Agenda

1. **Interpretation with transformed variables**
2. **Visualizing model predictions**
3. **Examining model fit**
4. **Overfitting and underfitting**
5. **Visualizing predictions in R**

Interpreting coefficients

Predicting: **Income** ← Using: **Age**

$$\text{Income}_i \sim \text{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \text{Age}_i$$

	Post. Mean
α	32813.3
β	185.7

$$\begin{aligned} E(\text{Inc.}|a_2) - E(\text{Inc.}|a_1) &= (\alpha + \beta a_2) - (\alpha + \beta a_1) \\ &= \beta(a_2 - a_1) \\ &= 185.7(a_2 - a_1) \end{aligned}$$

Interpreting coefficients

Predicting: **Income** ← Using: **Age**

$$\text{Income}_i \sim \text{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \text{Age}_i$$

	Post. Mean
α	32813.3
β	185.7

Units of age | Years

Units of income | Dollars

Interpreting β | For each year of age, the model predicts about $186 more income per year.

Standardized variables

Predicting: **St(Income)** ← Using: **St(Age)**

$$\mathrm{St}(\mathrm{Income}_i) \sim \mathrm{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \mathrm{St}(\mathrm{Age}_i)$$

	Post. Mean
α	0
β	0.065

Units of age	Standard deviations of age
Units of income	Standard deviations of income
Interpreting β	For each standard deviation of age, the model predicts an increase of about 0.065 standard deviations in income.

Log of outcome variable

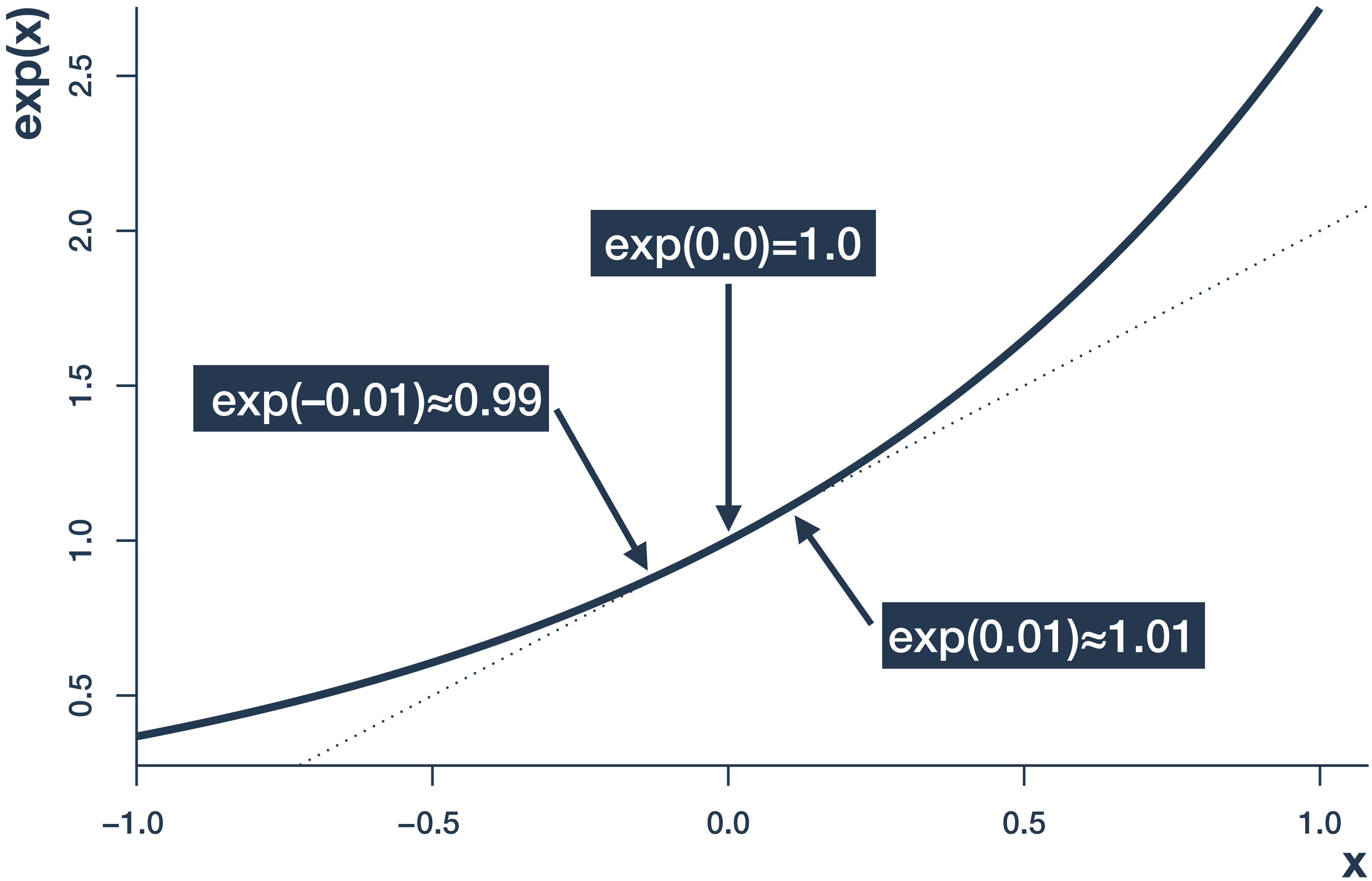
Predicting: **log(Inc)** $\longleftarrow$ Using: **Age**

$$\log(\text{Income}_i) \sim \text{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \text{Age}_i$$

	Post. Mean	exp(Mean)
α	9.648	15489.124
β	0.009	1.009

$$\begin{aligned}
\text{Inc}_2/\text{Inc}_1 &= \exp(\mathit{log}(\text{Inc}_2) - \log(\text{Inc}_1)) \\
&= \exp((\alpha + \beta a_2) - (\alpha + \beta a_1)) \\
&= \exp(\beta(a_2 - a_1))
\end{aligned}$$

Log of outcome variable



Log of outcome variable

Predicting: **log(Inc)** ← Using: **Age**

$$\log(\text{Income}_i) \sim \text{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \text{Age}_i$$

	Post. Mean	exp(Mean)
α	9.648	15489.124
β	0.009	1.009

Units of age | Years

Units of income | Log dollars

Interpreting β | For each year of age, the model predicts a 0.9% increase in income.

