Introduction to the Course

Sergio I. Garcia-Rios

Government 6029: Advanced Regression Analysis

Introductions

- Teaching team

Introductions

- Teaching team
- You!
  - Name
  - Research agenda
  - Previous projects, tools

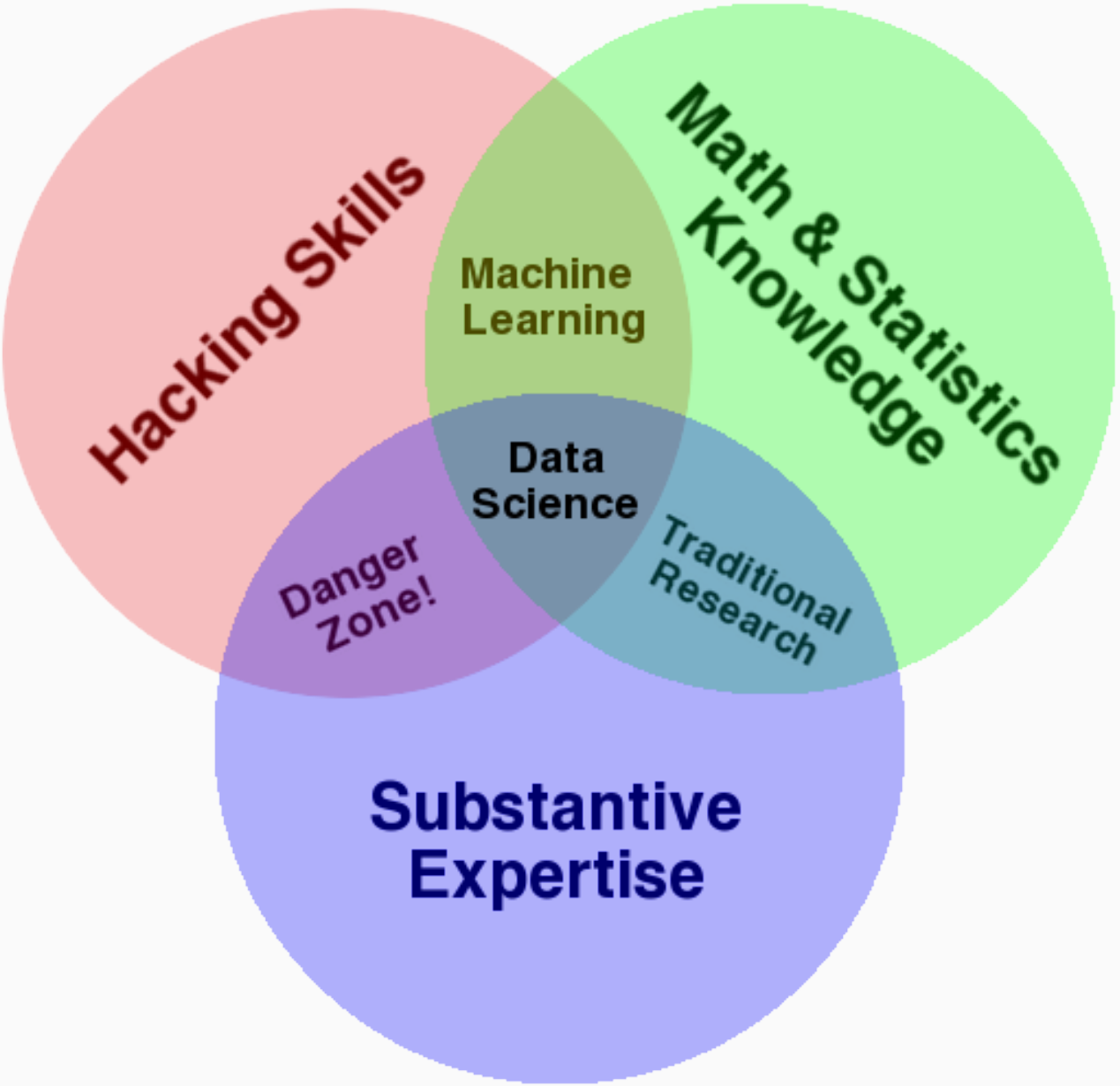
What are we doing here?

What are the course objectives?

At the end of this course, you will be able to ...

1. Estimate and interpret linear models
2. Identify and explain the assumptions of the linear model
3. Diagnose problems linear models and use appropriate solutions
4. Represent statistical models in matrix algebra and compute basic matrix operations
5. Use R to implement the statistical methods introduced here
6. Take other advanced stats courses

What are we covering in this course?



Why Linear Models?

The Linear Model

$$y = X\beta + \epsilon$$

Essentially, all models are wrong, but some are useful.

- Box, G. E. P. and Draper, N. P. (1987) Empirical Model Building and Response Surface

What else is new?

Recent Trends in (Quantitative Social) Science

- Reproducibility and Open Science
- NHST (Null Hypothesis Significance Test) doubts
- Causal inference
- Prediction
- Computation
- Data munging

Why the Tools we are Using?

Why R?

- R is free (as in free beer every first Friday)

Why R?

- R is free (as in free beer every first Friday)
- R is free (as in freedom)

Why R?

- R is free (as in free beer every first Friday)
- R is free (as in freedom)
- R is a language

Why R?

- R is free (as in free beer every first Friday)
- R is free (as in freedom)
- R is a language
- Graphics and data viz capabilities

Why R?

