Studying Social Inequality with Data Science

INFO 3370 / 5371
Spring 2024

Statistical Learning

Learning goals for today

By the end of class, you will be able to

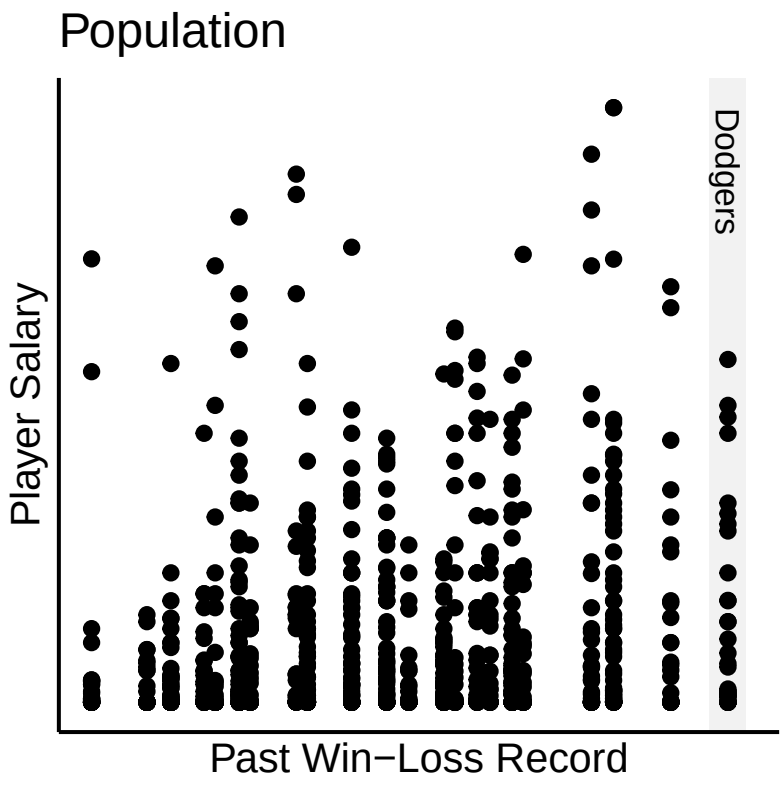
- use statistical learning to estimate when data are sparse
- work with models that are "wrong"

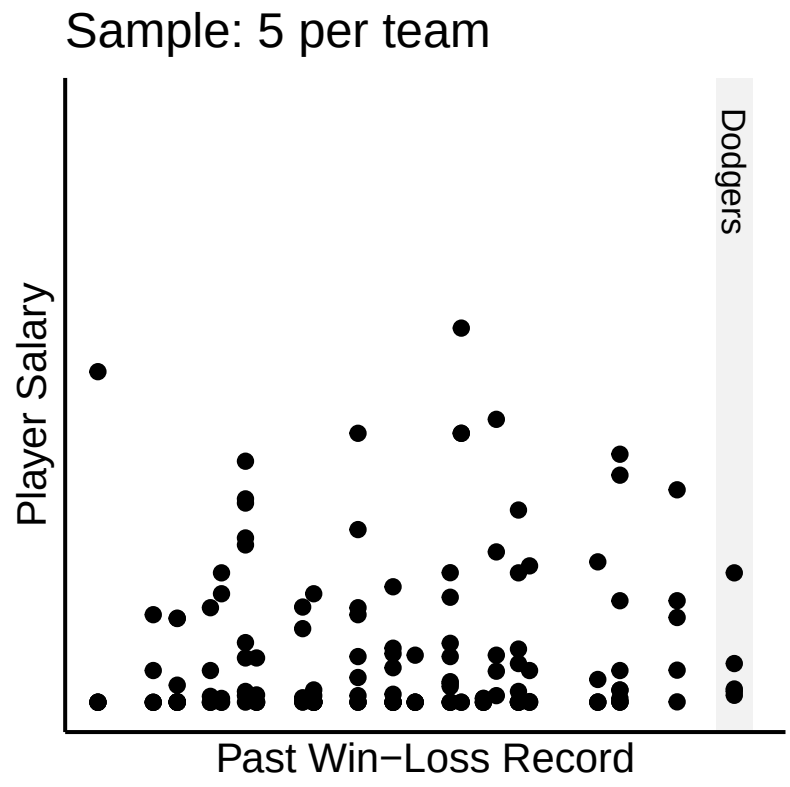
statistical learning: the idea

illustrated by a

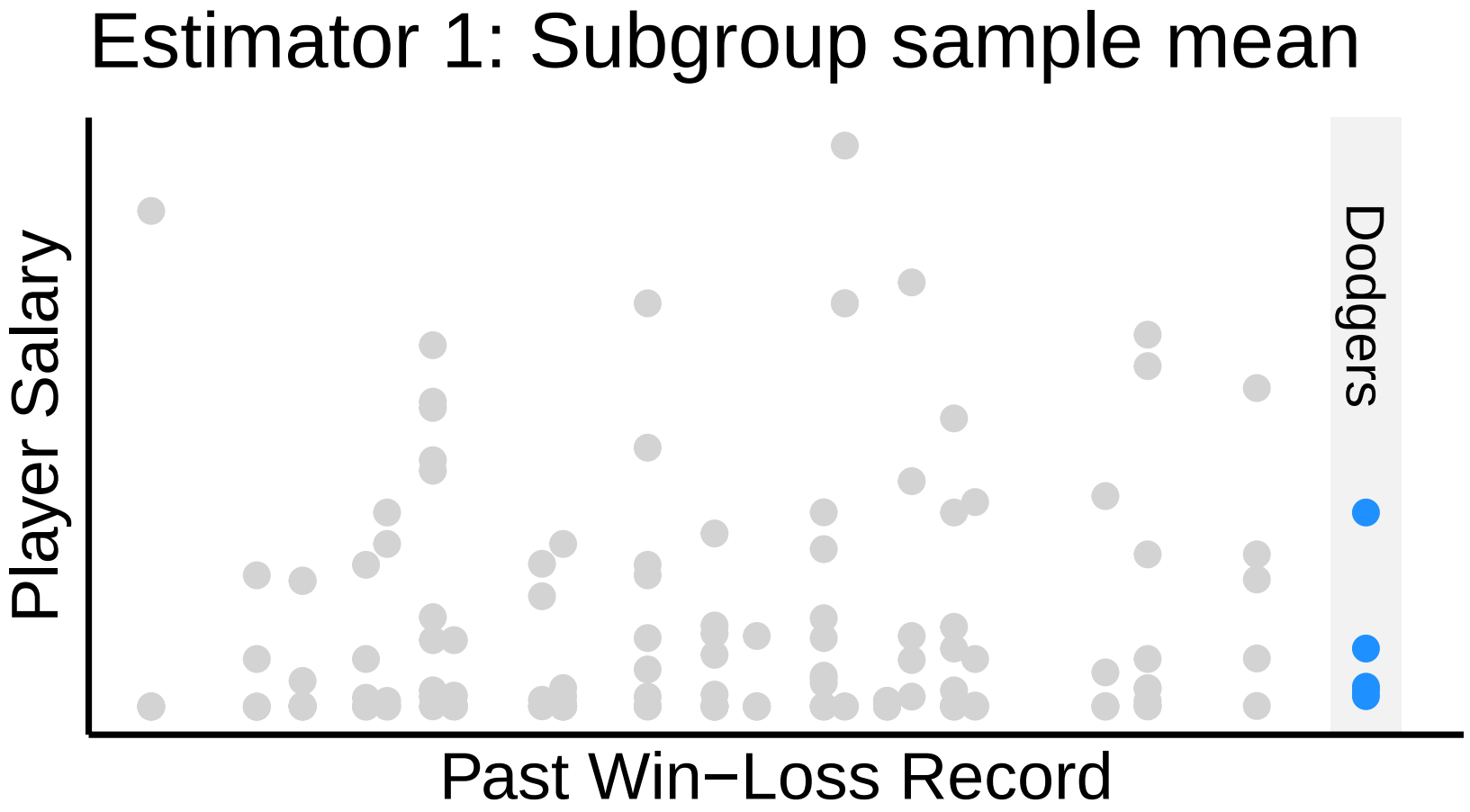
- discrete numeric predictor
- continuous numeric predictor

With only the sample, how would you estimate the mean salary of all the Dodgers?

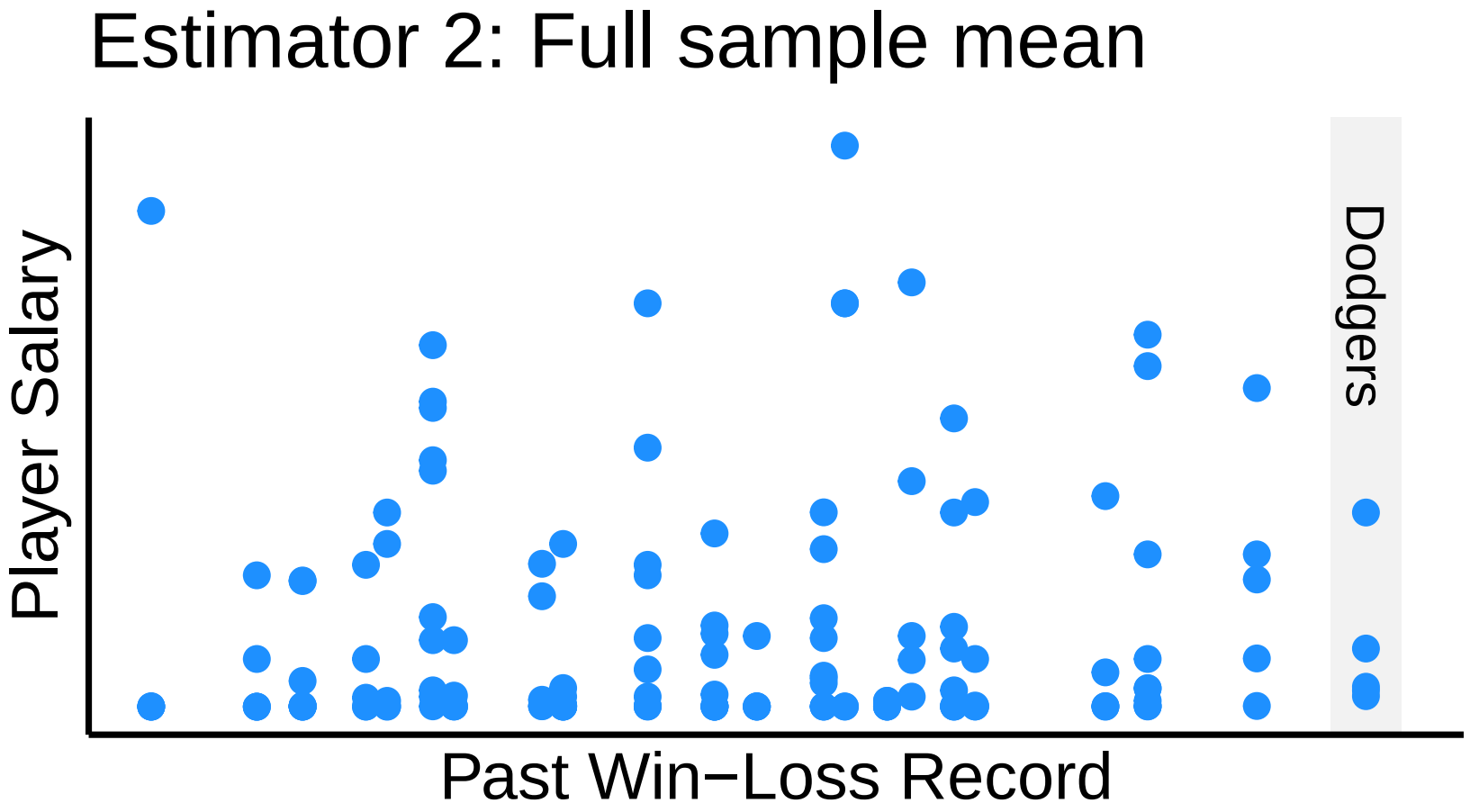




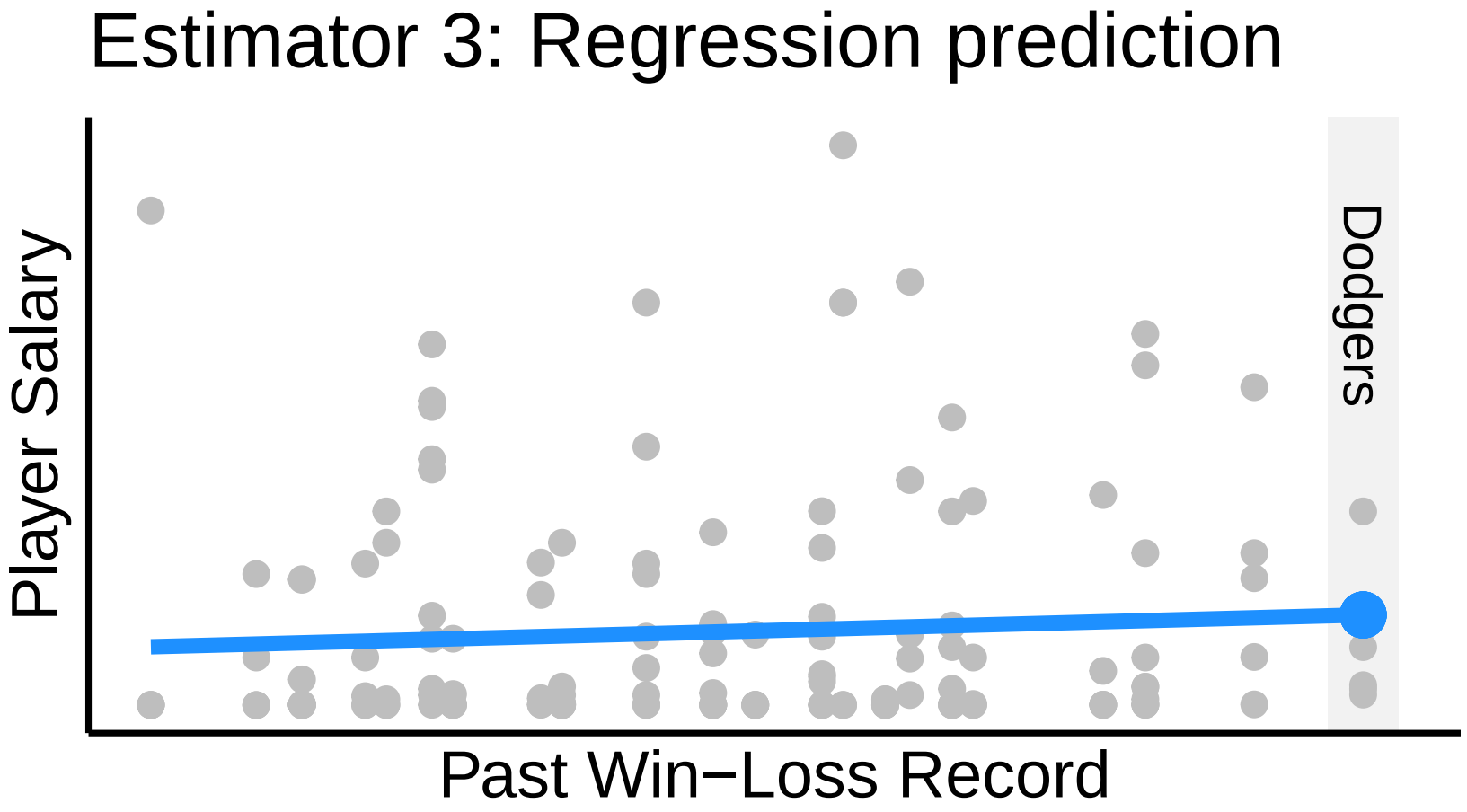
Three estimators for the Dodgers' mean salary



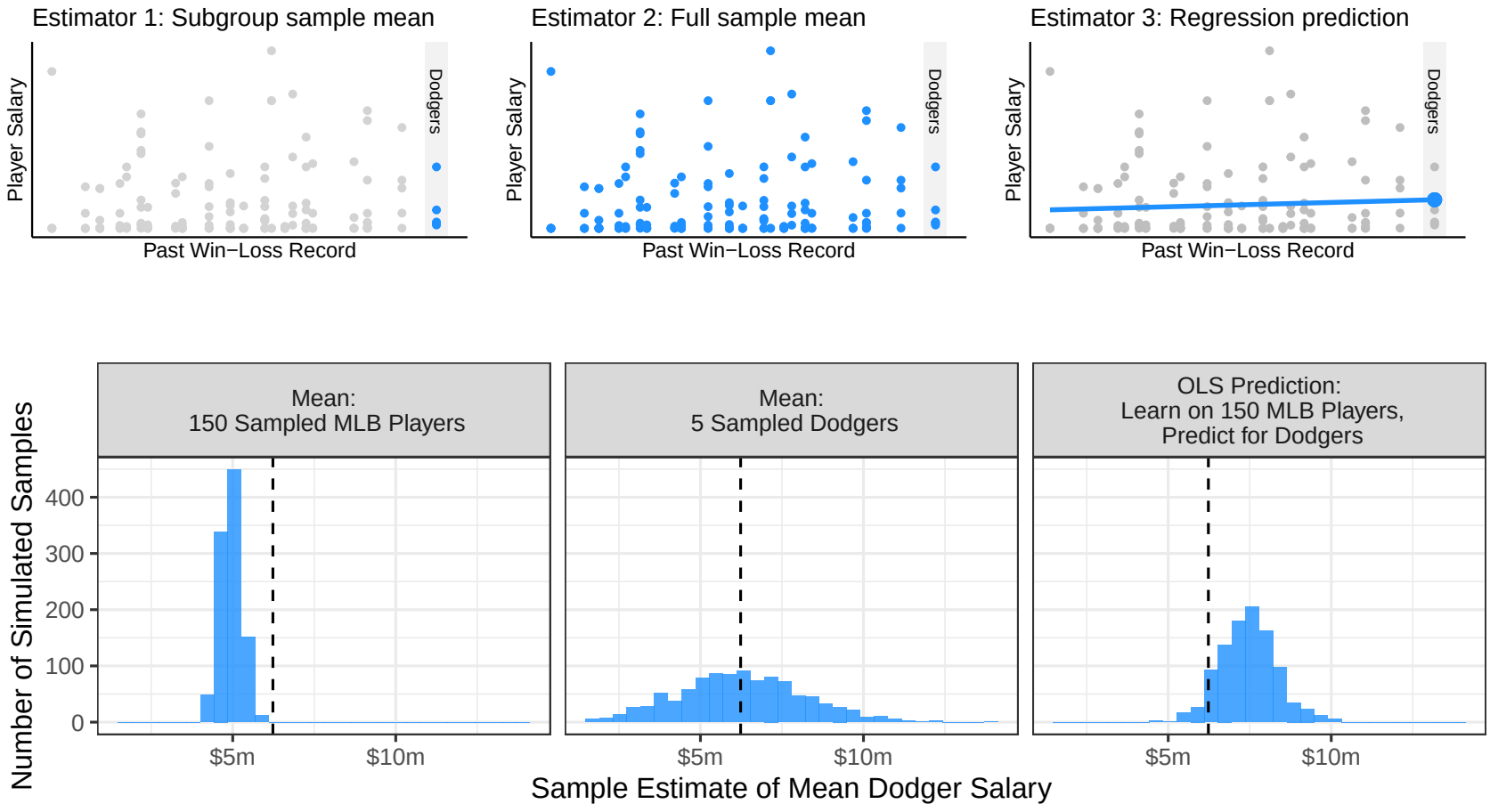
Three estimators for the Dodgers' mean salary



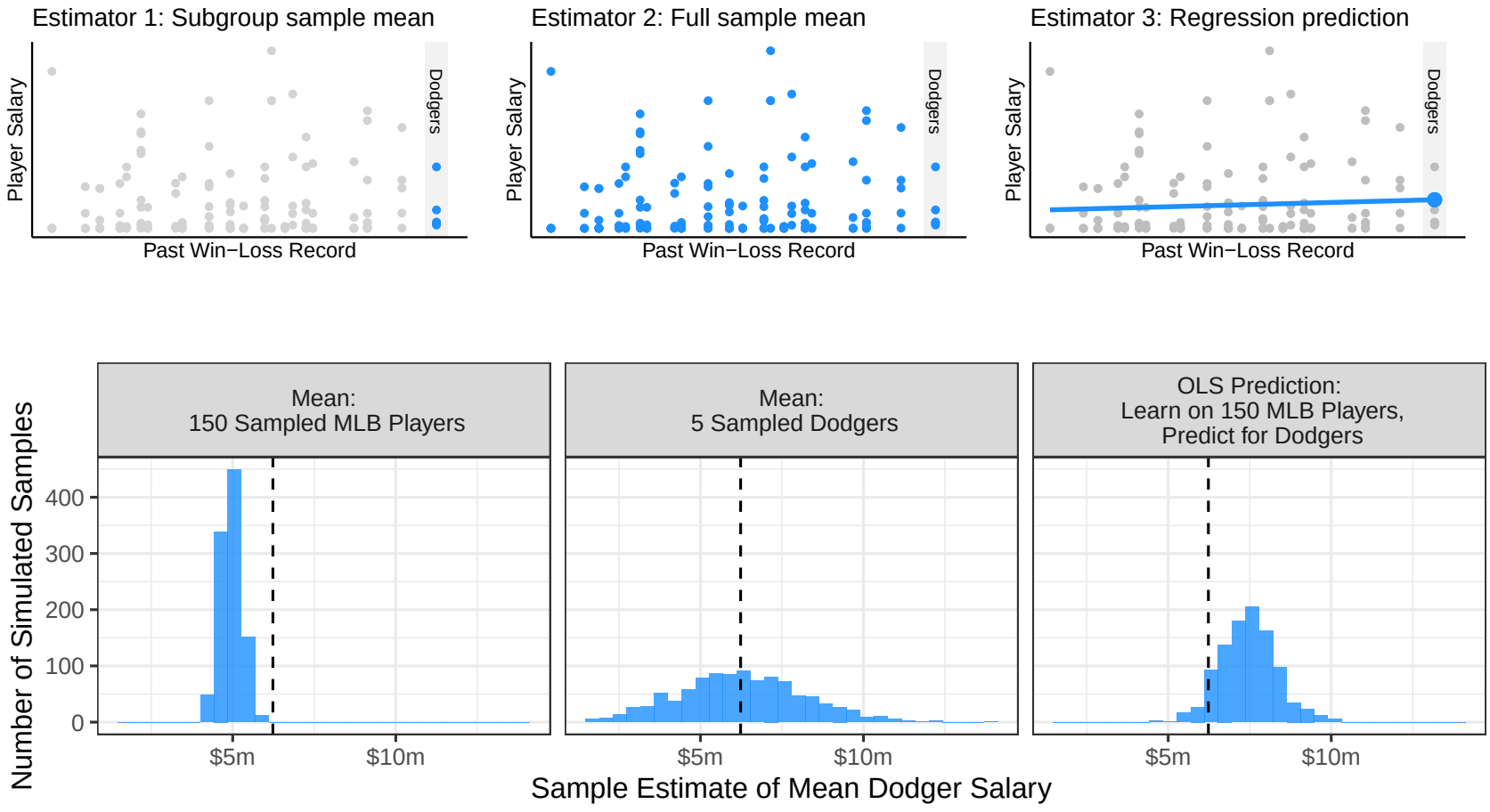
Three estimators for the Dodgers' mean salary



Three estimators for the Dodgers' mean salary



Three estimators for the Dodgers' mean salary



Which do you prefer? Why is your choice a little weird?

Statistical learning: A somewhat unusual view

Statistical learning: A somewhat unusual view

1. the entire goal of modeling is to solve sparse data
   - we sample very few Dodgers,
     so we use non-Dodgers to help our estimate

Statistical learning: A somewhat unusual view

1. the entire goal of modeling is to solve sparse data
   - we sample very few Dodgers,
     so we use non-Dodgers to help our estimate
2. in a huge sample, a model is unnecessary
   - estimate Dodger population mean
     by the Dodger sample mean

Statistical learning: A somewhat unusual view

1. the entire goal of modeling is to solve sparse data
   - we sample very few Dodgers,
     so we use non-Dodgers to help our estimate
2. in a huge sample, a model is unnecessary
   - estimate Dodger population mean
     by the Dodger sample mean
3. in a tiny sample, models may perform poorly
   - might even better to estimate a subgroup mean (Dodgers)
     by taking the mean of the whole sample (all MLB)

statistical learning: the idea

illustrated by a

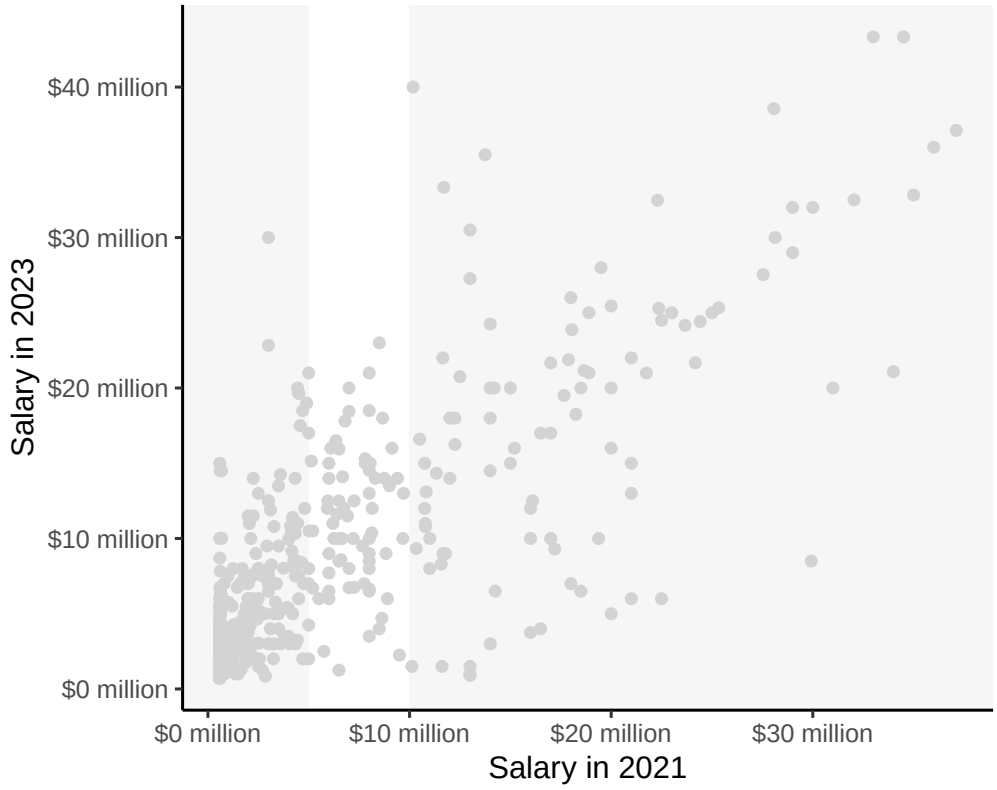
- discrete numeric predictor
- continuous numeric predictor

What is the mean 2023 salary among players
who in 2021 earned $5-10 million?

Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

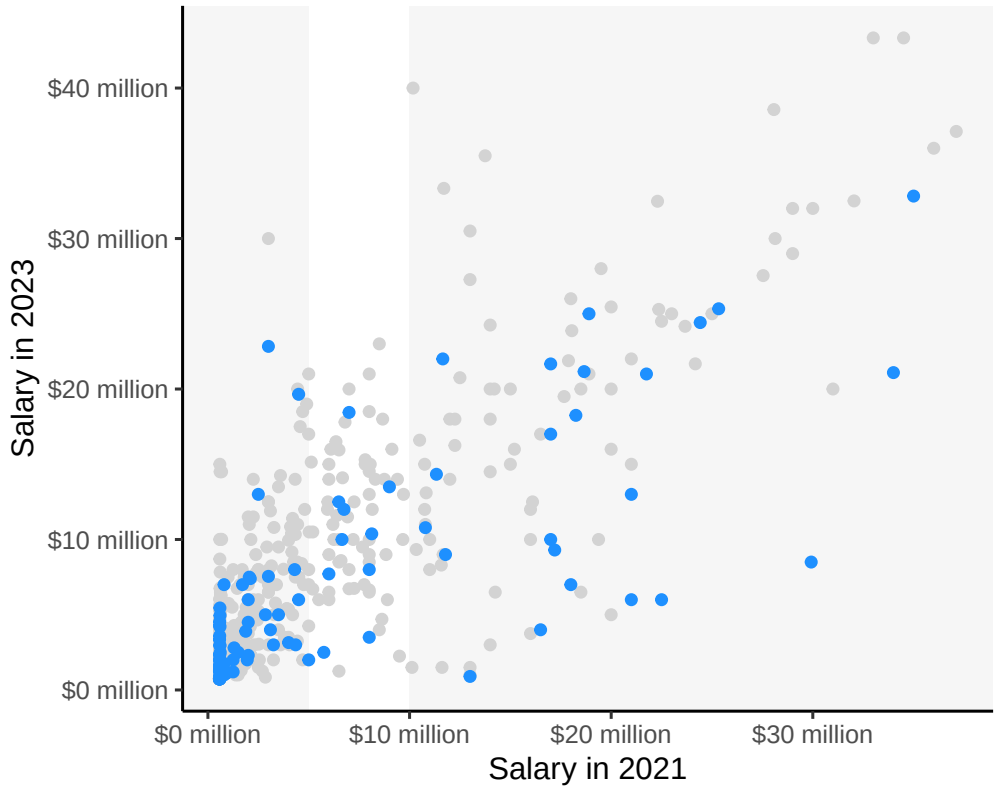
Begin with the population



Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

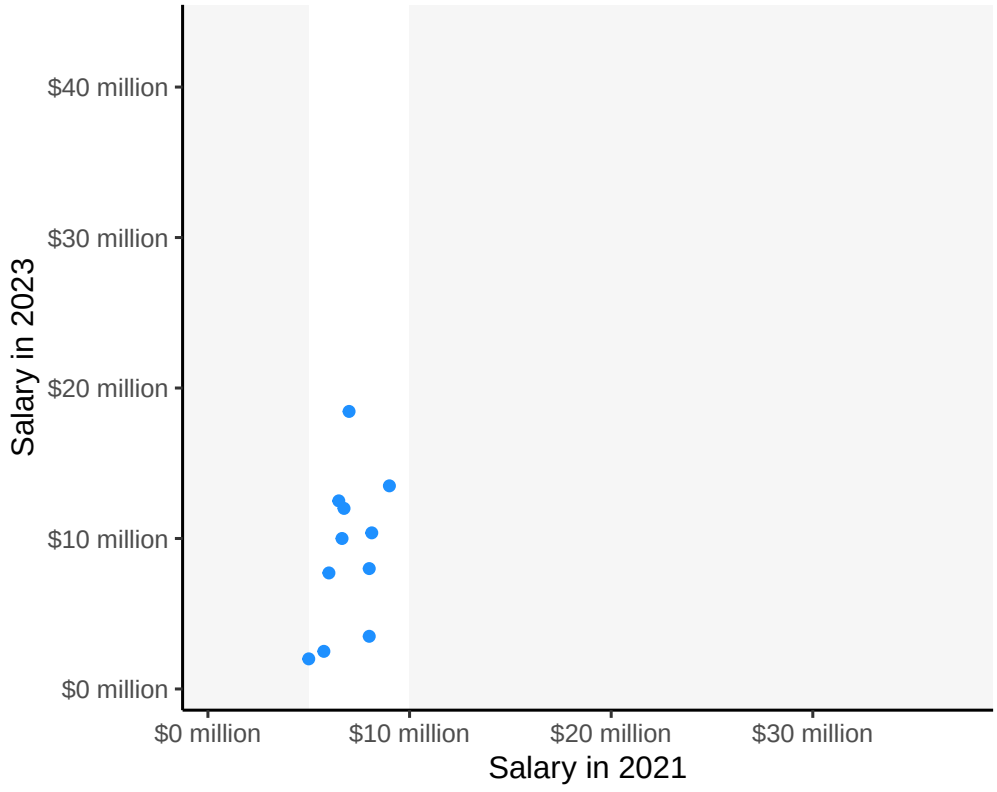
Sample



Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

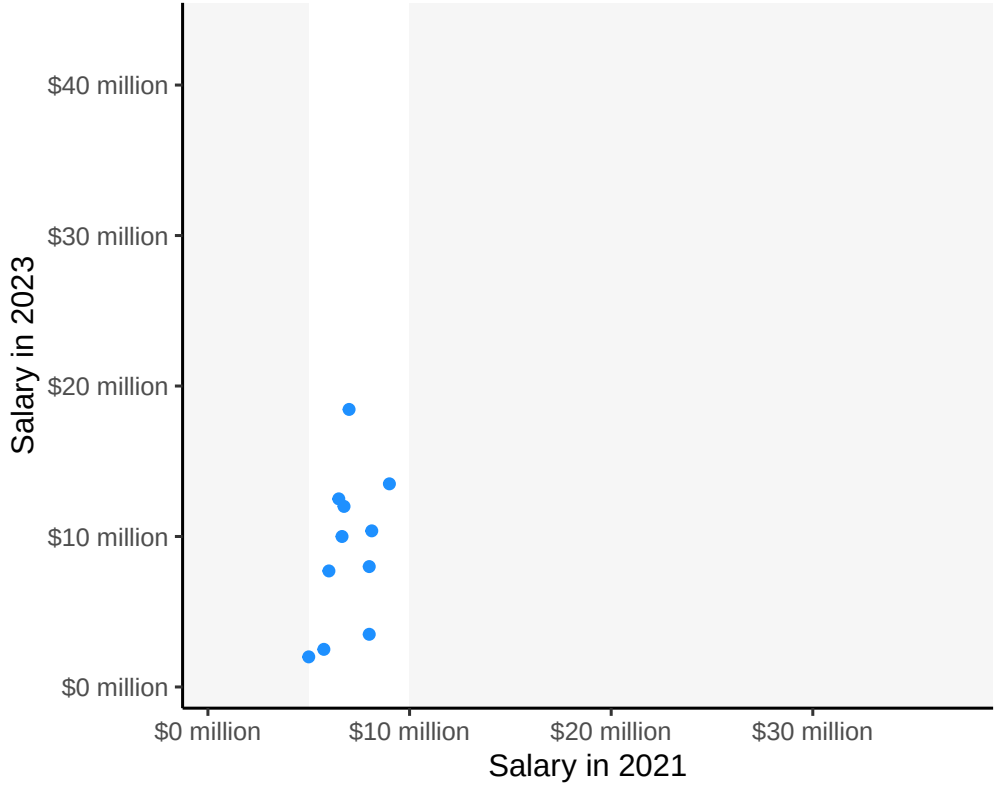
Sample



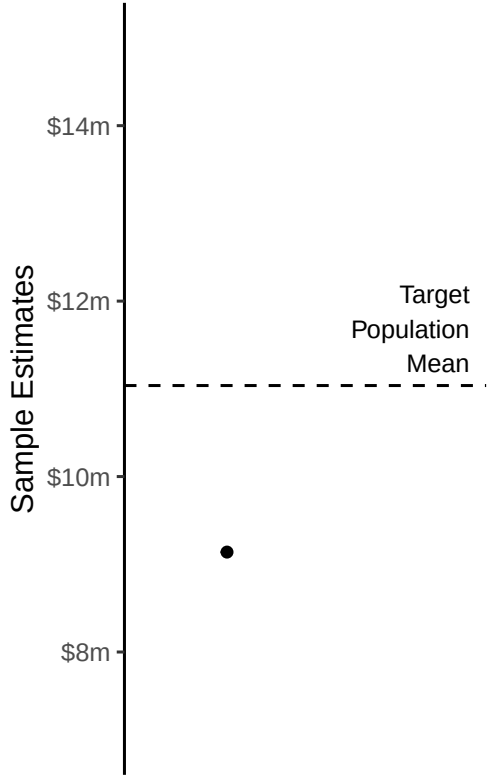
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



