Jan 31

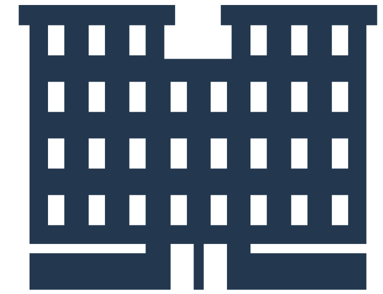
Parsimony and overfitting

1. Parsimony and Occam's razor
2. Overfitting and underfitting
3. Illustrating overfitting with test and training data
4. Information criteria as formal measures of (over)fit
5. Comparing criteria in R

Occam's razor

How many buildings?

Occam's razor



***M*₁: Four buildings**

***M*₂: Five buildings**

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

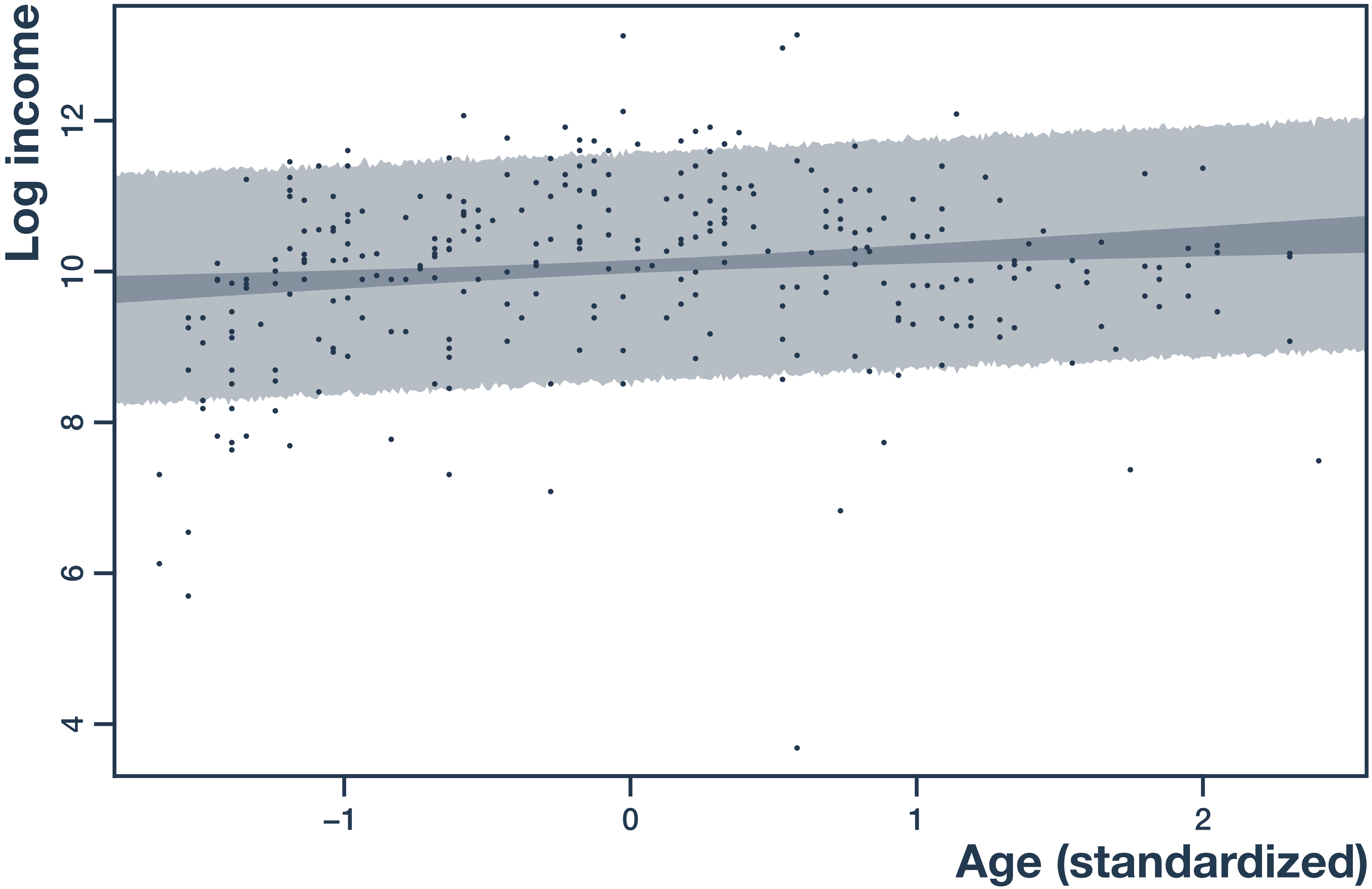
A priori | Simpler models are easier to interpret or more compelling