- R is free (as in free beer every first Friday)
- R is free (as in freedom)
- R is a language
- Graphics and data viz capabilities
- Widely used

R is popular and becoming more so

- r4stats.com, The Popularity of Data Analysis Software

R is popular and becoming more so

- r4stats.com, The Popularity of Data Analysis Software
- KDnuggets Languages for Analytics/Data Mining/Data Science

R is popular and becoming more so

- r4stats.com, The Popularity of Data Analysis Software
- KDnuggets Languages for Analytics/Data Mining/Data Science
- Kagglers' Favorite Tools

R is popular and becoming more so

- r4stats.com, The Popularity of Data Analysis Software
- KDnuggets Languages for Analytics/Data Mining/Data Science
- Kagglers' Favorite Tools
- TIOBE Index for March 2017. R is top 10 *all programming languages*

R is popular and becoming more so

- r4stats.com, The Popularity of Data Analysis Software
- KDnuggets Languages for Analytics/Data Mining/Data Science
- Kagglers' Favorite Tools
- TIOBE Index for March 2017. R is top 10 *all programming languages*
- The RedMonk Programming Language Rankings: January 2015. R is #13 of *all programming languages*

R Markdown

- R + markup language

R Markdown

- R + markup language
- Easier to combine code with results

R Markdown

- R + markup language
- Easier to combine code with results
- Increases reproducibility

How to Learn R

https://youtu.be/JxwxefRAu70?t=30m1s

An Example on the importance of (Good) Data Analysis and Presentation

The *Challenger* launch decision

In 1986, the Challenger space shuttle exploded moments after liftoff

Decision to launch one other most scrutinized in history

Failure of O-rings in the solid-fuel rocket boosters blamed for explosion

Could this failure have been foreseen?

The *Challenger* launch decision

Morton-Thiokol engineers made this table & worried about launching below 53 degrees (Why?)

Flights with O-ring damage	
Flt Number	Temp (F)
2	70
41b	57
41c	63
41d	70
51c	53
61a	79
61c	58

The *Challenger* launch decision

Morton-Thiokol engineers made this table & worried about launching below 53 degrees (Why?)

Flights with O-ring damage	
Flt Number	Temp (F)
2	70
41b	57
41c	63
41d	70
51c	53
61a	79
61c	58

O-ring would erode or have “blow-by” (2 ways to fail) in cold temp

The *Challenger* launch decision

Morton-Thiokol engineers made this table & worried about launching below 53 degrees (Why?)

Flights with O-ring damage	
Flt Number	Temp (F)
2	70
41b	57
41c	63
41d	70
51c	53
61a	79
61c	58

Failed to convince administrators there was a danger

The *Challenger* launch decision

Morton-Thiokol engineers made this table & worried about launching below 53 degrees (Why?)

Flights with O-ring damage	
Flt Number	Temp (F)
2	70
41b	57
41c	63
41d	70
51c	53
61a	79
61c	58

(Counter-argument: “damages at low and high temps”)

The *Challenger* launch decision

Morton-Thiokol engineers made this table & worried about launching below 53 degrees (Why?)

Flights with O-ring damage	
Flt Number	Temp (F)
2	70
41b	57
41c	63
41d	70
51c	53
61a	79
61c	58

Are there problems with this presentation? with the use of data?

The *Challenger* launch decision

Engineers did not consider successes, only failures;
selection on the dependent variable

The *Challenger* launch decision

Engineers did not consider successes, only failures; selection on the dependent variable

All flights, chronological order			
Damage?	Temp (F)	Damage?	Temp (F)
No	66	No	78
Yes	70	No	67
No	69	**Yes**	53
No	68	No	67
No	67	No	75
No	72	No	70
No	73	No	81
No	70	No	76
Yes	57	**Yes**	79
Yes	63	No	76
Yes	70	**Yes**	58

Other problems?

The *Challenger* launch decision

Engineers did not consider successes, only failures; selection on the dependent variable

All flights, chronological order			
Damage?	Temp (F)	Damage?	Temp (F)
No	66	No	78
Yes	70	No	67
No	69	**Yes**	53
No	68	No	67
No	67	No	75
No	72	No	70
No	73	No	81
No	70	No	76
Yes	57	**Yes**	79
Yes	63	No	76
Yes	70	**Yes**	58

Other problems? Why sort by launch number?

The *Challenger* launch decision

O-ring damage pre-Challenger, by temperature at launch

Damage?	Temp (F)	Damage?	Temp (F)
Yes	53	Yes	70
Yes	57	No	70
Yes	58	No	70
Yes	63	No	72
No	66	No	73
No	67	No	75
No	67	No	76
No	67	No	76
No	68	No	78
No	69	Yes	79
Yes	70	No	81

The *Challenger* launch decision

O-ring damage pre-Challenger, by temperature at launch			
Damage?	Temp (F)	Damage?	Temp (F)
Yes	53	Yes	70
Yes	57	No	70
Yes	58	No	70
Yes	63	No	72
No	66	No	73
No	67	No	75
No	67	No	76
No	67	No	76
No	68	No	78
No	69	Yes	79
Yes	70	No	81

The evidence begins to speak for itself.

The *Challenger* launch decision

O-ring damage pre-Challenger, by temperature at launch			
Damage?	Temp (F)	Damage?	Temp (F)
Yes	53	Yes	70
Yes	57	No	70
Yes	58	No	70
Yes	63	No	72
No	66	No	73
No	67	No	75
No	67	No	76
No	67	No	76
No	68	No	78
No	69	Yes	79
Yes	70	No	81

The evidence begins to speak for itself.

What if engineers had made this table before the launch?

The *Challenger* launch decision

Why didn't NASA make the right decision?

The *Challenger* launch decision

Why didn't NASA make the right decision?

Many answers in the literature:
bureaucratic politics; group think; bounded rationality, etc

The *Challenger* launch decision

Why didn't NASA make the right decision?

Many answers in the literature:
bureaucratic politics; group think; bounded rationality, etc

But Edward Tufte thinks it may have been a matter of presentation & modeling:

The *Challenger* launch decision

Why didn't NASA make the right decision?

