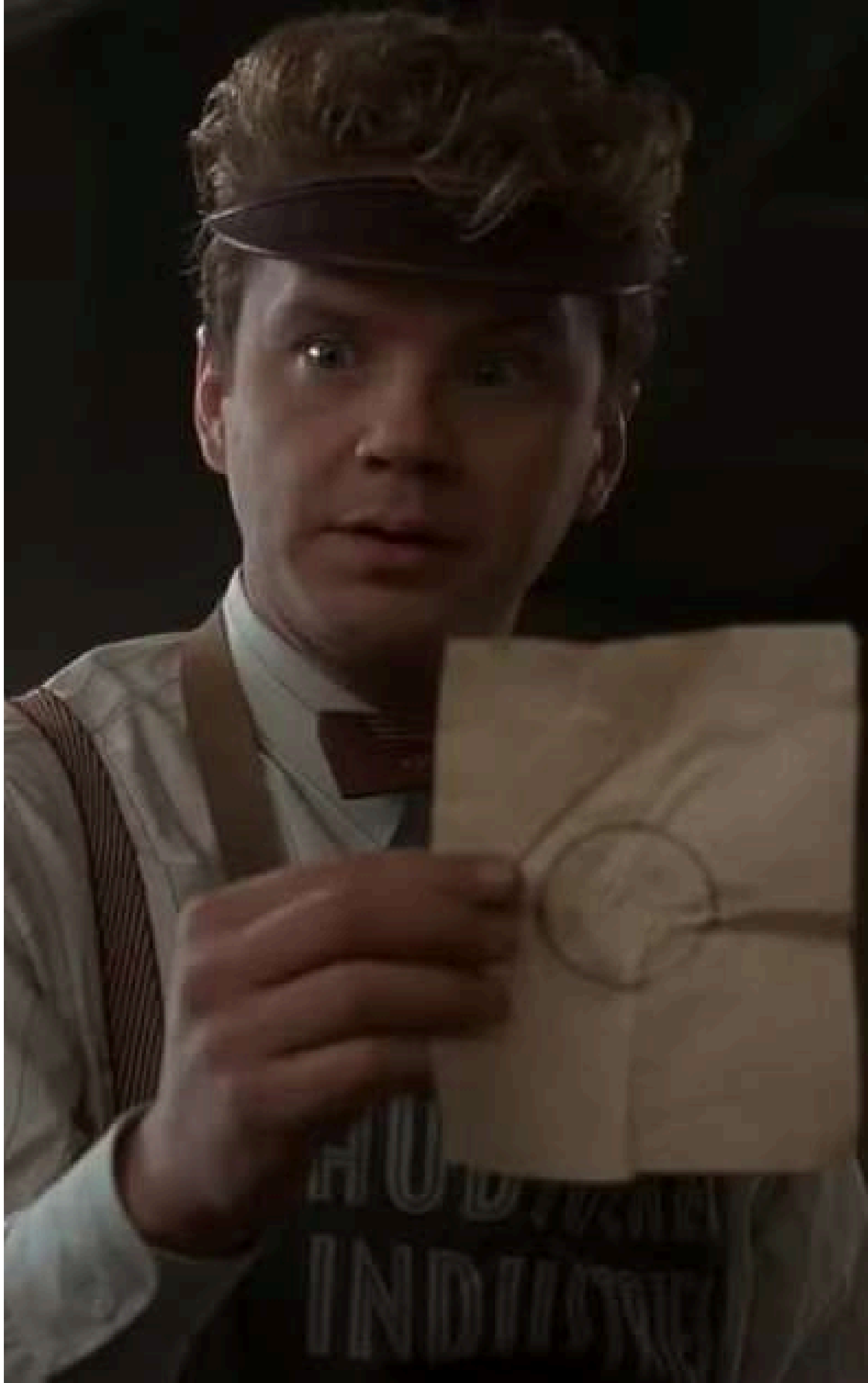
Agenda

Parsimony & overfitting

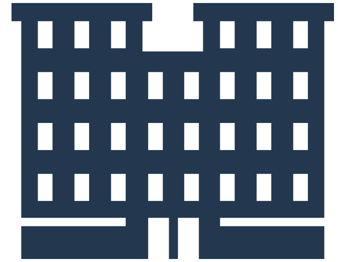
1. Administrative
2. Parsimony & Occam's Razor
3. Overfitting vs. underfitting
4. Test & training data
5. Information criteria
6. ***Hands on:*** *Comparing information criteria in R*

Parsimony & Occam's razor



How many buildings?

