Statistical Learning and Causal Inference

Jiaming Mao

Xiamen University

November 23, 2019

From Machine Learning to Econometrics

"What's in a name? that which we call a rose,
By any other name would smell as sweet." – *Juliet*

Machine Learning $\rightarrow$ Statistics $\rightarrow$ Econometrics

- Along this spectrum, the focus moves from prediction and pattern discovery to inference about causality and the underlying mechanisms that generate the observed data.

From Machine Learning to Econometrics









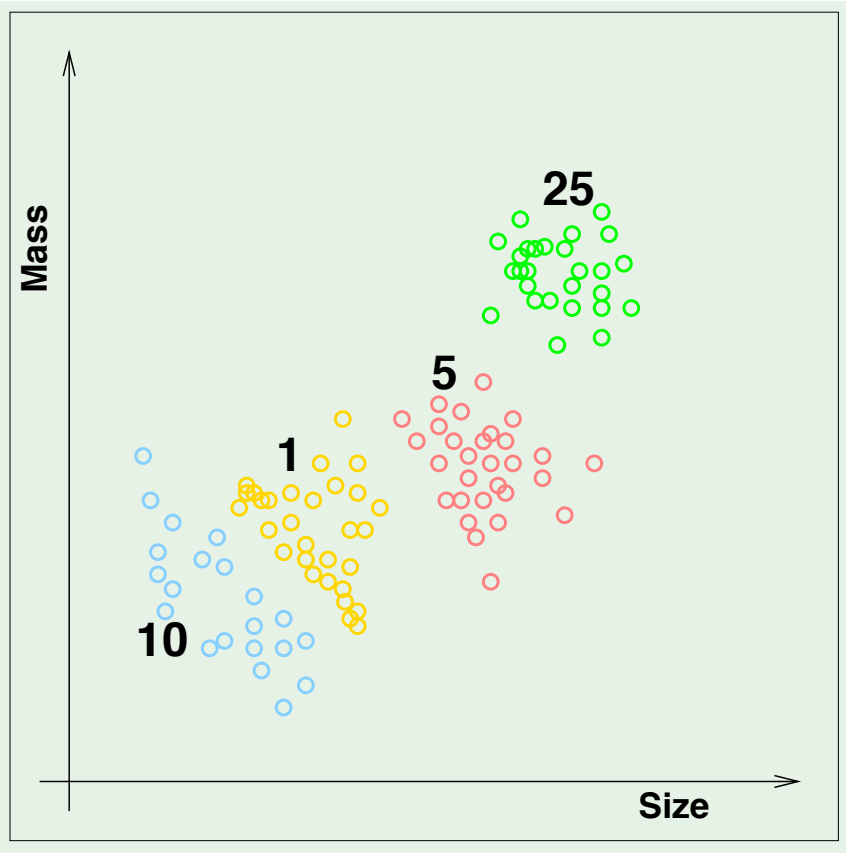
Classification

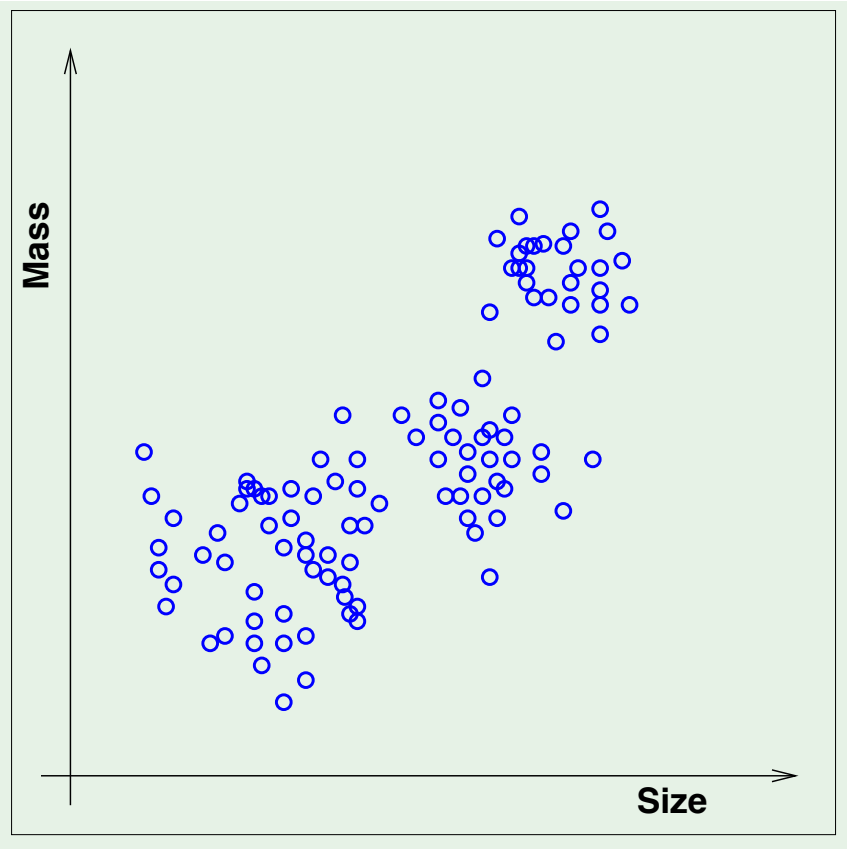
From Machine Learning to Econometrics

Discrete choice models:

- Which product will a consumer buy?
- Which market will a firm enter?
- Which political candidate will an individual vote for?

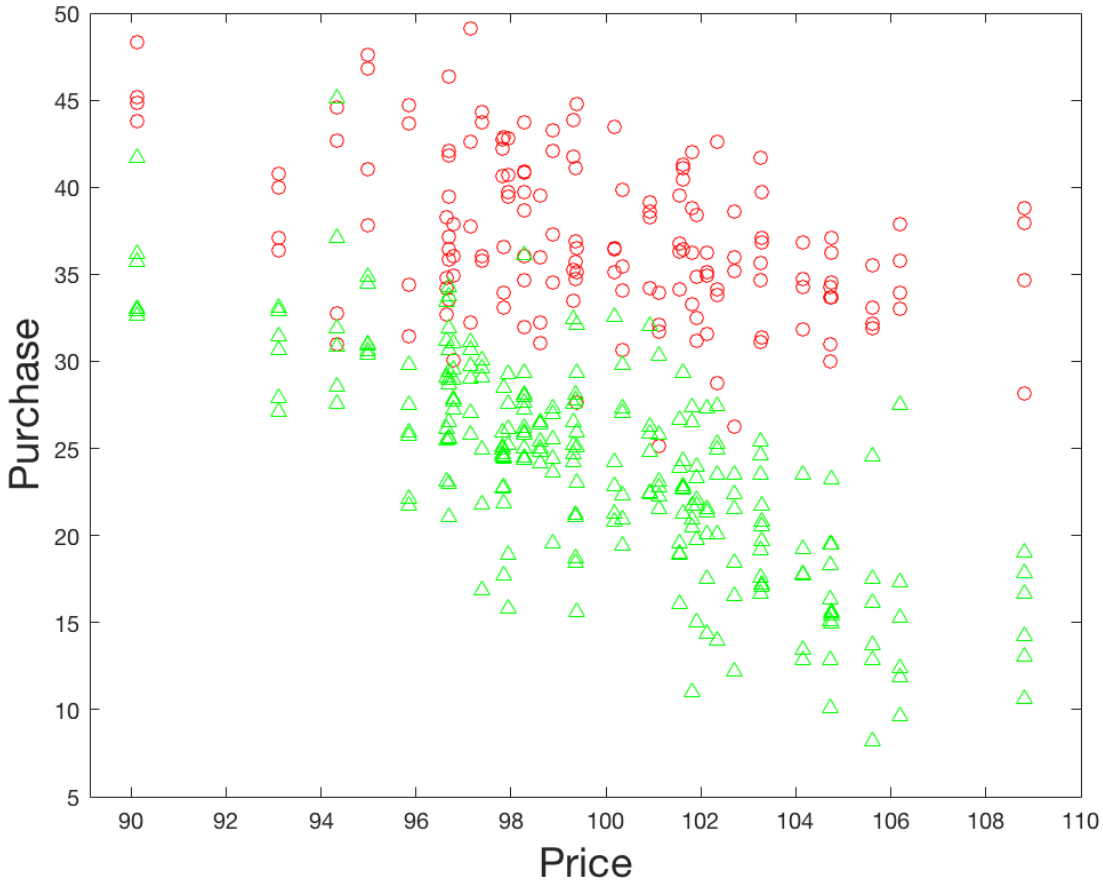
From Machine Learning to Econometrics





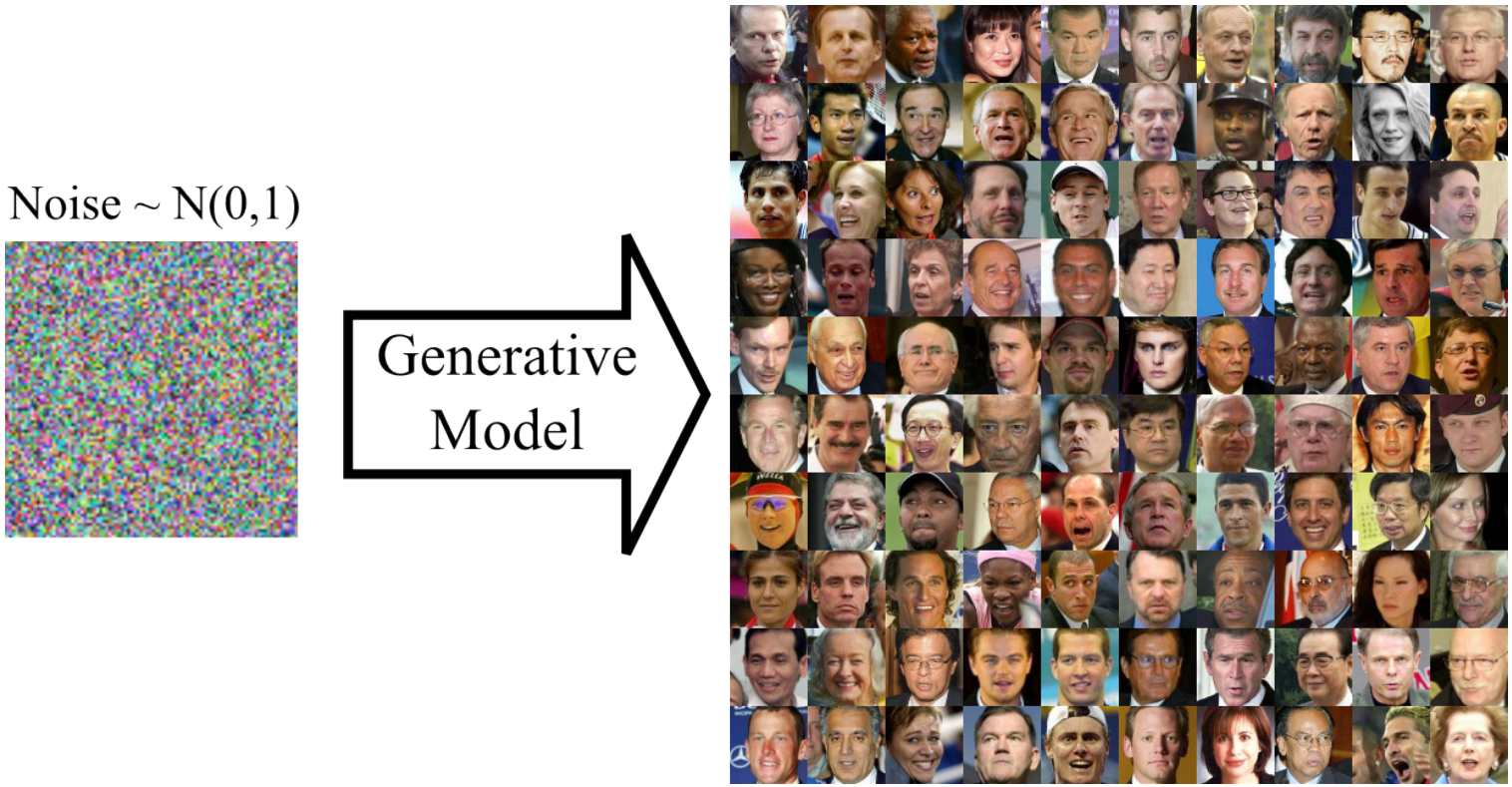
Unsupervised learning: coin recognition

From Machine Learning to Econometrics



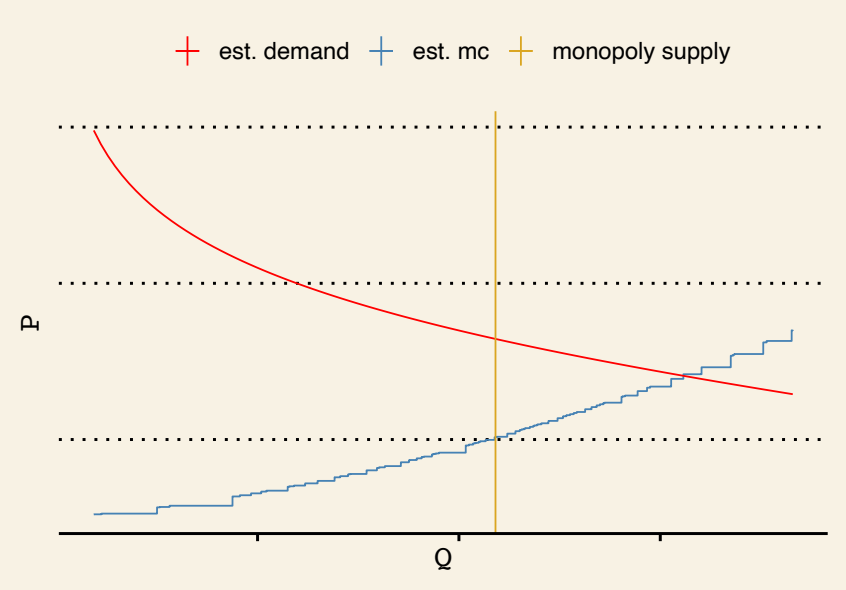
Unsupervised Learning: unobserved consumer types

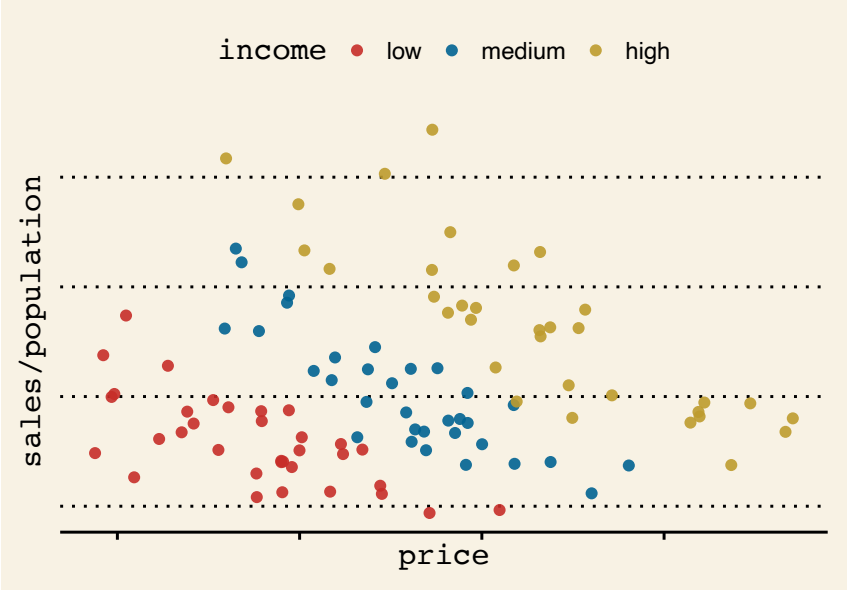
From Machine Learning to Econometrics



Generative Adversarial Network (GAN)

From Machine Learning to Econometrics





Structural econometric models

Machine Learning Methods for Economic Applications

"In this paper, we review and apply several popular methods from the machine learning literature to the problem of demand estimation ... we compare these methods to standard econometric models that are used by practitioners to study demand ... we used sales data on salty snacks [from] scanner panel data from grocery stores ... In our results, we find that the six models we use from the statistics and computer science literature predict demand out of sample in standard metrics much more accurately than a panel data or logistic model." – Bajari et al. (2015)

Machine Learning Methods for Economic Applications

	Validation		Out-of-Sample		
	RMSE	Std. Err.	RMSE	Std. Err.	Weight
Linear	1.169	0.022	1.193	0.020	6.62%
Stepwise	0.983	0.012	1.004	0.011	12.13%
Forward Stagewise	0.988	0.013	1.003	0.012	0.00%
Lasso	1.178	0.017	1.222	0.012	0.00%
Random Forest	0.943	0.017	0.965	0.015	65.56%
SVM	1.046	0.024	1.068	0.018	15.69%
Bagging	1.355	0.030	1.321	0.025	0.00%
Logit	1.190	0.020	1.234	0.018	0.00%
Combined	0.924		0.946		100.00%
# of Obs	226,952		376,980		
Total Obs	1,510,563				
% of Total	15.0%		25.0%		

Bajari et al. (2015)

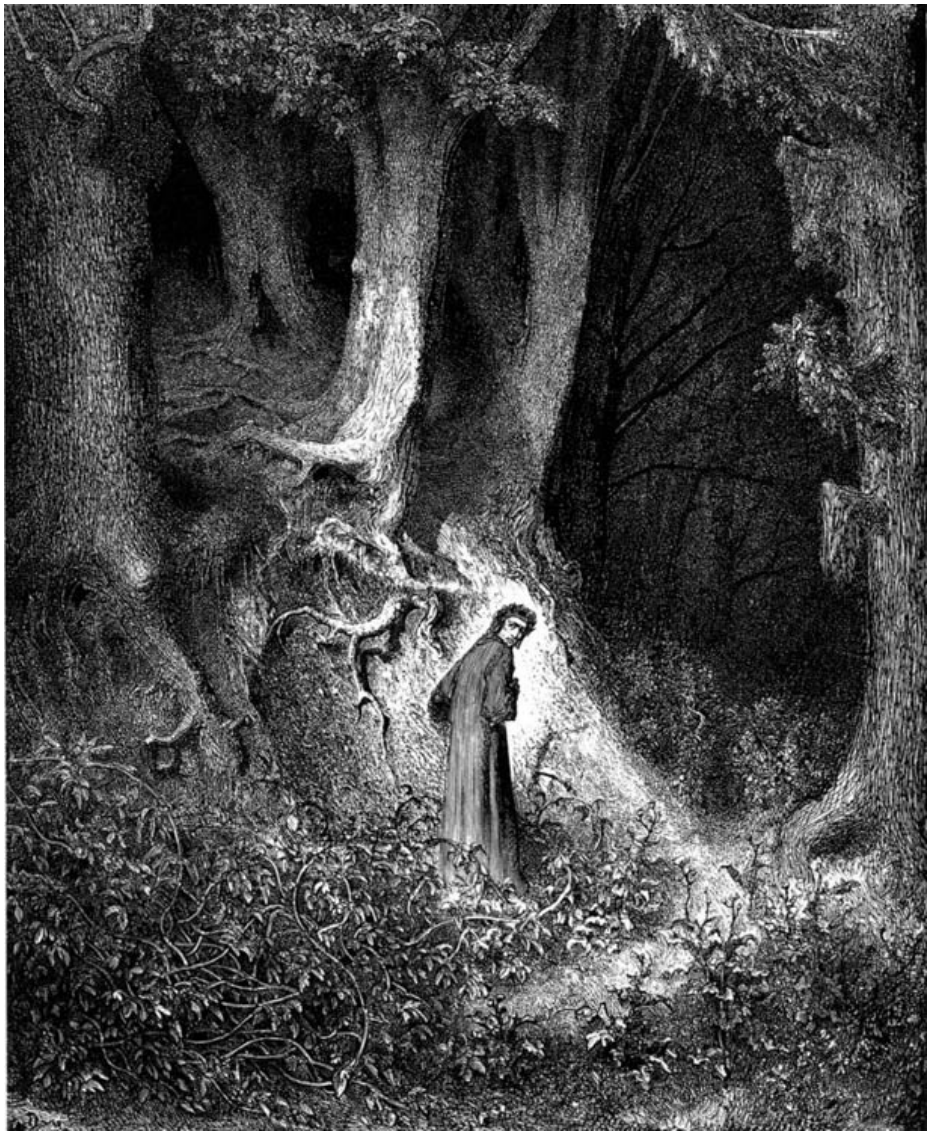
Road Map

1. Statistical Learning

2. Causal Inference

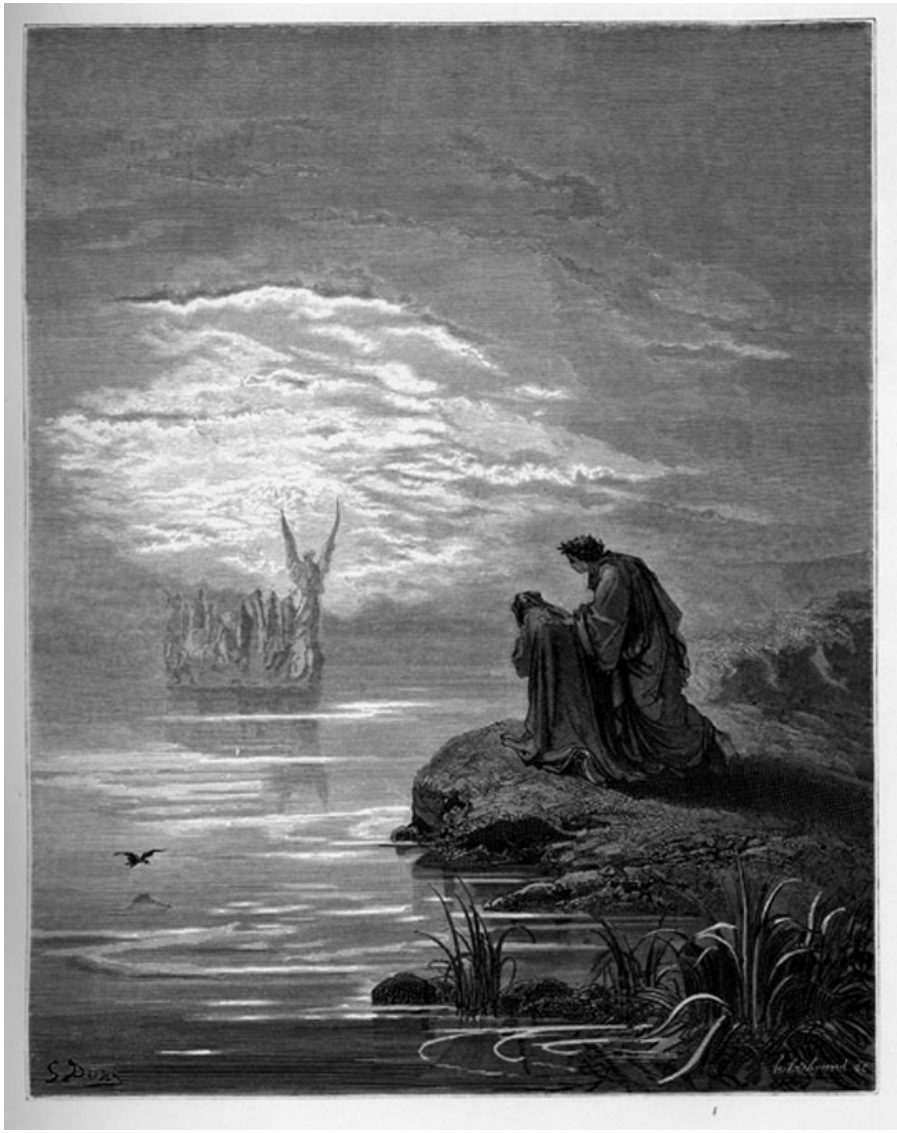
Road Map

Thematically, we follow the journey of a hero determined to seek knowledge from data, who departs the forest of ignorance,



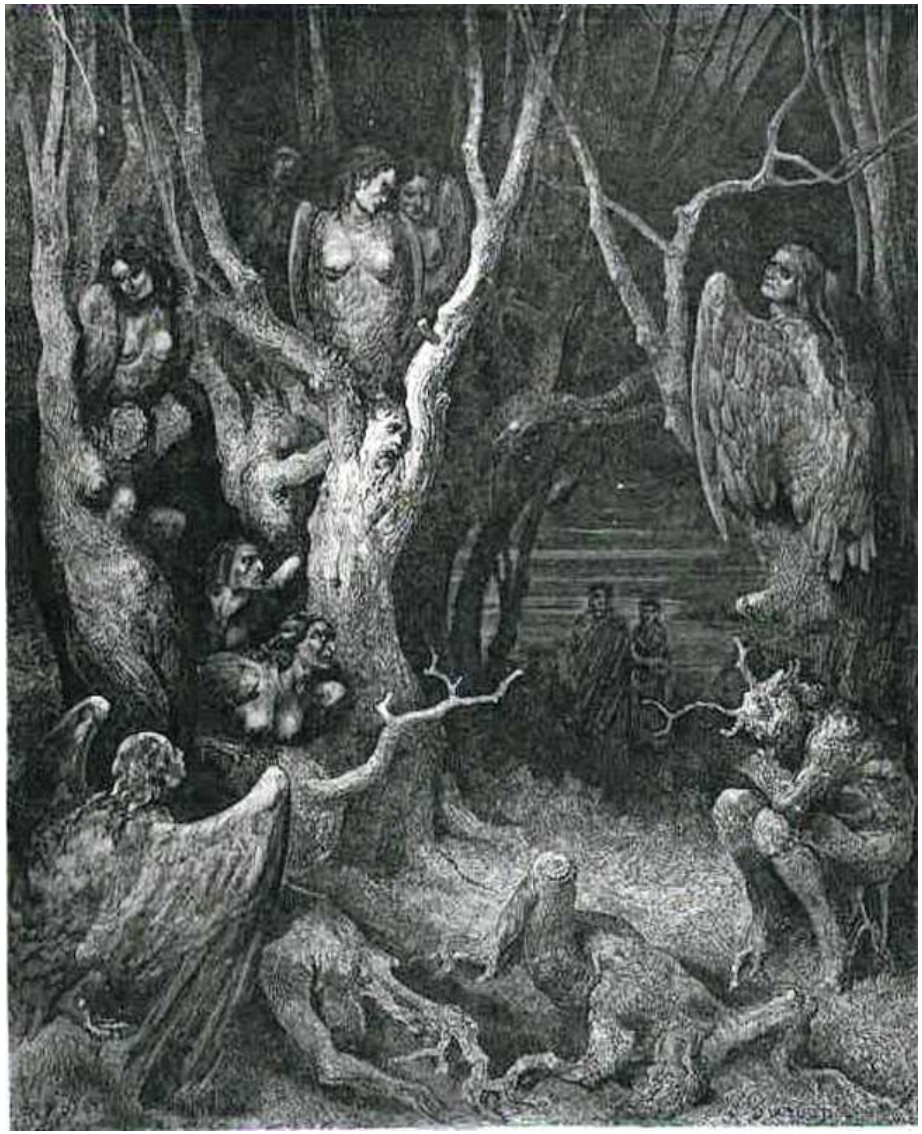
Road Map

... and journeys to the realm of patterns, where patterns in data are discovered and used to make predictions,



Road Map

... along the way he encounters the false prophets of correlation equals causation,



Road Map

... and then arrives at the land of causality, where people are serious about whether any two sets of observed phenomena are linked causally,

Road Map

... from where our hero finally reaches the mount of scientific discovery, where the mechanisms that generate the observed phenomena are investigated in the hope of attaining true knowledge about the world.

Statistical Learning

"All models are wrong but some are useful." – George Box

"The existence of a problem in knowledge depends on the future being different from the past, while the possibility of a solution of the problem depends on the future being like the past." – Frank Knight

Statistical Learning

- Given variables x and y, how do we characterize the statistical relationship between the two?
  - $p(x, y)$: joint distribution of x and y
- Oftentimes, we may not be interested in characterizing the full joint distribution $p(x, y)$. Instead, we are interested in predicting the value of y based on observed x.
  - We want to find a function $f(x)$ for predicting y given values of x.

Statistical Learning

Let

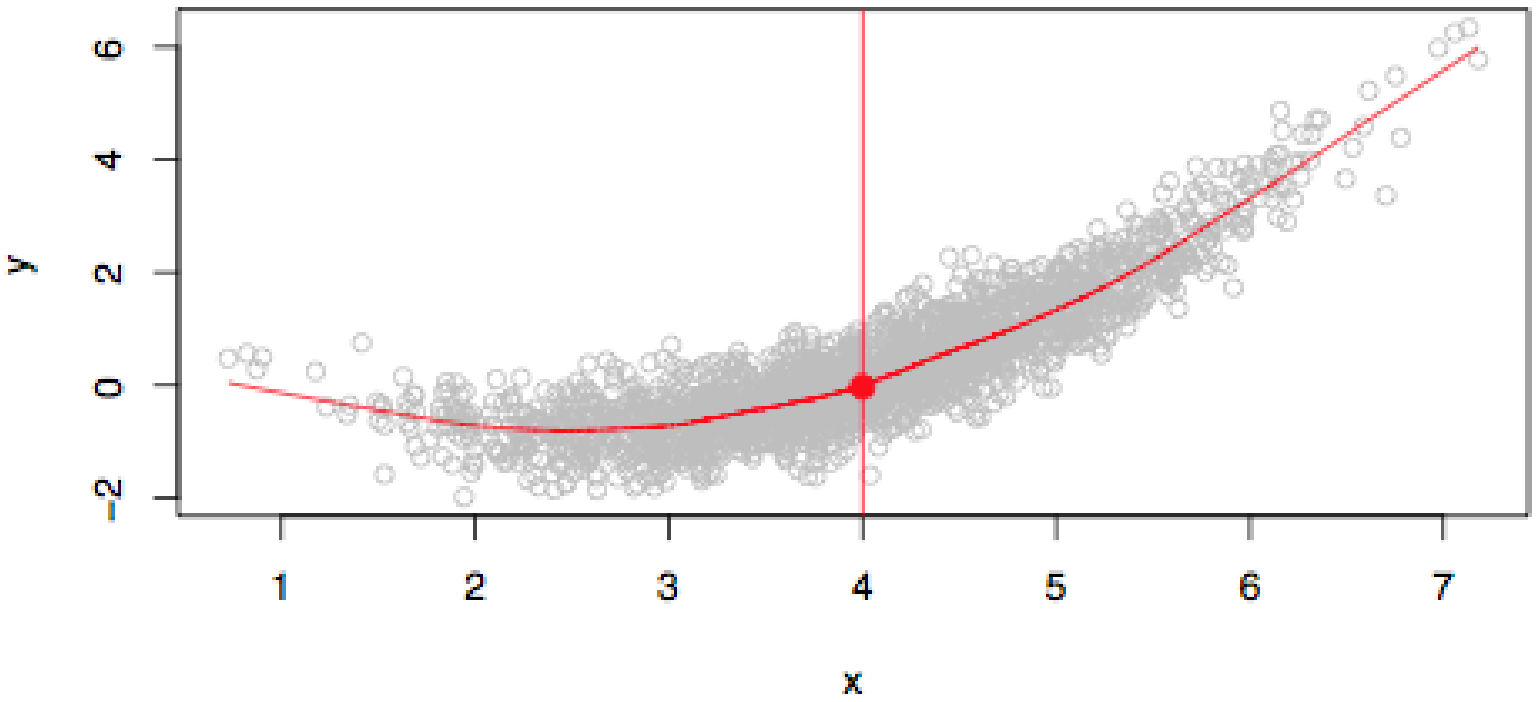
$$y = f(x) + e$$

, where e is an error term.

What is the function f that produces the best prediction of y given x?

- Depends on how we measure "best." Common choice: minimizing the expected squared-error loss $E\left[(y - f(x))^2\right] \Rightarrow f(x) = E[y|x]$.
- $f(x) = E[y|x]$ is the **target function** that we want to learn.

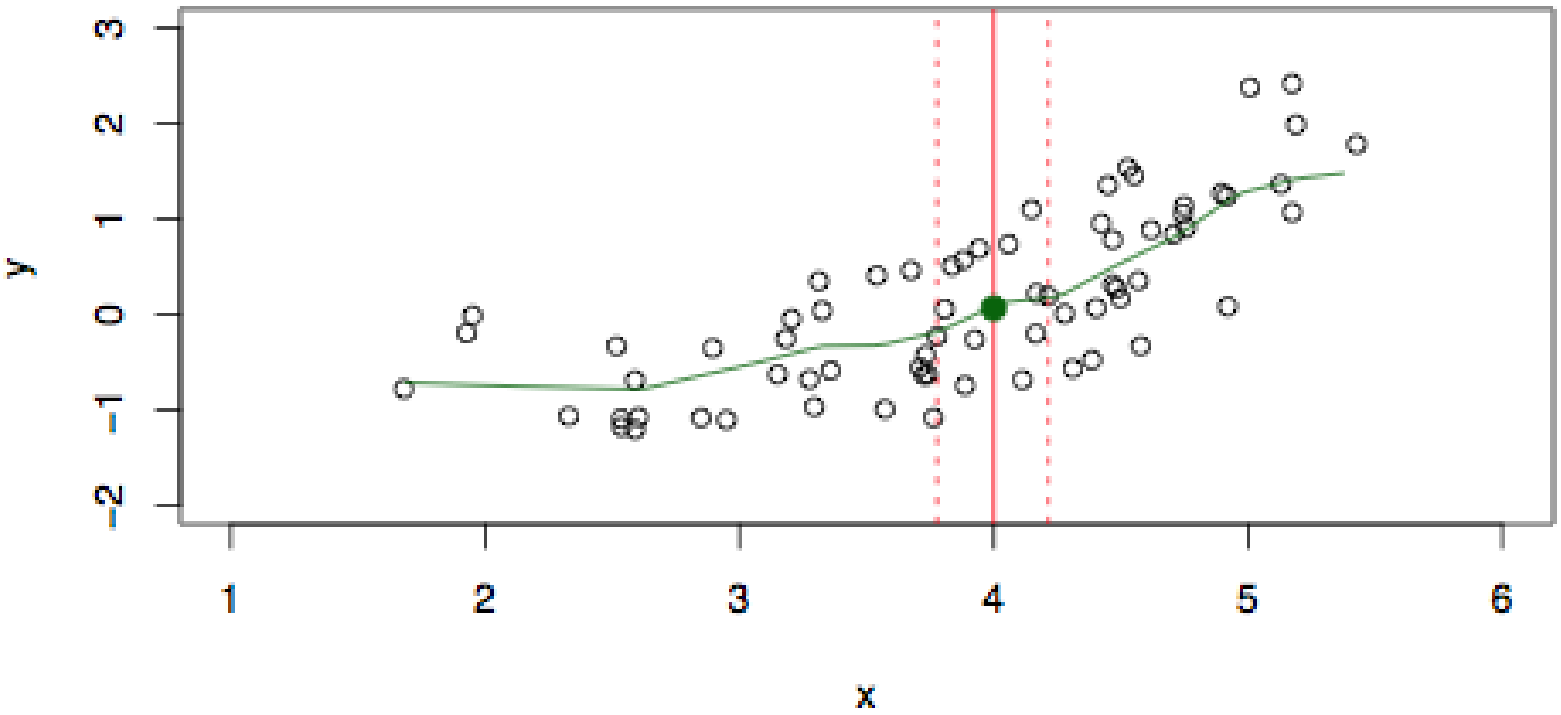
Learning f



$$\widehat{f}\left(x=4\right)=Ave\left(y|x=4\right)$$

Learning f

- Typically we have few if any data points at a specific value of x.
- One solution: relax the set of x over which y is averaged.



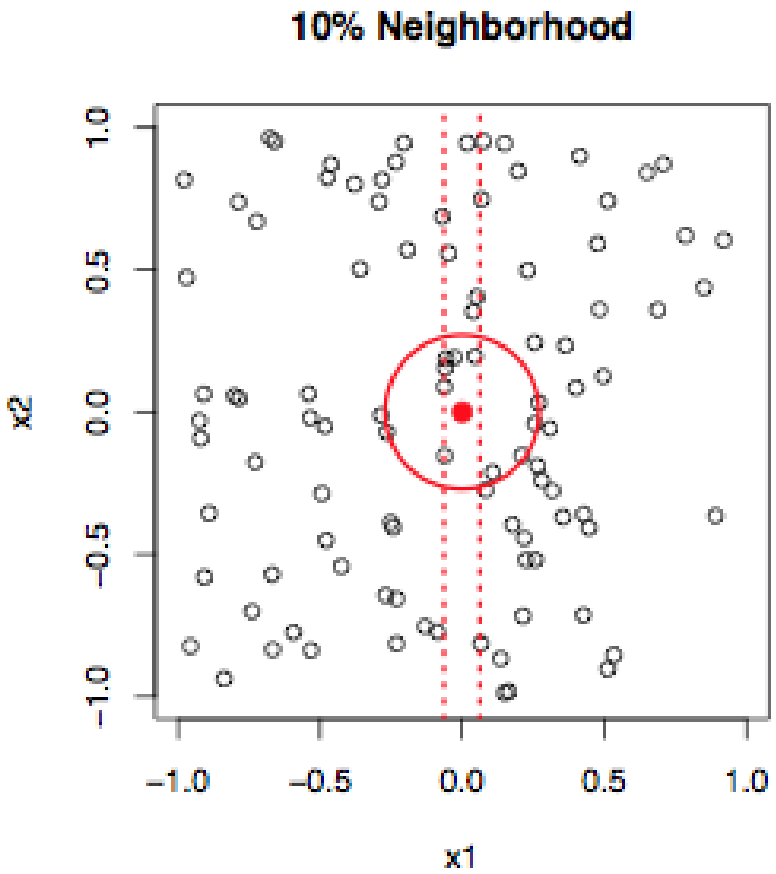
$$\widehat{f}(x=4) = Ave(y|x \in \mathcal{N}(x=4))$$

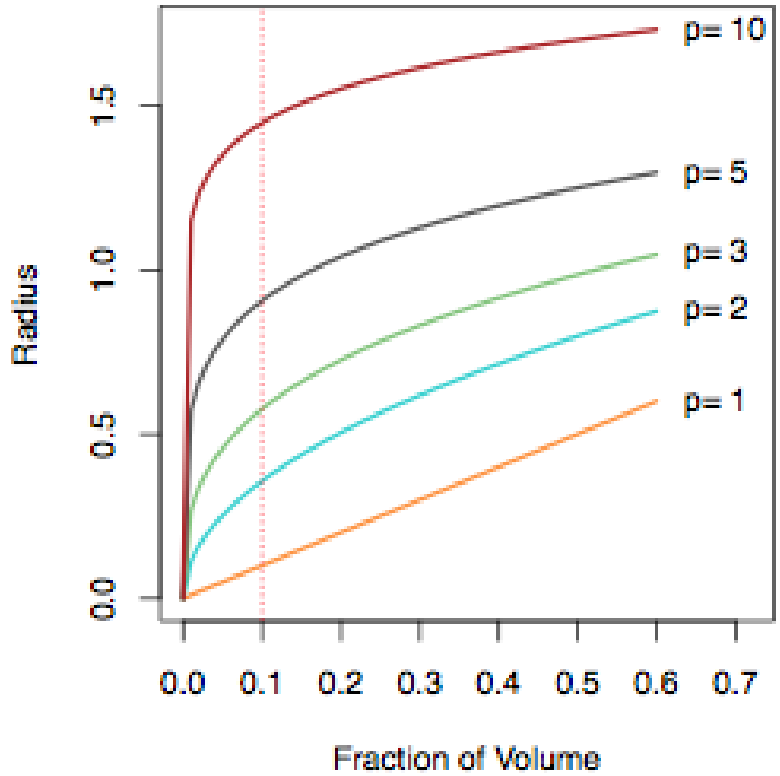
, where $\mathcal{N}(x)$ is some neighborhood of x.