Many answers in the literature:
bureaucratic politics; group think; bounded rationality, etc

But Edward Tufte thinks it may have been a matter of presentation & modeling:

- Never made the right tables or graphics
- Selected only failure data
- Never considered a simple statistical model

What do you think? How would you approach the data?

The *Challenger* launch decision

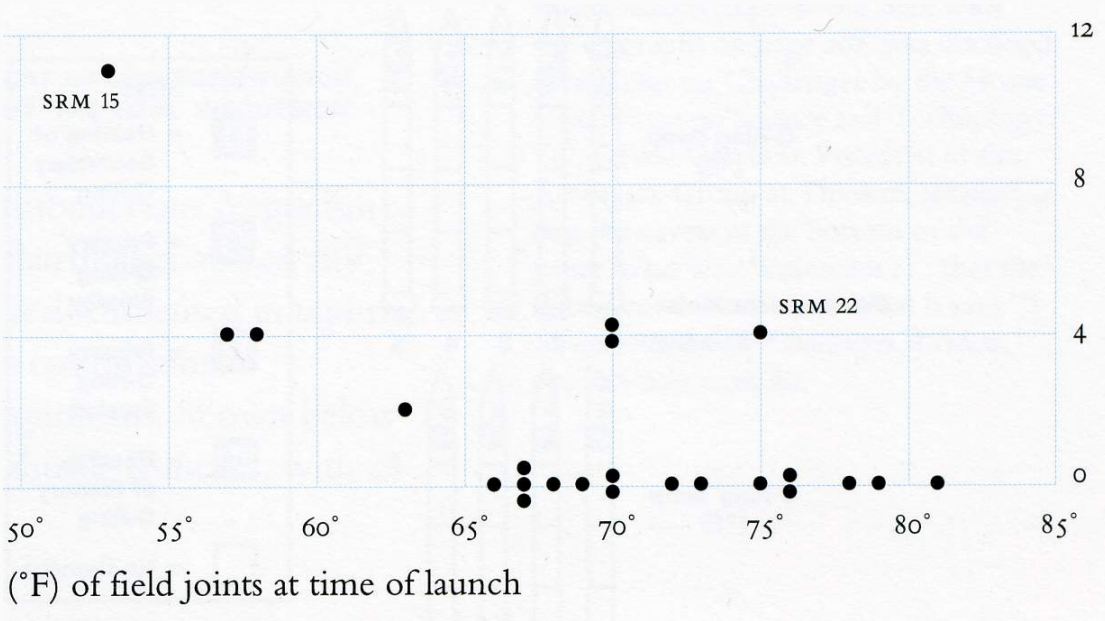
How about a scatterplot? Better for seeing relationships than a table.

Vertical axis is an O-ring damage index (due to Tufte, who made the plot)

The *Challenger* launch decision

How about a scatterplot? Better for seeing relationships than a table.

Vertical axis is an O-ring damage index (due to Tufte, who made the plot)

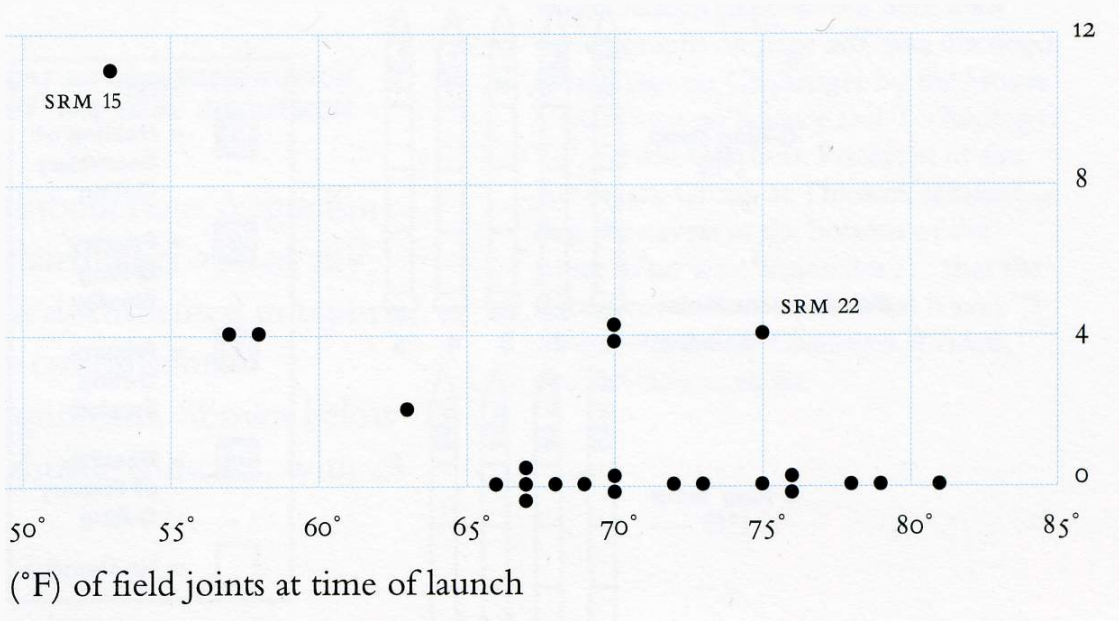


Suspicious.

The *Challenger* launch decision

How about a scatterplot? Better for seeing relationships than a table.

Vertical axis is an O-ring damage index (due to Tufte, who made the plot)



Suspicious. What was the forecast temperature for launch?