OCCAM'S RAZOR



***M*₁:** Four buildings

***M*₂:** Five buildings

$$\frac{\Pr(M_1|D)}{\Pr(M_2|D)} = \frac{\Pr(M_1)}{\Pr(M_2)} \frac{\Pr(D|M_1)}{\Pr(D|M_2)}$$

A-priori justification | Simpler models are easier to interpret or more compelling *on their own*

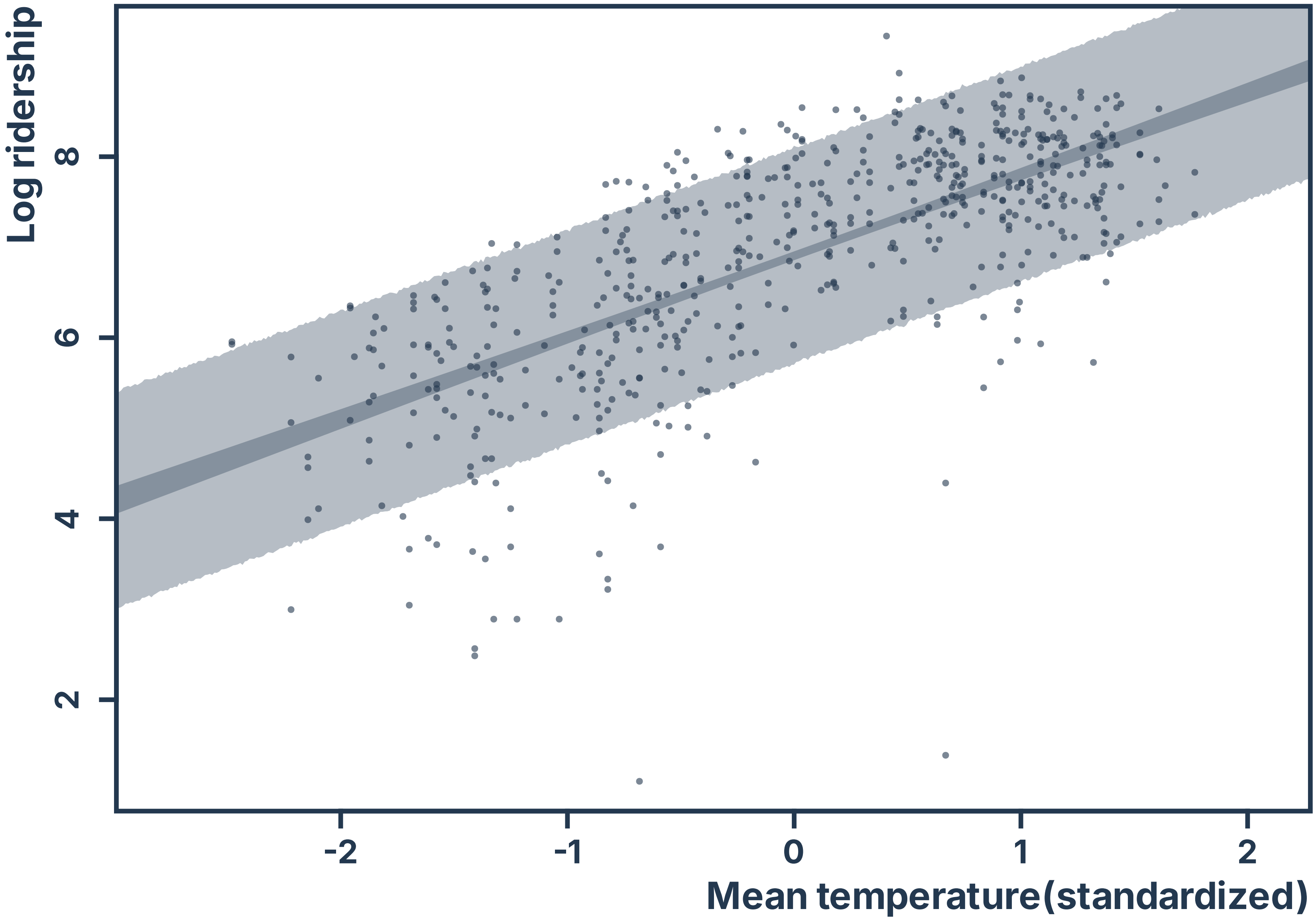
$$\frac{\Pr(M_1)}{\Pr(M_2)} > 1$$

Model likelihood justification | Simpler models rely *less on coincidence* to produce specific data