Log of predictor variable

Predicting: **Income** ← Using: **log(Age)**

$$\text{Income}_i \sim \text{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \log(\text{Age}_i)$$

	Post. Mean
α	–17586
β	15675

$$\begin{aligned}\text{Inc}_2 - \text{Inc}_1 &= (\alpha + \beta\log(a_2)) - (\alpha + \beta\log(a_1))) \\ &= \beta(\log(a_2) - \log(a_1))) \\ &= \beta(\log(a_2/a_1)))\end{aligned}$$

Log of predictor variable

Predicting: **Income** ← Using: **log(Age)**

$$\text{Income}_i \sim \text{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \log(\text{Age}_i)$$

	Post. Mean
α	–17586
β	15675

Units of age | Log years

Units of income | Dollars

Interpreting β | For each 10% increase in age, the model predicts an increase of $\beta \times \log(1.1) = 15{,}675 \times 0.095 = 1{,}494.07$ dollars in income.

Visualizing predictions

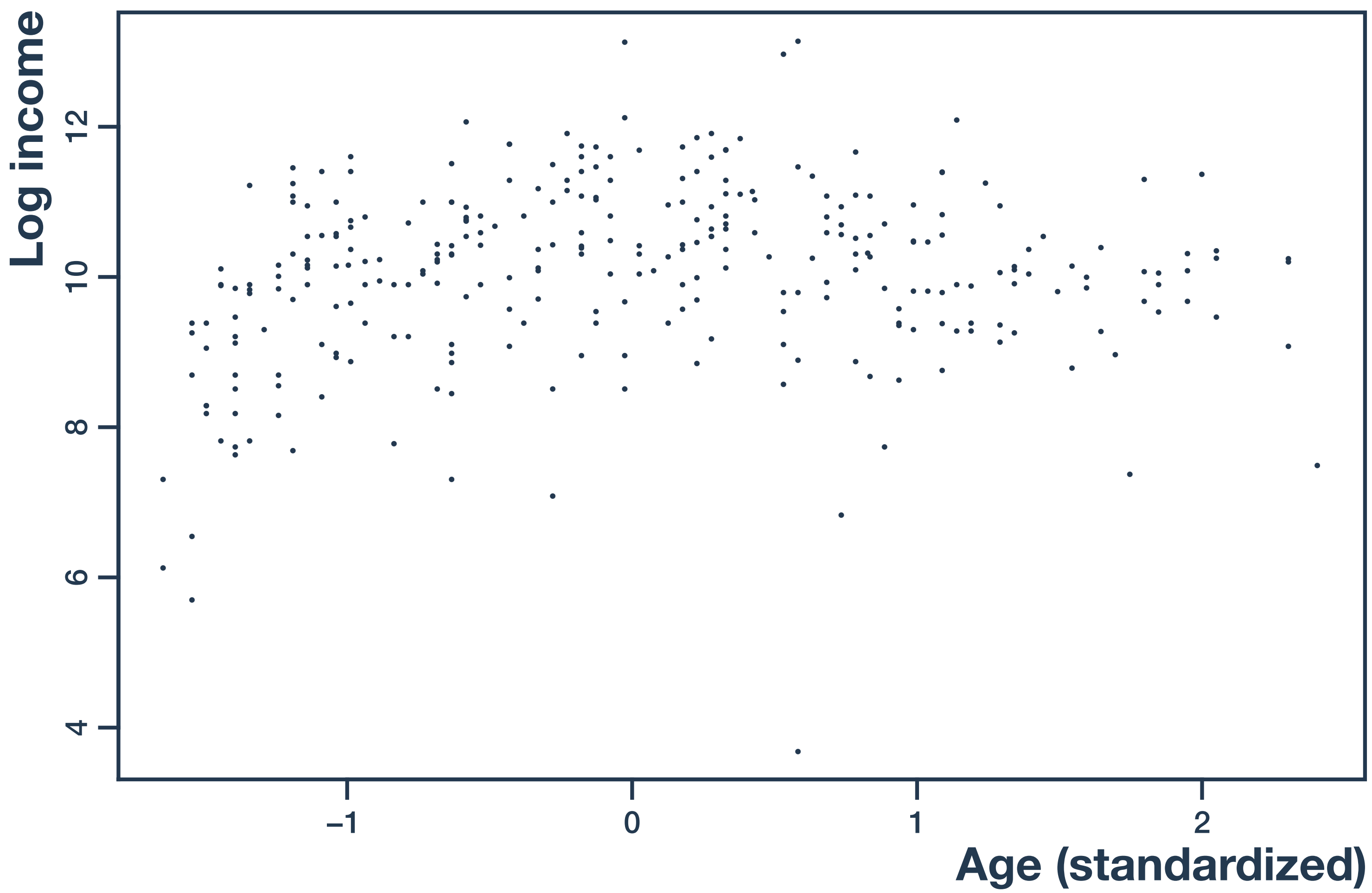
$$\log(\text{Inc}_i) \sim \text{Norm}(\mu_i, \sigma)$$