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

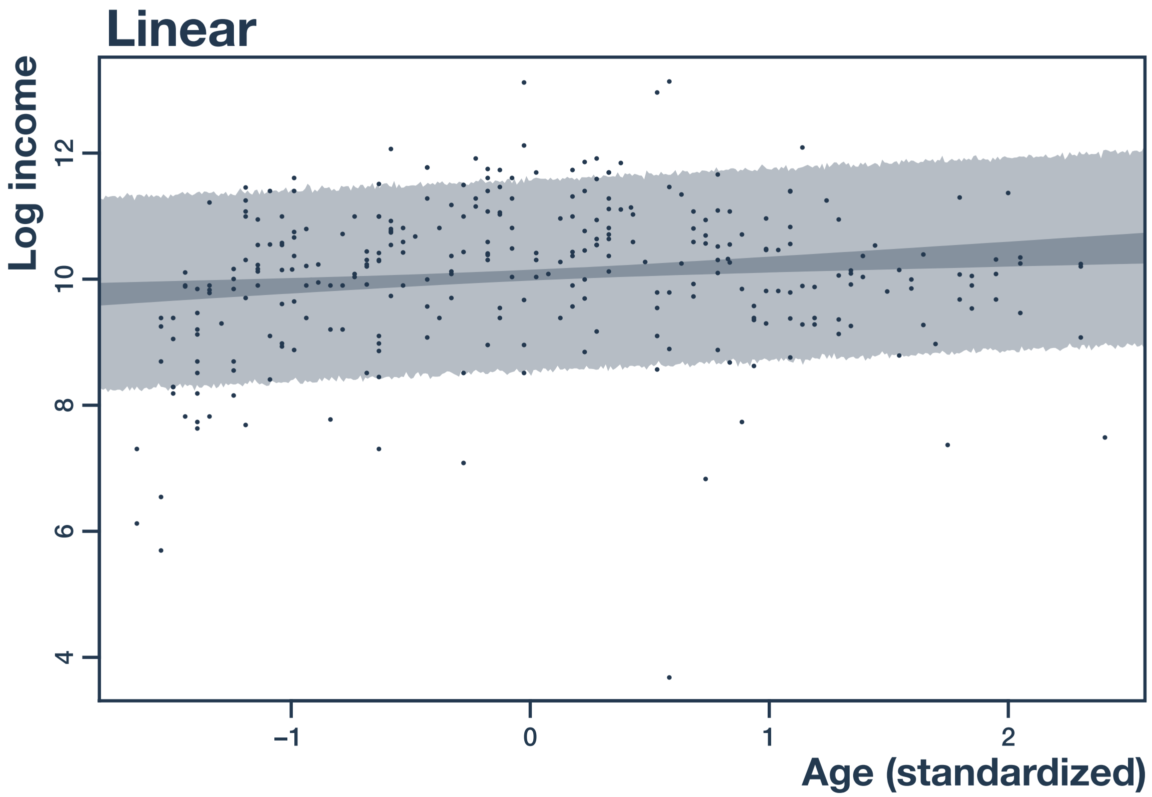
Model likelihood | Simpler models rely less on coincidence