Learning f

- When x is multi-dimensional, i.e. $x = (x_1, \ldots, x_p)$, nearest neighbor averaging can work well for small p and large N.
- Nearest neighbor methods can be lousy when p is large, because neighbors tend to be far away in high dimensions.
  - This is called the **curse of dimensionality.**

Learning f



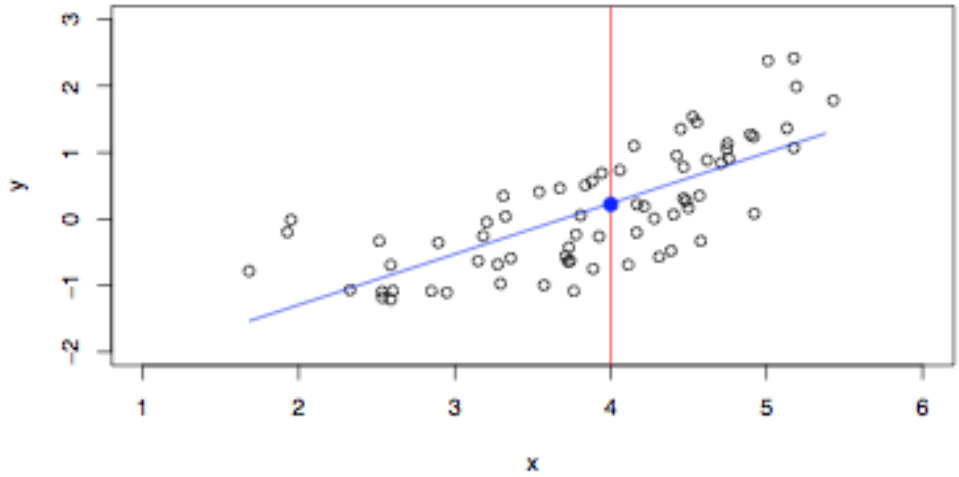


Nearest neighbor and the curse of dimensionality

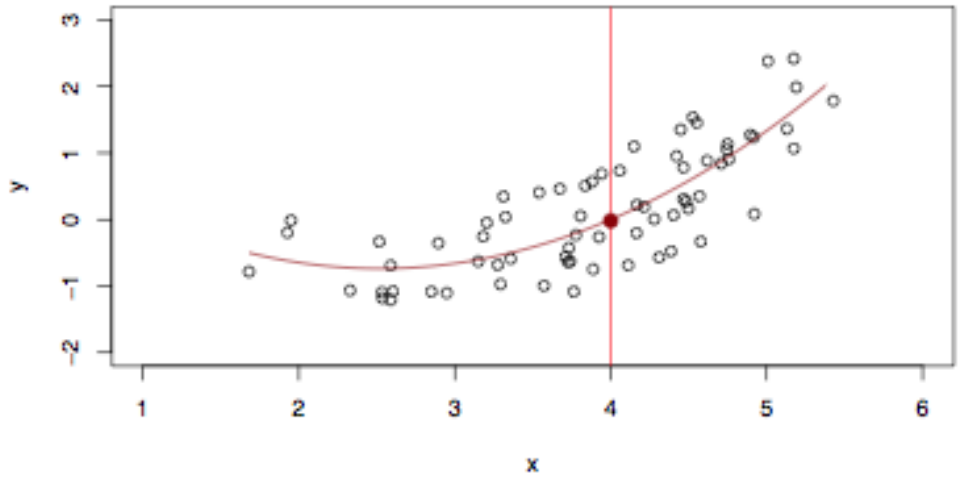
Learning f

- **Parametric methods** of estimating $f(x)$ assume a specific functional form with a fixed number of parameters.
  - Linear regression: $f(x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p = \beta' x$
- **Nonparametric methods** do not make explicit assumptions about the functional form of $f(x)$.
  - Nearest neighbor averaging is a nonparametric method.

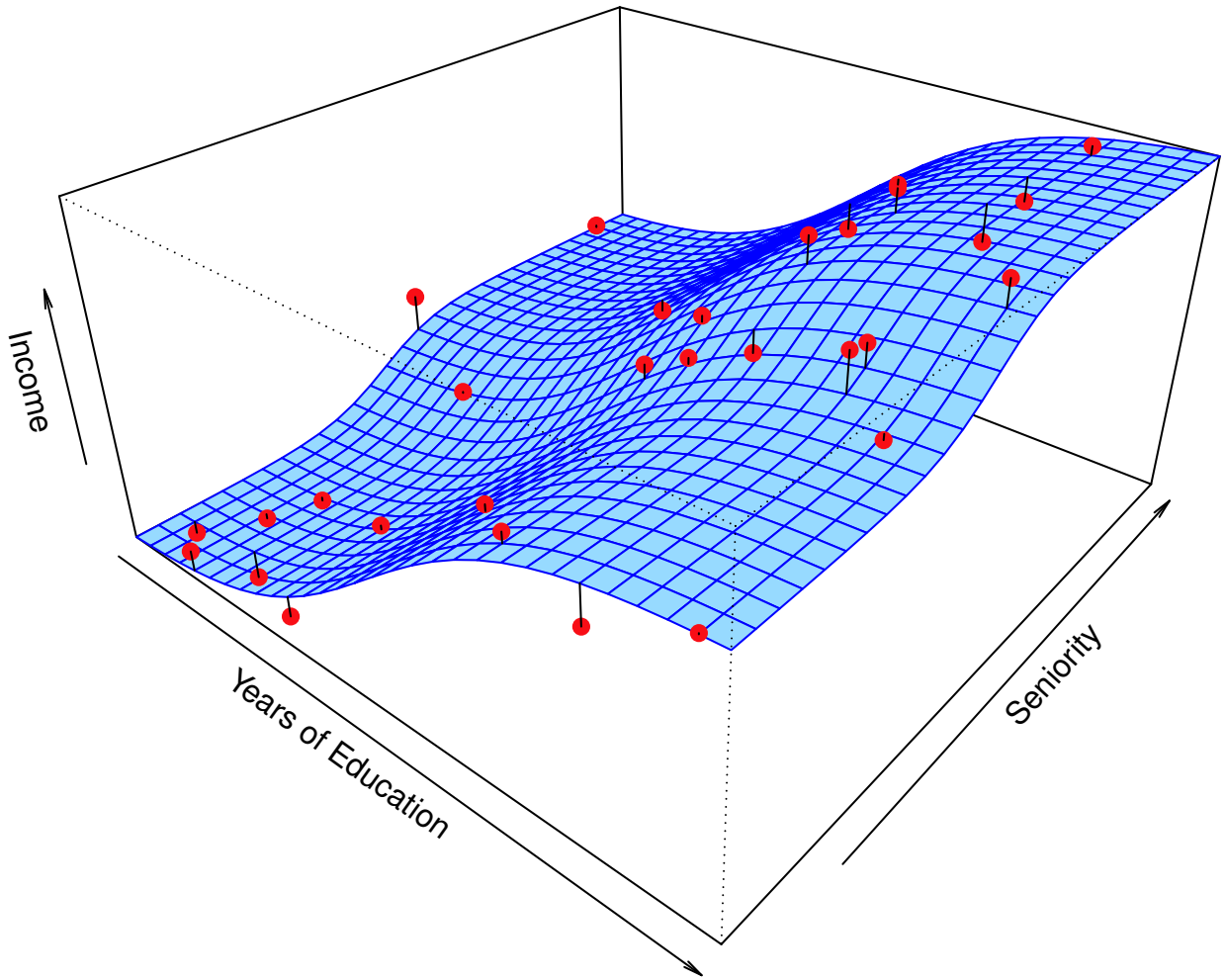
A linear model $\hat{f}_L(X) = \hat{\beta}_0 + \hat{\beta}_1 X$ gives a reasonable fit here



A quadratic model $\hat{f}_Q(X) = \hat{\beta}_0 + \hat{\beta}_1 X + \hat{\beta}_2 X^2$ fits slightly better.

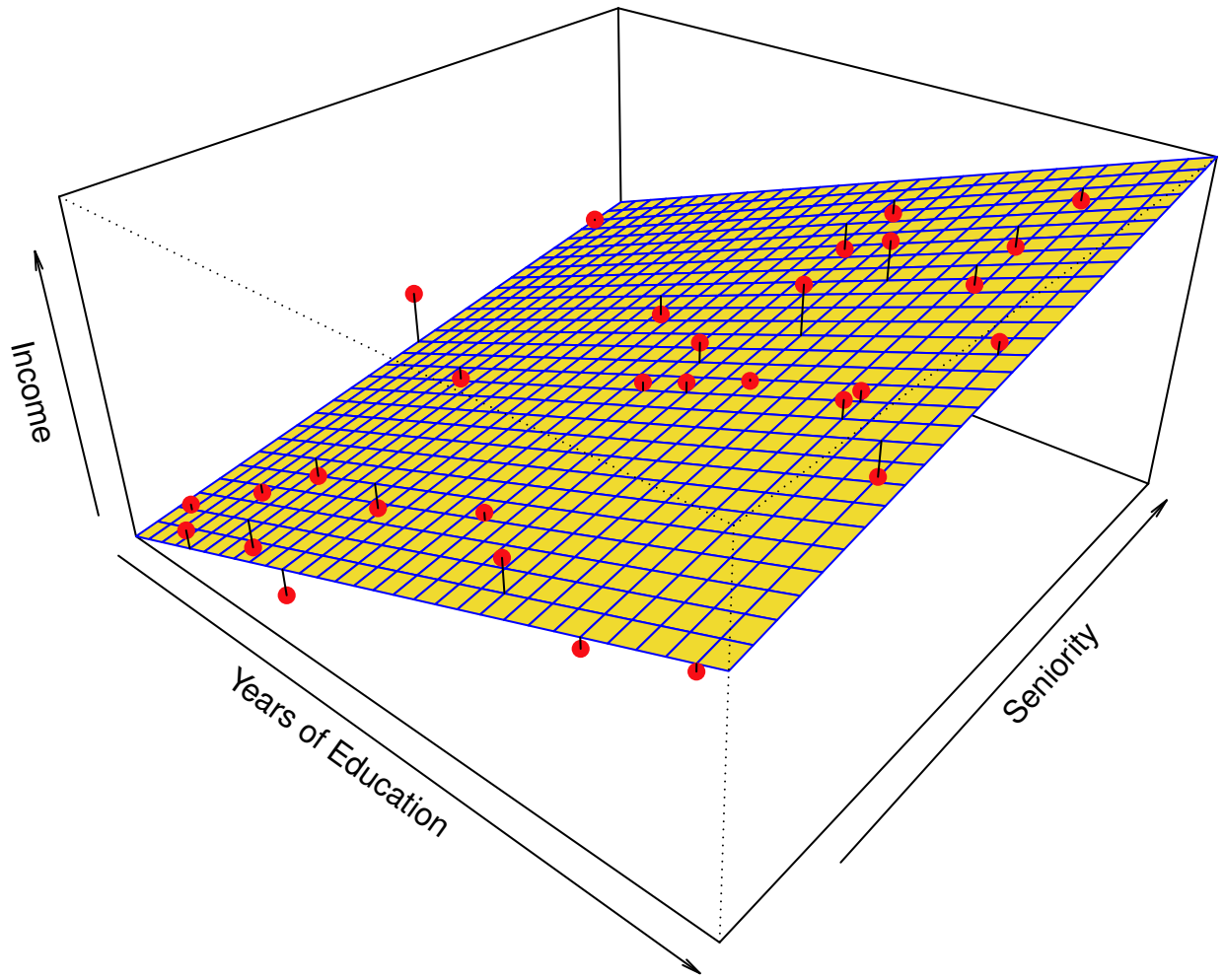


Learning f



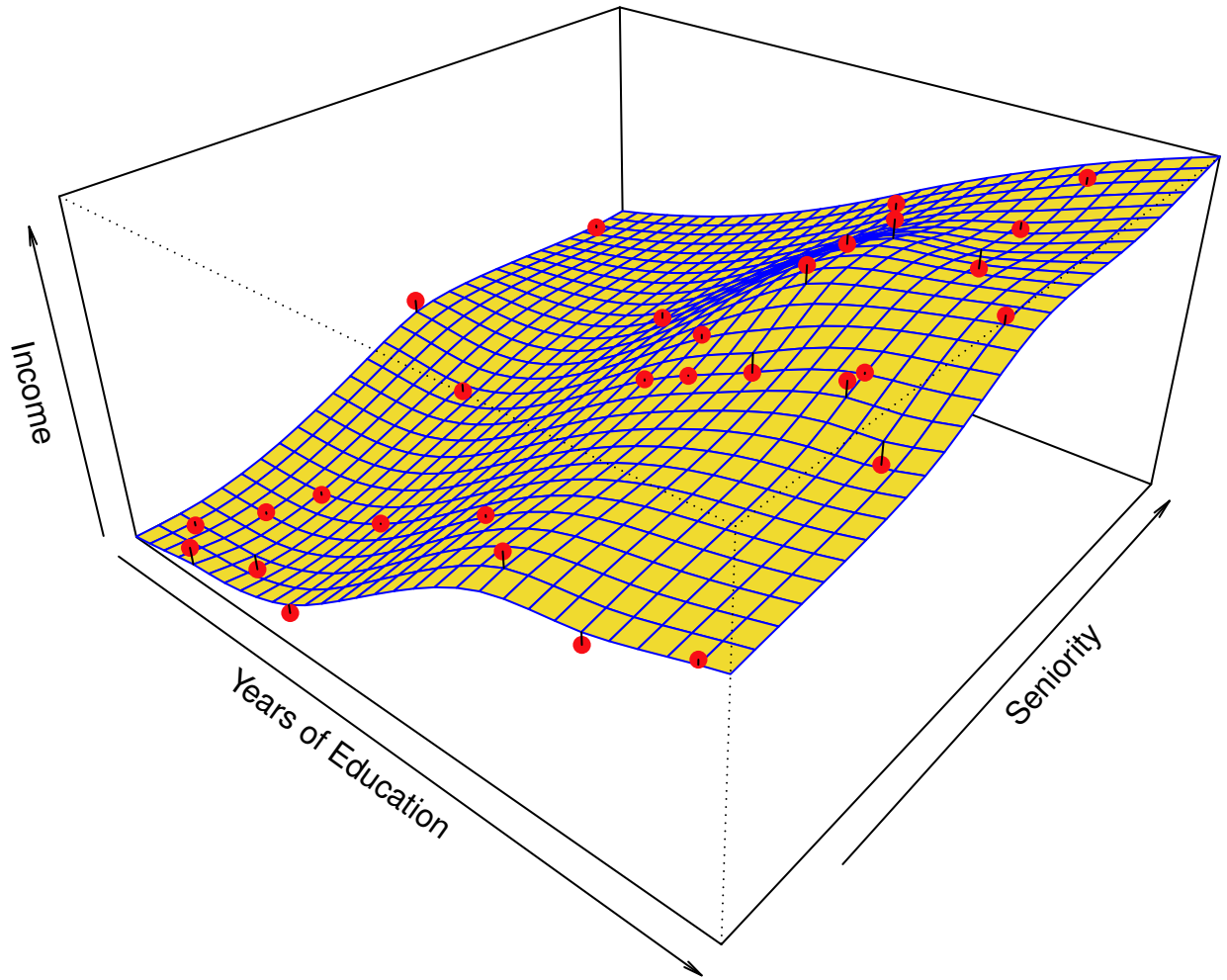
True f

Learning f



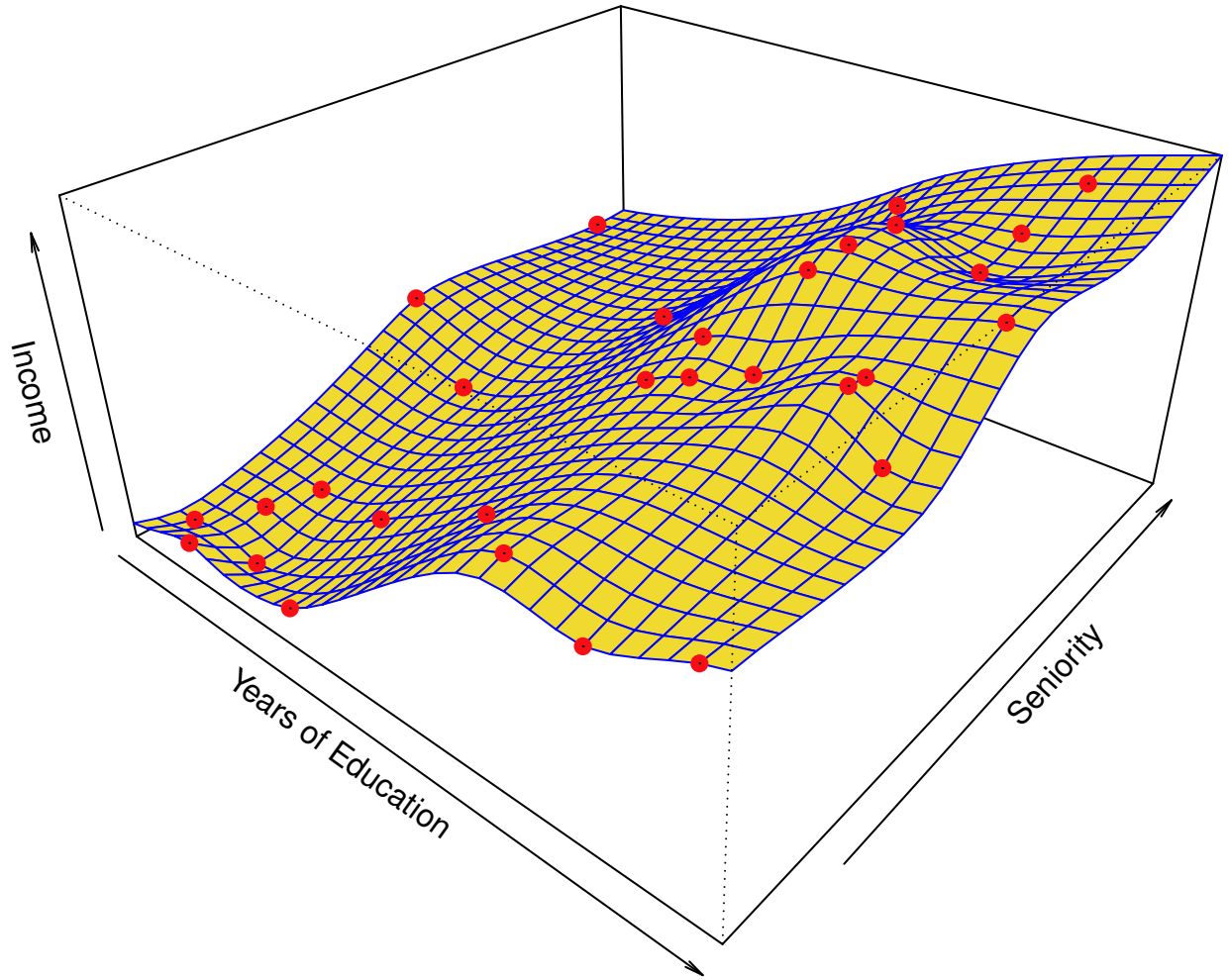
Linear Fit

Learning f



Thin-plate Spline Fit (Smooth)

Learning f



Thin-plate Spline Fit (Rough)

Here $\widehat{f}$ fits the data perfectly: $\widehat{f}(x)$ contains not only $f(x)$ but also e.

Assessing the Goodness of Fit

Let $\mathcal{D}_{TR} = \{(x_1, y_1), \ldots, (x_N, y_N)\}$ denote the data on which we estimate f. This is called **training data**.

We can assess how well $\widehat{f}$ fits the training data by calculating the **training error**:

$$\text{error}_{TR} = \frac{1}{N} \sum_{i \in \mathcal{D}_{TR}} \left(y_i - \widehat{f}(x_i) \right)^2$$

However, what we are really interested in is how well $\widehat{f}$ predicts previously unseen data.

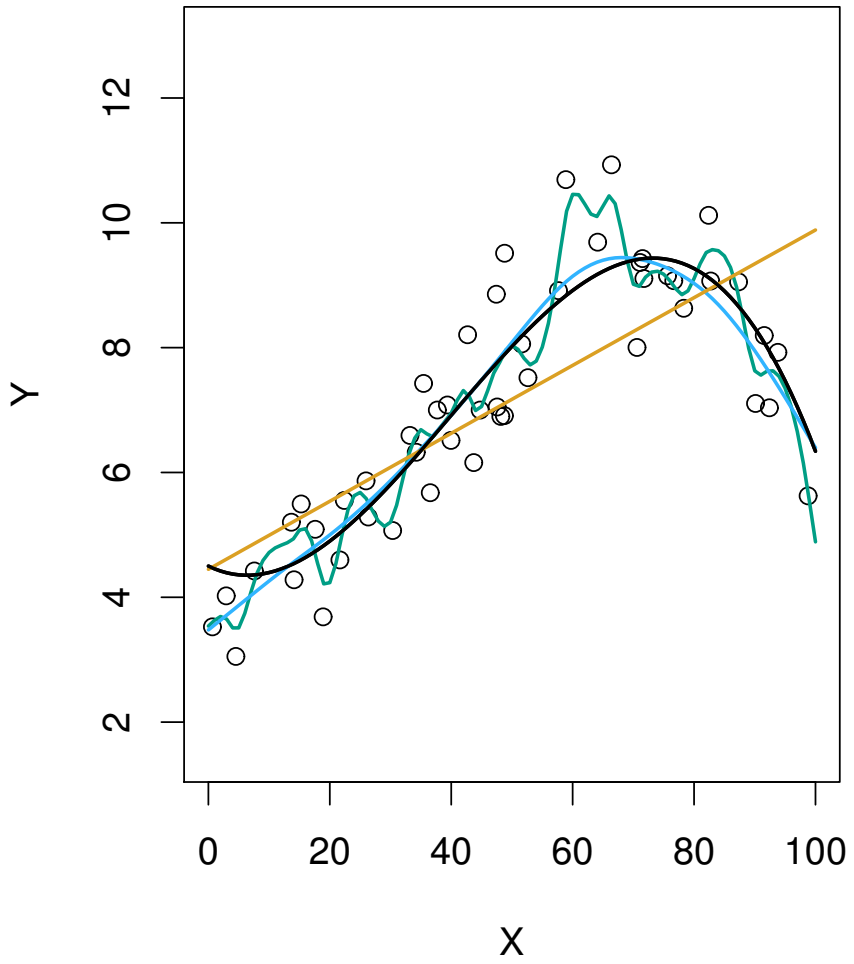
Assessing the Goodness of Fit

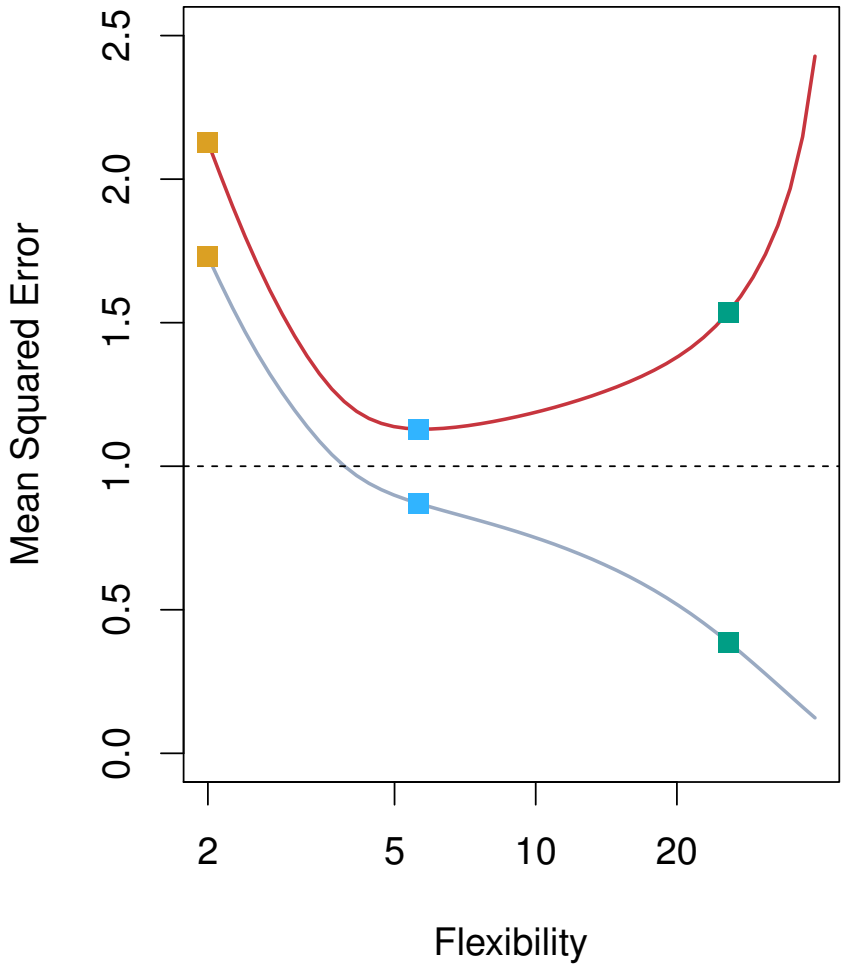
To this end, we can apply $\widehat{f}$ to a set of **test data**, $\mathcal{D}_{TE} = \{(x_1, y_1), \ldots, (x_M, y_M)\}$, and calculate the **test error**:

$$\text{error}_{TE} = \frac{1}{M} \sum_{i \in \mathcal{D}_{TE}} \left(y_i - \widehat{f}(x_i) \right)^2$$

When $M \to \infty$, $\text{error}_{TE} \to \underbrace{E\left[\left(y - \widehat{f}(x) \right)^2 \right]}_{\textbf{prediction error (true error)}}$.

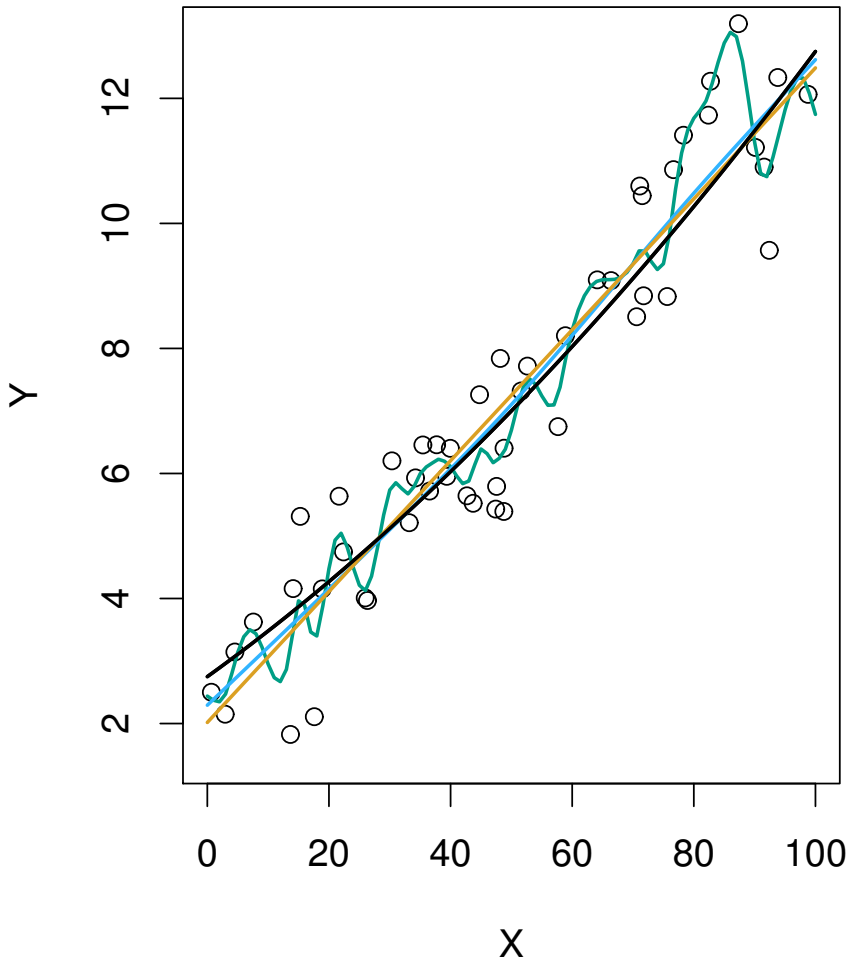
Assessing the Goodness of Fit

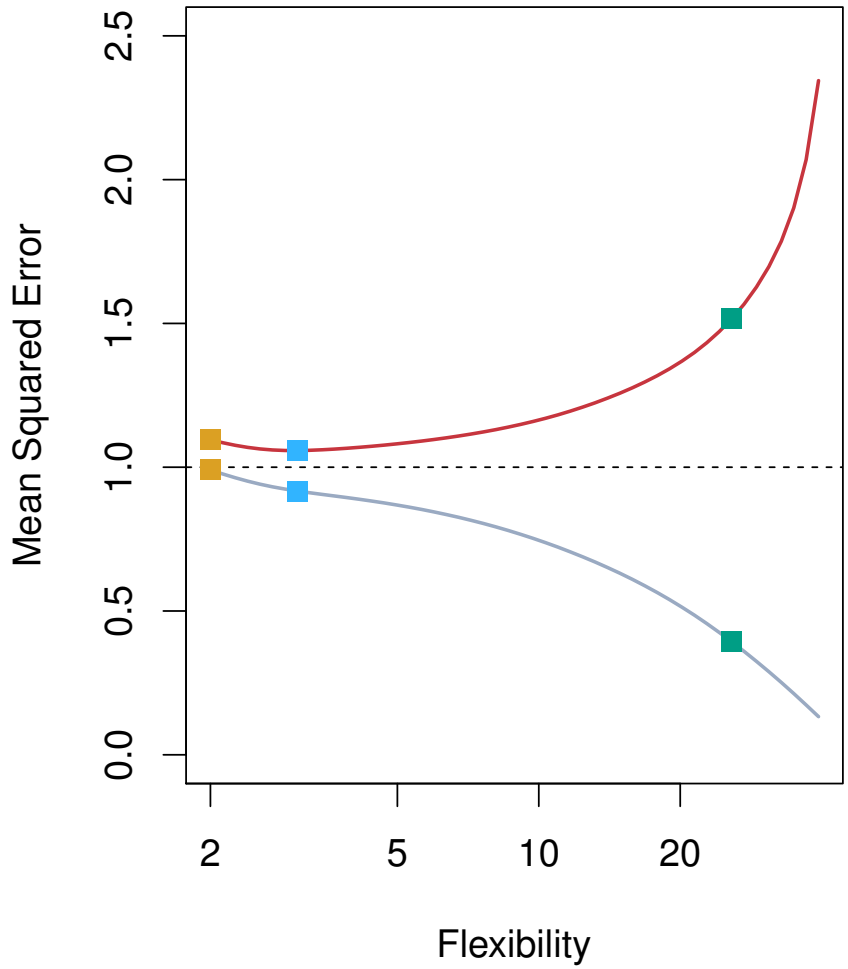




Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), Var(e) (dashed).

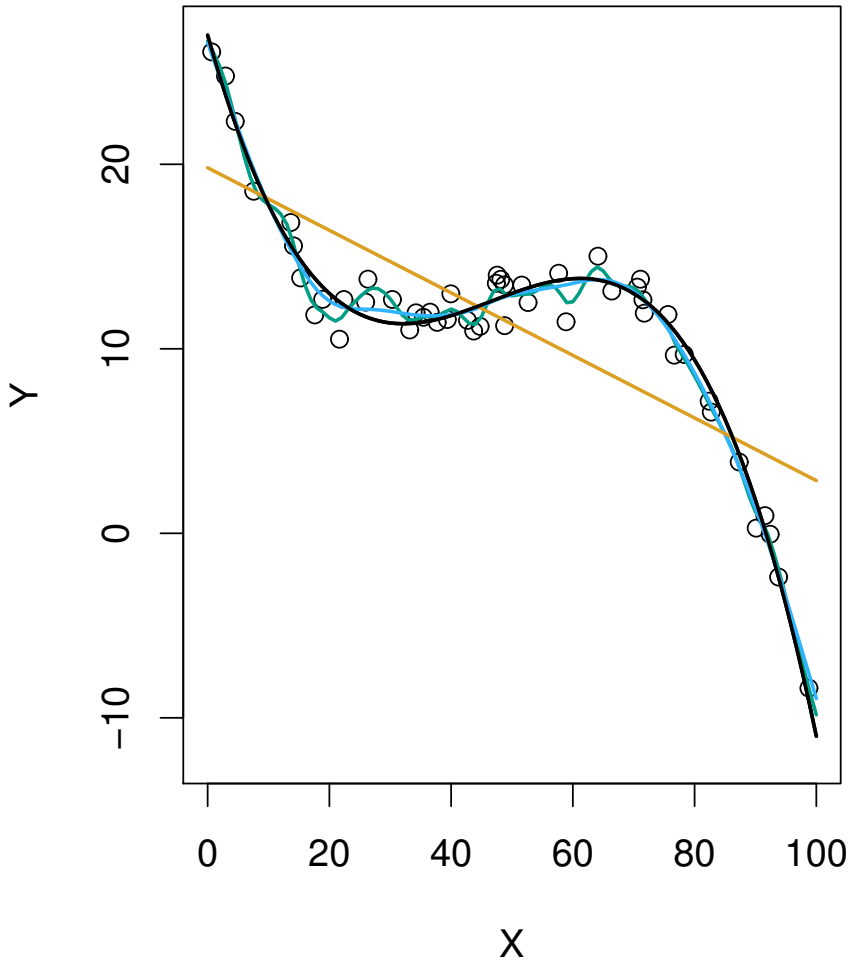
Assessing the Goodness of Fit

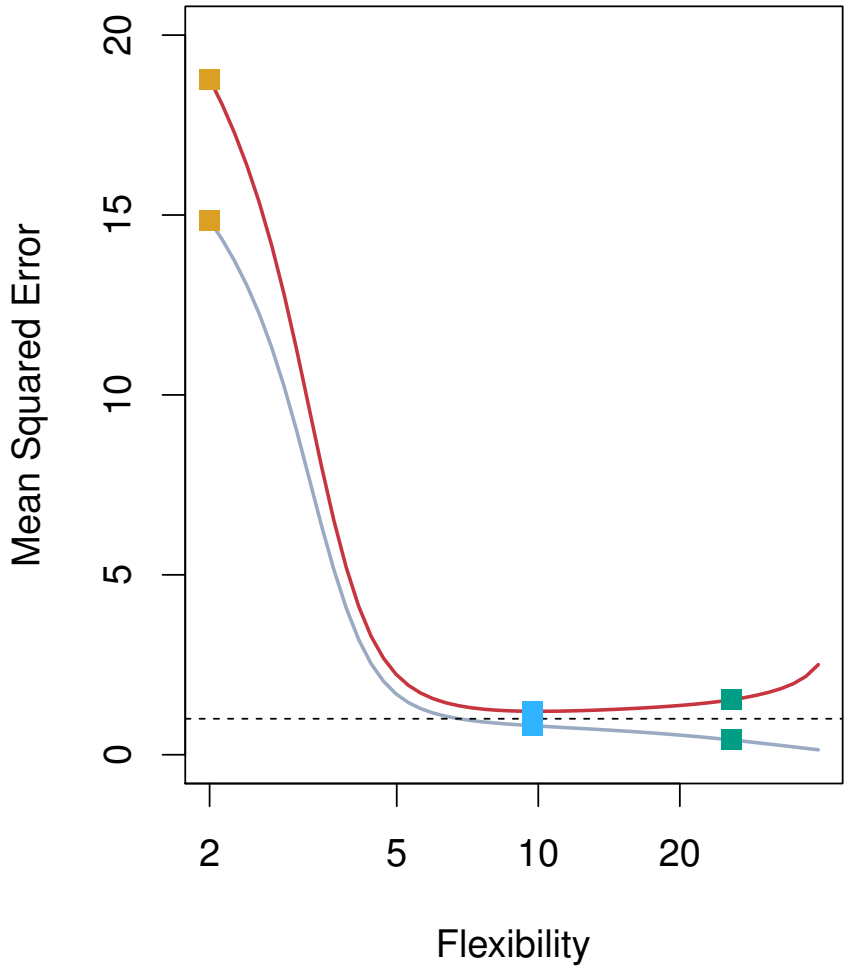




Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), Var(e) (dashed).

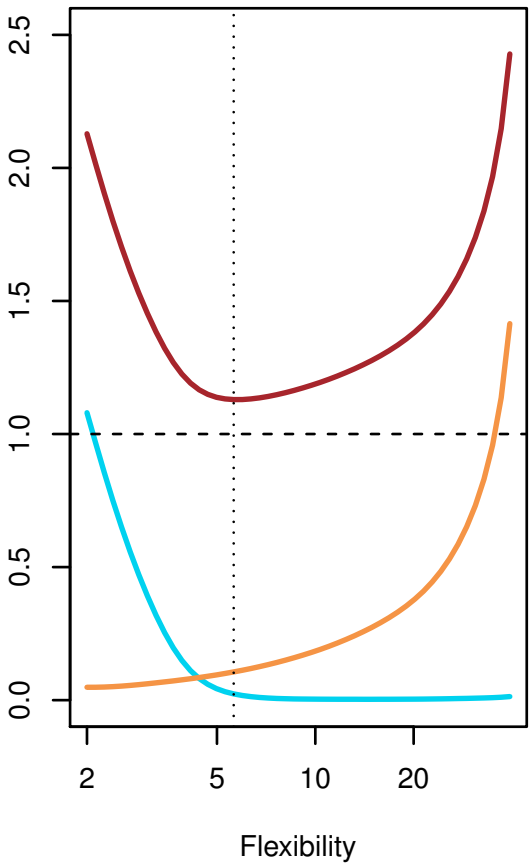
Assessing the Goodness of Fit

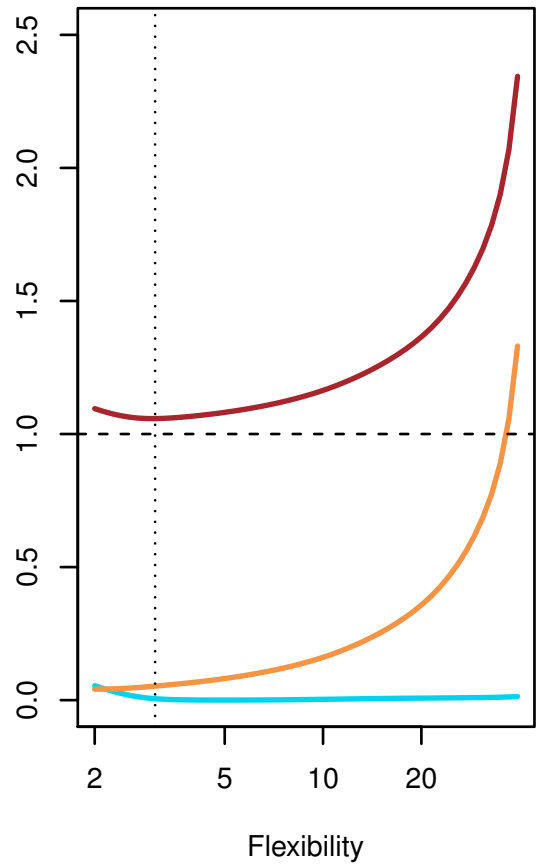


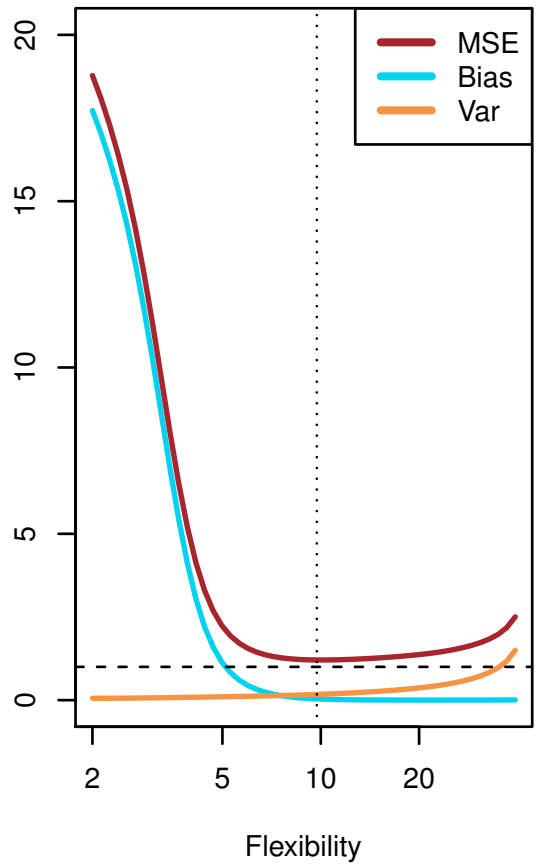


Left: true f (black), linear fit (orange), smoothing spline fits (blue & green).
Right: training error (grey), prediction error (red), Var(e) (dashed).