$$\mu_i = \alpha + \beta \text{St}(\text{Age}_i)$$

$$\alpha \sim \text{Norm}(10, 2)$$

$$\beta \sim \text{Norm}(0, 3)$$

$$\sigma \sim \text{Unif}(0, 5)$$

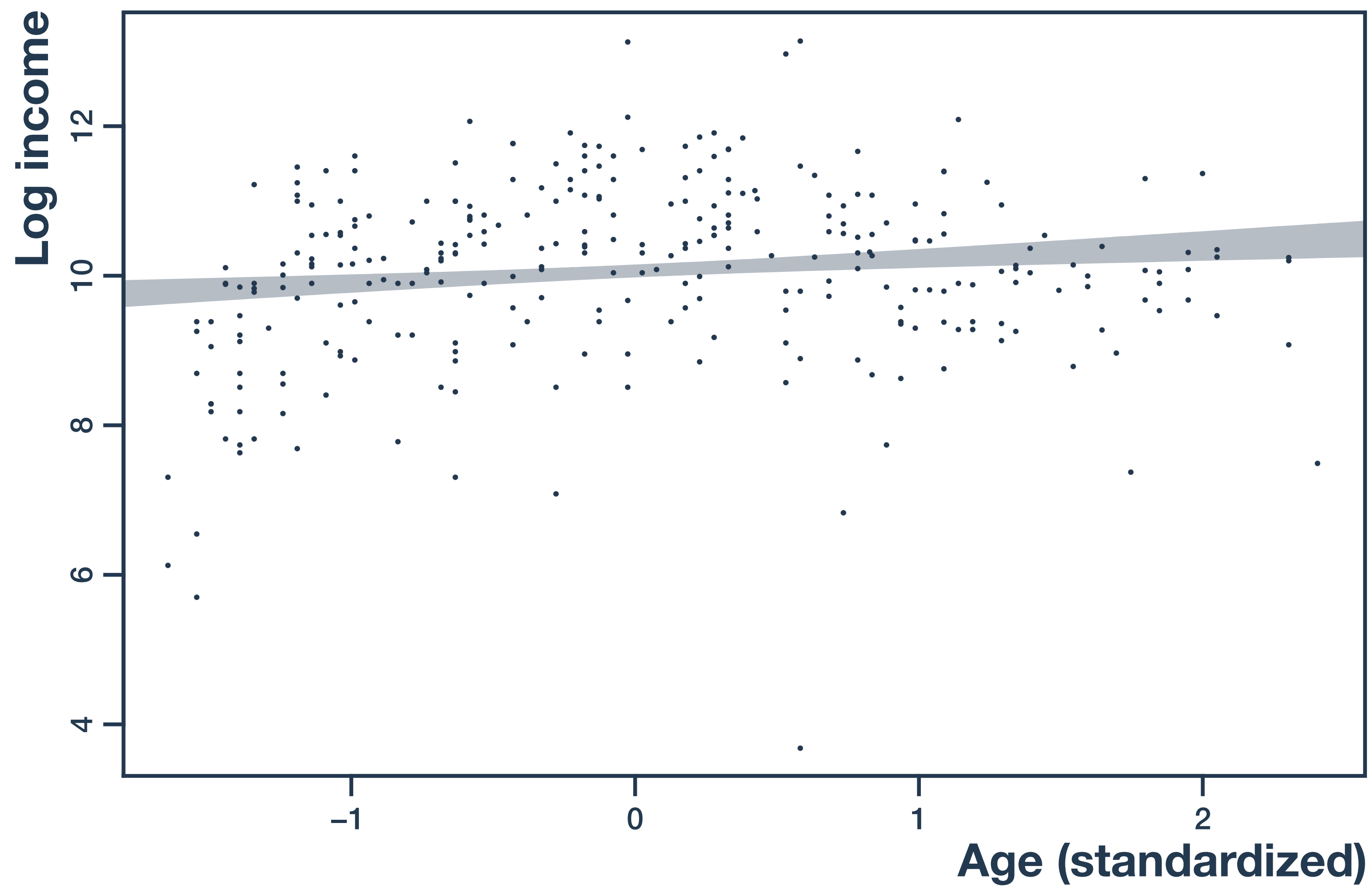


Visualizing predictions

$$\log(\mathrm{Inc}_i) \sim \mathrm{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \mathrm{St}(\mathrm{Age}_i)$$

$$\alpha \sim \mathrm{Norm}(10, 2)$$
$$\beta \sim \mathrm{Norm}(0, 3)$$
$$\sigma \sim \mathrm{Unif}(0, 5)$$

	Post. Mean	exp(Mean)
α	10.06	0.07
β	0.17	0.07
σ	1.18	0.05



Posterior distribution of mean:

$$\Pr(\mu | \mathrm{Age} = a)$$

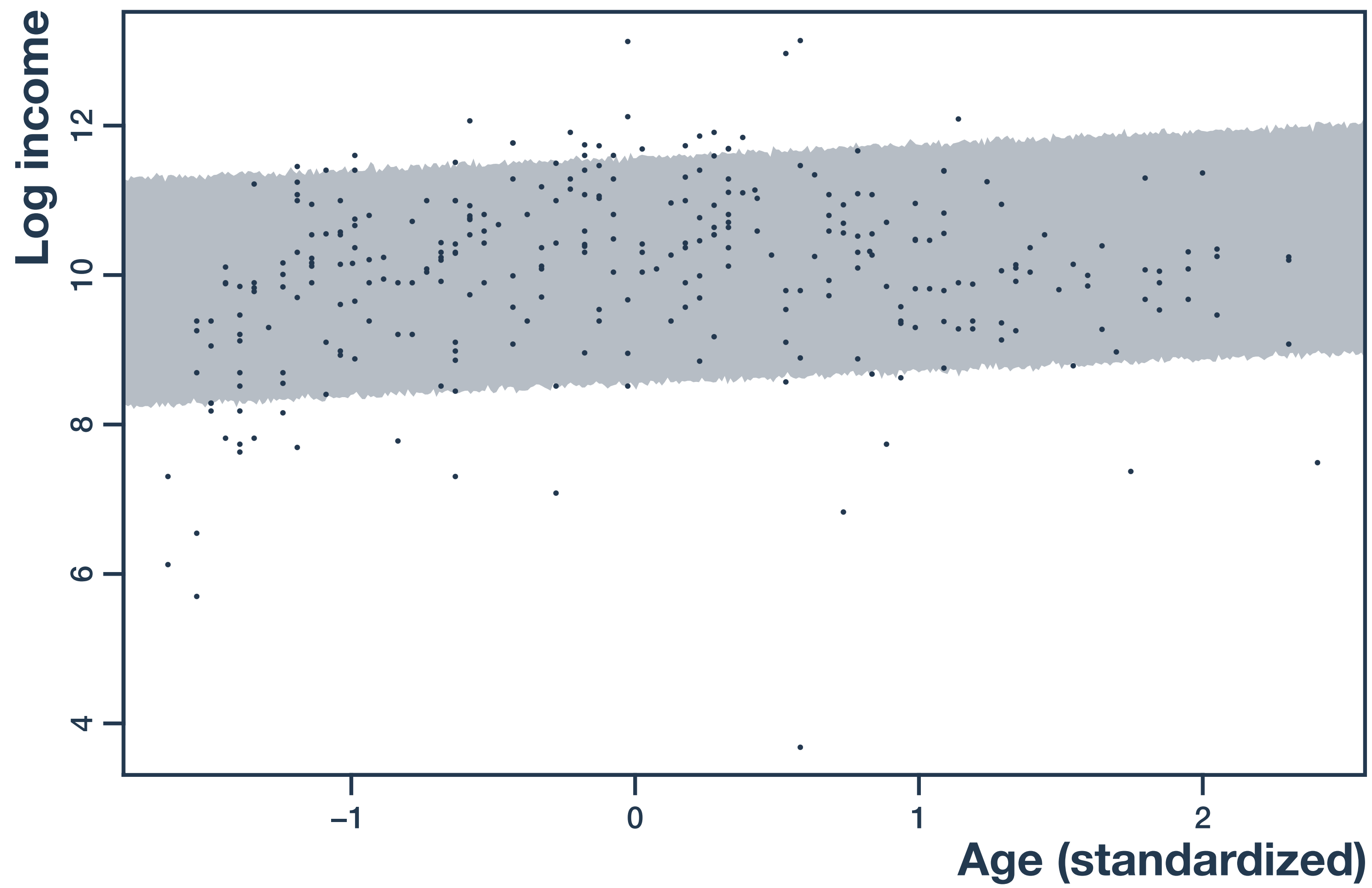
1. Take a sample of size N from posterior $\Pr(\alpha, \beta, \sigma | D)$.
2. For each value of Age a, calculate N values of $\mu = \alpha + \beta a$.
3. Calculate quantiles (say, 10% and 90%) for posterior of μ at each value of a.

