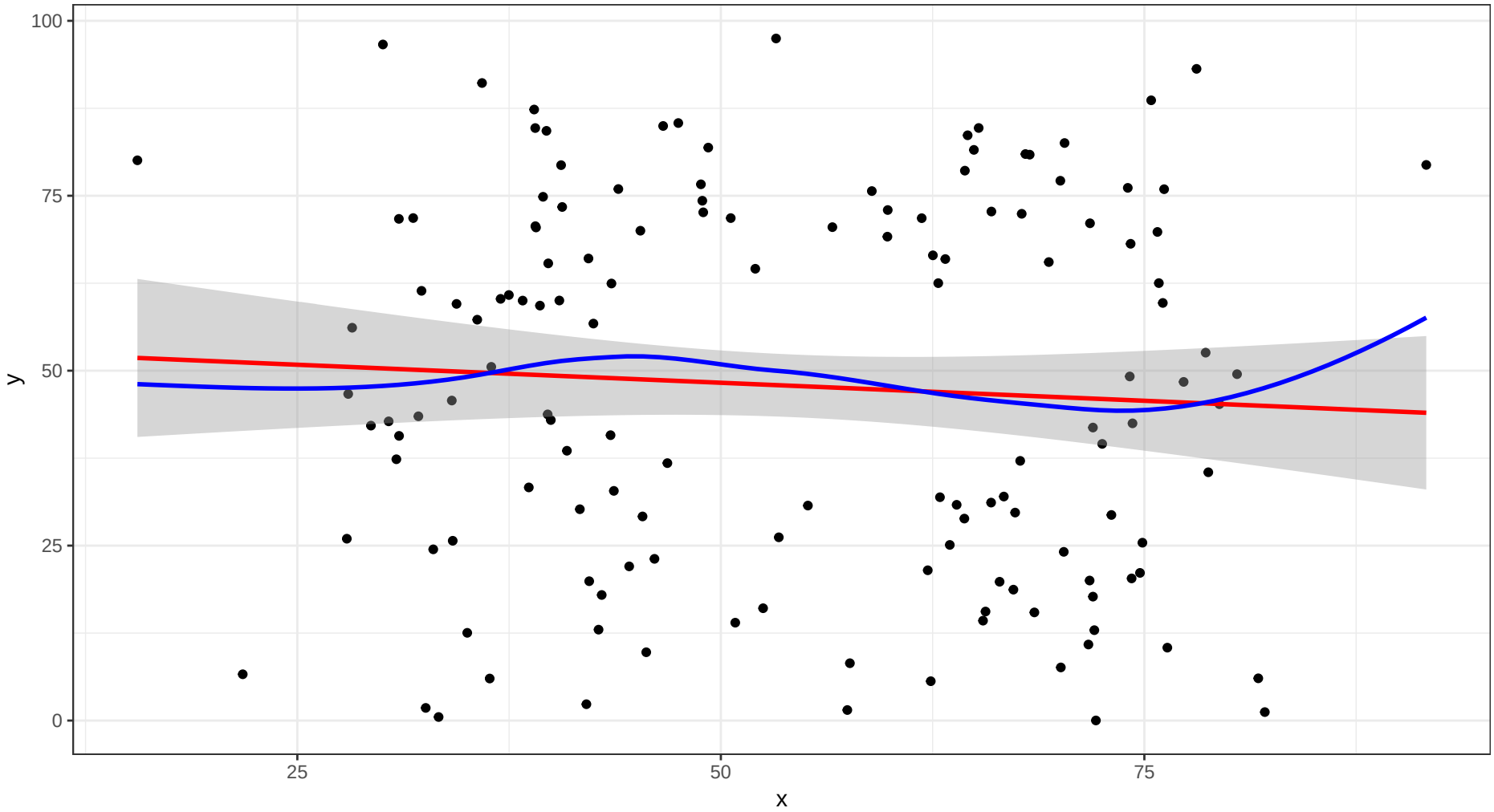


https://xkcd.com/1725/



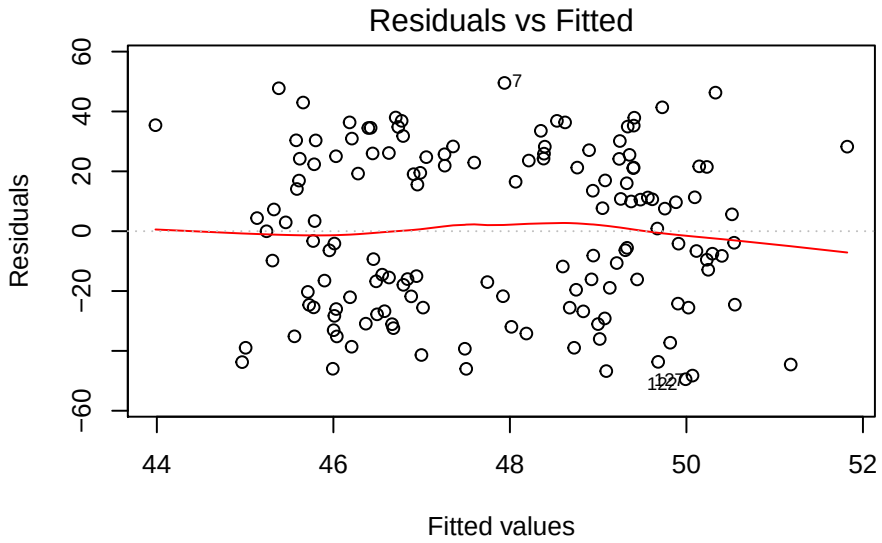
I DON'T TRUST LINEAR REGRESSIONS WHEN IT'S HARDER TO GUESS THE DIRECTION OF THE CORRELATION FROM THE SCATTER PLOT THAN TO FIND NEW CONSTELLATIONS ON IT.

