# Thinking and Confidence

Ned Augenblick
*UC Berkeley*

Matt Backus
*UC Berkeley*

Don A. Moore
*UC Berkeley*

Andrew Little
*UC Berkeley*

October 2023

# Overprecision

- *Overprecision* is excessive certainty in the accuracy of one's beliefs.
- Most pervasive/robust form of overconfidence (Moore, Tenney, and Haran, 2015)

# Overprecision

- *Overprecision* is excessive certainty in the accuracy of one's beliefs.
- Most pervasive/robust form of overconfidence (Moore, Tenney, and Haran, 2015)

- Bazillions of experiments (ref: Don)
- Chief financial officers only achieve 33 percent hit rates inside their 80 percent confidence intervals (Ben-David, Graham, and Harvey, 2013)
- Overconfidence is endemic to deep neural networks (Thulasidasan et al., 2019), Bayesian machine learning (Oelrich et al., 2020), and large language models (OpenAI, 2023)
- Survey of professional forecasters: Chief economists forecast with 53 percent confidence but are only 28 percent accurate (Campbell and Moore, 2023)

# Overprecision

- What causes overprecision?
- Are people just bad at understanding uncertainty at a fundamental level?
- Is there some natural tendancy to always believe the world is more certain than it is?
- This paper: new explanation with theory, experiment evidence, empirical data

# Overview

- Start with an observation:
- People make assumptions to understand complex systems (i.e., create a “model”)

# Overview

- Start with an observation:
- People make assumptions to understand complex systems (i.e., create a “model”)
- Models simplify by removing cognitively-challenging complicated scenarios

# Overview

- Start with an observation:
- People make assumptions to understand complex systems (i.e., create a "model")
- Models simplify by removing cognitively-challenging complicated scenarios

- Our hypothesis: this observation is intimately related to overprecision.

# Overview

- Start with an observation:
- People make assumptions to understand complex systems (i.e., create a "model")
- Models simplify by removing cognitively-challenging complicated scenarios

- Our hypothesis: this observation is intimately related to overprecision.

- The intuition: Simplified models ignore variation-inducing possibilities $\rightarrow$ people to disregard variation in their predictions $\rightarrow$ over-precise on average
- People are correct within the model constraints of their model (*within-model uncertainty*), but don't appreciate uncertainty outside their model (*across-model uncertainty*)

# Overview

- Example in social science: to make empirical progress, we make assumptions
  - Structure of the relationship between variables
  - Form of the error term
  - Stationarity, etc.

# Overview

- Example in social science: to make empirical progress, we make assumptions
  - Structure of the relationship between variables
  - Form of the error term
  - Stationarity, etc.
- We construct CI, which are correct conditional on assumptions (*within* the model)
- What are the correct unconditional CIs?
- Depends on chance model is wrong and CI conditional on being wrong...
- This across-model uncertainty is very hard to deal with (unknownable?), even for us
- Hypothesis: People broadly don't account for it.

# Outline

1. Overview
2. Theory
3. Experiment
4. Initial Empirics
5. Conclusion

# Setup

- People (indexed by $i$) make predictions about RV $Y$ given initial data: $x \in \mathcal{X}$

# Setup

- People (indexed by $i$) make predictions about RV $Y$ given initial data: $x \in \mathcal{X}$
- "Model uncertainty": DGP $m$ drawn from RV $M$

# Setup

- People (indexed by $i$) make predictions about RV $Y$ given initial data: $x \in \mathcal{X}$
- "Model uncertainty": DGP $m$ drawn from RV $M$
- For this talk, everyone sees same data so drop $|x$ (ie incorporate $x$ into prior)
- Prior over $y$: $\pi(y)$
- Prior over $m$: $\pi(m)$.

# Behavioral Assumption

- Main assumption: individual $i$ simplifies problem by *focusing* on subset of models.

# Behavioral Assumption

- Main assumption: individual $i$ simplifies problem by *focusing* on subset of models.

- Start with assumption that each person uses one model (easily relaxed)
- Assume person focuses on model $m$ with probability $f(m)$.
- Person's beliefs if focused on model $m$:

$$\hat{\pi}_i(y) = \pi(y|m)$$

# Notation

- Some useful notation:

| | Bayesian | Individual $i$'s Perception |
|---|---|---|
| Expected Value of Y | $\mu \equiv \sum \pi(y) \cdot y$ | $\hat{\mu}_i \equiv \sum \hat{\pi}_i(y) \cdot y$ |
| Variance of Y | $\sigma^2 \equiv \sum \pi(y) \cdot (y - \mu)^2$ | $\hat{\sigma}_i^2 \equiv \sum \hat{\pi}_i(y) \cdot (y - \hat{\mu}_i)^2$ |
| MSE of $i$'s estimate | $\text{MSE} \equiv \sum \pi(y) \cdot (y - \hat{\mu}_i)^2$ | $\widehat{\text{MSE}}_i \equiv \sum \hat{\pi}_i(y) \cdot (y - \hat{\mu}_i)^2$ |

# Important Assumption

- Crucial Q: which models do people focus on?
- If people can focus on any model, then absolutely anything can happen
- We want some way to
  1) allow disagreement across people
  2) have people focus on "reasonable" models.

# Important Assumption

- Crucial Q: which models do people focus on?
- If people can focus on any model, then absolutely anything can happen
- We want some way to
  1) allow disagreement across people
  2) have people focus on "reasonable" models.

- Start with a **baseline assumption**: $f(m) = \pi(m)$
- That is, people focus on models in relation to likelihood that model is right.

# Important Assumption

- Crucial Q: which models do people focus on?
- If people can focus on any model, then absolutely anything can happen
- We want some way to
  1) allow disagreement across people
  2) have people focus on “reasonable” models.

- Start with a **baseline assumption**: $f(m) = \pi(m)$
- That is, people focus on models in relation to likelihood that model is right.

- In paper, we discuss a set of possibilities: highest likelihood (=> no disagreement), sticky models (focus on previously-likely models), focus on simplier models.
- Again, allowing people to focus on any model will allow anything to happen...
- ...but we show that unless people focus on models that happen to imply much more variance than the “average” model, the main statements will hold.

# Results

- Some basic “results” given the setup and baseline assumption:
- Disagreement: Different people can hold different beliefs given same info
- Wisdom of the crowds: The average of peoples’ predictions is unbiased
- Stronger wisdom of crowds: The average of peoples’ beliefs are well-calibrated

# Results

- What about variance?