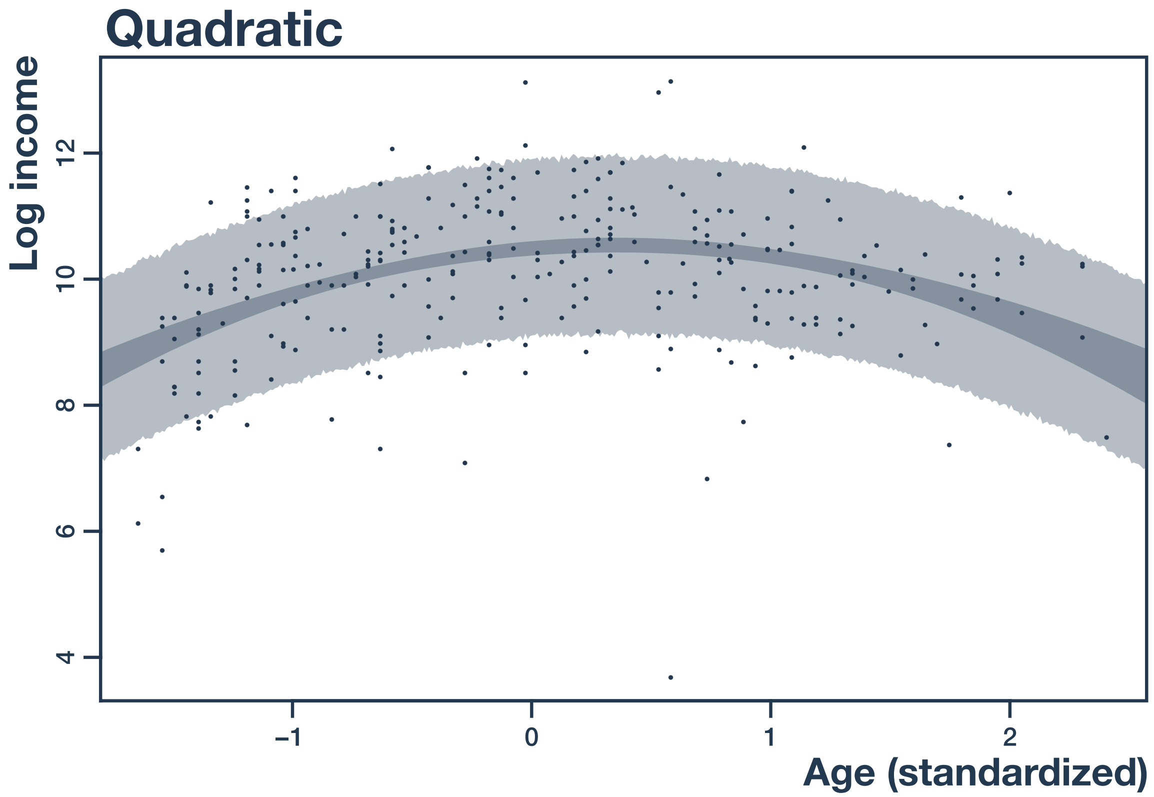
$$\frac{\Pr(D|M_1)}{\Pr(D|M_2)} > 1$$

Assessing fit



Assessing fit





$$\mu_i = \alpha + \beta_1 \text{Age}_i$$

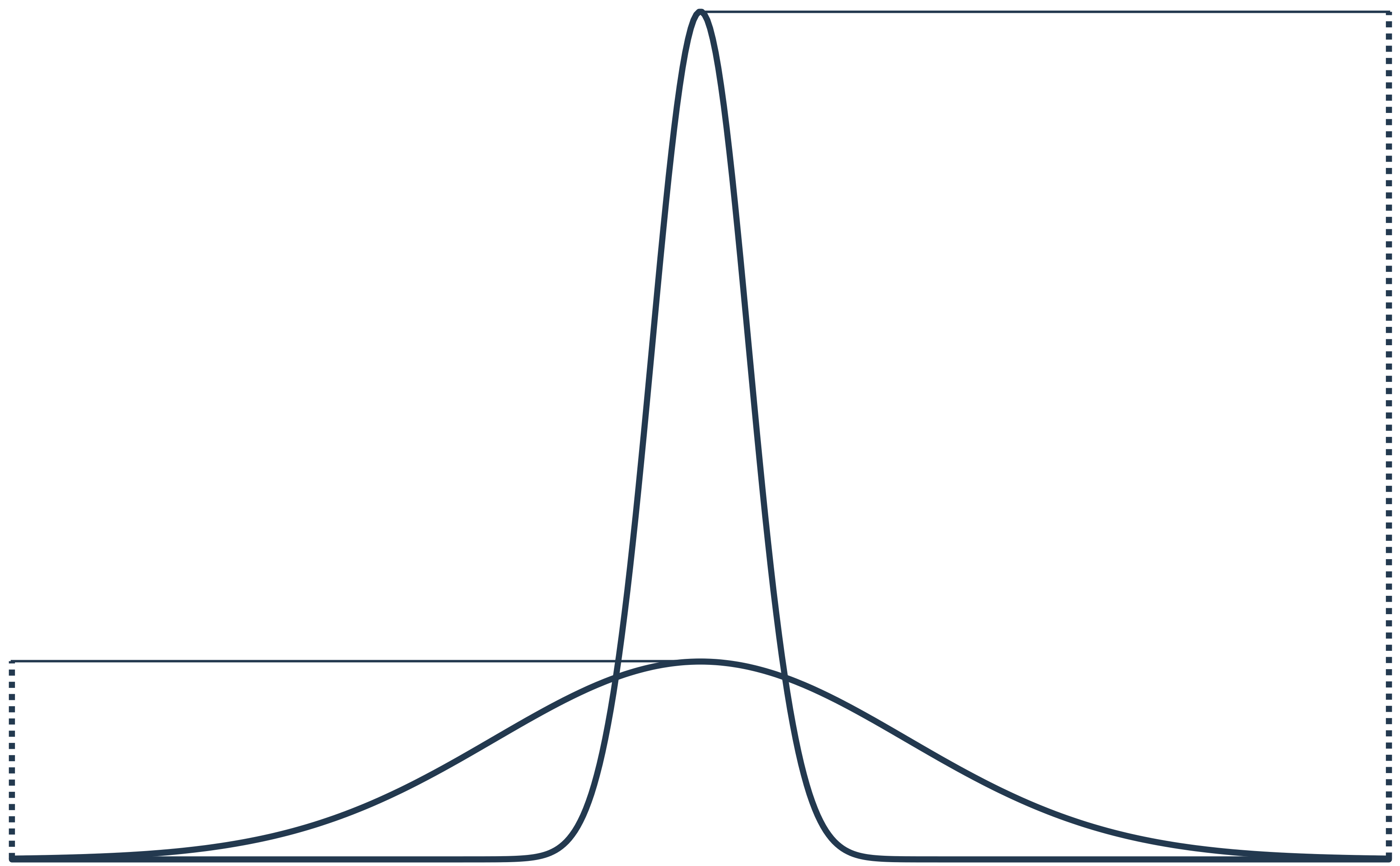
$$\mu_i = \alpha + \beta_1 \text{Age}_i + \beta_2 \text{Age}_i^2$$

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

But how can we measure that?