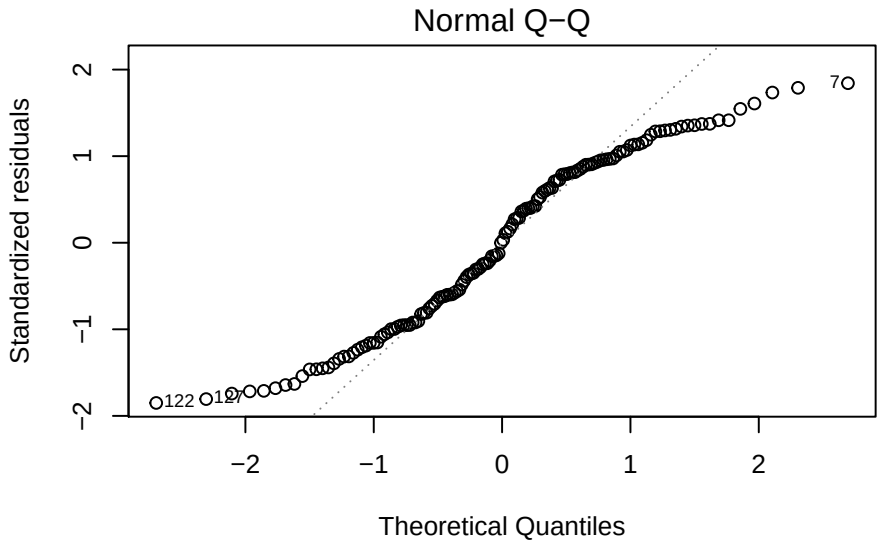
Plot for Set 1 (with linear model and loess smooth)

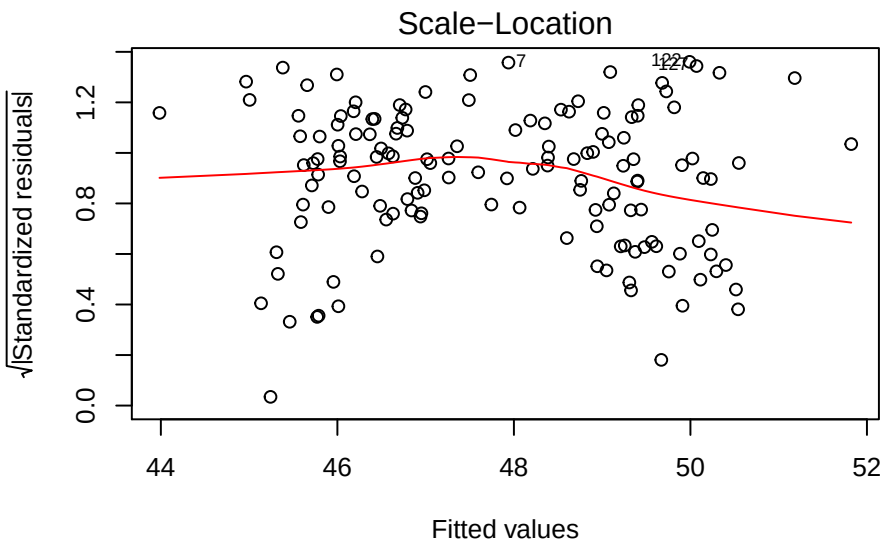


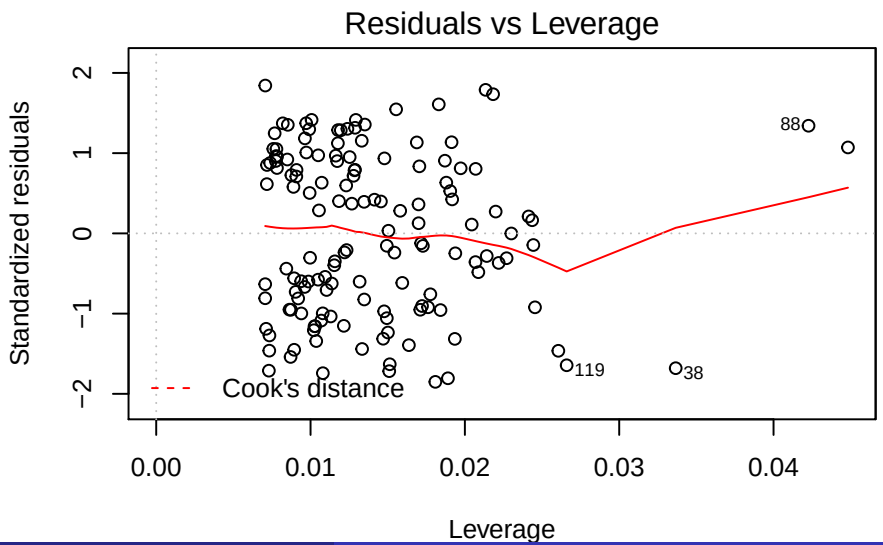
- Added `geom_smooth(method = "lm", col = "red")` and
- `geom_smooth(method = "loess", col = "blue", se = FALSE)`