Assessing the Goodness of Fit







Bias-variance trade-off for the three examples

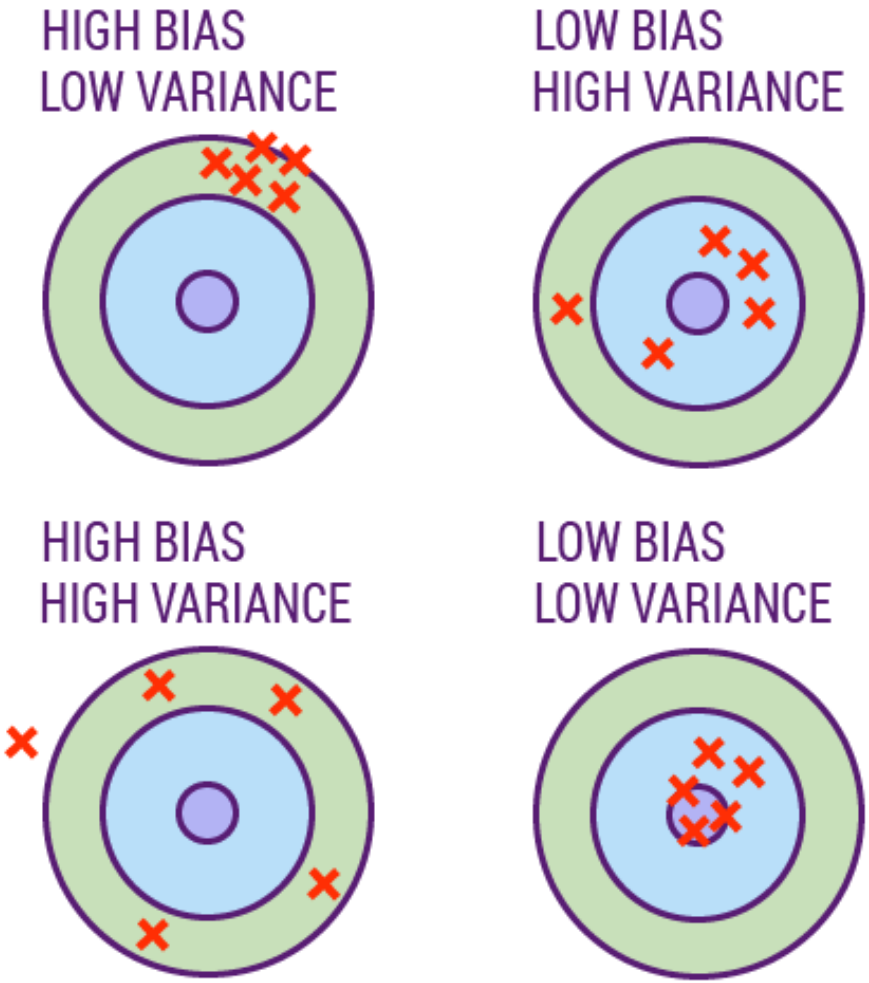
The Bias-Variance Trade-off

When $f(x) = E(y|x)$, for any estimate $\widehat{f}$ of f and $\widehat{y} = \widehat{f}(x)$,

$$\begin{aligned} E\left[(y-\widehat{y})^2\right] &= E\left[\left(f(x)-\widehat{f}(x)\right)^2\right] + Var(e) \\ &= Var\left(\widehat{f}(x)\right) + \left[\text{bias}\left(\widehat{f}(x)\right)\right]^2 + \underbrace{Var(e)}_{\text{Irreducible}} \end{aligned}$$

, where $\text{bias}\left(\widehat{f}(x)\right) \equiv E\left[f(x)-\widehat{f}(x)\right]$.

The Bias-Variance Trade-off



The Bias-Variance Trade-off

- $Var\left(\widehat{f}\right)$ refers to the amount by which $\widehat{f}$ would change if we estimate it using a different training data set.
- As a general rule, as model flexibility increases, bias $\left(\widehat{f}\right)$ will decrease and $Var\left(\widehat{f}\right)$ will increase.
  - More flexible models tend to have higher variance because they have the capacity to follow the data more closely. Thus changing any of the data points may cause the estimate $\widehat{f}$ to change considerably.

The Bias-Variance Trade-off

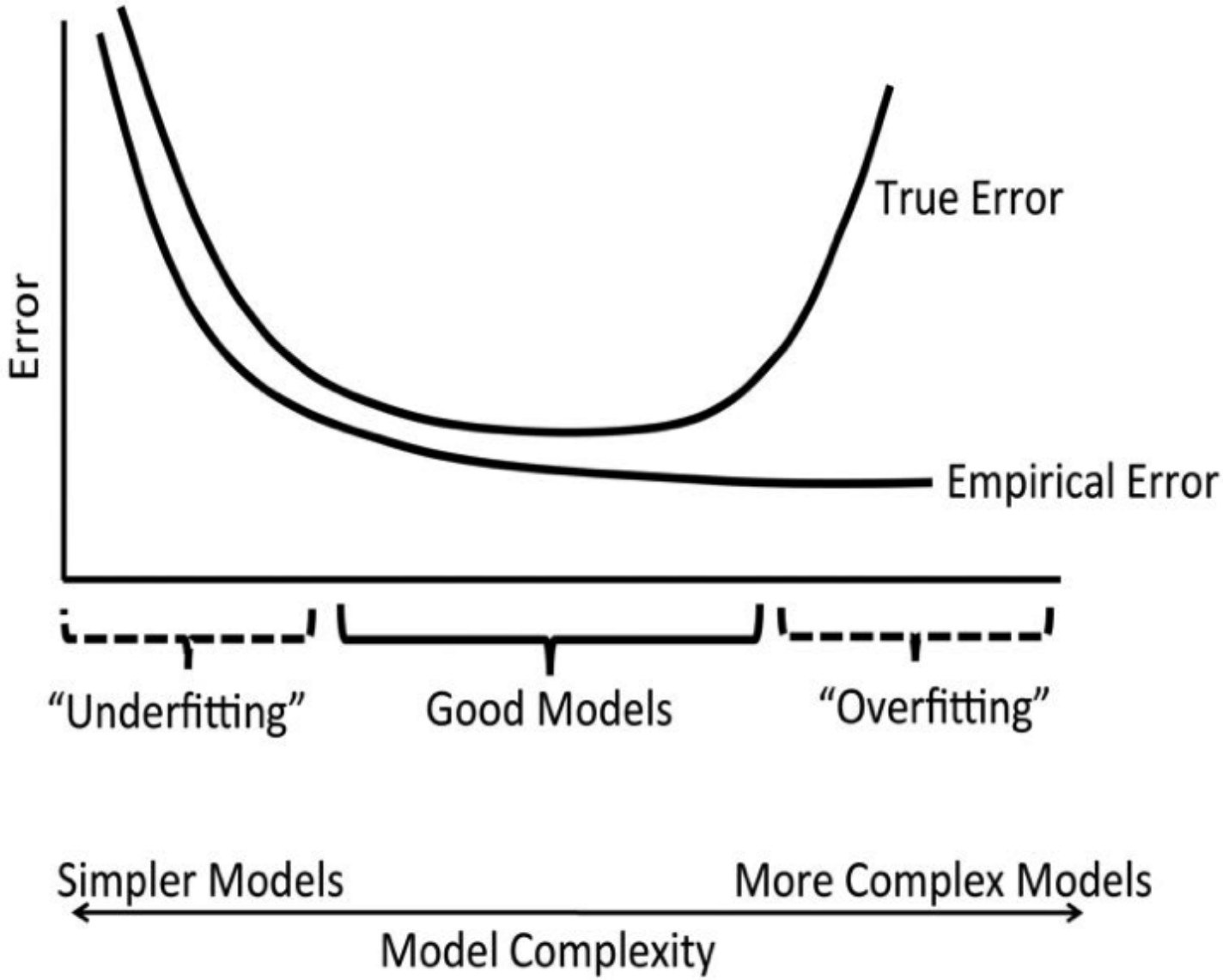
- As the flexibility of the model increases, we observe a monotone decrease in training error and a U-shape in prediction error.
- This is due to the **bias-variance trade-off**: as model flexibility increases, the bias tends to initially decrease faster than the variance increases. Then at some point increasing flexibility has little impact on the bias but starts to significantly increase the variance.

The Bias-Variance Trade-off

- The bias-variance trade-off is a trade-off because it is easy to have a model with extremely low bias but high variance (e.g., by drawing a curve that passes through every single training observation) or one with very low variance but high bias (e.g., by fitting a horizontal line to the data). The challenge lies in finding a model for which both the variance and the bias are low.

- **Overfitting** refers to the case in which a less flexible model would have yielded a smaller prediction error.

The Bias-Variance Trade-off

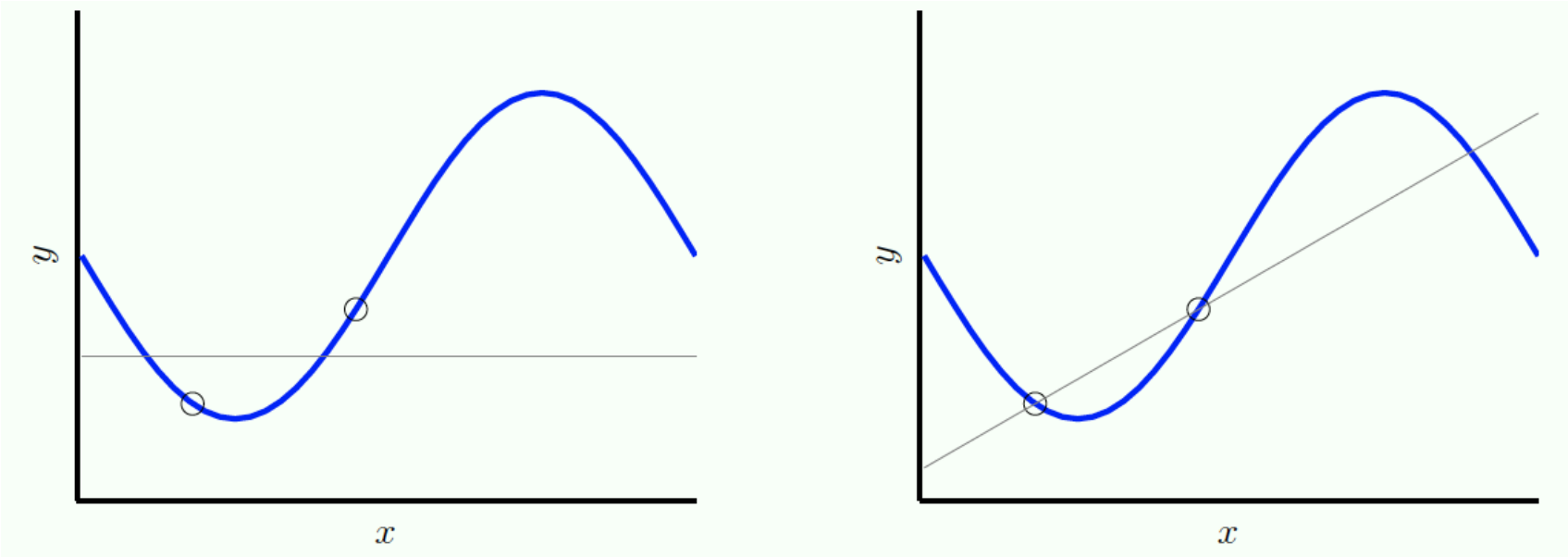


The Bias-Variance Trade-off

$$y = f(x) = \sin(\pi x)$$

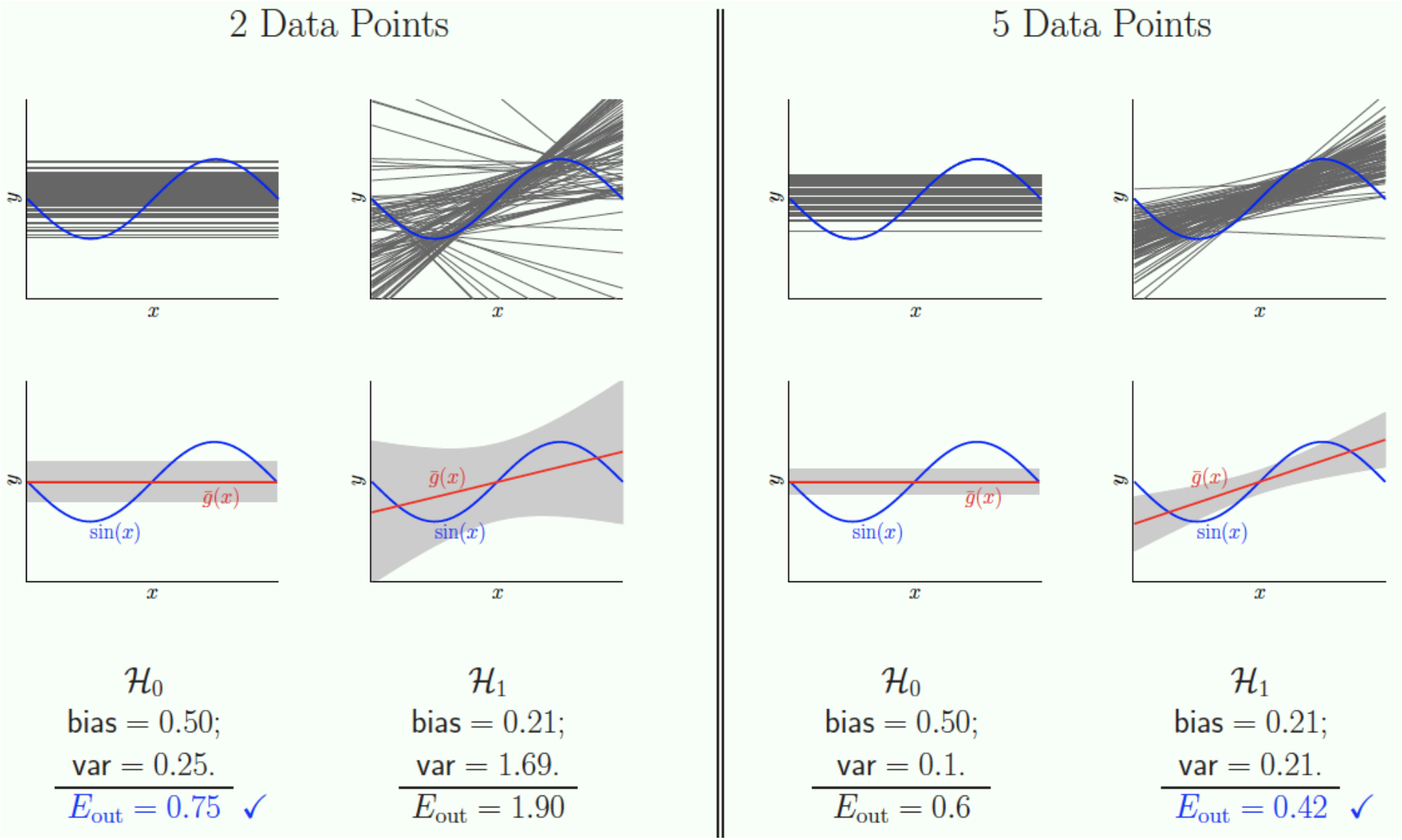
- Two models:

$$\mathcal{H}_0 : h(x) = b$$
$$\mathcal{H}_1 : h(x) = ax + b$$



2 data points

The Bias-Variance Trade-off



Regularization

- To choose the optimal complexity, we can *start with* a complex model and apply **regularization**: methods that constrain the model complexity in order to avoid overfitting and improve prediction error.
- Regularization methods solve the following problem:

$$\min \left\{ \sum_{i=1}^{N} \left(y_i - f\left(x_i \right) \right)^2 \right\} \tag{1}$$

$$\text{subject to} \quad \Omega\left(f \right) \leq C$$

, where $\Omega\left(f \right)$ is a measure of the complexity of f and is called a **regularizer**.

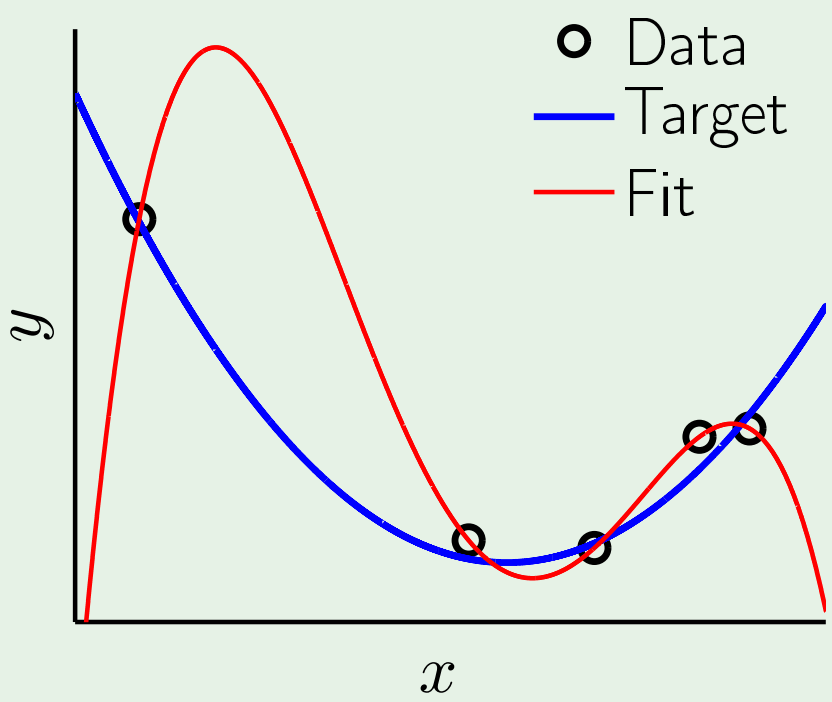
Regularization

- Equivalently, (1) can be written as

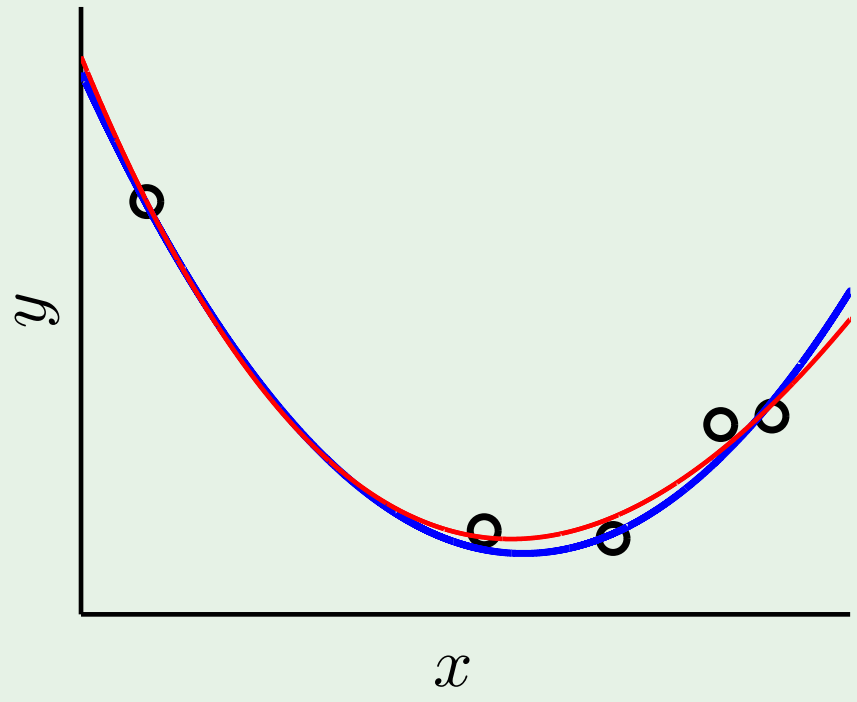
$$\min \left\{ \sum_{i=1}^{N} \left(y_i - f\left(x_i \right) \right)^2 + \lambda \Omega \left(f \right) \right\} \tag{2}$$

- (1) is called the **Ivanov form** and expresses regularization as *constrained minimization*.
- (2) is called the **Tikhonov form** and expresses regularization as *penalized minimization*.
- λ is called a **hyper-parameter** and controls the degree of regularization. To choose the optimal model complexity is to choose the optimal λ.

Regularization

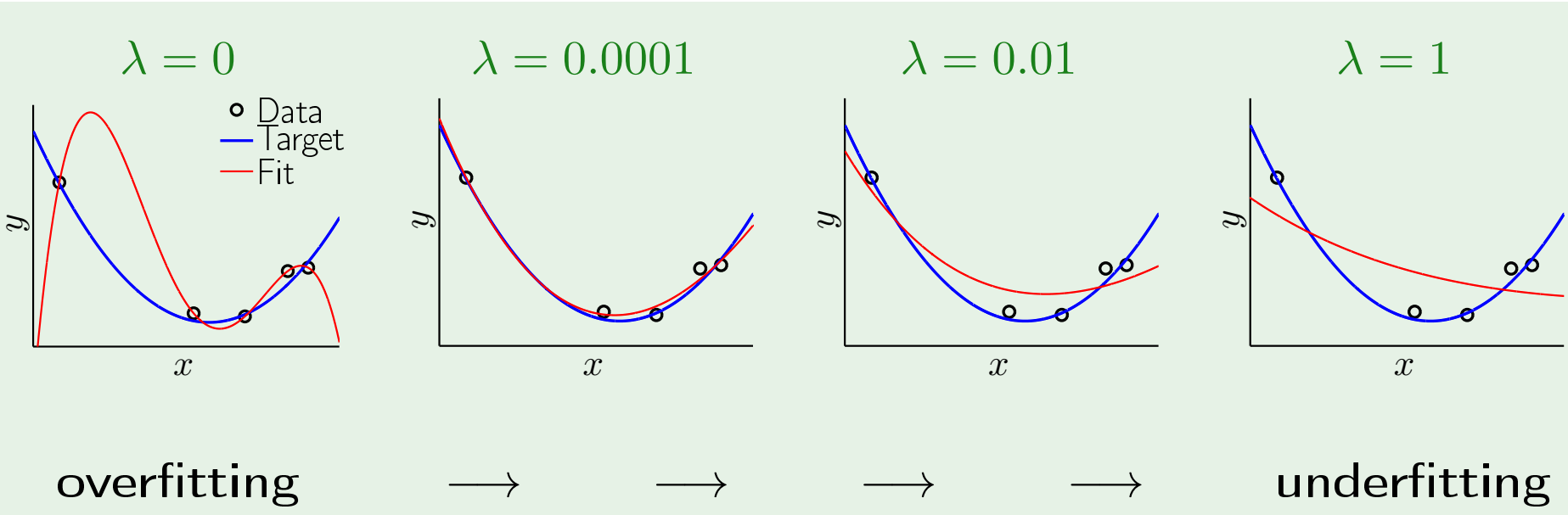


free fit



restrained fit

Regularization



Regularization

$f :[-1, 1]\to \mathbb{R}$ $\quad f(x) = \sin(\pi x)$

Only two training examples! $\quad N = 2$

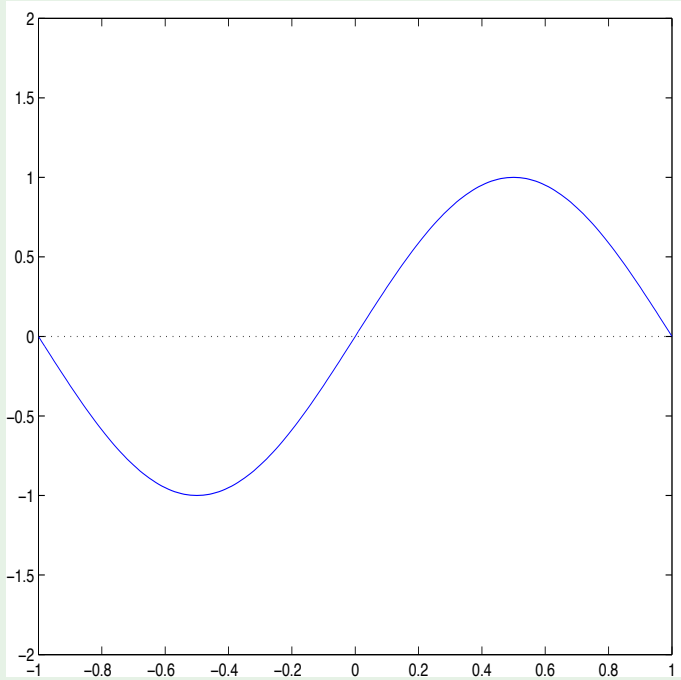
Two models used for learning:

$\mathcal{H}_0$: $\quad h(x) = b$

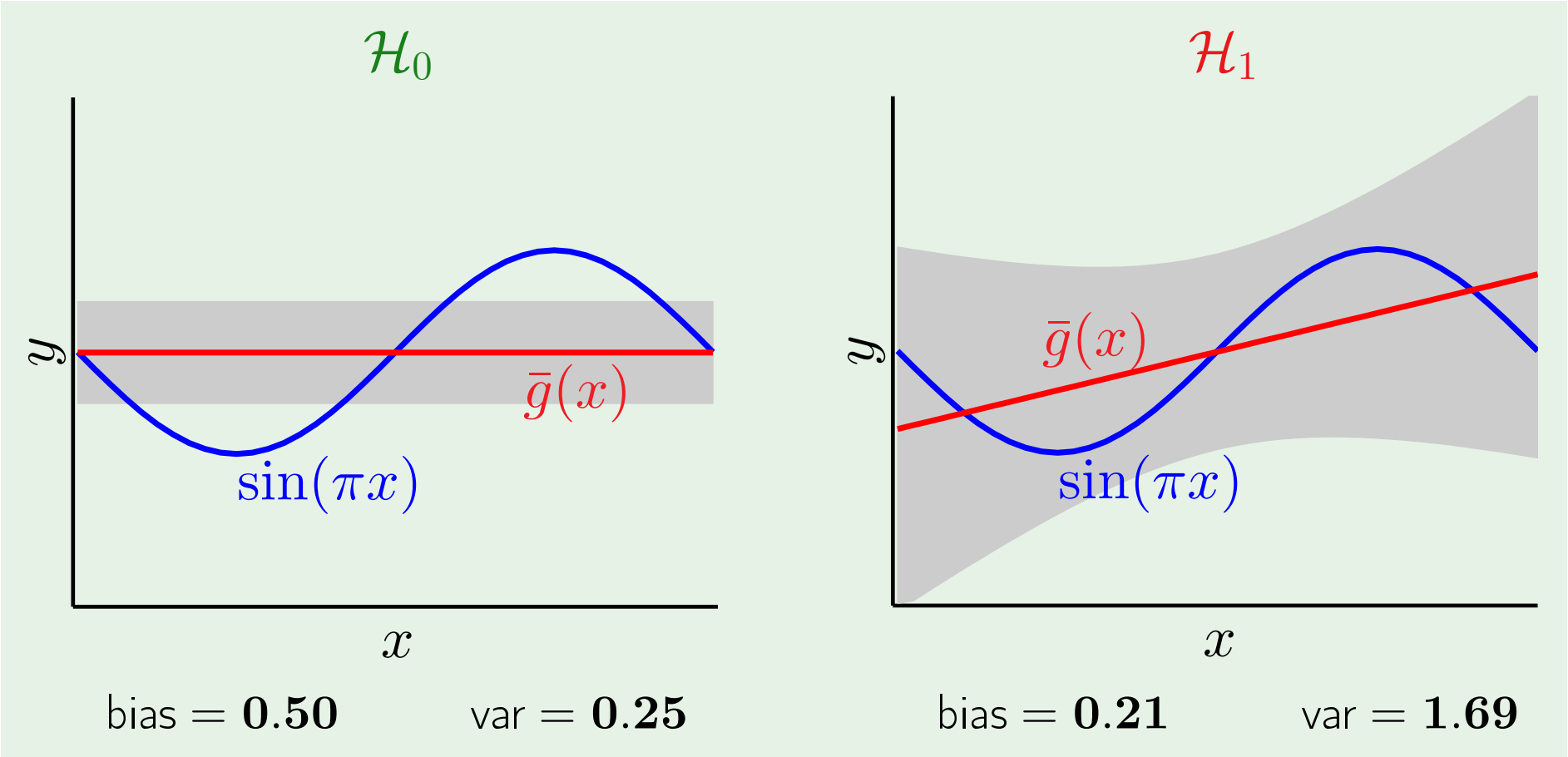
$\mathcal{H}_1$: $\quad h(x) = ax + b$

Which is better, $\mathcal{H}_0$ or $\mathcal{H}_1$?

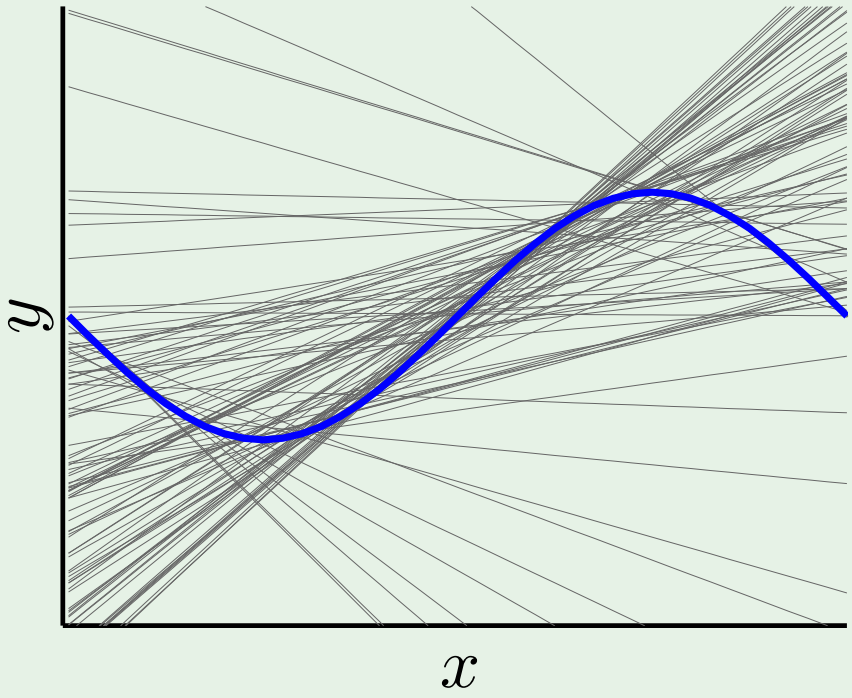
f



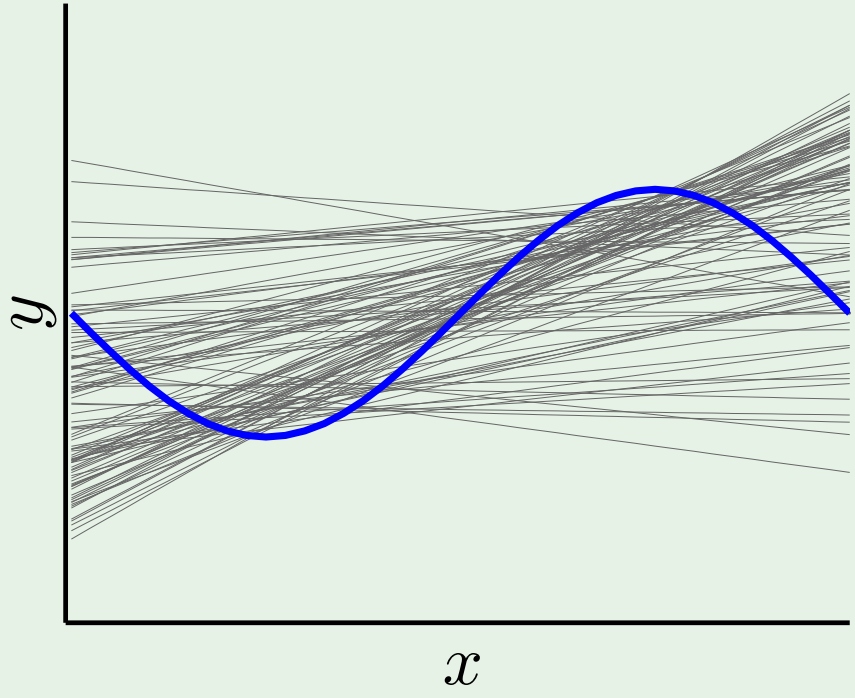
Regularization



Regularization



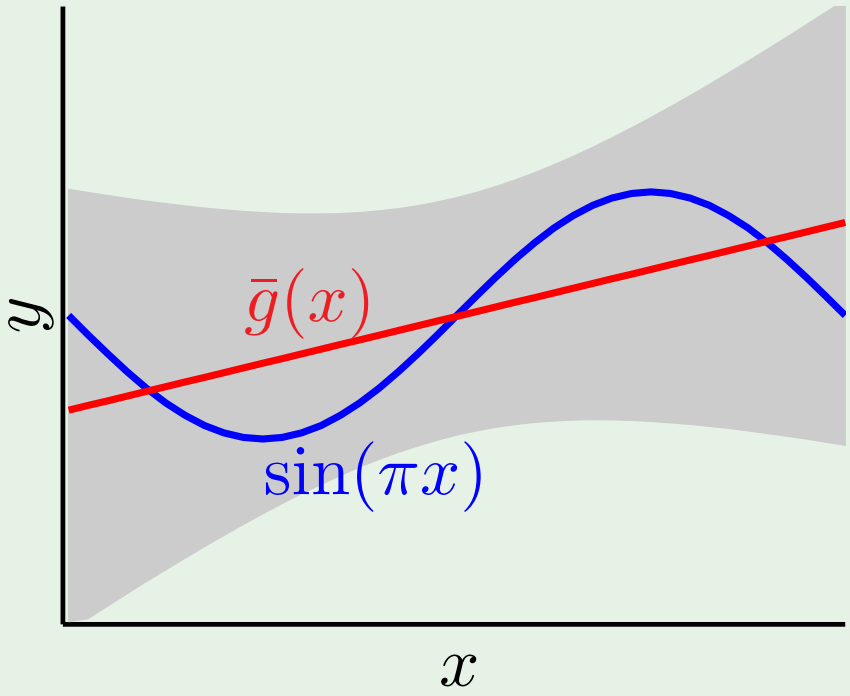
without regularization



with regularization

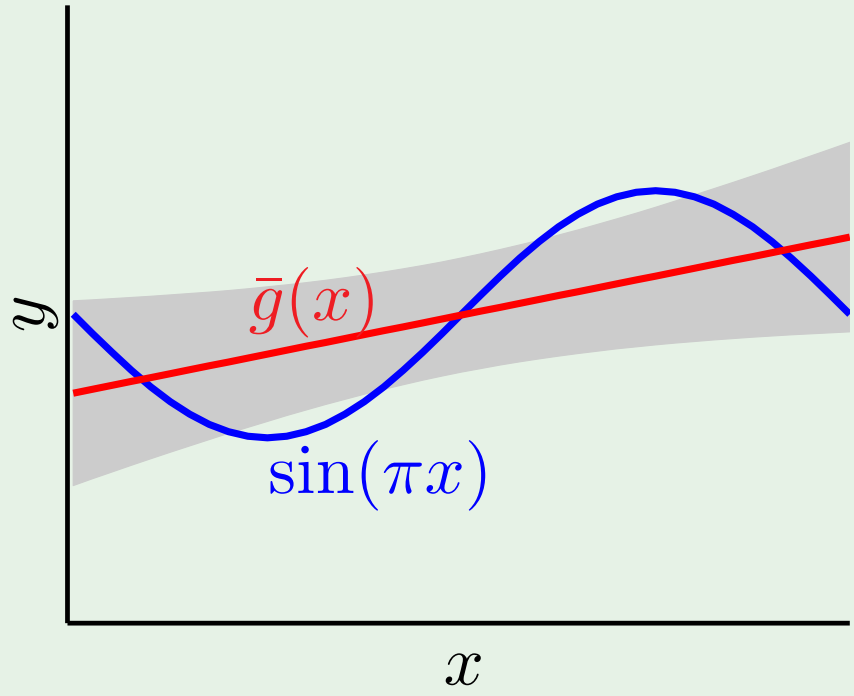
Regularization

without regularization



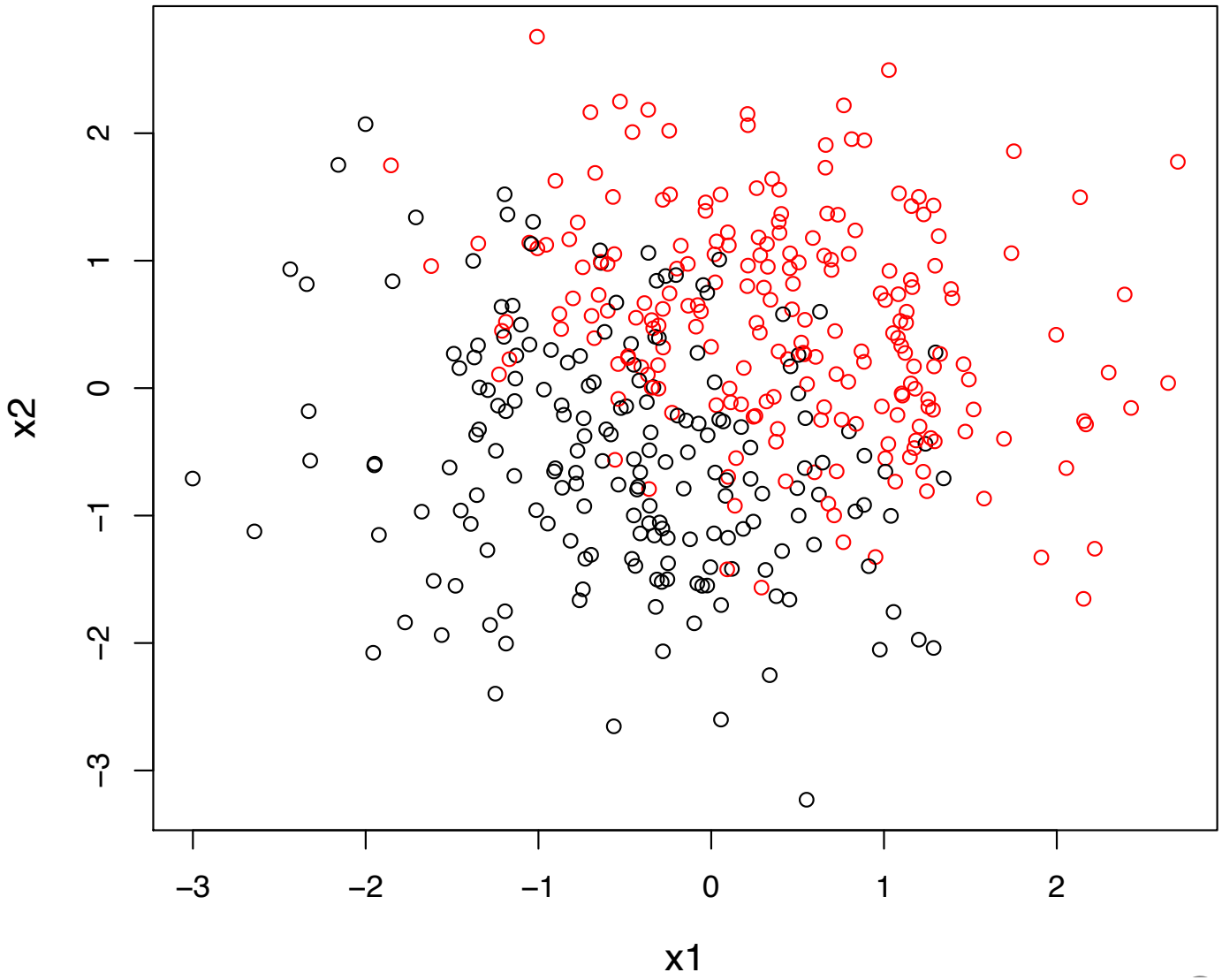
bias $= \mathbf{0.21}$ var $= \mathbf{1.69}$