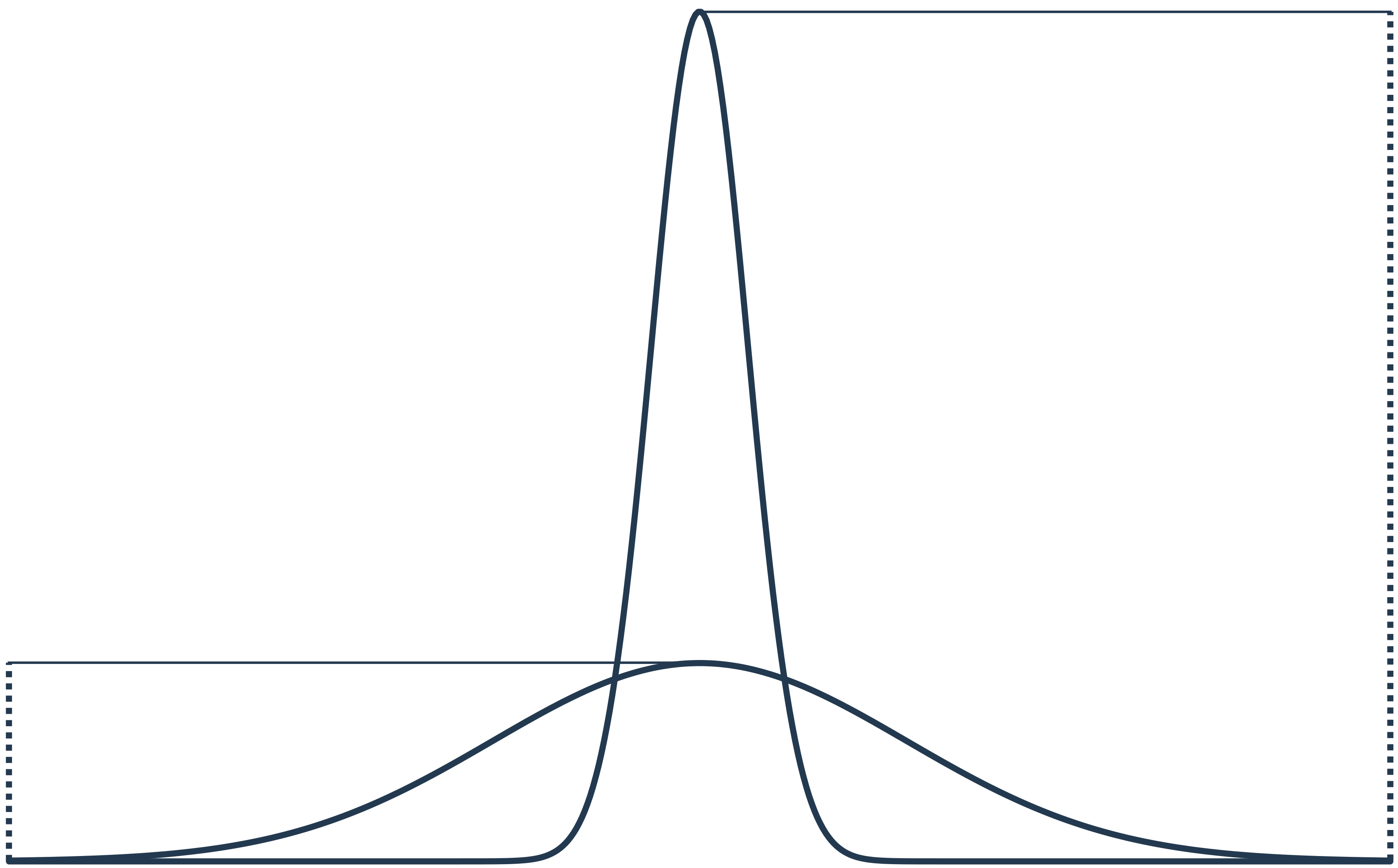
Assessing fit

$$\Pr(\theta \mid D) = \frac{\Pr(D \mid \theta)\Pr(\theta)}{\Pr(D)}$$

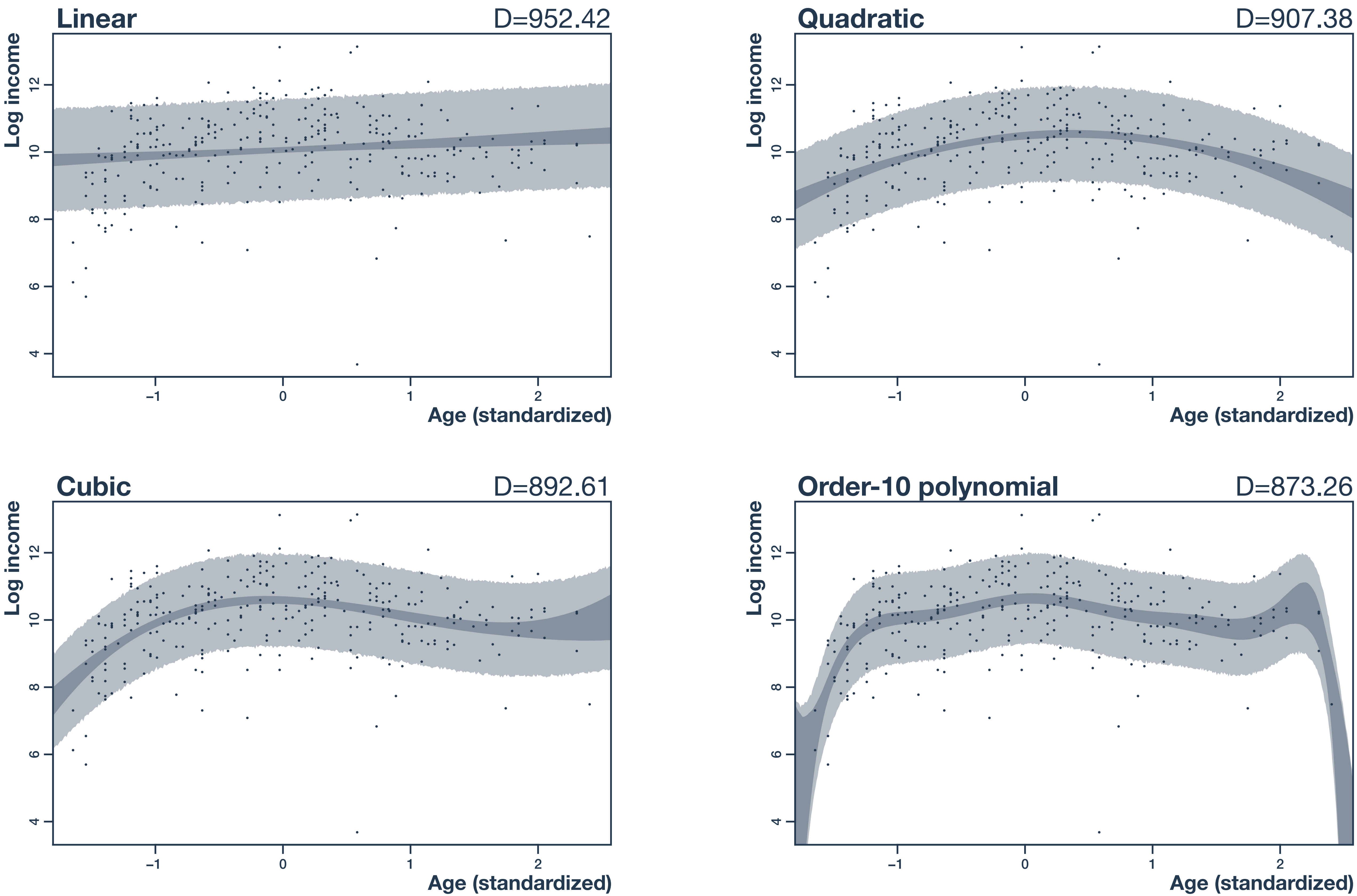


Deviance