Residuals Plots for Set 1 Model







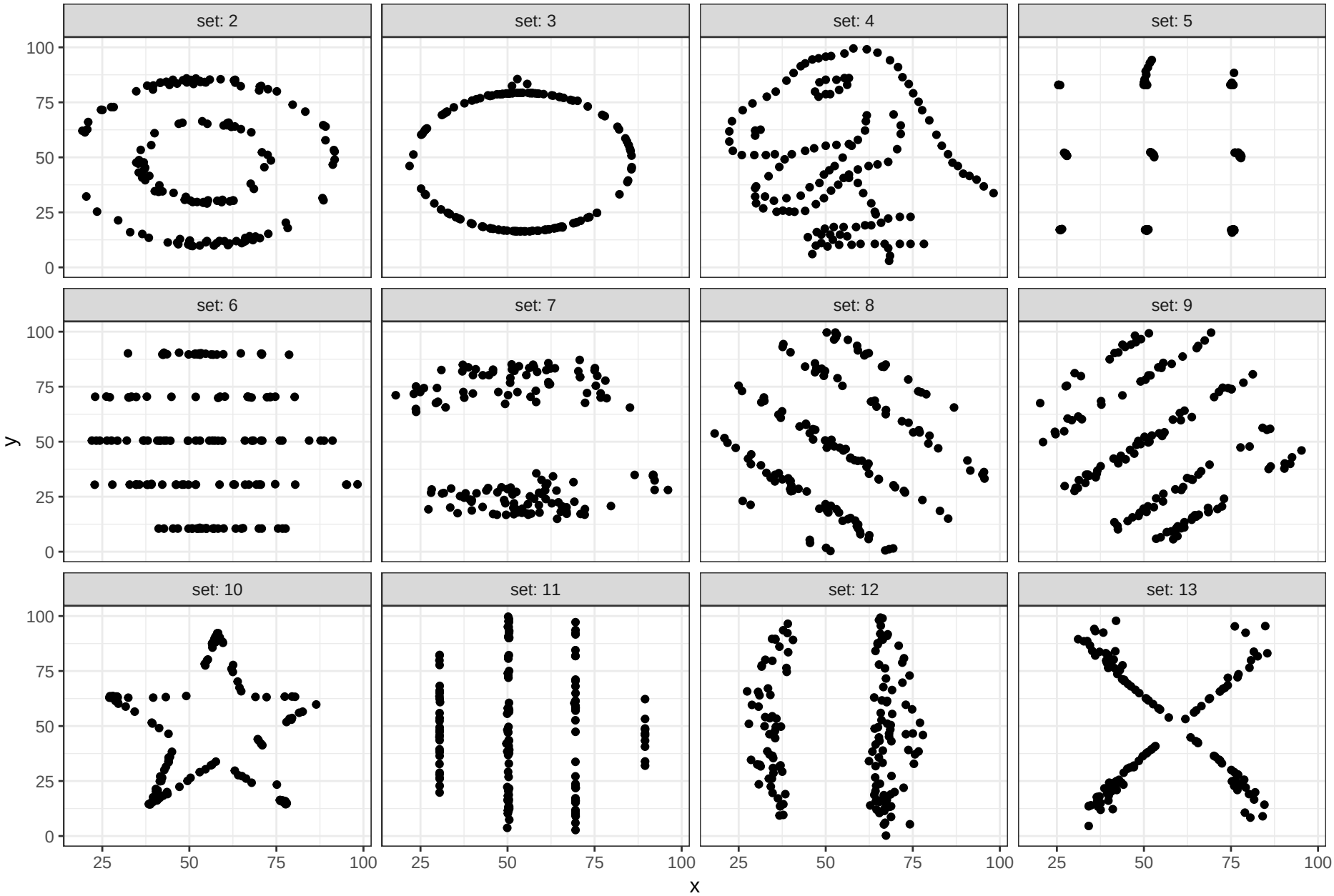


Models 2-13

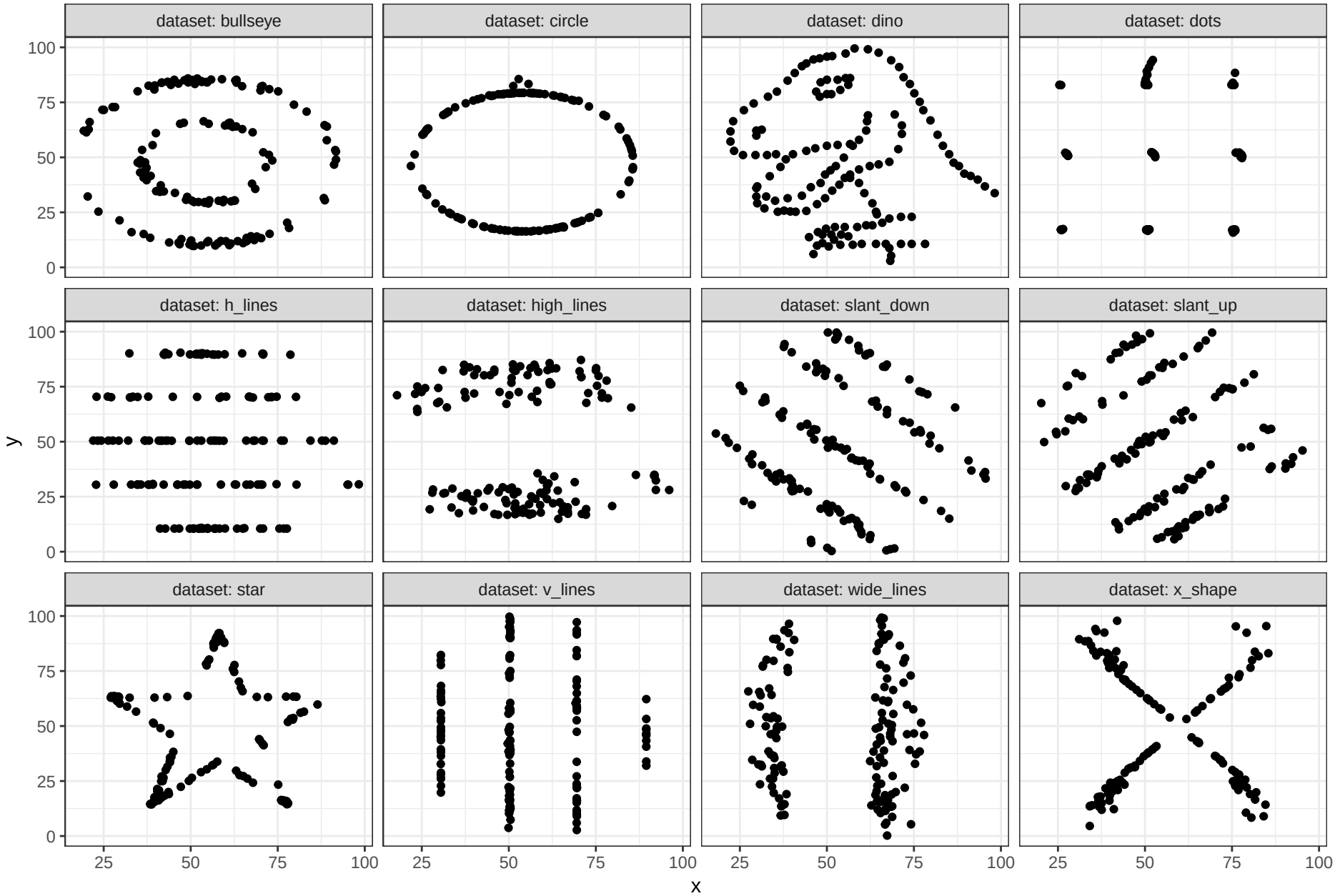
Models 2-13 all look about the same in terms of means, medians, correlations, regression models, but what happens if we plot the data?

```
d_long %>%
  filter(set != 1) %>%
  ggplot(., aes(x = x, y = y)) +
  geom_point() +
  theme_bw() +
  facet_wrap(~ set, labeller = "label_both")
```