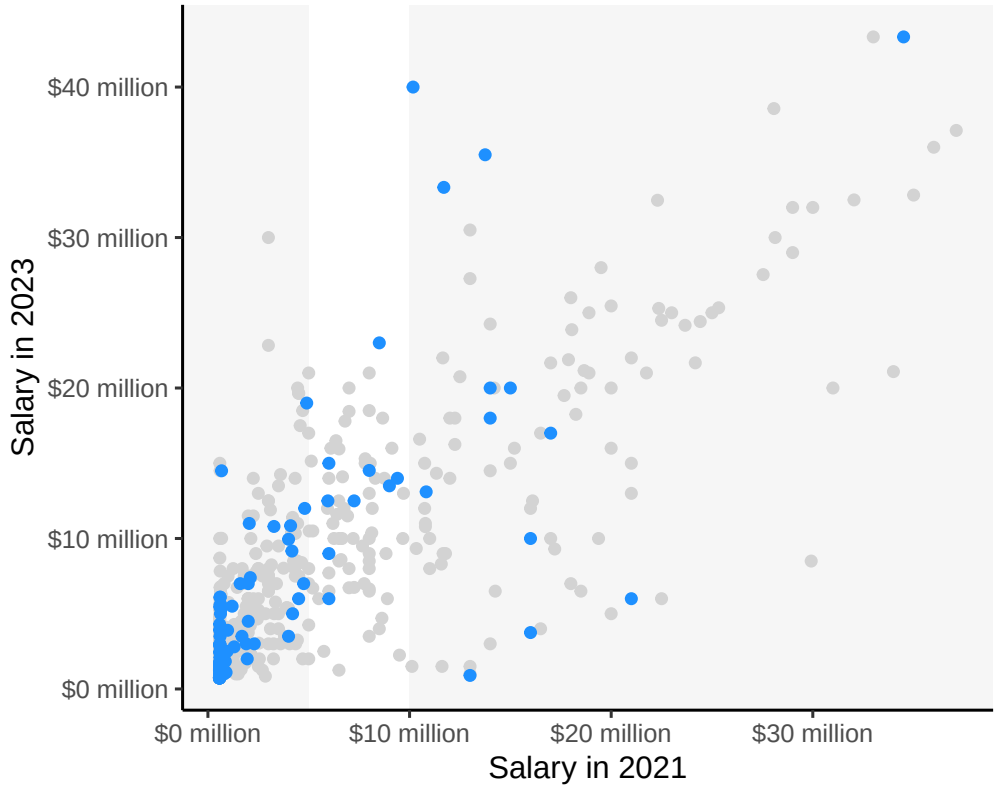
Sample average



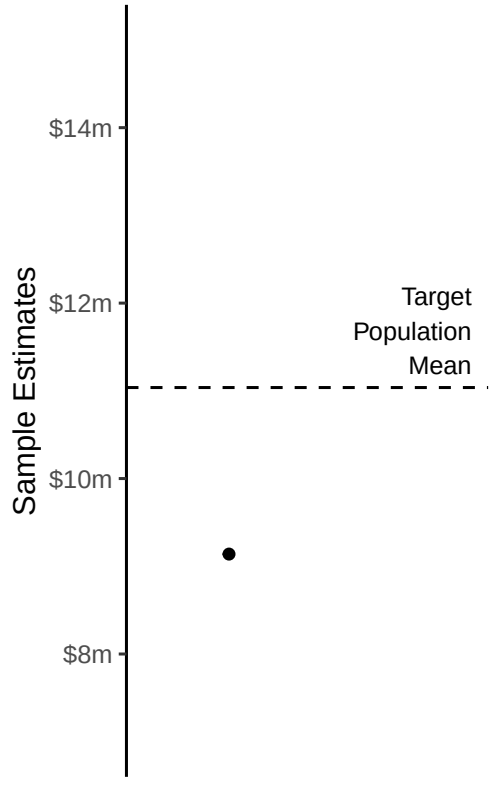
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



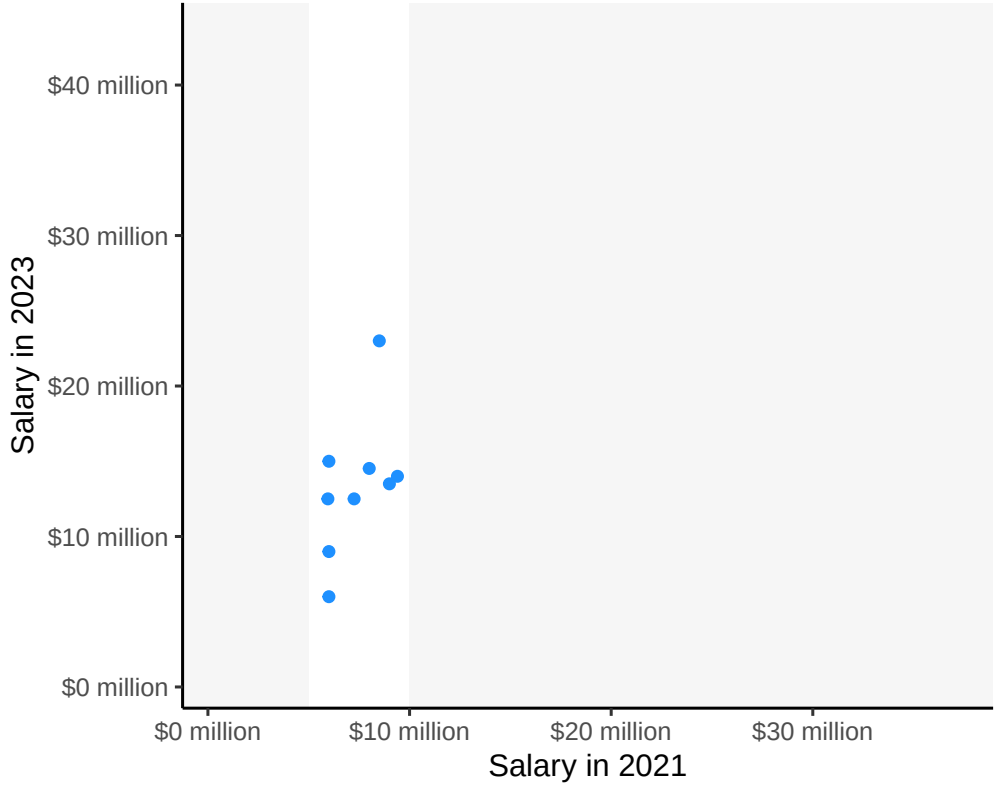
Sample average



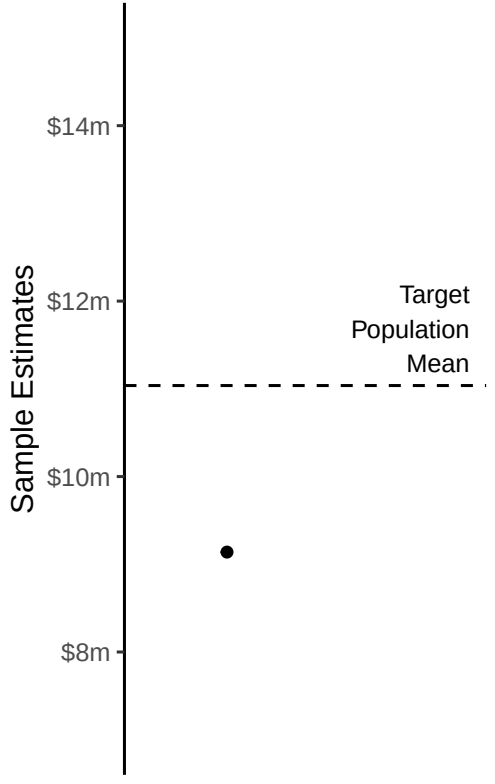
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



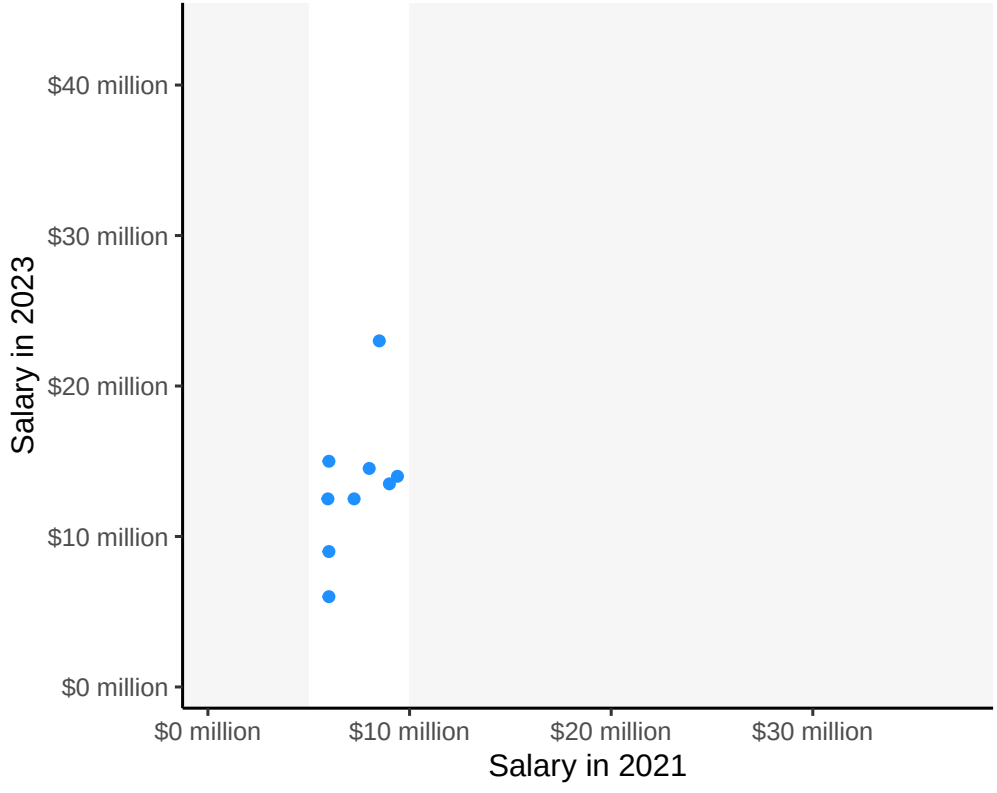
Sample average



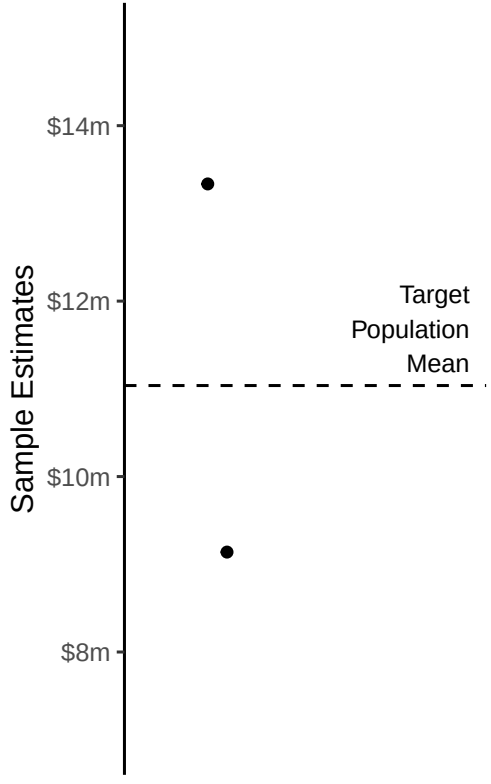
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



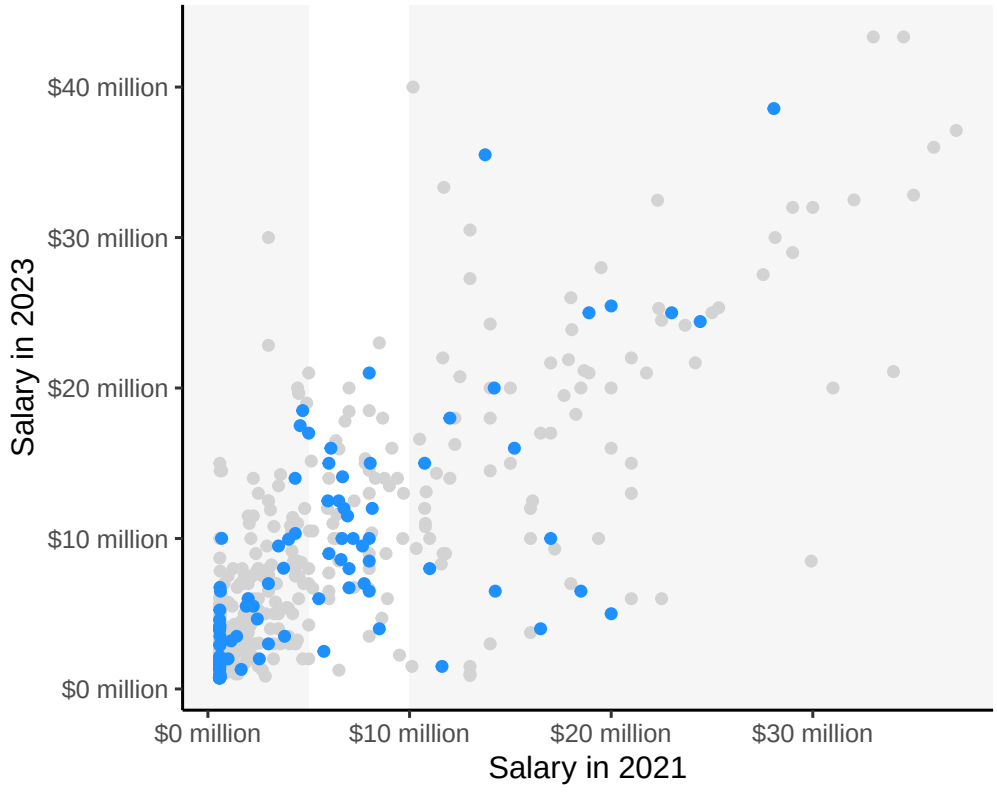
Sample average



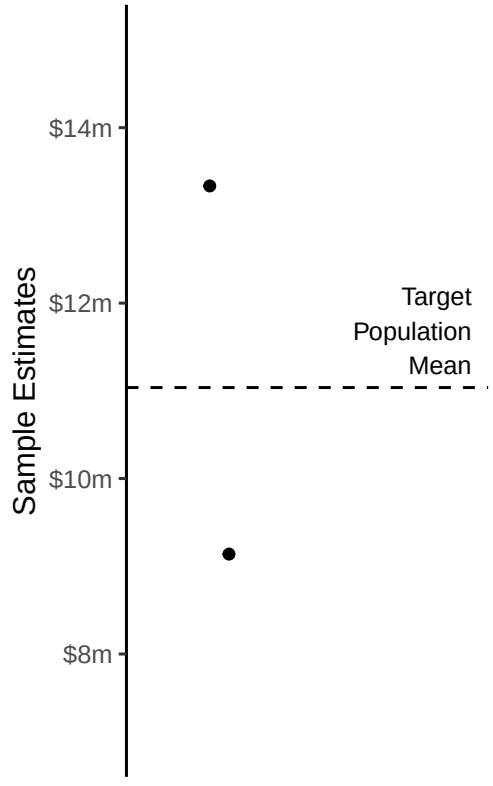
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



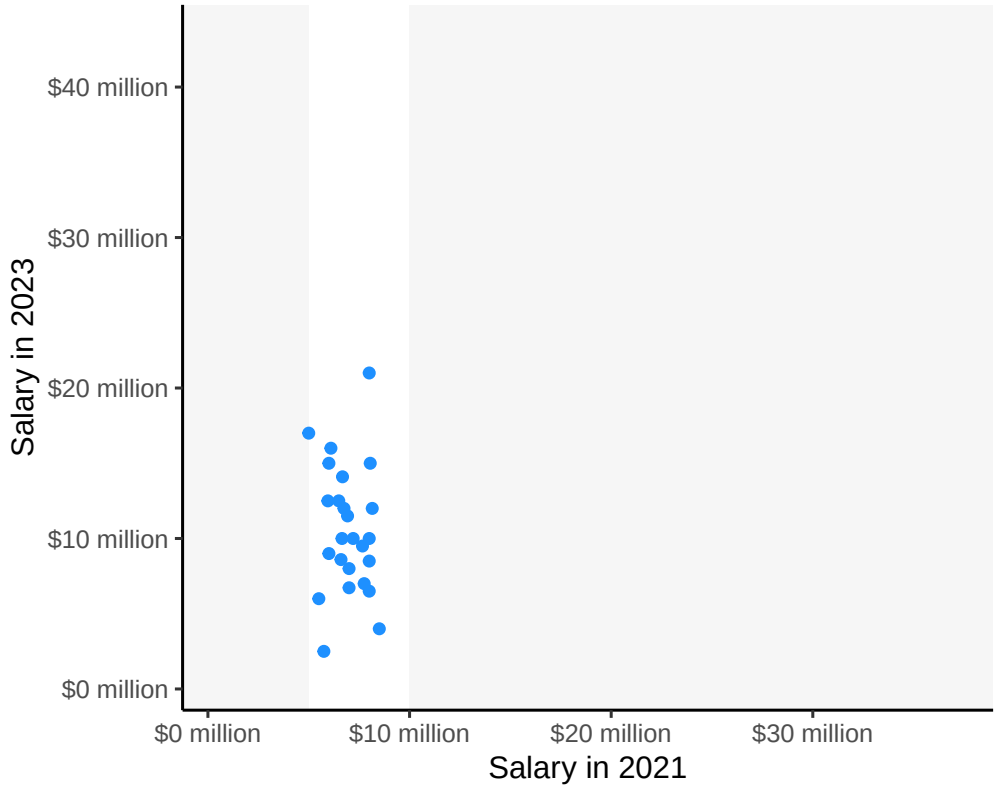
Sample average



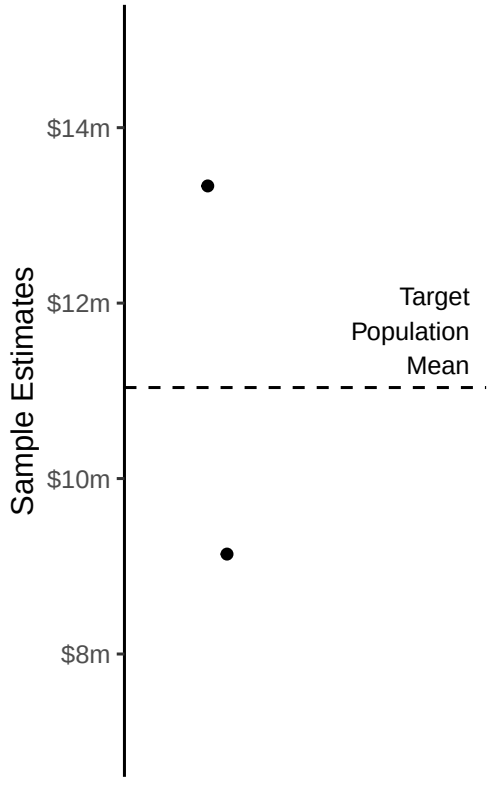
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



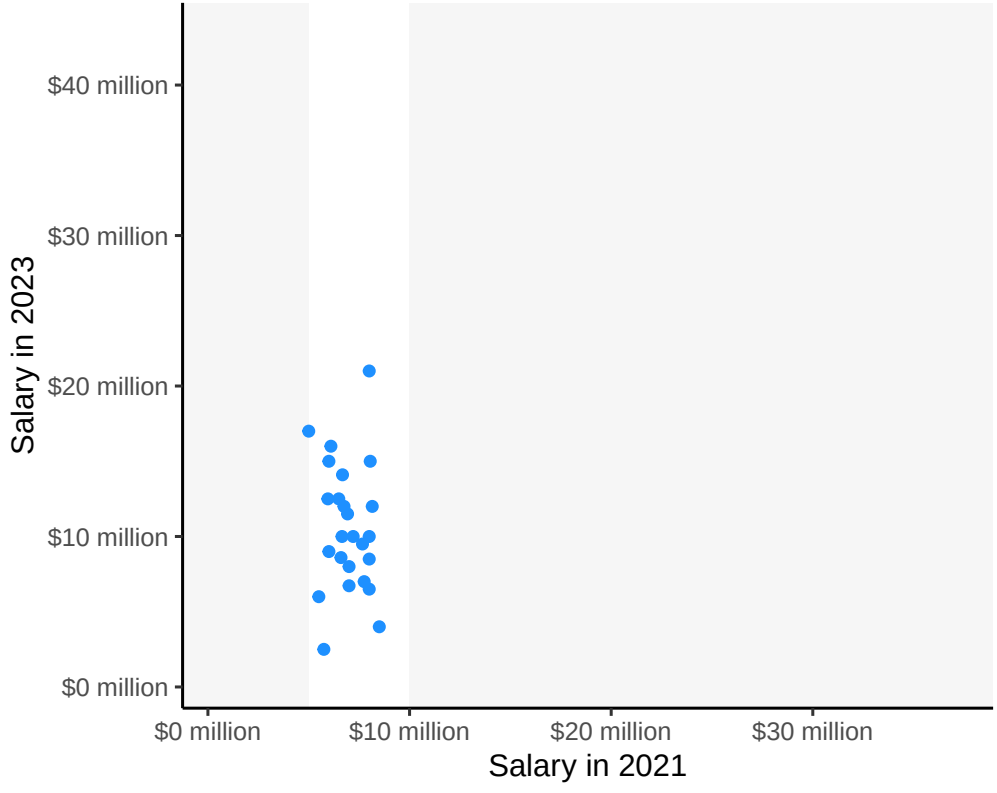
Sample average



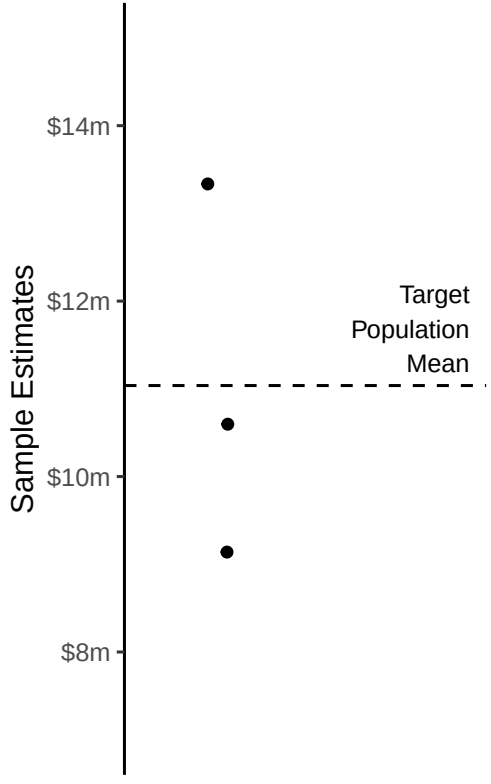
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



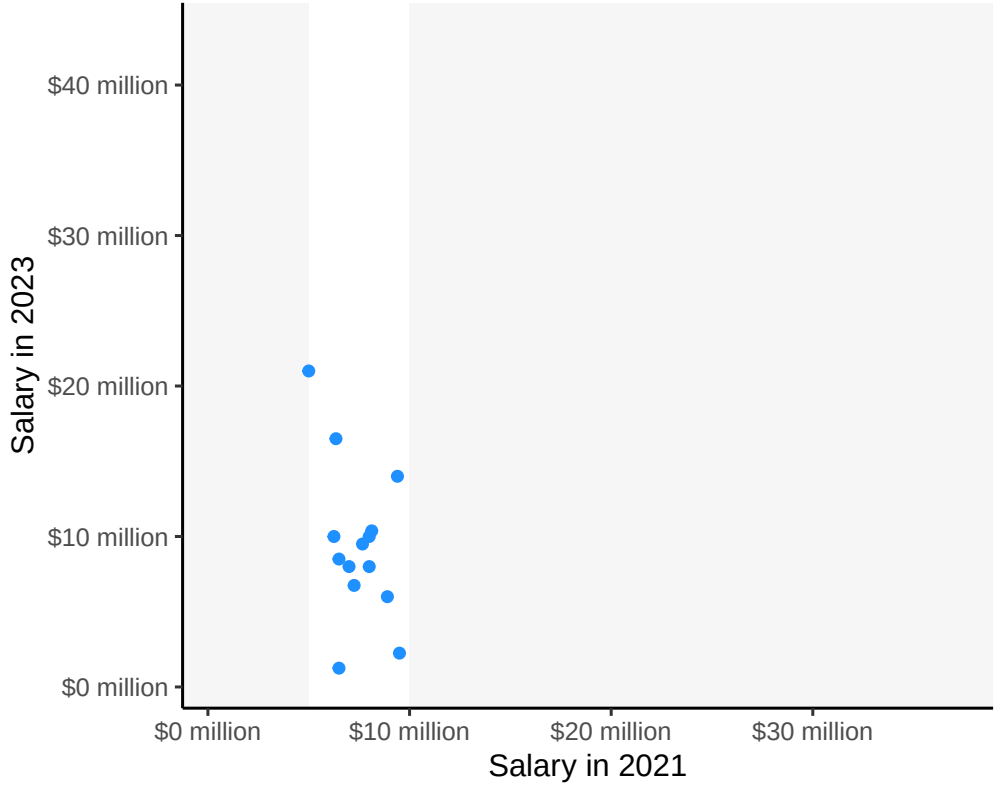
Sample average



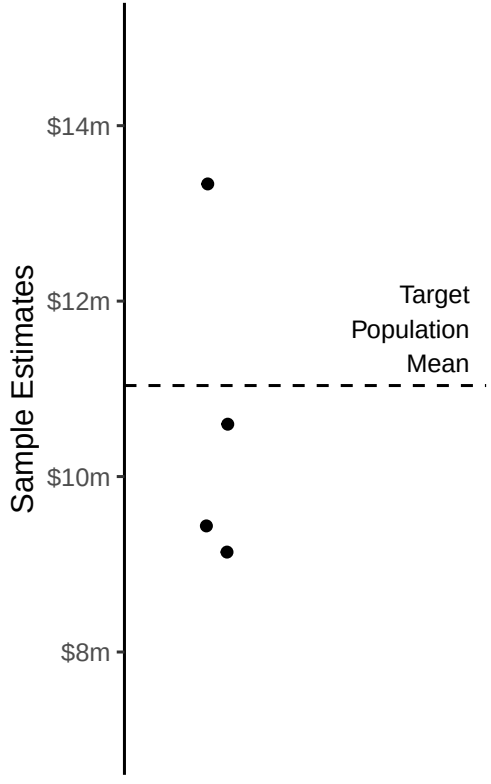
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



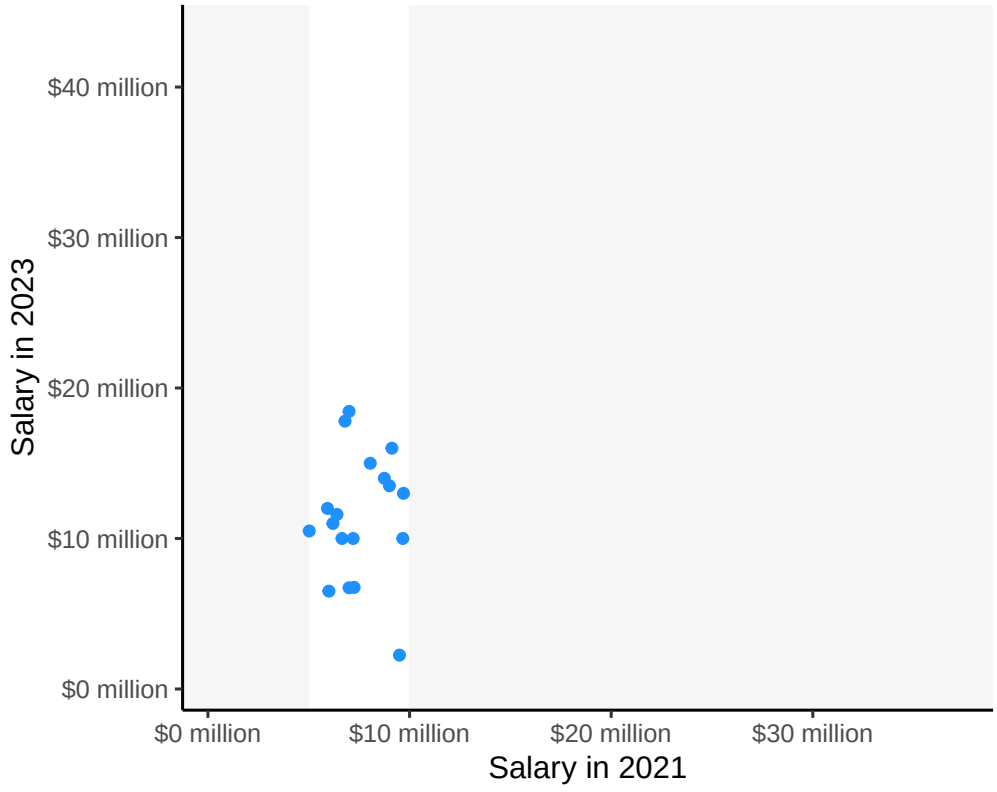
Sample average



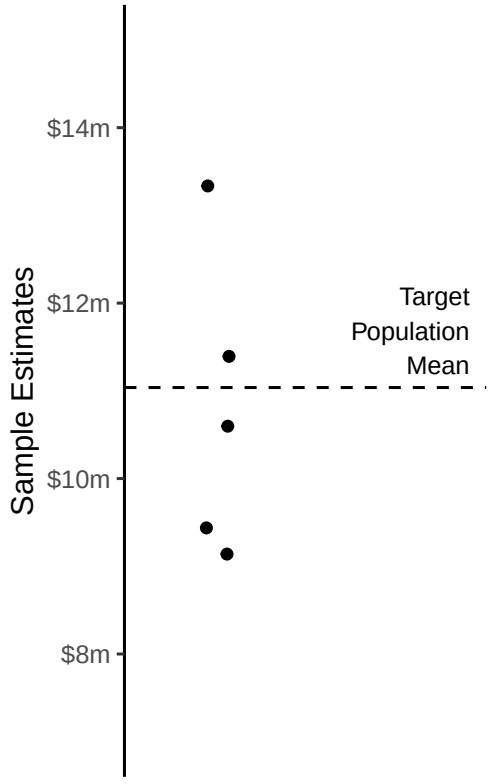
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



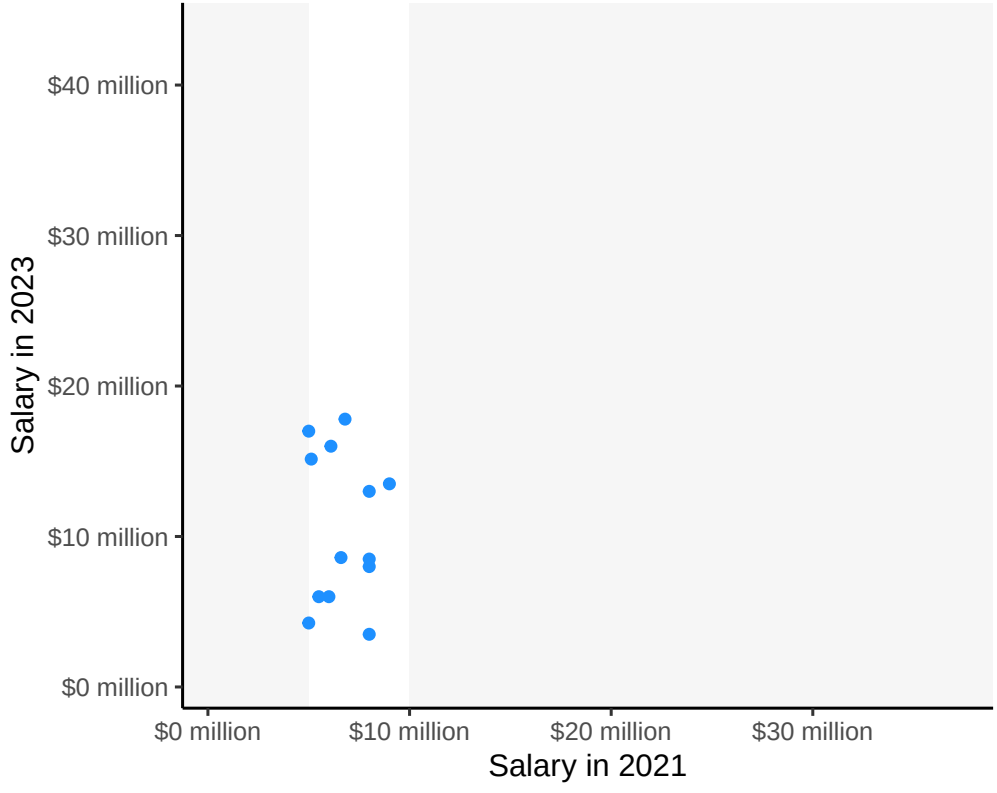
Sample average



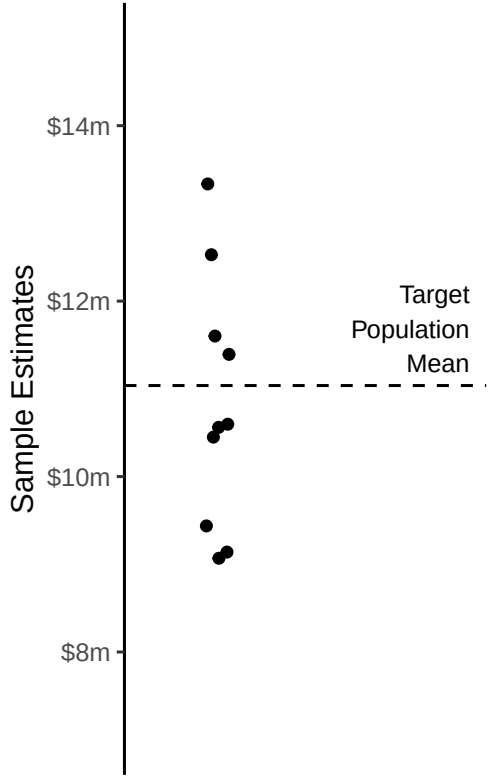
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