Visualizing predictions

$$\log(\text{Inc}_i) \sim \text{Norm}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta \text{St}(\text{Age}_i)$$

$$\alpha \sim \text{Norm}(10, 2)$$
$$\beta \sim \text{Norm}(0, 3)$$
$$\sigma \sim \text{Unif}(0, 5)$$

	Post. Mean	exp(Mean)
α	10.06	0.07
β	0.17	0.07
σ	1.18	0.05

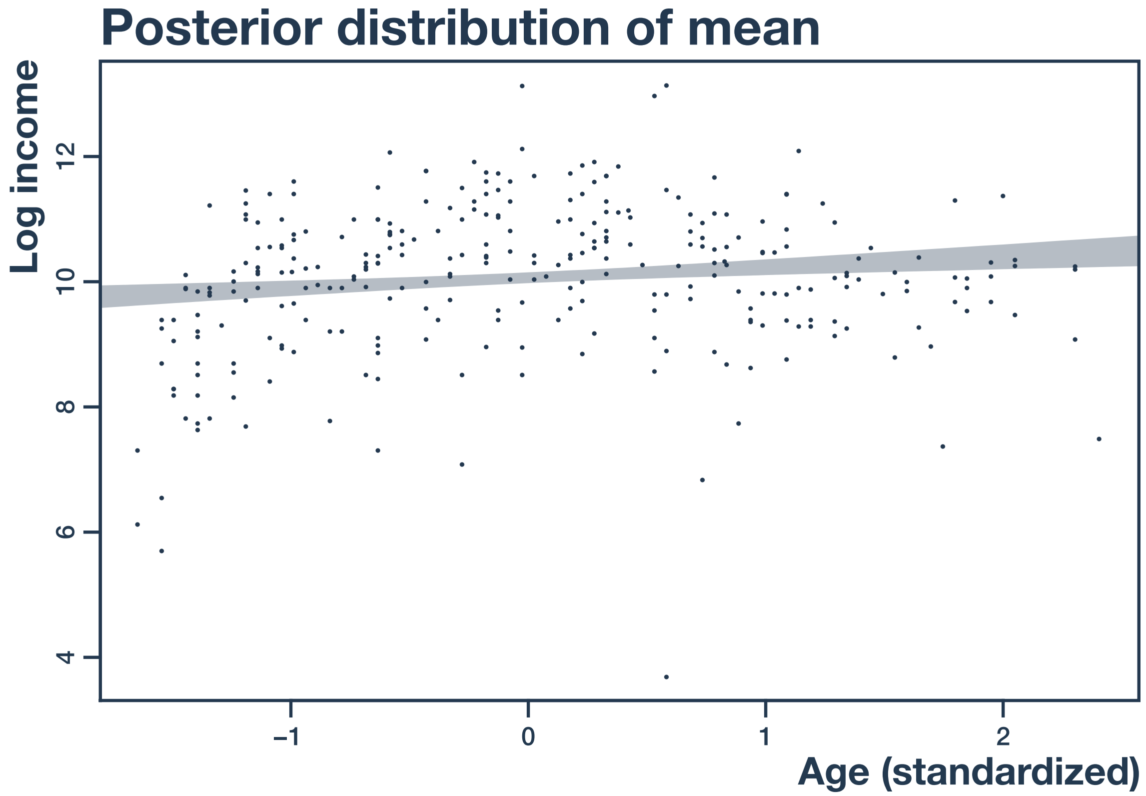


Posterior predictive distribution:

$$\Pr(\log(Inc) | \text{Age} = a)$$

1. Take a sample of size N from posterior Pr(α, β, σ|D).
2. For each value of Age a, calculate N values of $\mu = \alpha + \beta a$.
3. Draw from Norm(μ, σ) for each of the N posterior samples.
4. Calculate quantiles of these predicted outcomes.