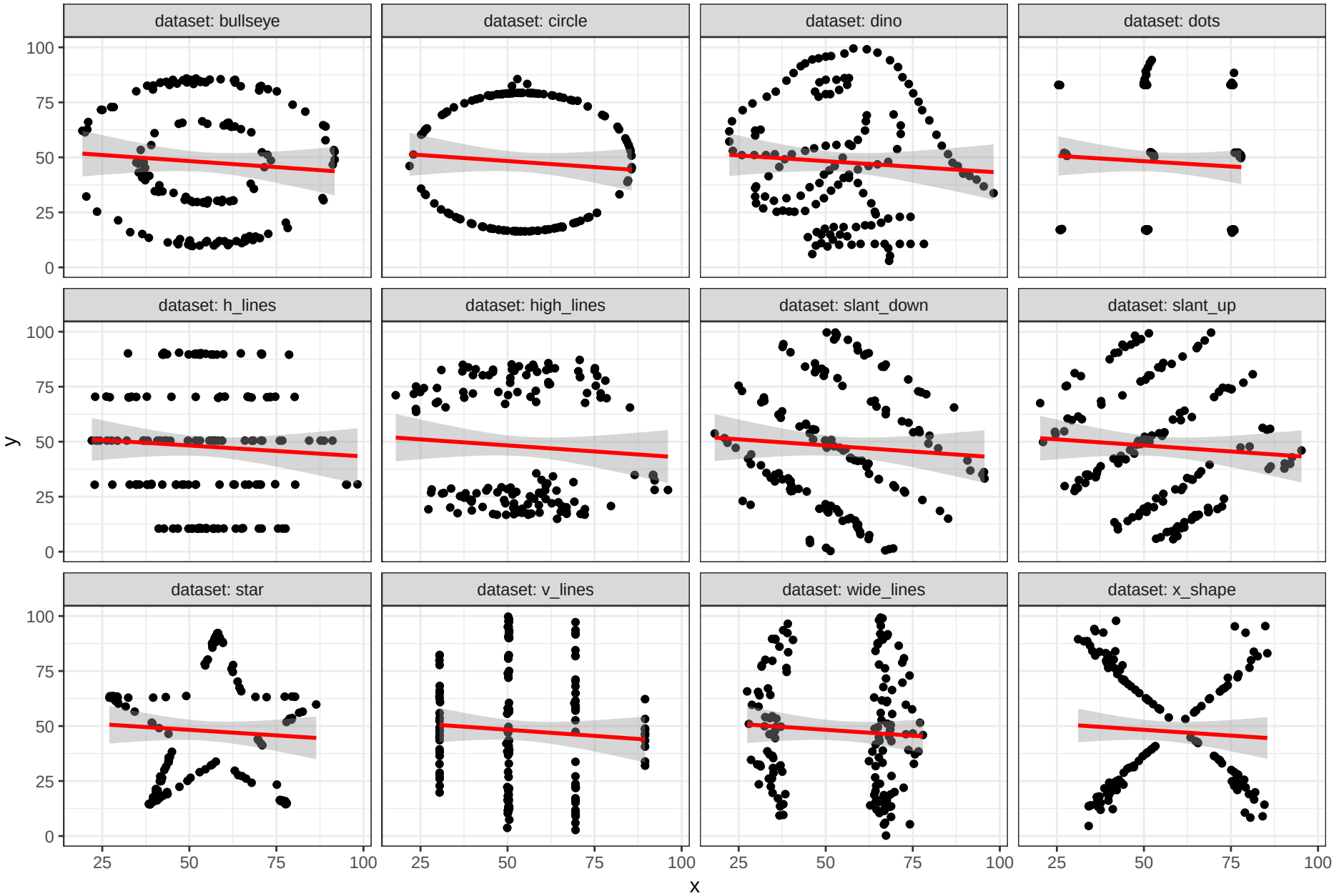
The Other 12 Data Sets



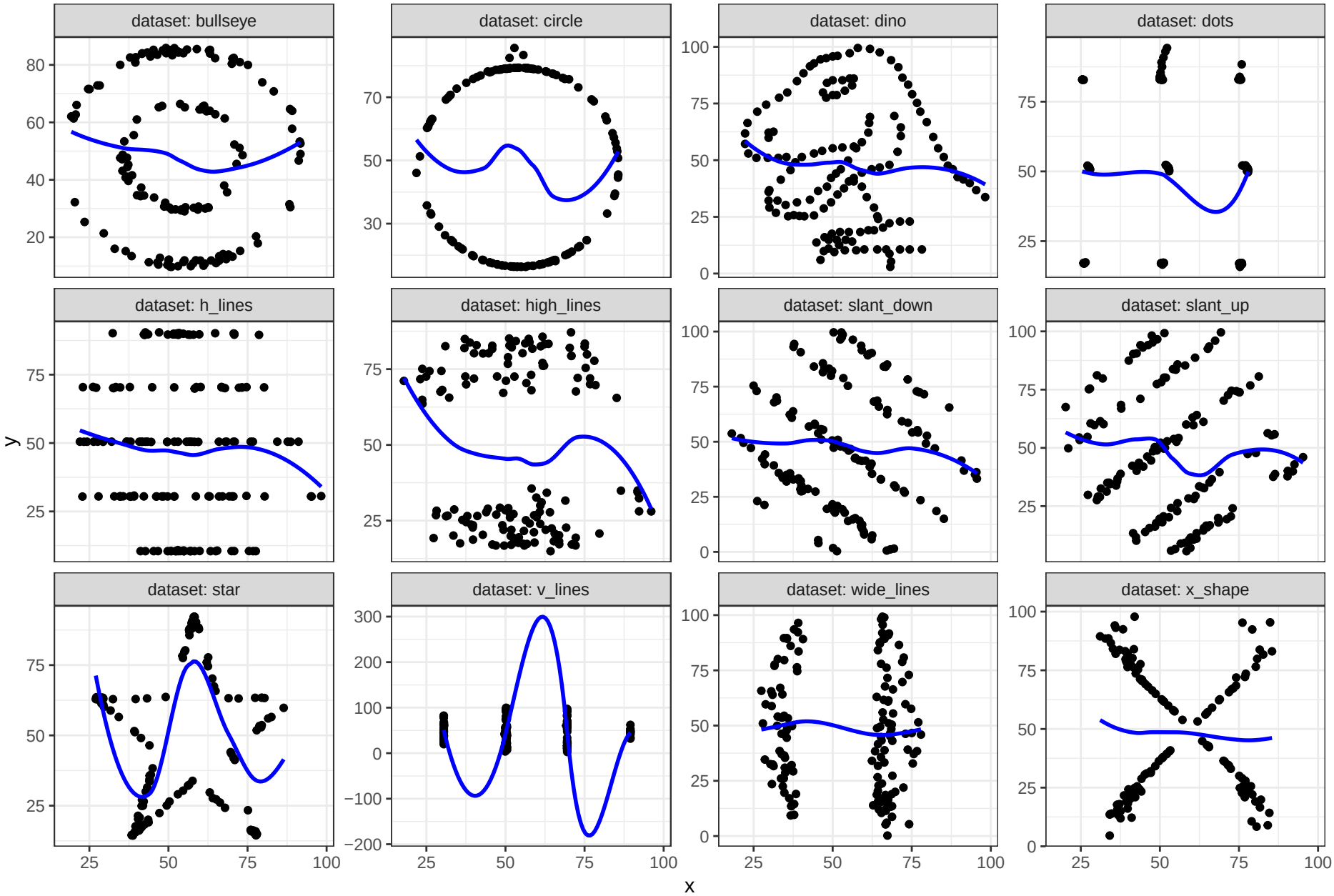
Actually, each of these sets has a name



And a linear model yields the same fit for each



And a loess smooth?