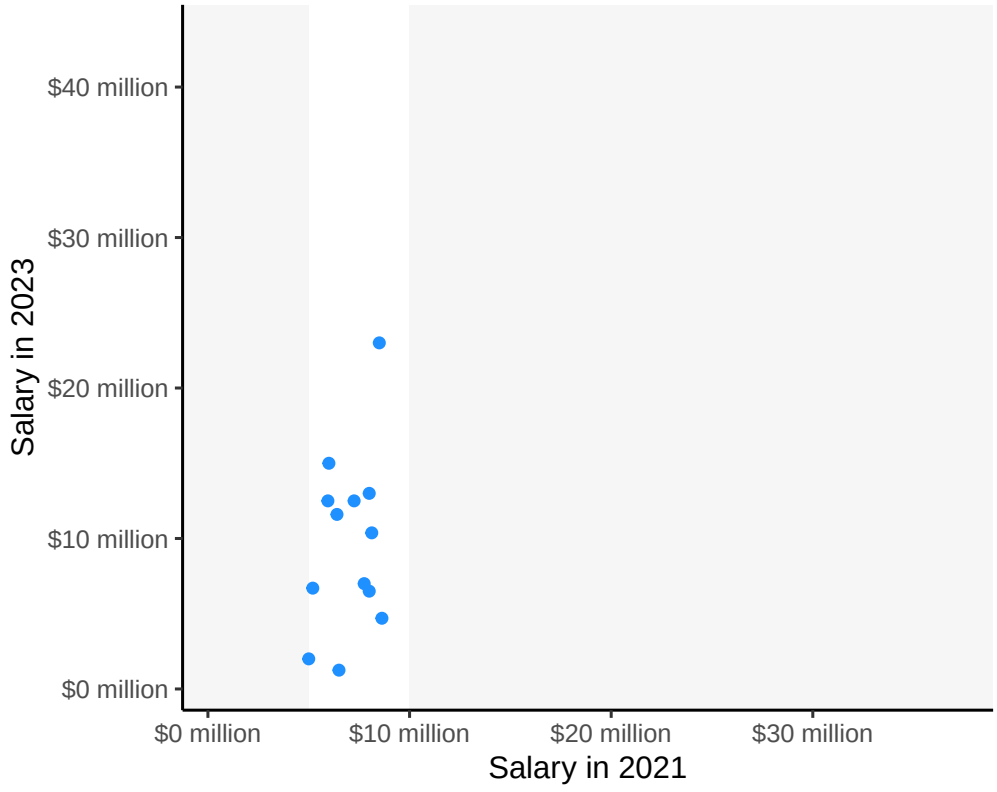
Sample average



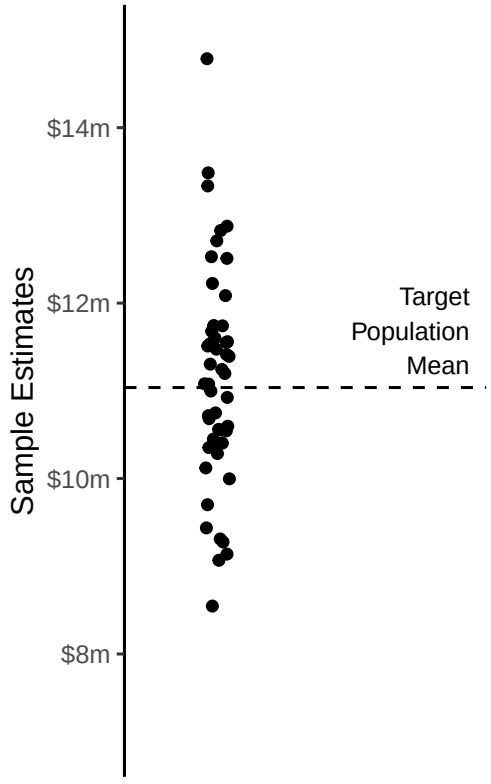
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



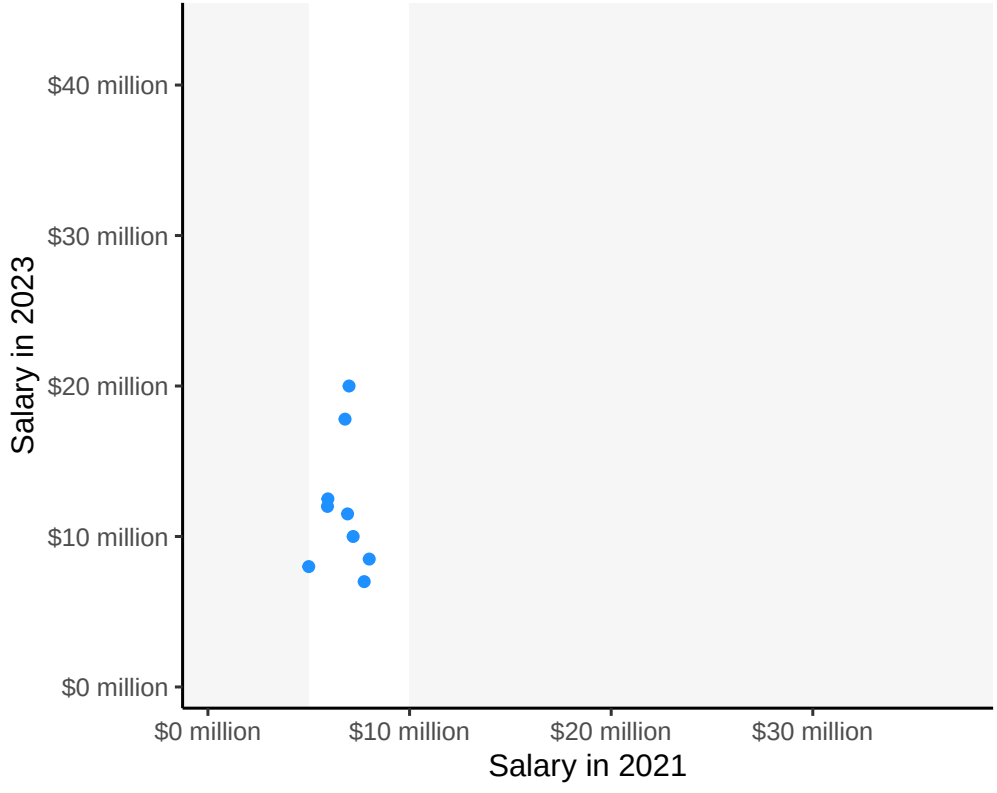
Sample average



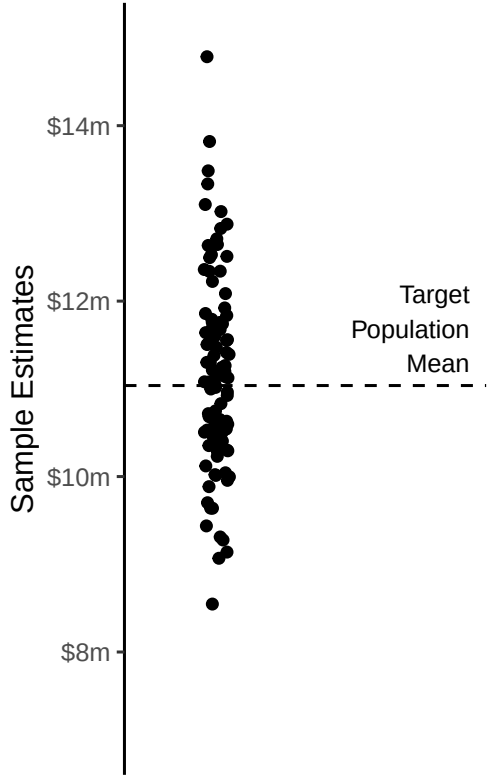
Goal: Estimate a target population mean from a sample

Method: Sample subgroup mean

Sample



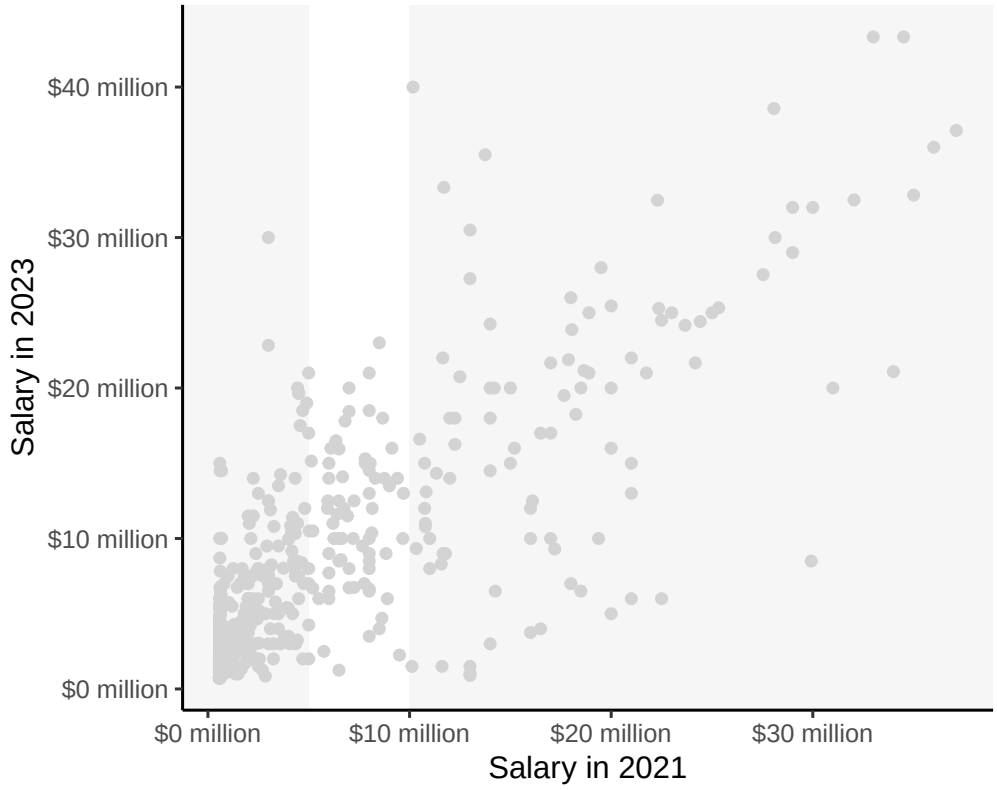
Sample average



Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

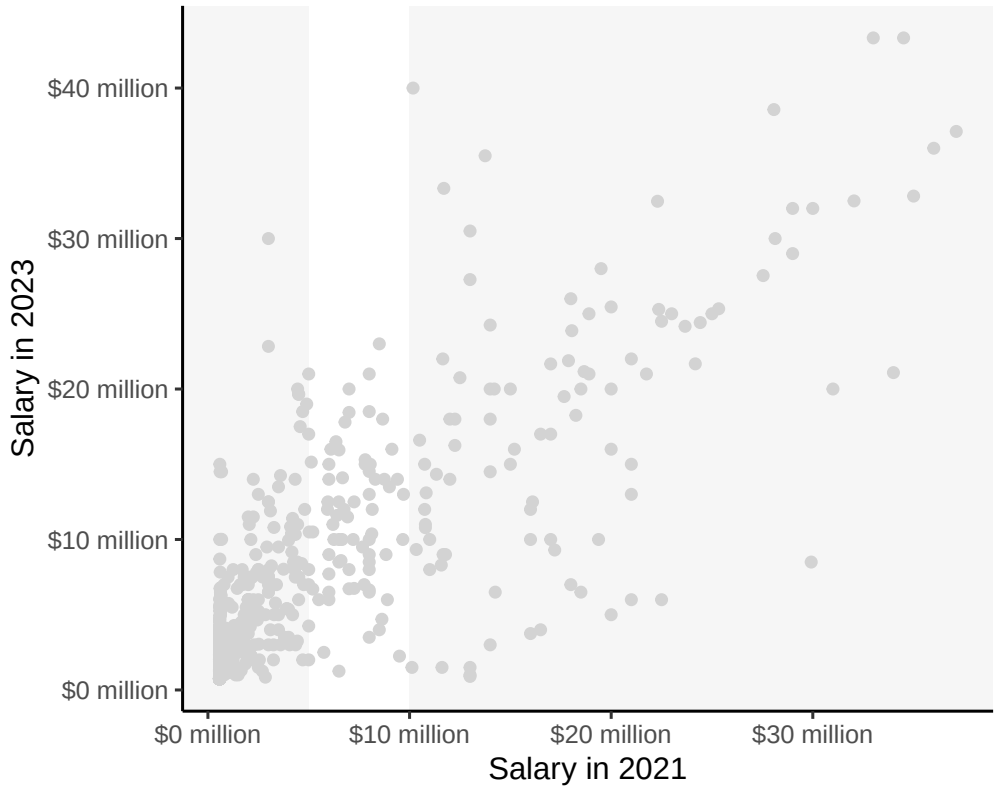
How would you use a model?



Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

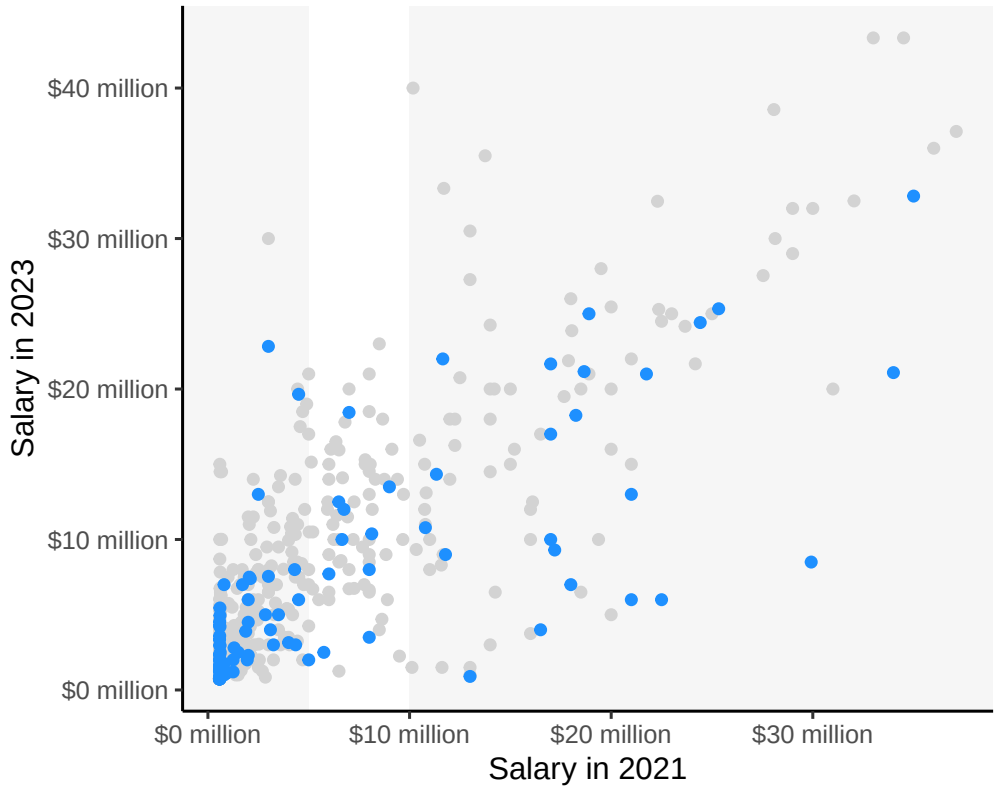
Begin with the population



Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

Draw a sample



Goal: Estimate a target population mean from a sample

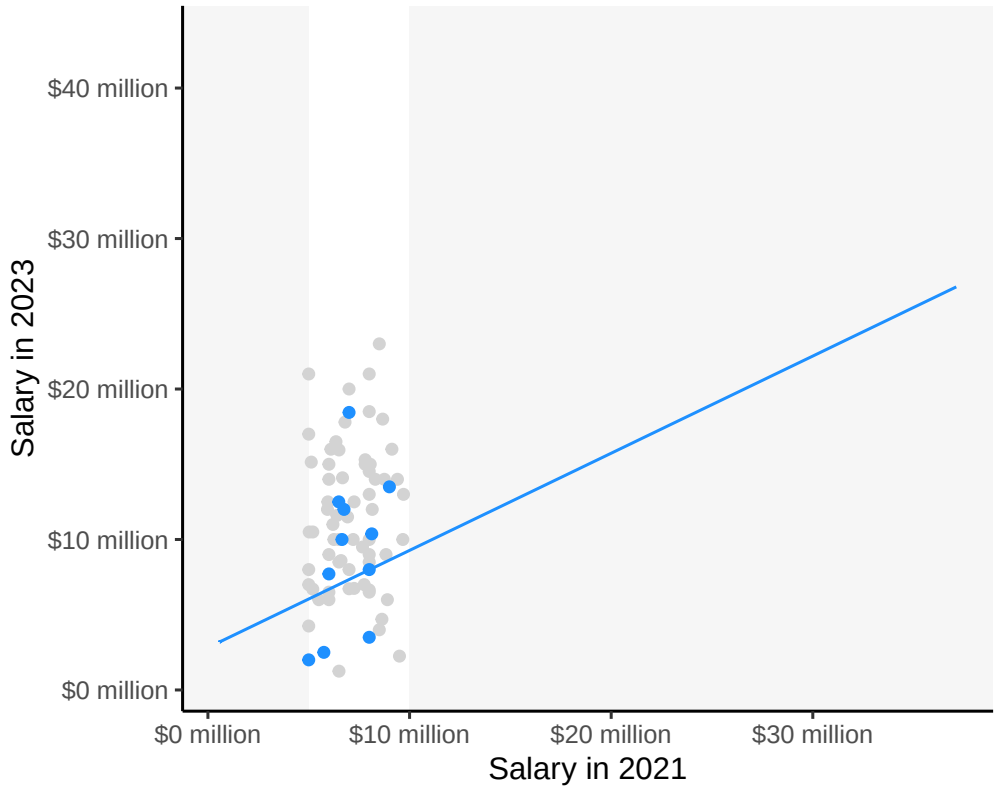
Method: Ordinary Least Squares prediction

Learn a model

Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

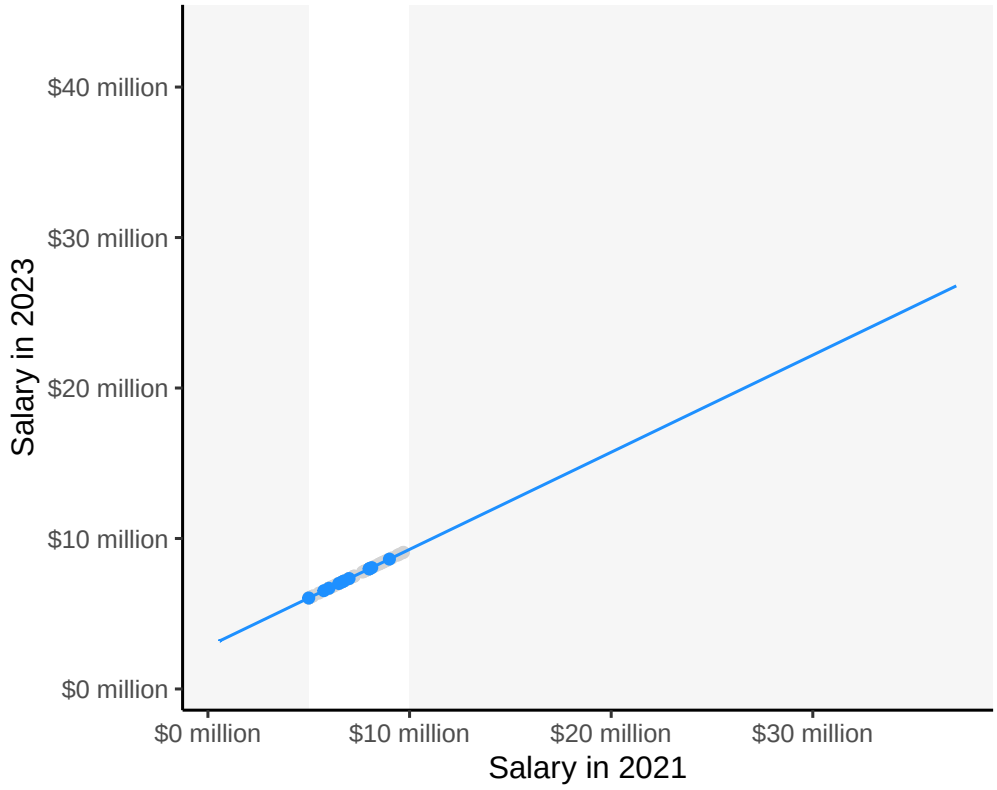
Focus on the target population



Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

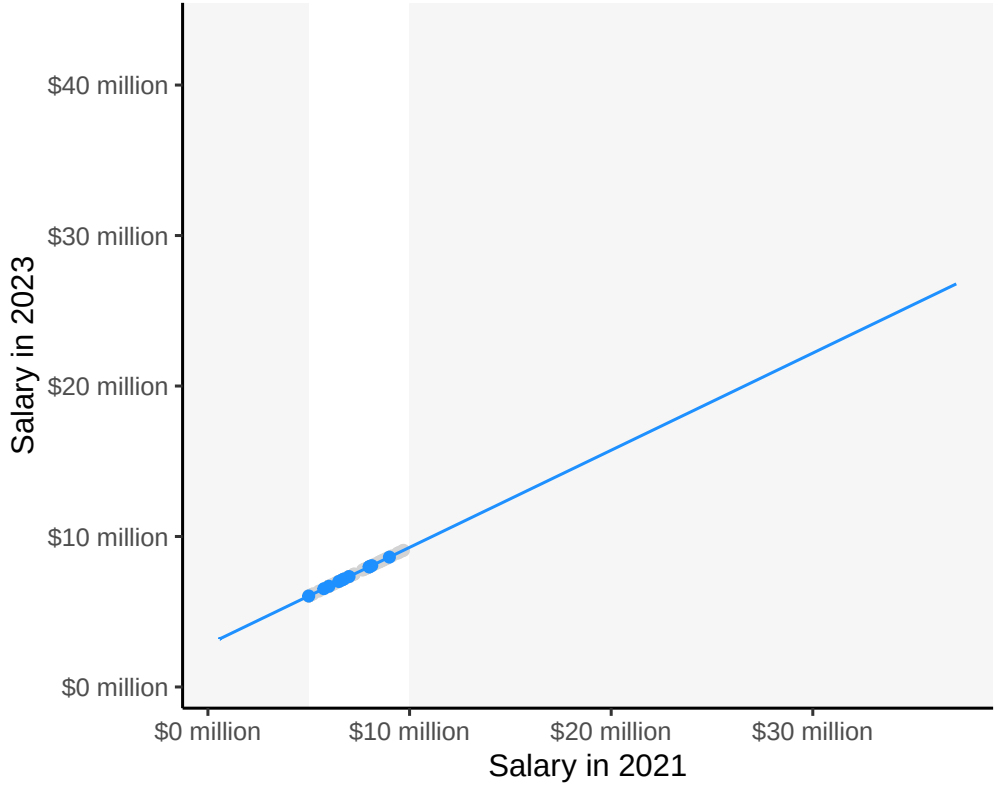
Predict



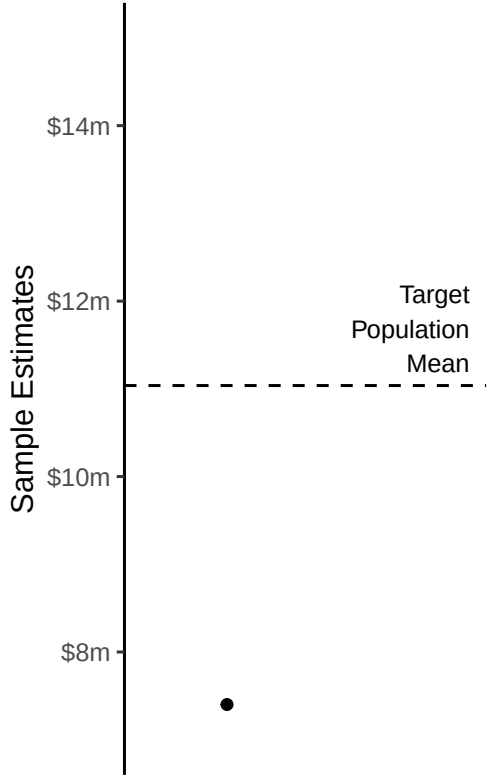
Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