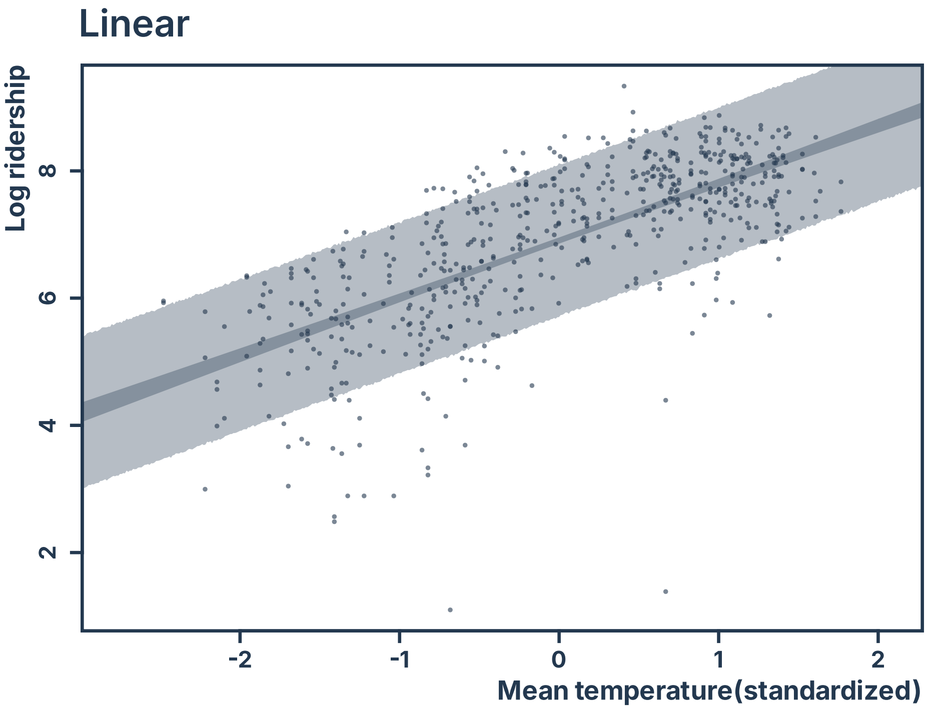
$$\frac{\Pr(D|M_1)}{\Pr(D|M_2)} > 1$$

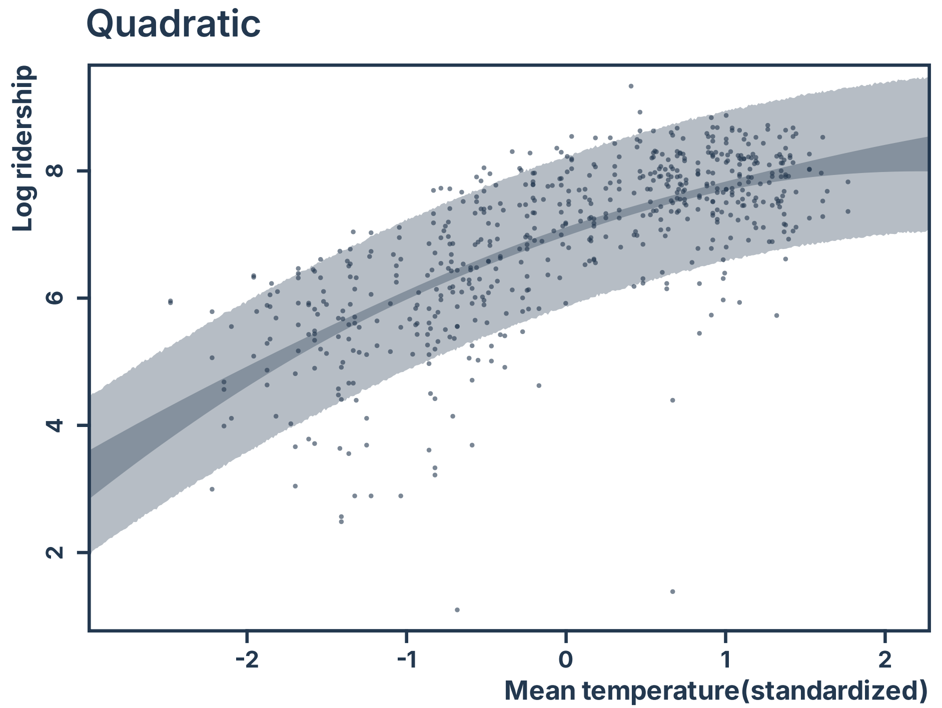
Assessing fit

ASSESSING FIT



ASSESSING FIT





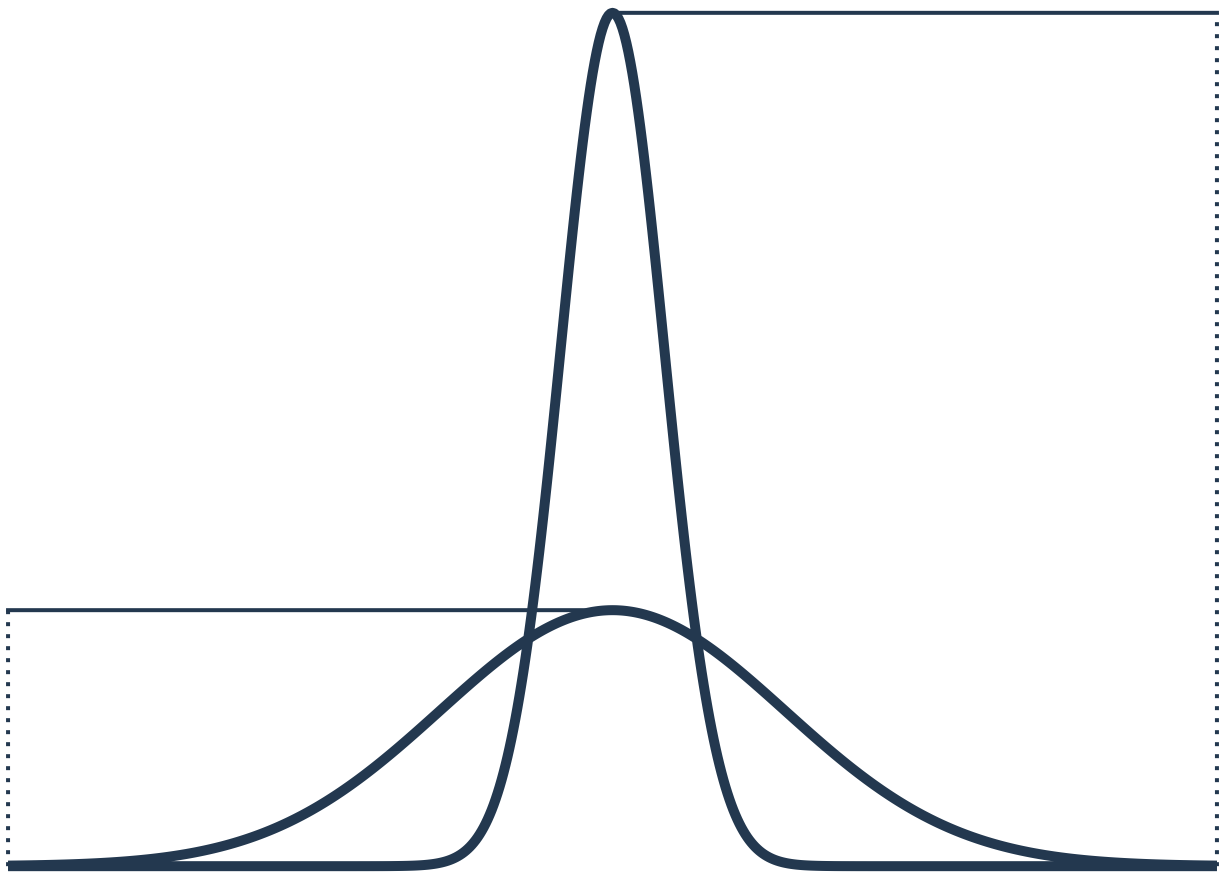
$$\mu_i = \alpha + \beta_1 T_i$$

$$\mu_i = \alpha + \beta_1 T_i + \beta_2 T_i^2$$

A quadratic model seems like it might be a better fit.

But how can we measure that?

$$\text{Prob}(\theta|\text{data}) = \frac{\text{Prob}(\text{data}|\theta)\text{Prob}(\theta)}{\text{Prob}(\text{data})}$$