And the data come from

These are the datasauRus dozen data sets, available in the datasauRus package, which you can install from CRAN.

```
library(datasauRus)
d_long <- datasaurus_dozen
```

A cool thing, available online. . .

Visit https://r-mageddon.netlify.com/post/reanimating-the-datasaurus/

```
library(datasauRus)
library(ggplot2)
library(gganimate)

ggplot(datasaurus_dozen, aes(x=x, y=y))+
  geom_point()+
  theme_minimal() +
  transition_states(dataset, 3, 1)
```

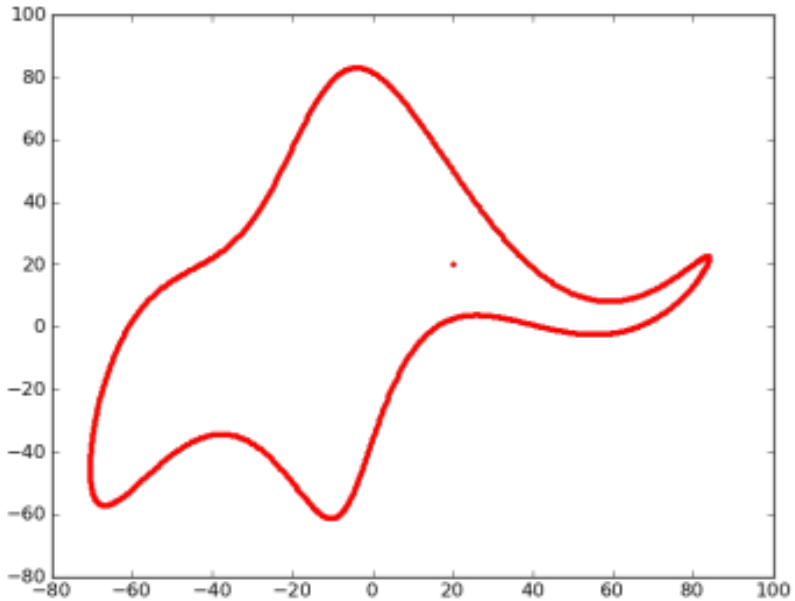
More on gganimate (I hope) in 432.

John von Neumann famously said

> With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

By this he meant that one should not be impressed when a complex model fits a data set well. With enough parameters, you can fit any data set.

It turns out you can literally fit an elephant with four parameters if you allow the parameters to be complex numbers.



p Hacking and "Researcher Degrees of Freedom"

Hack Your Way To Scientific Glory

https://fivethirtyeight.com/features/science-isnt-broken

Hack Your Way To Scientific Glory

You're a social scientist with a hunch: **The U.S. economy is affected by whether Republicans or Democrats are in office.** Try to show that a connection exists, using real data going back to 1948. For your results to be publishable in an academic journal, you'll need to prove that they are "statistically significant" by achieving a low enough p-value.

1 CHOOSE A POLITICAL PARTY

Republicans | Democrats

2 DEFINE TERMS

Which politicians do you want to include?

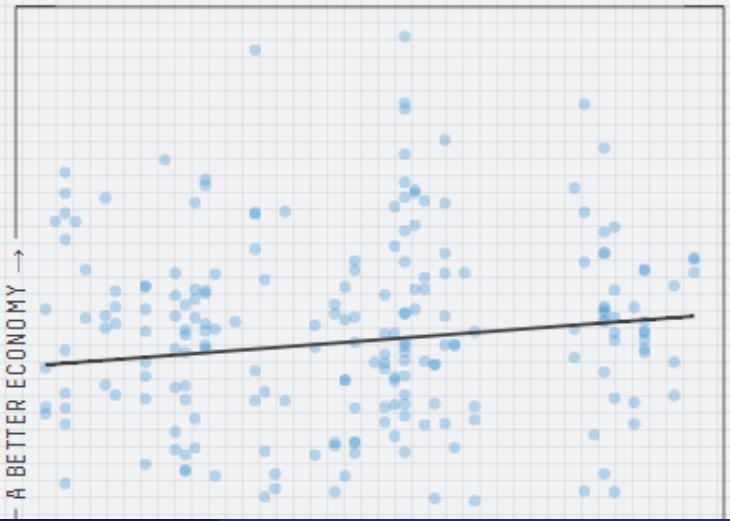
- [ ] Presidents
- [x] Governors
- [x] Senators
- [ ] Representatives

How do you want to measure economic performance?

- [ ] Employment
- [x] Inflation
- [x] GDP
- [x] Stock prices

3 IS THERE A RELATIONSHIP?

Given how you've defined your terms, does the economy do better, worse or about the same when more Democrats are in power? Each dot below represents one month of data.



4 IS YOUR RESULT SIGNIFICANT?

If there were no connection between the economy and politics, what is the probability that you'd get results at least as strong as yours? That probability is your p-value, and by convention, you need a **p-value of 0.05 or less** to get published.

Result: Almost

Your **0.06** p-value is close to the 0.05 threshold. Try tweaking your variables to see if you can push it over the line!

What can you get?

I was able to get

- $p < 0.01$ (positive effect of Democrats on economy)
- $p = 0.01$ (negative effect of Democrats)
- $p = 0.03$ (negative effect of Democrats)
- $p = 0.03$ (positive effect of Democrats)

but also . . .

- $p = 0.05, 0.06, 0.07, 0.09, 0.17, 0.19, 0.20, 0.22, 0.23, 0.47, 0.51$

without even switching parties, exclusively by changing my definitions of terms (section 2 of the graphic.)

"Researcher Degrees of Freedom", 1

[I]t is unacceptably easy to publish "statistically significant" evidence consistent with any hypothesis.