with regularization

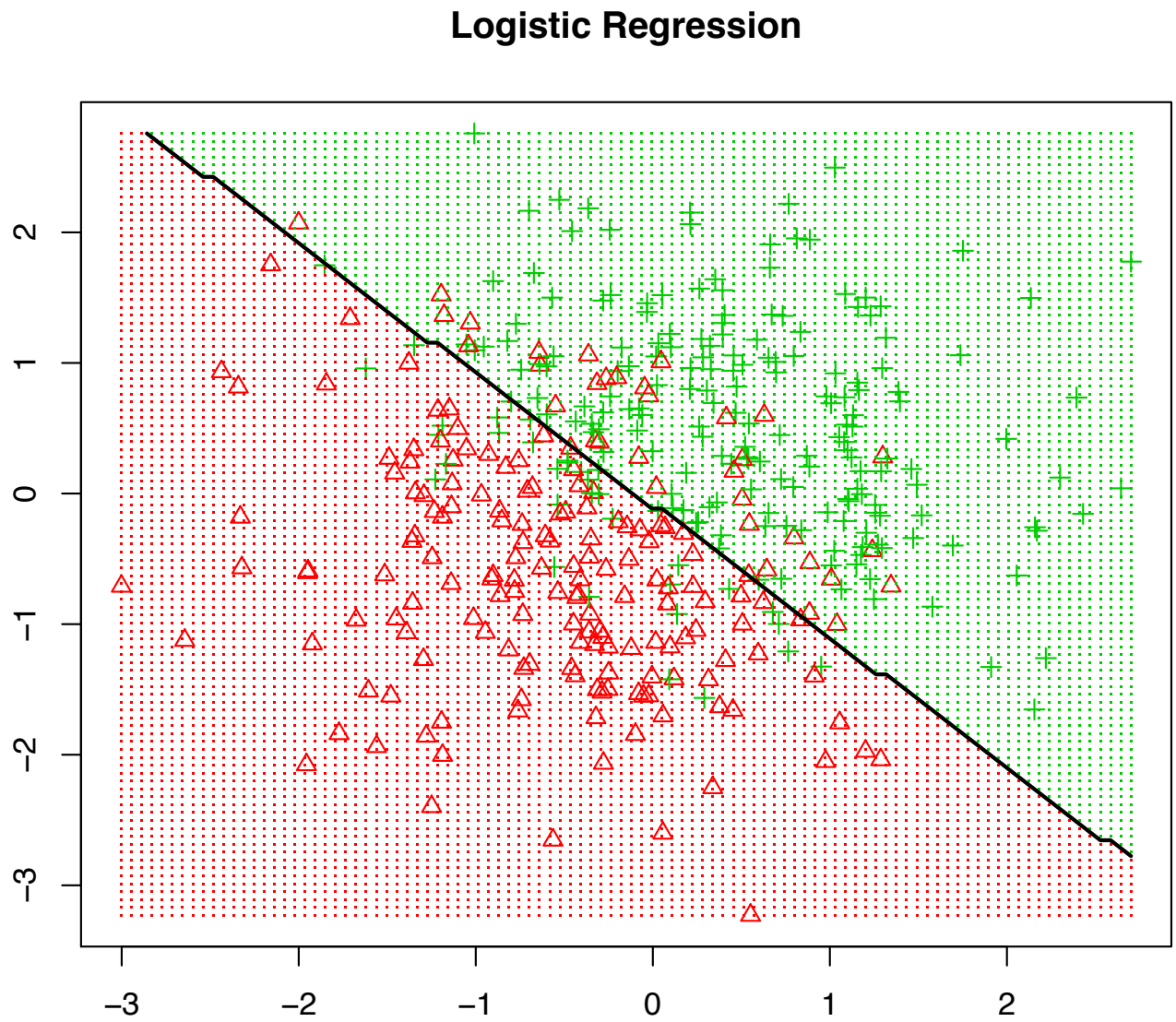


bias $= \mathbf{0.23}$ var $= \mathbf{0.33}$

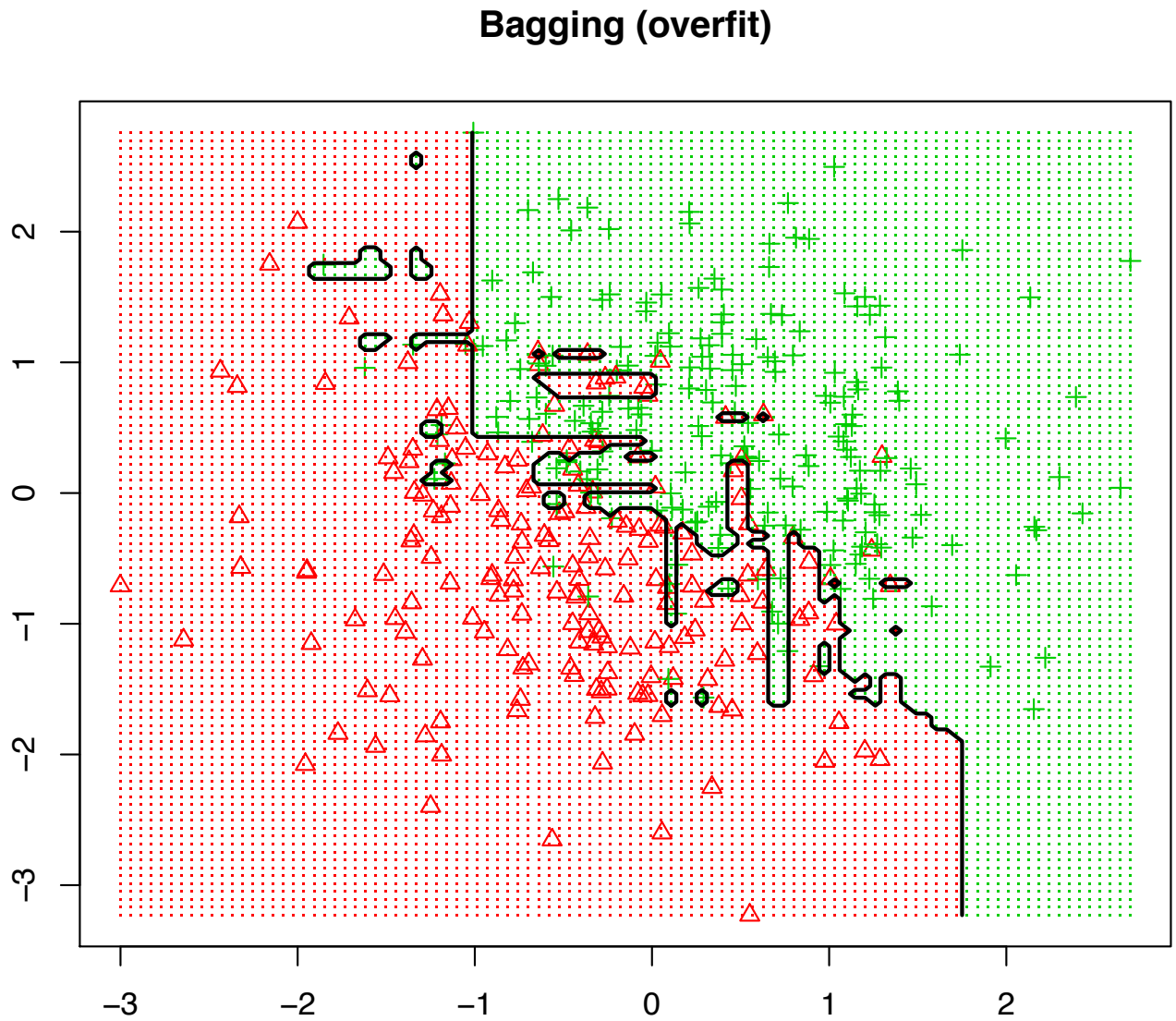
Regularization



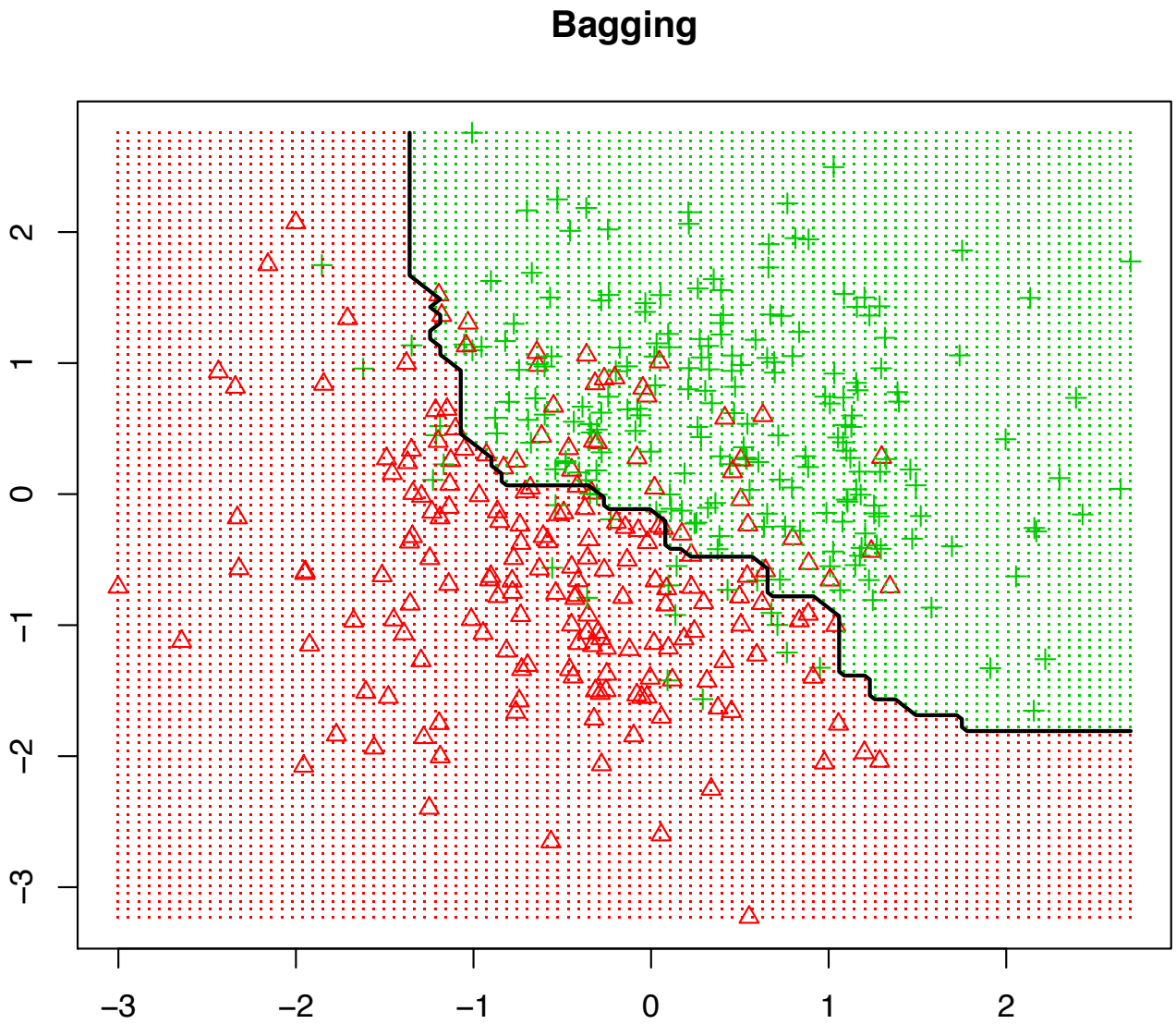
Regularization



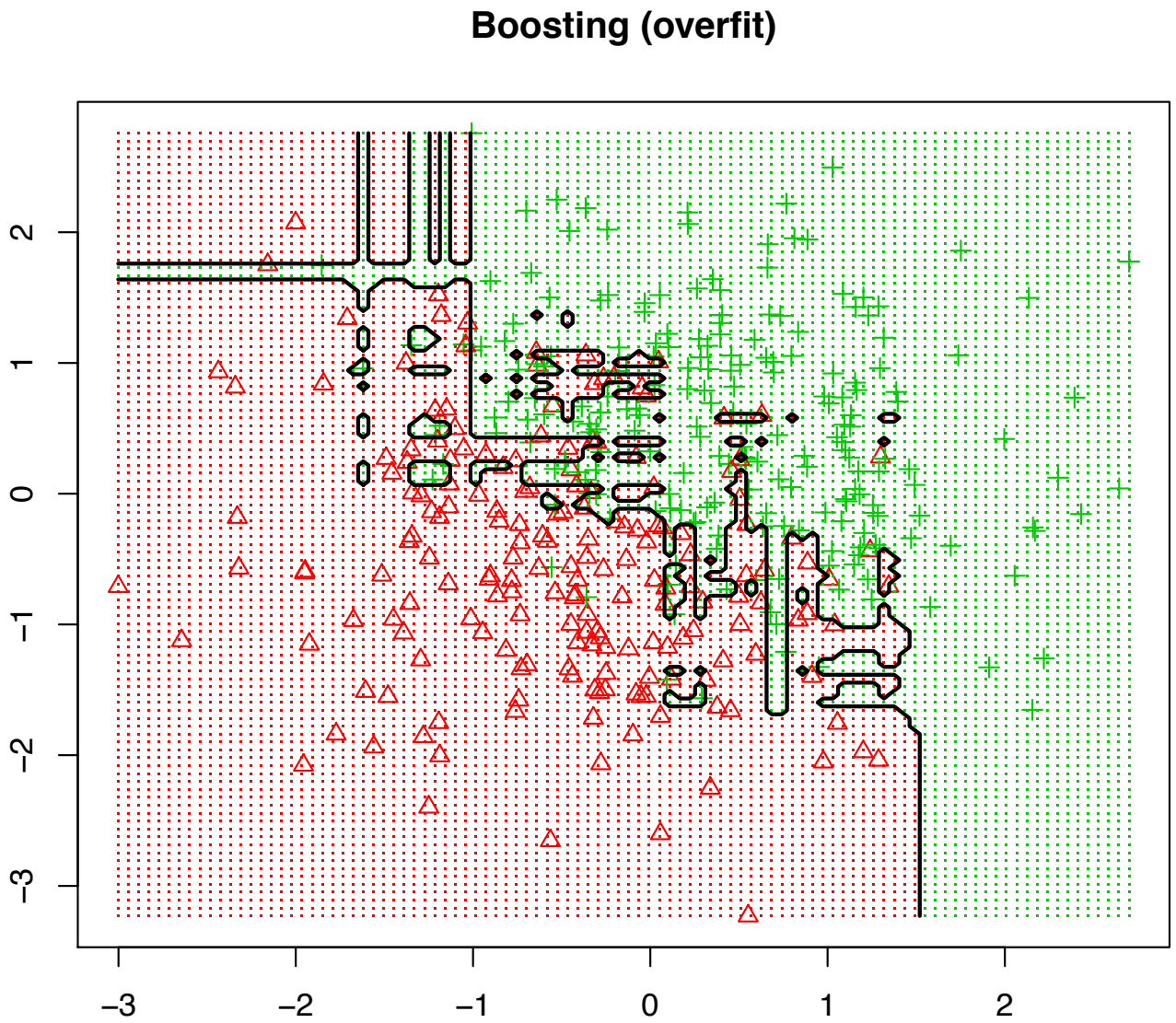
Regularization



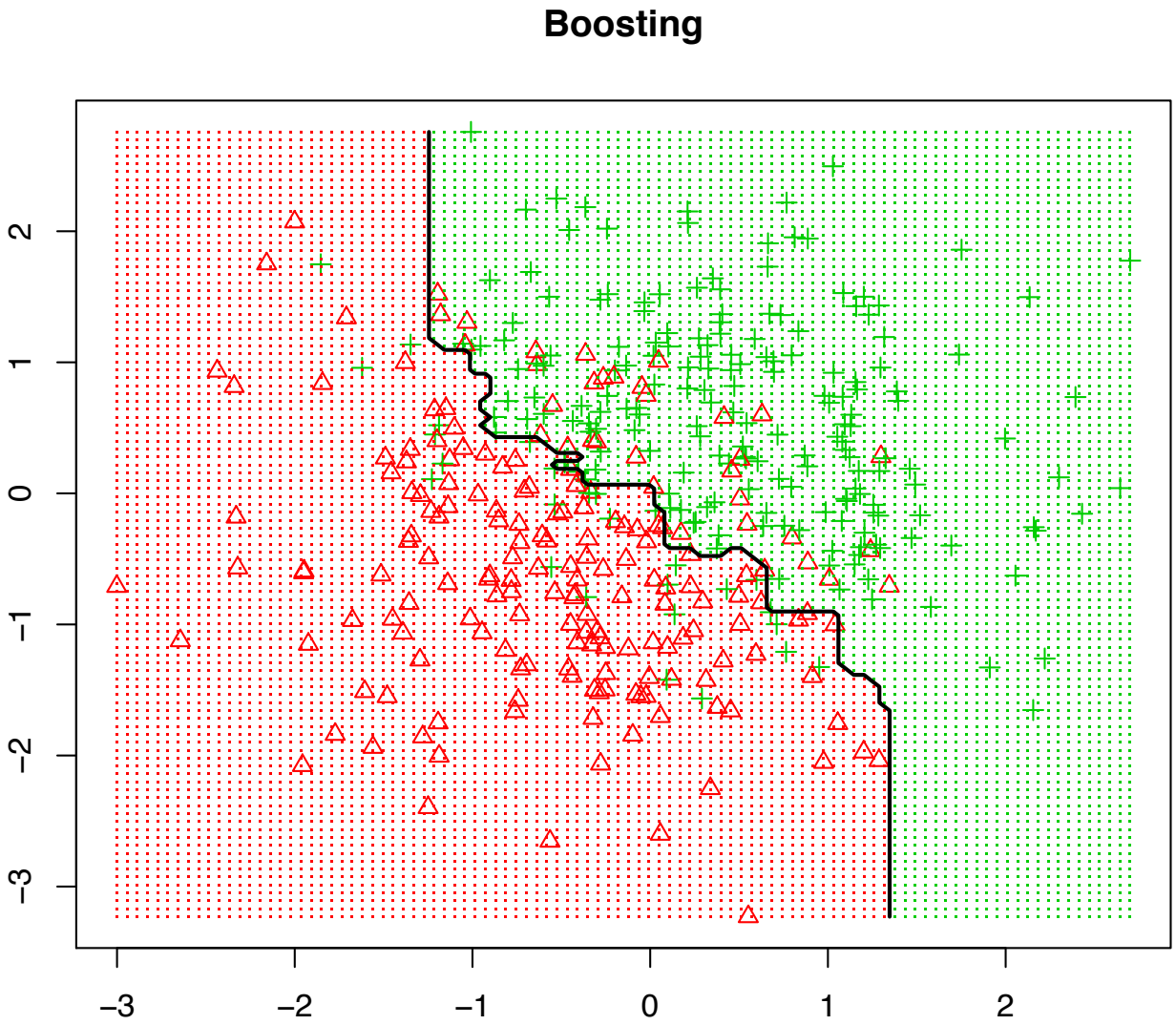
Regularization



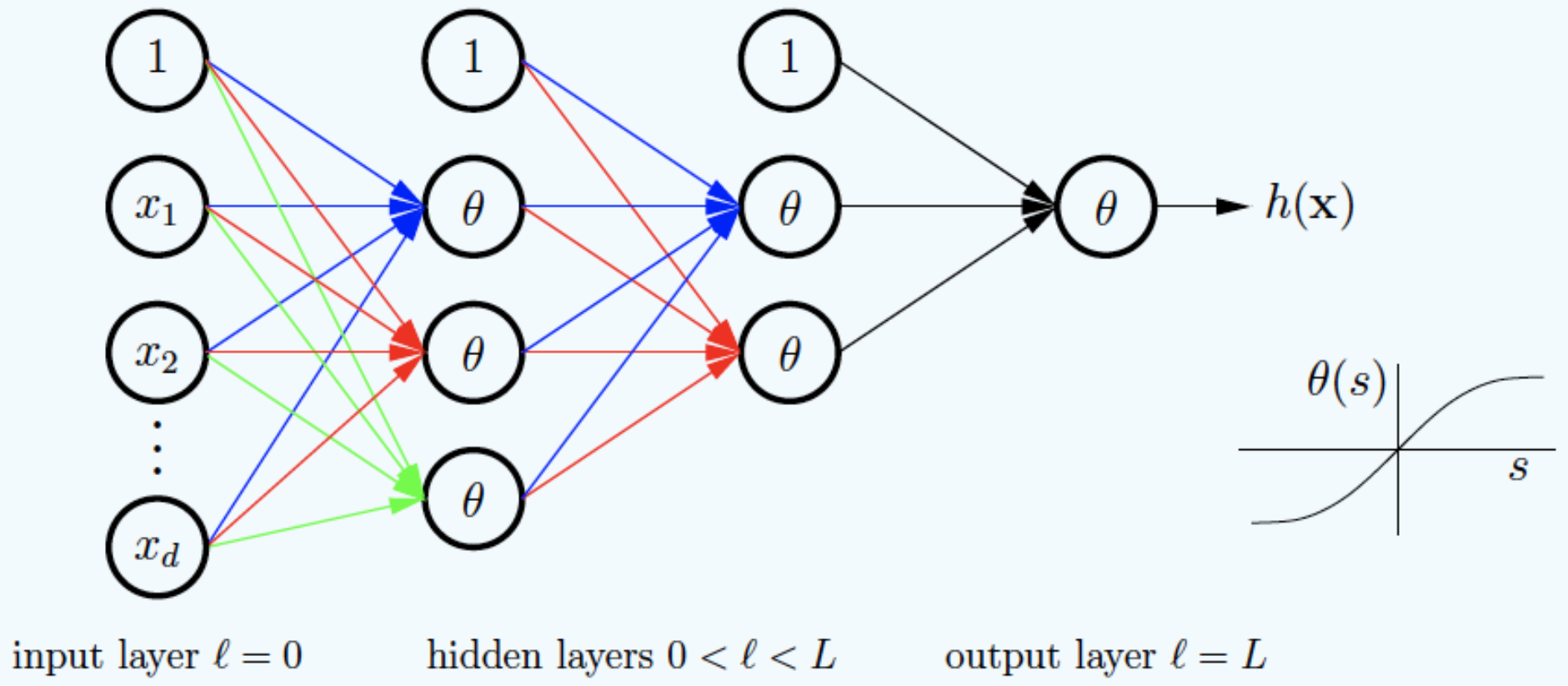
Regularization



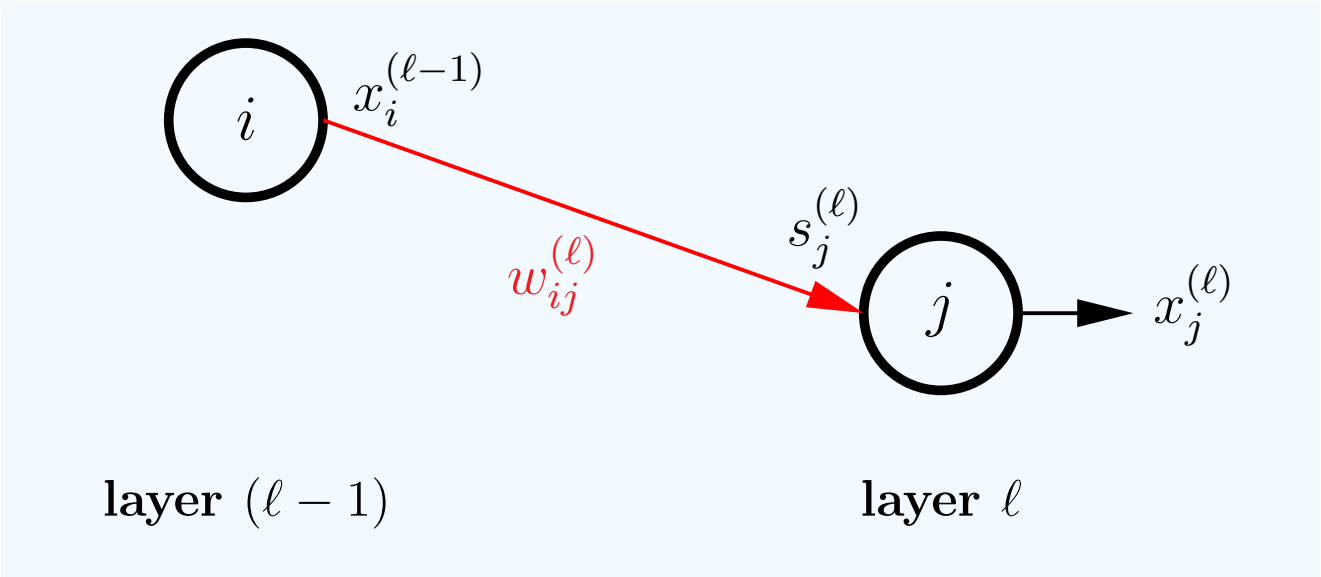
Regularization



The Neural Network

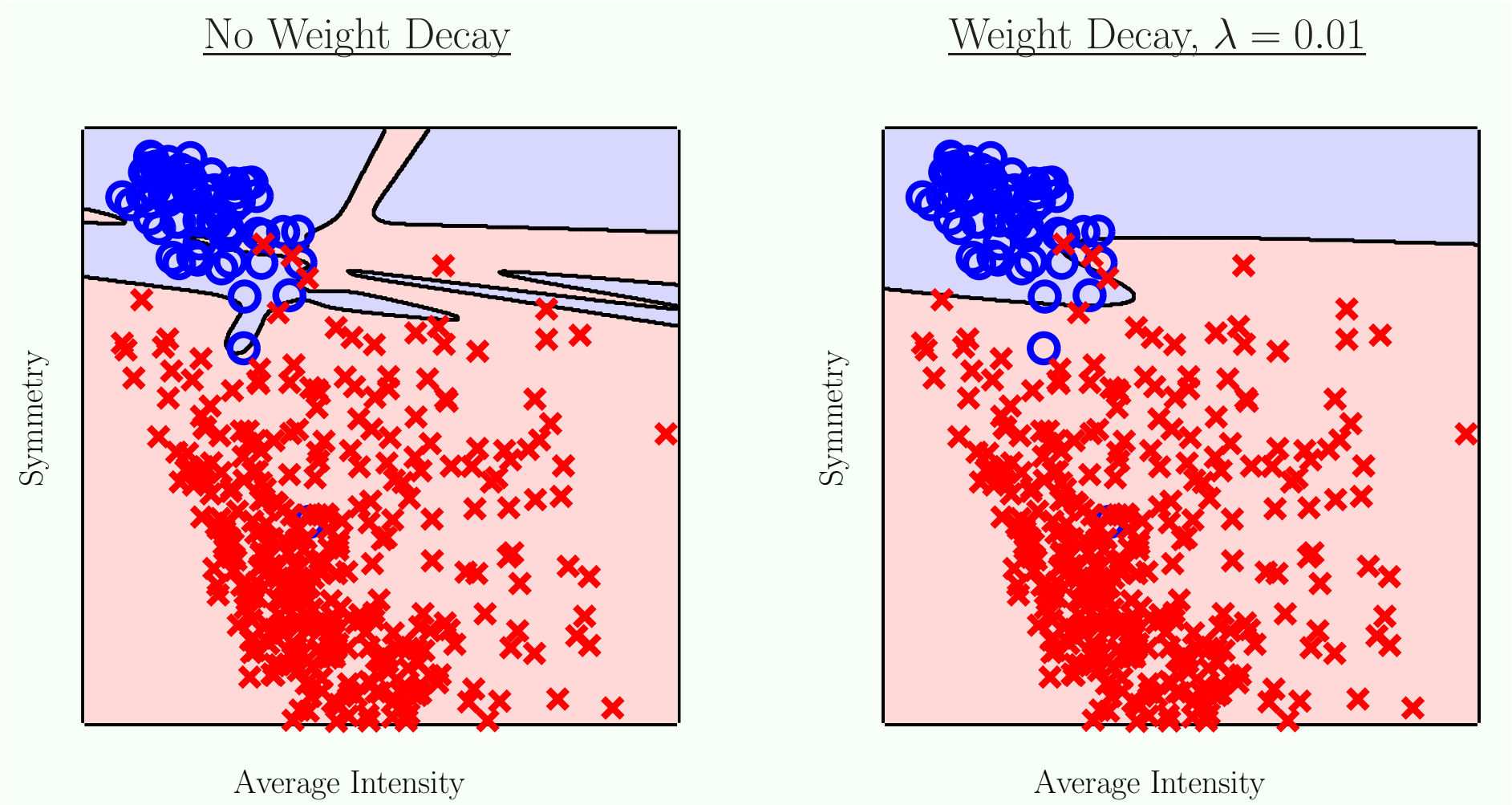


The Neural Network



$$\min_{w} \left\{ \sum_{i=1}^{N} \left(y_i - f\left(x_i; w \right) \right)^2 + \lambda \left\| w \right\|^2 \right\}$$

The Neural Network



A Marriage of Parametric and Nonparametric

- Like parametric models, neural networks have specific parameters to learn, but like nonparametric models, their complexity can grow with data.

- While classical nonparametric methods rely mainly on local averaging, modern machine and deep learning methods such as random forests and neural nets can vary almost continuously from local to global, depending on the need of the data, thus having the potential to out-perform both classical parametric and nonparametric methods in finite sample.

Supermarket Entry

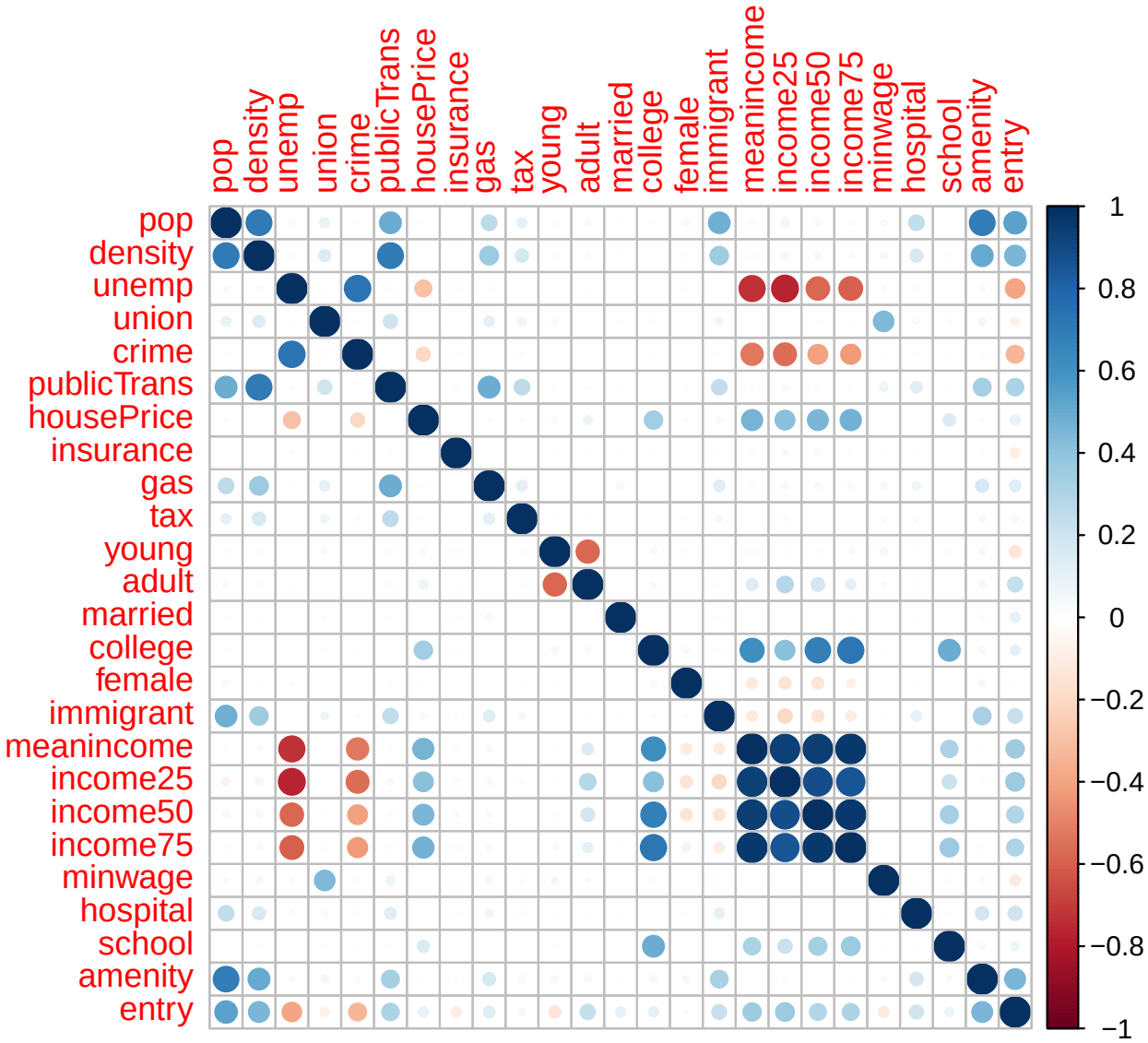
Supermarket entry in geographical markets. For each market, data include:

- total population
- population density
- average income
- percentage population with college degree
- minimum wage
- unemployment rate
- crime rate
- etc.

Supermarket Entry

```
# read data
data = read.csv("supermarket_entry.csv")
data[,1:24] = scale(data[,1:24]) # center & scale data
data$entry = as.factor(data$entry)

# create training and test sets
require(caret)
train =  createDataPartition(data$entry,p=0.5,list=F)
data_train = data[train,]
data_test = data[-train,]
```

Supermarket Entry



Supermarket Entry

```
########################
# Logistic Regression #
########################
fit = glm(entry~.,data_train,family="binomial")

# test err
ytrue = data_test$entry
phat = predict(fit,data_test,type="response")
yhat = as.numeric(phat > 0.5)
table(ytrue,yhat)

##       yhat
## ytrue   0   1
##     0 901  70
##     1  63 536

1-mean(yhat==ytrue)

## [1] 0.08471338
```

Supermarket Entry

```
require(AER)
coeftest(fit)
```

```
##
## z test of coefficients:
##
##                Estimate Std. Error  z value  Pr(>|z|)
## (Intercept) -2.309022   0.191259 -12.0727 < 2.2e-16 ***
## pop          2.569850   0.263613   9.7486 < 2.2e-16 ***
## density      1.596049   0.224054   7.1235 1.052e-12 ***
## unemp       -3.155528   0.552032  -5.7162 1.089e-08 ***
## union       -1.079137   0.147944  -7.2942 3.004e-13 ***
## crime       -0.805874   0.177419  -4.5422 5.567e-06 ***
## publicTrans  0.408581   0.177528   2.3015  0.021363 *
## housePrice  -0.892176   0.147556  -6.0463 1.482e-09 ***
## insurance   -0.964890   0.132385  -7.2885 3.135e-13 ***
## gas         -0.014365   0.121535  -0.1182  0.905912
## tax         -1.005974   0.138017  -7.2887 3.128e-13 ***
## young       -0.192683   0.181668  -1.0606  0.288857
## adult        1.761626   0.258098   6.8254 8.767e-12 ***
## married      1.254604   0.140756   8.9134 < 2.2e-16 ***
## college      1.774805   0.399058   4.4475 8.688e-06 ***
## female      -0.463359   0.155336  -2.9829  0.002855 **
## immigrant   -0.339721   0.171638  -1.9793  0.047784 *
## meanincome   0.784775   0.935196   0.8392  0.401382
## income25     1.578799   0.525607   3.0038  0.002667 **
## income50    -1.157721   0.536635  -2.1574  0.030977 *
## income75    -1.370595   0.770652  -1.7785  0.075324 .
## minwage     -0.949805   0.140431  -6.7635 1.347e-11 ***
## hospital     0.891519   0.133268   6.6897 2.237e-11 ***
```

Supermarket Entry

```
##################
# Random Forest #
##################
require(randomForest)
fit = randomForest(entry~.,data=data_train,mtry=6) # mtry selected by cv
```

```
# test err
yhat = predict(fit,data_test)
1-mean(yhat==ytrue)

## [1] 0.0433121
```

Supermarket Entry

```
############
# Boosting #
############
require(gbm)
data_boost = transform(data_train,entry=as.numeric(entry)-1)
fit = gbm(entry~.,data=data_boost,distribution="adaboost",
          n.trees=2000,
          interaction.depth=10,
          shrinkage = 0.01) # parameters selected by cv
```

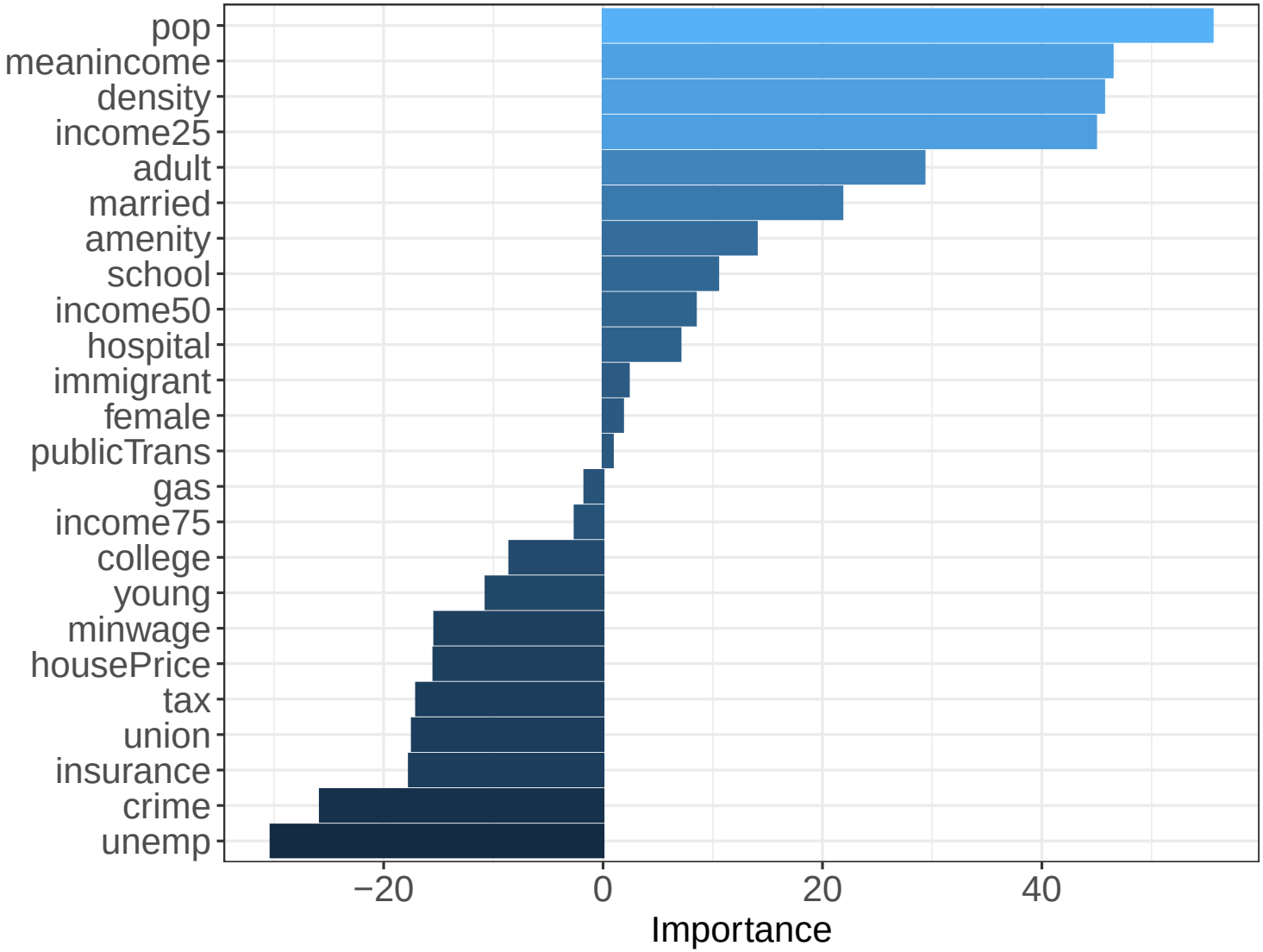
```
# test err
phat = predict(fit,data_test,n.trees=2000,type="response")
yhat = as.numeric(phat>0.5)
1-mean(yhat==ytrue)

## [1] 0.02802548
```

Supermarket Entry

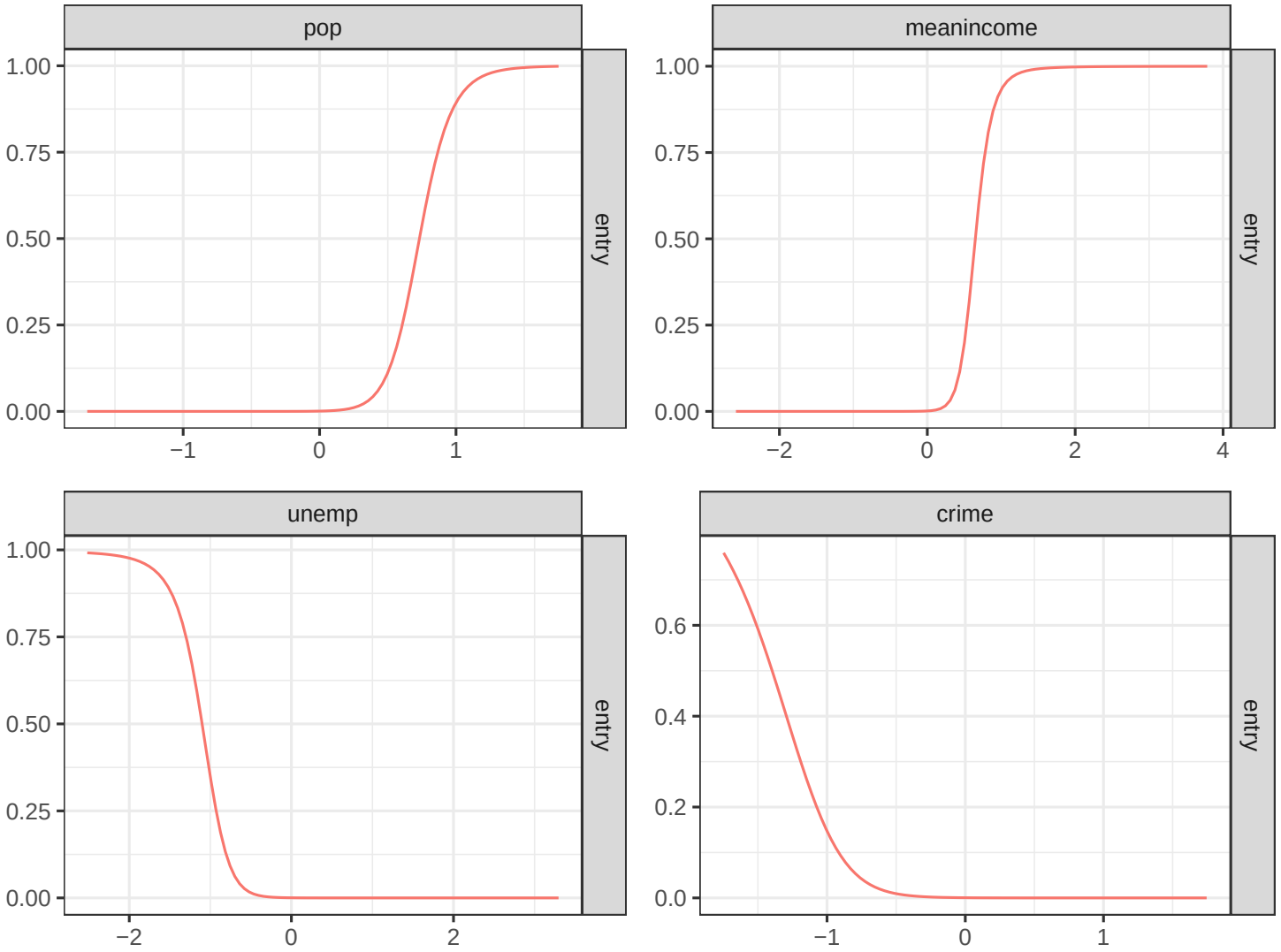
```
###################
# Neural Network #
###################
# Fit a single-hidden-layer network
require(nnet)
fit = nnet(entry ~.,data_train,maxit=10000,
           size=10,
           decay=0.1) # parameters selected by cv
```

```
# test err
yhat = predict(fit,data_test,type="class")
1-mean(yhat==ytrue)

## [1] 0.01910828
```

Supermarket Entry



Supermarket Entry



Partial dependence (other variables held at median)

High Dimensional Problems

- Most traditional statistical techniques for regression and classification are intended for the low-dimensional settings in which $N \gg p$[1].
- Settings in which p is large relative to N – in particular, $p > N$ – are often referred to as high-dimensional. In high dimensions, even linear models are *too flexible*.