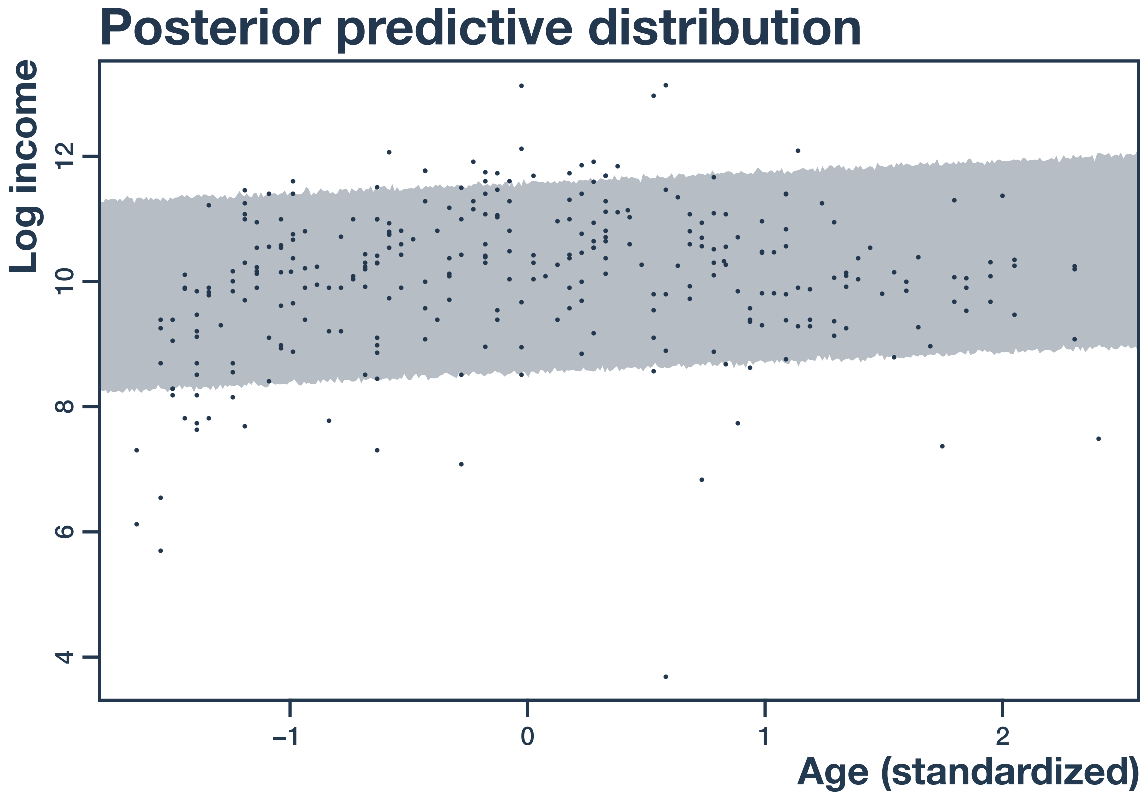
Mean versus prediction



For any given age, μ is the "expected" (mean) log income for people of that age.

The posterior distribution of μ describes our modeled uncertainty about the value of μ.

This distribution takes into account coefficients α and β, but not the standard deviation σ.



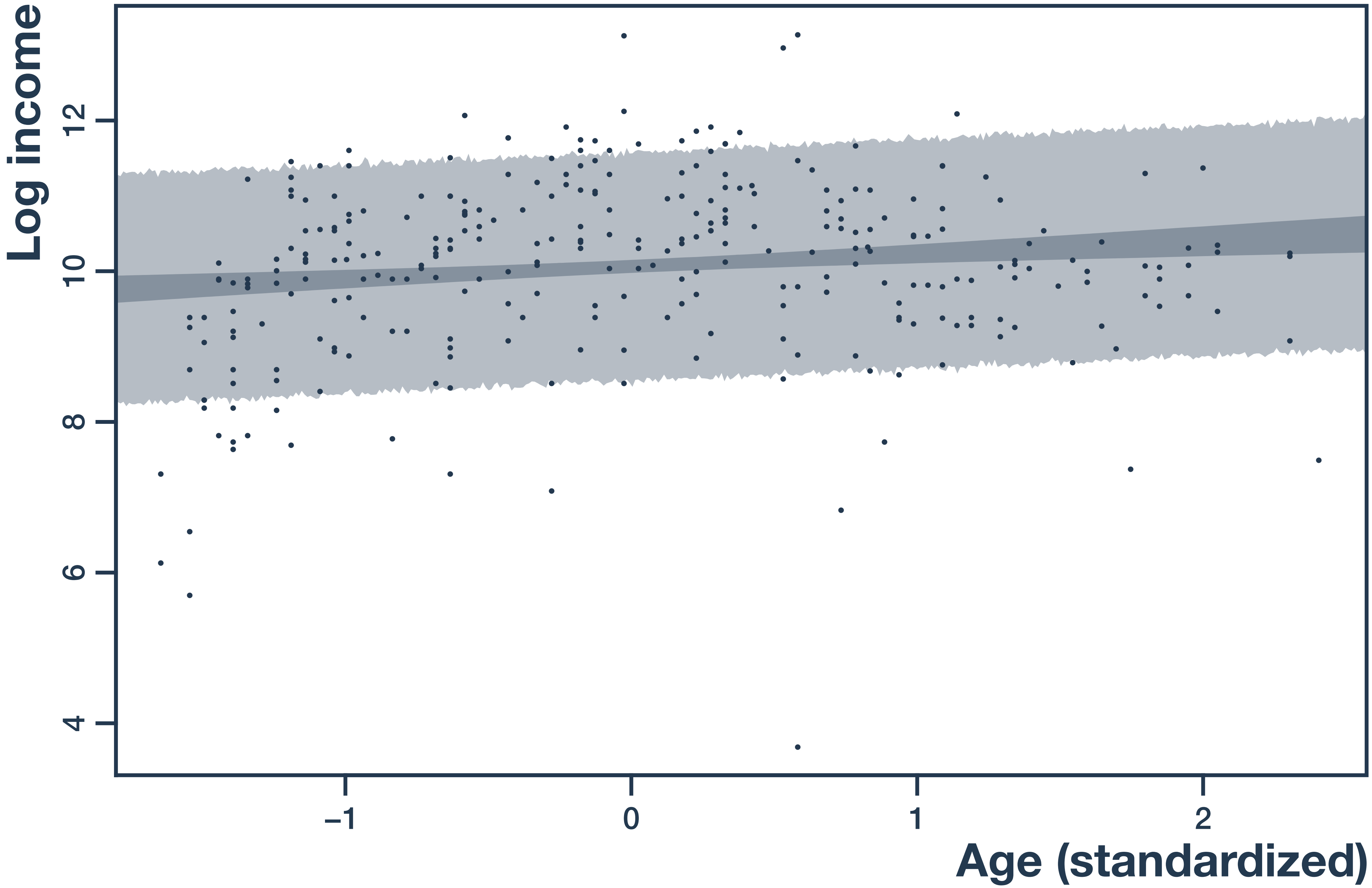
For any given age, the posterior predictive distribution predicts the log income for any individual person of that age.