Predict



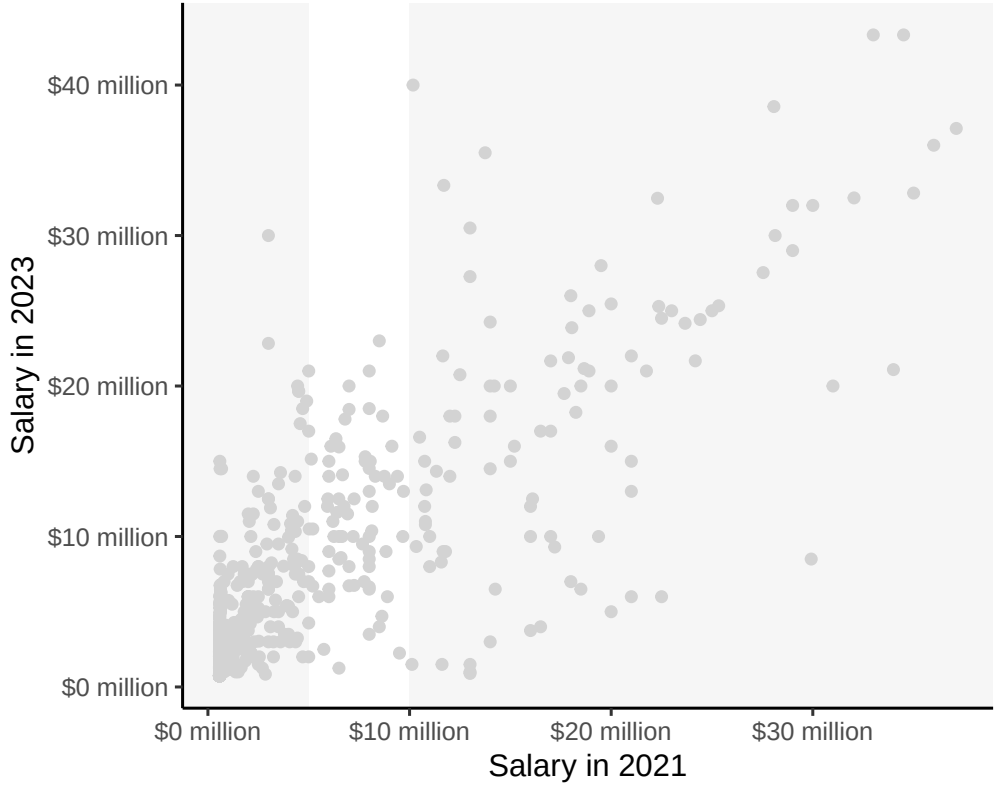
Record the average

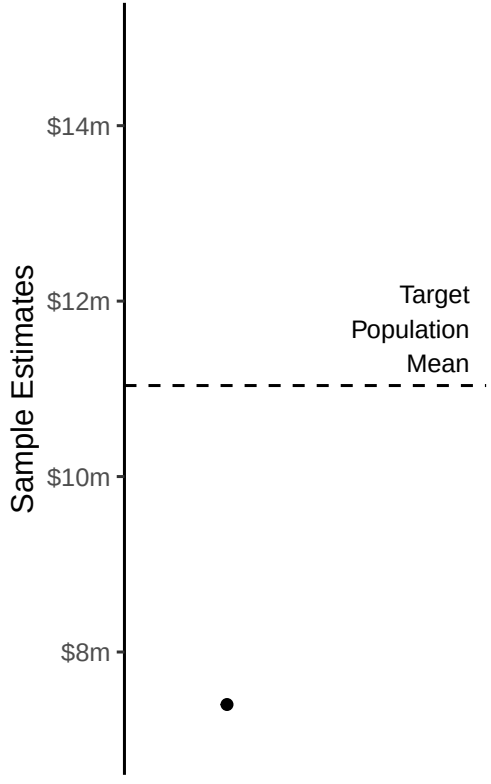


Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

Begin with the population

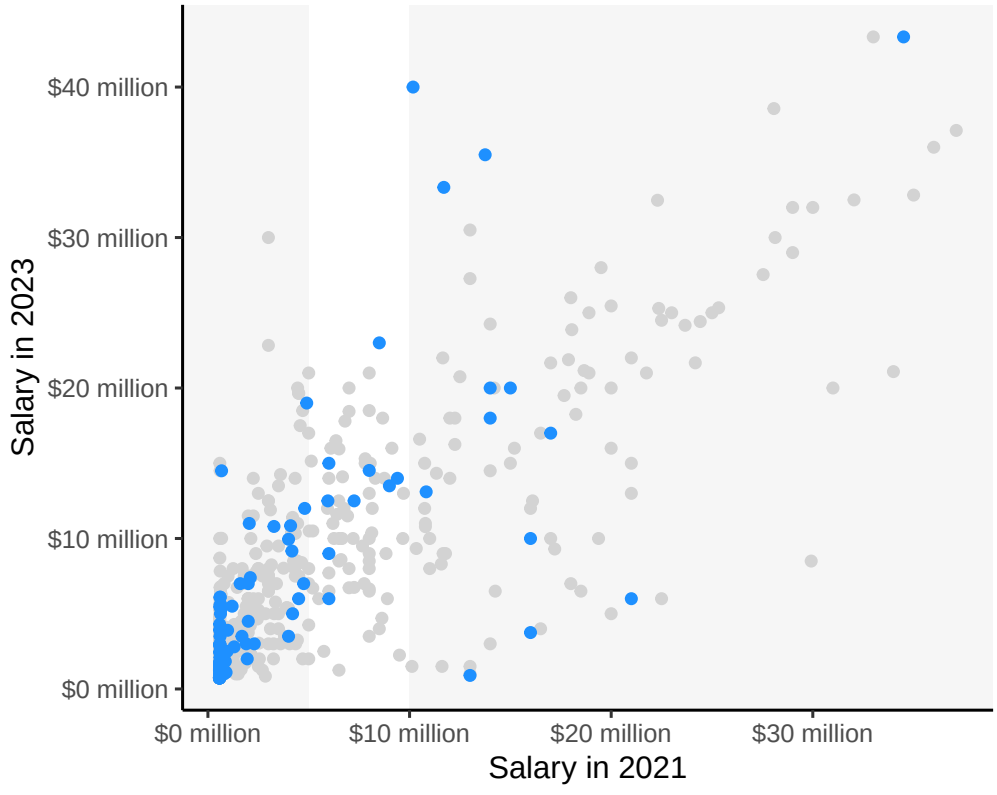


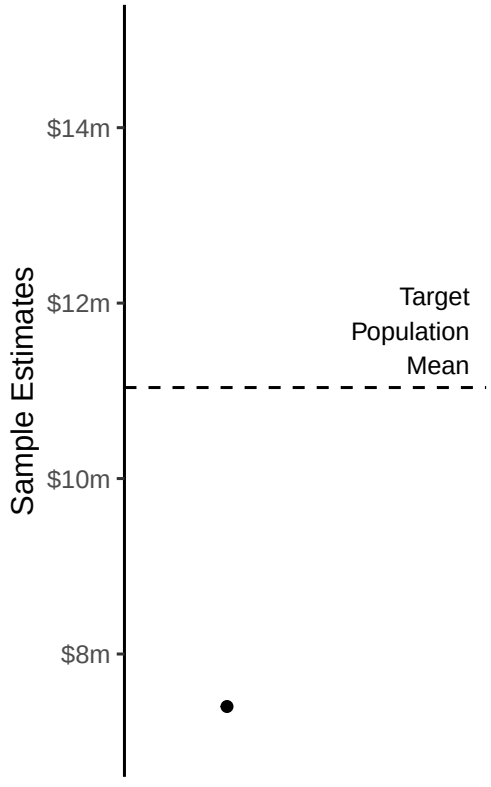


Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

Draw a sample

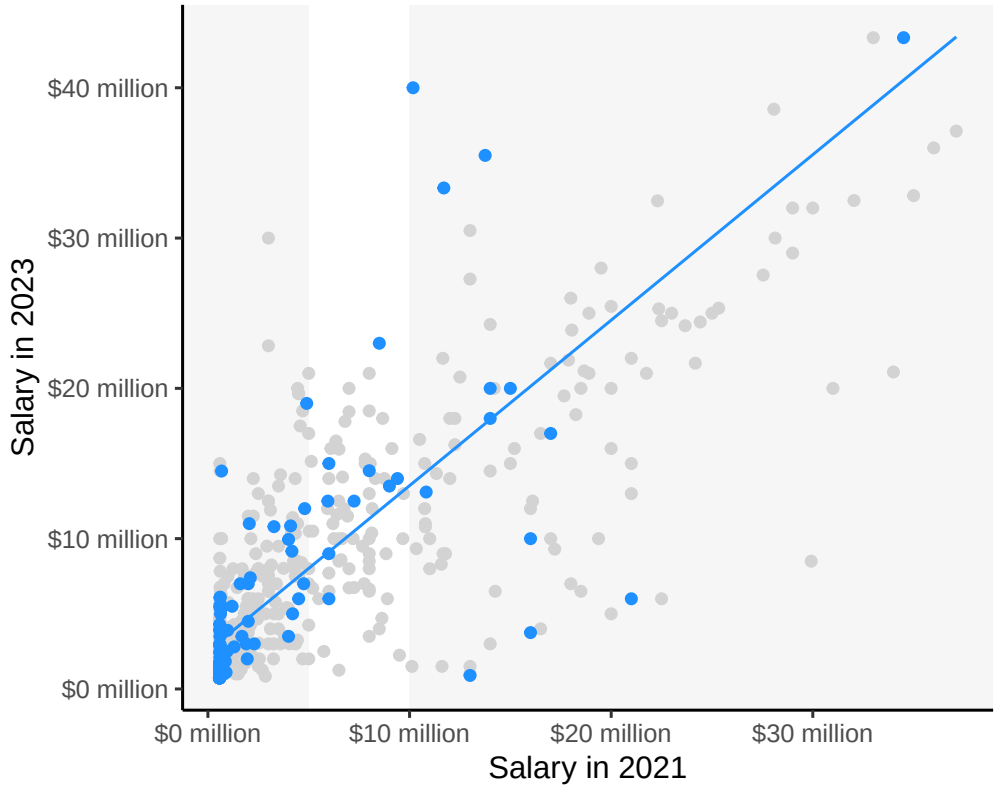


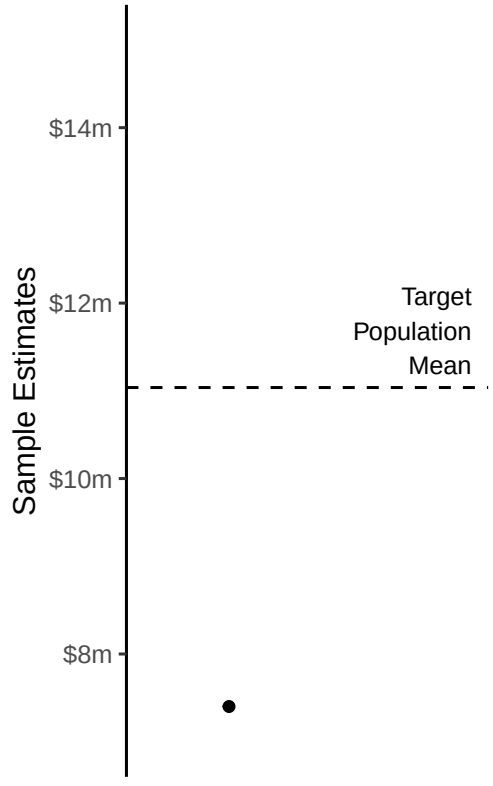


Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

Learn a model

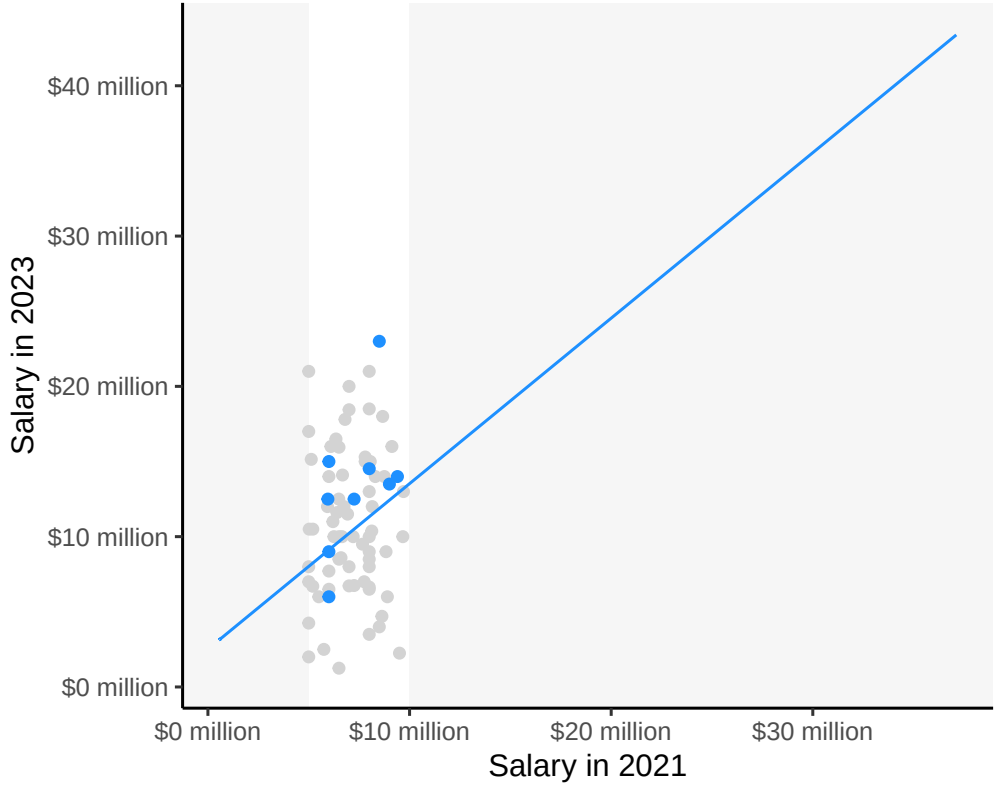


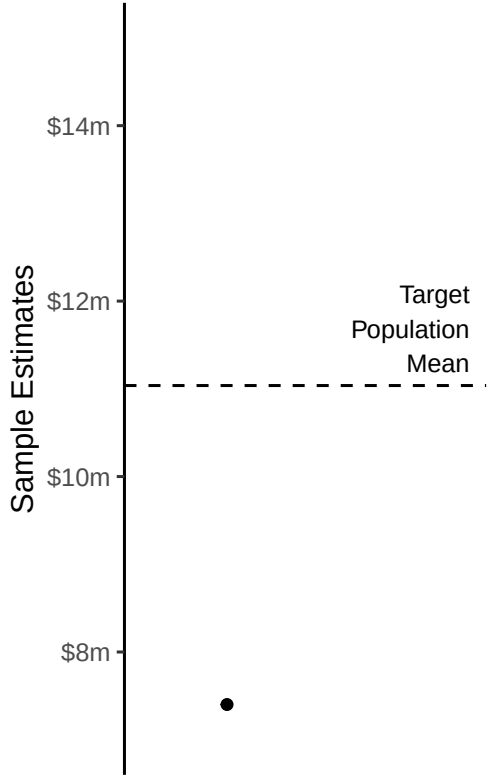


Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

Focus on the target population

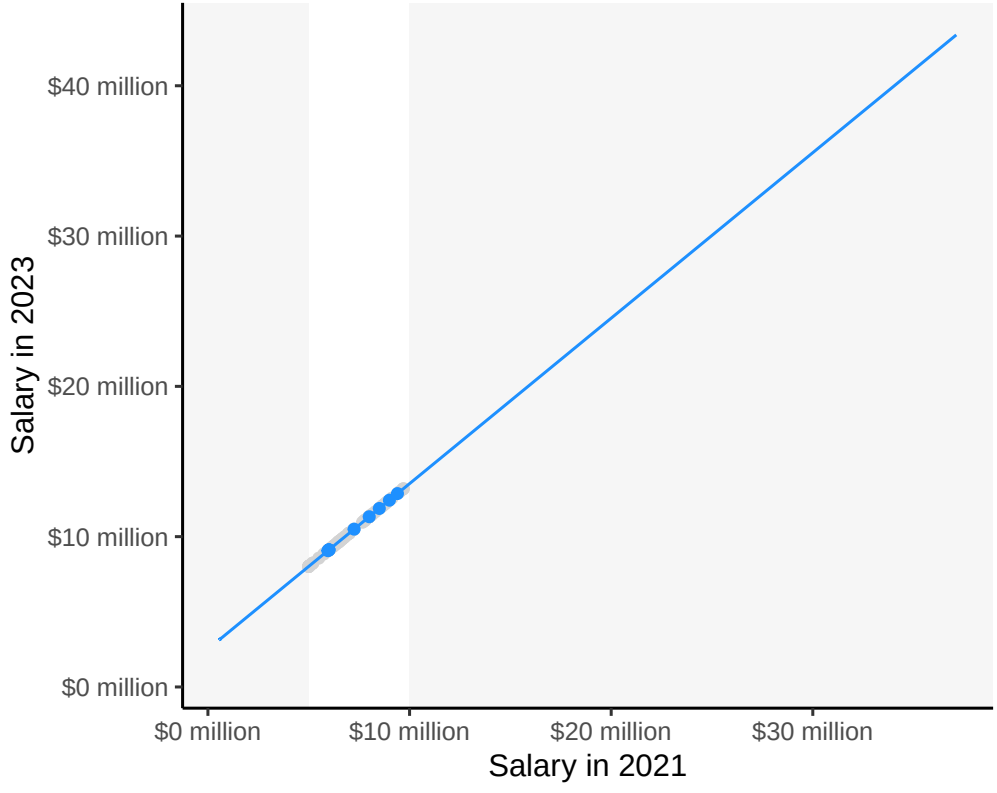


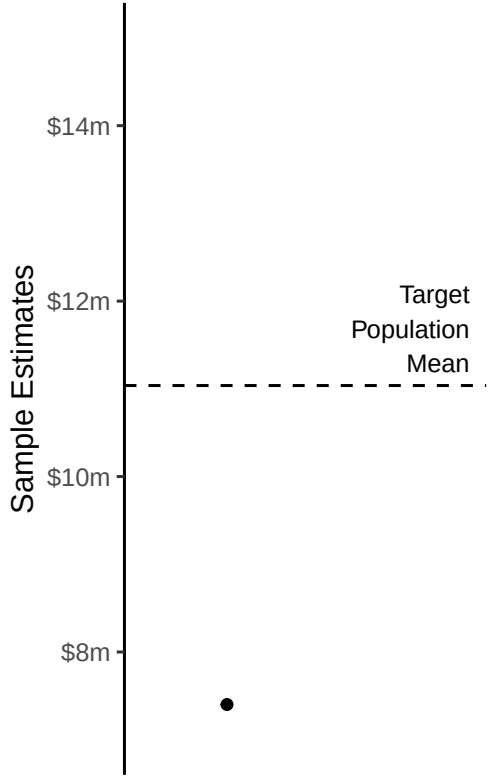


Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

Predict

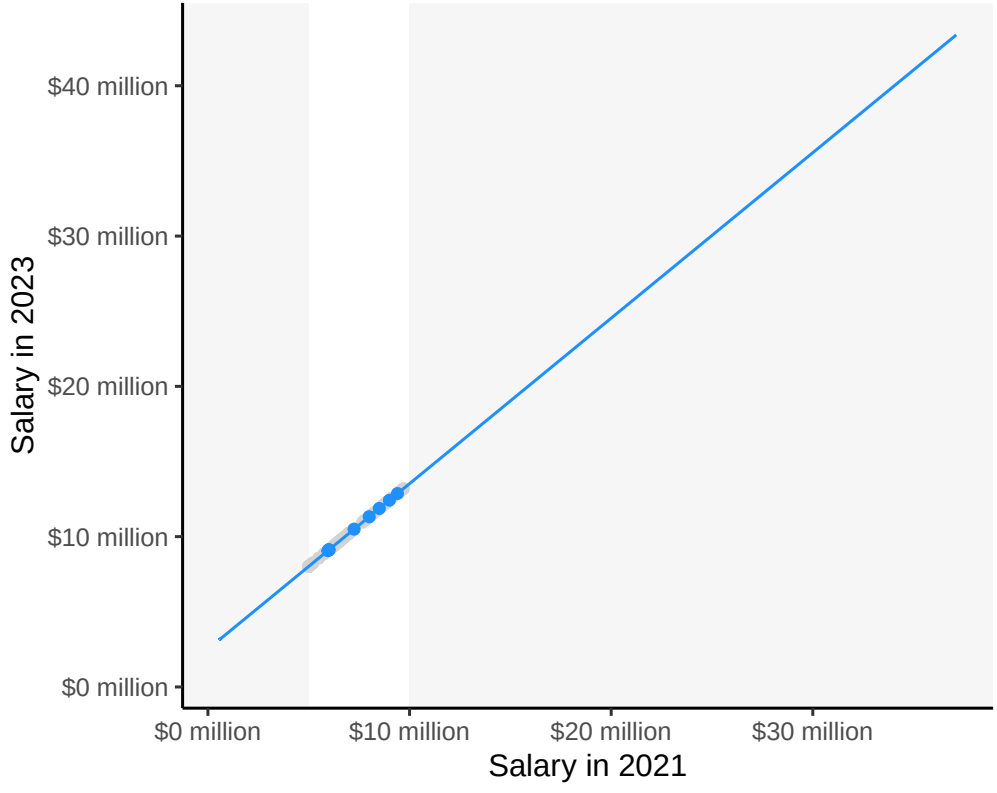




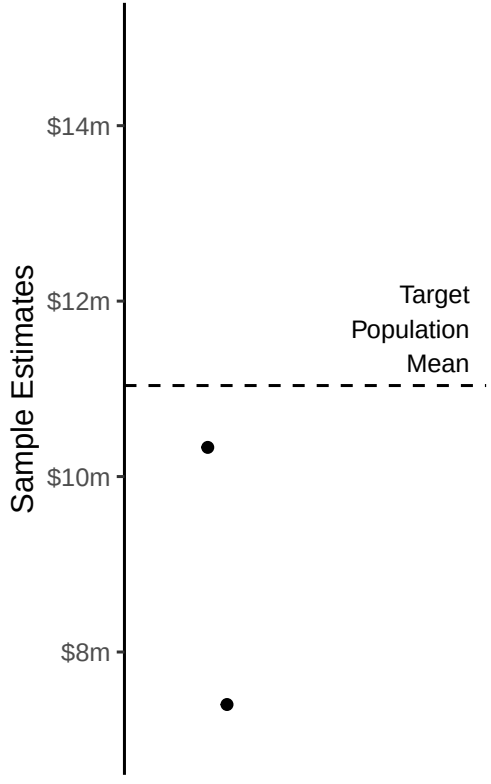
Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

Predict



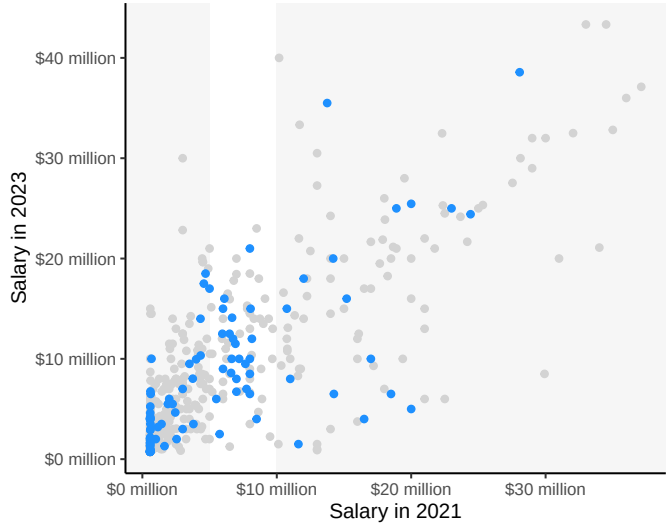
Record the average



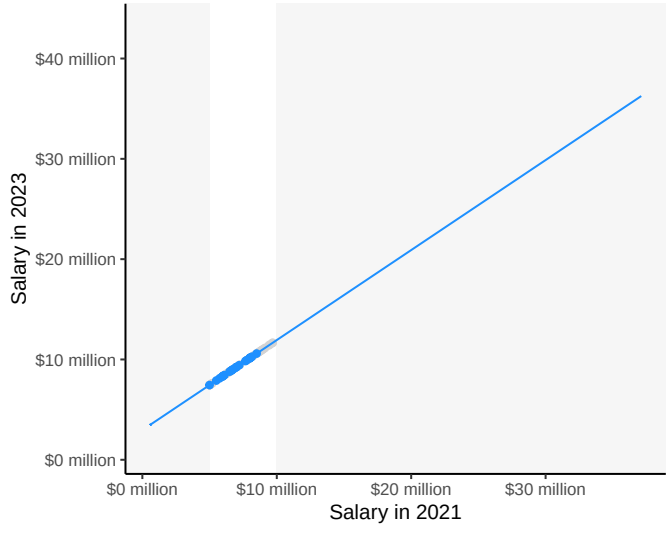
Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

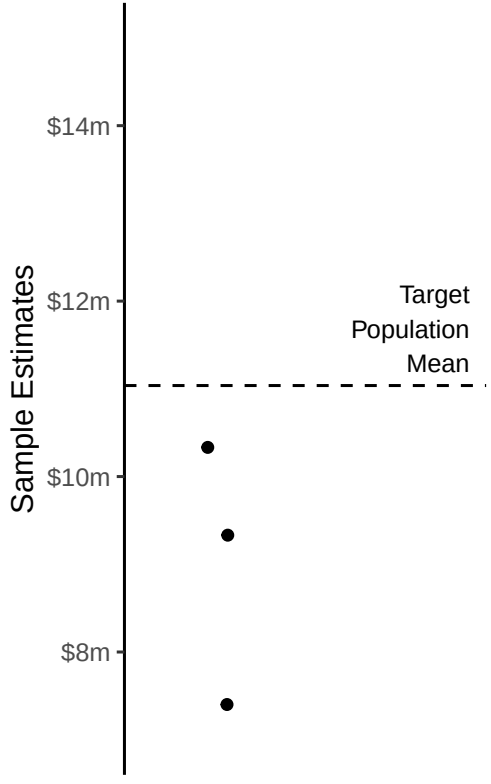
Sample



Learn



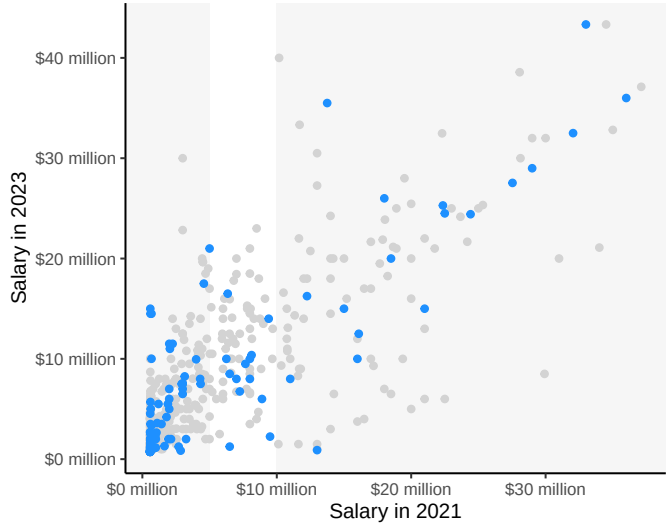
Record



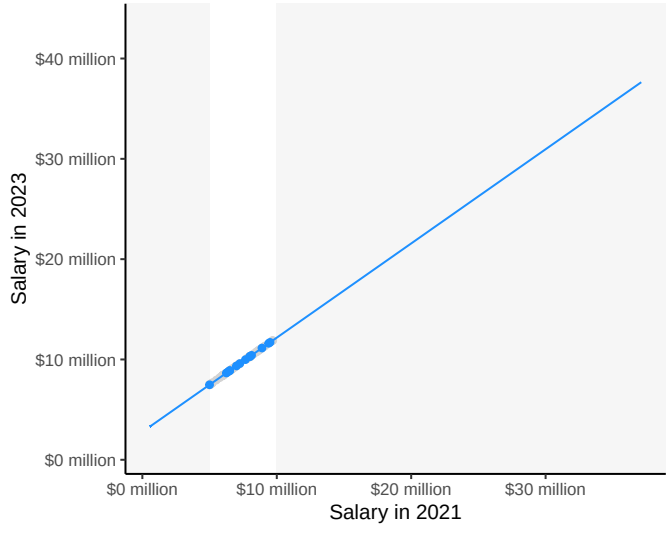
Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

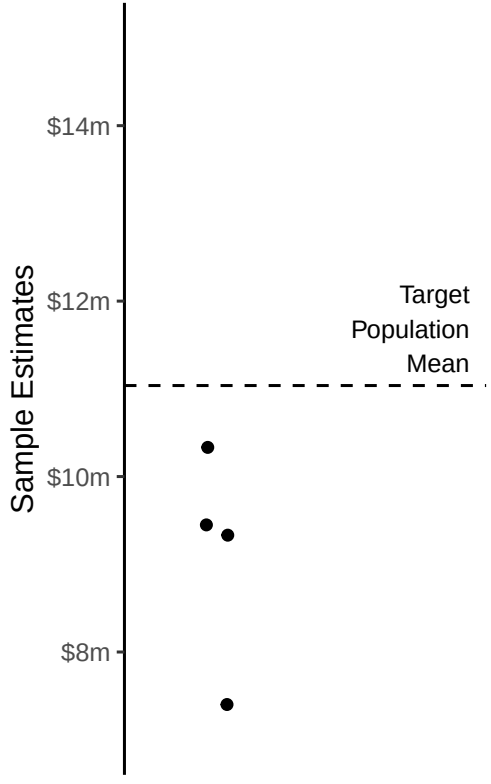
Sample



Learn



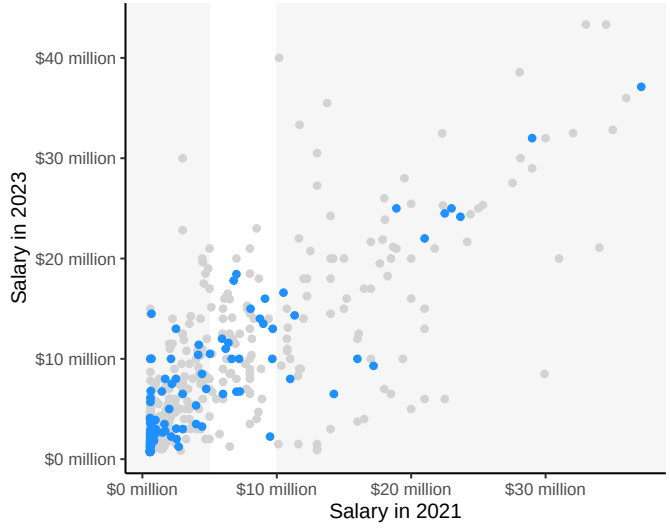
Record



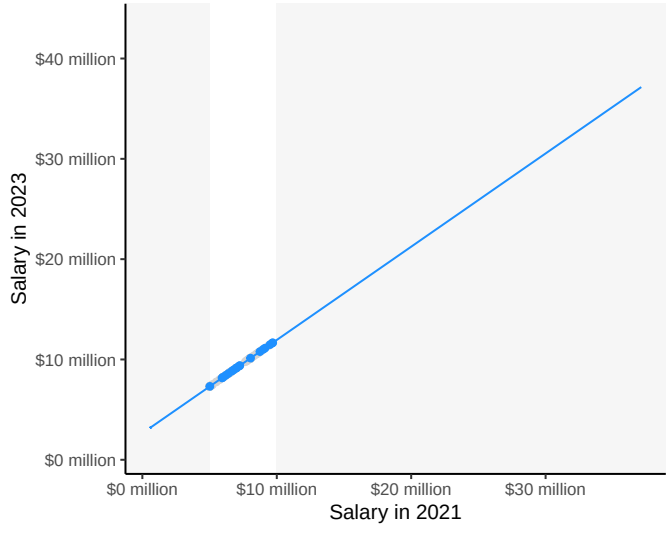
Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

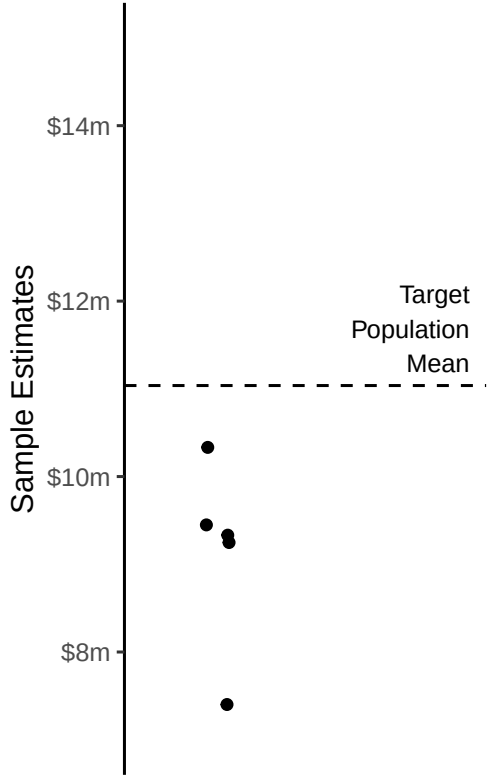
Sample



Learn



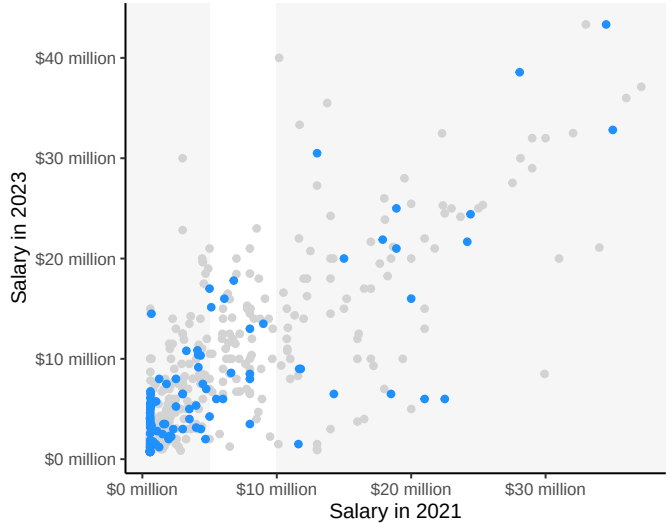
Record



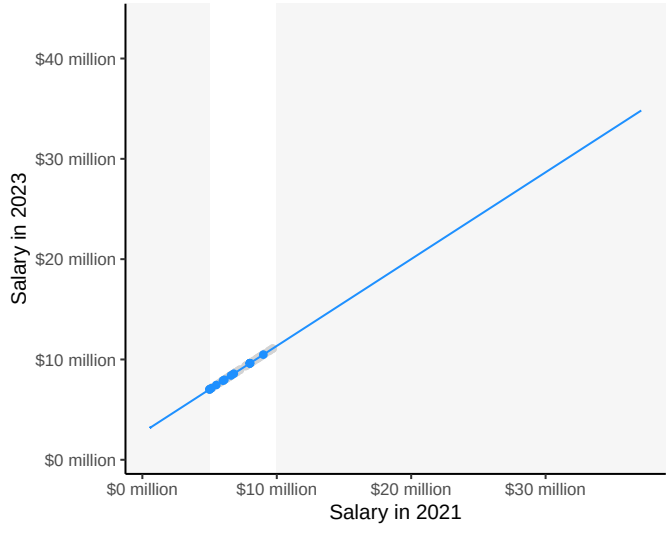
Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

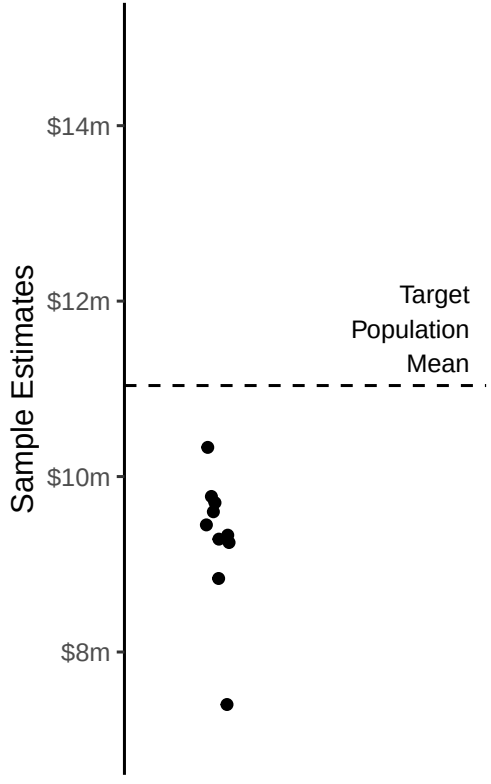
Sample



Learn



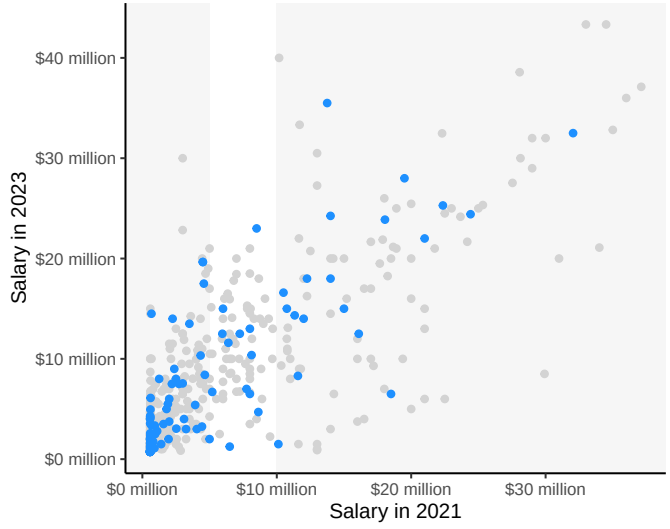
Record



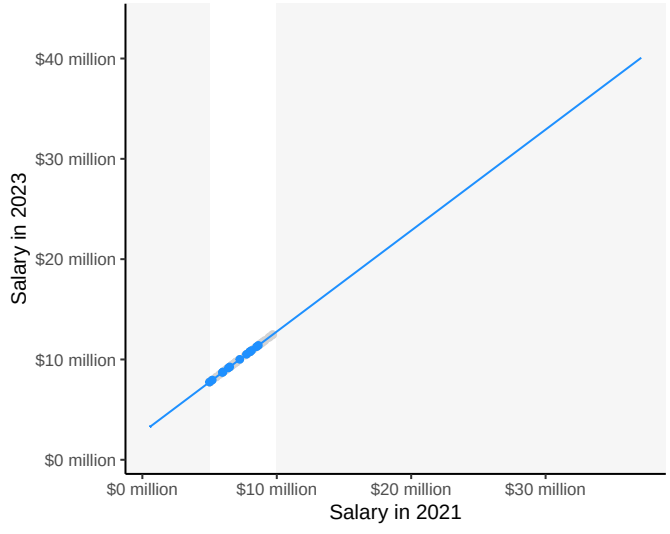
Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

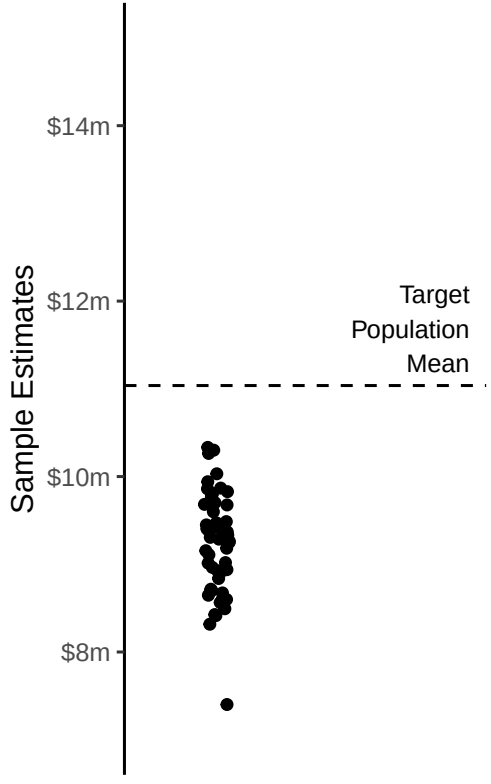
Sample



Learn



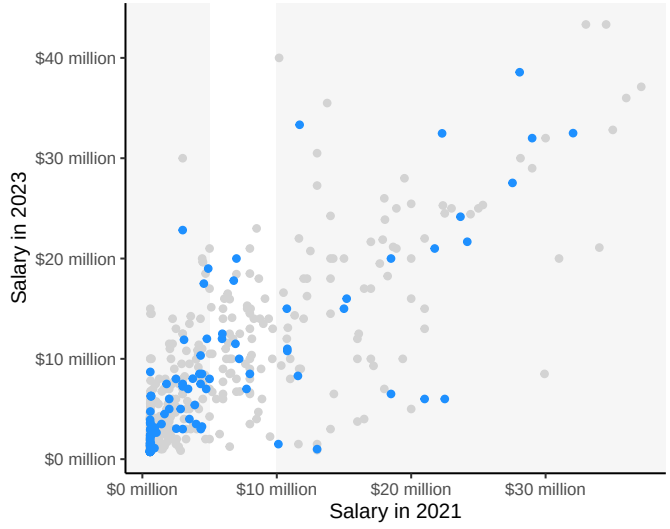
Record



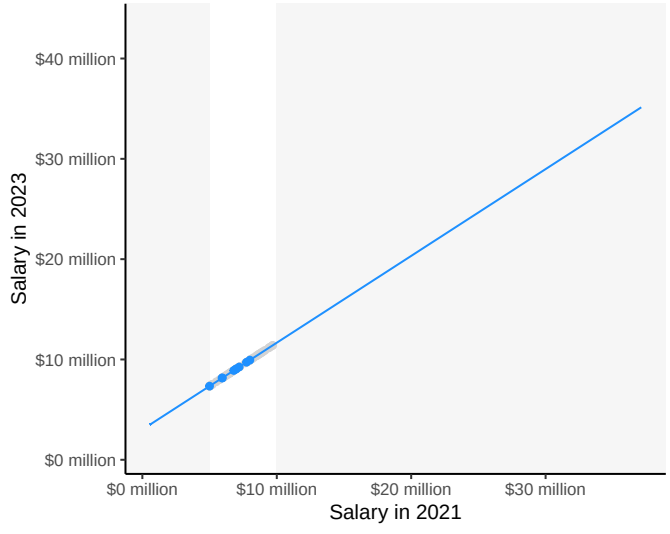
Goal: Estimate a target population mean from a sample

Method: Ordinary Least Squares prediction

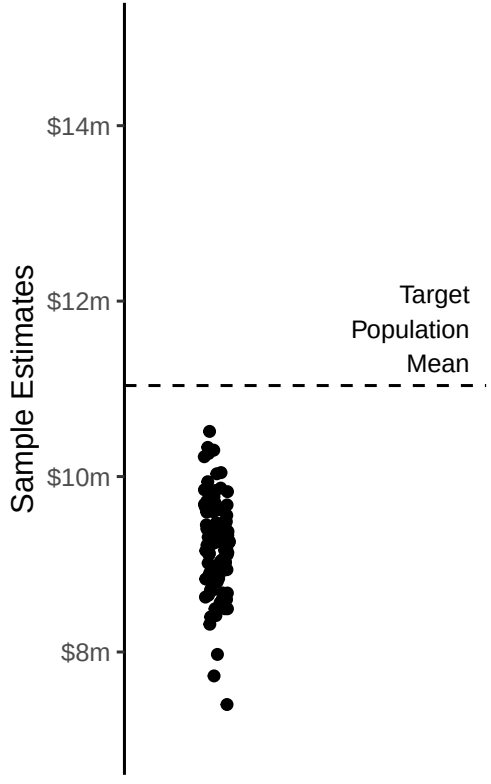
Sample



Learn



Record

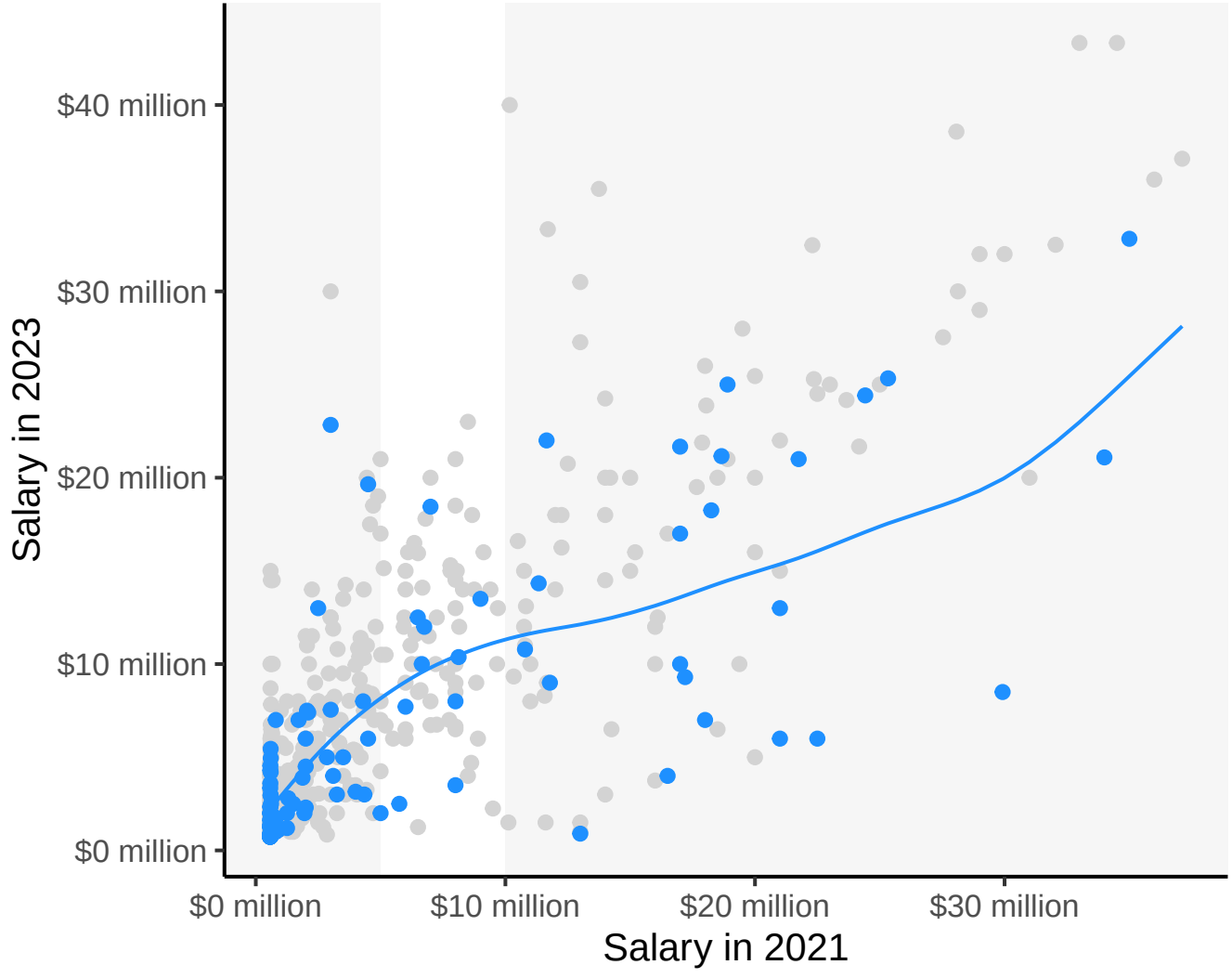


Ordinary Least Squares strategy:

1. Sample from the population
2. Learn a model
3. Record the average prediction in the target subgroup

Goal: Estimate a target population mean from a sample

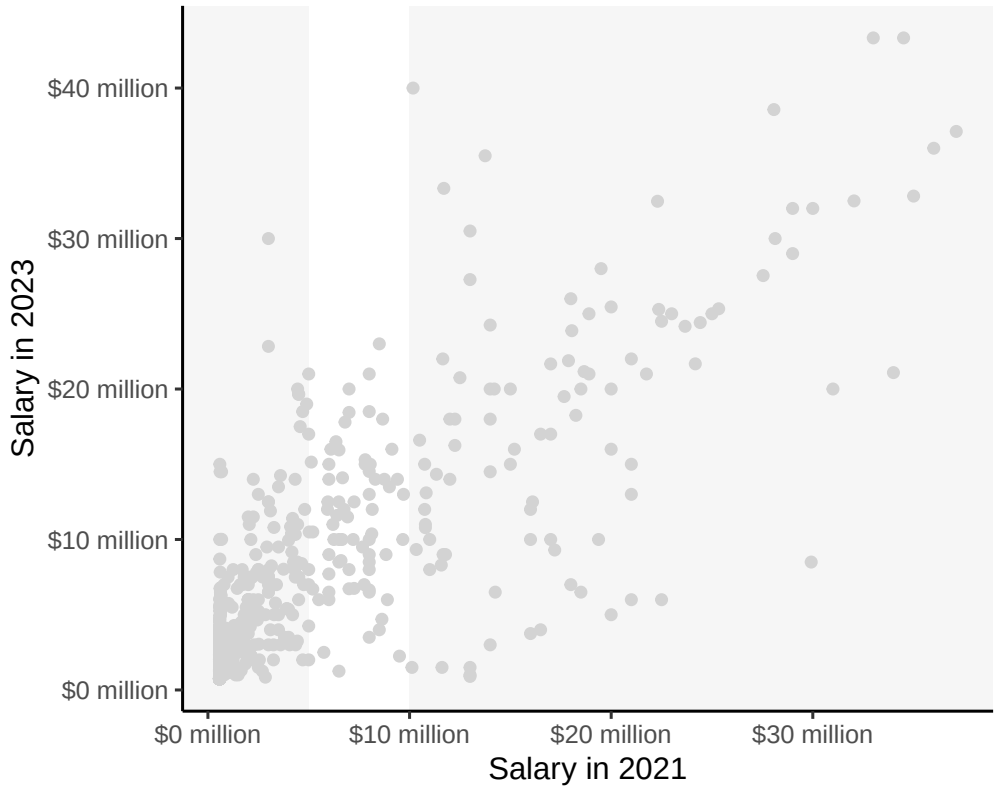
How would you do this with machine learning?



Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

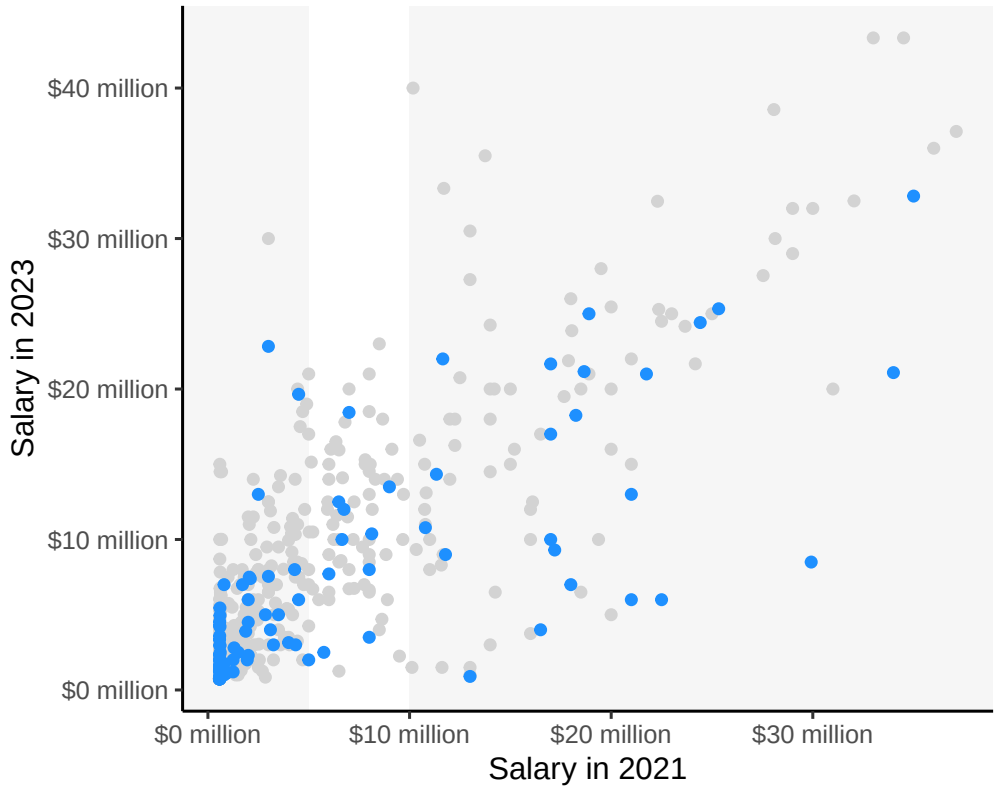
Begin with the population



Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

Draw a sample



Goal: Estimate a target population mean from a sample

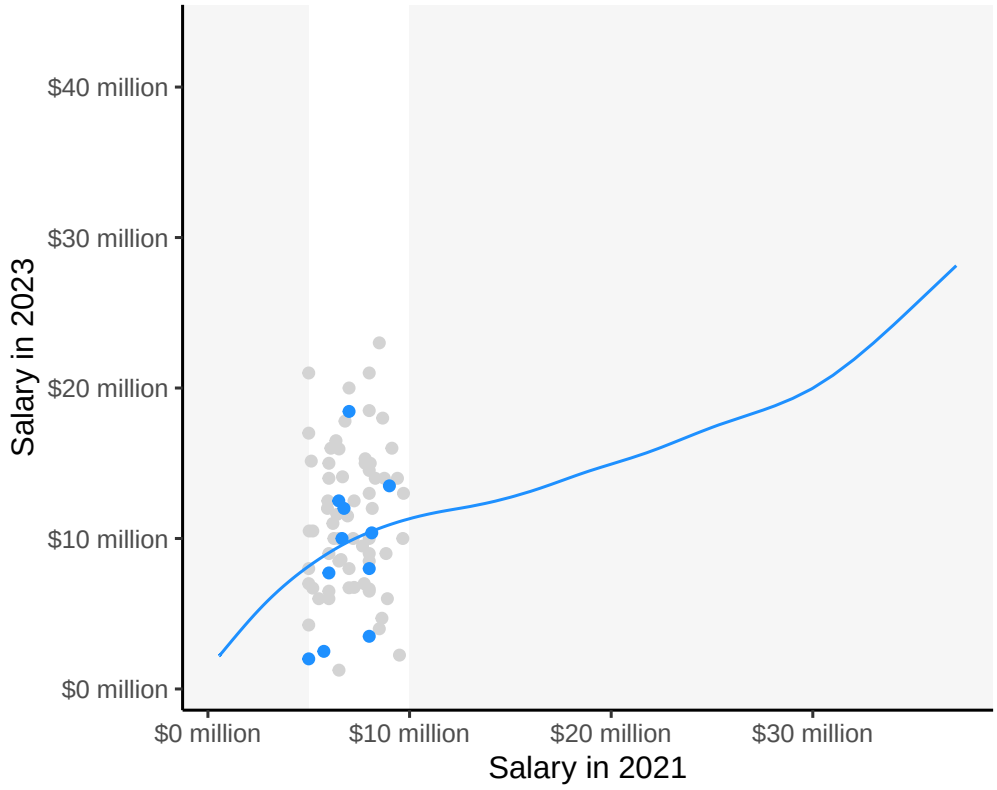
Method: Generalized Additive Model prediction

Learn a model

Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

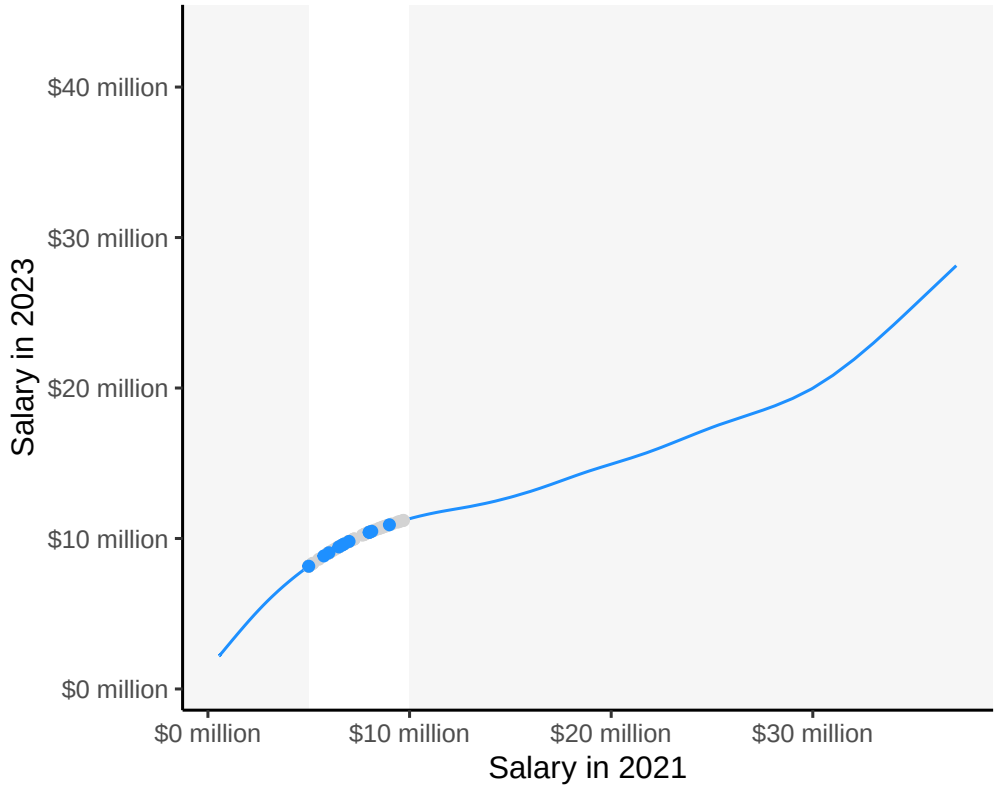
Focus on the target population



Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

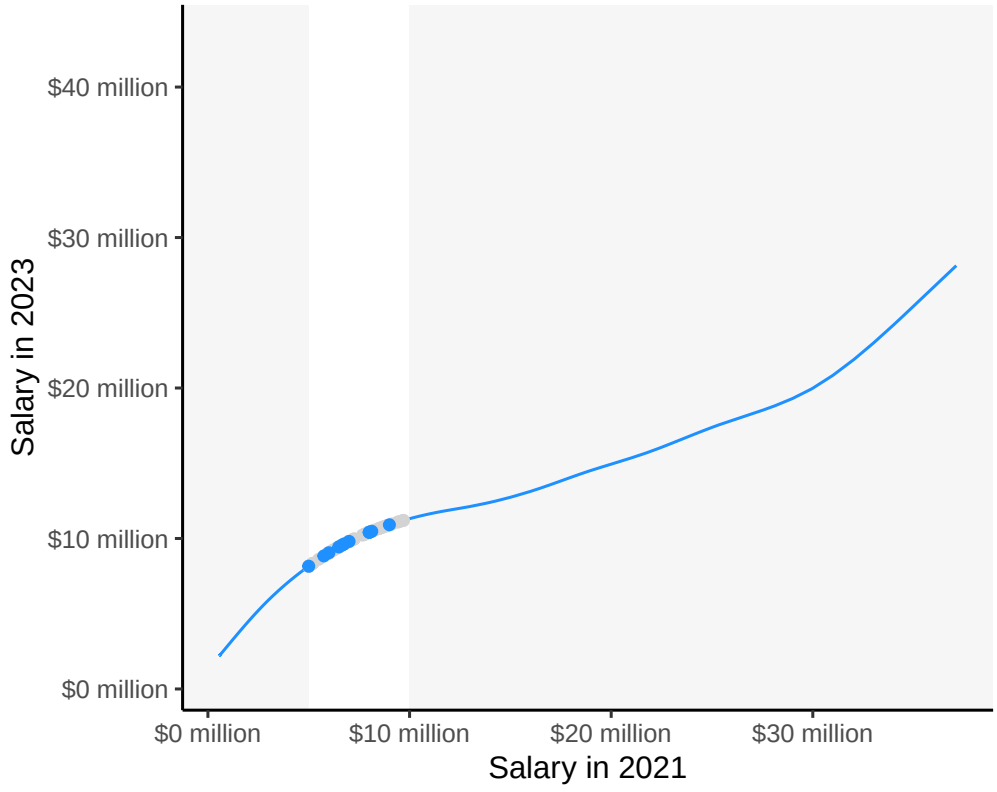
Predict



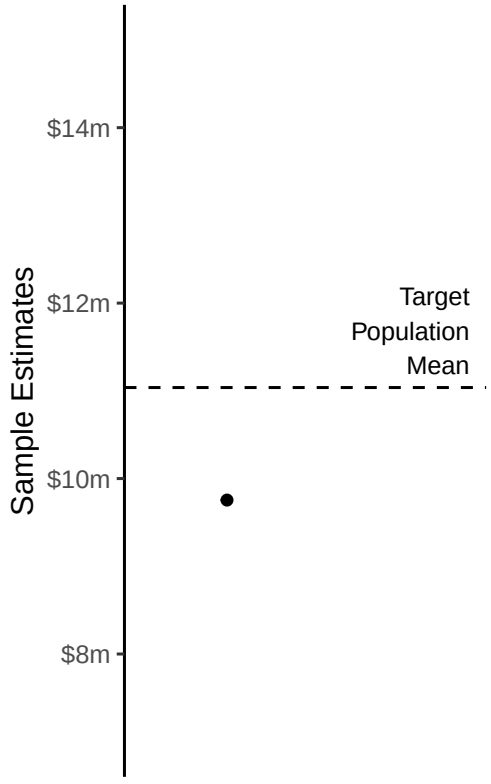
Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

Predict



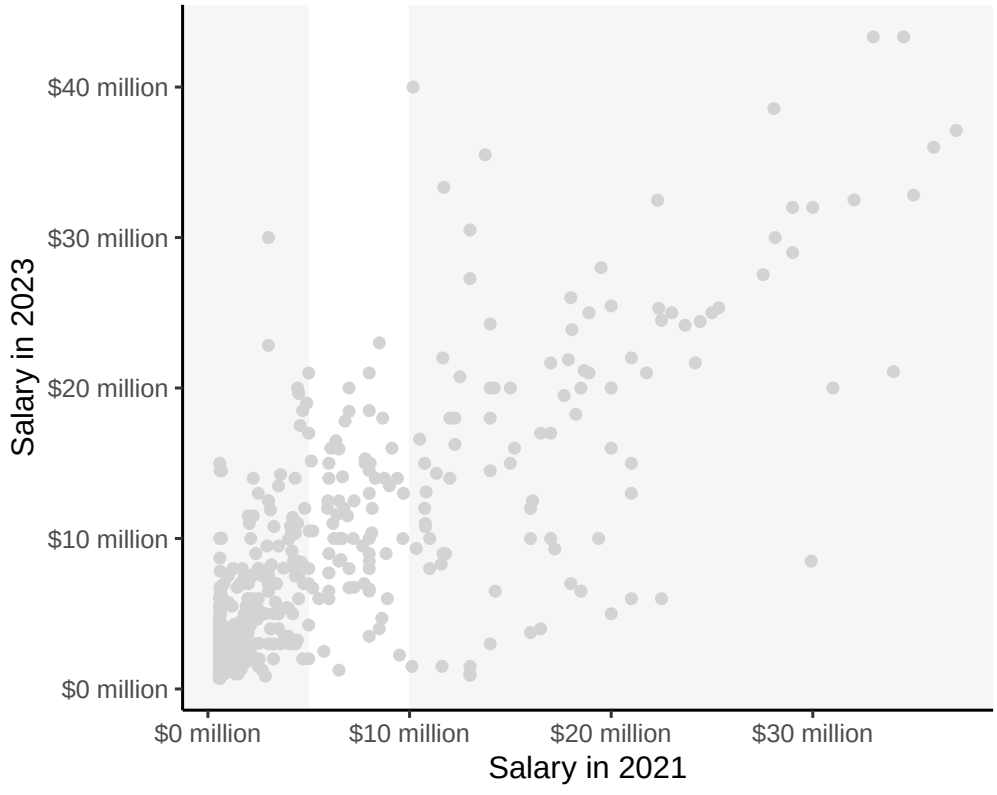
Record the average

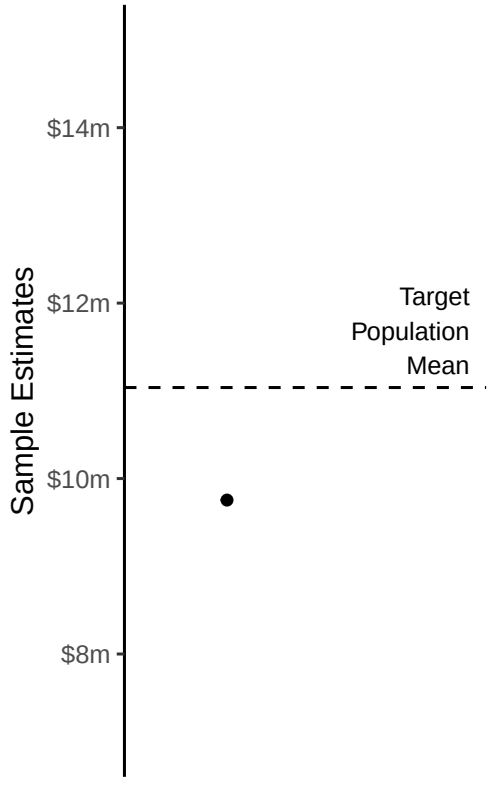


Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

Begin with the population

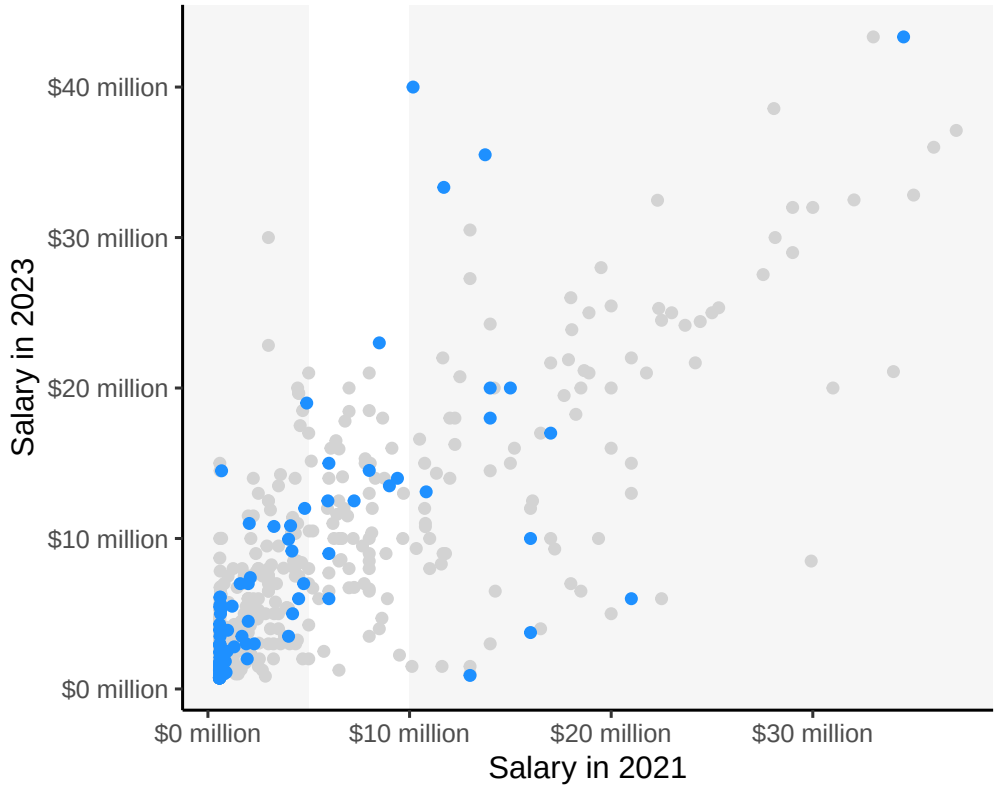


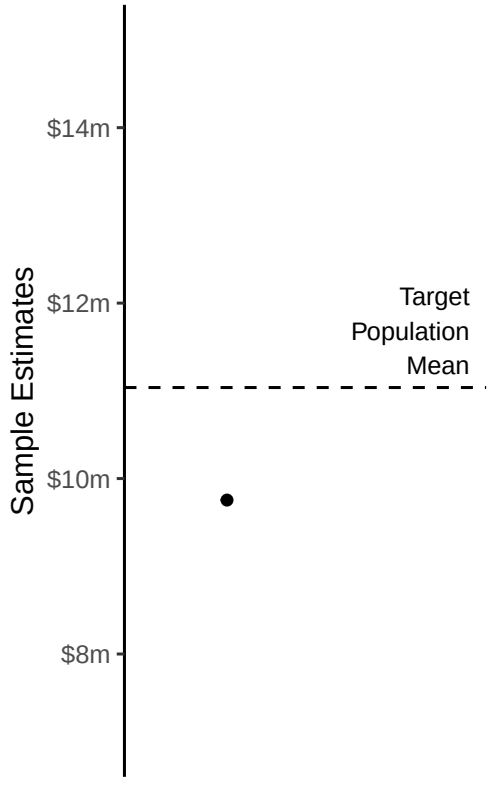


Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

Draw a sample

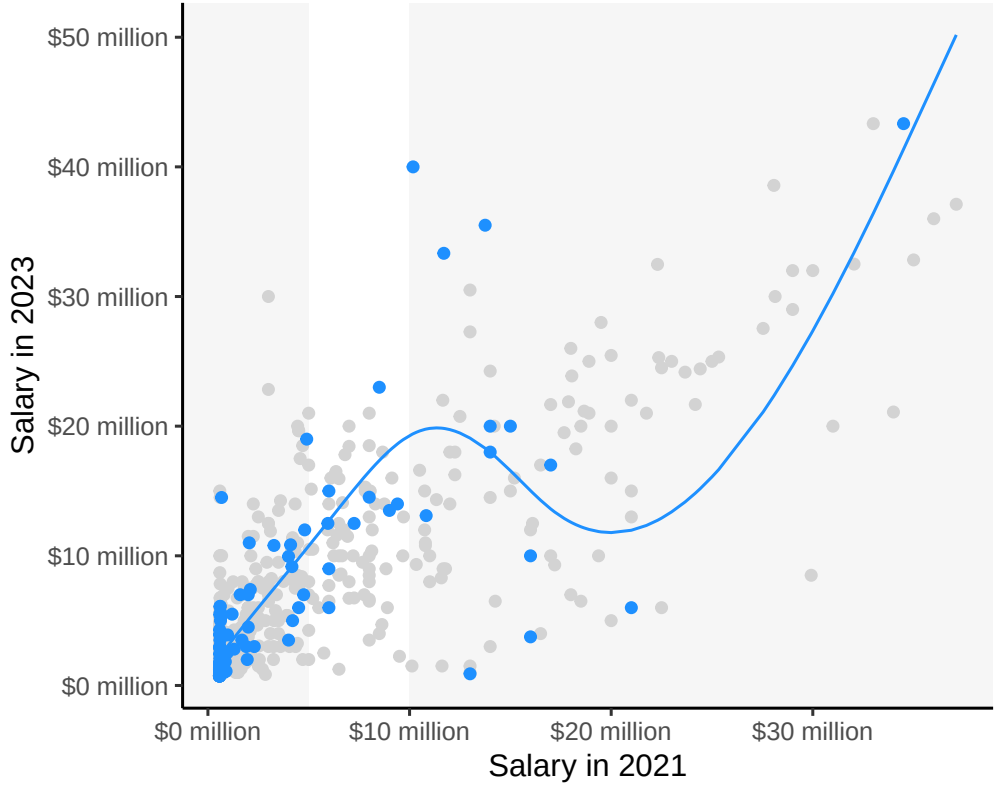


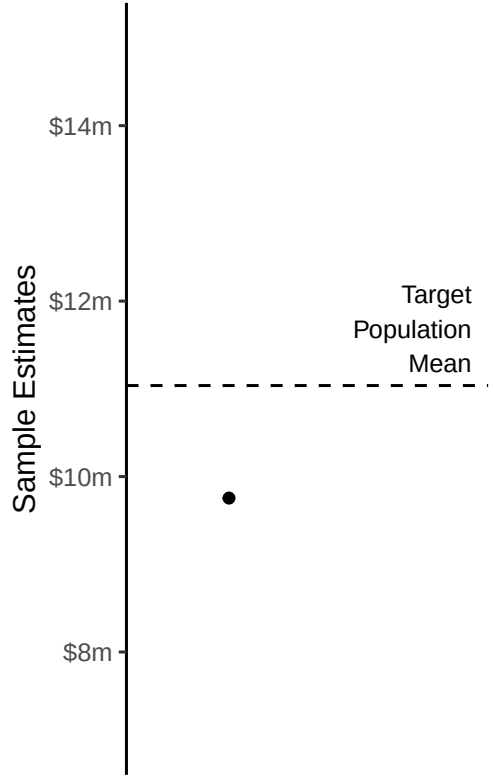


Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

Learn a model

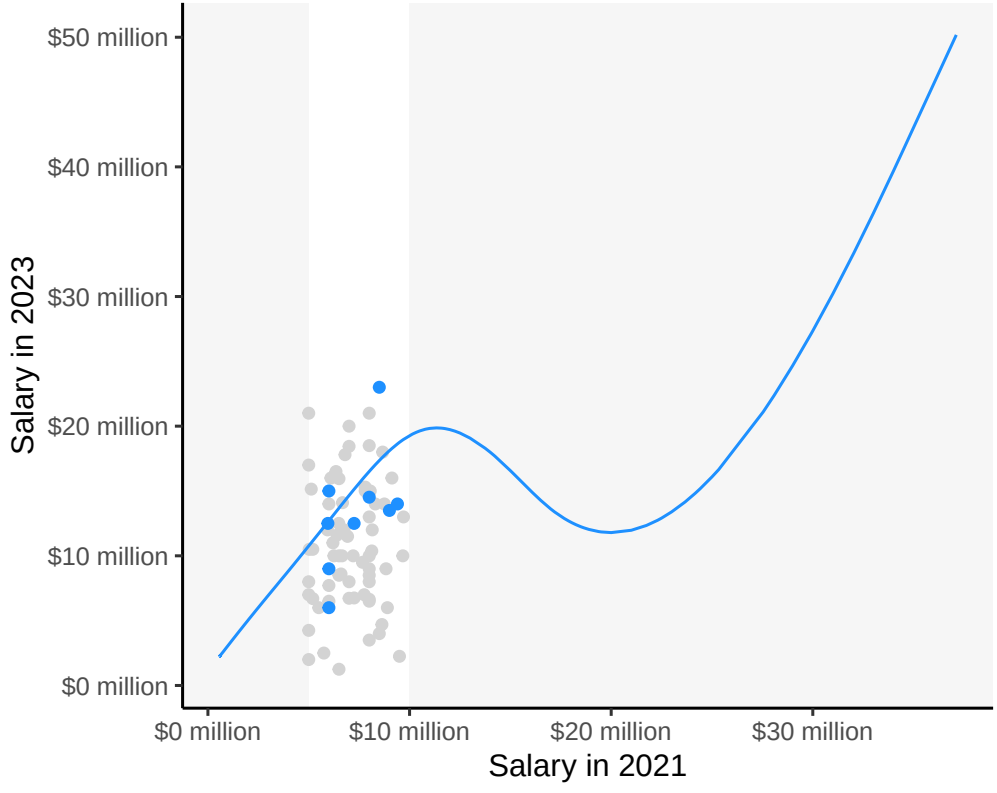


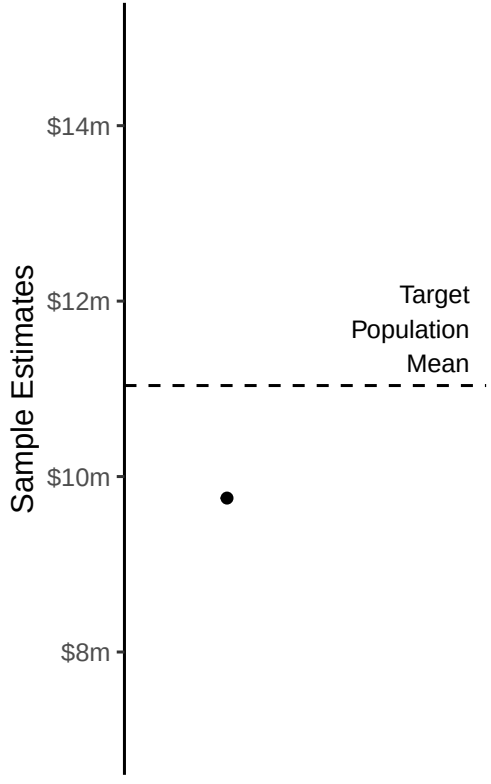


Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

Focus on the target population

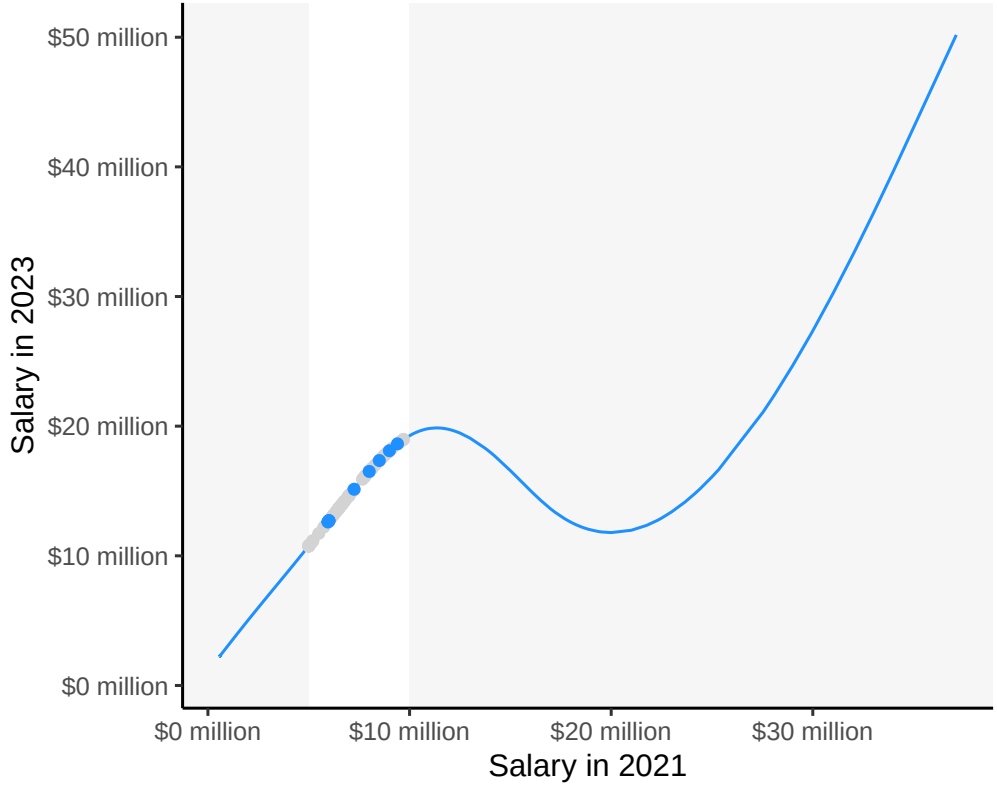


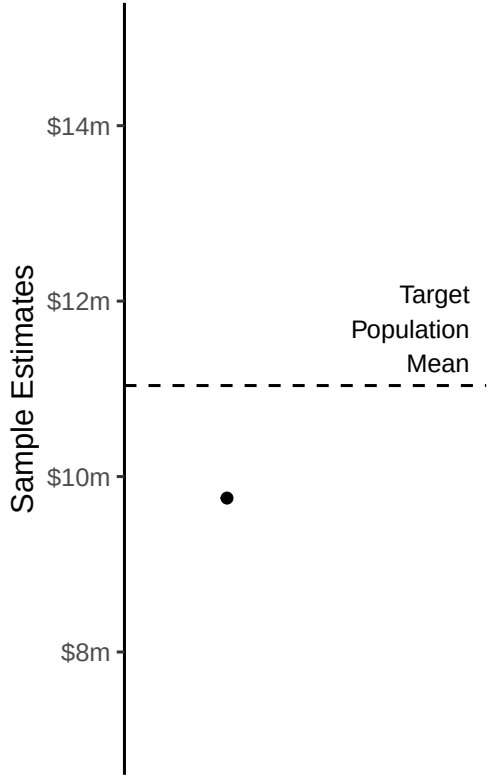


Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

Predict

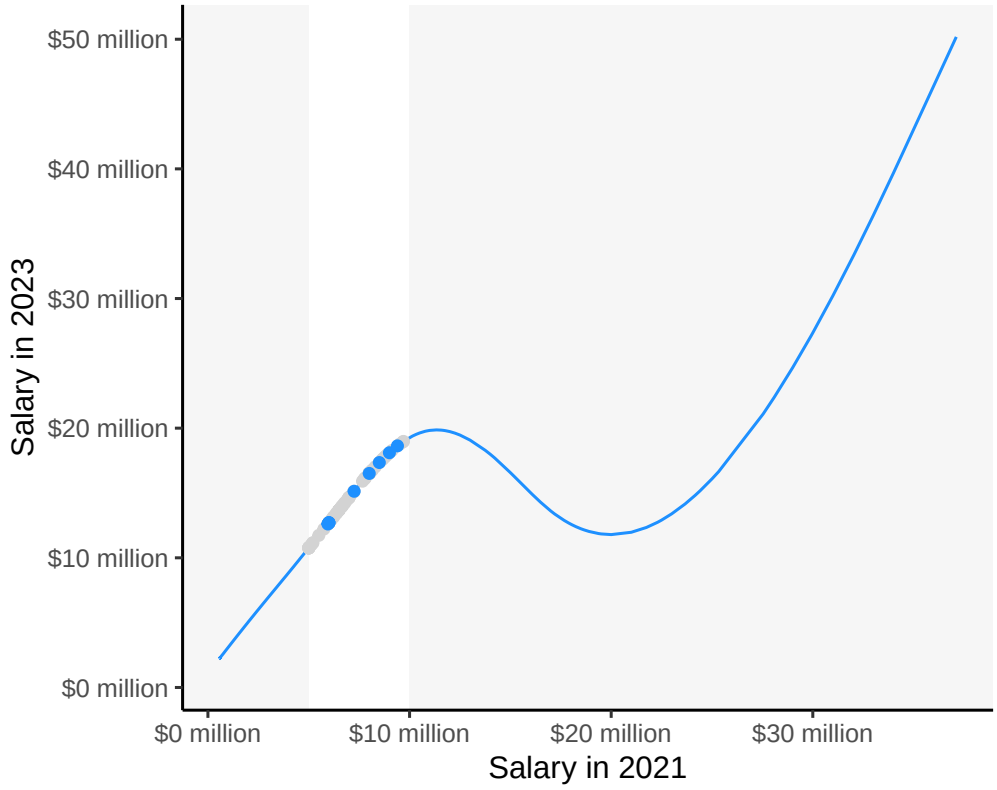




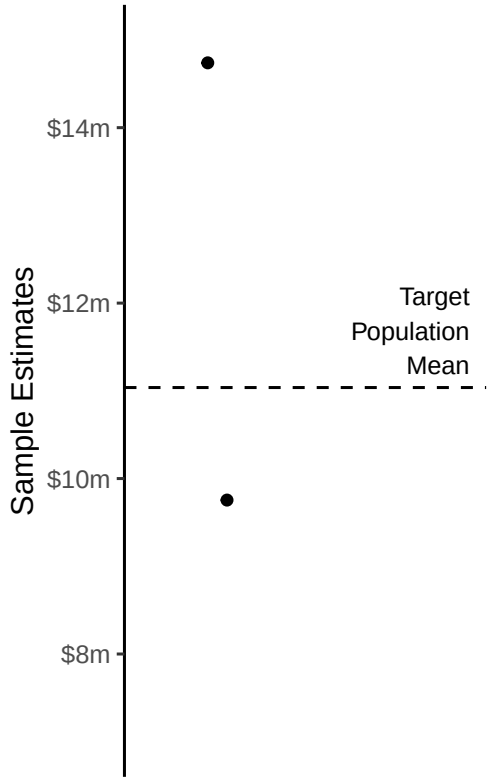
Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