The *Challenger* launch decision

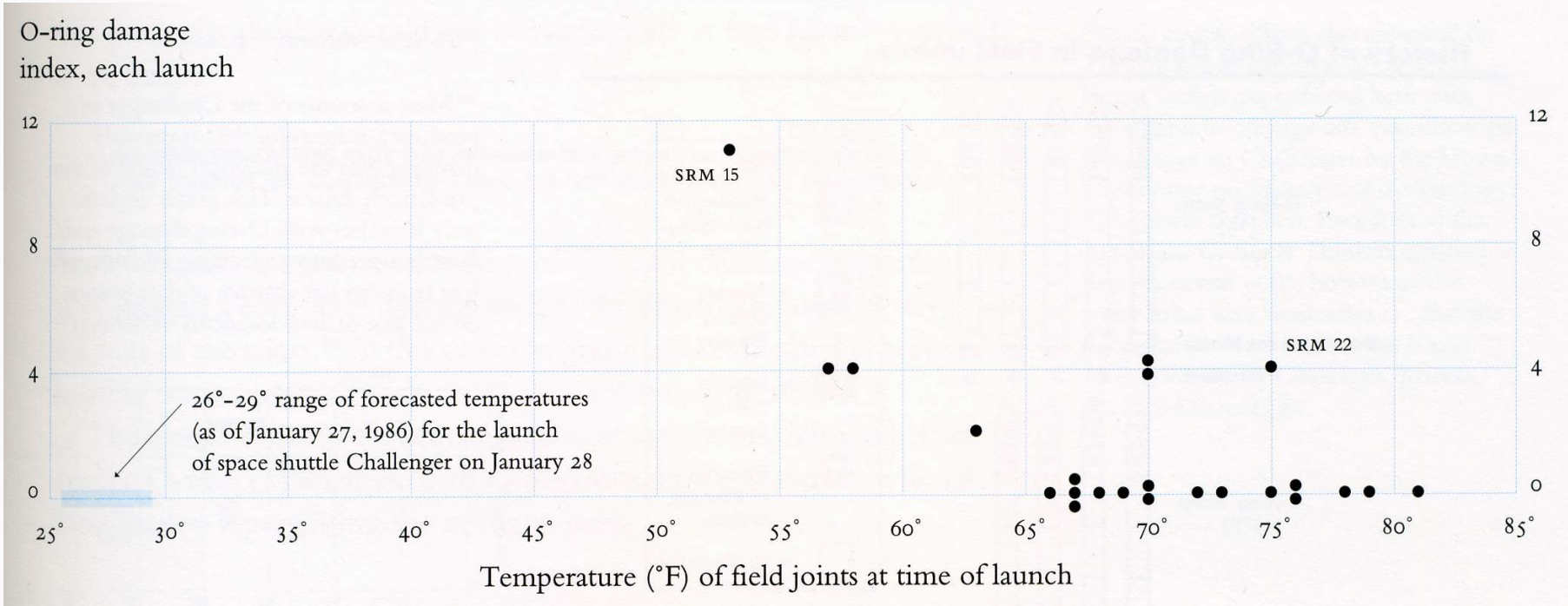
What was the forecast temperature for launch?

The *Challenger* launch decision

What was the forecast temperature for launch? 26 to 29 °F!

The *Challenger* launch decision

What was the forecast temperature for launch? 26 to 29 °F!



The shuttle was launched in unprecedented cold

The *Challenger* launch decision

Imagine you are the analyst making the launch recommendation.

You've made the scatterplot above. What would you add to it?

Put another way, what do you is the first question you expect from your boss?

The *Challenger* launch decision

Imagine you are the analyst making the launch recommendation.

You've made the scatterplot above. What would you add to it?

Put another way, what do you is the first question you expect from your boss?

"What's the chance of failure at 26 °F?"

The *Challenger* launch decision

Imagine you are the analyst making the launch recommendation.

You've made the scatterplot above. What would you add to it?

Put another way, what do you is the first question you expect from your boss?

"What's the chance of failure at 26 °F?"

The scatterplot suggests the answer is "high", but that's vague

The *Challenger* launch decision

Imagine you are the analyst making the launch recommendation.

You've made the scatterplot above. What would you add to it?

Put another way, what do you is the first question you expect from your boss?

"What's the chance of failure at 26 °F?"

The scatterplot suggests the answer is "high", but that's vague

But what if the next launch is at 58 °F? Or 67 °F?

The *Challenger* launch decision

Imagine you are the analyst making the launch recommendation.

You've made the scatterplot above. What would you add to it?

Put another way, what do you is the first question you expect from your boss?

"What's the chance of failure at 26 °F?"

The scatterplot suggests the answer is "high", but that's vague

But what if the next launch is at 58 °F? Or 67 °F?

Clearly, we want a more precise way to state the probability of failure

We need a *model*, and a way to convey that model to the public.

The *Challenger* launch decision

Model the probability of O-ring damage as a function of temperature

We can use a statistical tool called “logit” for this purpose

The model is nonlinear: $\Pr(\text{damage}) = (1 - \exp(-\beta_0 - \beta_1 \text{temperature}))^{-1}$

The *Challenger* launch decision

Model the probability of O-ring damage as a function of temperature

We can use a statistical tool called “logit” for this purpose

The model is nonlinear: $\Pr(\text{damage}) = (1 - \exp(-\beta_0 - \beta_1 \text{temperature}))^{-1}$

R gives us this lovely logit output...