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



Deviance

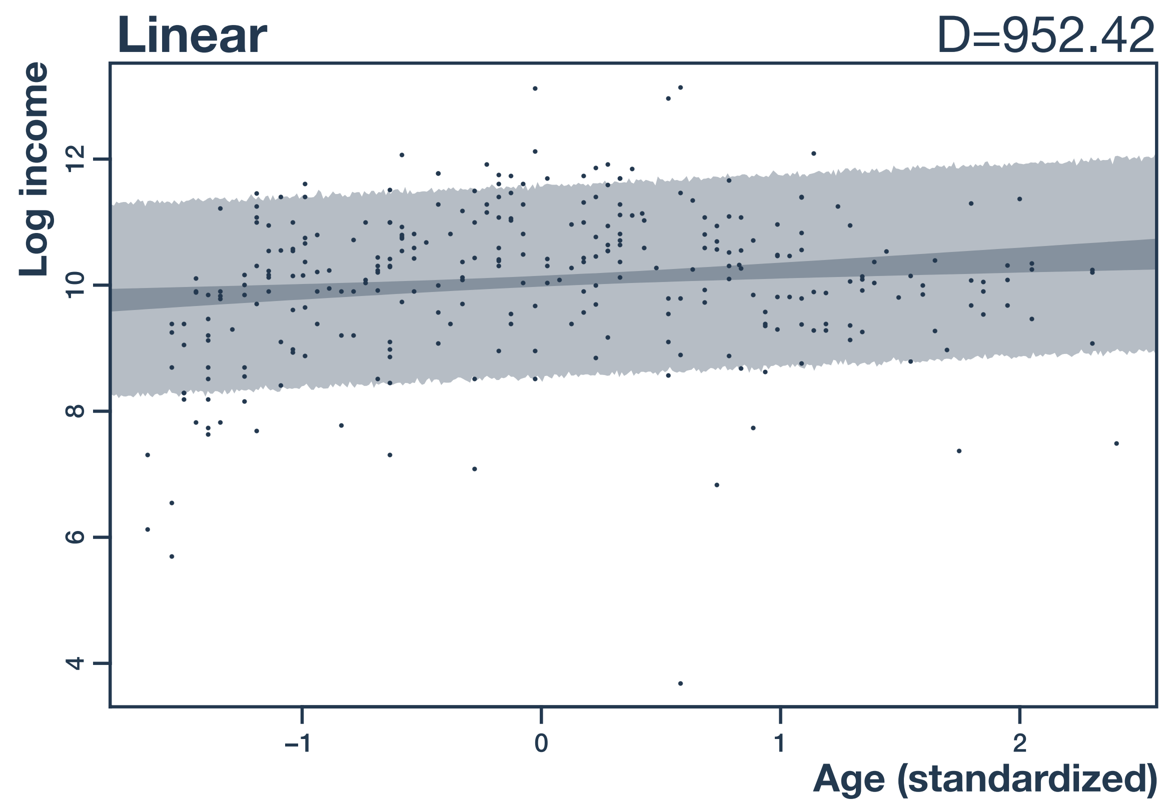

Linear
D=952.42
Quadratic
D=907.38
Cubic
D=892.61
Order-10 polynomial
D=873.26
Log income
Age (standardized)
4
6
8