# Results

- What about variance?
- People are right on average about the mean...
- ...but the average of variance estimate is biased below the true variance:

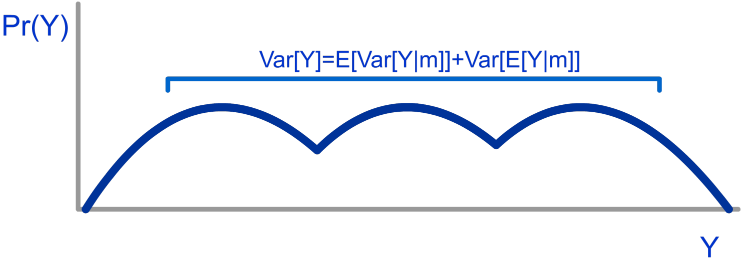
$$\underbrace{E_i[\hat{\sigma}_i^2]}_{\substack{\text{Mean of Perceived} \\ \text{Variance of Y}}} = \underbrace{\sigma^2}_{\substack{\text{True} \\ \text{Variance of Y}}} - \underbrace{\text{Var}_m[E[Y|m]]}_{\substack{\text{(+) Ignore Model} \\ \text{Uncertainty}}}$$

# Results

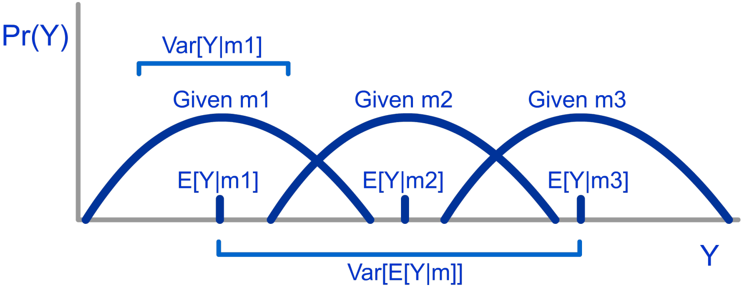
- What about variance?
- People are right on average about the mean...
- ...but the average of variance estimate is biased below the true variance:

$$\underbrace{E_i[\hat{\sigma}_i^2]}_{\substack{\text{Mean of Perceived} \\ \text{Variance of Y}}} = \underbrace{\sigma^2}_{\substack{\text{True} \\ \text{Variance of Y}}} - \underbrace{\text{Var}_m[E[Y|m]]}_{\substack{\text{(+) Ignore Model} \\ \text{Uncertainty}}}$$

# Results



Pr(Y)
Var[Y]=E[Var[Y|m]]+Var[E[Y|m]]
Y

Pr(Y)
Var[Y|m1]
Given m1
Given m2
Given m3
E[Y|m1]
E[Y|m2]
E[Y|m3]
Y
Var[E[Y|m]]

# Results

$$\underbrace{E_i[\hat{\sigma}_i^2]}_{\substack{\text{Mean of Perceived} \\ \text{Variance of Y}}} = \underbrace{\sigma^2}_{\substack{\text{True} \\ \text{Variance of Y}}} - \underbrace{\mathrm{Var}_m[E[Y|m]]}_{\substack{\text{(+) Ignore Model} \\ \text{Uncertainty}}}$$

- Prediction: As model uncertainty rises, overprecision rises.
- Aggressive prediction: overprecision connected 1-to-1 to across-model uncertainty.

# Results

- Aggressive: if the baseline assumption is correct, then:

$$\underbrace{\text{Var}[\hat{\mu}_i]}_{\substack{\text{Individual}\\ \text{Disagreement}}} = \underbrace{\text{Var}_m[E[Y|m]]}_{\substack{\text{Model}\\ \text{Uncertainty}}}$$

- Prediction: model uncertainty related to individual disagreement
- Aggressive Prediction: model uncertainty equal to disagreement
- ...but then:

$$\underbrace{E_i[\hat{\sigma}_i^2]}_{\substack{\text{Mean of Perceived}\\ \text{Variance of Y}}} = \underbrace{\sigma^2}_{\substack{\text{True}\\ \text{Variance of Y}}} - \underbrace{\text{Var}[\hat{\mu}_i]}_{\substack{\text{Individual}\\ \text{Disagreement}}}$$

# Results

- Aggressive: if the baseline assumption is correct, then:

$$\underbrace{\text{Var}[\hat{\mu}_i]}_{\substack{\text{Individual}\\ \text{Disagreement}}} = \underbrace{\text{Var}_m[E[Y|m]]}_{\substack{\text{Model}\\ \text{Uncertainty}}}$$

- Prediction: model uncertainty related to individual disagreement
- Aggressive Prediction: model uncertainty equal to disagreement
- ...but then:

$$\underbrace{E_i[\hat{\sigma}_i^2]}_{\substack{\text{Mean of Perceived}\\ \text{Variance of Y}}} = \underbrace{\sigma^2}_{\substack{\text{True}\\ \text{Variance of Y}}} - \underbrace{\text{Var}[\hat{\mu}_i]}_{\substack{\text{Individual}\\ \text{Disagreement}}}$$

- Prediction: problems with more mean disagreement will have more avg. overprecision
- Aggressive Prediction: Amount of variance misestimation equal to disagreement

# Results

- But, being right about variance does not imply calibration / correct precision.
- Example: if a person says $\hat{\mu}_i = 100$, $\hat{\sigma}^2 = 2$ but Bayesian says $\mu = 110$, $\sigma^2 = 2$

# Results

- But, being right about variance does not imply calibration / correct precision.
- Example: if a person says $\hat{\mu}_i = 100$, $\hat{\sigma}^2 = 2$ but Bayesian says $\mu = 110$, $\sigma^2 = 2$
- Person's variance estimate matches the Bayesian estimate...
- ...but person $\sim$95% confident answer in $[96, 104]$ but very unlikely $\rightarrow$ overprecision.

# Results

- But, being right about variance does not imply calibration / correct precision.
- Example: if a person says $\hat{\mu}_i = 100$, $\hat{\sigma}^2 = 2$ but Bayesian says $\mu = 110$, $\sigma^2 = 2$

- Person's variance estimate matches the Bayesian estimate...
- ...but person $\sim$95% confident answer in $[96, 104]$ but very unlikely $\rightarrow$ overprecision.

- This "being wrong about the mean" effect is captured in mean squared error (MSE).