The *Challenger* launch decision

Model the probability of O-ring damage as a function of temperature

We can use a statistical tool called "logit" for this purpose

The model is nonlinear: $\Pr(\text{damage}) = (1 - \exp(-\beta_0 - \beta_1 \text{temperature}))^{-1}$

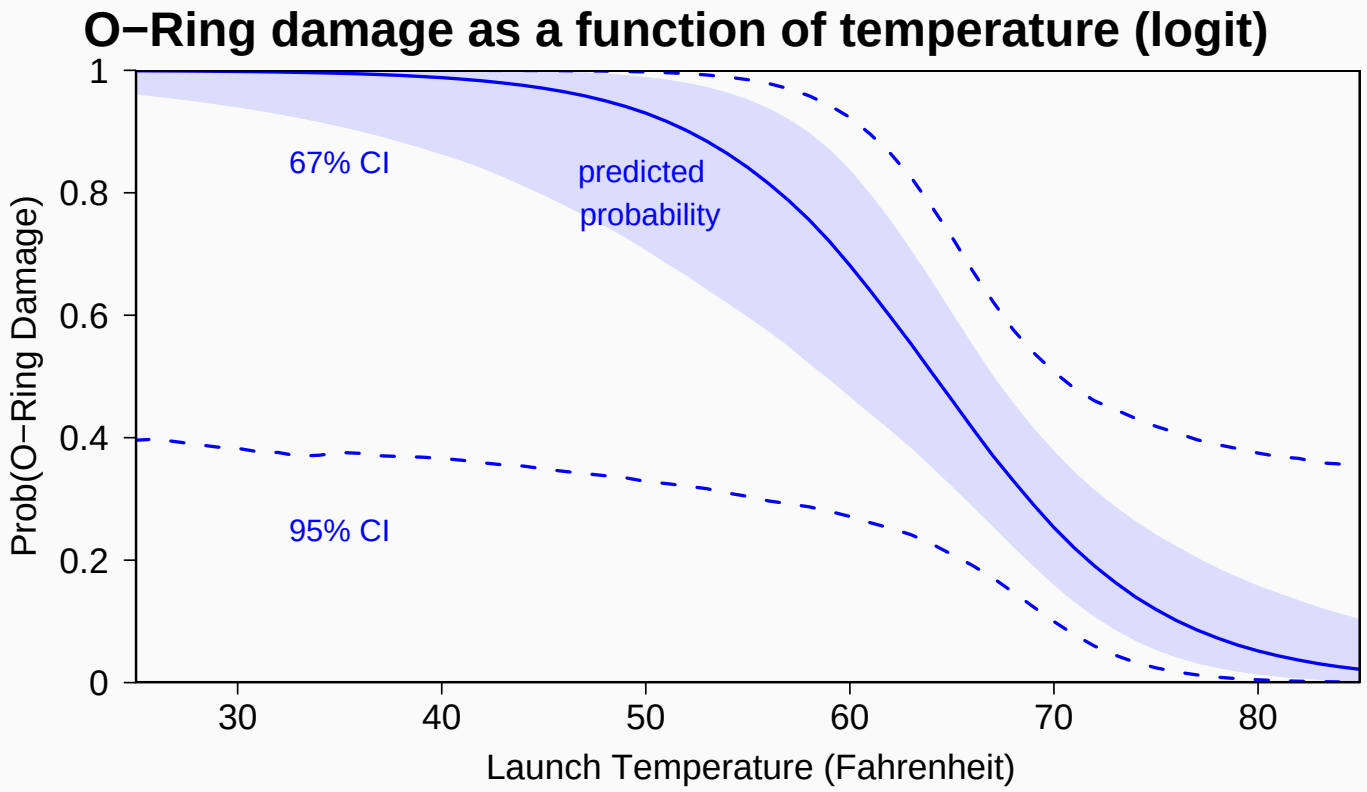
R gives us this lovely logit output...

Variable	est.	s.e.	*p*
Temperature (F)	-0.18	0.09	0.047
Constant	11.9	6.34	0.062
N	22		
log-likelihood	-10.9		

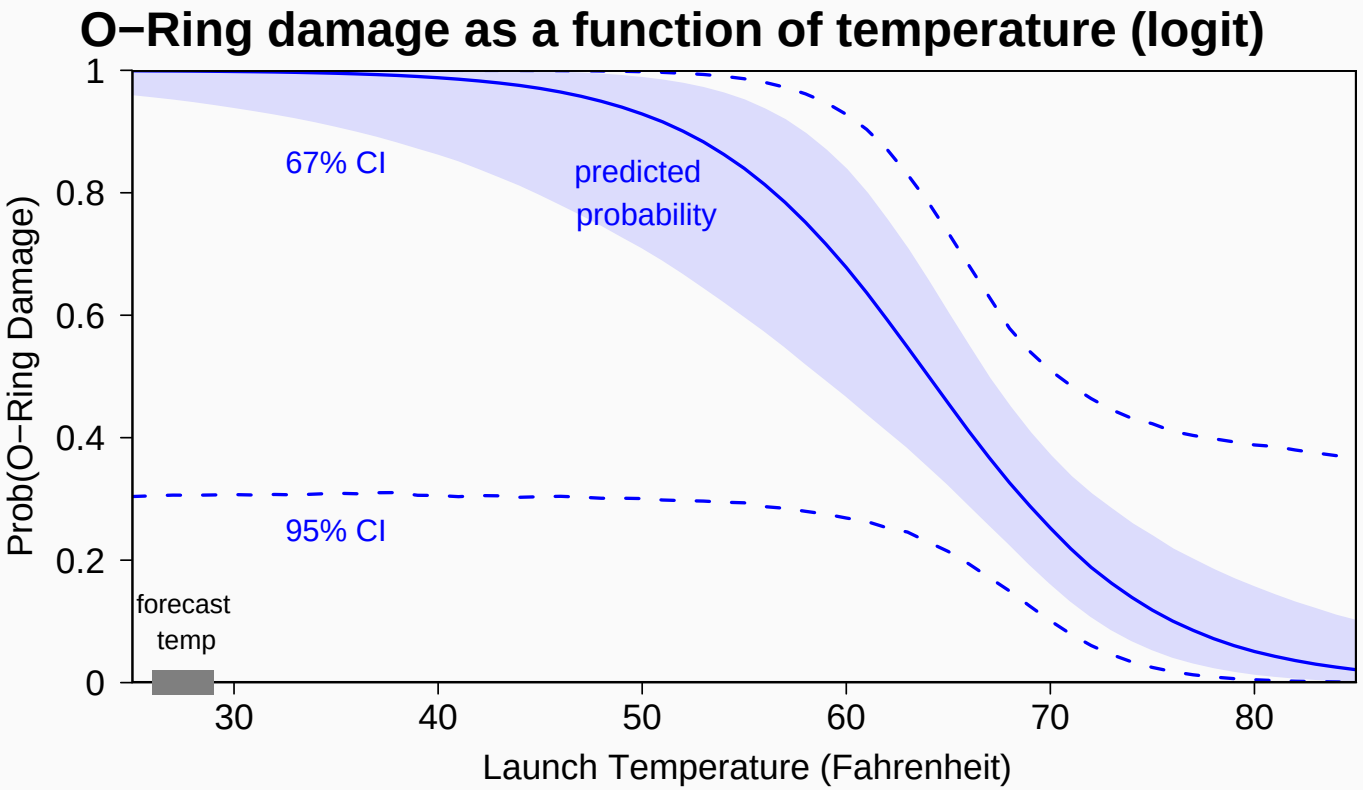
which most social scientists read as "a statistically significant negative relationship b/w temperature and probability of damage"