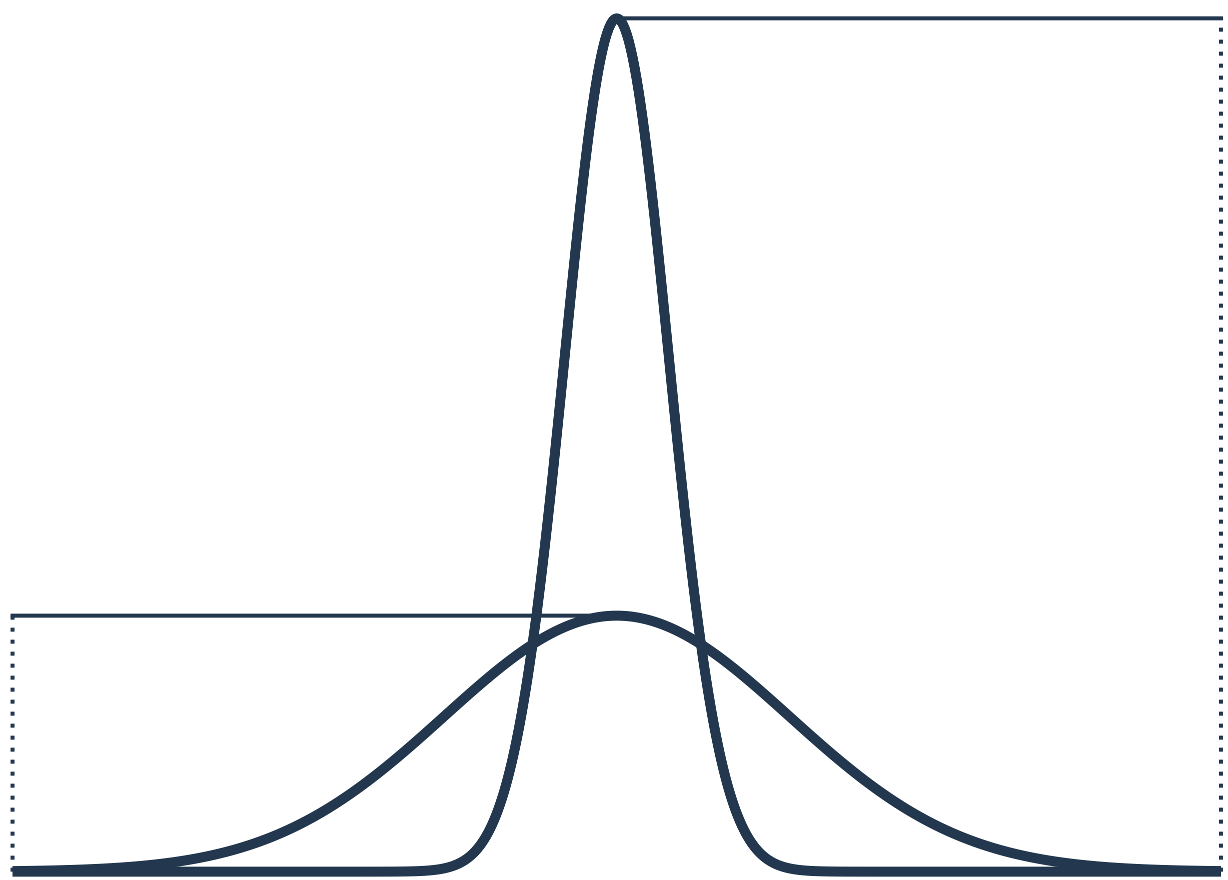


ASSESSING FIT: DEVIANCE

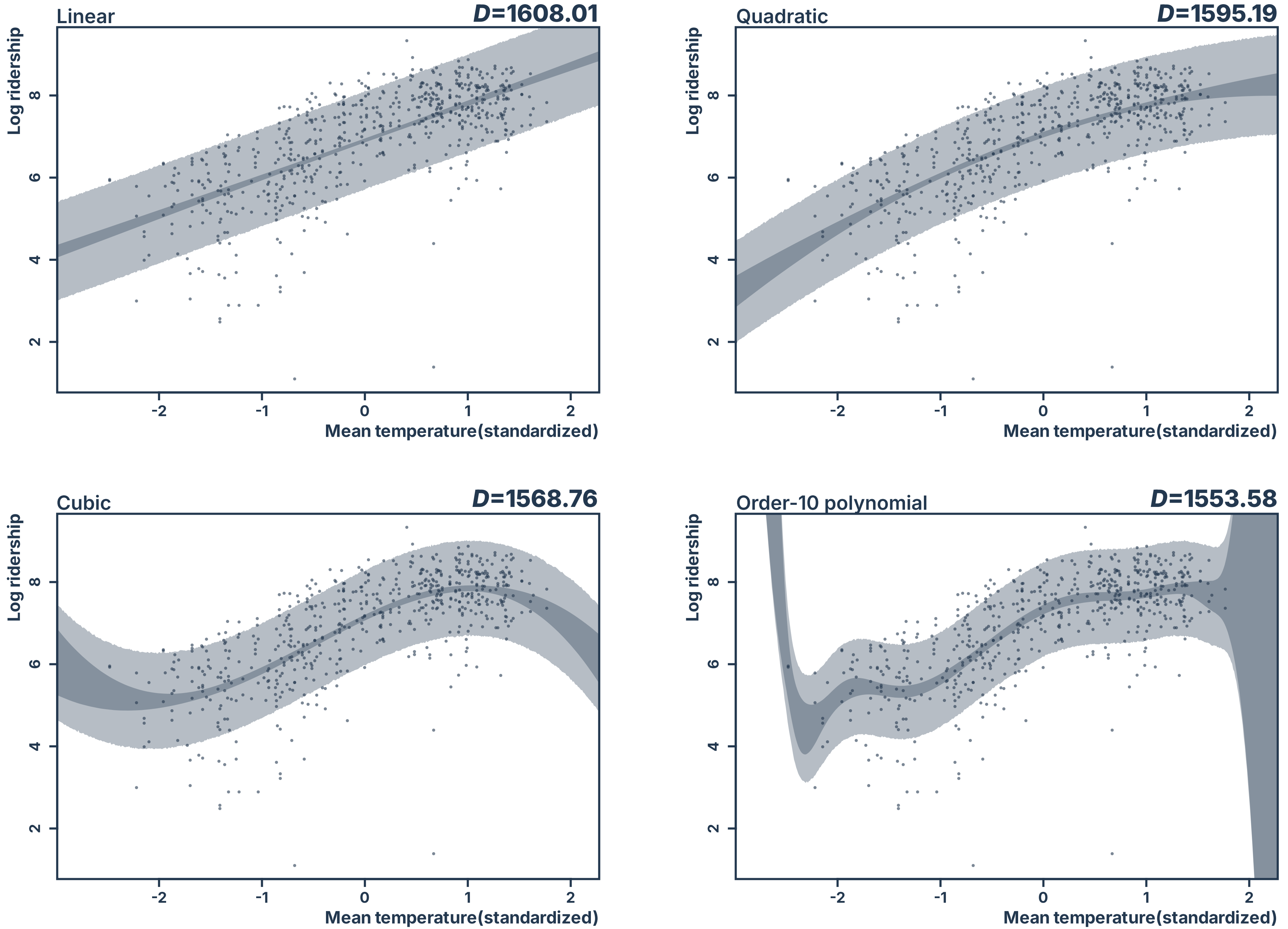
Deviance **(D)* is minus two times the *log likelihood* of the data, given the model and a point estimate for the model parameters ($\hat{\theta}$):**