[1] p is the dimension of x.

Shrinkage Estimators

- **Ridge regression**:

$$\widehat{\beta}^R = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2 \right\}$$

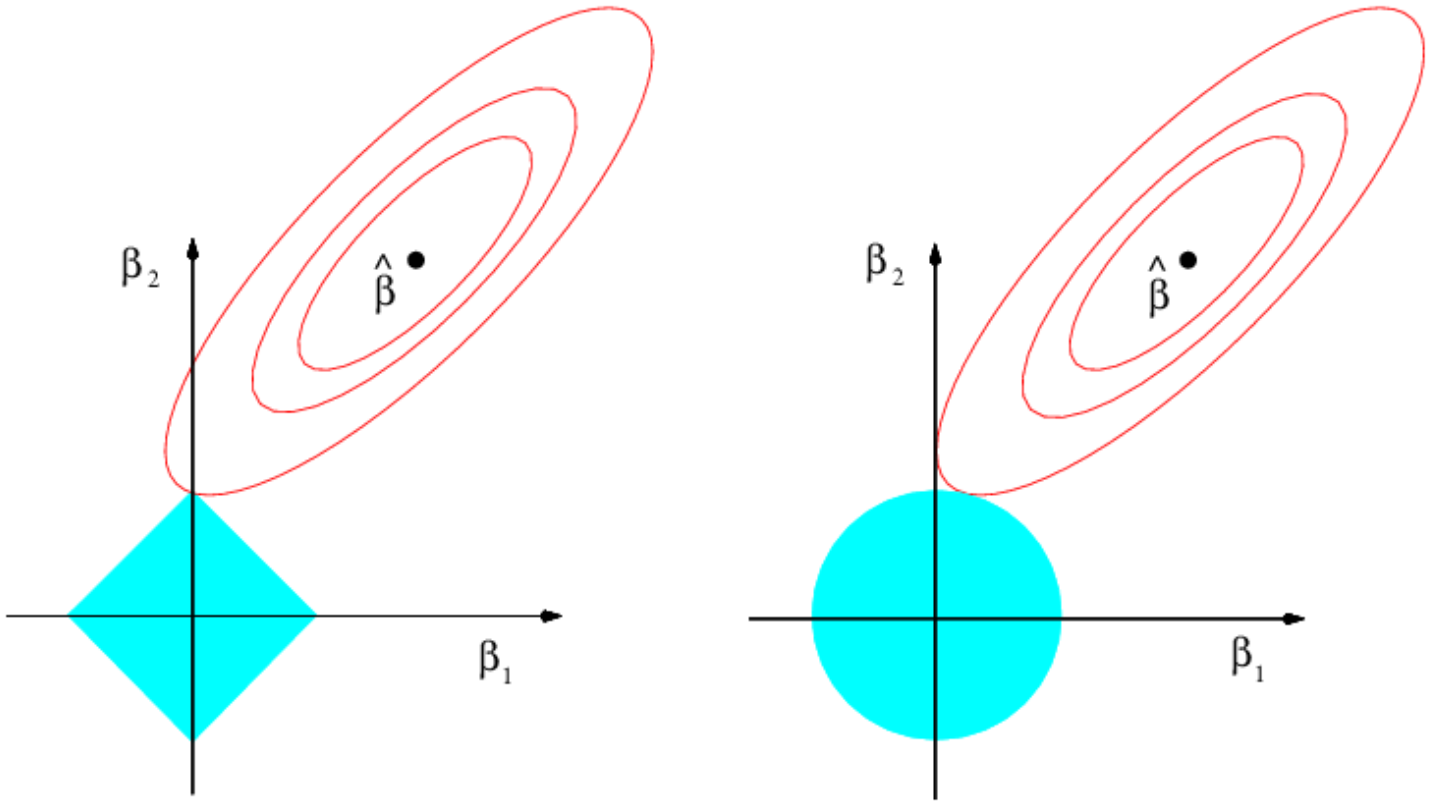
- **The lasso**:

$$\widehat{\beta}^L = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - x_i'\beta\right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Shrinkage Estimators

- The lasso and the ridge regularize linear models by shrinking the coefficient estimates towards zero.
- In the case of the lasso, the ℓ_2 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter is sufficiently large.
- Hence, the lasso also performs model selection. We say that the lasso yields **sparse models** — models that involve only a subset of the variables.

Shrinkage Estimators



Contours of the error and constraint functions for the lasso (left) and ridge regression (right). The solid blue areas are the constraint regions, $|\beta_1| + |\beta_2| \leq C$ and $\beta_1^2 + \beta_2^2 \leq C$, while the red ellipses are the contours of the RSS.

U.S. Election

- We want to predict whether a person voted for Clinton or Trump in the 2016 U.S. presidential election using demographic variables taken from the 2018 General Social Survey (GSS).
- Selected variables include: age, sex, income, work, marriage, family information, political party affiliation, etc.

U.S. Election

```
gss <- read.csv("GSS2018.csv")
dim(gss)

## [1] 455  50

names(gss)

##  [1] "age"      "sex"      "sibs"     "size"     "adults"   "childs"
##  [7] "class"    "coninc"   "actssoc"  "attend"   "born"     "cappun"
## [13] "courts"   "degree"   "earnrs"   "ethnum"   "famgen"   "finalter"
## [19] "finrela"  "fund"     "hhrace"   "madeg"    "marital"  "mobile16"
## [25] "natmass"  "natpark"  "othlang"  "natchld"  "sexornt"  "partyid"
## [31] "phone"    "polviews" "pres16"   "prestg10" "natsci"   "quallife"
## [37] "race"     "raclive"  "rank"     "relig"    "reliten"  "satfin"
## [43] "satsoc"   "sei10"    "vetyears" "vote12"   "happy"    "weekswrk"
## [49] "wrkslf"   "wrkstat"
```

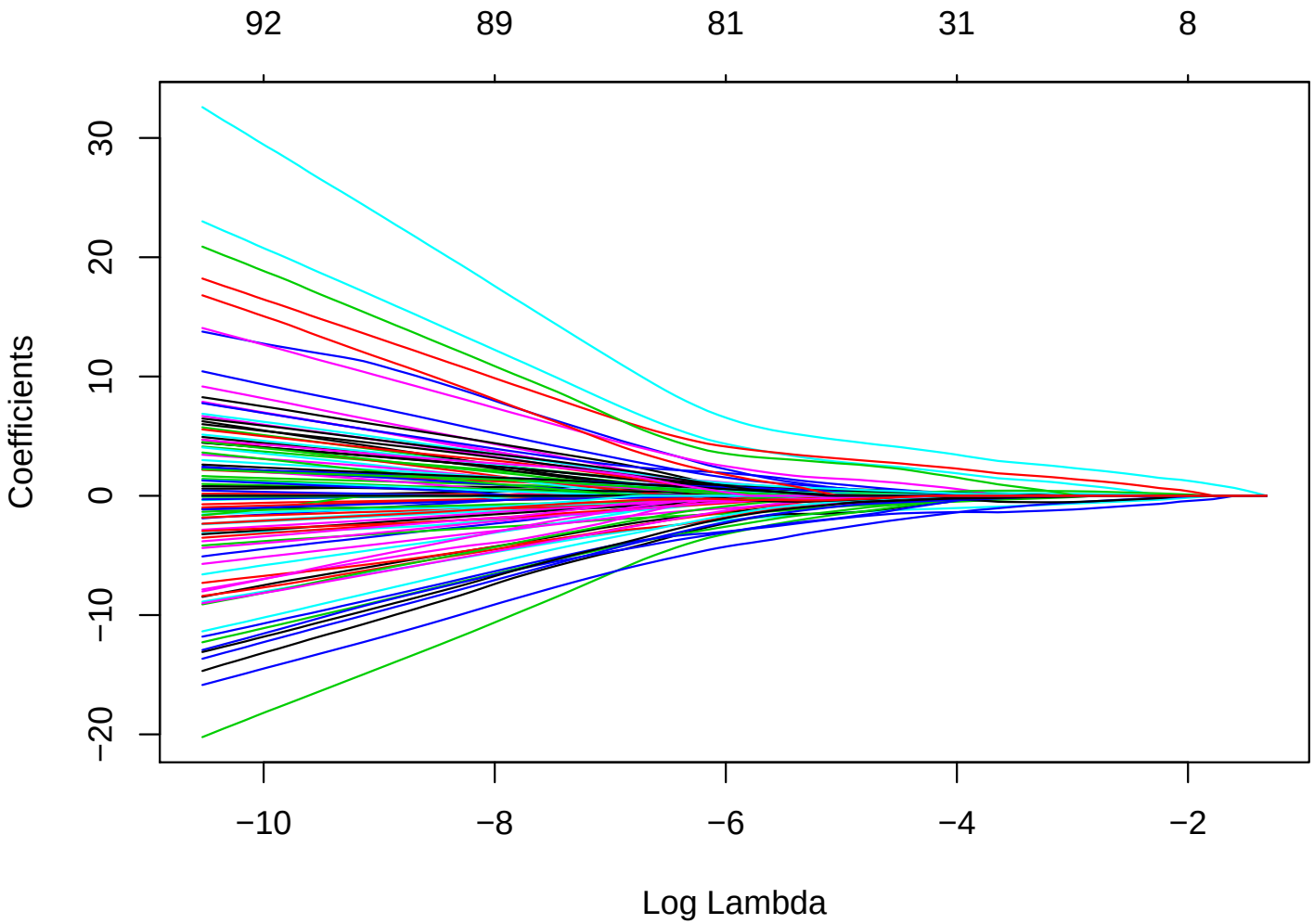
U.S. Election

```
Y <- gss$pres16 # GSS Q: Did you vote for Clinton or Trump?
summary(Y)

## Clinton   Trump
##     260     195

X <- model.matrix(pres16 ~.,gss)[,-1] # create dummies for categorical vars
dim(X)

## [1] 455 128
```

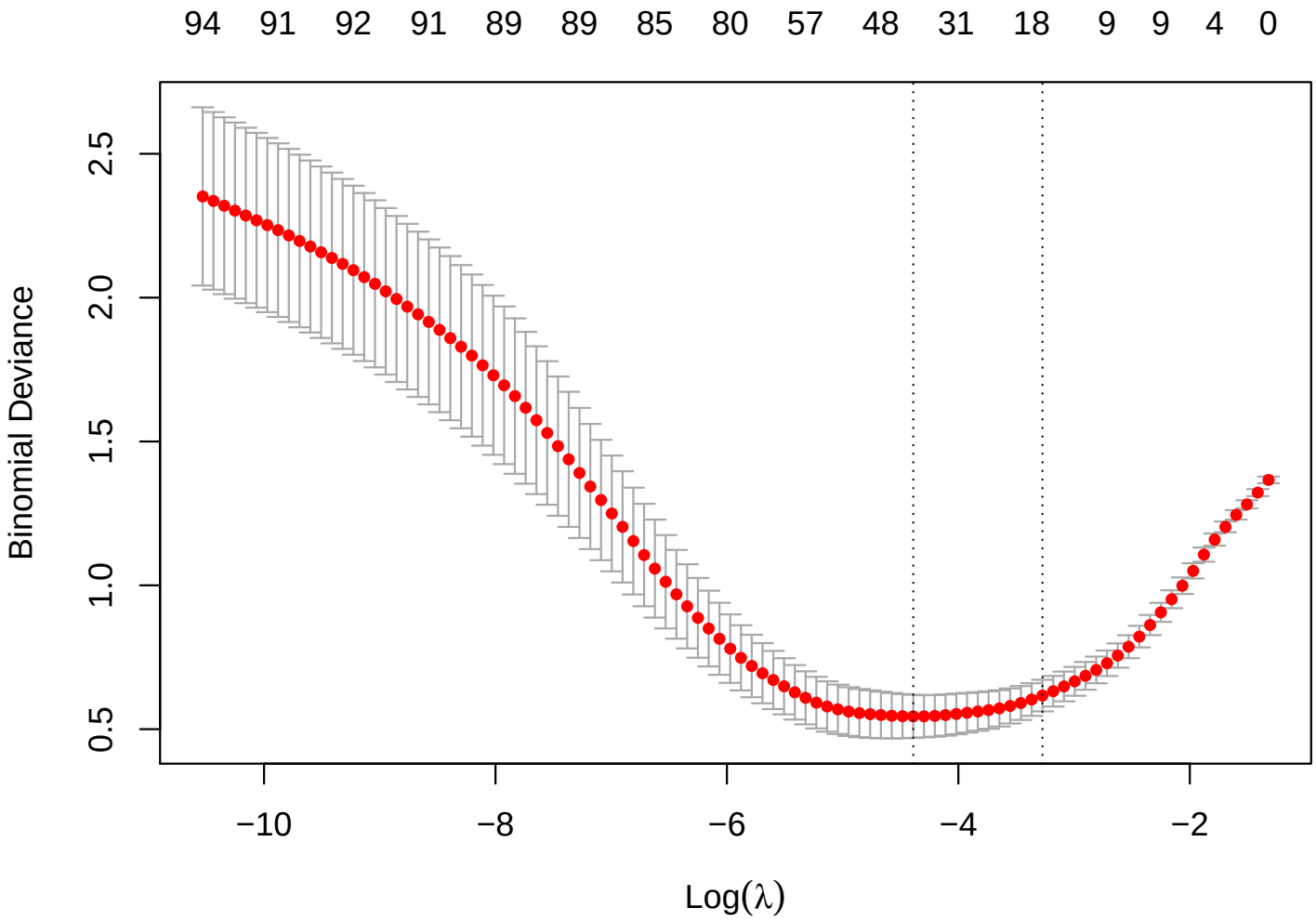
U.S. Election

```
################################
# Logistic Lasso Regression #
################################
require(glmnet)
cv.lasso <- cv.glmnet(X,Y,alpha=1,family="binomial") # cross-validation
# choose lambda associated with min CV error (or plus 1se)
lambda.star <- cv.lasso$lambda.1se # Alternatively: cv.lasso$lambda.min
# fit the lasso with optimal lambda
fit = glmnet(X,Y,alpha=1,lambda=lambda.star,family="binomial")
```

U.S. Election



U.S. Election


94 91 92 91 89 89 85 80 57 48 31 18 9 9 4 0
Binomial Deviance
0.5
1.0
1.5
2.0
2.5
−10
−8
−6
−4
−2
Log(λ)

U.S. Election

```
source('lassosummary.R') # external function
lassosummary(fit)
```

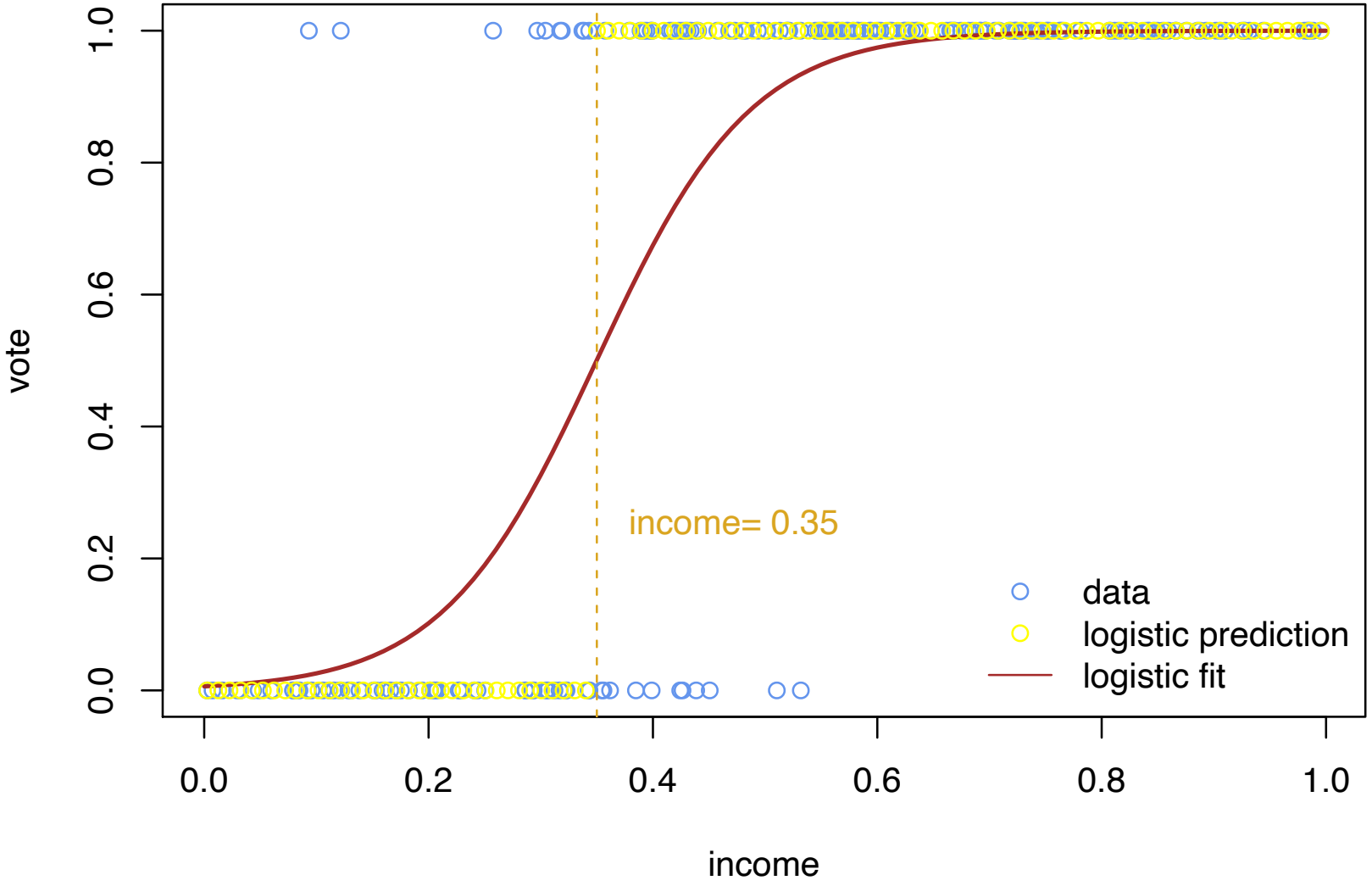
```
##                                 beta
## sexmale                   0.01785683
## cappunoppose             -0.77766891
## hhraceblack              -0.73658259
## hhracewhite               0.20921740
## maritalmarried            0.06177141
## maritalnever married     -0.21100431
## natchldtoo little        -0.11405573
## partyidind,near rep       1.22815699
## partyidnot str democrat  -0.54035075
## partyidnot str republican 1.63637730
## partyidother party        0.49325142
## partyidstrong democrat   -1.26057401
## partyidstrong republican  2.60413814
## polviewsliberal          -0.03666117
## polviewsslightly liberal -0.04752788
## racewhite                 0.41245491
## religprotestant           0.05175686
## satfinsatisfied           0.02233638
```

The Three Musketeers

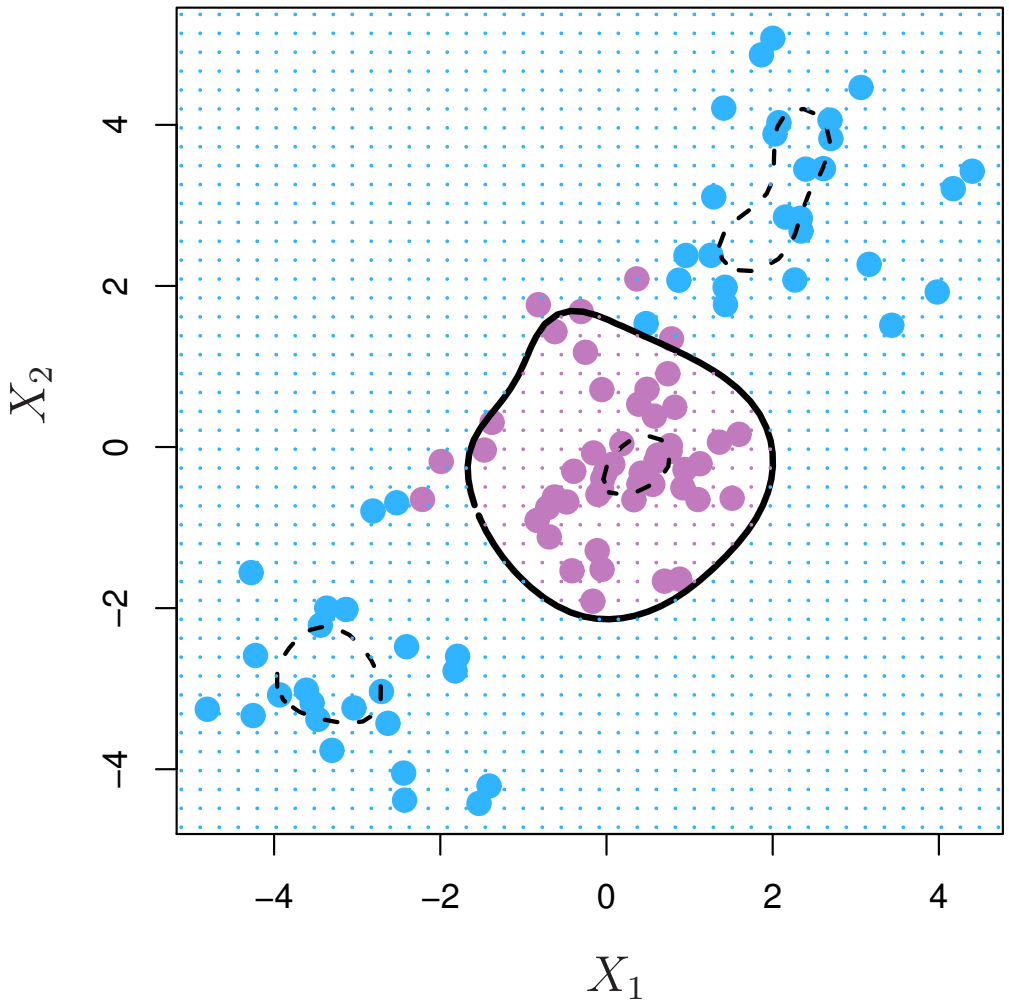
In general, if the goal is to predict y based on x, there are distinct approaches:

1. Find a deterministic function f such that $y = f(x) + e$, and learn the target function f.
2. Learn $p(y|x)$ using a probabilistic model and use the estimated $p(y|x)$ to predict y – in this case we have a *target distribution* rather than a target function.
3. Let $p(x, y)$ be the target distribution, from which we can calculate $p(y|x) = \frac{p(x,y)}{p(x)}$.

Logistic Regression



Support Vector Machine



Generative vs. Discriminative Models

- Models of the joint distribution $p(x, y)$ are called **generative models**, while models of $p(y|x)$ or $f(x)$ are called **discriminative models**.

- While discriminative models are mainly used for prediction tasks, generative models allows us to do more than just making predictions of y given x. We can, for example, generate new data points $\{(x_i, y_i)\}$ by drawing from $\widehat{p}(x, y)$. These new data points are called **synthetic data**, since they are not real, observed data. The process of generating synthetic data is called **simulation**.

Scientific Models

- **Scientific models** (**causal models**) are an important type of generative models that describe the **causal mechanisms** that generate $p(x, y)$.
- While scientific models can be used for prediction, the goal of learning causal mechanisms is distinct from the goal of prediction.

Scientific Models

Scientific vs. Statistical Model

If you want to predict where Mars will be in the night sky[a], you may do very well with a model in which Mars revolves around the Earth. You can estimate, from data, how fast Mars goes around the Earth and where it should be tonight. But the estimated model does not describe the actual causal mechanisms. Nor does it need to: if our only goal is prediction, then we often do not need a scientific model.

[a] This example is taken from Shalizi (2019).

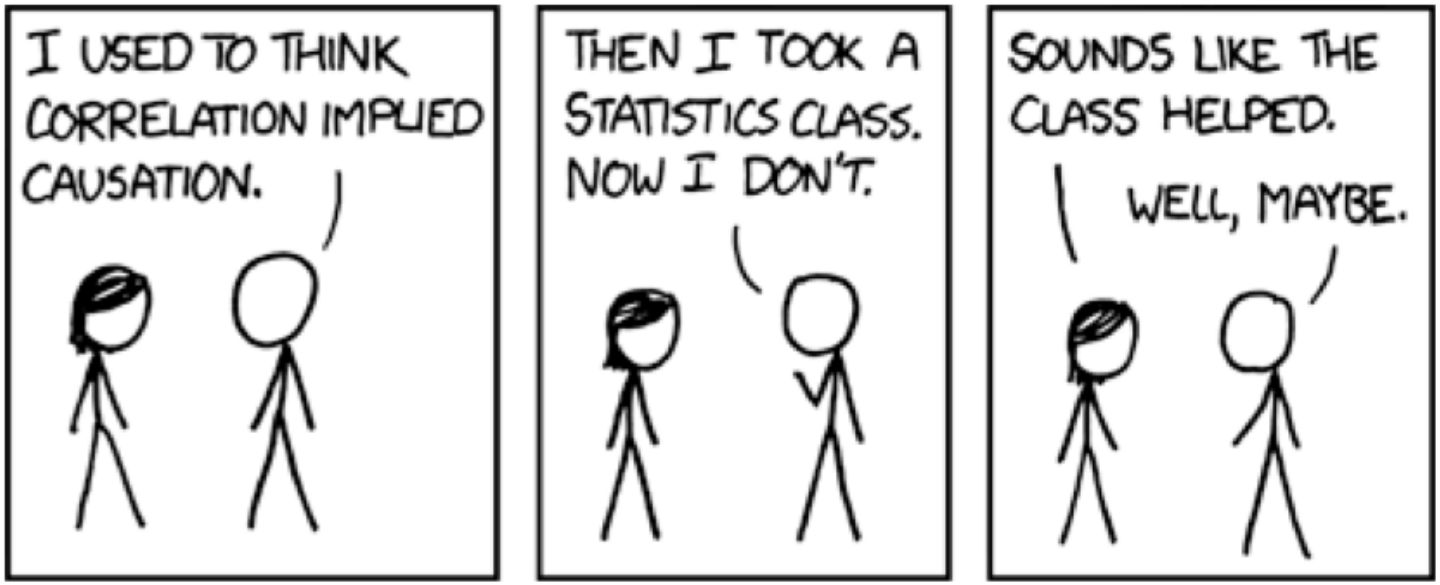
Scientific Models

- Because scientific models describe causal mechanisms, what we learn from one set of data $\mathcal{D} \sim p(x, y)$ can be potentially used to explain and predict data drawn from another distribution, say $p(u, v)$, if $\{x, y\}$ and $\{u, v\}$ share similar underlying causal mechanisms.
  - In other words, what we learn from one observed phenomenon can be used to explain and predict other related phenomena.
  - For example, we can learn individuals' risk aversion from their investment behavior, which in turn, can help explain and predict their career choices.

Scientific Models

- Good scientific models (Quantum Mechanics!) can potentially deliver better predictive performance than statistical models trained on single data sets, because they can be learned from a combination of data from various sources that share the same underlying causal mechanisms.
    - Apples falling down trees and the earth orbiting around the sun both inform us of the gravitational constant.

Causal Inference



Causal Inference

"Causa latet: vis est notissima (The cause is hidden, but the result is known)" – Ovid, Metamorphoses, IV. 287.

"Felix qui potuit rerum cognoscere causas (happy be the man who has been able to learn the causes of things)" – Virgil, Georgics, II, 490.

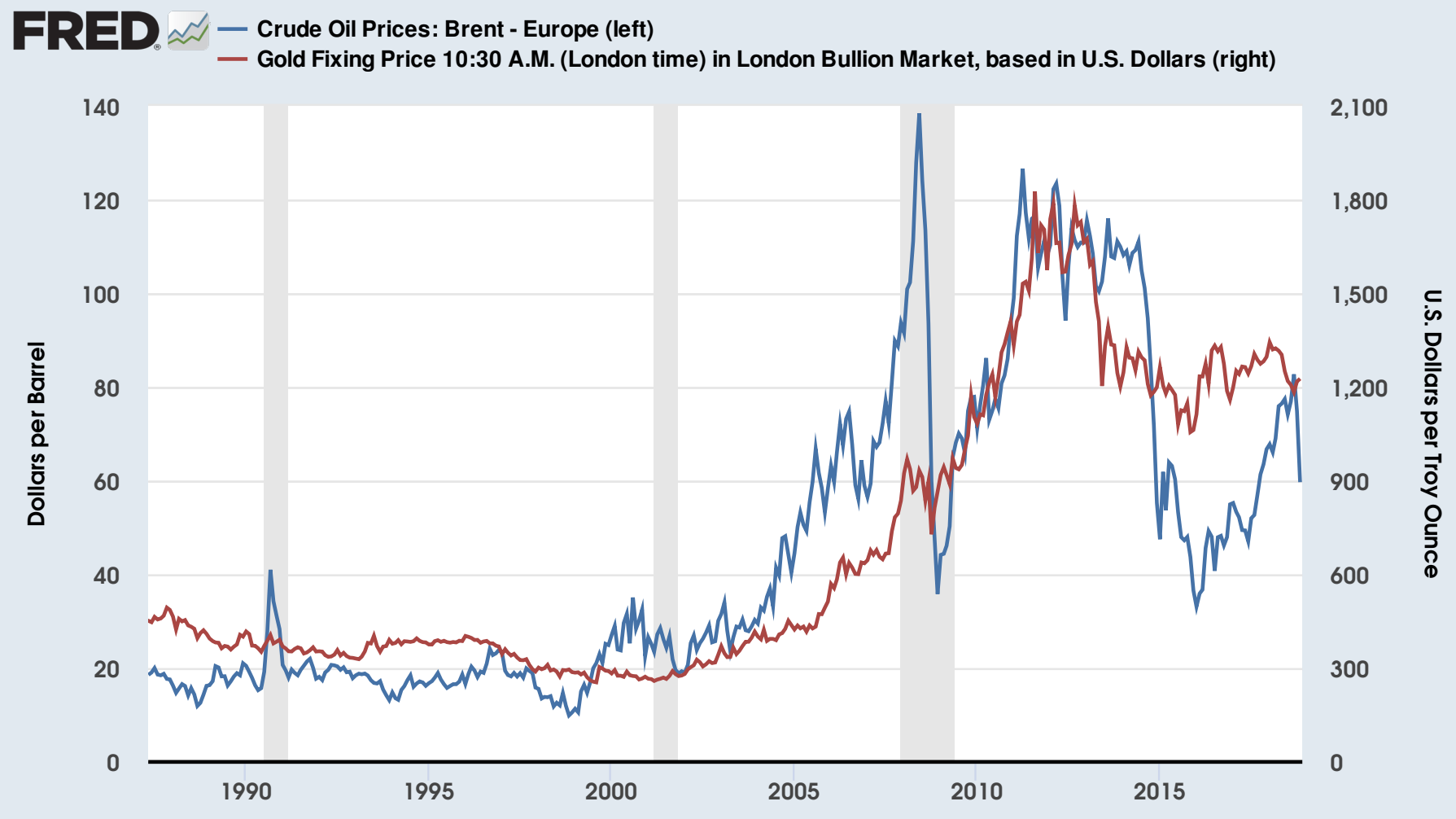
"Shallow men believe in luck or in circumstance. Strong men believe in cause and effect." — Ralph Waldo Emerson

Causal Inference

Causal inference is concerned with the following questions:

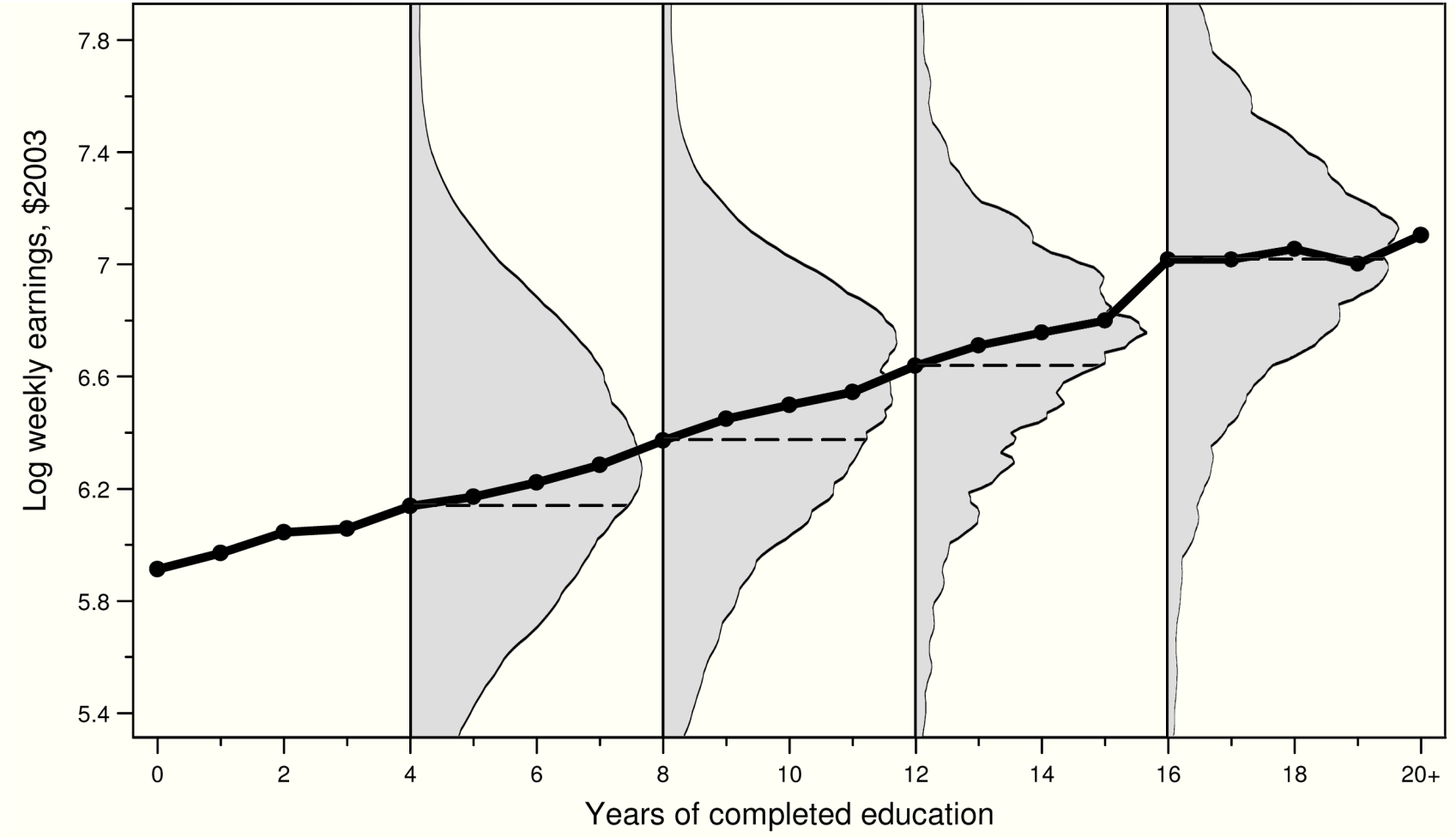
1. Does x have a causal effect on y? If so, how large is the effect? (**causal effect learning**)
2. If a causal effect exists, what is the mechanism by which it occurs? (**causal mechanism learning**)

Causal Inference



Do oil and gold prices cause each other to move?