Predict



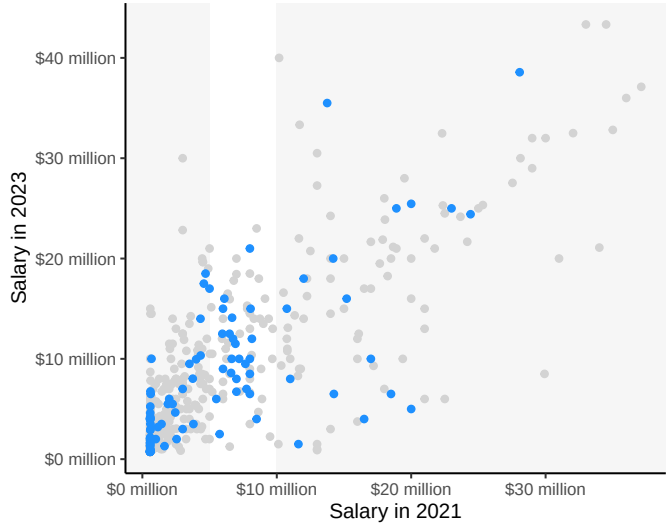
Record the average



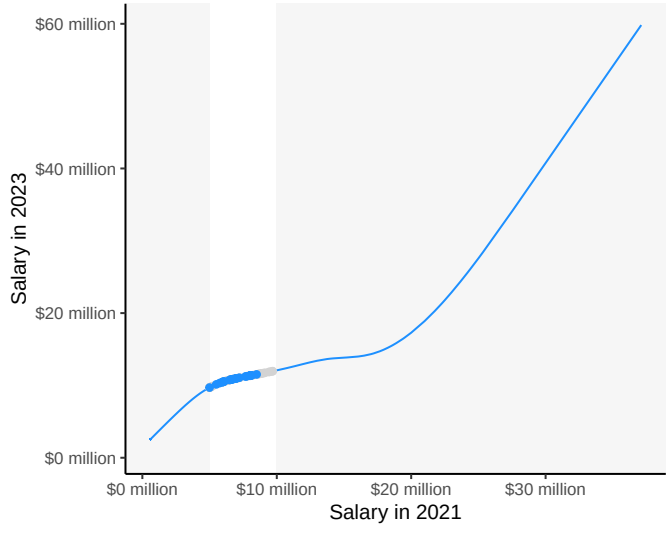
Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

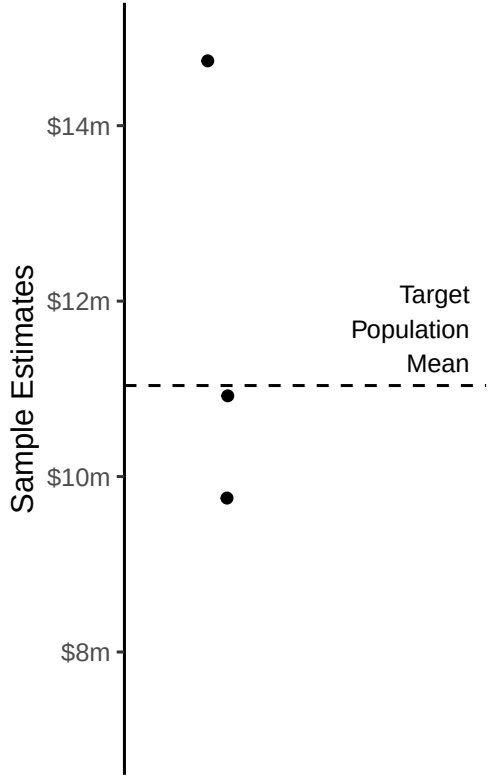
Sample



Learn



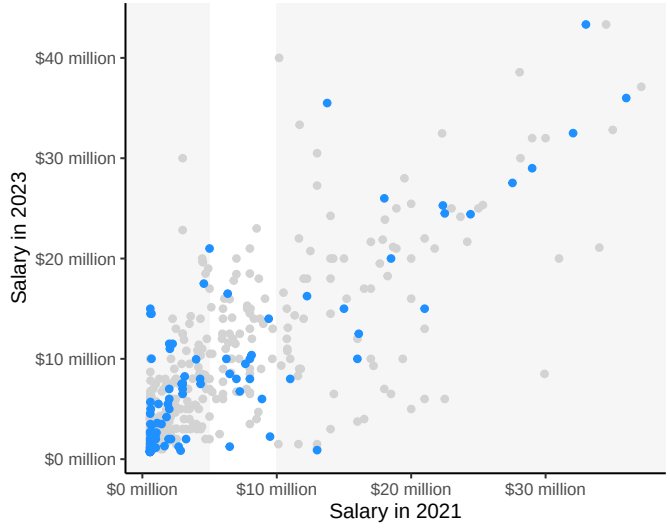
Record



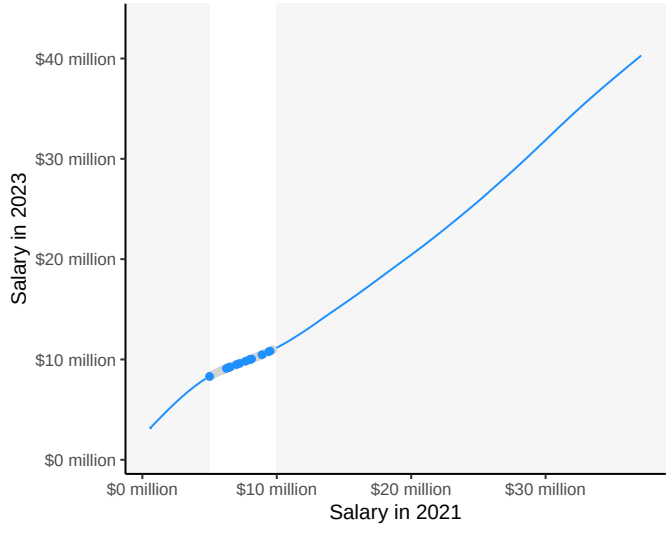
Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

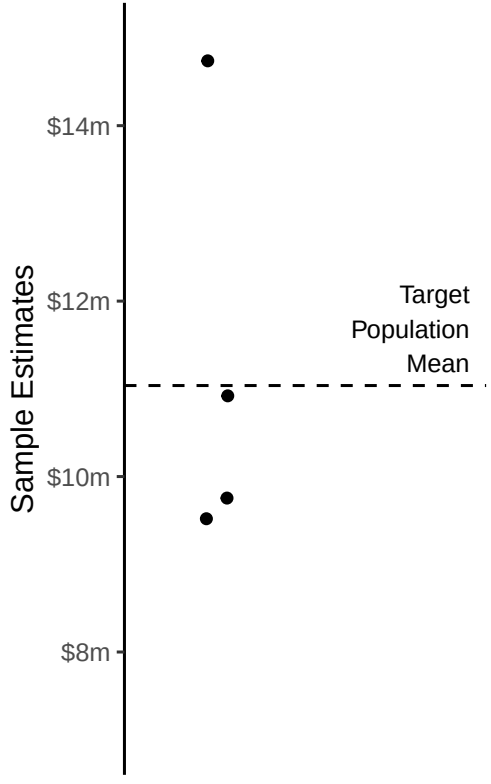
Sample



Learn



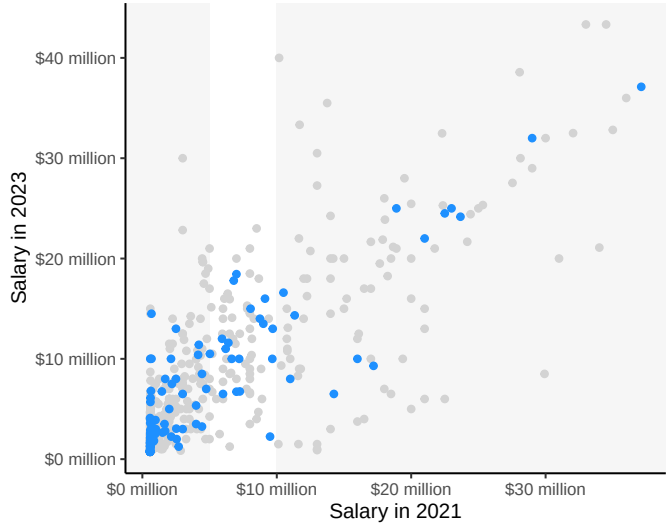
Record



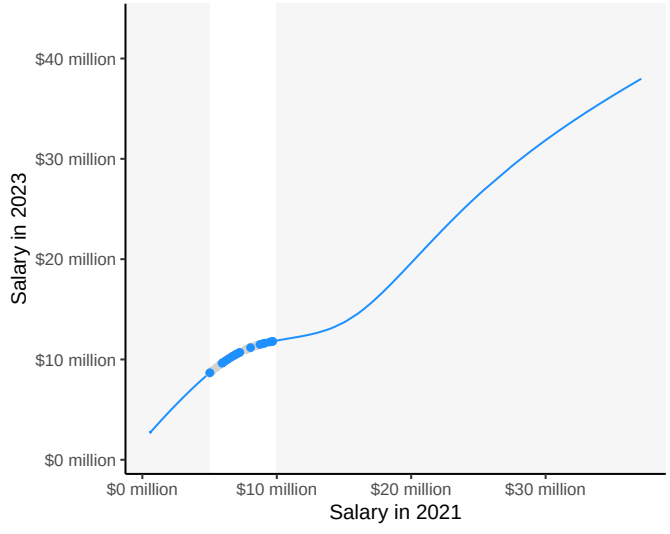
Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

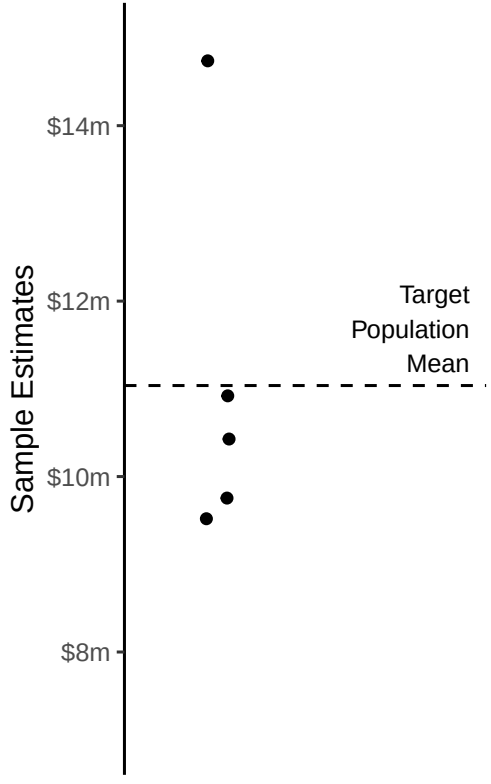
Sample



Learn



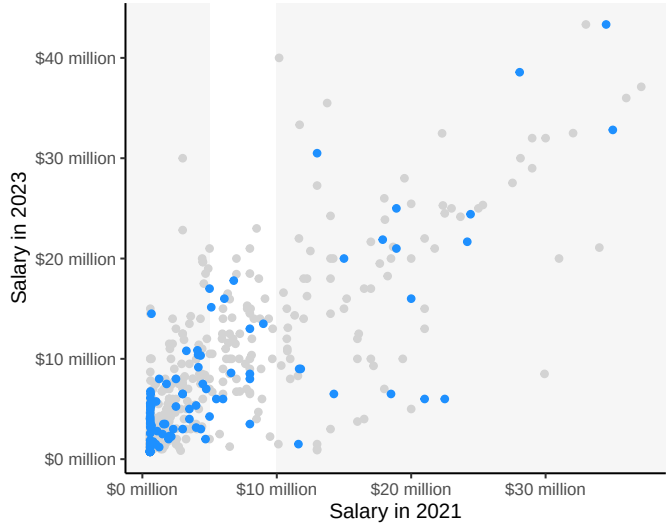
Record



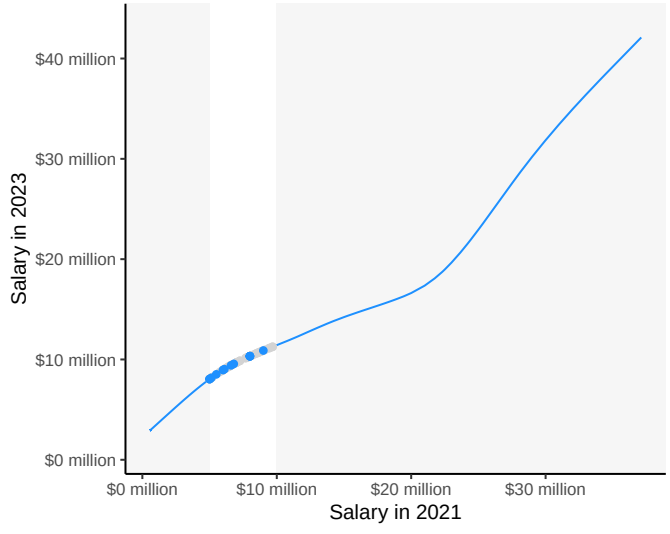
Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

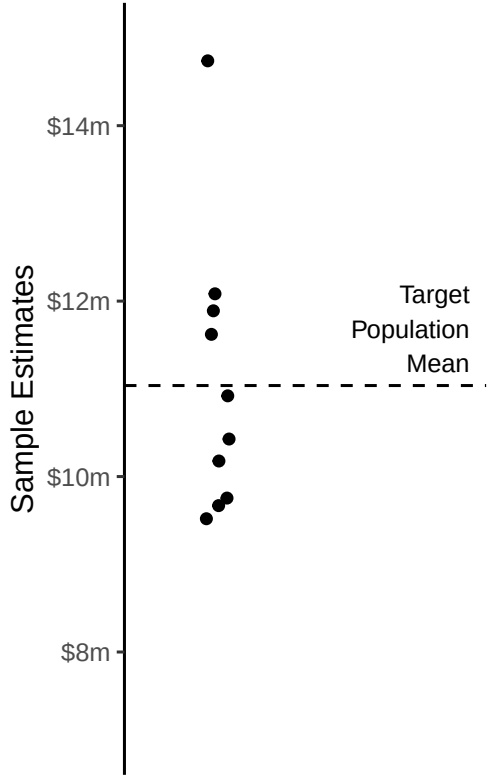
Sample



Learn



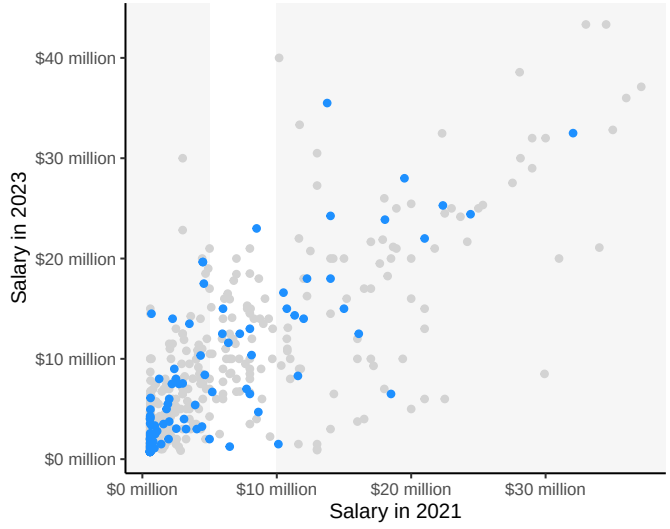
Record



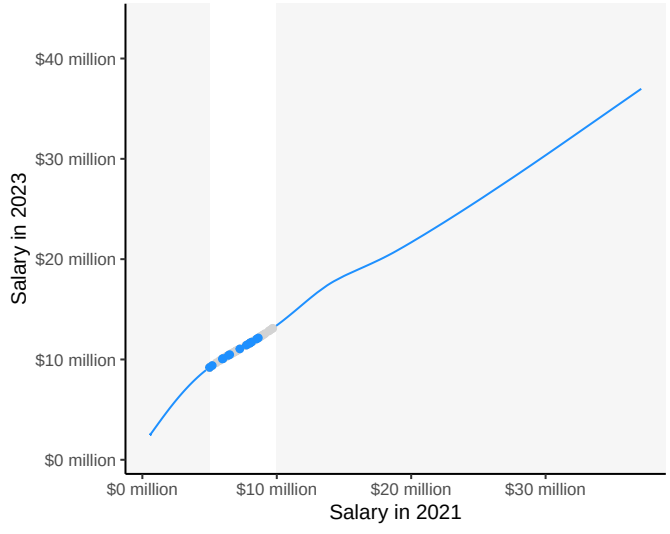
Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

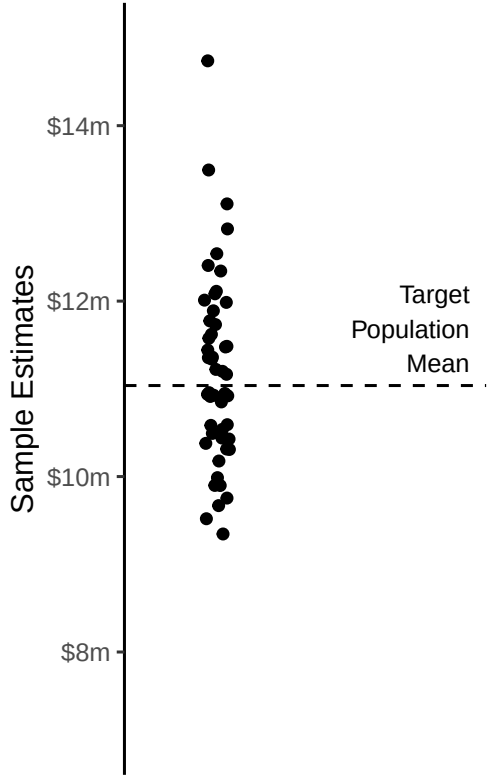
Sample



Learn



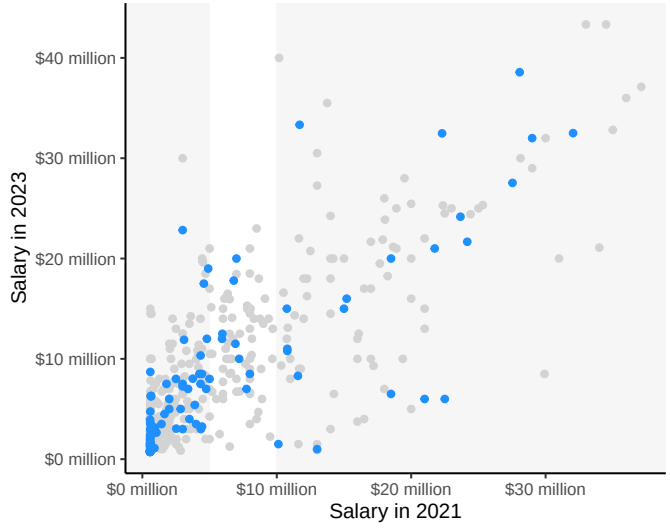
Record



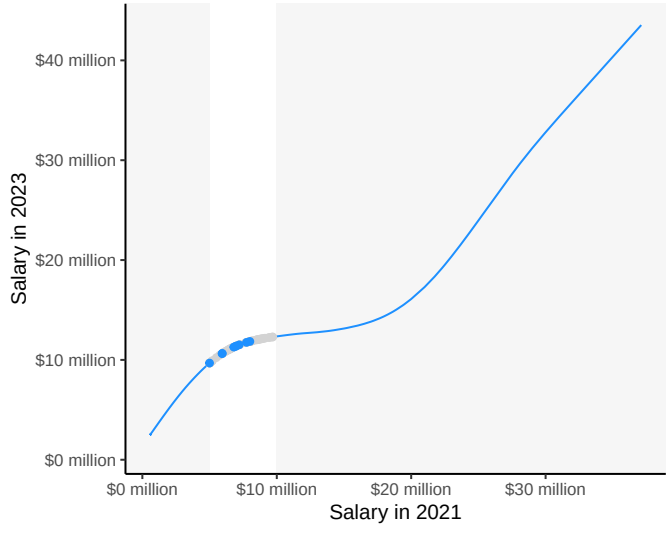
Goal: Estimate a target population mean from a sample