The culprit is a construct we refer to as ***researcher degrees of freedom****. In the course of collecting and analyzing data, researchers have many decisions to make: Should more data be collected? Should some observations be excluded? Which conditions should be combined and which ones compared? Which control variables should be considered? Should specific measures be combined or transformed or both?*

Simmons et al. *link*

"Researcher Degrees of Freedom", 2

> *. . . It is rare, and sometimes impractical, for researchers to make all these decisions beforehand. Rather, it is common (and accepted practice) for researchers to explore various analytic alternatives, to search for a combination that yields statistical significance, and to then report only what worked. The problem, of course, is that the likelihood of at least one (of many) analyses producing a falsely positive finding at the 5% level is necessarily greater than 5%.*

For more, see

- Gelman's blog $2012 - 11 - 01$ "Researcher Degrees of Freedom",
- Paper by *Simmons* and others, defining the term.

And this is really hard to deal with. . .

The garden of forking paths: Why multiple comparisons can be a problem, even when there is no fishing expedition or p-hacking and the research hypothesis was posited ahead of time

> *Researcher degrees of freedom can lead to a multiple comparisons problem, even in settings where researchers perform only a single analysis on their data. The problem is there can be a large number of potential comparisons when the details of data analysis are highly contingent on data, without the researcher having to perform any conscious procedure of fishing or examining multiple p-values. We discuss in the context of several examples of published papers where data-analysis decisions were theoretically-motivated based on previous literature, but where the details of data selection and analysis were not pre-specified and, as a result, were contingent on data.*

- *Link* to the paper from Gelman and Loken

Benjamin et al 2017 Redefine Statistical Significance

We propose to change the default P-value threshold for statistical significance for claims of new discoveries from 0.05 to 0.005.

- 0.005 is stringent enough to "break" the current system - makes it very difficult for researchers to reach threshold with noisy, useless studies.

Visit the main *article*. Visit an explanatory piece in *Science*.

Lakens et al. Justify Your Alpha

"In response to recommendations to redefine statistical significance to $p \leq .005$, we propose that researchers should transparently report and justify all choices they make when designing a study, including the alpha level." Visit *link*.


P > 0.05
GAME OVER, TRY AGAIN
imgflip.com

Why not post hoc power analysis?

So you collected data and analyzed the results. Now you want to do an after data gathering (post hoc) power analysis.

1. What will you use as your "true" effect size?
   - Often, point estimate from data - yuck - results very misleading - power is generally seriously overestimated when computed on the basis of statistically significant results.
   - Much better (but rarer) to identify plausible effect sizes based on external information rather than on your sparkling new result.
2. What are you trying to do? (too often)
   - get researcher off the hook (I didn't get $p < 0.05$ because I had low power - an alibi to explain away non-significant findings) or
   - encourage overconfidence in the finding.

The Impact of Study Design (Gelman)

Applied statistics is hard.

- Doing a statistical analysis is like playing basketball, or knitting a sweater. You can get better with practice.
- Incompetent statistics does not necessarily doom a research paper: some findings are solid enough that they show up even when there are mistakes in the data collection and data analyses. But we've also seen many examples where incompetent statistics led to conclusions that made no sense but still received publication and publicity.
- We should be thinking not just about data analysis, but also data quality.

"Let's shrink Big Data into Small Data ...
and hope it magically becomes Great Data."

So, what have we learned so far?

Ten Simple Rules for Effective Statistical Practice

From *PLoS Comput Biol* *link*

1. Statistical Methods Should Enable Data to Answer Scientific Questions
2. Signals Always Come with Noise
3. Plan Ahead, Really Ahead
4. Worry About Data Quality
5. Statistical Analysis Is More Than a Set of Computations
6. Keep it Simple
7. Provide Assessments of Variability
8. Check Your Assumptions
9. When Possible, Replicate!
10. Make Your Analysis Reproducible

On Planning Ahead - Really Ahead...

"To consult the statistician after an experiment is finished is often merely to ask him to conduct a post mortem examination. He can perhaps say what the experiment died of."