10
12
−1
0
1
2

Goodness of fit

Underfit

Errs in prediction in a systematic way

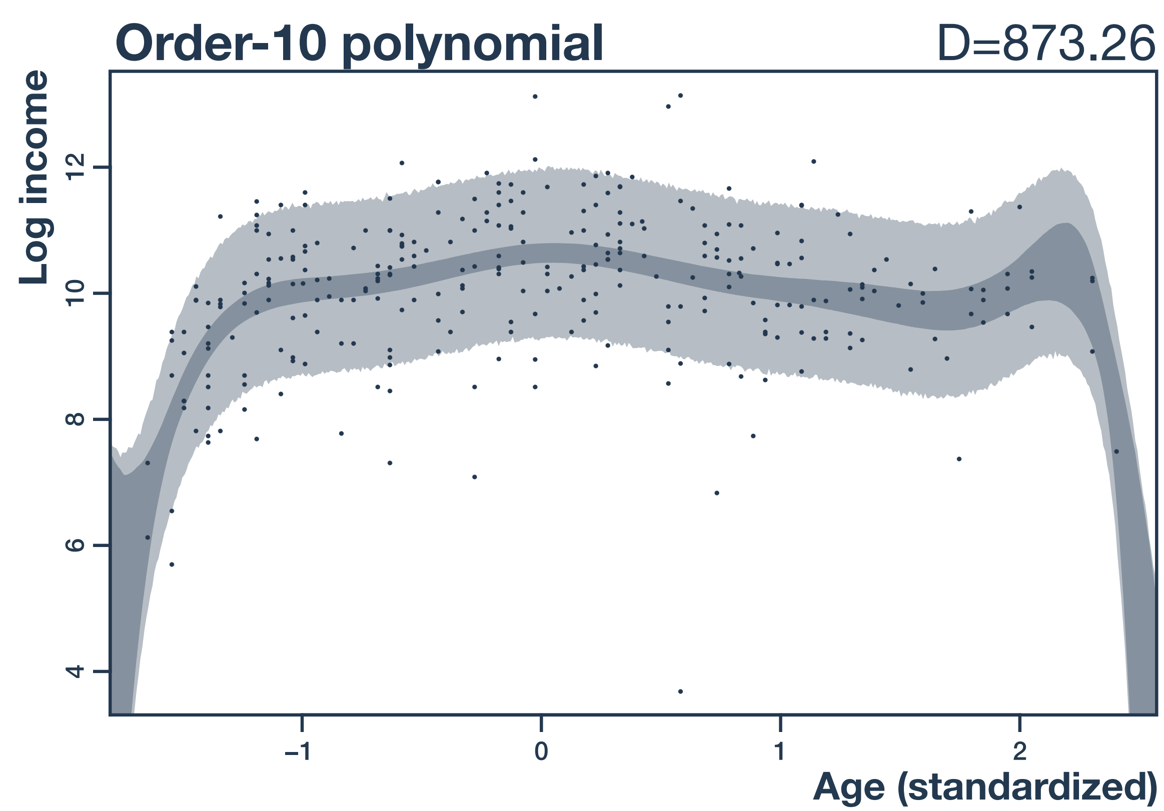
Misses important aspects of relationship between predictor(s) and outcome