The posterior distribution of μ describes our modeled uncertainty about the value of log(Inc).

This distribution takes into account coefficients α, β, and σ.

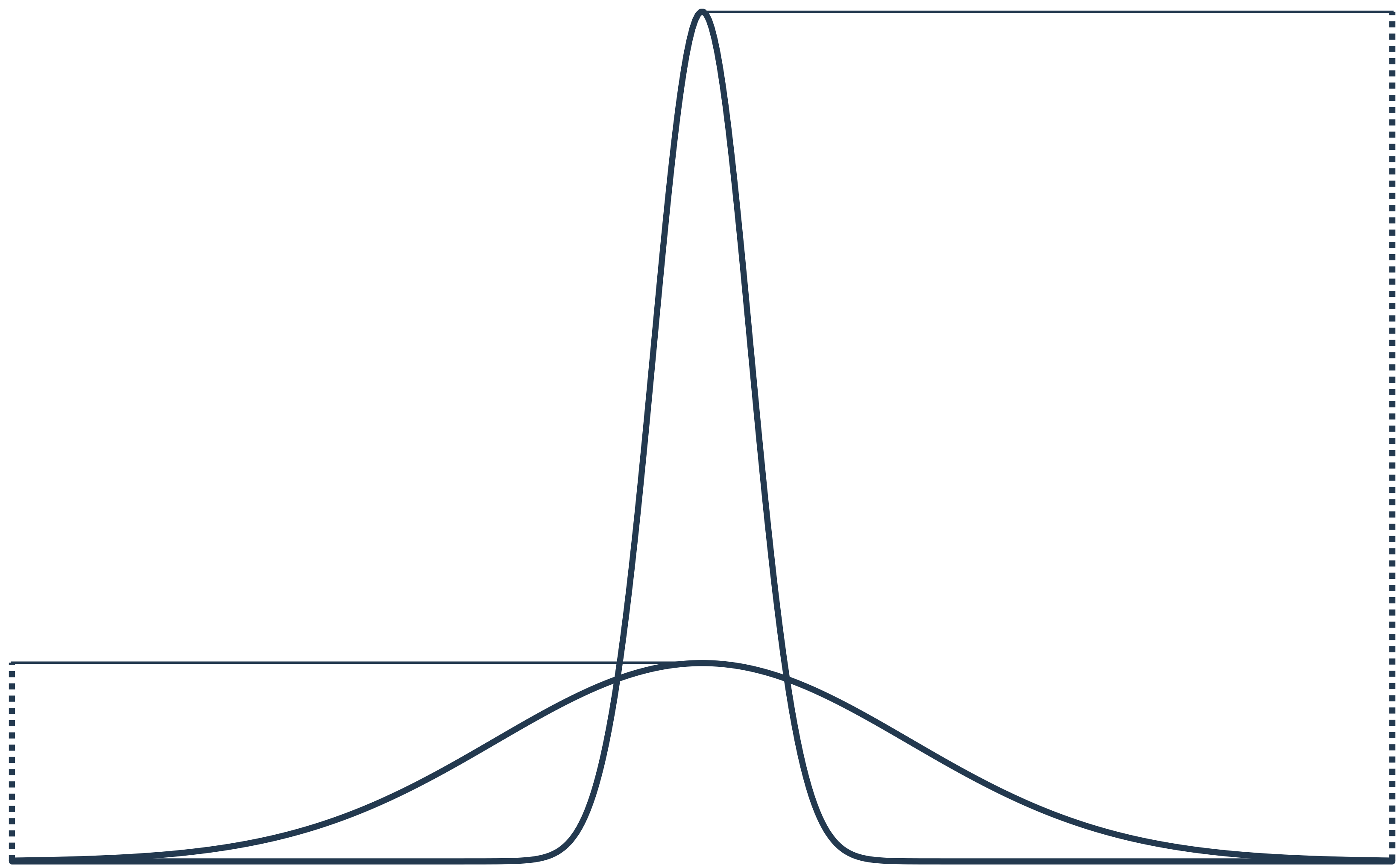
The 80% posterior interval should contain about 80% of the data.