- Sir Ronald Fisher

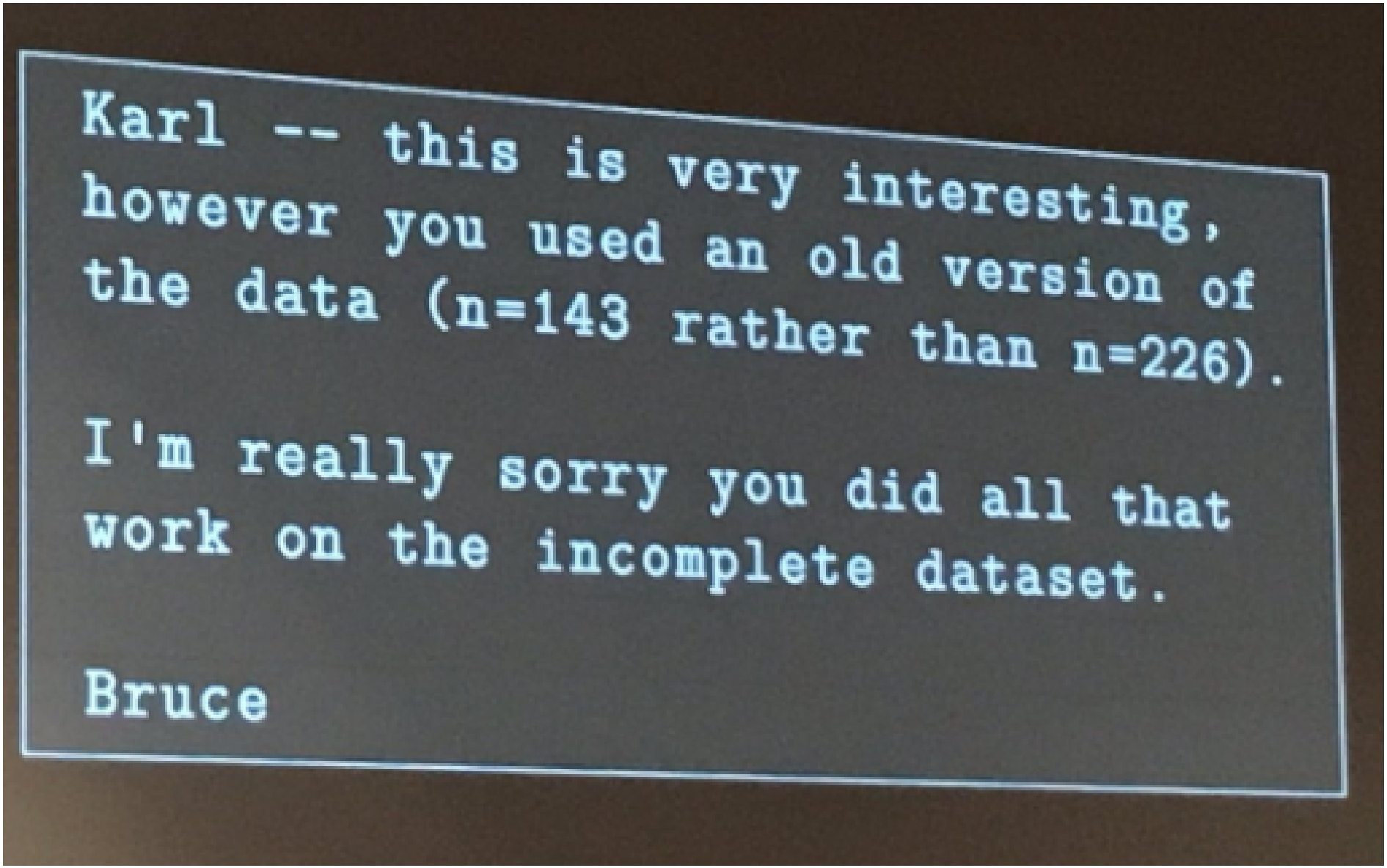
Goals of Reproducible Research

The goal of reproducible research is to tie specific instructions to data analysis so that scholarship can be recreated, better understood and verified. This is usually facilitated by literate programming: a document that combines content and data analytic code. Software? R and R Studio, mostly.

1. Be able to reproduce your own results and allow others to reproduce your results
2. Reproduce an entire report / manuscript / thesis / book / website with a single system command when changes occur (in operating system, statistical software, graphics engines, source data, derived variables, analysis, interpretation).
3. Save time.
4. Provide the ultimate documentation of work done.

Vanderbilt *Tutorial*

Why I Do This. . .


Karl -- this is very interesting,
however you used an old version of
the data (n=143 rather than n=226).
I'm really sorry you did all that
work on the incomplete dataset.
Bruce

Five Practical Tips

1. Document everything.
2. Everything is a (text) file.
3. All files should be human-readable.
4. Explicitly tie your files together.
5. Have a plan to organize, store and make your files available.

The papers/slideshows/abstracts are not the research. The research is the full software environment, code and data that produced the results. (Donoho, 2010). Separating research from its advertisement makes it hard for others to verify or reproduce our findings.

Your closest collaborator is you, six months ago, but you don't respond to emails. (Holder via Broman)

Karl Broman, Steps Towards Reproducible Research *link*

Build Tidy Data Sets

- Each variable you measure should be in one column.
- Each different observation of that variable should be in a different row.
- There should be one table for each "kind" of variable.
- If you have multiple tables, they should include a column in the table that allows them to be linked.
- Include a row at the top of each data table that contains real row names. `Age_at_Diagnosis` is a much much better name than `ADx`.
- Build useful codebooks.

Jeff Leek: "How to share data with a statistician" *link*

Wisdom from DL Donoho (2010) re: Open-Source

But other people will use my data and code to compete with me?

- True.

Wisdom from DL Donoho (2010) re: Open-Source

But other people will use my data and code to compete with me?

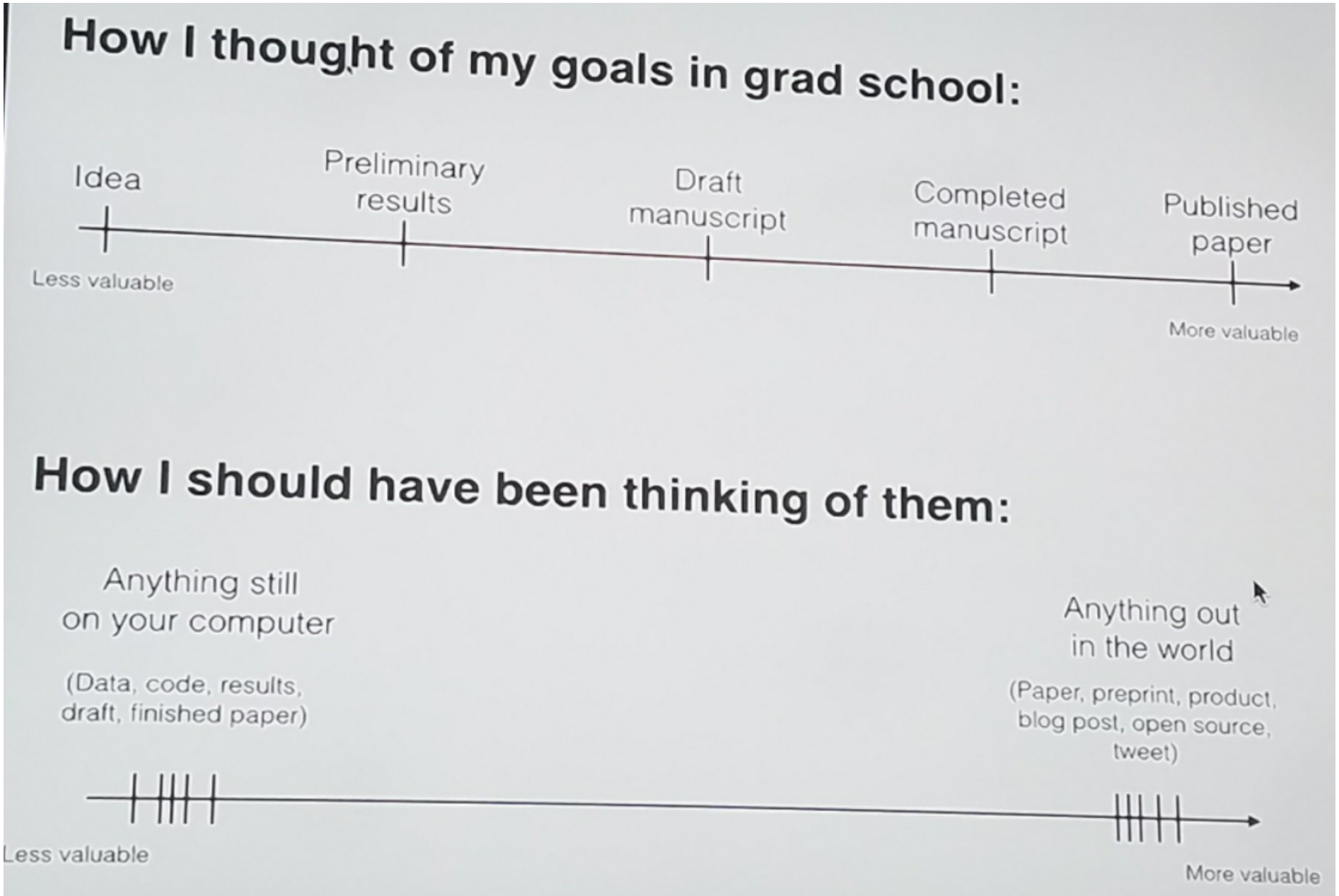
- True.
- But competition means that strangers will read your work, try to learn from you, cite you, and try to do things even better.