# Results

- Their MSE is even more wrong. That is:

$$\underbrace{E_i[\hat{\text{MSE}}_i]}_{\substack{\text{Mean of Perceived}\\ \text{MSE of Estimate}}} = \underbrace{E_i[\text{MSE}]}_{\substack{\text{True Mean}\\ \text{MSE of Estimate}}} - \underbrace{\text{Var}_m[E[Y|m]]}_{\substack{\text{(+) Ignore Model}\\ \text{Uncertainty}}} - \underbrace{E[(\hat{\mu}_i - \mu)^2]}_{\substack{\text{(+) Ignore Mean}\\ \text{Is Wrong}}}$$

- Aggressive: if the baseline assumption is correct, then:

$$\underbrace{E_i[\hat{\text{MSE}}_i]}_{\substack{\text{Mean of Perceived}\\ \text{MSE of Estimate}}} = \underbrace{E_i[\text{MSE}]}_{\substack{\text{True Mean}\\ \text{MSE of Estimate}}} -2 \cdot \underbrace{\text{Var}[\hat{\mu}_i]}_{\substack{\text{Individual}\\ \text{Disagreement}}}$$

- Aggressive Prediction: Amount of MSE misestimation equal to 2 times disagreement

# Results

- Other results:

- Overprecision 1: People place too much prob on events that they see as most-likely
- Overprecision 2: People place too little prob on events that they see as least-likely

- Private Info Private info also leads to disagreement.
  Bias coefficients attenuated based on disagreement due to model focus

- Bayesian Learning On avg, Bayesians more confident / more accurate with more info
- Learning On avg, modelers more confident with more info. Accuracy unclear.
- Anti-Learning On avg, modelers less confident / more accurate when loosen assumptions

- Different model focusing Bias terms from above all remain. Is possible for underprecision but requires focus on model with very-above-average variance

# Outline

1. Overview
2. Theory
3. Experiment
4. Initial Empirics
5. Conclusion

# Setup

- Basic predictions we would like to test:
- People are overprecise (underestimate variance)
- People appreciate within-model uncertainty but not across-model uncertainty.
- Model uncertainty related to variance underestimates
- More aggressive: Model uncertainty related to disagreement. Therefore, disagreement related to variance underestimates

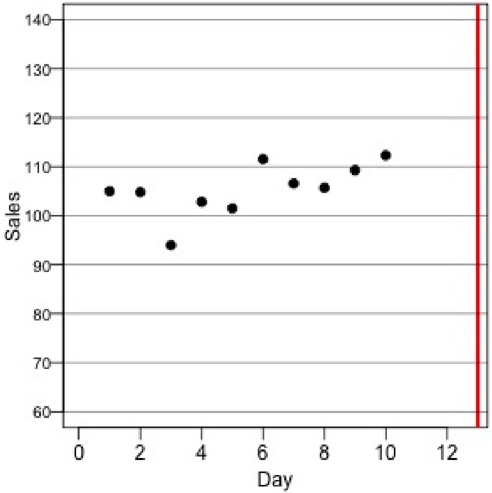
# Setup

- We really struggled to find a good experimental setting.
- Big tradeoff between 1) being understandable and 2) complete control of DGP
- To really test our model, need to:
  - Explain outcomes
  - Give prior over outcomes
  - Explain all models
  - Give prior over models
  - Explain distribution of data for all models
  - Give data
  - Get variance estimates from people
  - Know true answers
- But, this is hard to do without confusing people!

# Experimental Design

- Idea: present people with a scatter plot - “sales data”
- Generated with linear trend + normal noise
- Tell them to make a statistic prediction about future sales

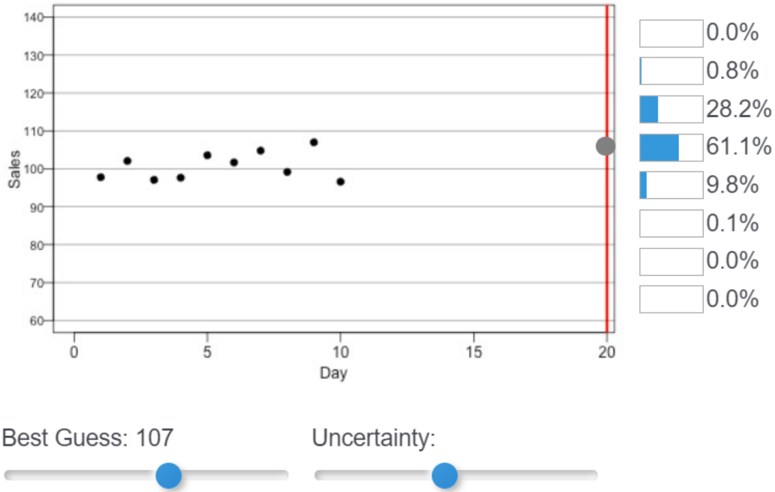
# Screenshot



# Experimental Design

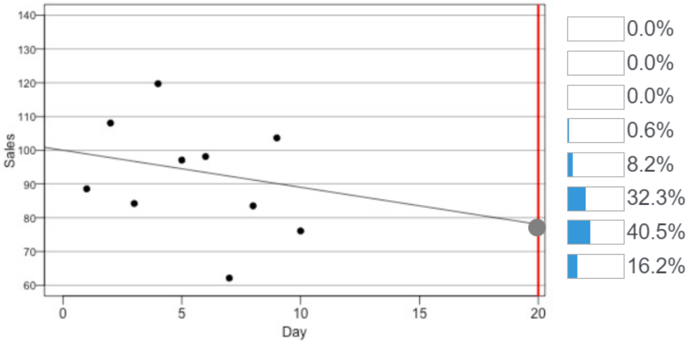
- Explain with many examples (with feedback) that trend is linear and randomly chosen
- (Ask lots of comprehension Qs - people seem to understand)
- One design choice to make easier for subjects: rather than asking people to fill out probabilities for each bin, have them use slider for mean and variance (truth=normal) and bins fill in automatically.

# Experimental Design



# Screenshot

Here is another example. Try your best to get the right answer - then click the button to see if you would have gotten close enough to the right answer to get a bonus.