But that's pretty vague too.

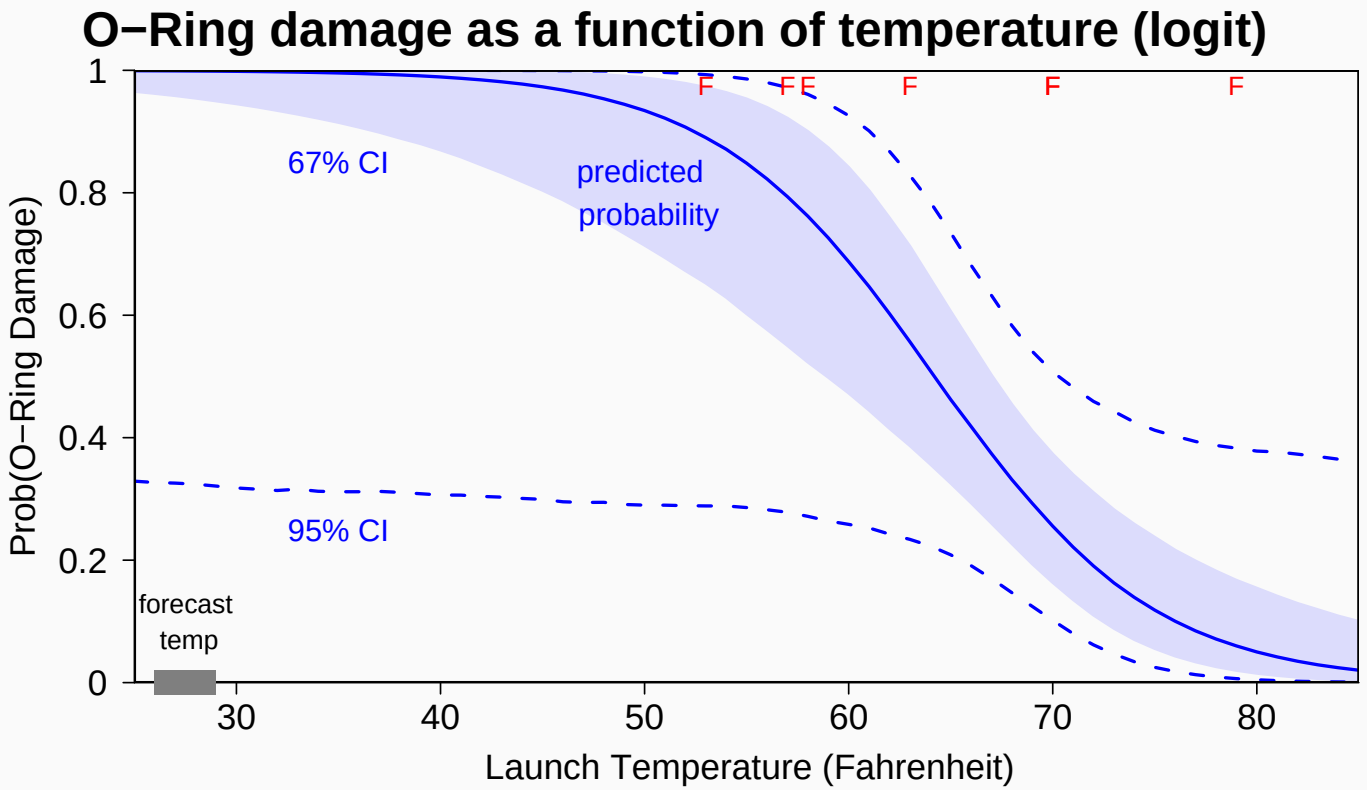
Is there a more persuasive/clear/useful way to present these results?