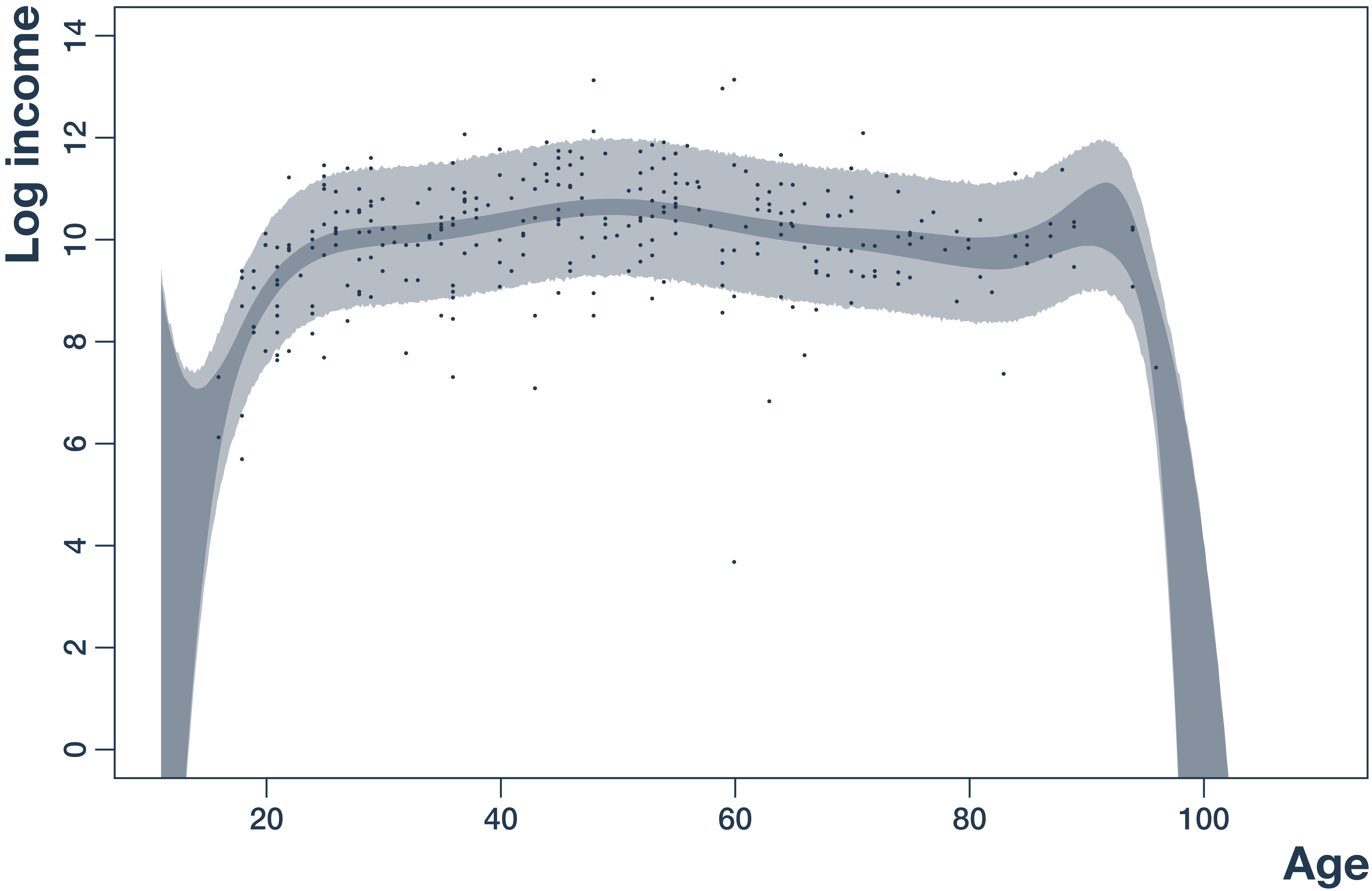
Overfit

Takes random variation to be systematic

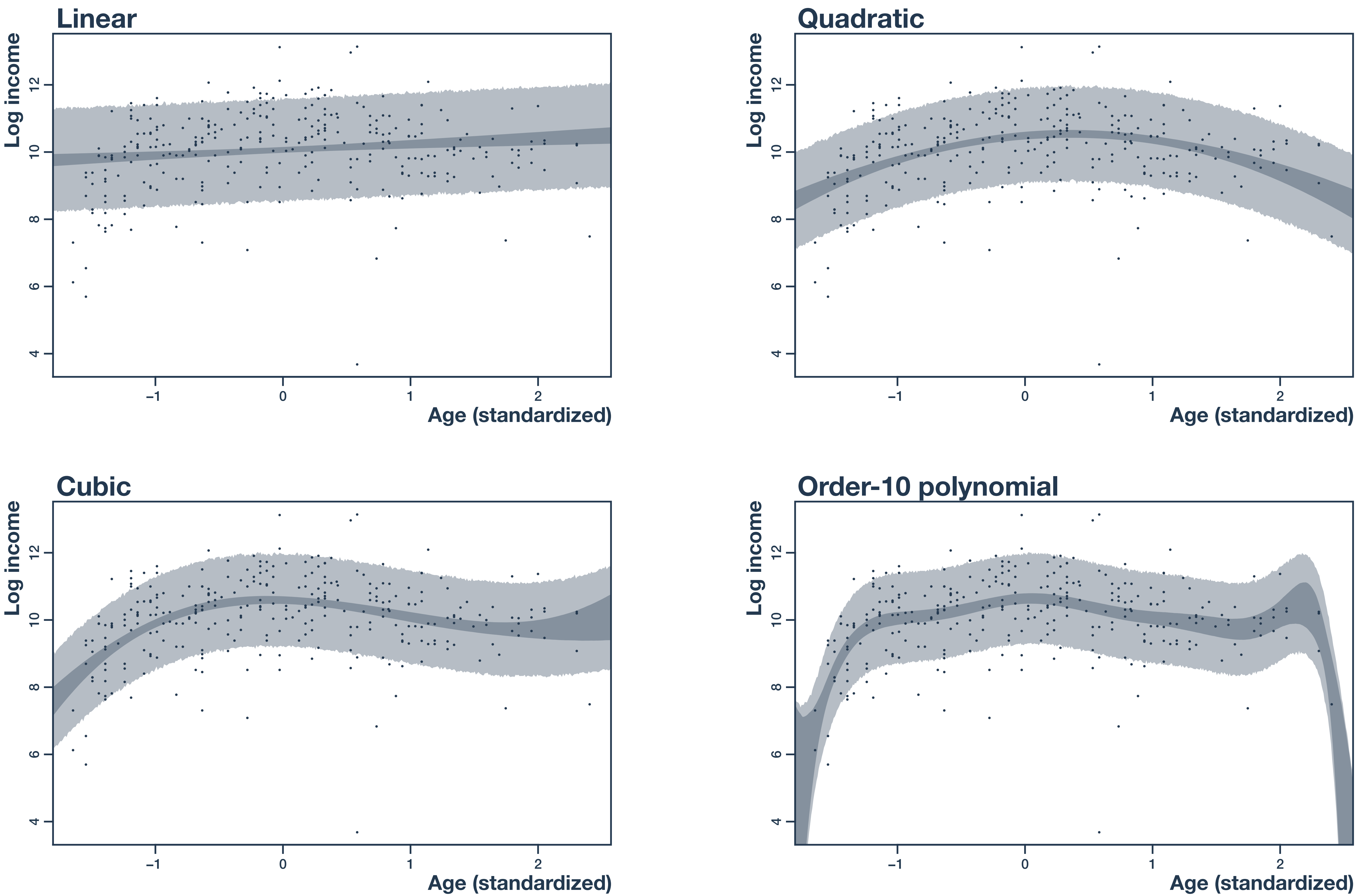
Predicts cases in the sample well, but tends to predict new data very poorly