Method: Generalized Additive Model prediction

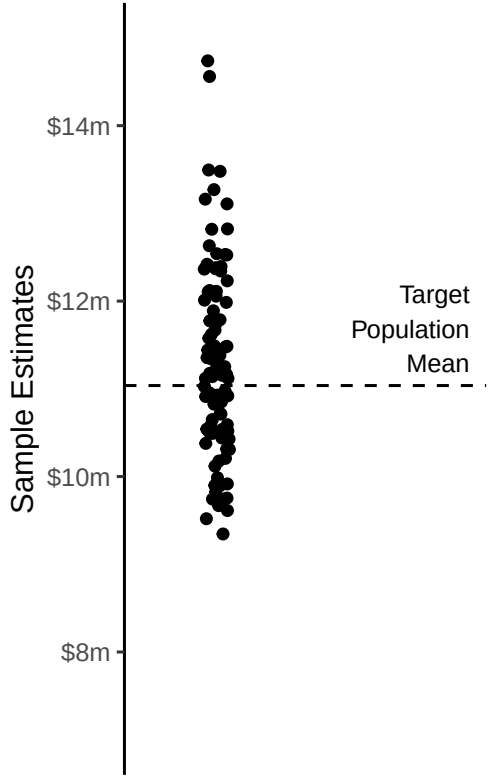
Sample



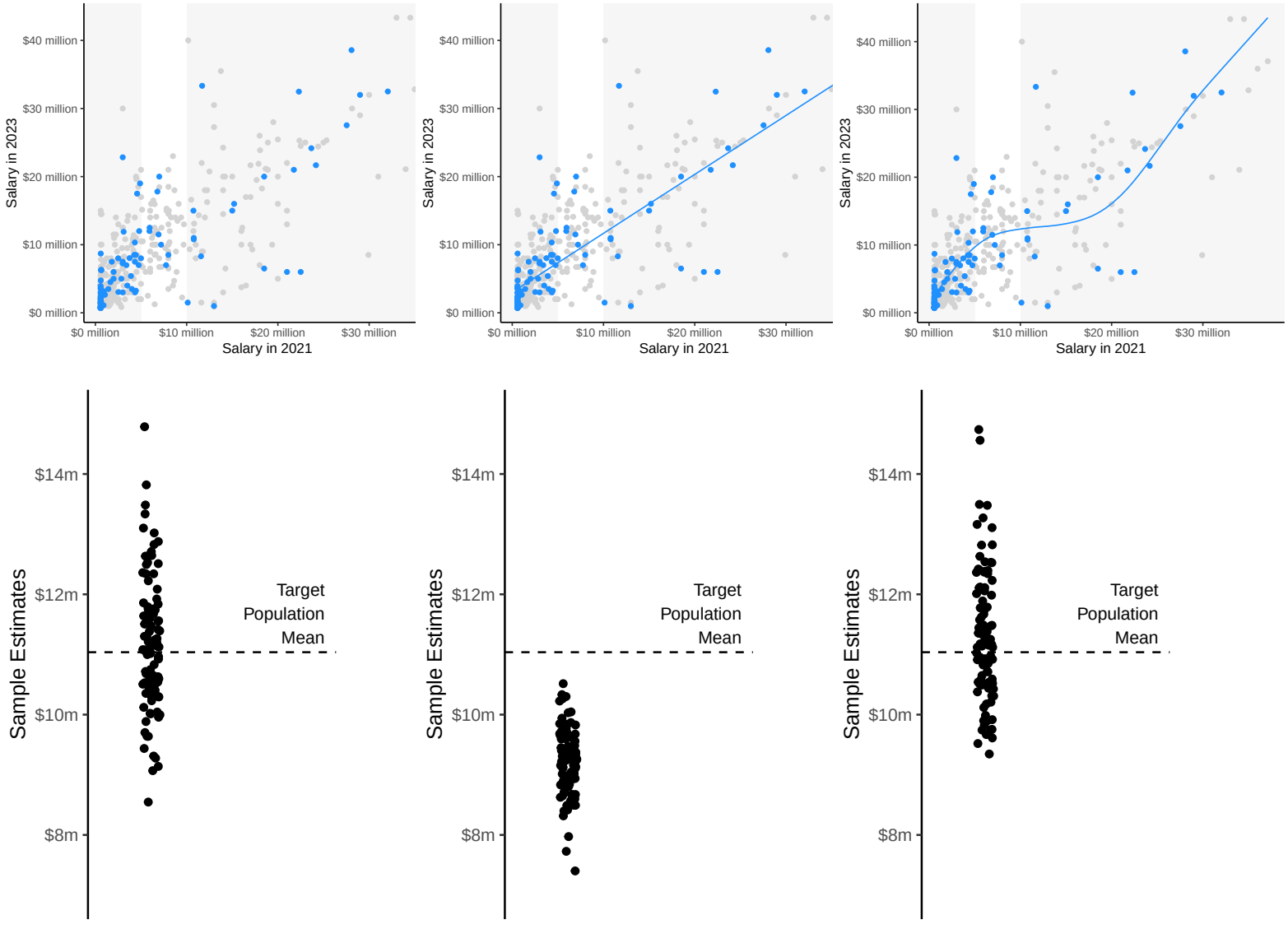
Learn



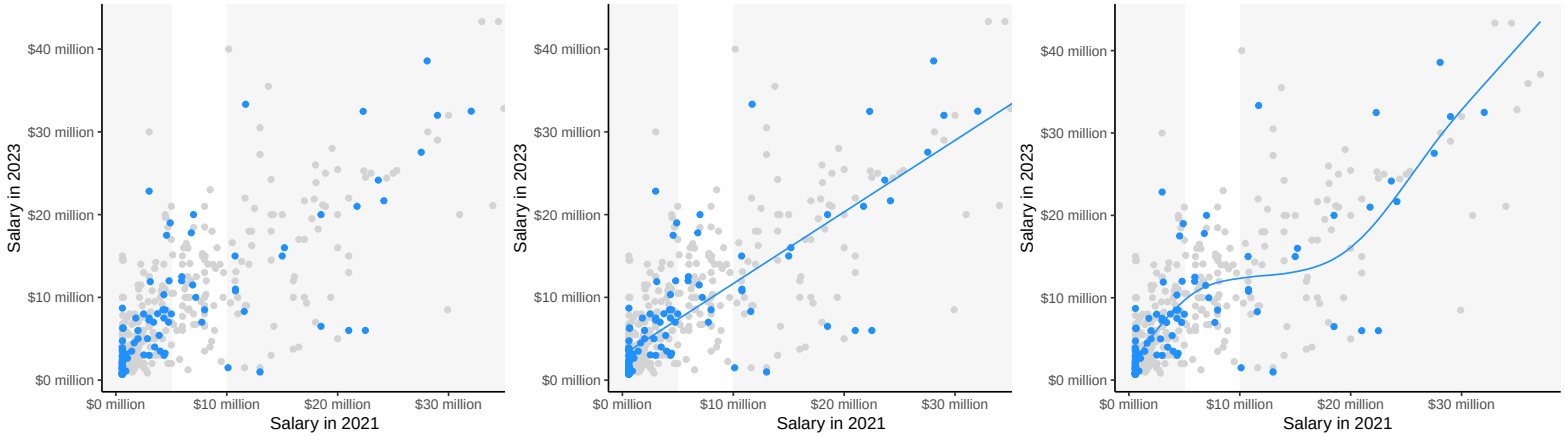
Record

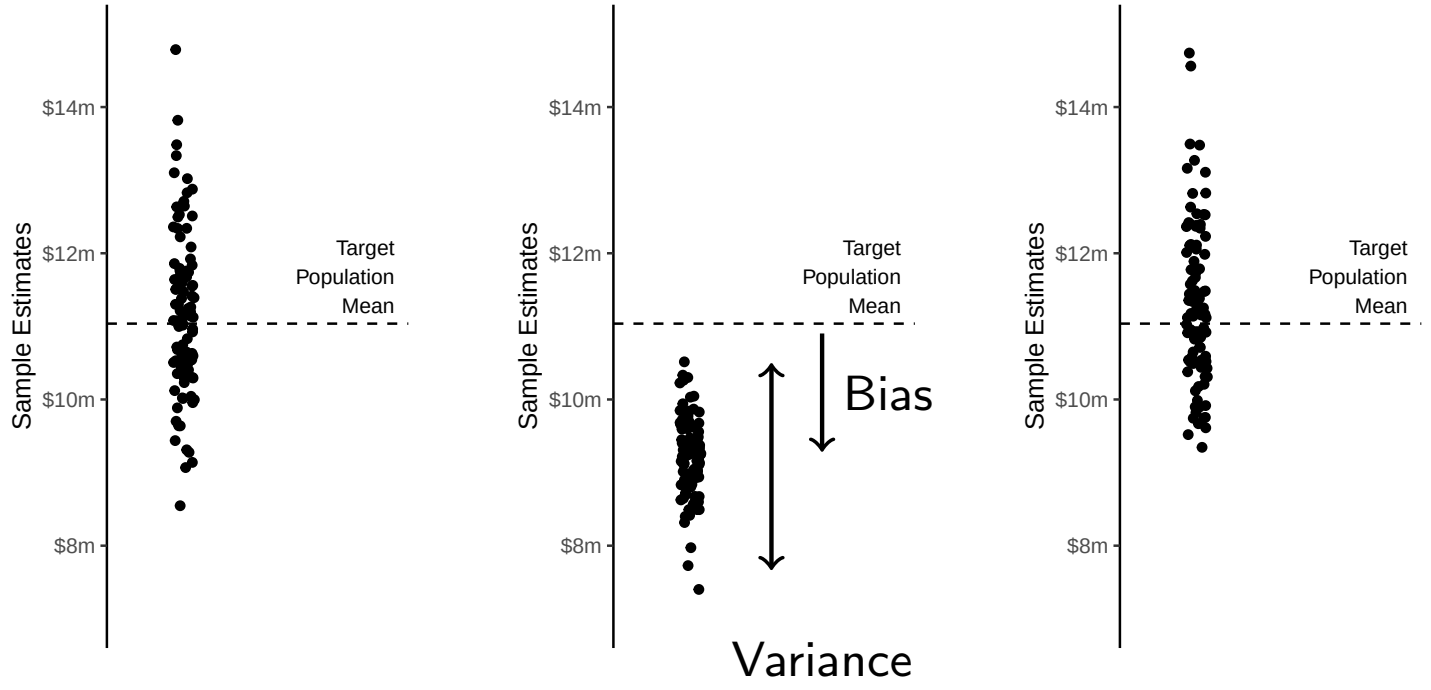


Comparing the estimators

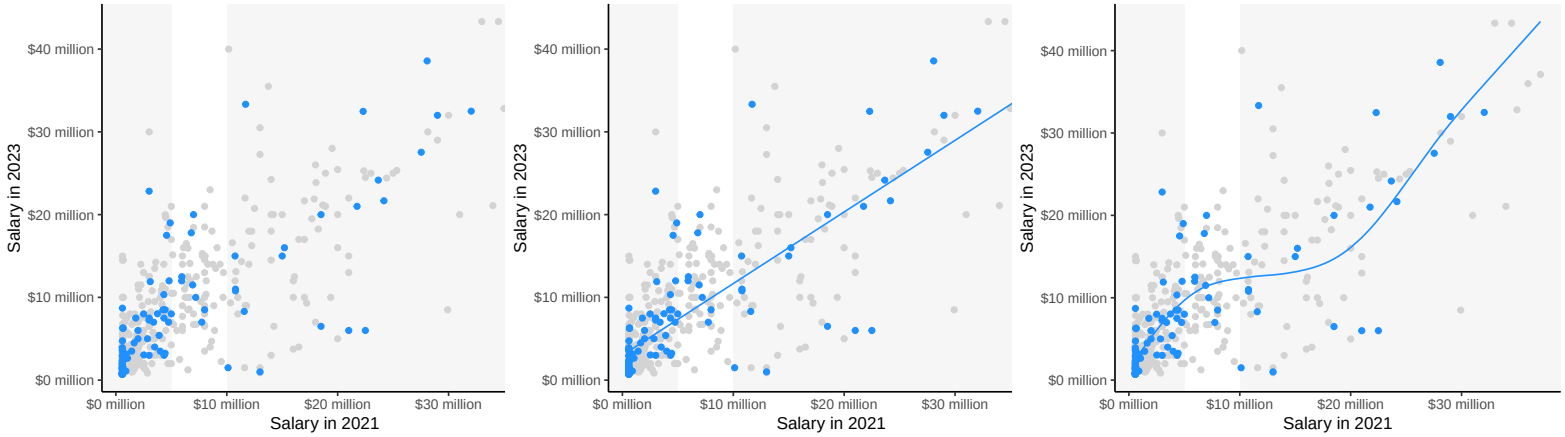


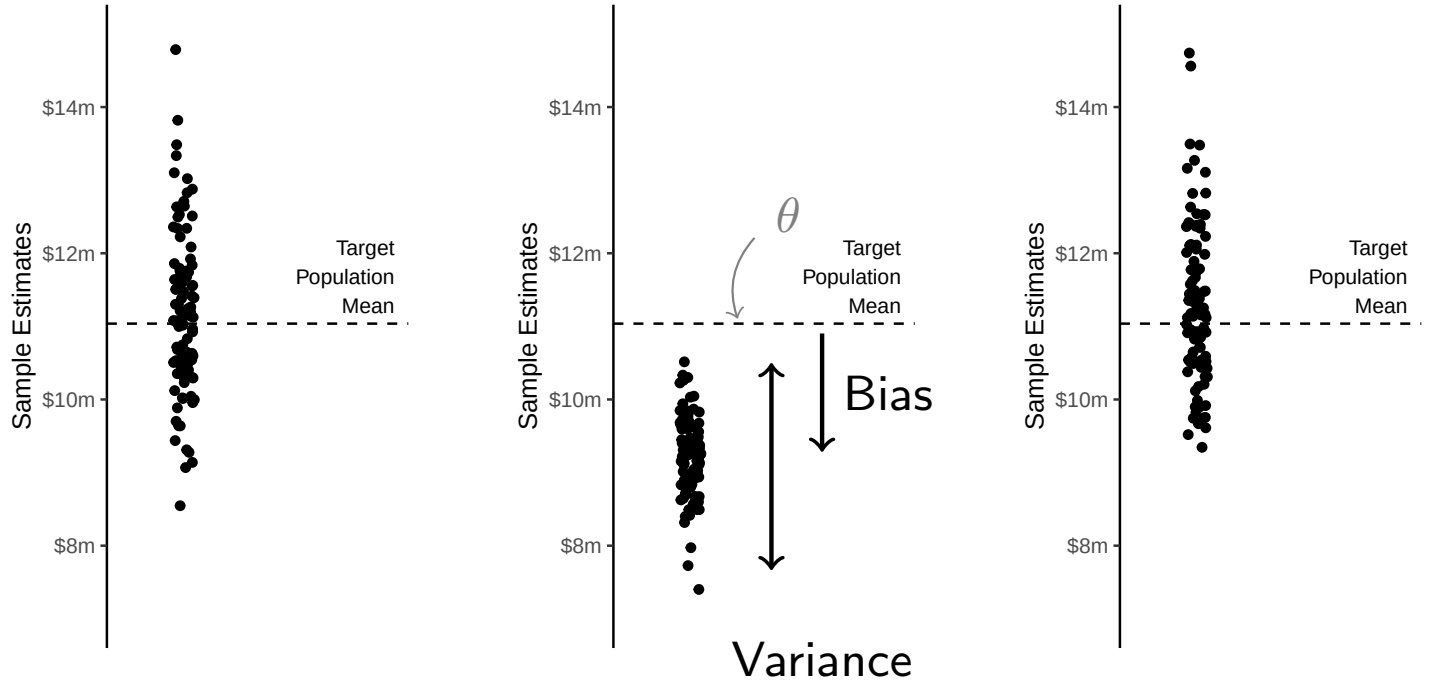
Comparing the estimators



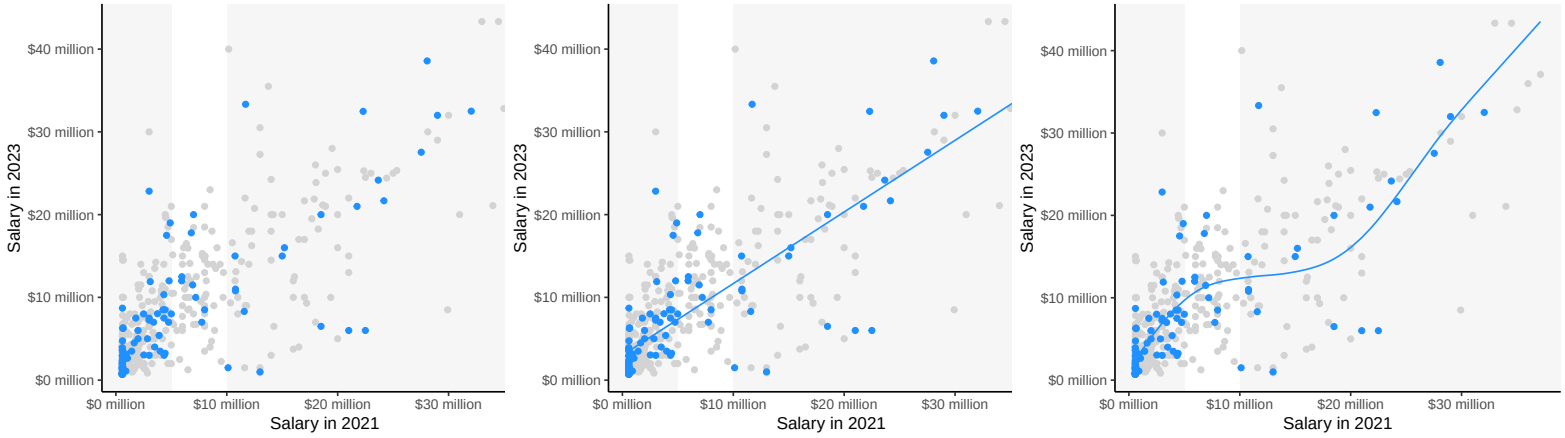


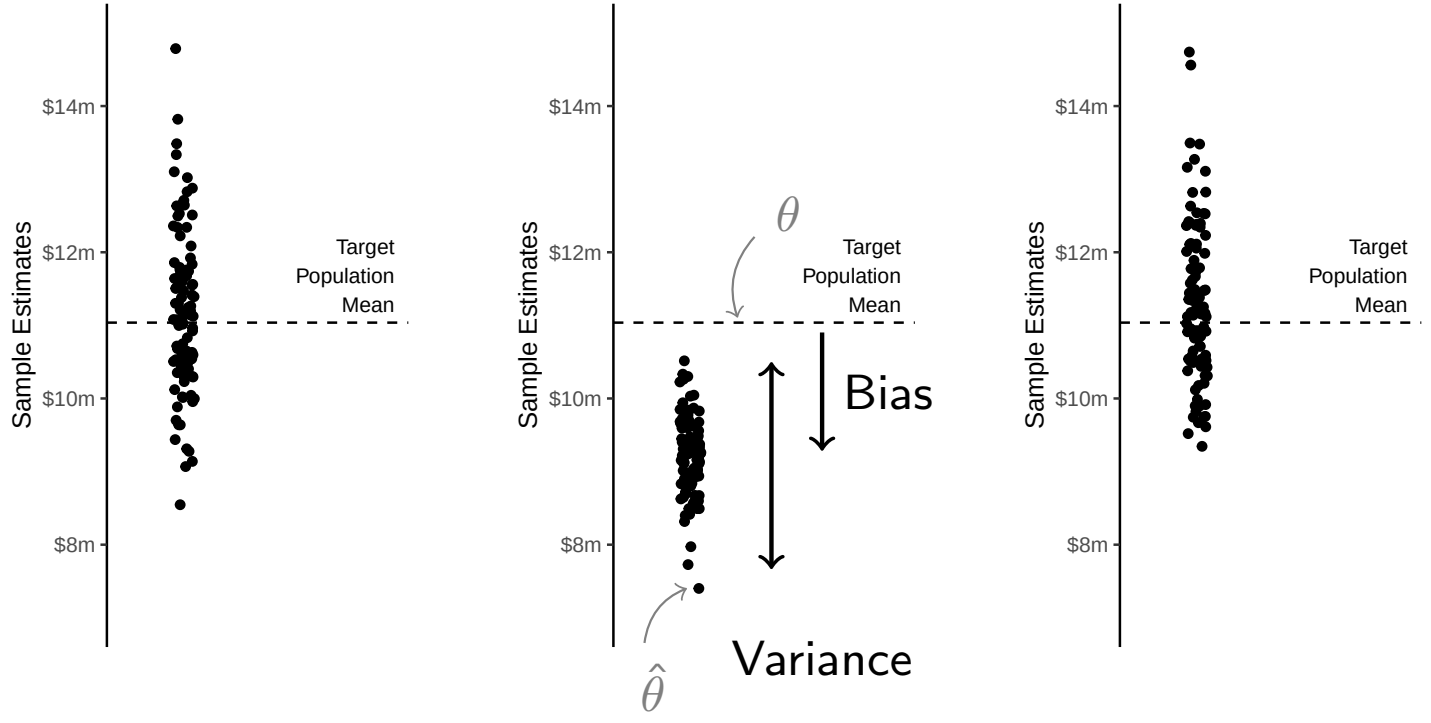
Comparing the estimators



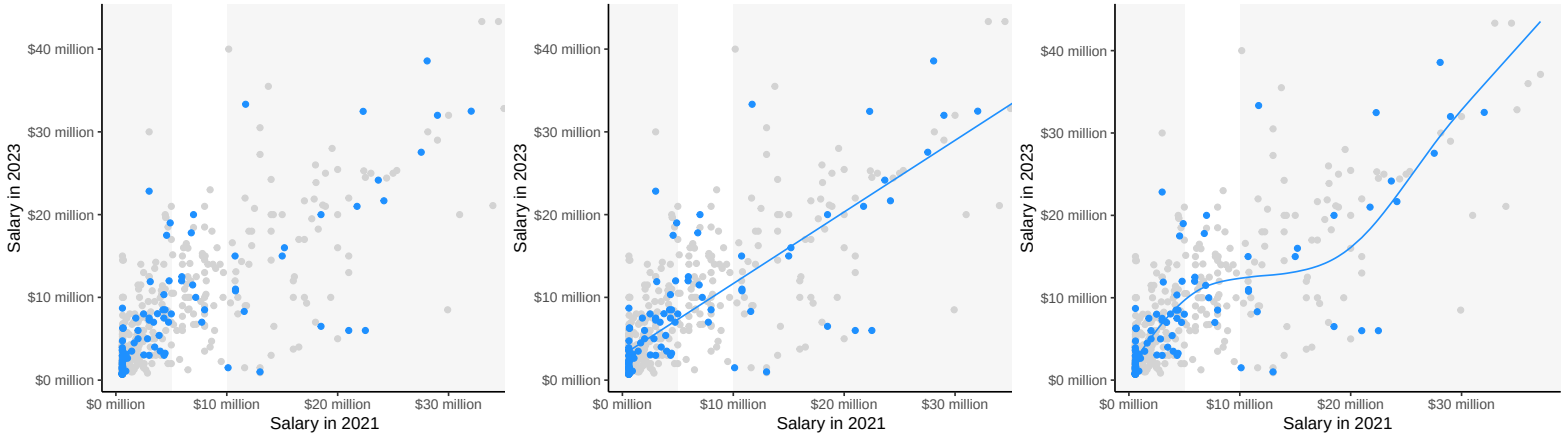


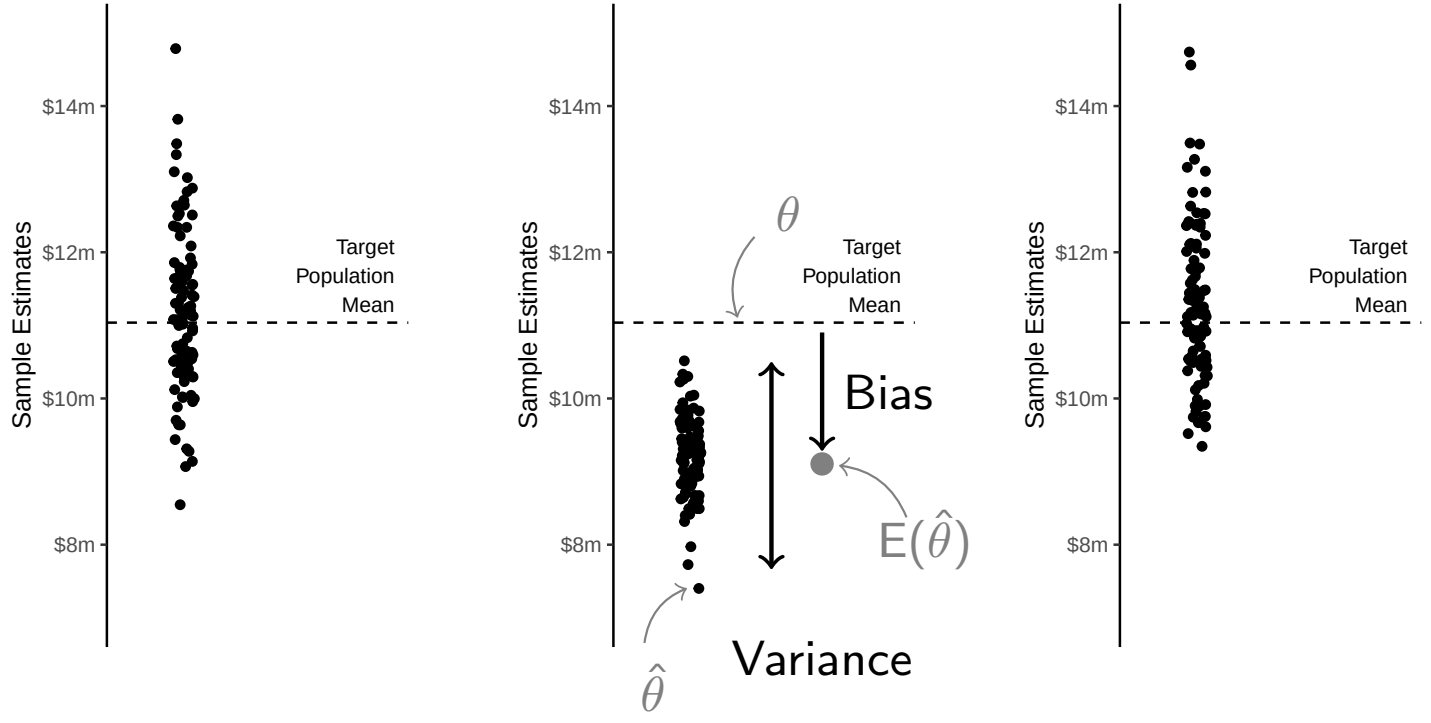
Comparing the estimators





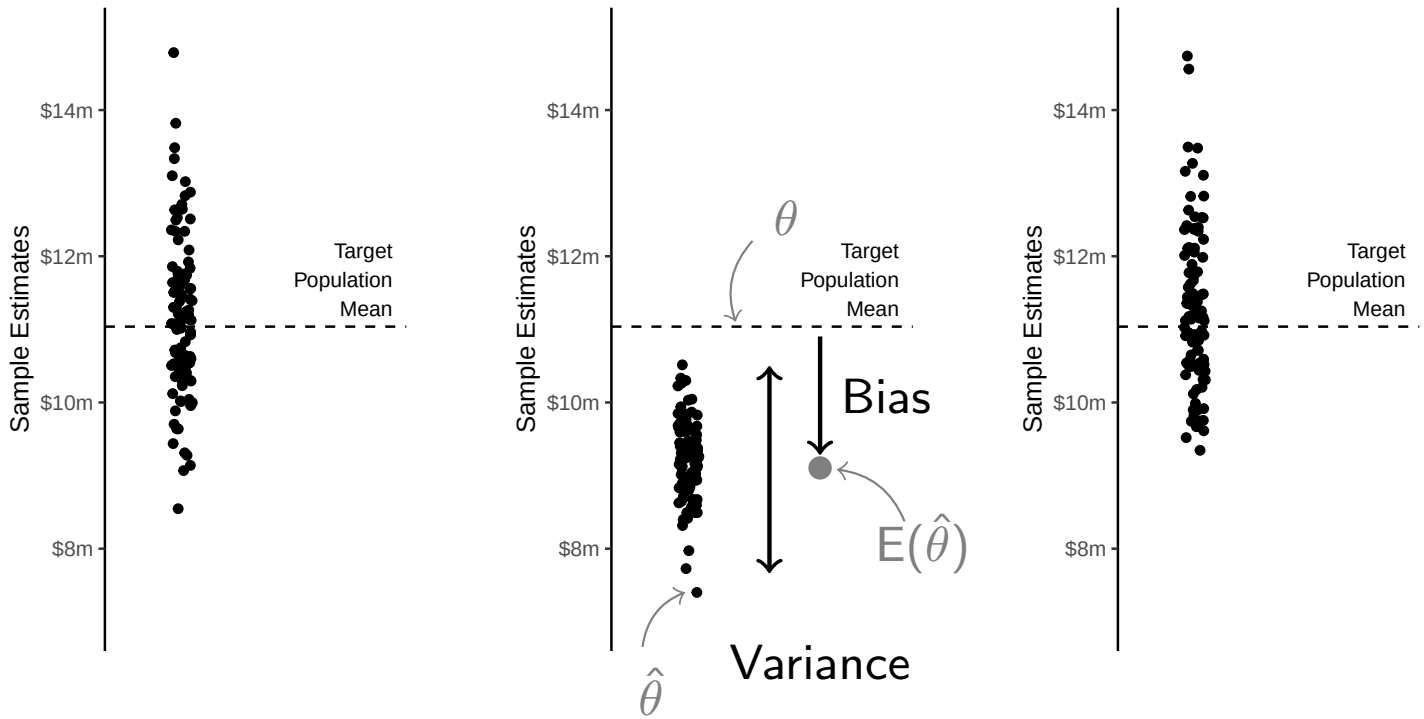
Comparing the estimators





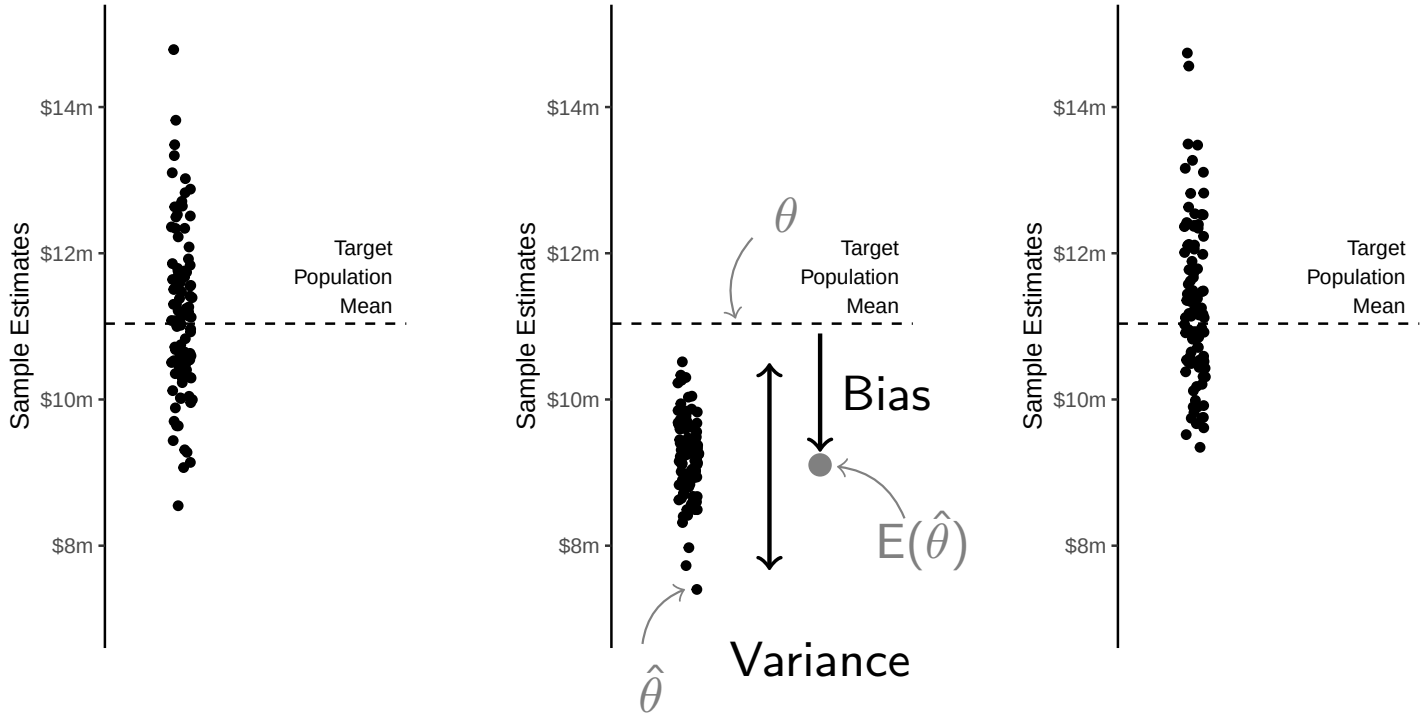
Comparing the estimators

$$\left(\hat{\theta} - \theta\right) = \left(\hat{\theta} - \mathsf{E}\left(\hat{\theta}\right)\right) + \left(\mathsf{E}\left(\hat{\theta}\right) - \theta\right)$$



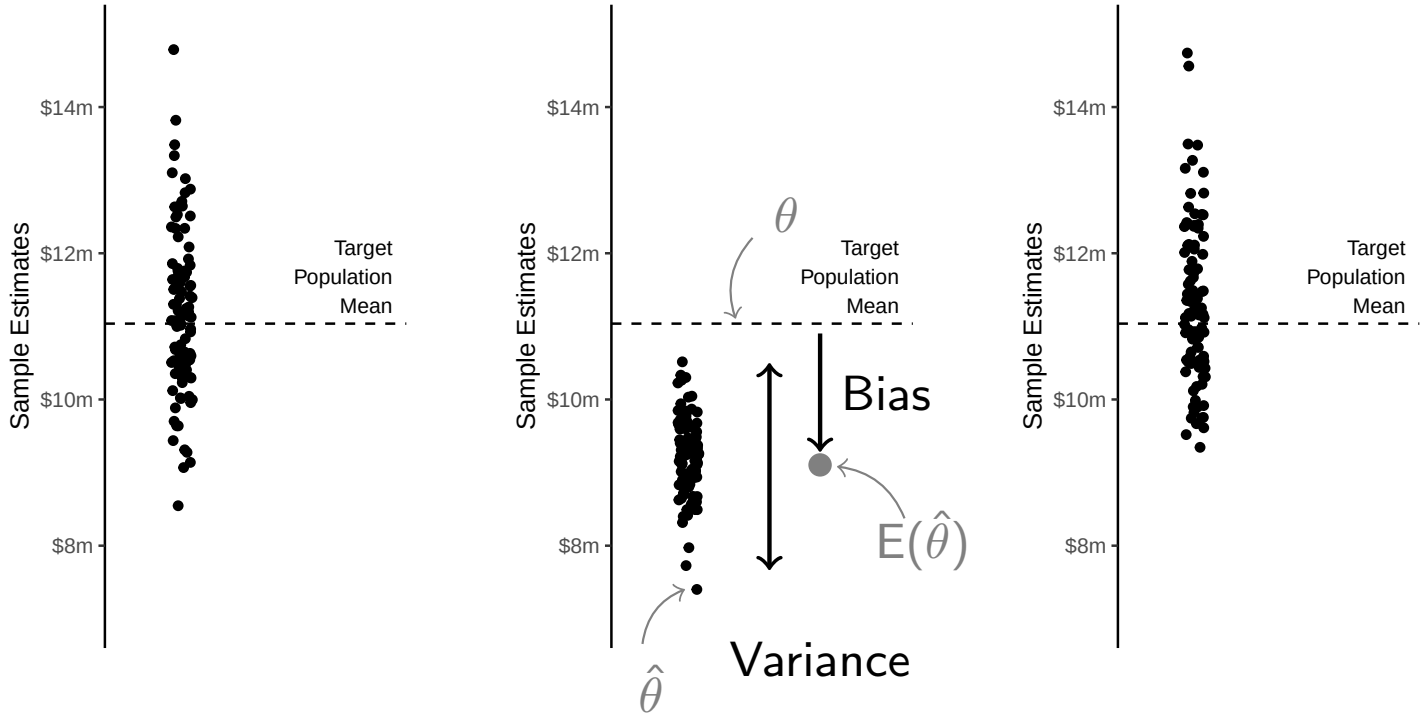
Comparing the estimators

$$\left(\hat{\theta} - \theta\right)^2 = \left(\hat{\theta} - \mathsf{E}\left(\hat{\theta}\right)\right)^2 + \left(\mathsf{E}\left(\hat{\theta}\right) - \theta\right)^2 + 2\left(\hat{\theta} - \mathsf{E}\left(\hat{\theta}\right)\right)\left(\mathsf{E}\left(\hat{\theta}\right) - \theta\right)$$



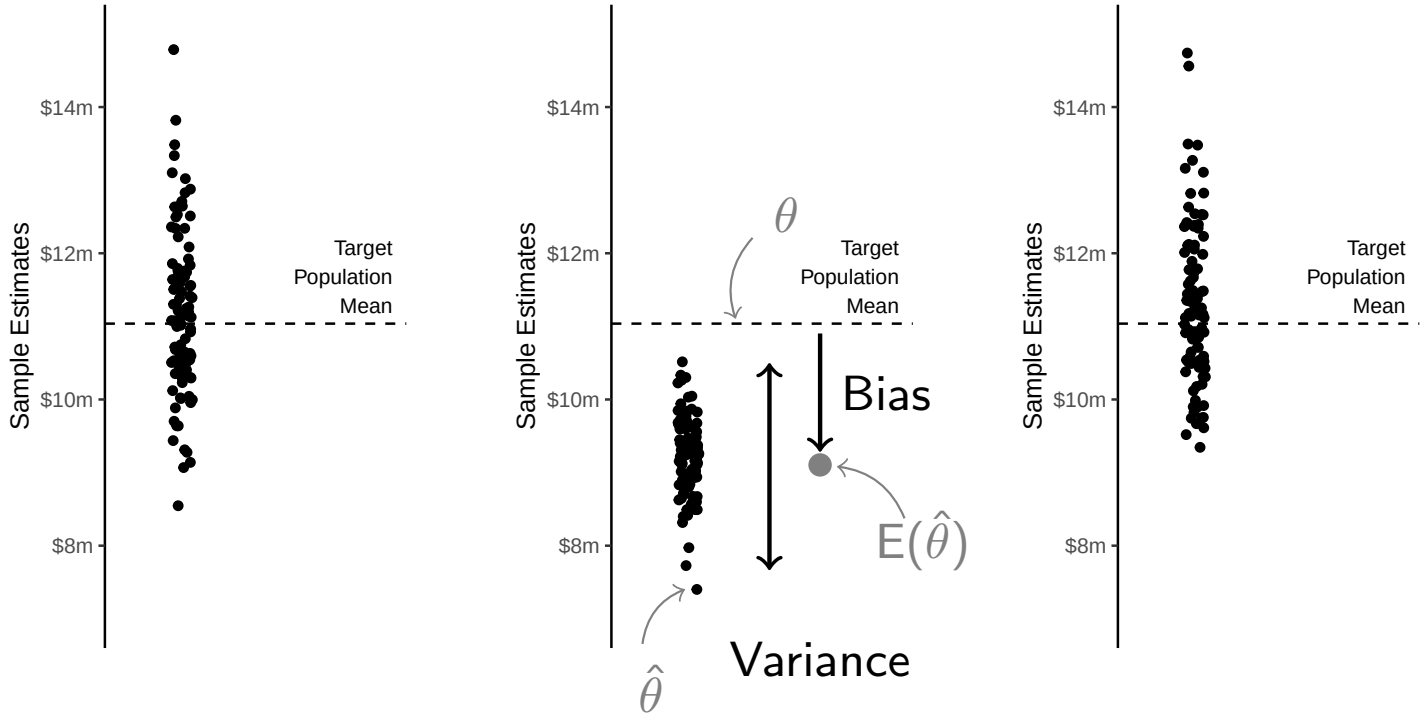
Comparing the estimators

$$\mathrm{E}\left[\left(\hat{\theta}-\theta\right)^2\right] = \mathrm{E}\left[\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)^2\right] + \mathrm{E}\left[\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2\right] + 2\mathrm{E}\left[\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)\right]$$



Comparing the estimators

$$\mathrm{E}\left[\left(\hat{\theta}-\theta\right)^2\right] = \mathrm{E}\left[\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)^2\right] + \mathrm{E}\left[\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2\right] + \cancel{2\mathrm{E}\left[\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)\right]} = 0$$



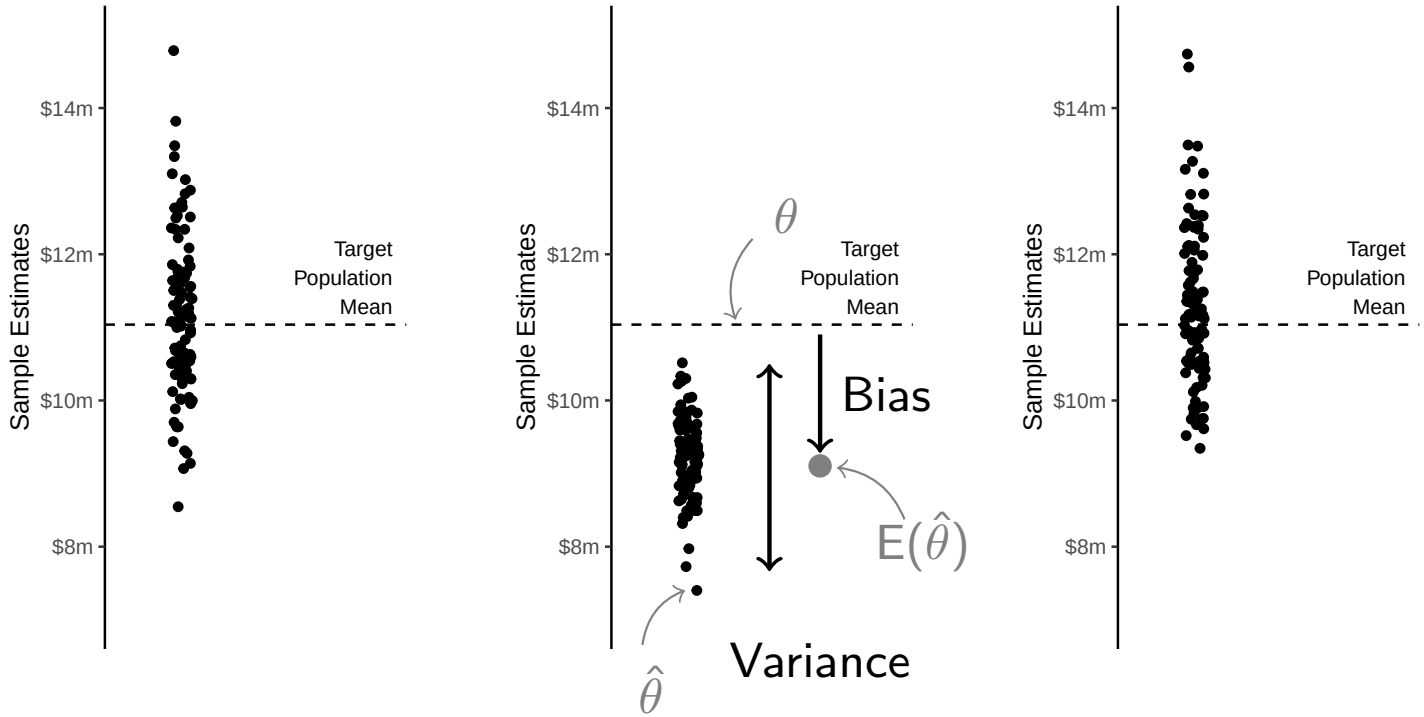
Comparing the estimators

$$\mathrm{E}\left[\left(\hat{\theta}-\theta\right)^2\right] = \mathrm{E}\left[\left(\hat{\theta}-\mathrm{E}\left(\hat{\theta}\right)\right)^2\right] + \mathrm{E}\left[\left(\mathrm{E}\left(\hat{\theta}\right)-\theta\right)^2\right]$$

Mean Squared Error

Variance

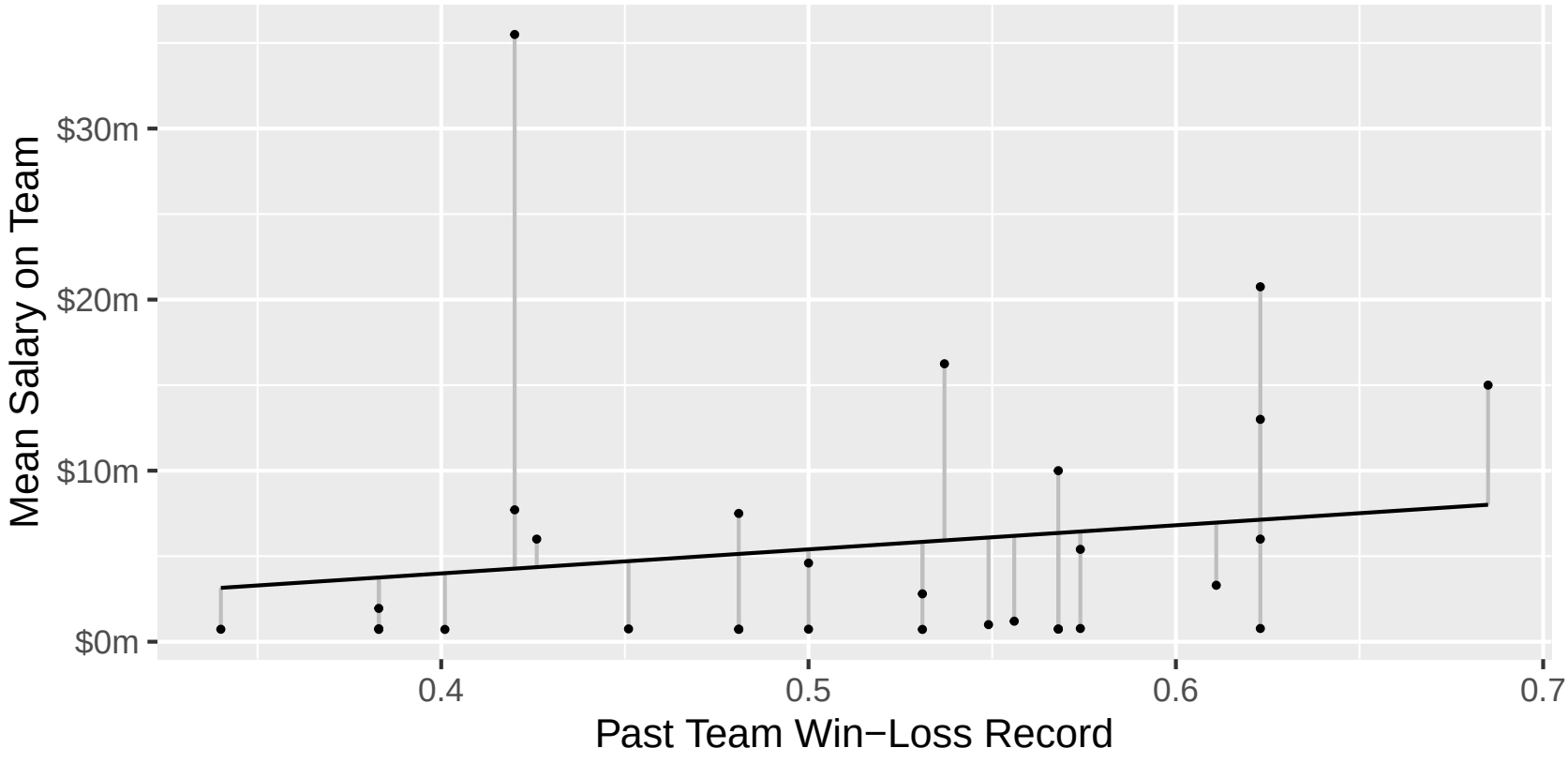
Bias^2



some statistical learning algorithms

Ordinary Least Squares

$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ with $\hat{\alpha}$ and $\hat{\beta}$ chosen to minimize $\underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}_{\text{Squared Error}}$