Causal Inference



Does receiving more education make you earn more?

Program Evaluation

Evaluating and predicting the effects of government programs and economic policies is a central problem in applied economic research:

- Effect of worker training programs on employment
- Effect of early childhood interventions on adult outcomes
- Effect of negative income taxes on labor supply
- Effect of environmental regulations on pollution emission
- ...

Why Causal Inference

- Learning patterns in the data is not enough. We want understanding.
  - the focus of ☼science☼.
- Understanding how things work makes a big difference in how we act: if the rooster's crow causes the sun to rise, we could make the night shorter by waking up our rooster earlier.
- Ultimately, every question related to the effect of actions must be decided by causal considerations. Statistical information alone is insufficient.
- True understanding enables predictions under a wide range of circumstances, including new hypothetical situations.

Russell's Chicken

Russell's Chicken

Bertrand Russell[a] told the following cautionary tale of the perils of not understanding causal mechanisms:

> *A chicken infers, on repeated evidence, that when the farmer comes in the morning, he feeds her. The inference serves her well until Christmas morning, when he wrings her neck and serves her for dinner.*

[a] Russell (1912), via Deaton and Cartwright (2018).

Simpson's Paradox

	Drug	No drug
Men	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Women	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Should a doctor prescribe this drug?

- Any statistical relationship between two variables may be reversed by including additional factors in the analysis.
- Causal relationships are more stable than statistical relationships.

Artificial Intelligence

Research on causal inference methodologies has taken on new importance with the development of artificial intelligence (AI).

- How should a robot acquire causal information through interaction with its environment?
- How should a robot receive causal information from humans?

Causal Inference Frameworks

- The potential outcomes framework[2]
- Causal graphical model

[2]Also called the Rubin causal model (RCM). The RCM uses the language of statistical analysis of experiments to model causality by conceptualizing observed data as if they were outcomes of experiments, conducted either by the researcher – as in actual experiments, or by the subjects of the research themselves – as in observational studies.

Seeing vs. Doing

The do operator:

$$\text{do}(x = a) : \text{set } x = a$$

- Barometer readings are useful for predicting rain:

$$\Pr(\text{rain} \mid \text{barometer} = \text{low}) > \Pr(\text{rain} \mid \text{barometer} = \text{high})$$

- But hacking a barometer won't change the probability of raining:

$$\Pr(\text{rain} \mid \text{do}(\text{barometer} = \text{low})) = \Pr(\text{rain} \mid \text{do}(\text{barometer} = \text{high}))$$

Seeing vs. Doing

- Doing: if x has a causal effect on y, then we can change x and expect it to cause a change in y.
- Seeing: If x is correlated with y but does not have a causal effect on y, then we can only observe the correlation without the ability to change y by manipulating x.

Causal vs. Statistical Predictions

- **Causal prediction**: What will y be if I *set* $x = a$?
  - $E[y|\text{do}(x = a)]$[3]
- **Statistical prediction**: What will y be if I *observe* $x = a$?
  - $E[y|x = a]$

[3] Assuming we minimize the expected L2 loss in prediction.

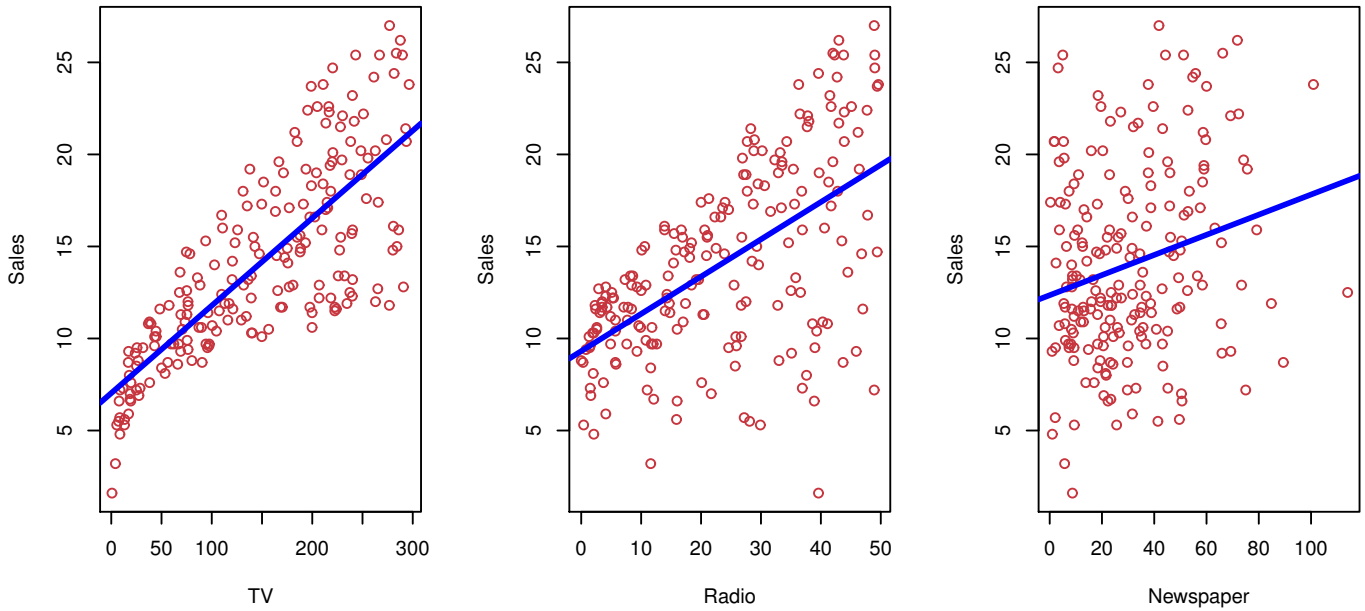
Hospitalization and Health

Average health (assigning a 1 to poor health and a 5 to excellent health) contrasting those who have been an inpatient in the past 12 months and those who have not (tabulated from the 2005 NHIS):

Group	Sample Size	Mean health status	Std. Error
Hospital	7774	3.21	0.014
No Hospital	90049	3.93	0.003

- Q1: what is the expected health status of someone who has received hospitalization? (statistical prediction)
- Q2: what will my health status be if I receive hospitalization? (causal prediction)

Advertising and Sales



- Q1: what is the expected sales of a company with a given amount of TV ad spending? (statistical prediction)
- Q2: how much will my sales increase if I increase my TV ad spending by a certain amount? (causal prediction)

Causal Effect Learning

- To learn $f(x) = E[y|\text{do}(x)]$, the simplest way is to "just do it".
- Let a be a possible value of x. Randomly select individual units, *set* their $x = a$, and observe the resulting y. In this way, we can *generate data* from $p(y|\text{do}(x))$.
  - This is in essence what a randomized experiment does.
- A nonparametric estimator for $f(x)$ is then

$$\widehat{f}(x = a) = Ave(y|x = a)$$

Causal Effect Learning

Randomized Experiment

- Consider $x \in \{0, 1\}$. Suppose we are interested in learning the causal effect of $x = 1$ on y.

- Given a set of experimental units, a **randomized controlled trial** (**RCT**) randomly selects a subset of individual units – call them the **treatment group** – to receive $x = 1$, and assign $x = 0$ to the rest of the experimental units – called them the **control group**.

Randomized Experiment

Using the experimental language, x is called **treatment** and y is called **outcome**. The **average treatment effect** (**ATE**) is defined as

$$\begin{aligned}\text{ATE} &= E\left[y|\text{do}\left(x=1\right)\right] - E\left[y|\text{do}\left(x=0\right)\right] \\ &\overset{[1]}{=} E\left[y|x=1\right] - E\left[y|x=0\right]\end{aligned}$$

, where [1] follows because randomized experiments generate data from $p\left(y|\text{do}\left(x\right)\right)$, therefore $E\left[y|x\right] = E\left[y|\text{do}\left(x\right)\right]$.

For data generated by randomized experiments, correlation implies causation.

Randomized Experiment

The
Design of Experiments

By

Sir Ronald A. Fisher, Sc.D., F.R.S.

Honorary Research Fellow, Division of Mathematical Statistics, C.S.I.R.O., University of Adelaide; Foreign Associate, United States National Academy of Sciences; and Foreign Honorary Member, American Academy of Arts and Sciences; Foreign Member of the Swedish Royal Academy of Sciences, and the Royal Danish Academy of Sciences and Letters; Member of the Pontifical Academy; Member of the German Academy of Sciences (Leopoldina); formerly Galton Professor, University of London, and Arthur Balfour Professor of Genetics, University of Cambridge.

HAFNER PRESS
A Division of Macmillan Publishing Co., Inc.
New York
COLLIER MACMILLAN PUBLISHERS
London

Observational Studies

- For many problems, RCTs are impossible or impractical to run:
  - ethical reasons
  - cost and duration
  - high-dimensionality
- How do we learn $E[y|\text{do}(x)]$ from observational data?

Observational Studies

- For observational data, correlation no longer implies causation. Consider the example of hospitalization and health: the fact that hospitalization is associated with worse health outcomes may not be due to the adverse effect of hospitalization on health, but to the fact the people with worse health received hospitalization in the first place. This is called **self-selection effect** or **self-selection bias**.
- Self-selection bias is a central concern to causal inference based on observed socio-economic data generated by individual choices.
- When individuals choose their own treatments (*self-selection*), those who choose to receive a treatment can be *systematically* different than those who choose not to, leading to a correlation between treatment and outcome that is not due to direct causation.

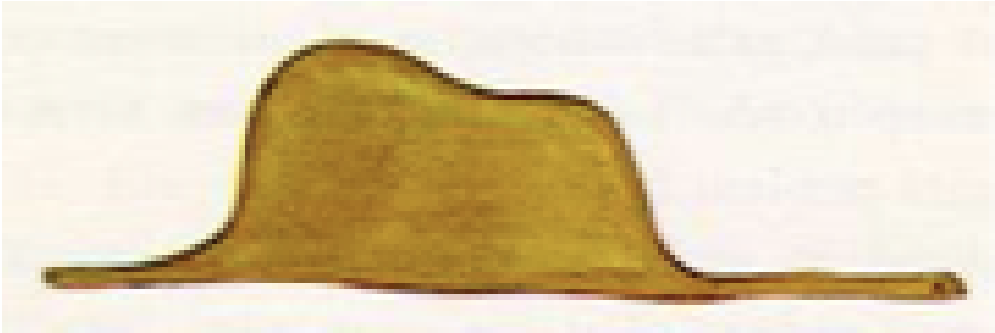
Causal Diagrams

- To learn causal effects from observational data, we need to have an understanding of the **causal mechanism** that generates the data[4].
- **Causal diagrams** are graphs that can be used to represent causal relationships and therefore describe our qualitative knowledge about a causal mechanism.

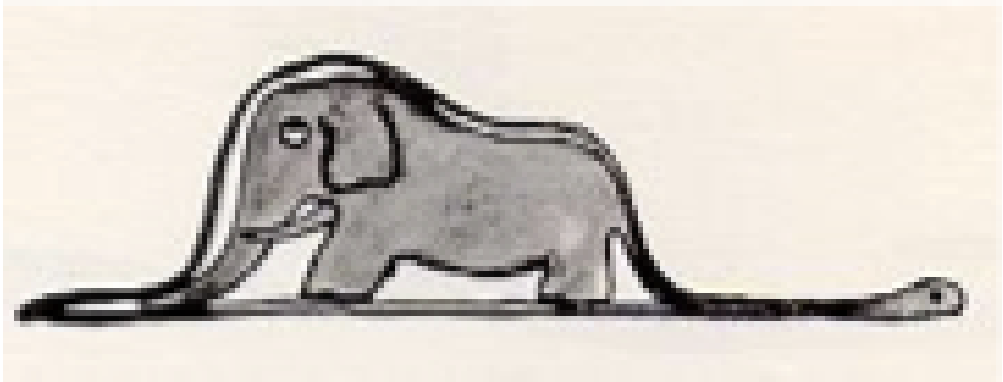
[4] In fact, such understanding is also necessary for interpreting and using experimental results. Without an understanding of – or making assumptions on – the underlying causal mechanism, any causal effect estimate is meaningless!

Causal Diagrams

For example, suppose we see a boa constrictor that looks like this:



Our theory of the causal mechanism that leads to the boa constrictor looking like this is that it has just swallowed a baby elephant:



Causal Diagrams

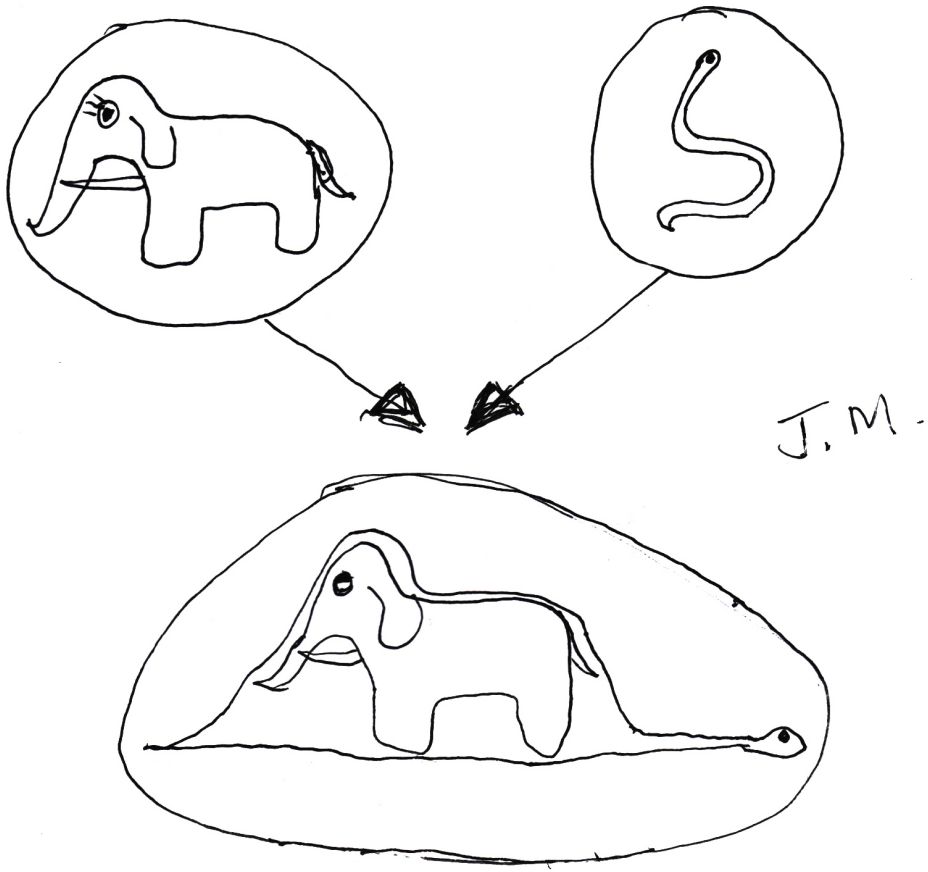
How do we represent our theory? We can write out a full causal model:

$$\begin{aligned}
\log h^e &\sim \mathcal{N}(0, 0.1) \\
\log \ell^e &\sim \mathcal{N}(0.5, 0.1) \\
\log \ell^b &\sim \mathcal{N}(1.5, 0.2) \\
a \left| \ell^e, \ell^b, \ell^b > \ell^e \right. &\sim U\left(0, \ell^b - \ell^e\right) \\
y &= \begin{cases} h^e \mathcal{I}(a \leq x \leq a + \ell^e) \mathcal{I}(E = 1) & \ell^e < \ell^b \\ 0 & \ell^e \geq \ell^b \end{cases}
\end{aligned}$$

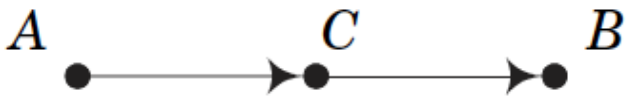
, where y is the height of the boa constrictor, x is the distance along the body of the boa constrictor from its head, (h^e, ℓ^e) are respectively the height and length of the baby elephant, ℓ^b is the length of the boa constrictor, and $E \in \{0, 1\}$ is the event that the boa constrictor has swallowed the elephant.

Causal Diagrams

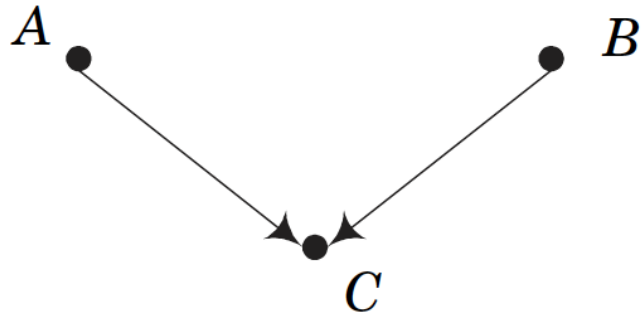
Or we can draw a causal diagram:



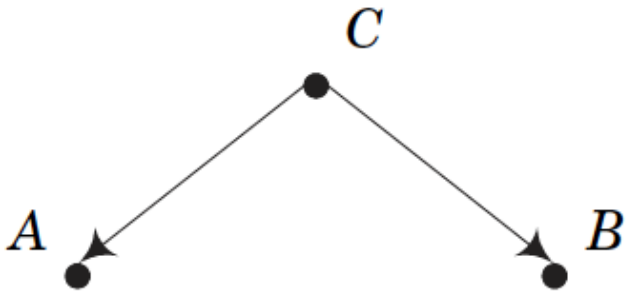
Causal Diagrams



(a) Mediation



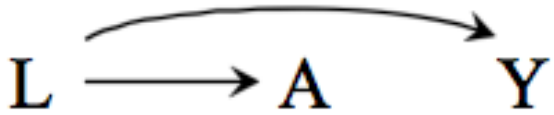
(c) Mutual causation



(b) Mutual dependence

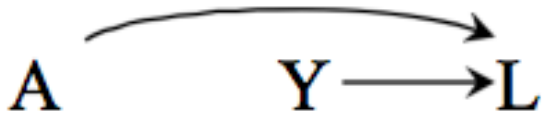
Basic patterns of causal relationships among three variables

Association and Causation



- L has a causal effect on both A and Y. A does not have a causal effect on Y. A depends on L and on *no other causes* of Y.
- L is a **common cause** to A and Y.
- A and Y are associated: having information about A improves our ability to predict Y, even though A does not have a causal effect on Y.
- E.g., A : carrying a lighter; Y : lung cancer; L : smoking

Association and Causation



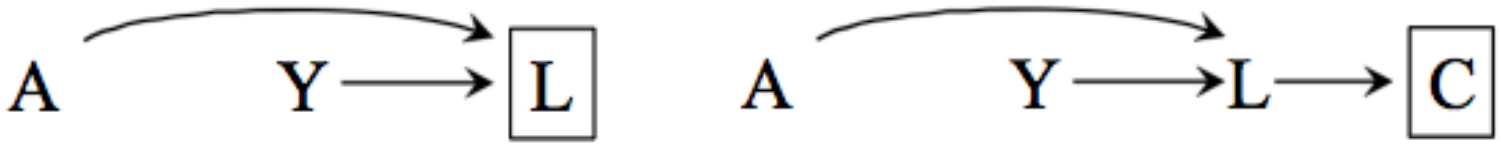
- Both A and Y have a causal effect on L. A does not have a causal effect on Y.
- L is a **common effect** of A and Y.
- L is called a **collider** on the path $A \rightarrow L \leftarrow Y$ and is said to *block* the path.
- A and Y are independent: the only path between them, $A \rightarrow L \leftarrow Y$, is blocked.
- E.g., A : family heart disease history; Y : smoking; L : heart disease

Association and Causation

Box indicates conditioning

- A and Y are conditionally independent after conditioning on B and L, even though they are marginally associated in both graphs.
- Conditioning on B and L *blocks* the paths $A \to B \to Y$ and $A \leftarrow L \to Y$.
- E.g. (left), A : smoking; B : tar deposits in lung; Y : lung cancer

Association and Causation



- A and Y are conditionally associated after conditioning on L and C, even though they are marginally independent.
- Conditioning on collider L or its descendent C *opens* the path $A \to L \leftarrow Y$, which is blocked otherwise.
- E.g. (right), A : family heart disease history; Y : smoking; L : heart disease; C : taking heart disease medication

Association and Causation

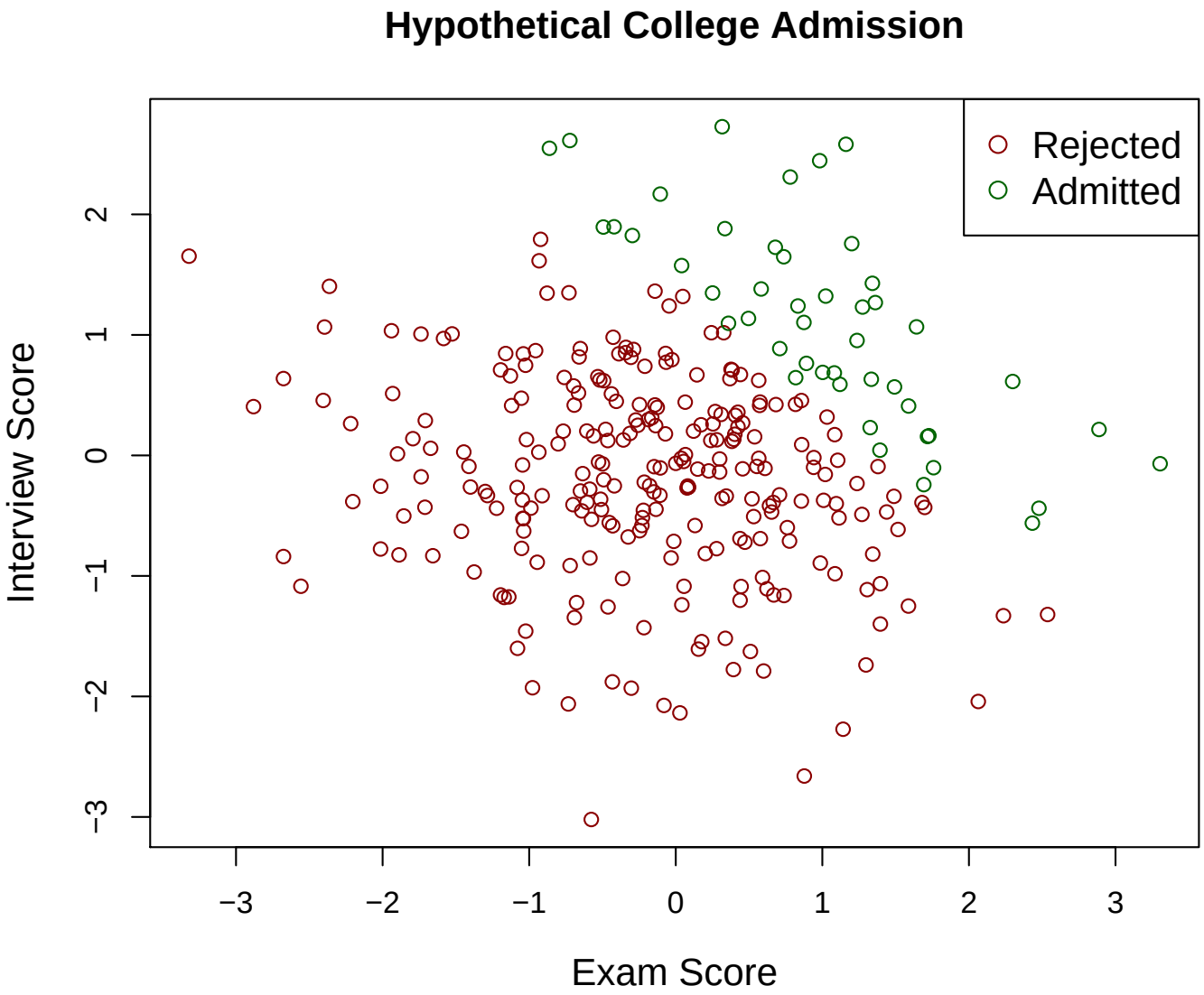
Conditioning on Colliders

Suppose a light bulb (C) is controlled by two on/off switches (A and B). The states of A and B are independent. C is lit up only if both A and B are in the "on" state. Then the causal diagram is:

$$A \rightarrow C \leftarrow B$$

- A and B are independent: the state of A tells you nothing about the state of B.
- A and B are dependent conditional on C: conditional on the light being off, A must be off if B is on, and vice versa.

Association and Causation



Association and Causation

In summary, there are three structural reasons why two variables may be associated:

1. One causes the other[5]
2. They share common causes
3. The analysis is conditioned on their common effects[6]

[5] either directly or through mediating variables.

[6] or the consequences of the common effects.

Confounding

- When two variables share common causes, they are correlated even if they do not cause each other. This makes it harder for us to learn the causal effect one has on the other. We call this problem **confounding**. The common causes are called **confounders**.

- Self-selection bias is an important type of confounding. Consider treatment x and outcome y. When x is selected based on the values of z, if z also has a causal effect on y, then z is a confounder and there is self-selection bias.

Causal Effect Learning

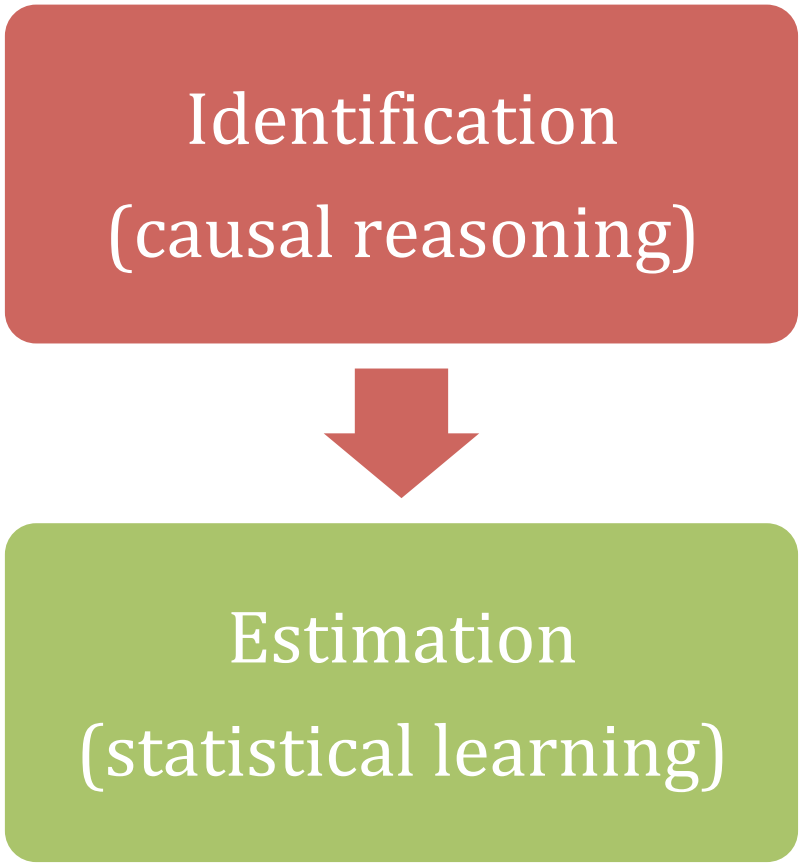
A basic strategy to learn the causal effect of x on y is to condition on their common causes[7] (while avoiding to condition on any of their common effects).

- Conditioning on common causes make two variables independent if they do not have direct causal effects on each other.
- Therefore, any association between two variables after their common causes have been conditioned on should be due to causation.

Therefore, if x, y, and all their common causes are observed, then we say the causal effect of x on y is identifiable.

[7] When we condition on a variable, we also say we **control for** the variable.

Causal Effect Learning: Two Stages



Causal Effect Learning: Identification

- The identification question: is it *possible* to learn the causal effect of interest from our observed variables[8]? What causal assumptions do we need in order to do so?

- More rigorously, suppose we are interested in learning $p(y|\text{do}(x))$ and v is the set of observed variables[9]. The identification problem is whether we can use $p(v)$ to (uniquely) determine $p(y|\text{do}(x))$ – equivalently, whether $p(y|\text{do}(x))$ can be expressed uniquely as a function of $p(v)$.

[8] i.e., given infinite data on our observed variables – if we actually know their true distribution – can we learn our causal effect of interest?

[9] v can include $\{x, y, \ldots\}$.

Causal Effect Learning: Identification

- If we can express $p(y|\text{do}(x))$ in terms of $p(v)$ without making any parametric assumptions on the relationships among the variables, then we say $p(y|\text{do}(x))$ is **nonparametrically identified**.
  - Nonparametric identification is based on knowledge of the causal diagram *only* and does not rely on any statistical or functional form assumptions.
- If statistical and functional form assumptions are involved in expressing $p(y|\text{do}(x))$ in terms of $p(v)$, then $p(y|\text{do}(x))$ is **parametrically identified**.

Causal Effect Learning: Estimation

- Once we have expressed $p(y|\text{do}(x))$ in terms of $p(v)$, say $p(y|\text{do}(x)) = g(p(v))$, then we can estimate $g(p(v))$ from the observed data $\mathcal{D} \sim p(v)$ using any appropriate statistical models – parametric or nonparametric[10].
- The estimation question: how to learn the identified causal effect from *finite* sample.
- Herein lies the connection between statistical learning and causal effect learning: once we have established identification using causal reasoning based on causal diagrams, we are left with a pure statistical learning problem.

[10] Hence, we could have nonparametrically identified–nonparametrically estimated causal effect, nonparametrically identified–parametrically estimated causal effect, etc.

The Back-Door Criterion

- The **back-door criterion** provides *sufficient* conditions for the *nonparametric* identification of causal effects.
- A **back-door path** is a path between treatment x and outcome y that has an arrow *into* x.
  - These are the paths that, if left open, induce association between x and y that is not a result of x causing y.
- A set of variables z satisfies the back-door criterion if (i) conditioning on z blocks every back-door path from x to y, and (ii) no variable in z is a descendant of x.

The Back-Door Criterion

If we observe a set of variables z that meets the back-door criterion, then $p(y|\text{do}(x))$ is nonparametrically identifiable[11]:

$$p(y|\text{do}(x), z) = p(y|x, z) \tag{3}$$

$$p(y|\text{do}(x)) = \int p(y|x, z)\, p(z)\, dz$$

[11] Note that in general,

$$p(y|\text{do}(x)) = \int p(y|x, z)\, p(z)\, dz$$

$$\neq \int p(y|x, z)\, p(z|x)\, dz = p(y|x)$$

The Back-Door Criterion

- x is said to be **exogenous** to y if there is no open back-door path from x to y, in which case $p\left(y|\text{do}\left(x\right)\right) = p\left(y|x\right)$.
- x is **conditionally exogenous** to y if x is exogenous to y after conditioning on z, in which case $p\left(y|\text{do}\left(x\right), z\right) = p\left(y|x, z\right)$.

The Back-Door Criterion

Let $s^{\mathscr{Y}}$ be the set of direct causes of y other than x. Condition (i) of the back-door criterion is equivalent to the requirement that $s^{\mathscr{Y}} \perp x \,|\, z$.

- If there exists an open back-door path from x to a variable $s \in s^{\mathscr{Y}}$, such that $s \not\perp x$, then there exists an open back-door path from x to y.
- Therefore, if conditioning on z blocks all back-door paths from x to y, then we have $s \perp x \,|\, z \;\; \forall s \in s^{\mathscr{Y}}$.

Therefore, the back-door criterion can be equivalently stated as requiring that conditional on z, no direct causes of y are correlated with x.

Equivalently, x is exogenous to y conditioning on z if $s^{\mathscr{Y}} \perp x \,|\, z$. Otherwise, x is endogenous[12].

[12] This is the way the econometrics literature defines endogeneity. See Appendix for a discussion on the econometric approach to causal reasoning and its limitations.

The Back-Door Criterion

The back-door criterion shows that we do not need to observe and condition on all confounders, but only a *(minimally) sufficient* set of variables that renders all back-door paths blocked[13].

[13] When all common causes are fully observed, we say that there is **no unmeasured confounding**, or that we have **selection on observables**. As the back-door criterion makes clear, this is sufficient but not necessary for causal effect identification.

Causal Effect Estimation under Sufficient Control for Confounding

Assume we observe a set of variables z that satisfies the back-door criterion. Then:

$$\begin{aligned} E\left[y|\text{do}\left(x=a\right)\right] &= \int E\left[y|x=a,z\right]p\left(z\right)dz \\ &\approx \frac{1}{N}\sum_{i} E\left[y_i|\, x_i=a,z_i\right] \end{aligned} \tag{4}$$

Therefore, if we can estimate $E\left[y|x,z\right]$, then we have an estimate of $E\left[y|\text{do}\left(x\right)\right]$.

Causal Effect Estimation under Sufficient Control for Confounding

Once we have an estimate of $E\left[y|\text{do}\left(x\right)\right]$, we can use it to compute the average treatment effect:

$$\text{ATE}\left(x\right)=\frac{dE\left[y|\text{do}\left(x\right)\right]}{dx}$$

If $x\in\left\{0,1\right\}$ is a binary treatment, then

$$\text{ATE}=E\left[y|\text{do}\left(x=1\right)\right]-E\left[y|\text{do}\left(x=0\right)\right]$$

Regression for Causal Inference

- With a sufficient set of observed variables z that controls for confounding, the causal effect learning problem boils down to the statistical learning problem of estimating the target function $E[y|x,z]$.
- This can be achieved using *any* appropriate statistical and machine learning models.

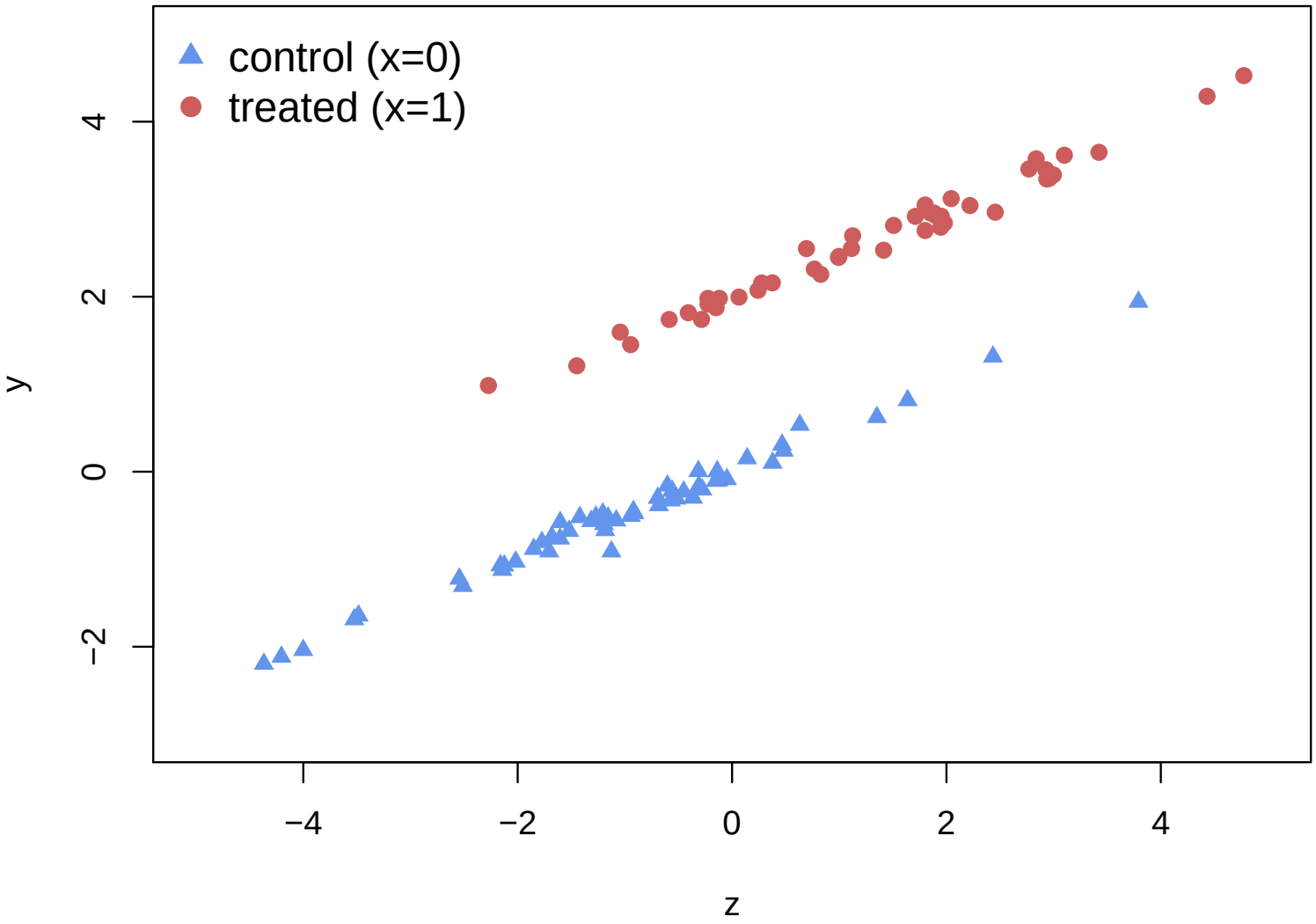
Simulation 1

Let's generate some data with binary treatment x and outcome y:

$$z \sim \mathcal{N}(0, 4)$$
$$x = \text{Bernoulli}\left((1 + \exp(-z))^{-1}\right)$$
$$y = 0.5z + 2x + e, \quad e \sim \mathcal{N}(0, 0.01)$$

```
# Simulation
require(sigmoid)
n = 100
z = 2*rnorm(n)
x = rbinom(n,1,sigmoid(z))
y = 0.5*z + 2*x + 0.1*rnorm(n)
```

Simulation 1



Simulation 1

Here the the treatment effect is homogeneous: $\tau = 2$.

Naive comparison of $E[y|x = 1]$ and $E[y|x = 0]$ gives biased estimate:

```
mean(y[x==1]) - mean(y[x==0])
```

```
## [1] 3.107458
```

Need to control for confounder z ...