Best Guess: 78

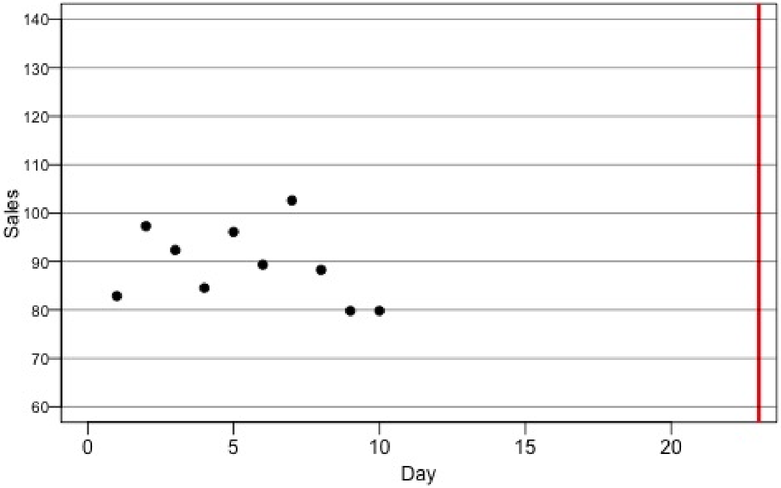
Uncertainty:

Click here to show the right answer on the sliders

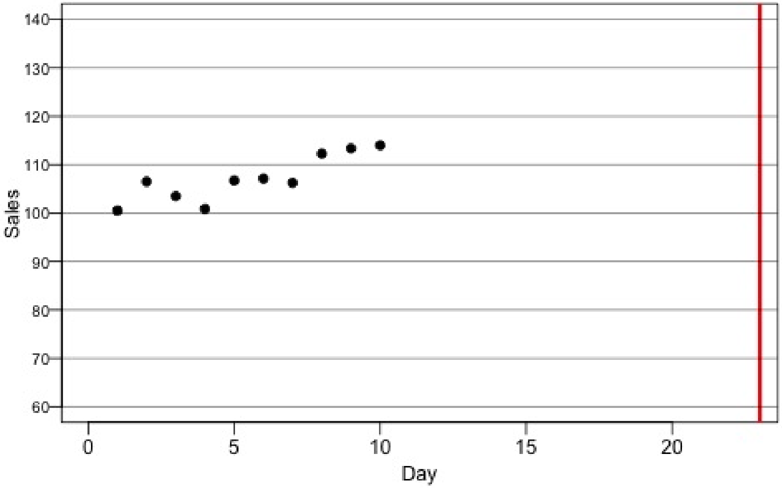
# Experimental Design

- Show people $\sim 6$ plots with various (unknown) true trends and levels of variance

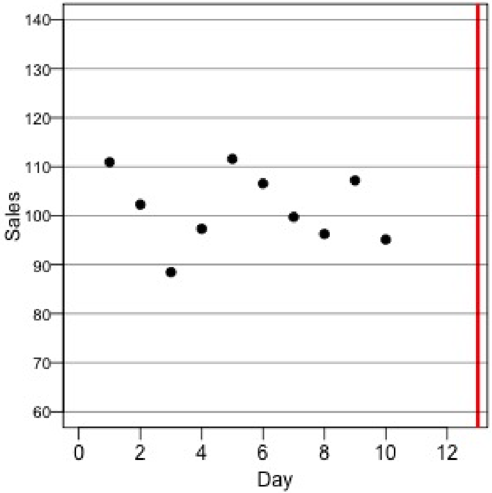
# Experimental Design



# Experimental Design

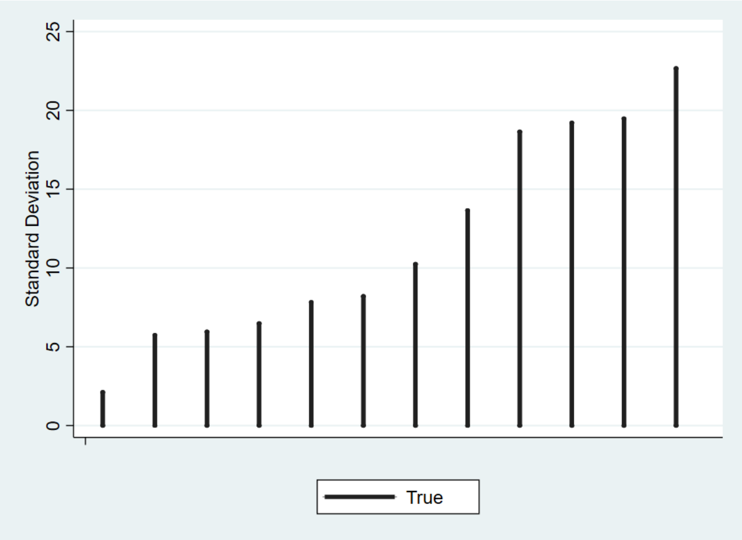


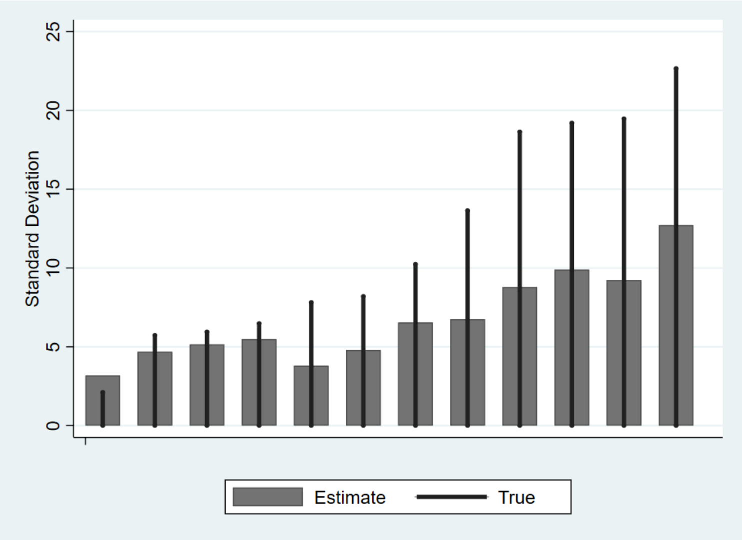
# Experimental Design



# Results: Overprecision

- People are overprecise (as with many other tasks in the lit)
- Avg true variance is: 133. People's estimate is: 88 (66%)
- People's most likely bin has more probability than reality. True most likely bin has less.