$$D = -2 \log \left(\mathrm{Prob}(\mathrm{data} | \hat{\theta}) \right)$$

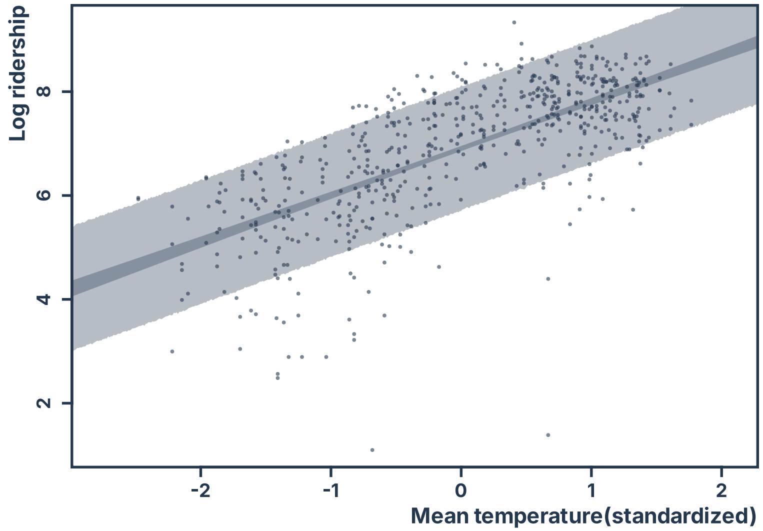
* Note: a common definition of deviance requires a comparison to a 'saturated' model. For clarity, we use this simpler definition.