Assessing fit



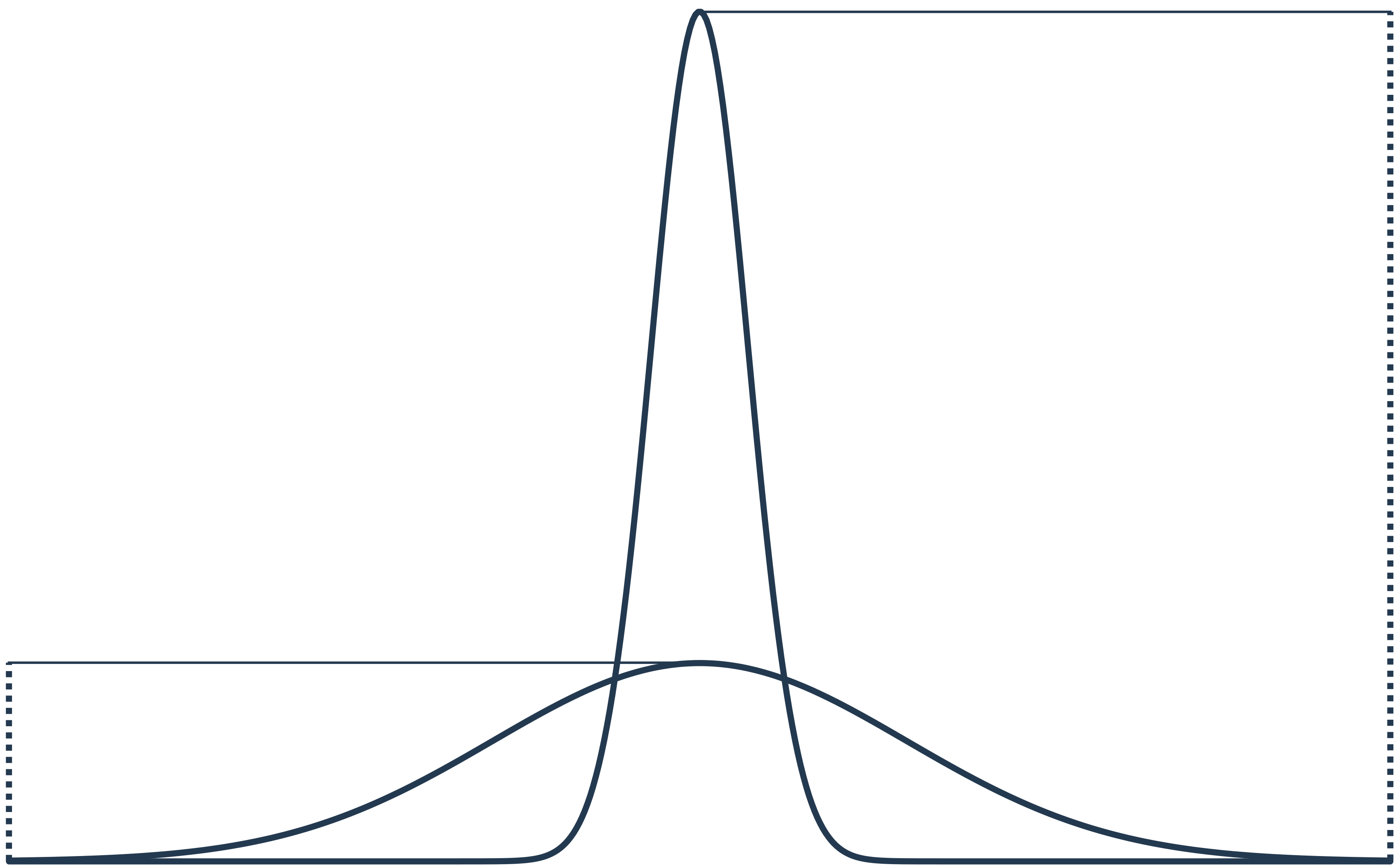
Assessing fit

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

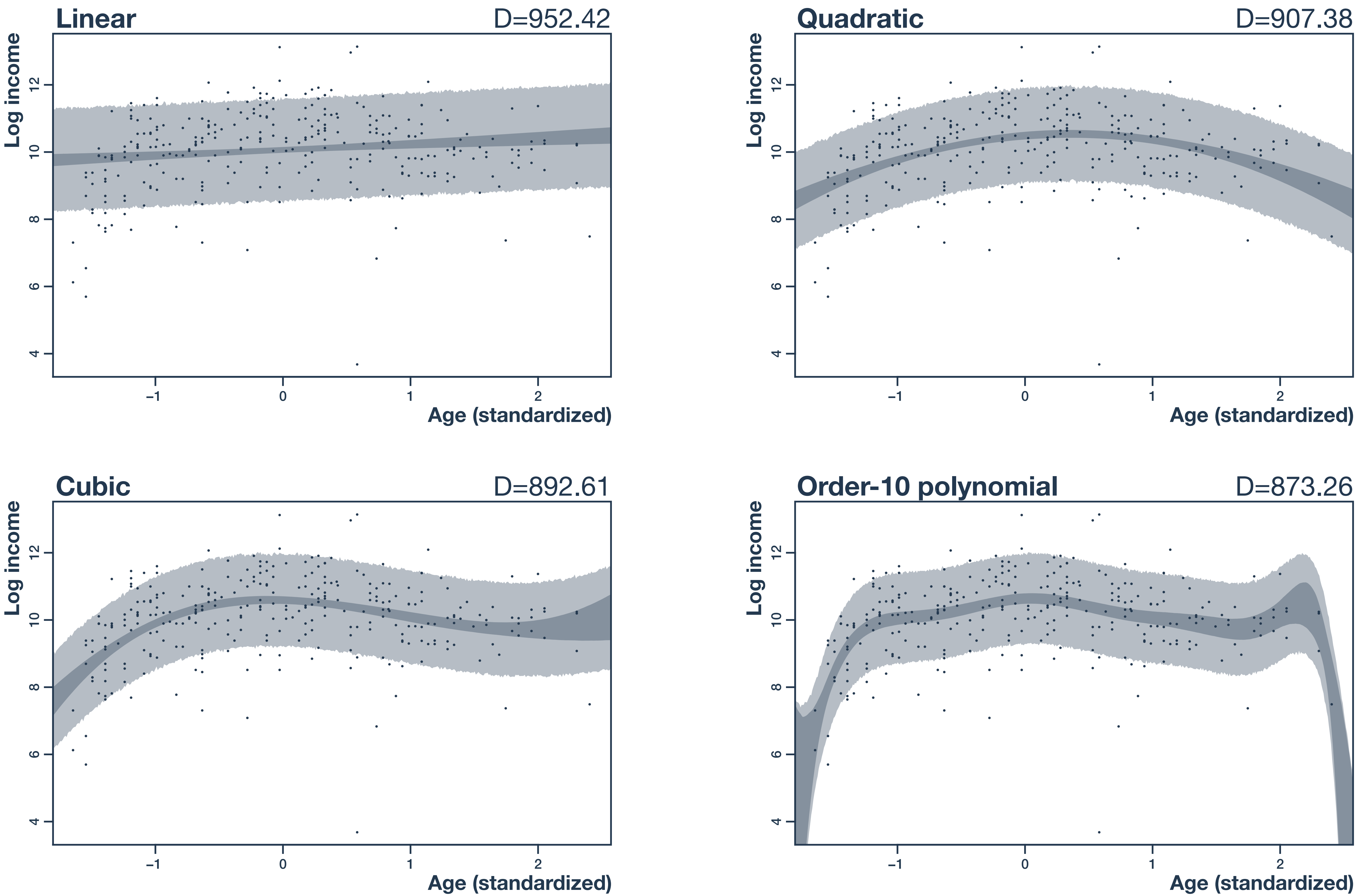


Deviance

$$D = -2\log\Big(\Pr\big(\theta \big| D\big)\Big)$$



Deviance

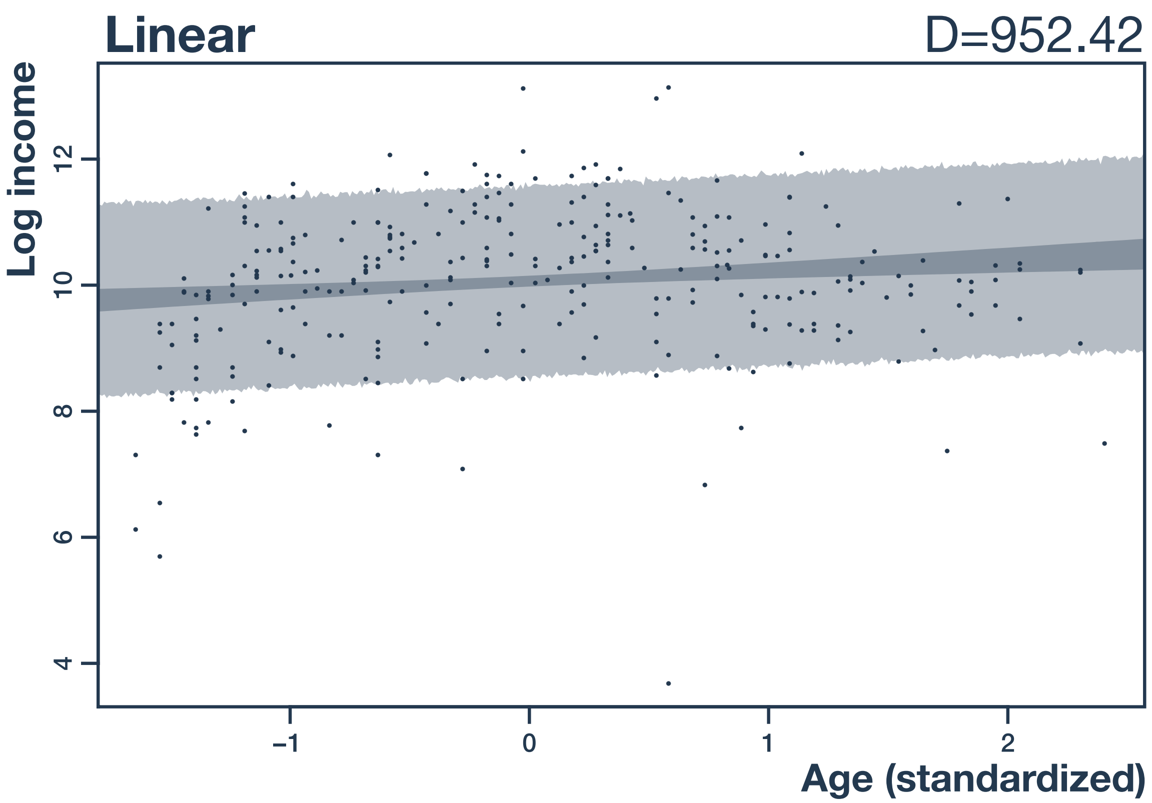

Linear
D=952.42
Quadratic
D=907.38
Cubic
D=892.61
Order-10 polynomial
D=873.26