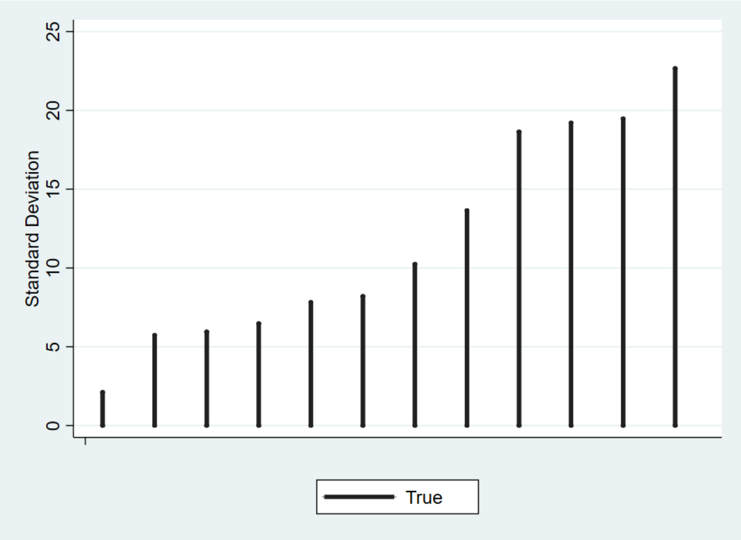

Standard Deviation
0
5
10
15
20
25
True



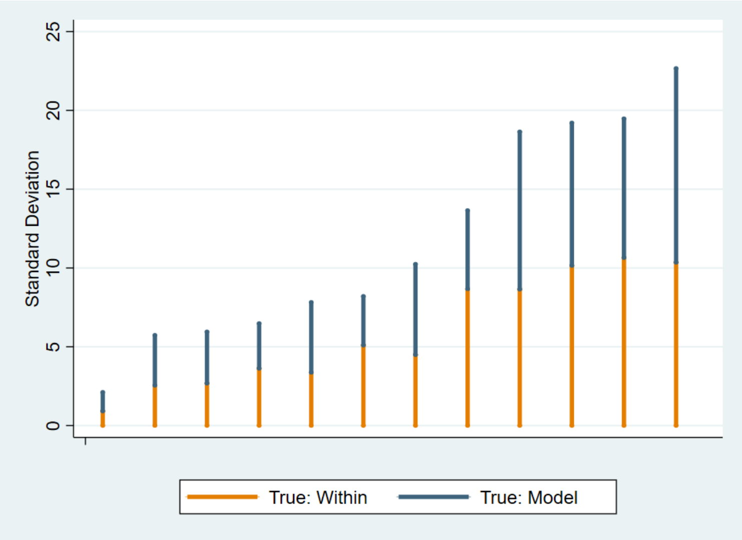
Standard Deviation
0
5
10
15
20
25
Estimate
True

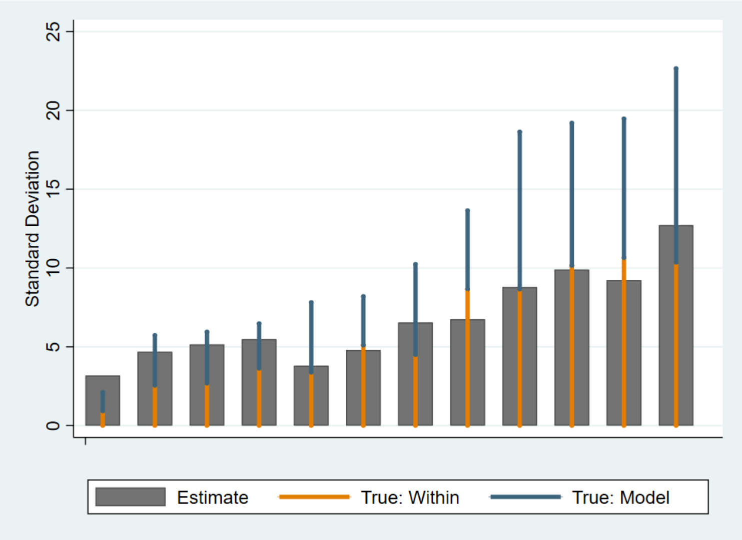
# Results

- But theory is more specific
- Idea: people will focus on a subset (one?) of possible lines (ie make an assumption/focus on a model) and answer as if that is the true model.
- They will then account for uncertainty within model (ie given that line) but not for the across-model uncertainty (ie uncertainty about lines) This is what leads to overprecision.



25
20
15
10
5
0
Standard Deviation
True

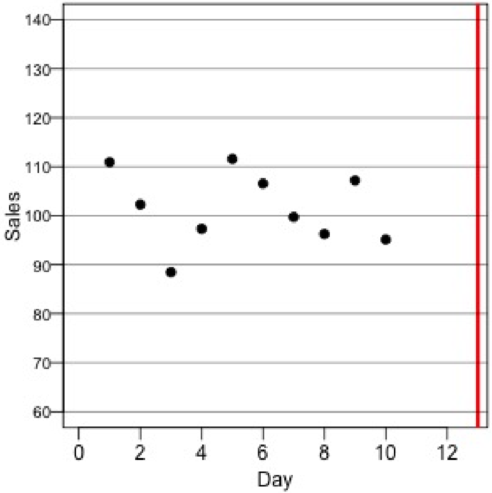

25
20
15
10
5
0
Standard Deviation
True: Within
True: Model


Standard Deviation
0
5
10
15
20
25
Estimate
True: Within
True: Model

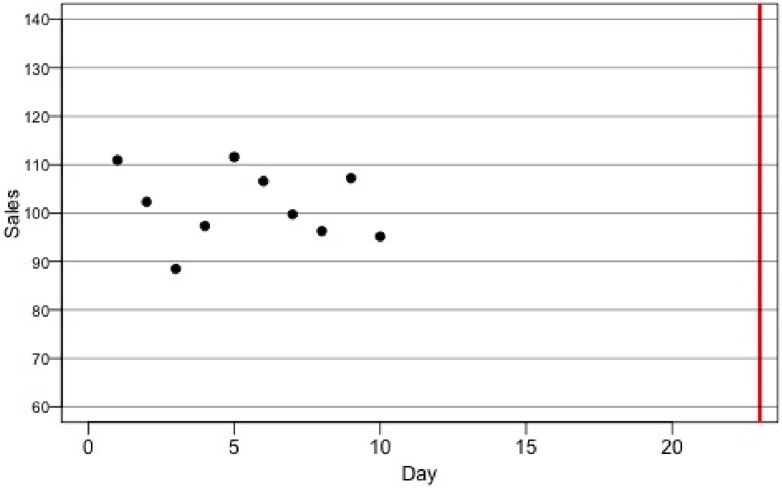
# Results

- But: within-model and across-model uncertainty very correlated - as error variance rises, increases both...
- Would like to vary these independently
- One way: vary whether prediction is for a closer or farther date
- Idea: as prediction date gets farther away, the impact of across-model uncertainty rises. But, the within-model uncertainty remains constant.