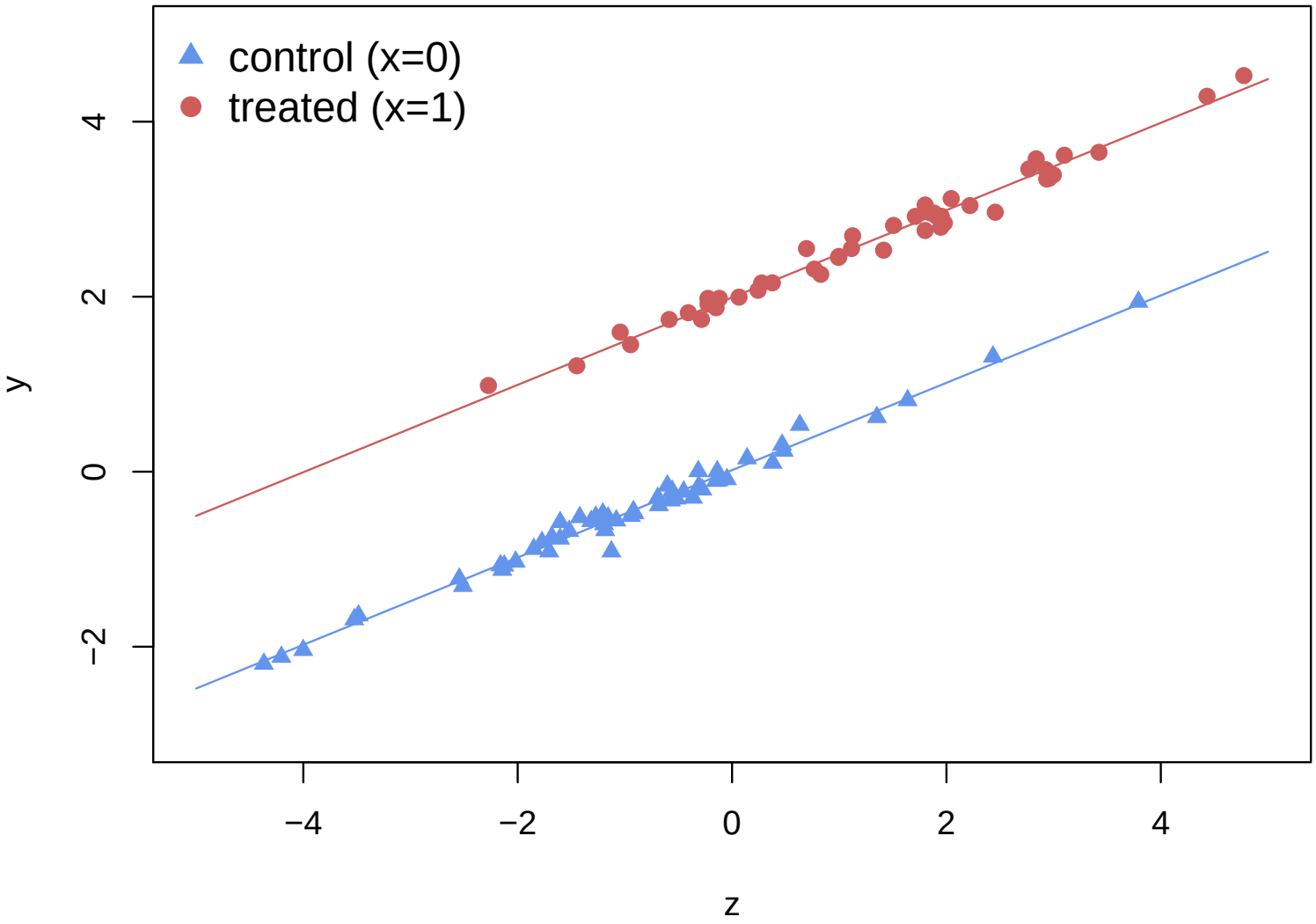
Simulation 1

Linear regression:

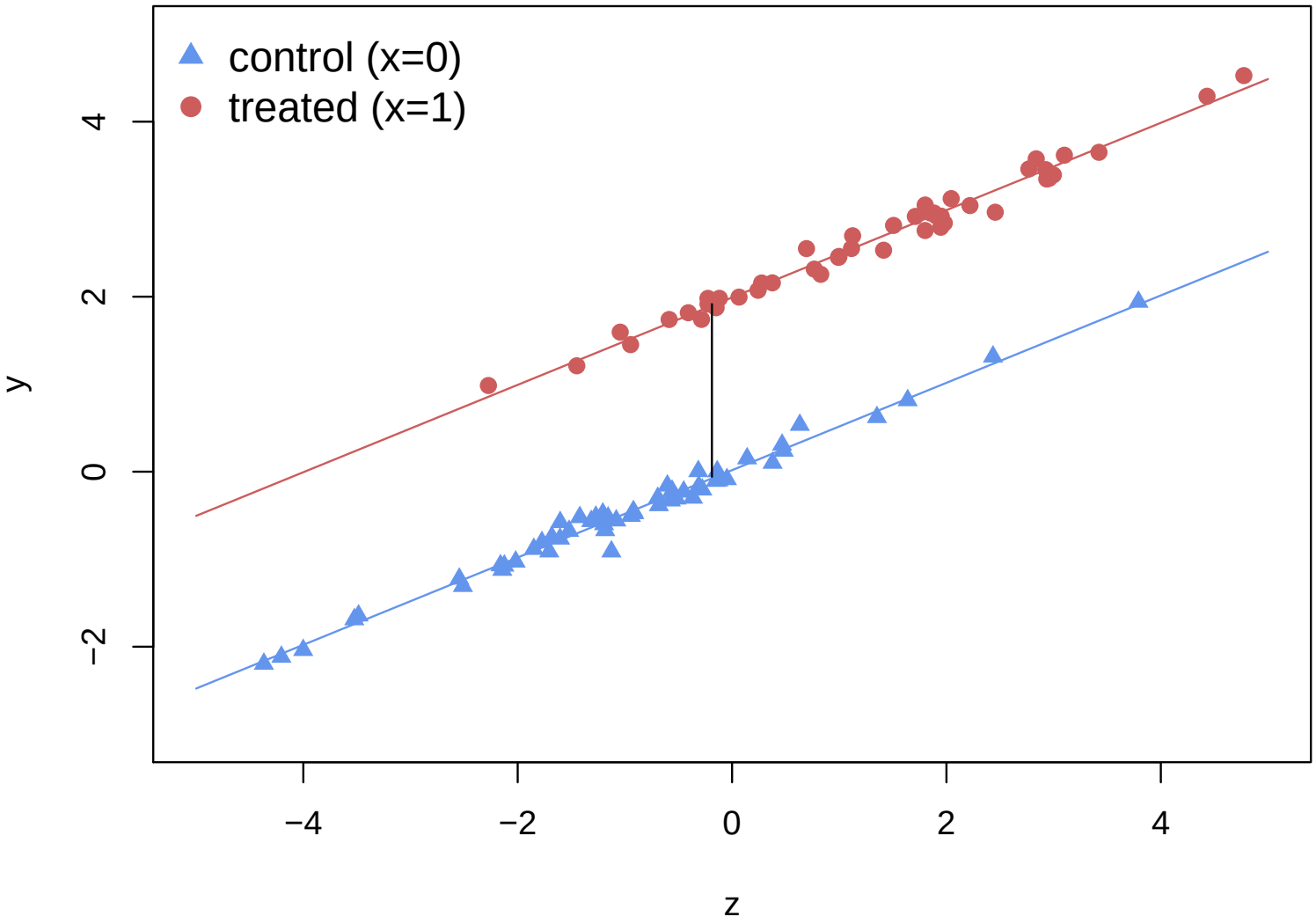
$$y = \beta_0 + \beta_1 x + \beta_2 z + e \tag{5}$$

```
require(AER)
data = data.frame(y=y,x=x,z=z)
fit = lm(y~.,data)
coeftest(fit)

##
## t test of coefficients:
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.0177336  0.0149535  1.1859   0.2386
## x           1.9727686  0.0248930 79.2499   <2e-16 ***
## z           0.4990824  0.0065954 75.6710   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simulation 1



Simulation 1



Simulation 1

In model (5), $\widehat{\beta_1}$ *is* our estimate of the ATE:

$$\begin{aligned}
\widehat{\text{ATE}} &= \widehat{E}\left[y|\text{do}\left(x=1\right)\right] - \widehat{E}\left[y|\text{do}\left(x=0\right)\right] \\
&= \int \left(\widehat{E}\left[y|x=1,z\right] - \widehat{E}\left[y|x=0,z\right]\right) p\left(z\right) dz \\
&= \int \left[\left(\widehat{\beta_0} + \widehat{\beta_1} + \widehat{\beta_2} z\right) - \left(\widehat{\beta_0} + \widehat{\beta_2} z\right)\right] p\left(z\right) dz \\
&= \widehat{\beta_1}
\end{aligned}$$

Simulation 1

In general, if the regression model for $E[y|x,z]$ is additive in x:

$$E[y|x,z] = \phi_1(x) + \phi_2(z) \tag{6}$$

, i.e., if we assume *no interaction* between x and z, then

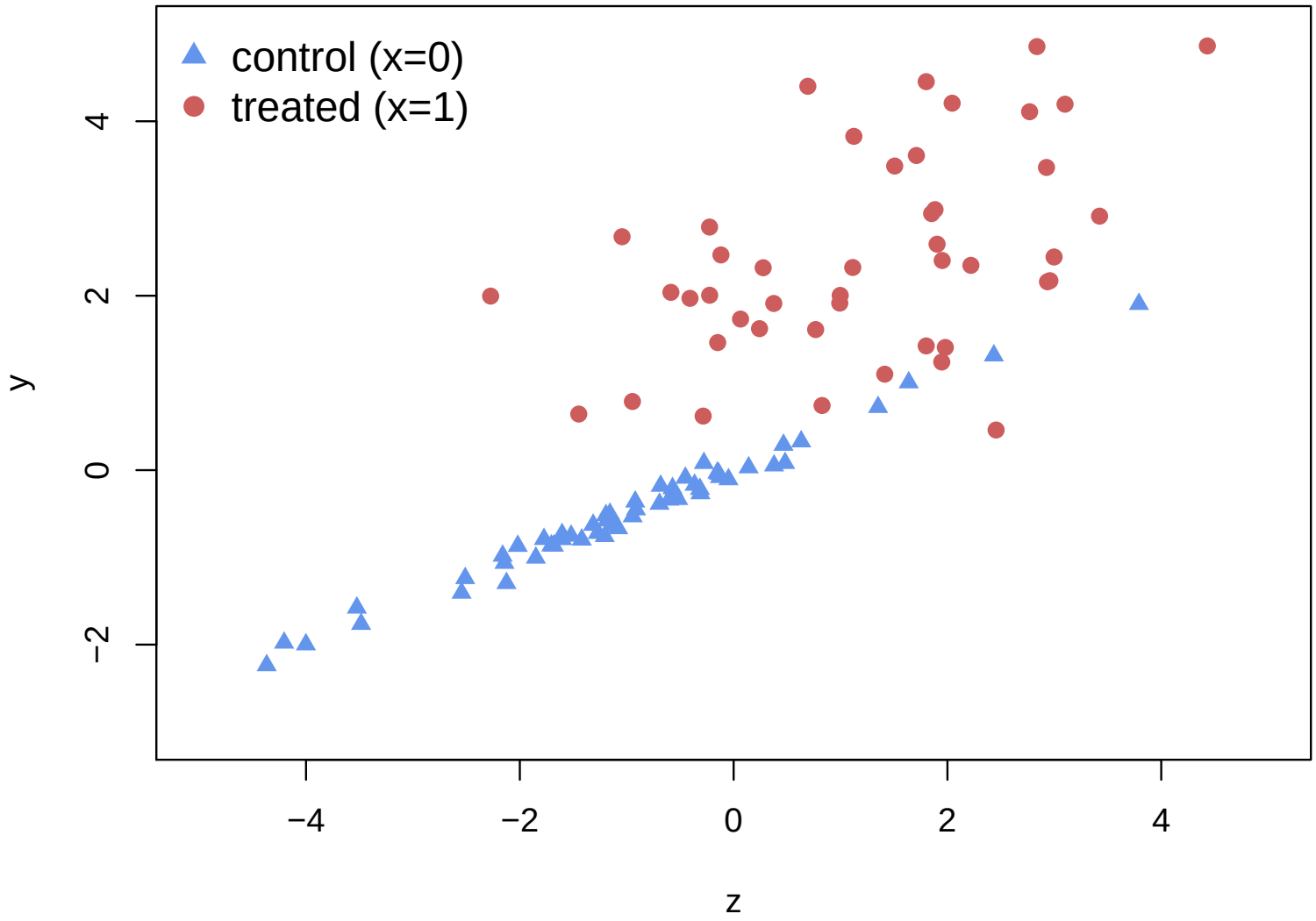
$$\begin{aligned}
\text{ATE}(x) &= \frac{dE[y|\text{do}(x)]}{dx} \\
&= \int \frac{\partial E[y|x,z]}{\partial x} p(z)\, dz \\
&= \int \phi_1'(x)\, p(z)\, dz \\
&= \phi_1'(x) = \frac{\partial E[y|x,z]}{\partial x}
\end{aligned}$$

Simulation 2

$$z \sim \mathcal{N}(0, 4)$$
$$x = \text{Bernoulli}\left((1 + \exp(-z))^{-1}\right)$$
$$y = 0.5z + \alpha x + e, \quad \alpha \sim \mathcal{N}(2, 1), \quad e \sim \mathcal{N}(0, 0.01)$$

```
# Simulation
n = 100
z = 2*rnorm(n)
x = rbinom(n,1,sigmoid(z))
y = 0.5*z + (2+1*rnorm(n))*x + 0.1*rnorm(n)
```

Simulation 2



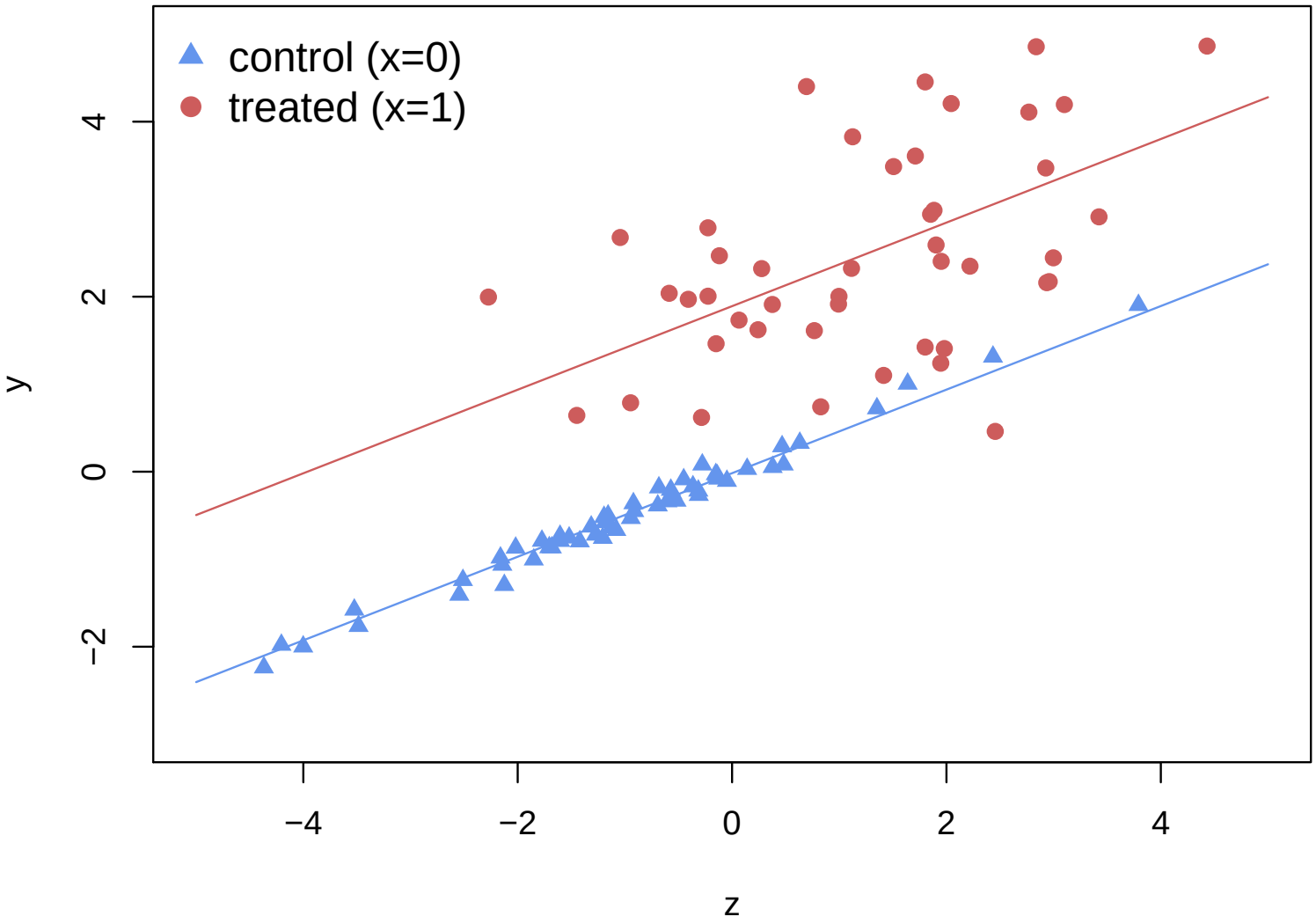
Simulation 2

Linear regression:

$$y = \beta_0 + \beta_1 x + \beta_2 z + e \tag{7}$$

```
data = data.frame(y=y,x=x,z=z)
fit = lm(y~.,data)
coeftest(fit)

##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.01704    0.10899 -0.1564   0.8761
## x            1.90869    0.18143 10.5203   <2e-16 ***
## z            0.47746    0.04807  9.9326   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simulation 2



Simulation 2

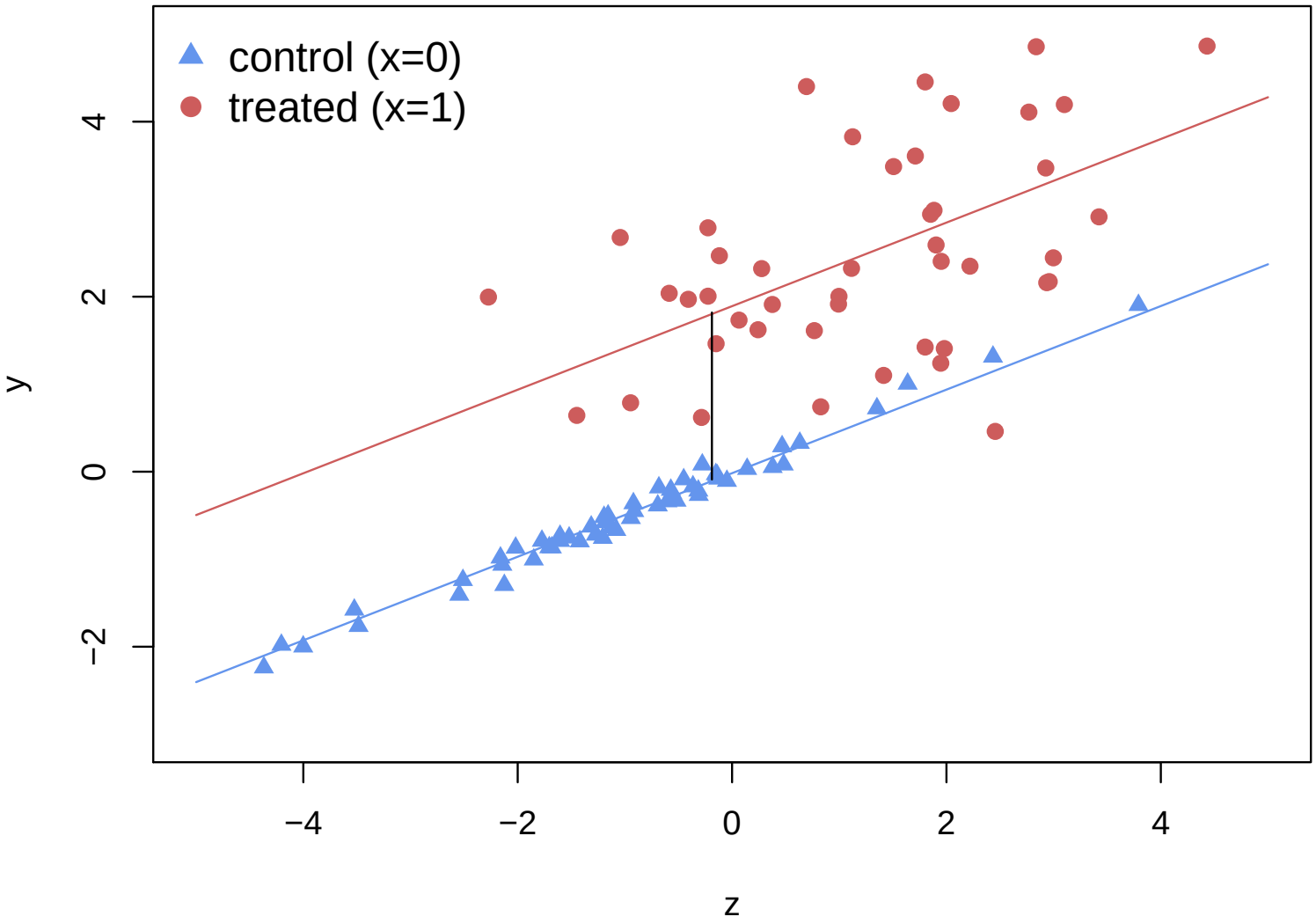
Here the the treatment effect is heterogeneous: $\tau_i = \alpha_i$, but does not vary with z.

The average treatment effect

$$\begin{aligned} \text{ATE} &= E(\tau) = E(\alpha) \\ &= E[y|x=1,z] - E[y|x=0,z] \end{aligned}$$

Hence in (7), $\widehat{\beta_1}$ is again our estimate of the ATE.

Simulation 2

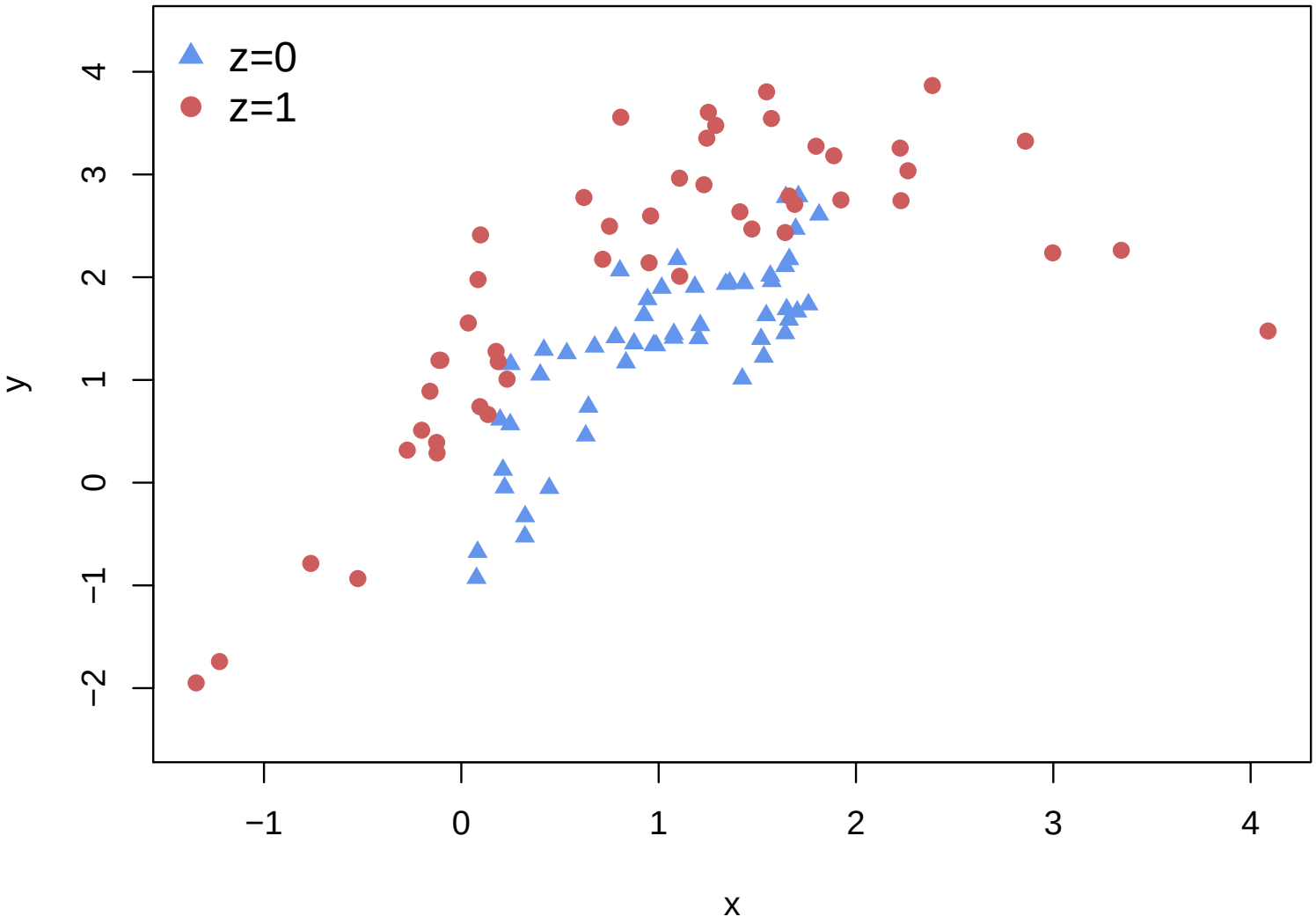


Simulation 3

$$z \sim \text{Bernoulli}\,(0.5)$$
$$x = z \times \mathcal{N}\,(0,1) + U\,(0,2)$$
$$y = z + 2x - 0.5x^2 + e, \quad e \sim \mathcal{N}\,(0, 0.25)$$

```
# Simulation
n = 100
z = rbinom(n,1,0.5)
x = 2*runif(n) + z*rnorm(n)
y = z + 2*x - 0.5*x^2 + 0.5*rnorm(n)
```

Simulation 3



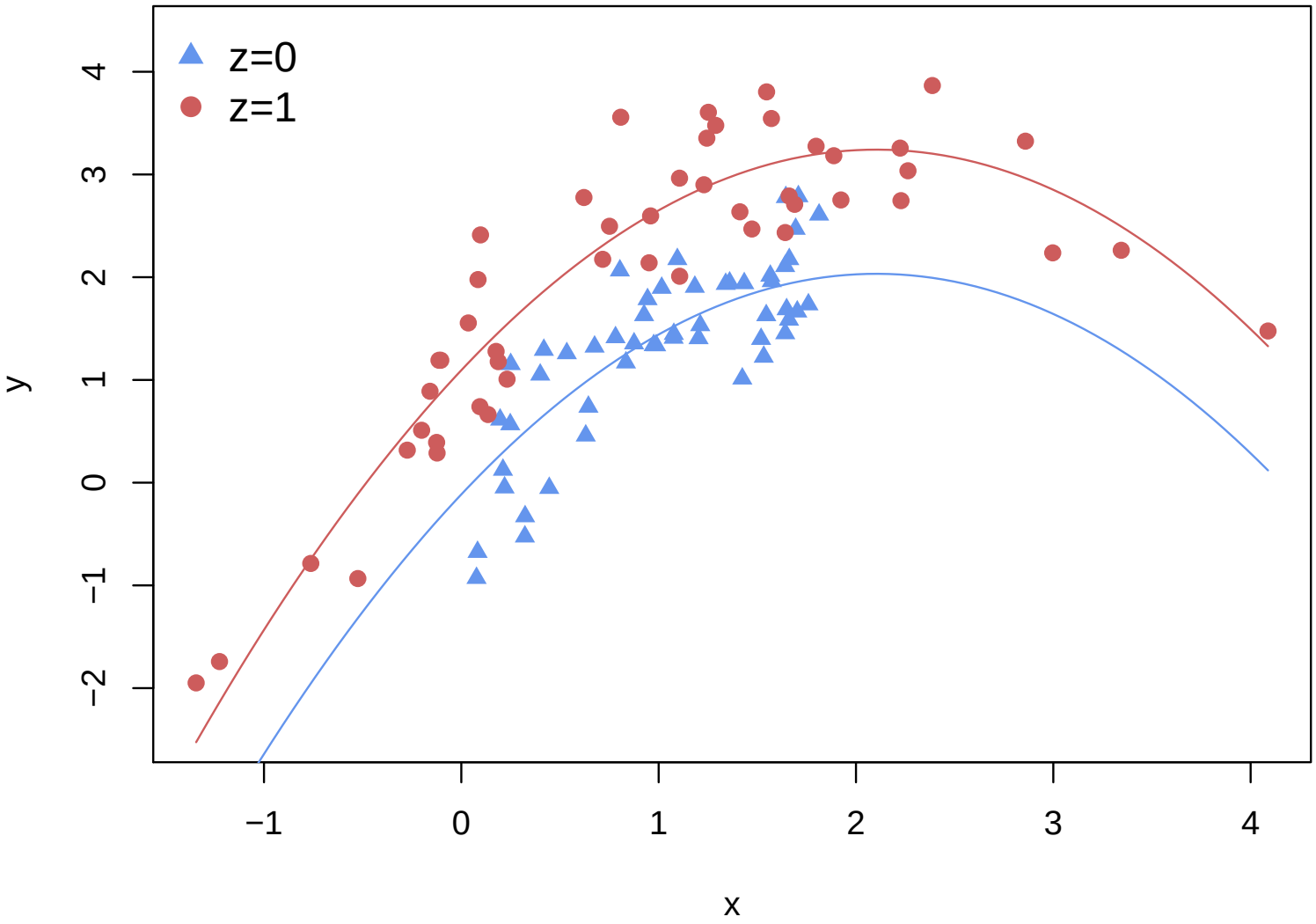
Simulation 3

Polynomial regression in x:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2 + \beta_3 z + e \tag{8}$$

```
fit = lm(y ~ poly(x,2,raw=T) + z)
coeftest(fit)

##
## t test of coefficients:
##
##                          Estimate Std. Error  t value Pr(>|t|)
## (Intercept)            -0.114061   0.098290  -1.1605   0.2487
## poly(x, 2, raw = T)1  2.042029   0.105225  19.4062   <2e-16 ***
## poly(x, 2, raw = T)2 -0.485534   0.039532 -12.2820   <2e-16 ***
## z                      1.208719   0.108421  11.1484   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simulation 3



Simulation 3

Here the treatment effect is homogeneous but non-constant (in x).

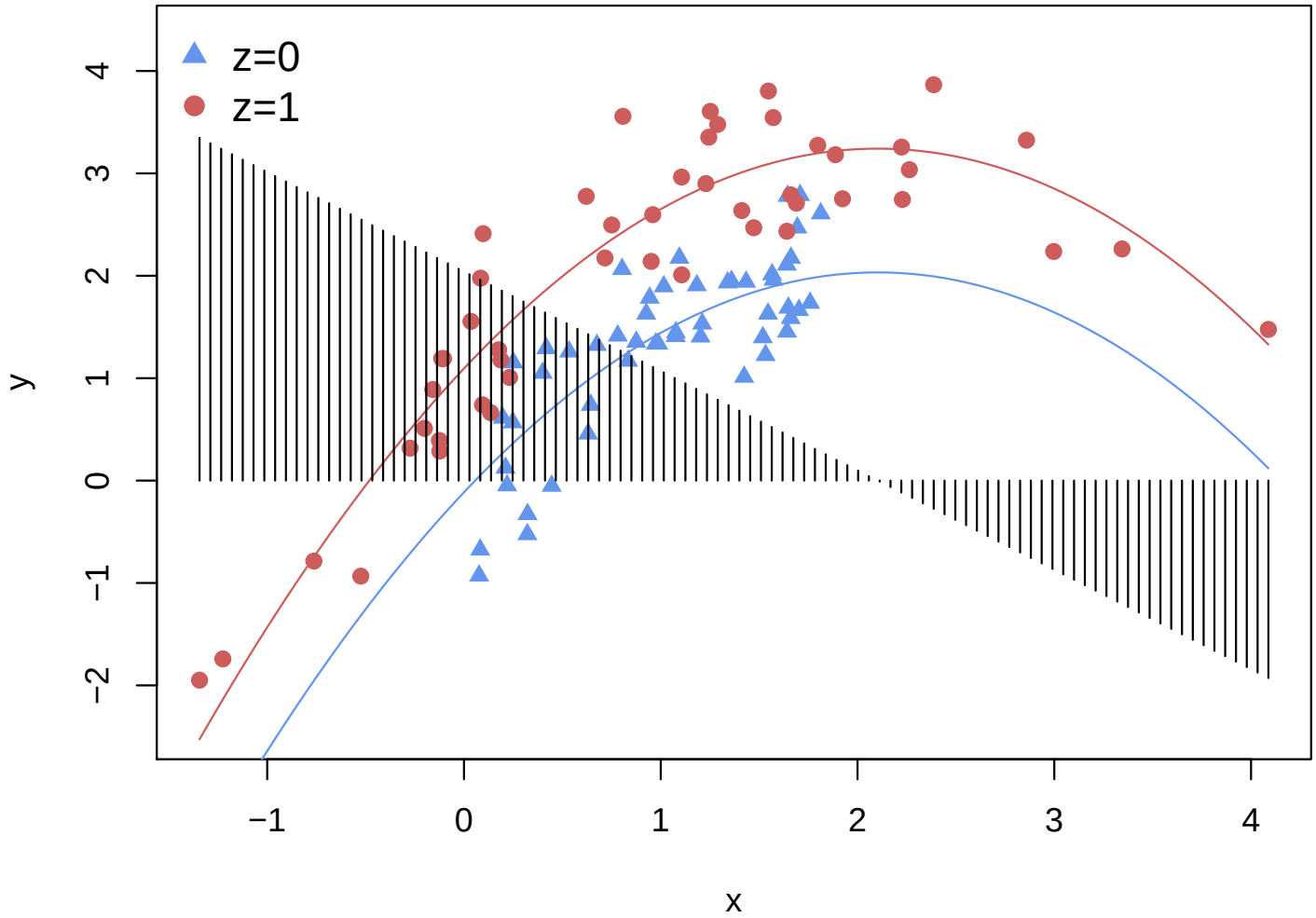
Given the additive structure of (6),

$$\text{ATE}(x) = \phi_1'(x) = 2 - x$$

Given (8),

$$\widehat{\text{ATE}}(x) = \widehat{\beta_1} + 2\widehat{\beta_2}x$$

Simulation 3



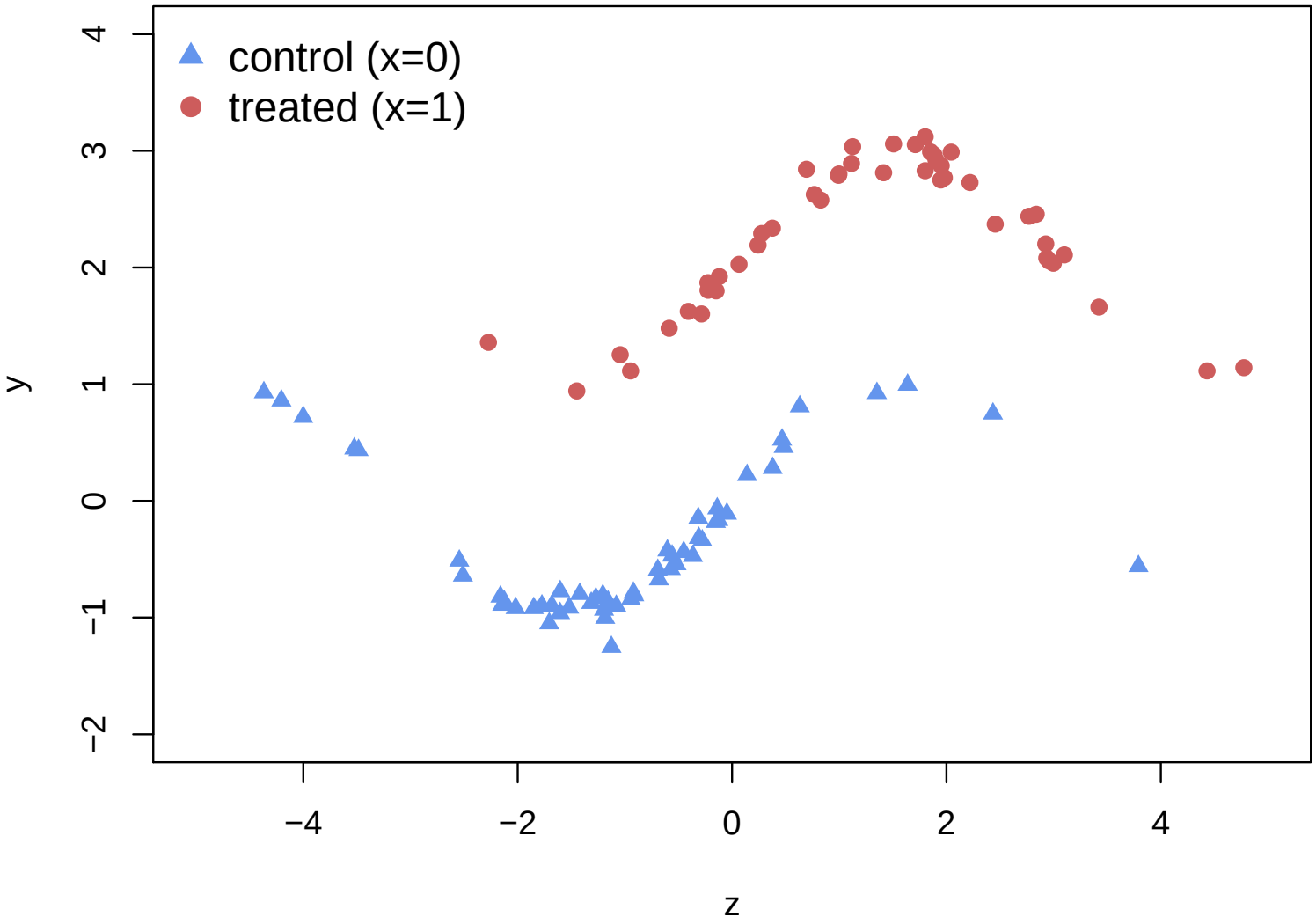
Vertical lines represent $\widehat{\beta}_1 + 2\widehat{\beta}_2 x$: the estimated ATE.

Simulation 4

$$z \sim \mathcal{N}(0, 4)$$
$$x = \text{Bernoulli}\left((1 + \exp(-z))^{-1}\right)$$
$$y = \sin(z) + 2x + e, \quad e \sim \mathcal{N}(0, 0.01)$$