ASSESSING FIT: DEVIANCE

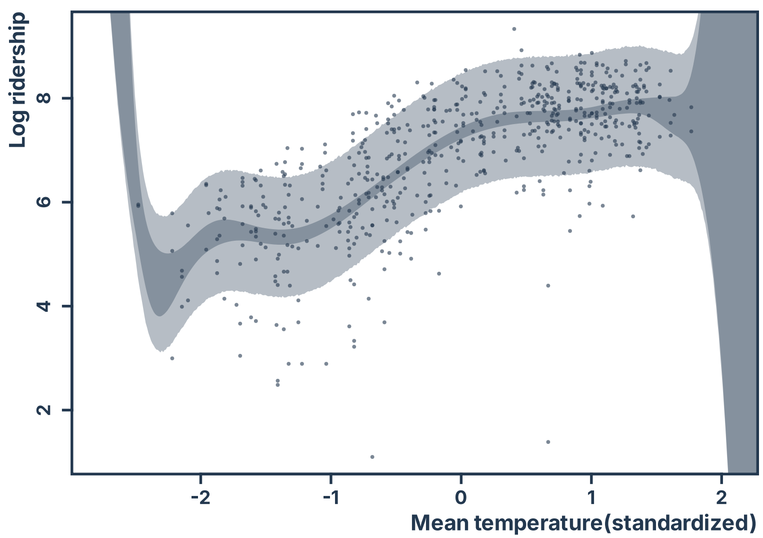


GOODNESS OF FIT



Underfit

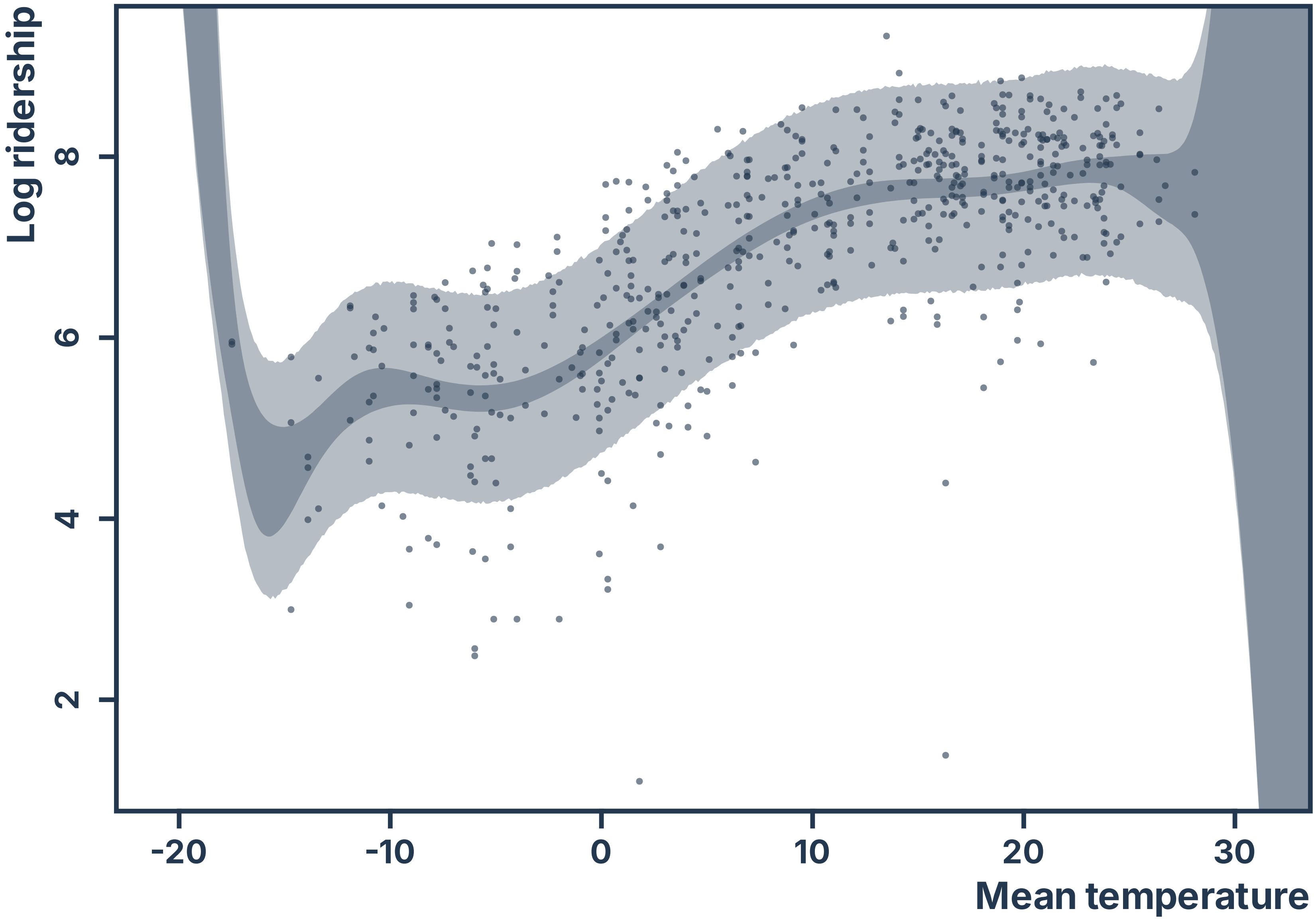
- Predictions err in systematic ways
- Misses meaningful patterns in the relationship between predictor(s) and outcome



Overfit

- Takes random variation to be systematic
- Predicts cases in the sample well, but tends to predict new data very poorly