# Far vs. Close

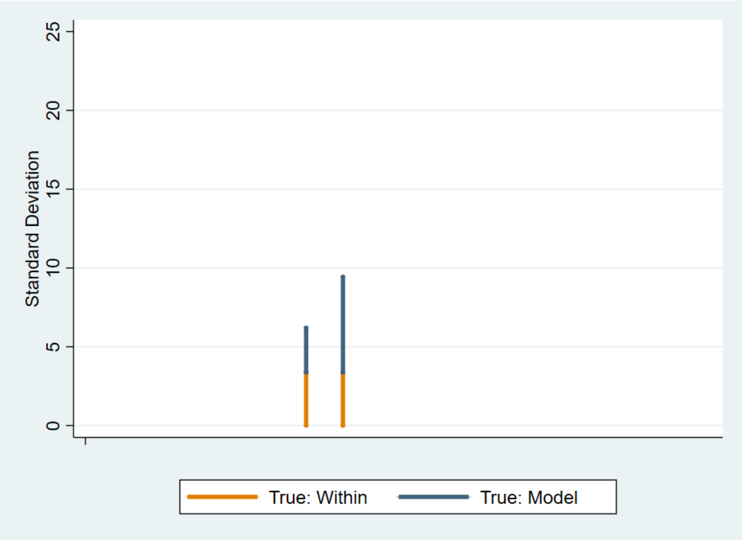


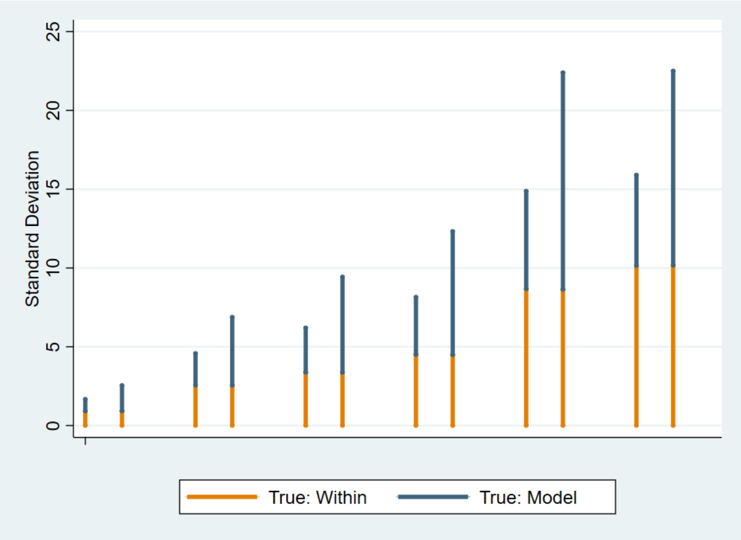
# Far vs. Close



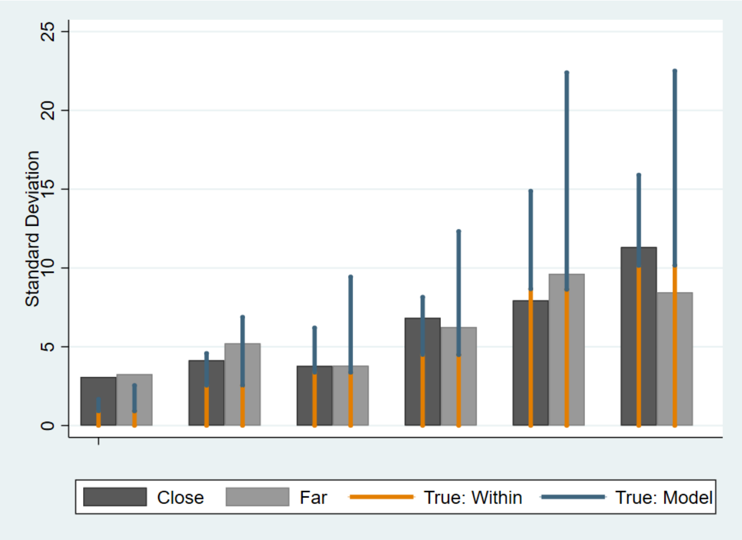
# Results

- Prediction: people will not adjust to increased model uncertainty (close vs. far).

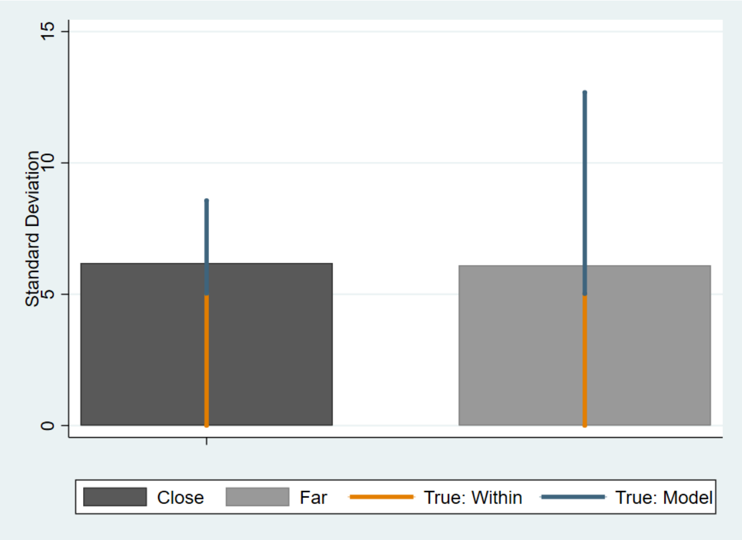

Standard Deviation
0
5
10
15
20
25
True: Within
True: Model



25
20
15
10
5
0
Standard Deviation
True: Within
True: Model



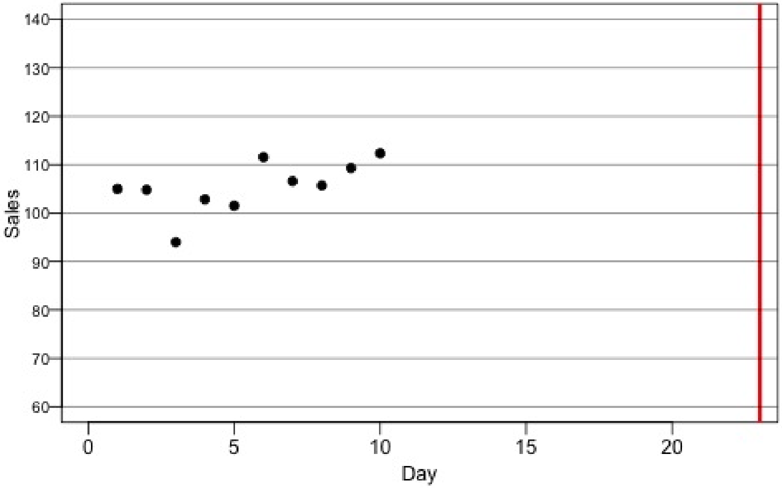
Standard Deviation
0
5
10
15
20
25
Close
Far
True: Within
True: Model