Wisdom from DL Donoho (2010) re: Open-Source

But other people will use my data and code to compete with me?

- True.
- But competition means that strangers will read your work, try to learn from you, cite you, and try to do things even better.
- If you prefer obscurity, why are you publishing?

A Most Important Tip (@drob)



From Kass et al.'s Ten Simple Rules...

"A central and common task for us as research investigators is to decipher what our data are able to say about the problems we are trying to solve. Statistics is a language constructed to assist this process, with probability as its grammar. While rudimentary conversations are possible without good command of the language (and are conducted routinely), principled statistical analysis is critical in grappling with many subtle phenomena to ensure that nothing serious will be lost in translation and to increase the likelihood that your research findings will stand the test of time."

> *Among the many articles reporting on the ASA's statement on p-values, we particularly liked a quote from biostatistician Andrew Vickers:* ***Treat statistics as a science, not a recipe.***

The Signal and The Noise

- Nature's laws do not change very much.
- There is no reason to conclude that the affairs of men are becoming more predictable. The opposite may well be true.

Thinking Probabilistically, and using the Bayesian way of thinking about prediction

- Don't fall into the comforting trap of binary thinking. Expressions of uncertainty are not admissions of weakness.
- Know Where You're Coming From - state explicitly how likely we believe an event is to occur *before* we begin to weigh the evidence.
- The volume of information is increasing exponentially. But the signal-to-noise ratio may be waning. We need better ways of distinguishing the two.

Our bias is to think that we are better at prediction than we really are.

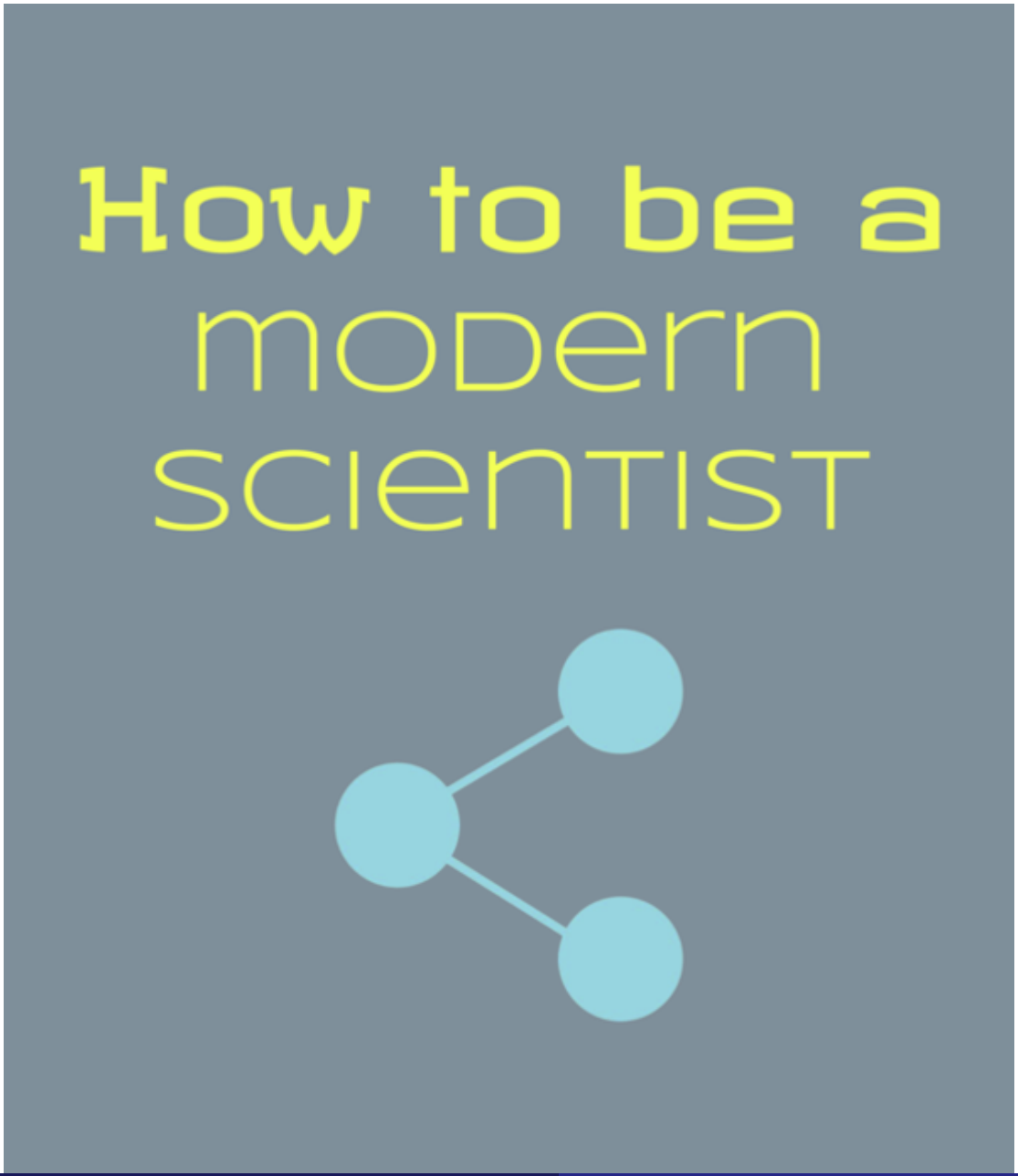
The Course So Far

1. Statistics is too important to be left to statisticians.
2. Models and visualization are the big takeaways, but don't forget about methods for making statistical inferences.
3. Reproducible research is the current wave.
4. Things are changing quickly. We live in interesting times.

What About 432?

- What do you want to know?

https://leanpub.com/modernscientist




That's all Folks!