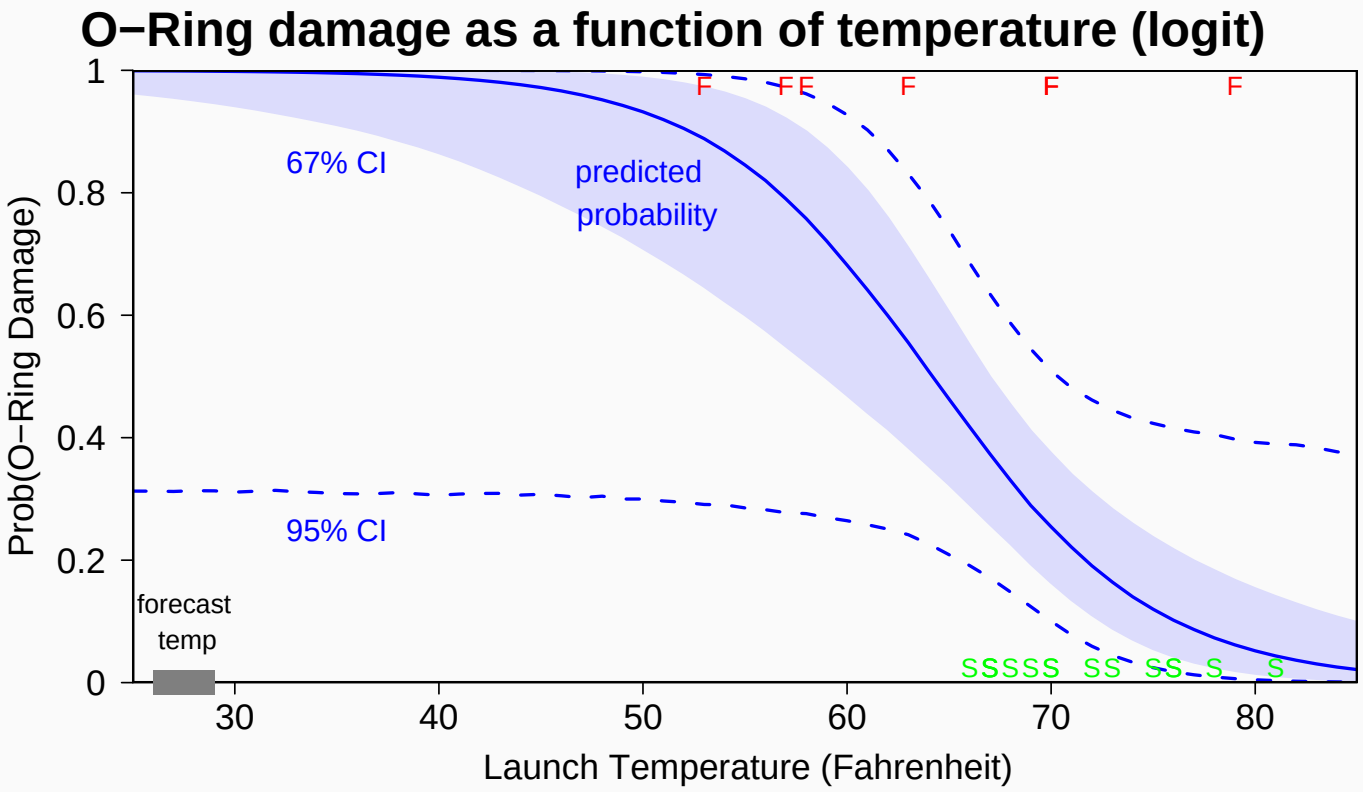
A picture clearly shows non-linear model predictions *and* uncertainty



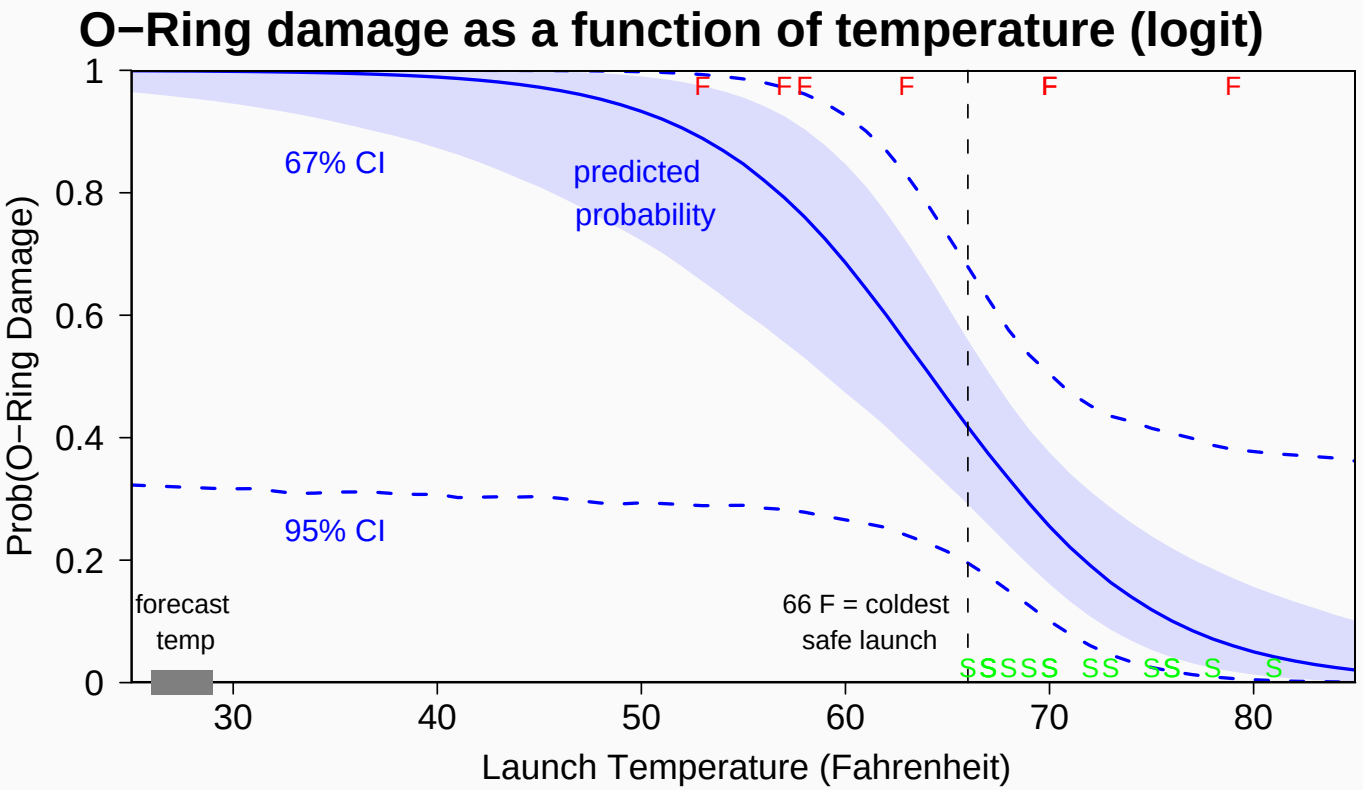
And gives a more precise sense of how foolhardy launching at 29 F is.