Standard Deviation
15
10
5
0
Close
Far
True: Within
True: Model

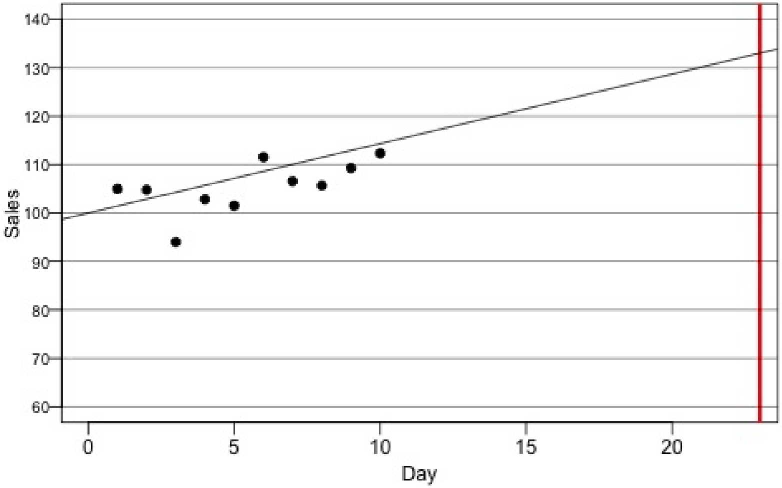
# Results

- Alternatively, we could remove all model uncertainty providing the line. This leaves within-model uncertainty.

# Line vs. No Line

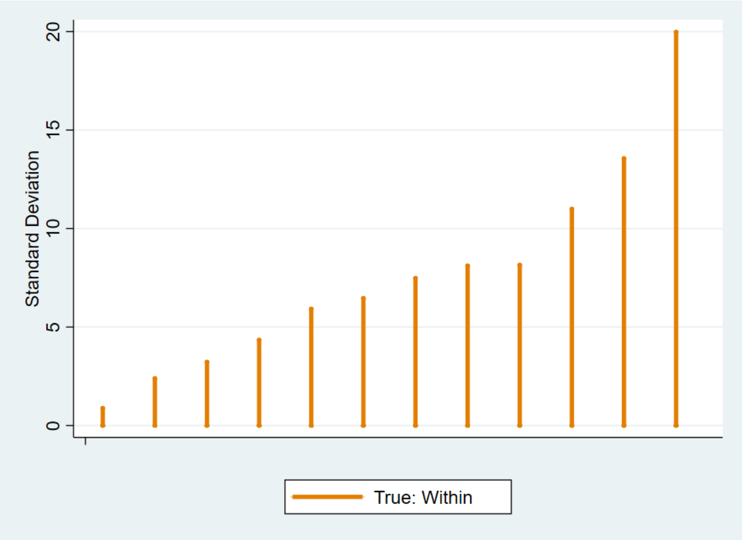


# Line vs. No Line

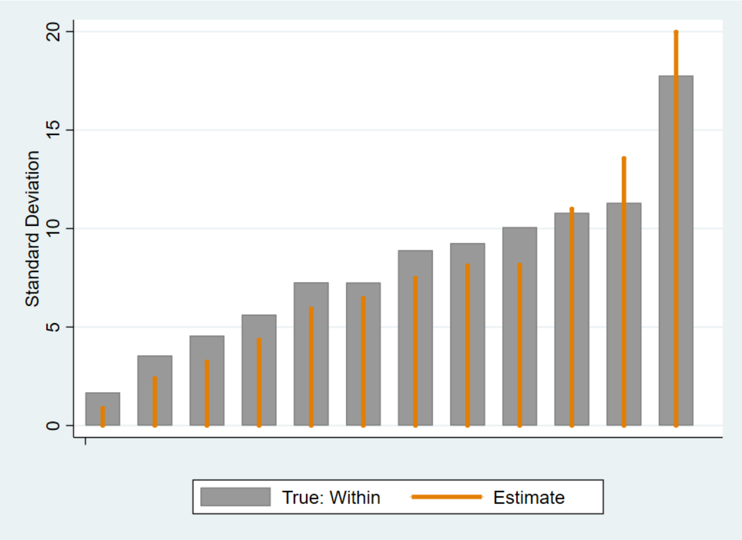


# Results

- Prediction: people will be generally correct on these problems.



Standard Deviation
0
5
10
15
20
True: Within


Standard Deviation
0
5
10
15
20
True: Within
Estimate

# Results

Putting it all together, we can estimate how much people account for both types of uncertainty (standard errors clustered at both subject and problem):

# Results

Putting it all together, we can estimate how much people account for both types of uncertainty (standard errors clustered at both subject and problem):

| DV: Var Estimate | Bayesian | Our Model | Experiment |
|---|---|---|---|
| Within-model Variance | 1 | 1 | 0.82 |
| | | | (0.05) |
| Across-Model Variance | 1 | 0 | 0.06 |
| | | | (0.04) |
| Constant | 0 | 0 | 29.80 |
| | | | (4.77) |
| $R^2$ | | | 0.33 |
| $N$ | | | 1164 |

That is, people account for much (82%) of the within-model uncertainty, but account very little for the model uncertainty...