Penalized regression

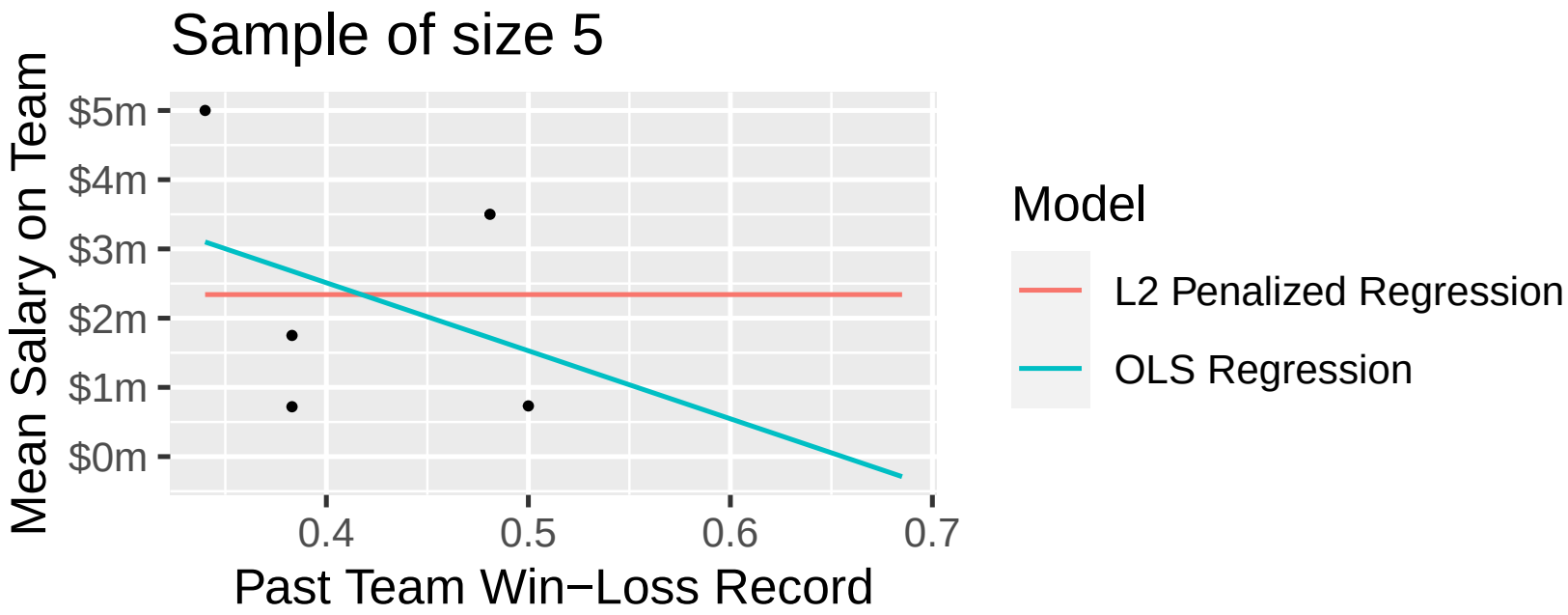
$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ with $\hat{\alpha}$ and $\hat{\beta}$ chosen to minimize

$$\underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}_{\text{Squared Error}} + \underbrace{\lambda \beta^2}_{\text{Penalty}}$$

Penalized regression

$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ with $\hat{\alpha}$ and $\hat{\beta}$ chosen to minimize

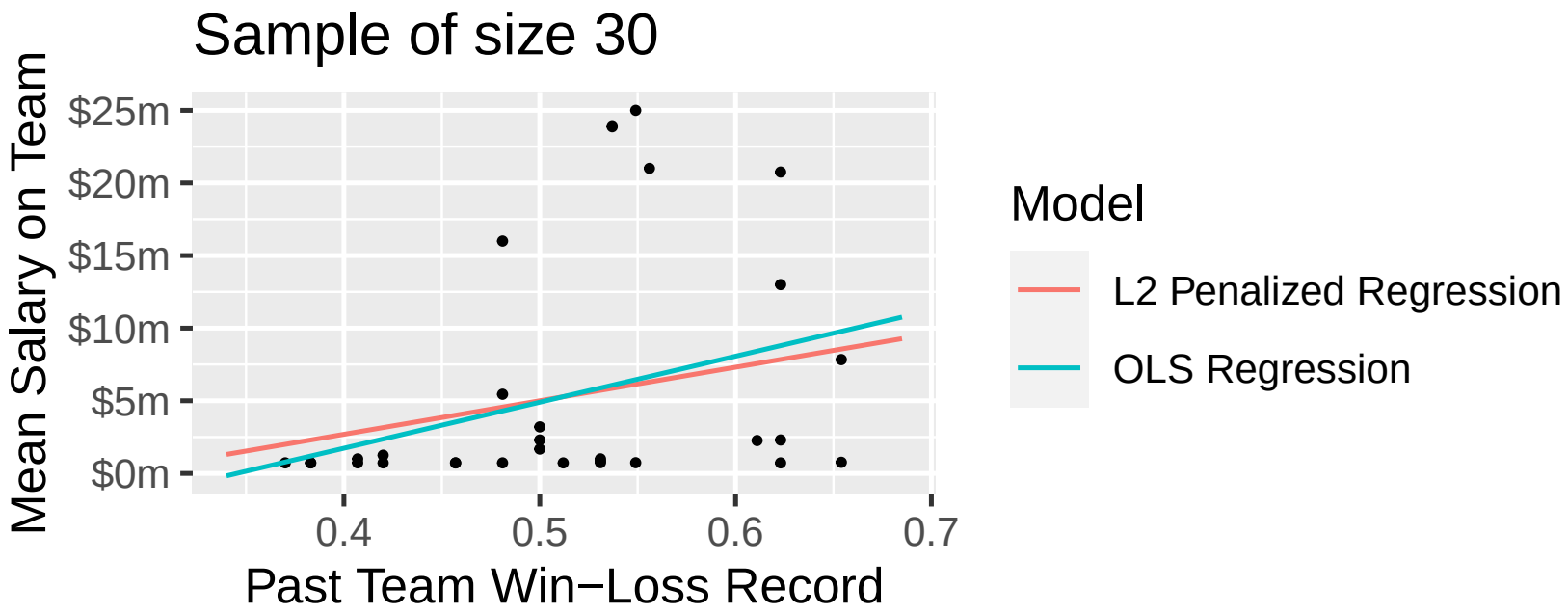
$$\underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}_{\text{Squared Error}} + \underbrace{\lambda \beta^2}_{\text{Penalty}}$$



Penalized regression

$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ with $\hat{\alpha}$ and $\hat{\beta}$ chosen to minimize

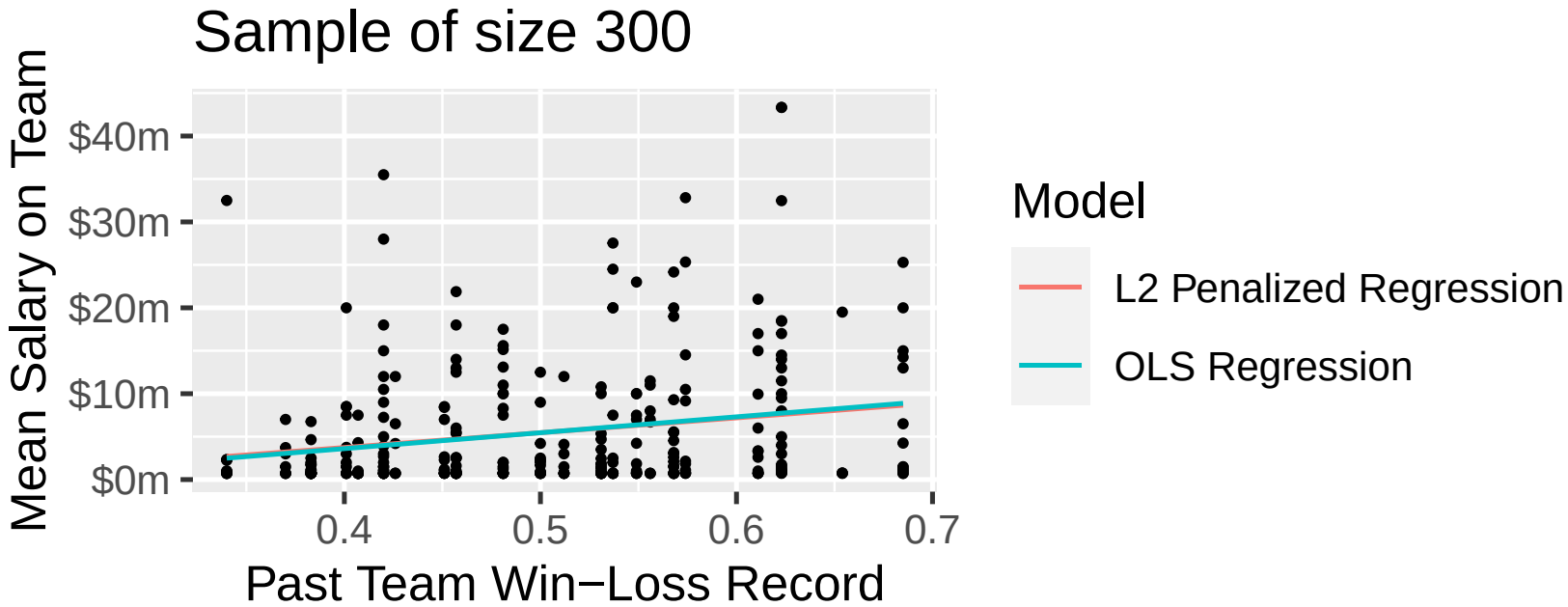
$$\underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}_{\text{Squared Error}} + \underbrace{\lambda \beta^2}_{\text{Penalty}}$$



Penalized regression

$\hat{Y}_i = \hat{\alpha} + \hat{\beta} X_i$ with $\hat{\alpha}$ and $\hat{\beta}$ chosen to minimize

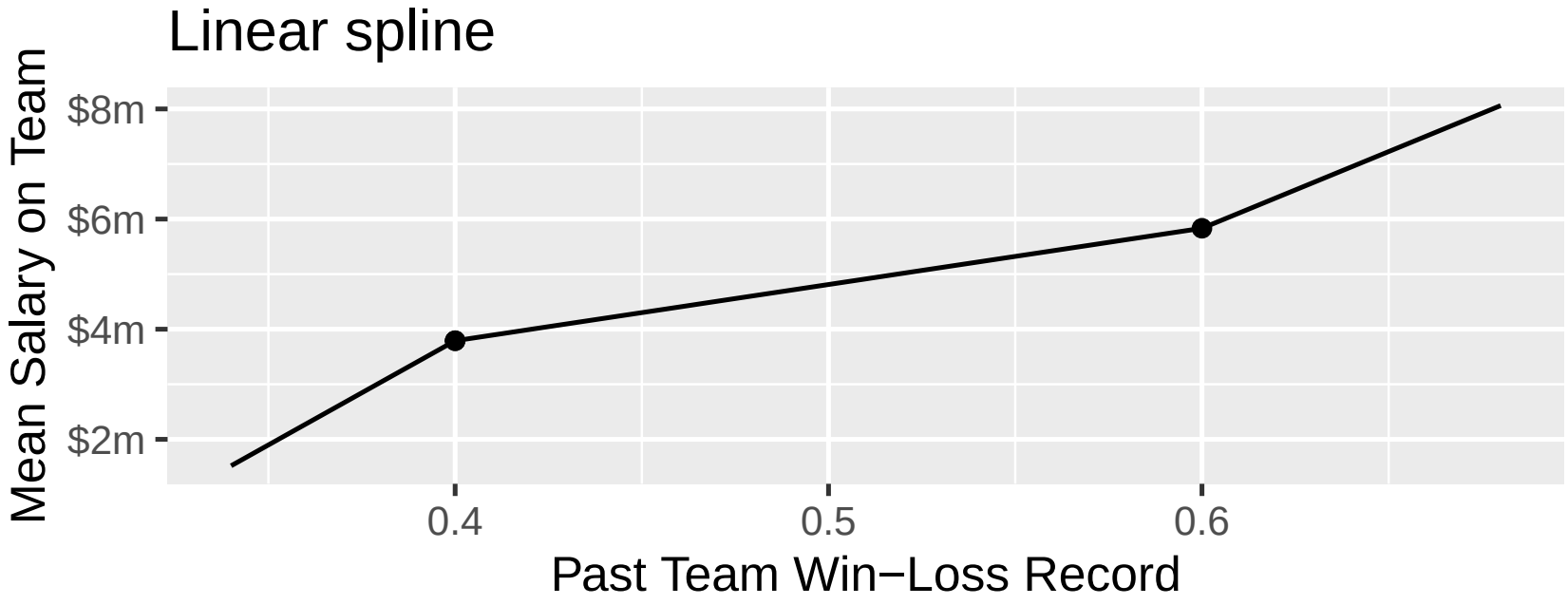
$$\underbrace{\sum_{i=1}^{n} \left(Y_i - \hat{Y}_i \right)^2}_{\text{Squared Error}} + \underbrace{\lambda \beta^2}_{\text{Penalty}}$$



ols regression	standard tool
penalized regression	OLS with reduced variance

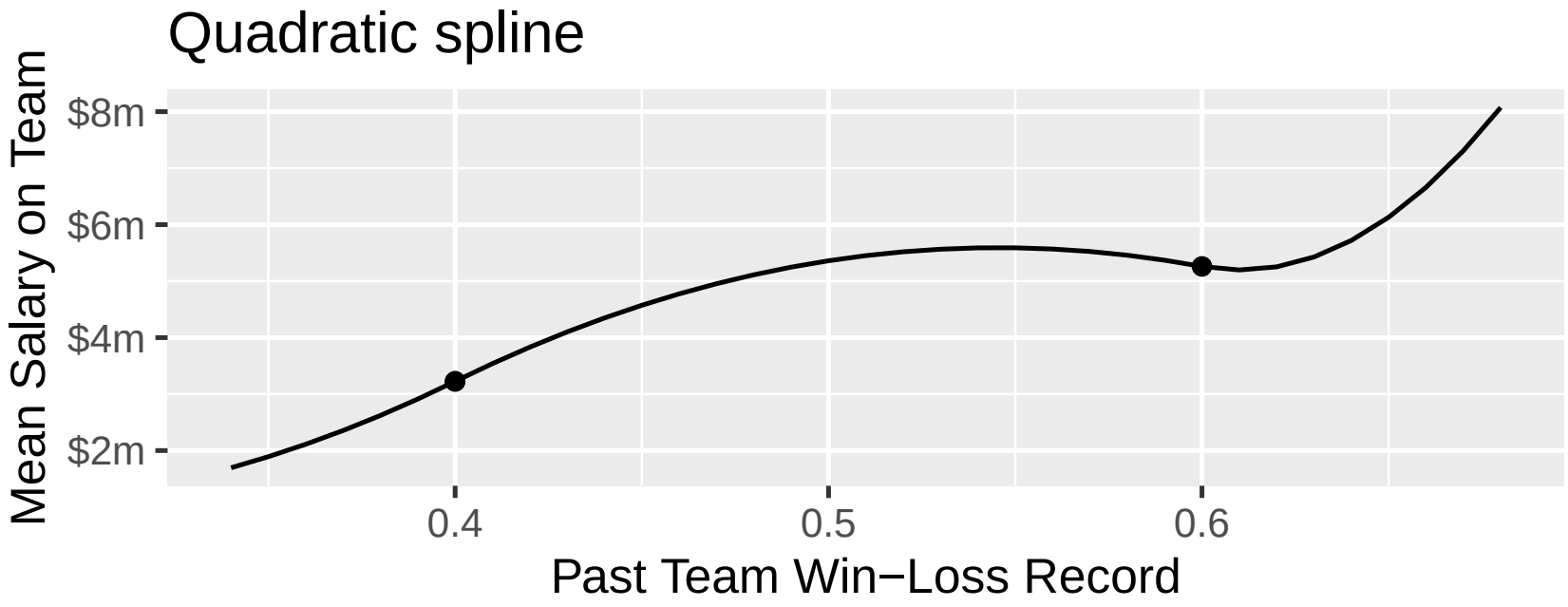
Splines

Regression with some terms estimated locally in regions of the data separated by **knots**



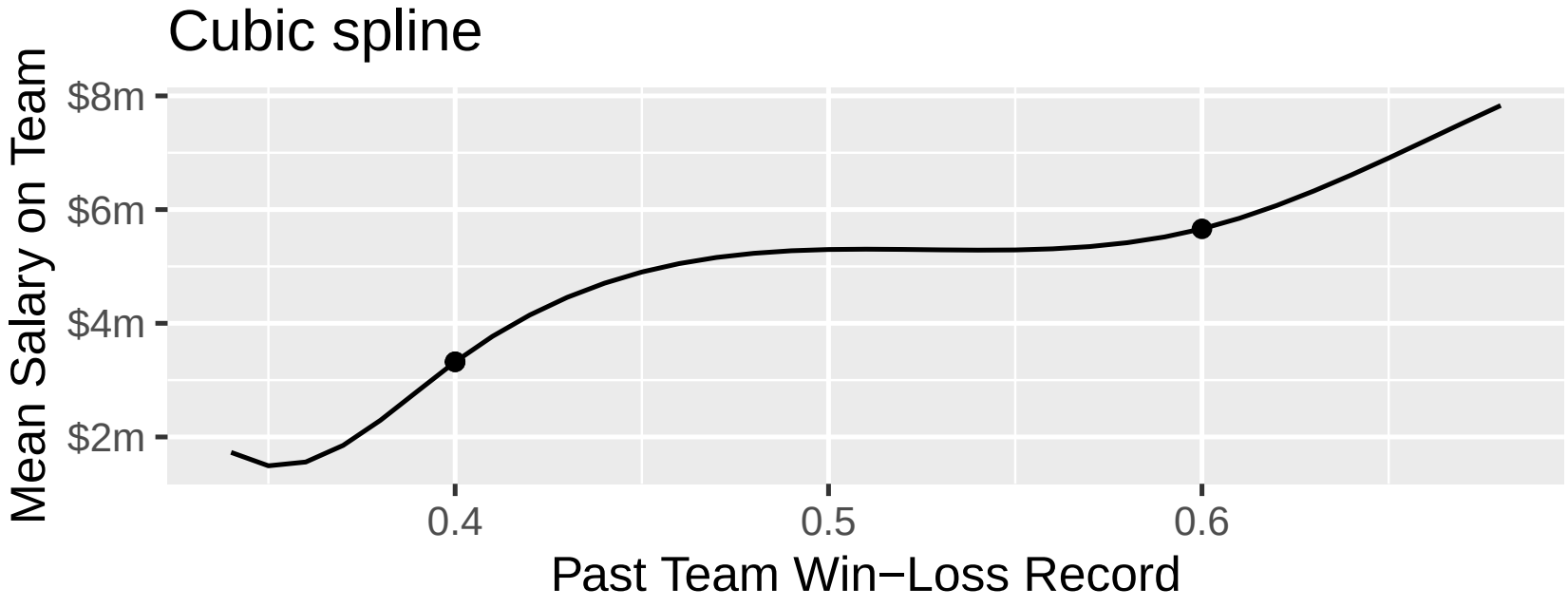
Splines

Regression with some terms estimated locally
in regions of the data separated by **knots**



Splines

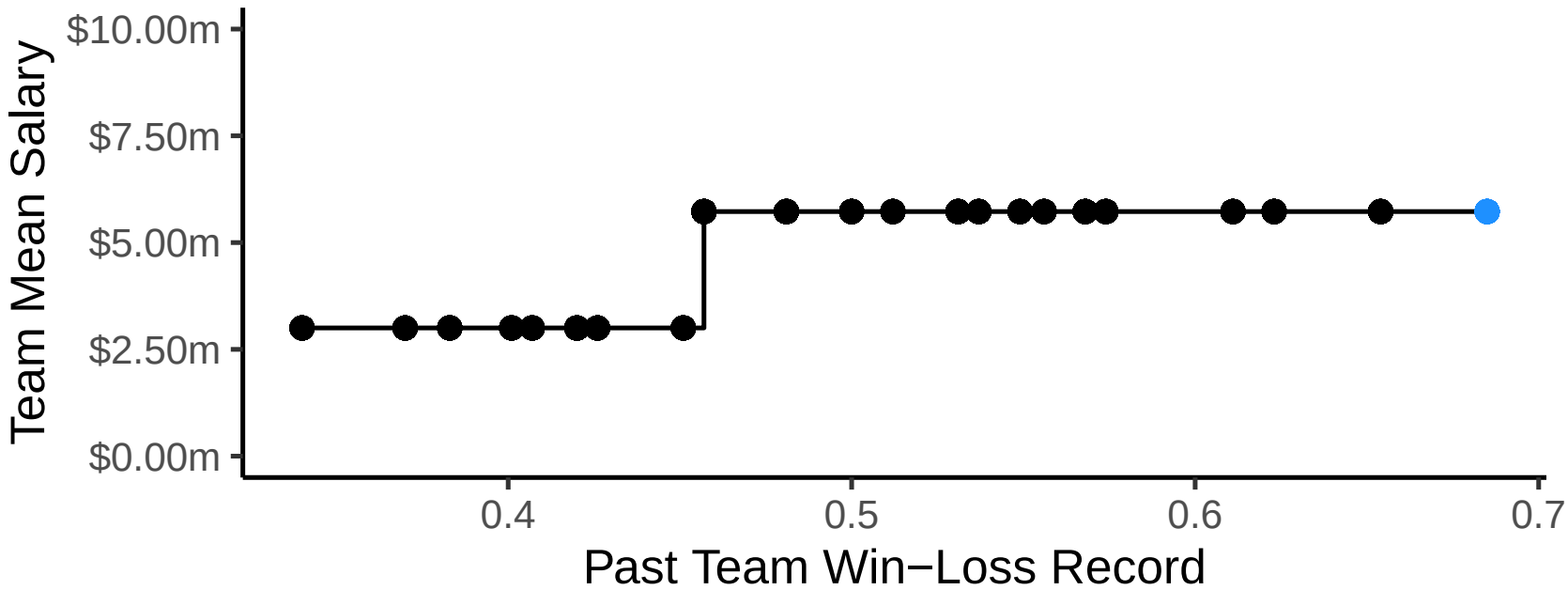
Regression with some terms estimated locally
in regions of the data separated by **knots**



ols regression	standard tool
penalized regression	OLS with reduced variance
splines	capture smooth nonlinearity

Decision tree

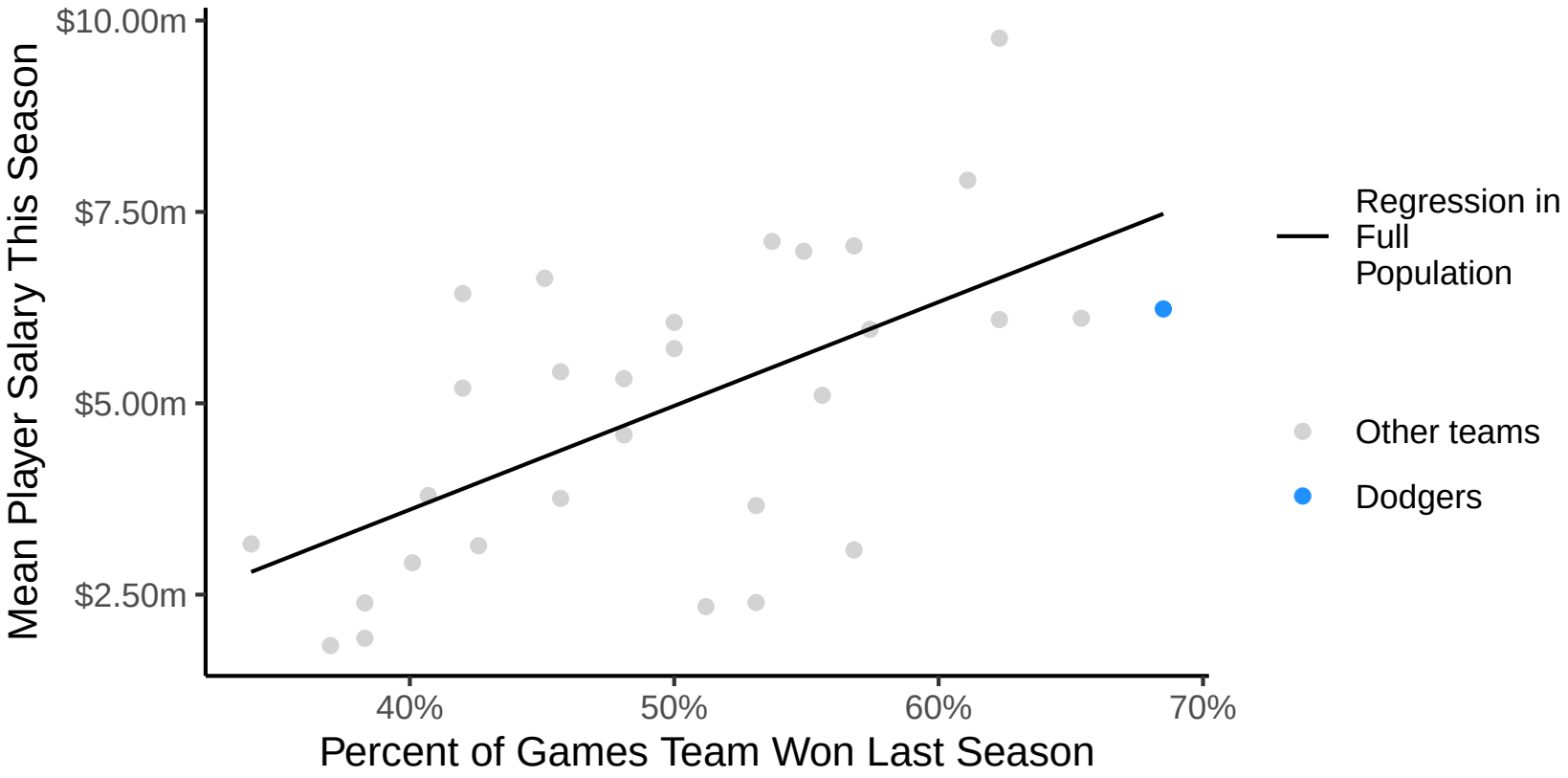
Assume the response is locally flat
Find places where it jumps

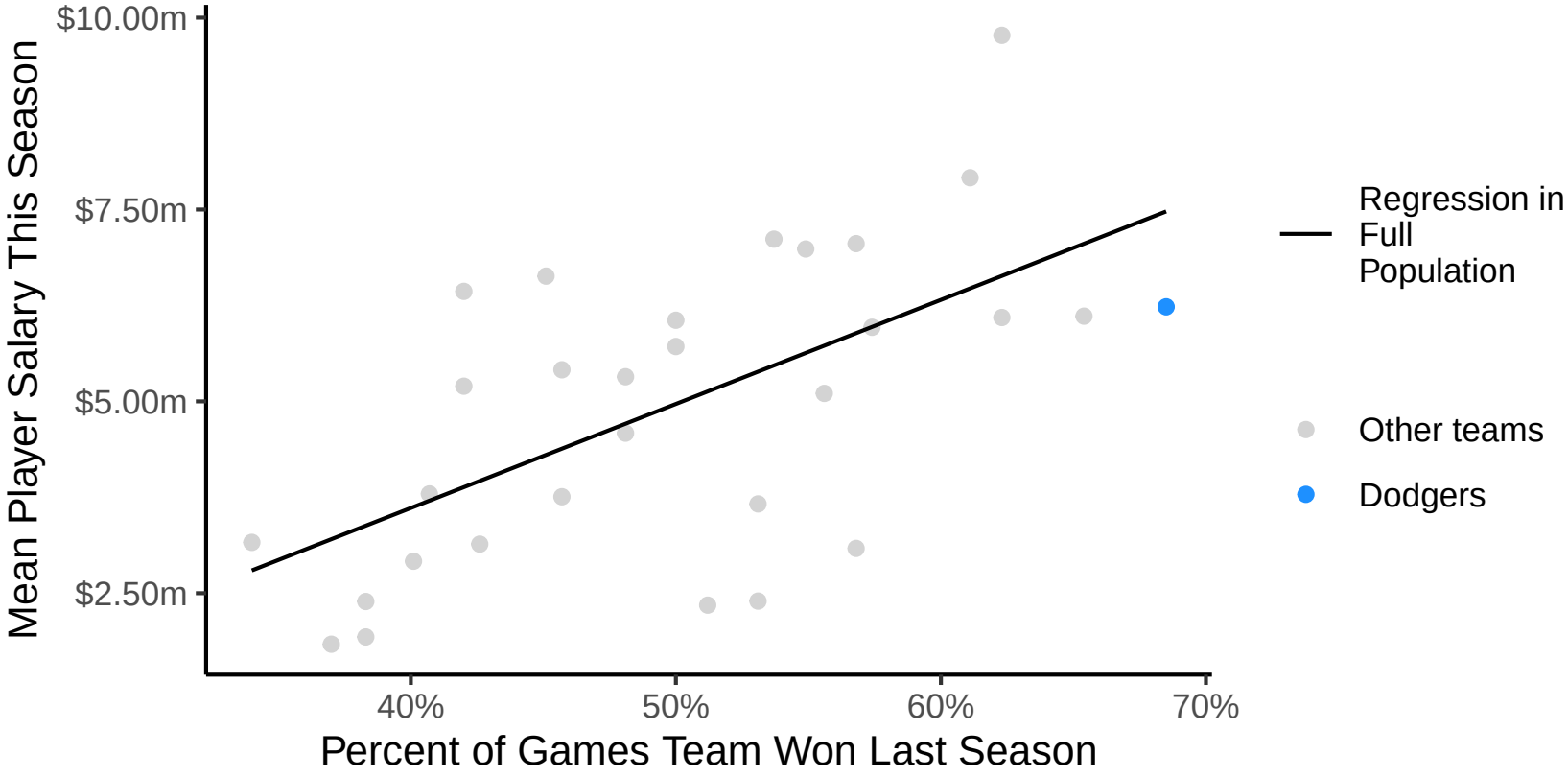


ols regression	standard tool
penalized regression	OLS with reduced variance
splines	capture smooth nonlinearity
trees	capture discrete nonlinearity

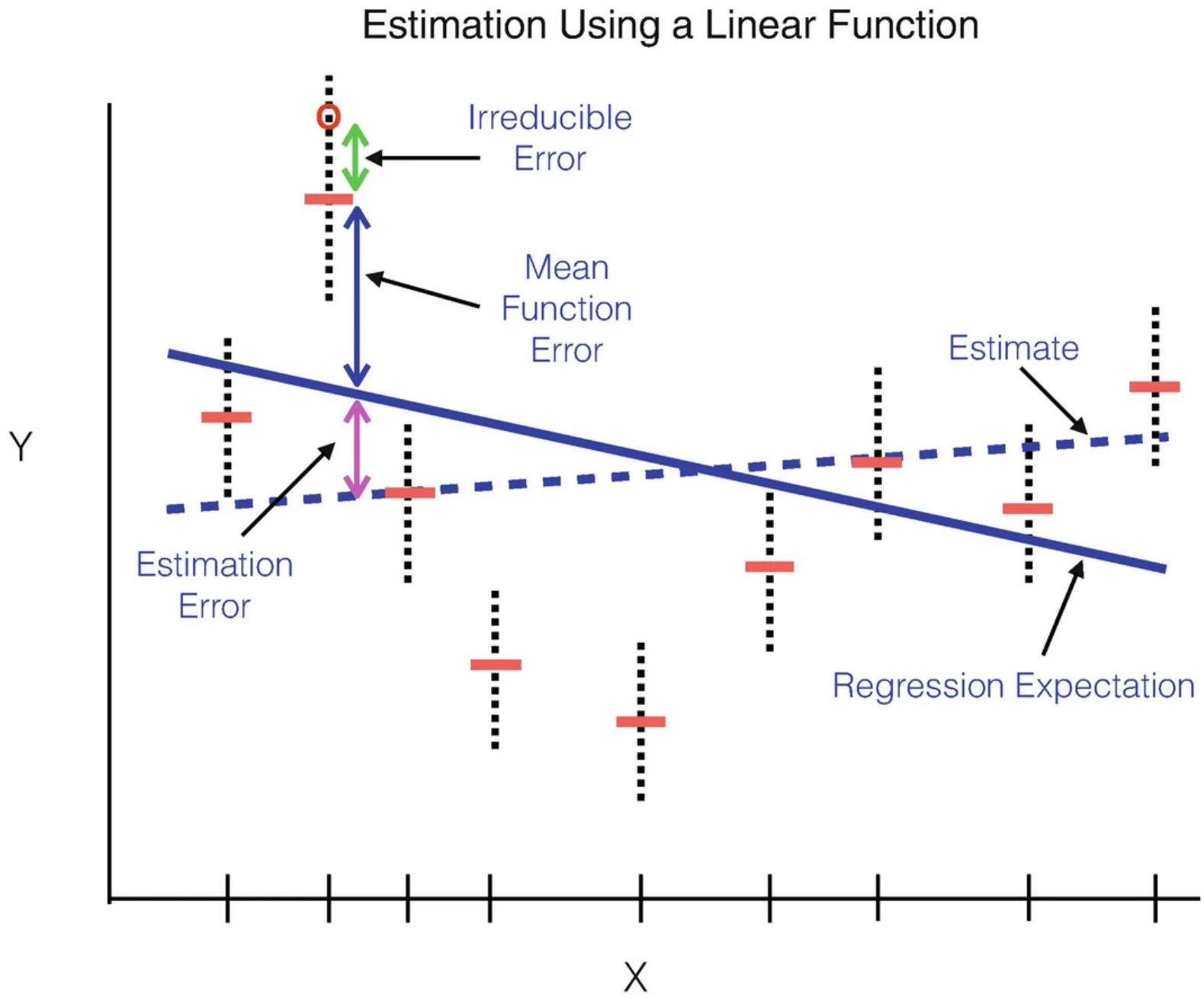
working with imperfect models

Drawing on Berk 2020.
Statistical Learning from a Regression Perspective


$10.00m
$7.50m
$5.00m
$2.50m
Mean Player Salary This Season
40%
50%
60%
70%
Percent of Games Team Won Last Season
Regression in Full Population
Other teams
Dodgers



The model is wrong. Why might we still use it?


Estimation Using a Linear Function
Irreducible Error
Mean Function Error
Estimate
Estimation Error
Regression Expectation
Y
X

Learning goals for today

By the end of class, you will be able to

- use statistical learning to estimate when data are sparse
- work with models that are "wrong"