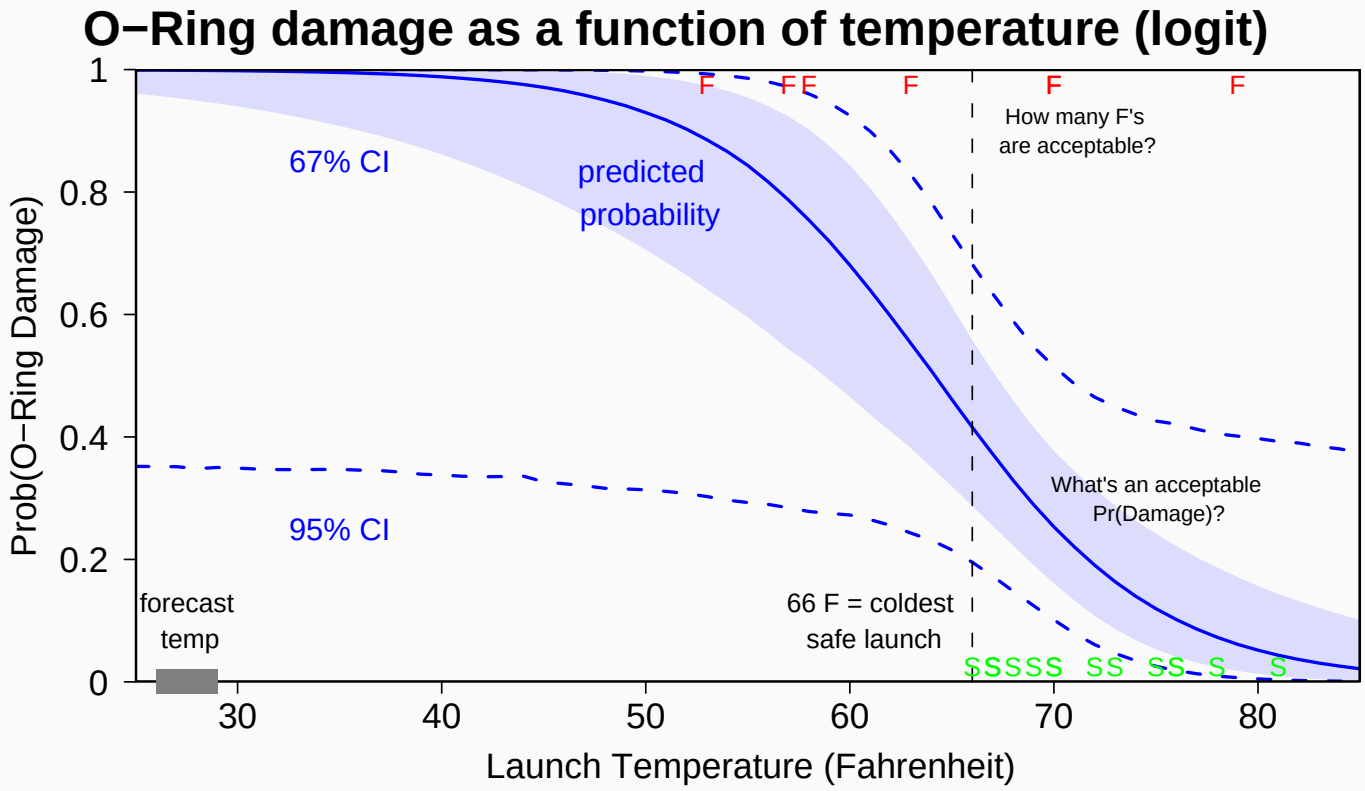
It's also good to show the data giving rise to the model.



Remembering that the Failures are only meaningful compared to Successes



Looking just at the data might show that launches under 66 F likely O-ring failures.



This inference is based on an unstated model.

In a hearing, Richard Feynmann dramatically showed O-rings lose resilence when cold by dropping one in his ice water.

Experiment cut thru weeks of technical gibberish concealing flaws in the O-ring

But it shouldn't have taken a Nobel laureate:
any scientist with a year of statistical training could have used the launch record to reach the same conclusion

And it would take no more than a single graphic to show the result

Outline of this Course

References

- Drew Conway, “The Data Science Venn Diagram”, http://drewconway.com/zia/2013/3/26/the-data-science-venn-diagram, CC-BY-NC
- R Logo. Tobias Wolf. http://developer.r-project.org/Logo/Rlogo.pdf CC-SA
- `http://commons.wikimedia.org/wiki/File:Under_Construction.jpegCC-BY-SA`
- Challenger example inspired by Edward Tufte, *The Visual Display of Quantitative Information*
- Idea for using the Challenger example in this course from Christopher Adolph, “Introduction to the Course and R”, *POLS/CSSS 221: Advanced Quantitative Political Methodology*, Spring 2014. <http://faculty.washington.edu/cadolph/503/topic1.pw.pdf>