Overfitting

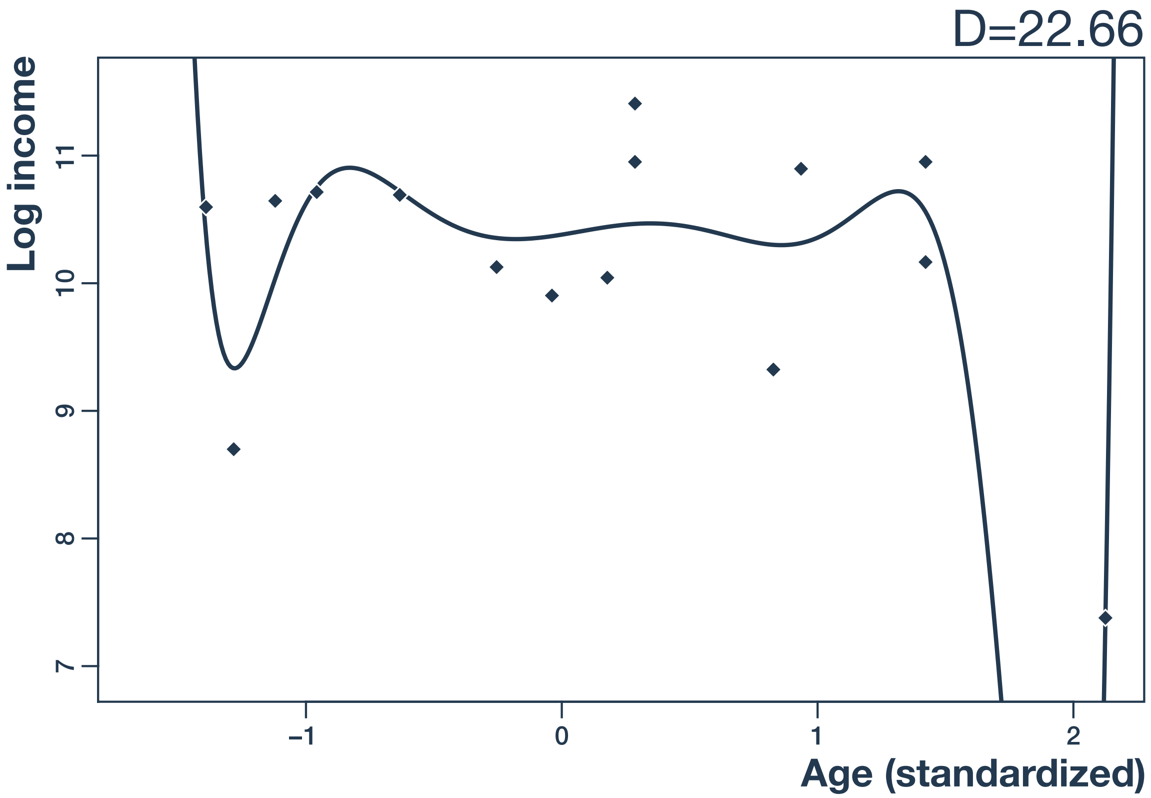


Overfitting

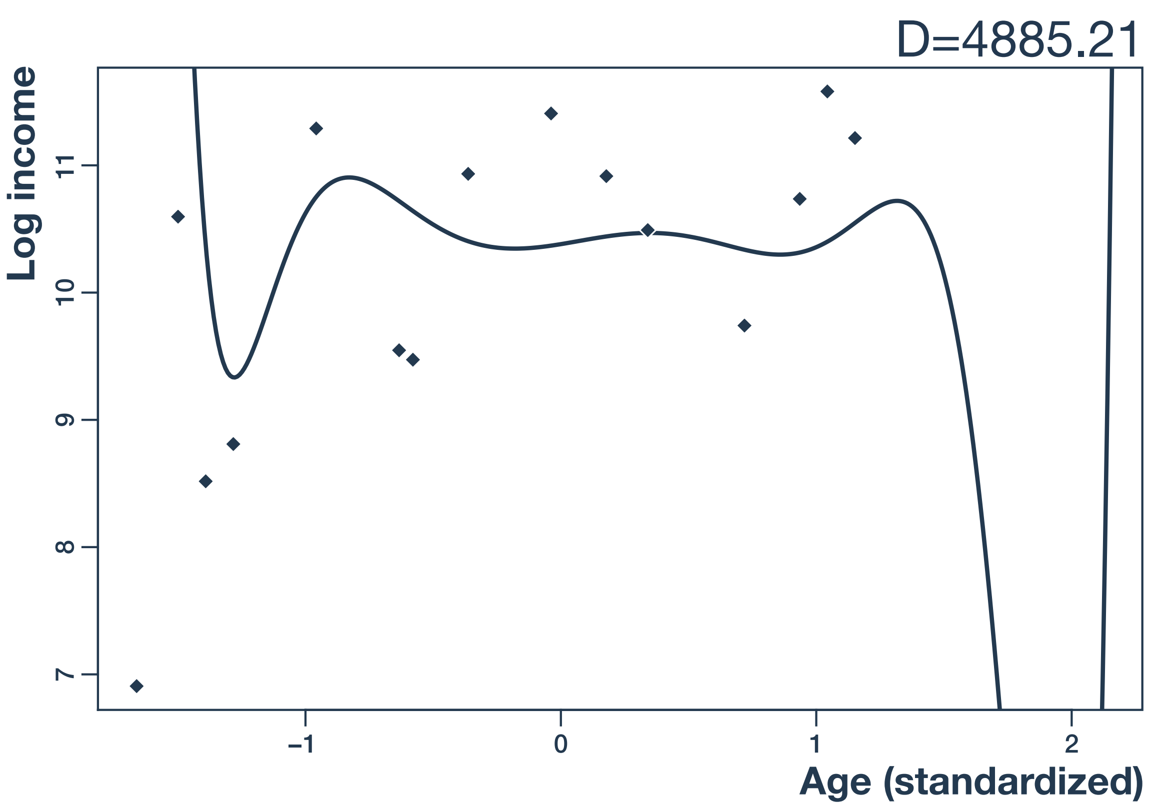


Test and training data

Training data | Fit model on half of the data.



Test data | Assess fit on the other half of the data.



Akaike information criterion (AIC)

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

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

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.

Sample size ≫ number of parameters (k)
Priors have minimal influence (flat or lots of data)
Posterior is approximately (multivariate) normal

Akaike information criterion (AIC)

$$\mathrm{AIC} = \underbrace{\boxed{-2\log\Big(\Pr\big(\mathrm{data}\,\big|\,\theta\big)\Pr\big(\theta\big)\Big)}}_{\text{Measure of model fit}} + \underbrace{\boxed{2k}}_{\text{Penalty for model complexity}}$$

Measure of model fit

Penalty for model complexity

Information criteria

Criterion	Fit	Penalty
Akaike Information Criterion (AIC)	Deviance at MAP estimate (usually)	Number of parameters
“Bayesian” Information Criterion (BIC)	Deviance at MAP estimate	#parameters times 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 and obs.)

Using information criteria

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

Building linear models

Considerations when choosing covariates

Theoretical relevance

Independent variables chosen address theoretical concerns

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

Predictive accuracy

Independent variables chosen to maximize predictive power

Accuracy of out-of-sample predictions; Interpretation of models with many moving parts

Information criteria are for this