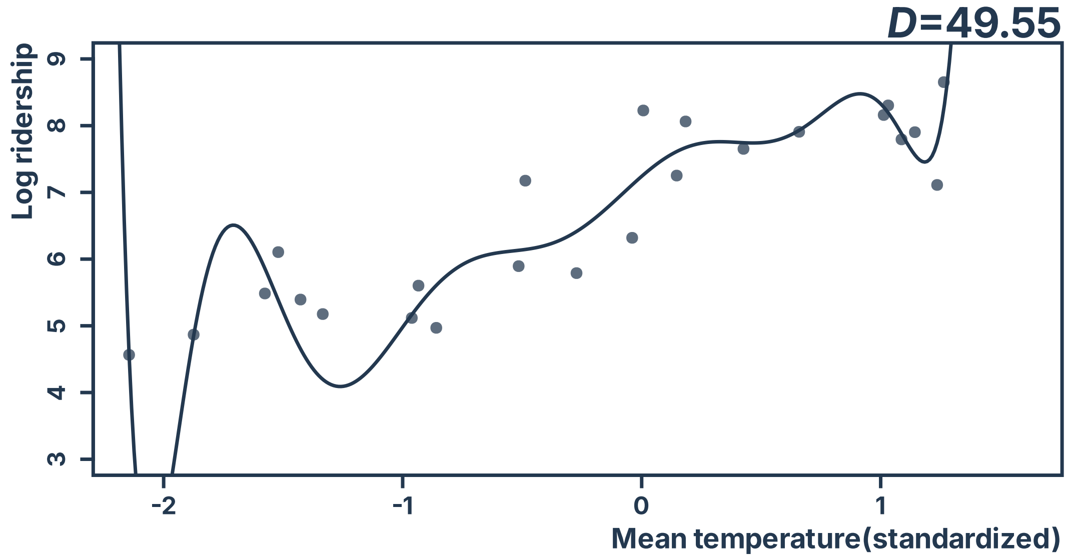
OVERFITTING



TEST AND TRAINING DATA

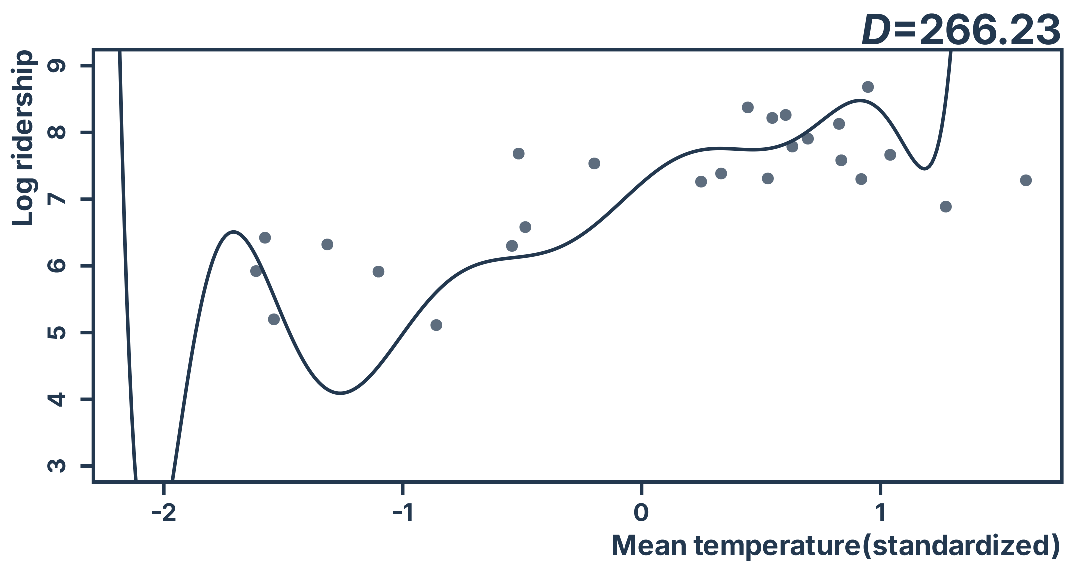
Training data

Fit the model on a subset of the data (e.g. 50%)



Test data

Asses model fit on the held-out portion of the data



AKAIKE INFORMATION CRITERION (AIC)

$$D = -2\log\left(\Pr(\text{data}|\hat{\theta})\right)$$

$$\begin{aligned} AIC &= -2\log\left(\Pr(\text{data}|\hat{\theta})\right) + 2k \\ &= D + 2k \end{aligned}$$

Interpretation 1 | Penalize deviance score for each added parameter by some 'reasonable' value.

Interpretation 2 | Model the average difference in deviance between training and test data.

Assumptions:

- Sample size ≫ number of parameters (k)
- Posterior is approximately (multivariate) normal

INFORMATION CRITERIA

Criterion	Fit	Penalty
Akaike Information Criterion (AIC)	Deviance at the MAP/ML estimate (usually)	#parameters
"Bayesian" Information Criterion (BIC)	Deviance at the MAP/ML estimate	#parameters × log(#observations)
Deviance Information Criterion (DIC)	Deviance averaged across posterior	"Effective" #parameters (posterior)
Widely Applicable Information Criterion (WAIC)	Deviance averaged across posterior and observations	"Effective" #parameters (posterior & obs.)

Strategy 1

Pick the model with the lowest value

WAIC(M_1) = 209.0; WAIC(M_2) = 208.1
→ M_2 is the winner

Strategy 2

Report several models along with values

Multi-model table showing estimates for different combinations of coefficients, along with WAIC

Strategy 3

Average predictions across models

Simultaneous posterior predictions of new data from all models, weighted by WAIC

Considerations when building a model (i.e. choosing covariates)

Theoretical relevance

- **Independent variables chosen to address theoretical concerns**
- *E.g. test theoretical predictions, account for theorized connections*

Causal inference

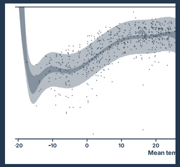
- **Independent variables chosen to make robust causal claims**
- *Worry about including confounders, omitting colliders, and thinking through role of moderating and mediating variables*

Information criteria are for this →

Predictive accuracy

- **Independent variables chosen to maximize predictive power**
- *Accuracy of out-of-sample predictions; Interpretation of models with many moving parts*

Image credit



Figures by Peter McMahan (source code)

Still from The Hudsucker Proxy (1994)

David Byrne by Deborah Feingold