Log income
Age (standardized)
4
6
8
10
12
−1
0
1
2

Goodness of fit

Underfit

Errs in prediction in a systematic way

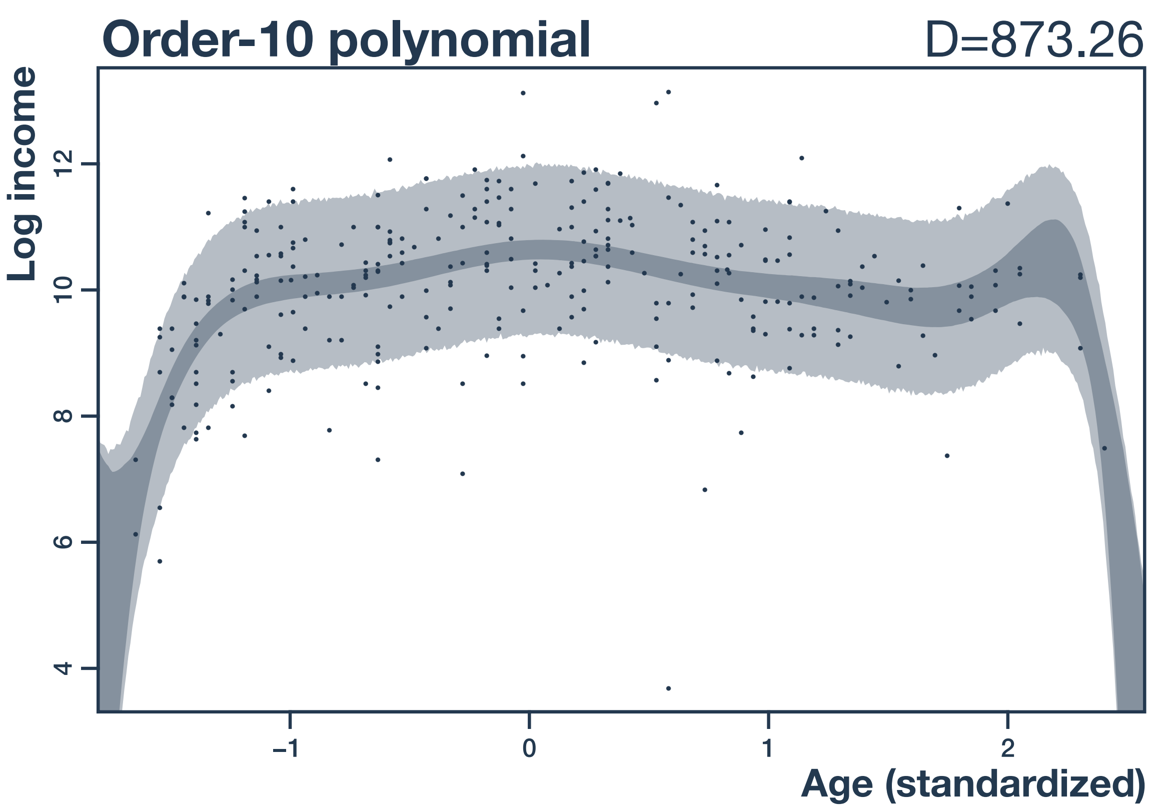
Misses important aspects of relationship between predictor and outcome



Overfit

Takes random variation to be systematic

Predicts data from sample well, but tends to predict new data very poorly



Overfitting