# Results

Another way: Can we predict error?

$$\underbrace{E_i[\hat{\sigma}_i^2]}_{\substack{\text{Mean of Perceived}\\ \text{Variance of Y}}} = \underbrace{\sigma^2}_{\substack{\text{True}\\ \text{Variance of Y}}} - \underbrace{\mathrm{Var}_m[E[Y|m]]}_{\substack{\text{(+) Ignore Model}\\ \text{Uncertainty}}}$$

$$\underbrace{\sigma^2 - \hat{\sigma}_i^2}_{\substack{\text{Error in}\\ \text{Variance Est of Y}}} = \underbrace{\mathrm{Var}_m[E[Y|m]]}_{\substack{\text{Model}\\ \text{Uncertainty}}} + \epsilon$$

| DV: Error Var Estimate | Bayesian | Our Model | Experiment |
|---|---|---|---|
| Across-Model Variance | 0 | 1 | 0.68 |
| | | | (0.08) |
| Constant | 0 | 0 | -45.047 |
| | | | (7.93) |
| $R^2$ | | | 0.47 |
| $N$ | | | 1164 |

# Results

- At this point, we have only been accounting for the mistakes in variance.
- Recall that we should also account for the wrong-mean effect that increases MSE.
- Given this, people are even more wrong...
- Account for $\sim$44% of MSE on average

# Results

Adding to the regression suggests that people don't respond to estimate being wrong (as might be expected)?

| DV: MSE Estimate | Bayesian | Our Model | Experiment |
|---|---|---|---|
| Within-model Variance | 1 | 1 | 0.82 |
| | | | (0.05) |
| Across-Model Variance | 1 | 0 | 0.059 |
| | | | (0.04) |
| Being-Wrong Variance | 1 | 0 | 0.019 |
| | | | (0.02) |
| Constant | 0 | 0 | 28.80 |
| | | | (4.80) |
| $R^2$ | | | 0.336 |
| $N$ | | | 1164 |

# Results: What about other predictions?

- How do the implications of our baseline assumption look?
- Wisdom of Crowds:
  Absolute deviation of average of mean estimates from truth averages $\sim 4\%$
- Wisdom of Crowds Aggressive:
  Predicted Var of average distribution averages 90% true distribution

# Results: What about other predictions?

- Prediction: dispersion in means connected to model uncertainty (corr=0.55 (p=.005))
- Key prediction: overprecision related to disagreement in means
- Aggressive prediction: underestimate of variance rises 1-to-1 with disagreement
- Aggressive prediction: underestimate of MSE rises 2-to-1 with disagreement

# Results

| DV: Error In Var Estimate | Bayesian | Our Model | Data |
|---|---|---|---|
| Disagreement $(-i)$ | 0 | 1 | 2.49 |
| | | | (0.77) |
| Constant | 0 | 0 | -6.51 |
| | | | (37.52) |
| $R^2$ | | | 0.18 |
| $N$ | | | 570 |

| DV: Error In MSE Estimate | Bayesian | Our Model | Data |
|---|---|---|---|
| Disagreement $(-i)$ | 0 | 2 | 3.48 |
| | | | (0.70) |
| Constant | 0 | 0 | 28.80 |
| | | | (42.82) |
| $R^2$ | | | 0.18 |
| $N$ | | | 570 |

# Summary

- Data seems fairly consistent with our model (certainly better than Bayesianism)
- People are overprecise
- Overprecision related to not appreciating 1) model uncertainty and 2) being wrong
- ...But, people seem decent at within-model uncertainty
- Model uncertainty related to disagreement across people
- Overestimates vary positively with disagreement

# Quick note

- Can run study with more naturalistic data (real stock market trends)
- Benefit: feels less abstract
- Problem: unknown model uncertainty
- Find overprecision
- Next step: far/near

# Outline

# SPF

- Look at the survey of professional forecasters
- Making forecasts about a variety of economic outcomes in the future
- As with model, likely have same information but different models/assumptions

# SPF

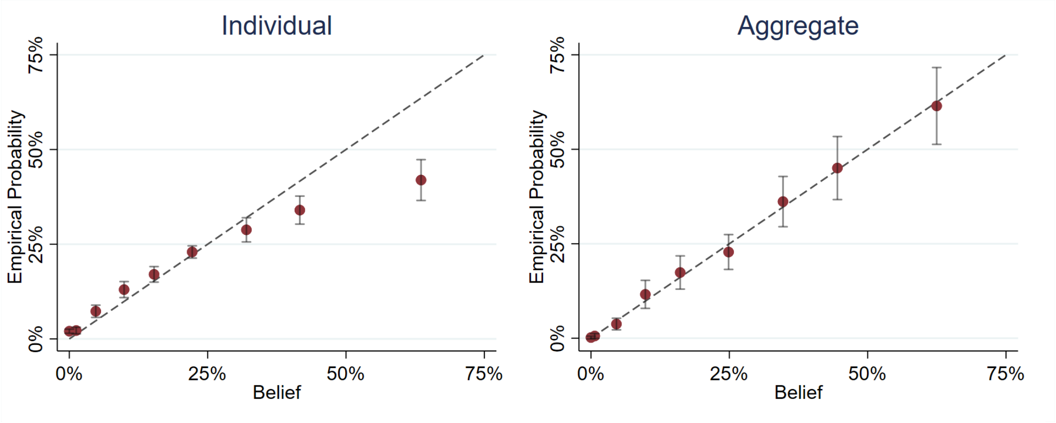
- They are overprecise:
- Pr in highest bin averages 53.5%. But, bin actually occurs 36.4% of the time. ($p<0.001$).
- "Border bins" in which people place 0% occur 4.9% of the time

# SPF

- They are overprecise:
- Pr in highest bin averages 53.5%. But, bin actually occurs 36.4% of the time. ($p<0.001$).
- "Border bins" in which people place 0% occur 4.9% of the time

- But, strong wisdom of the crowds:
- Averaging over all people:
- Highest bins 46.0% and occur 46.5% ($p=0.90$)
- "Border bins" in which people place 0% occur 0.5% of the time

# SPF



Individual
Aggregate
Empirical Probability
0%
25%
50%
75%
Belief

## SPF

- In MSE terms:
- Individuals
  Avg($\hat{\text{MSE}}_i$)=.81 but
  Avg($\text{MSE}_i$)=1.44 (p=0.004)
- Next prediction: Individual's error is larger when mean disagreement is larger

# SPF

| DV: Error in MSE Estimate | Bayesian | Our Model | Data |
|---|---|---|---|
| Disagreement ($-i$) | 0 | 2 | 1.785 |
| | | | (0.306) |
| Constant | 0 | 0 | -0.1972 |
| | | | (0.1631) |
| $R^2$ | | | 0.094 |
| Observations | | | 38,449 |

# Outline

# Summary

- As before...
- People are overprecise
- Our idea: comes from assumptions/model focus
- Model predicts overprecision and average...
- ...but also predicts how overprecision depends on within-model uncertainty, model uncertainty, and error from wrong model.
- ...and predicts overestimates with rise with disagreement in means
- We find broad evidence in support but