```
# Simulation
n = 100
z = 2*rnorm(n)
x = rbinom(n,1,sigmoid(z))
y = sin(z) + 2*x + 0.1*rnorm(n)
```

Simulation 4



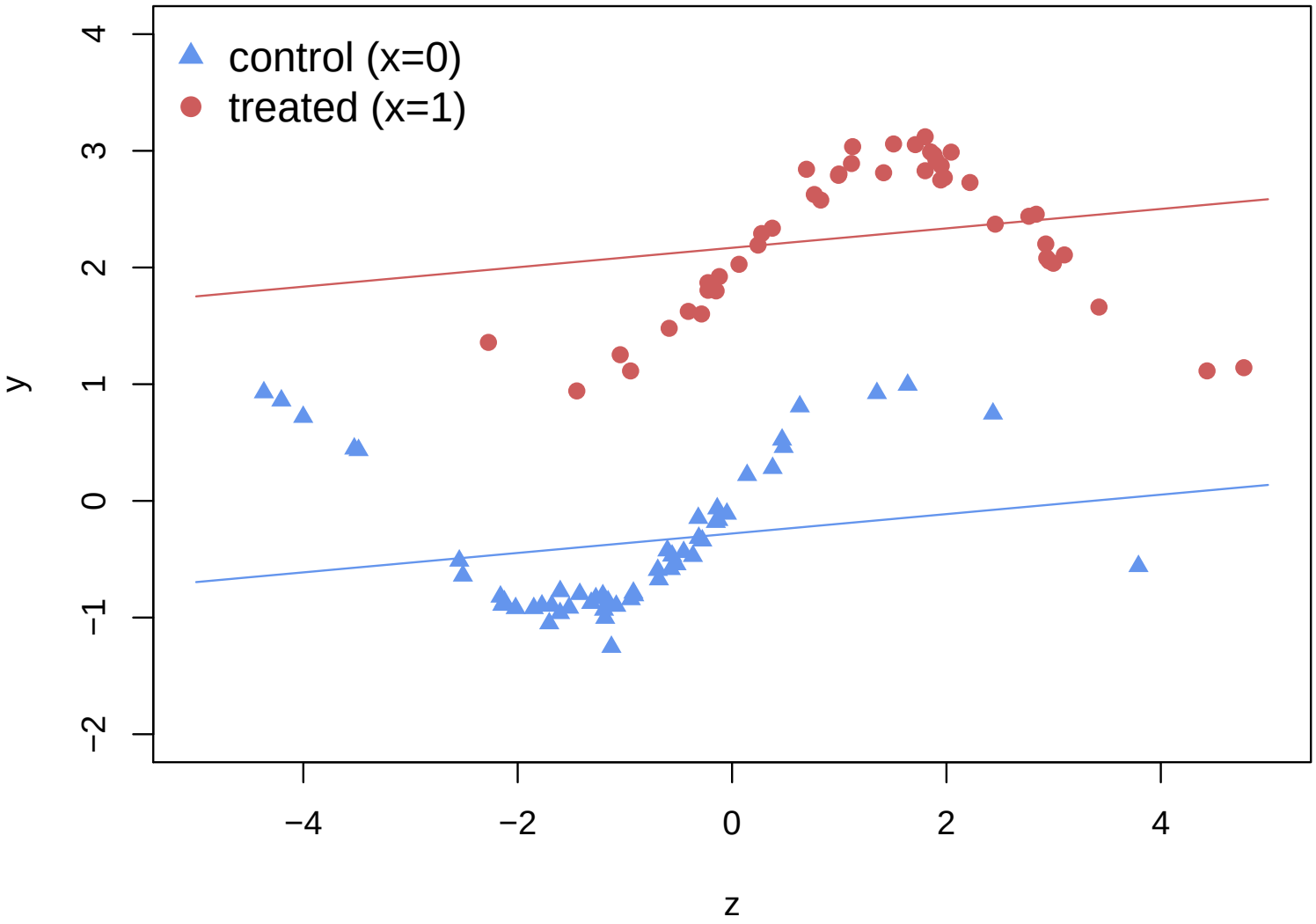
Simulation 4

Linear regression:

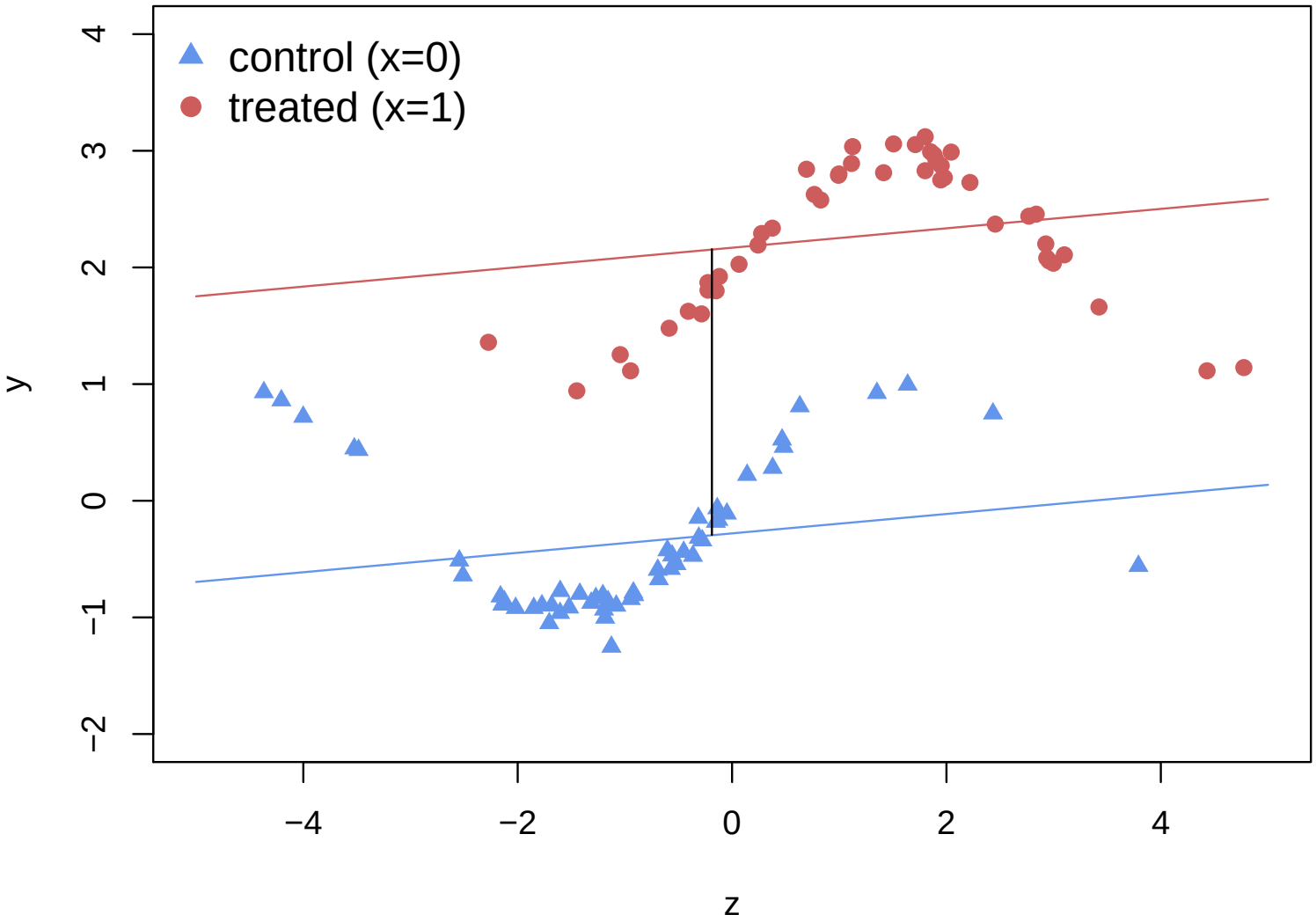
$$y = \beta_0 + \beta_1 x + \beta_2 z + e$$

```
data = data.frame(y=y,x=x,z=z)
fit = lm(y~.,data)
coeftest(fit)

##
## t test of coefficients:
##
##              Estimate Std. Error t value  Pr(>|t|)
## (Intercept) -0.279858   0.094458 -2.9628  0.003835 **
## x            2.448736   0.157244 15.5729 < 2.2e-16 ***
## z            0.083252   0.041662  1.9983  0.048487 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Simulation 4



Simulation 4



Simulation 4

Semi-parametric generalized additive model:

$$y = \beta_0 + \beta_1 x + g(z) + e$$

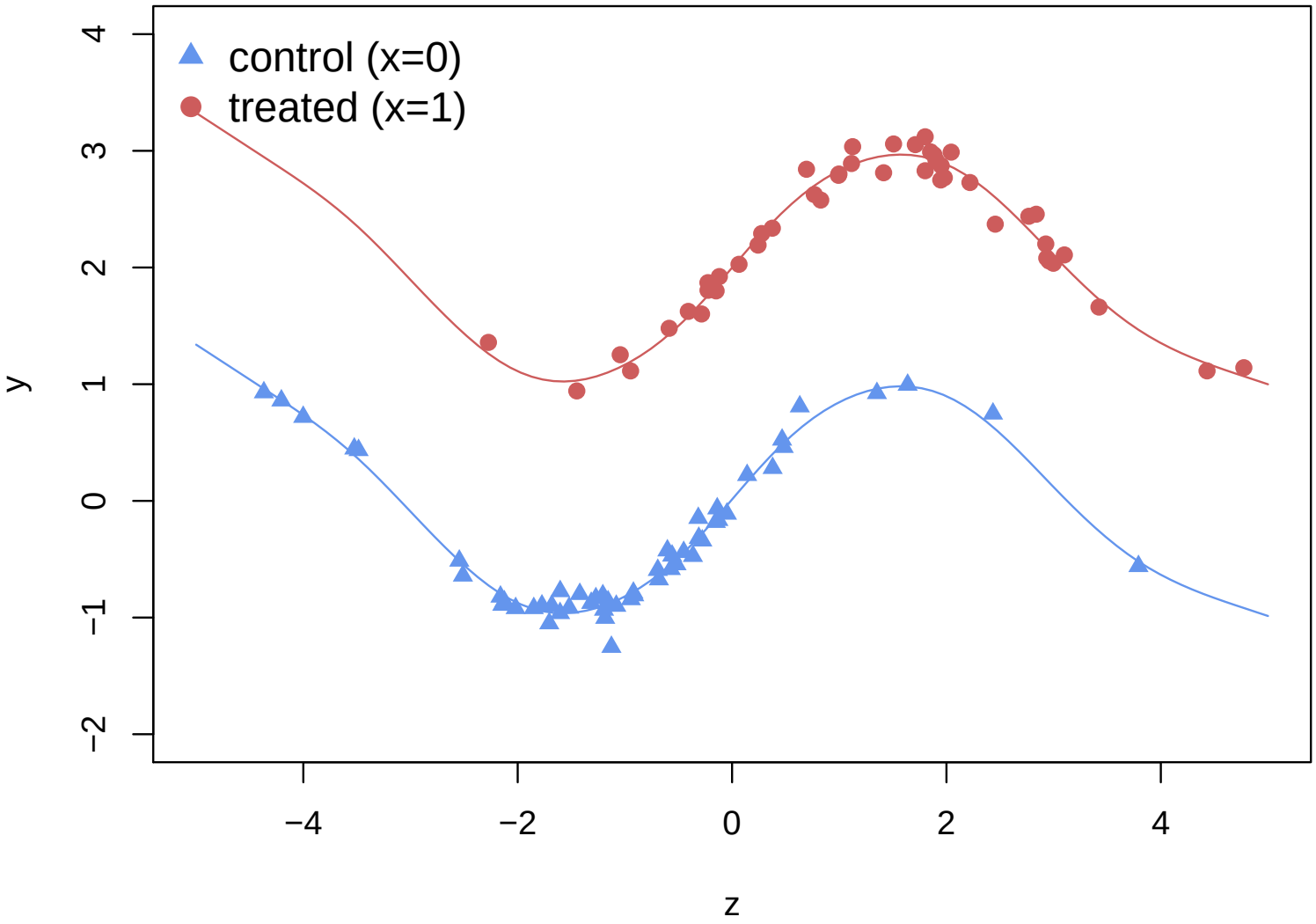
, where $g(z)$ is a smoothing spline.

```
library(mgcv)
fit = gam(y ~ x + s(z),data,family=gaussian)
```

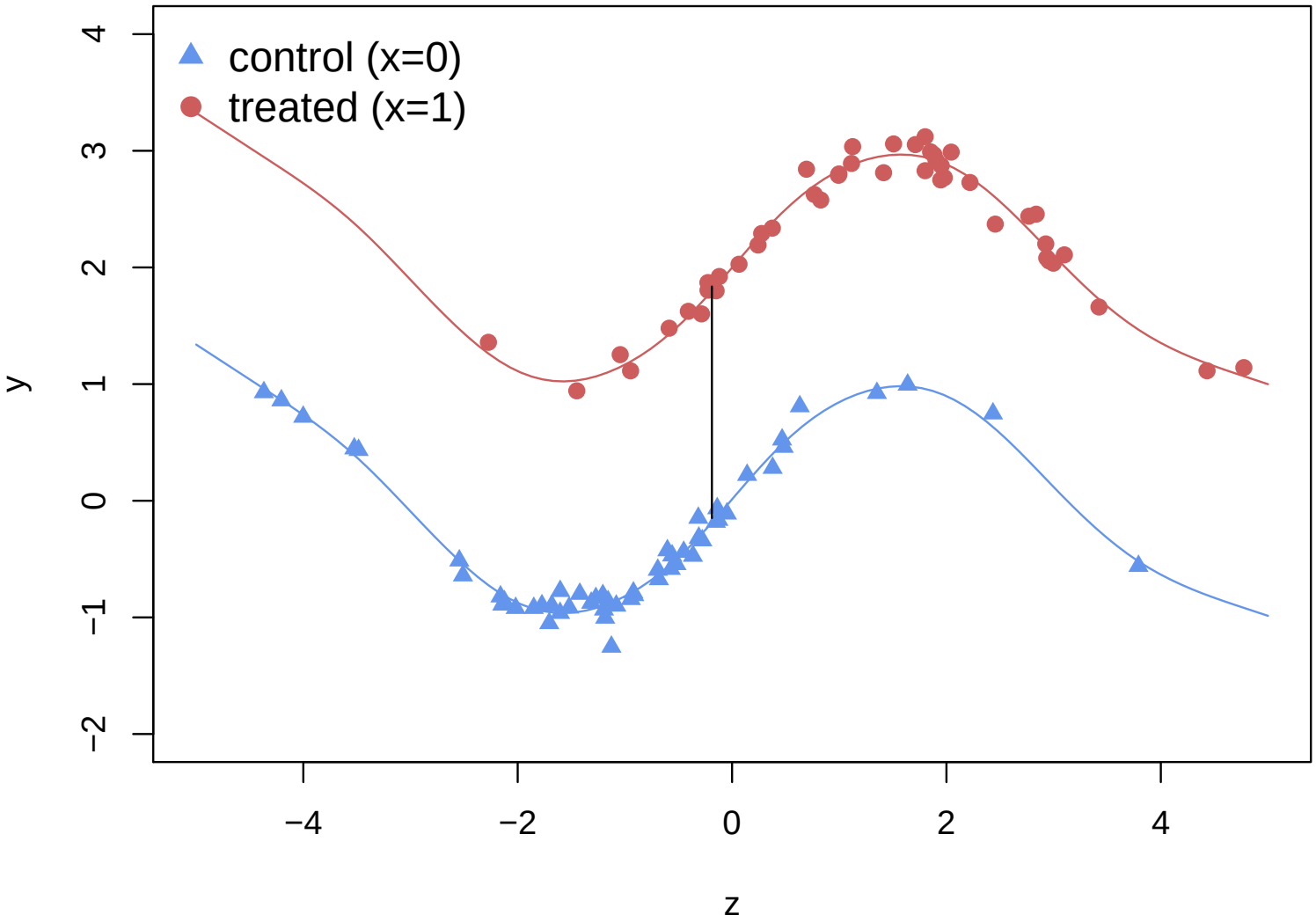
Simulation 4

```
summary(fit)

##
## Family: gaussian
## Link function: identity
##
## Formula:
## y ~ x + s(z)
##
## Parametric coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -0.06169    0.01589  -3.883 0.000197 ***
## x            1.98592    0.02665  74.509  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Approximate significance of smooth terms:
##        edf Ref.df     F p-value
## s(z) 8.455  8.914 422.5  <2e-16 ***
## ---
```

Simulation 4



Simulation 4

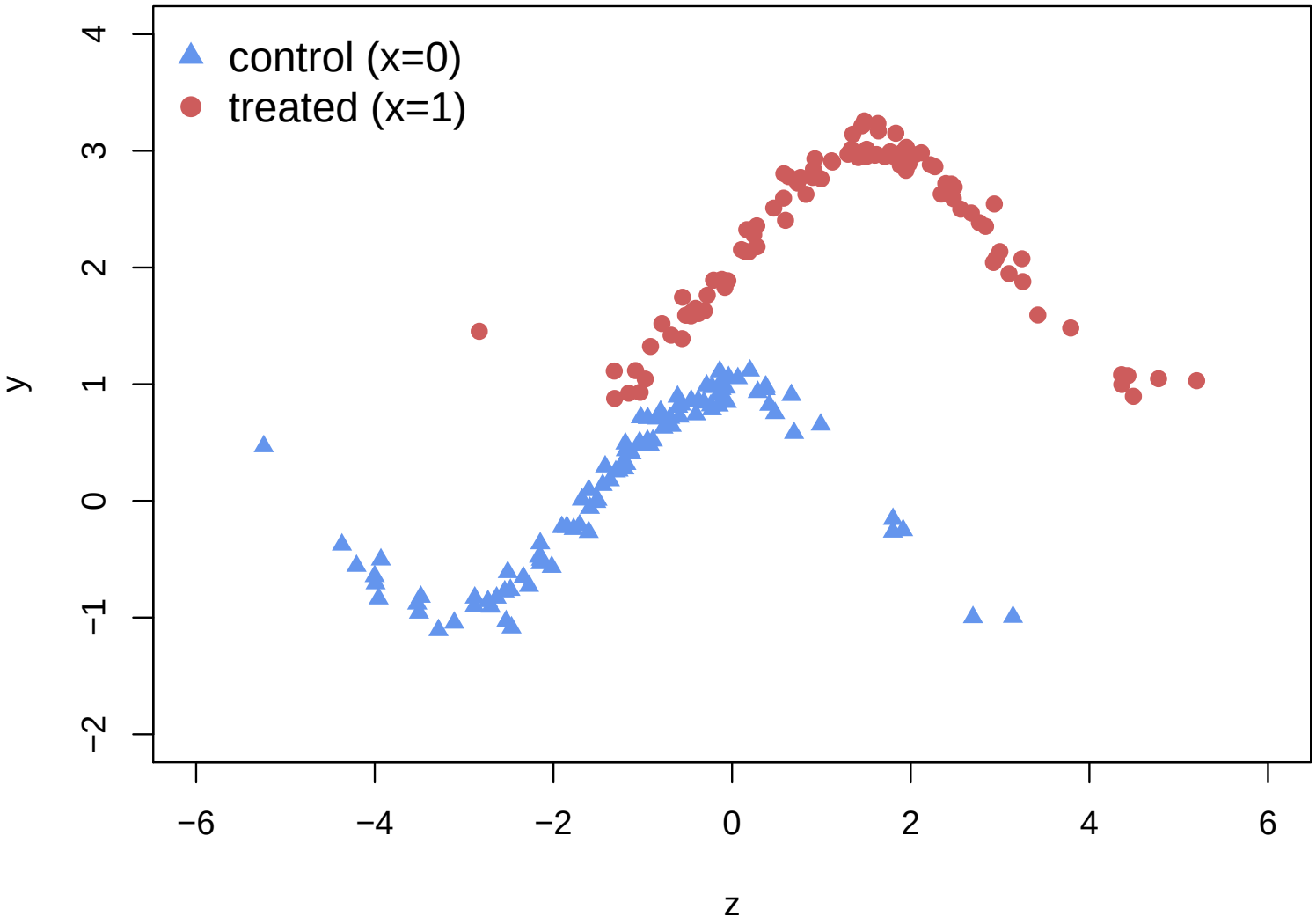


Simulation 5

$$
\begin{aligned}
z &\sim \mathcal{N}(0, 4) \\
x &= \text{Bernoulli}\left((1+\exp(-z))^{-1}\right) \\
\mathcal{Y}^0 &= \cos(z) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.01) \\
\mathcal{Y}^1 &= 2 + \sin(z) + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, 0.01) \\
y &= x\mathcal{Y}^1 + (1-x)\mathcal{Y}^0
\end{aligned}
$$

```
# Simulation
n = 200
z = 2*rnorm(n)
x = rbinom(n,1,sigmoid(z))
y0 = cos(z) + 0.1*rnorm(n)
y1 = 2 + sin(z) + 0.1*rnorm(n)
y = y0*(x==0) + y1*(x==1)
```

Simulation 5



Simulation 5

Here the treatment effect is heterogeneous and varies with z:

$$\begin{aligned}
\text{ATE} &= E\left[\mathcal{Y}^1 - \mathcal{Y}^0\right] \\
&= \int \left(E\left[y|x=1,z\right] - E\left[y|x=0,z\right]\right) p\left(z\right) dz \\
&\approx \frac{1}{N}\sum_i \left(E\left[y_i|\, x_i=1,z_i\right] - E\left[y_i|\, x_i=0,z_i\right]\right)
\end{aligned}$$

```
# ATE
ate = mean(y1 - y0)
ate

## [1] 1.869096
```

Simulation 5

To model the interaction between x and z, we can run the following linear regression:

$$y = \beta_0 + \beta_1 x + \beta_2 z + \beta_3 x \cdot z + e \tag{9}$$

Or more flexibly, we can run the following set of regressions[14]:

$$\begin{cases} y = \beta_0 + \beta_1 z + \epsilon & \text{if } x = 0 \\ y = \beta_2 + \beta_3 z + \varepsilon & \text{if } x = 1 \end{cases} \tag{10}$$

Estimating (10) $\Rightarrow \widehat{E}[y|x=0,z]$ and $\widehat{E}[y|x=1,z]$, from which we obtain:

$$\widehat{\text{ATE}} \approx \frac{1}{N} \sum_i \left(\widehat{E}[y_i | x_i = 1, z_i] - \widehat{E}[y_i | x_i = 0, z_i] \right)$$

[14] (9) and (10) are equivalent when x is binary.

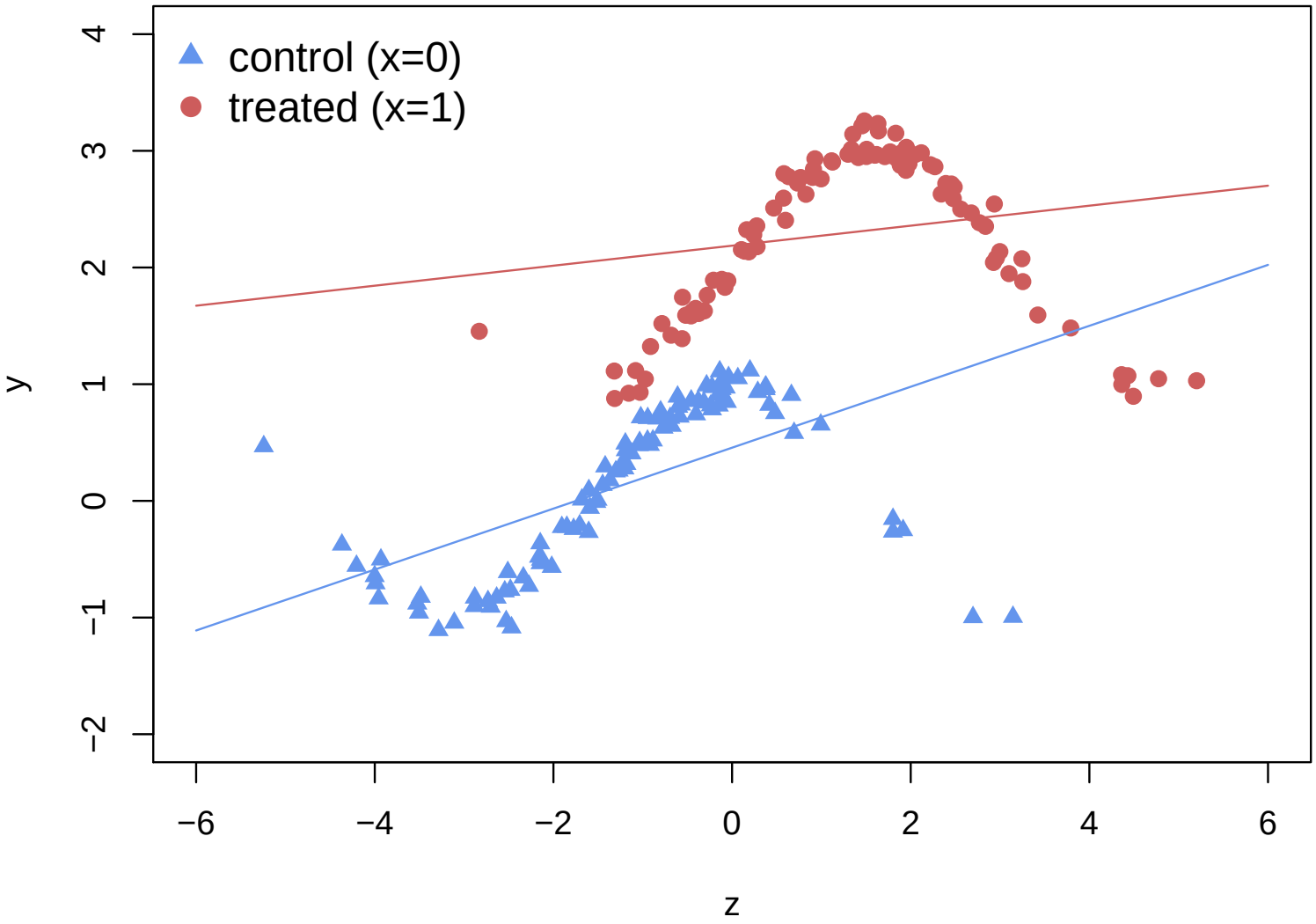
Simulation 5

```r
# Linear Regression
data = data.frame(y=y,x=x,z=z)
fit0 = lm(y ~ z,data,subset=(x==0))
fit1 = lm(y ~ z,data,subset=(x==1))

# Estimated ATE
y0hat = predict(fit0,data)
y1hat = predict(fit1,data)
atehat = mean(y1hat - y0hat)
atehat

## [1] 1.722792
```

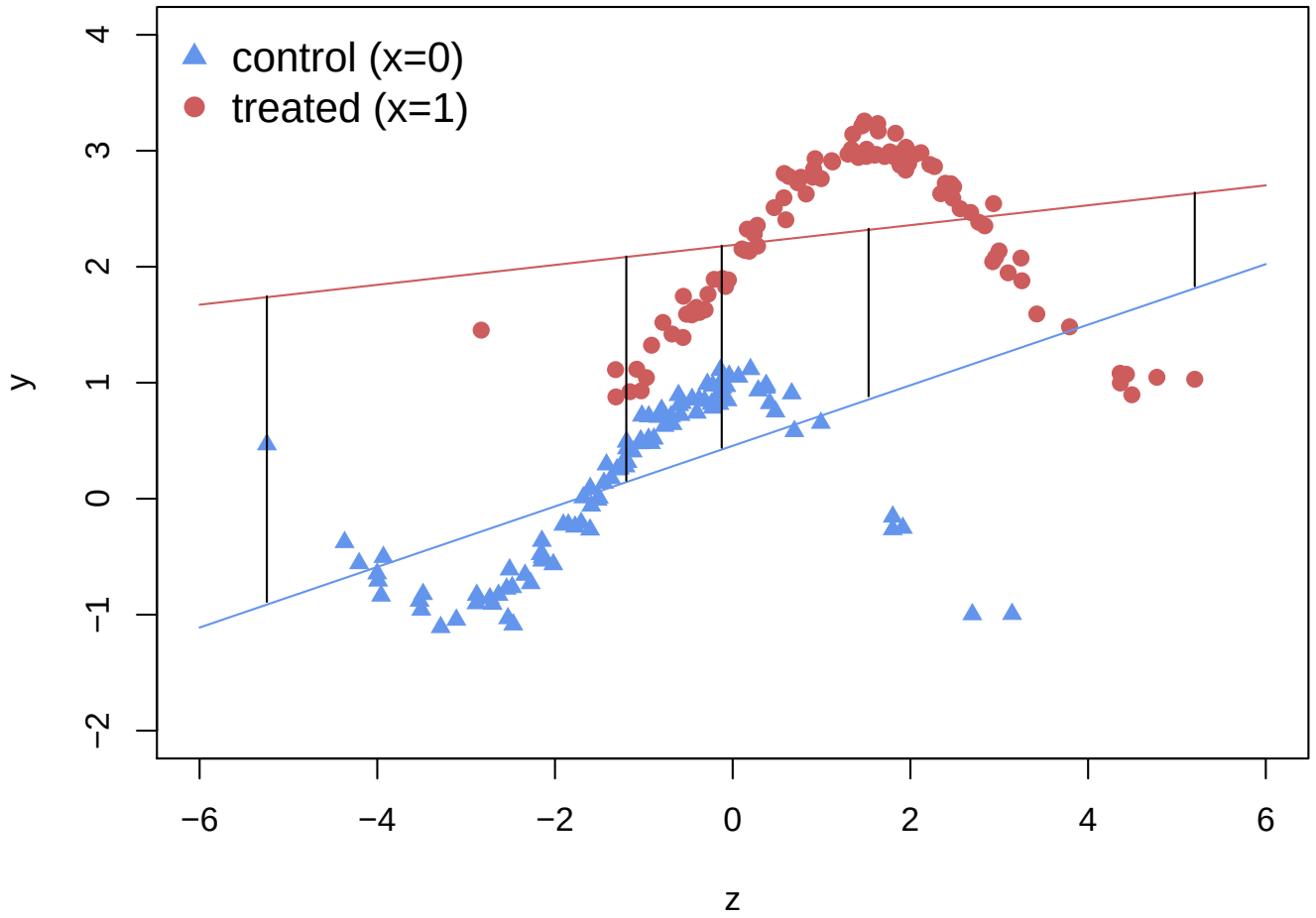
Simulation 5

```
coeftest(fit0)

##
## t test of coefficients:
##
##             Estimate Std. Error t value  Pr(>|t|)
## (Intercept) 0.455760   0.074658  6.1047 2.123e-08 ***
## z           0.261113   0.038104  6.8527 6.698e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

coeftest(fit1)

##
## t test of coefficients:
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 2.186996   0.090796  24.087  < 2e-16 ***
## z           0.085725   0.045190   1.897  0.06074 .
## ---
```

Simulation 5



Simulation 5



Vertical lines represent estimated ATEs conditional on z at 1%, 25%, 50%, 75%, and 99% percentiles.

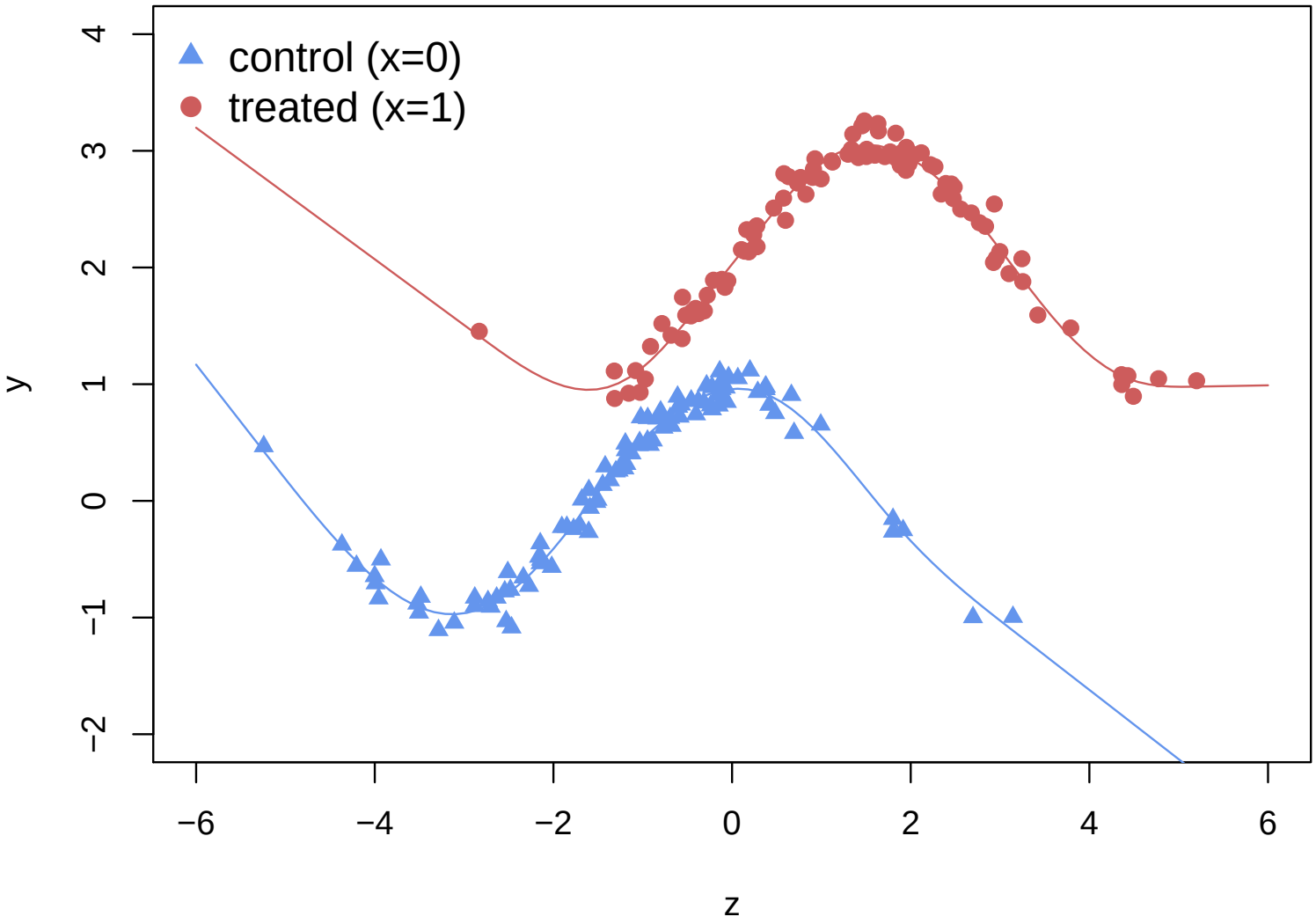
Simulation 5

Smoothing splines:

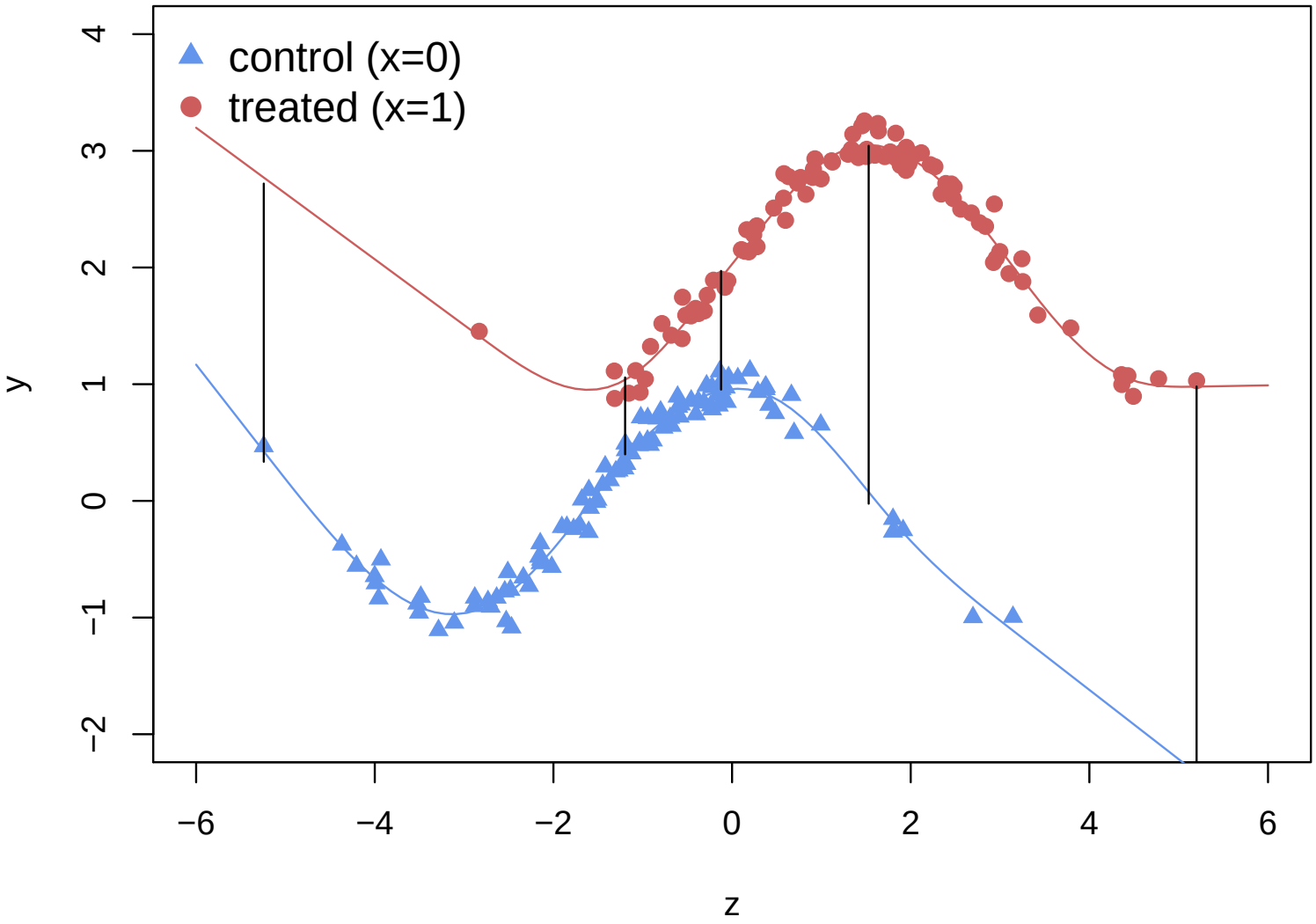
$$\begin{cases} y = g_1(z) + \epsilon & \text{if } x = 0 \\ y = g_0(z) + \varepsilon & \text{if } x = 1 \end{cases}$$

```
# Smoothing Splines
fit0 = gam(y ~ s(z),data,subset=(x==0),family=gaussian)
fit1 = gam(y ~ s(z),data,subset=(x==1),family=gaussian)

# Estimated ATE
y0hat = predict(fit0,data)
y1hat = predict(fit1,data)
atehat = mean(y1hat - y0hat)
atehat

## [1] 1.877149
```

Simulation 5



Simulation 5

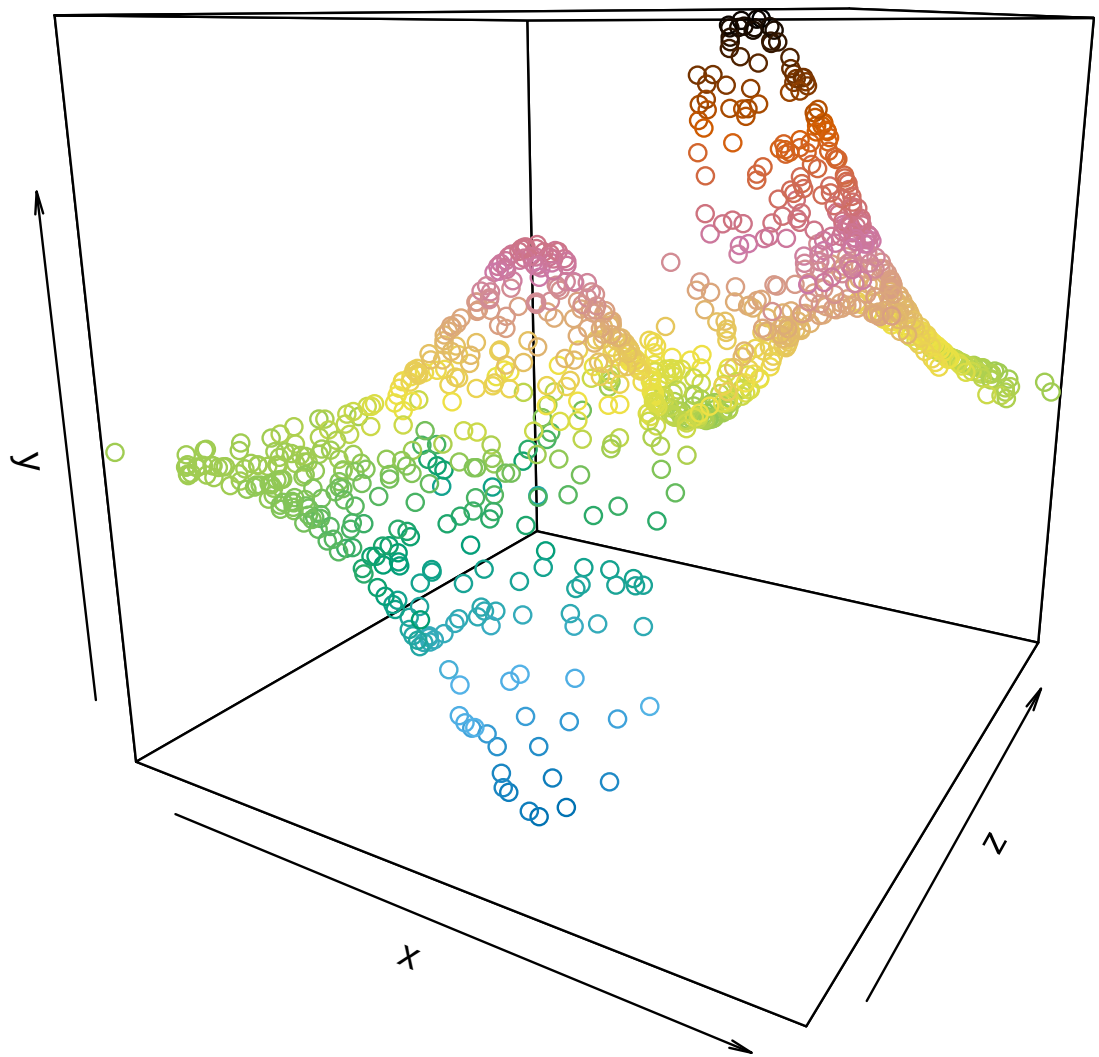


Simulation 6

$$z \sim U\left(-\frac{5}{2}, \frac{5}{2}\right)$$
$$x = \frac{1}{2}z + \frac{1}{2}\mathcal{N}(0,1)$$
$$y = 3(1-x)^2 \exp\left(-x^2 - (1+z)^2\right) - 10\left(\frac{1}{5}x - x^3 - z^5\right)$$
$$\times \exp\left(-x^2 - z^2\right) - \frac{1}{3}\exp\left(-(1+x)^2 - z^2\right)$$

```
# Simulation
n = 1000
z = 5*runif(n) - 2.5
x = 0.5*z + 0.5*rnorm(n)
g = expression((3*(1-x)^2)*exp(-(x^2)-(z+1)^2)-
               10*(x/5-x^3-z^5)*exp(-x^2-z^2)-1/3*exp(-(x+1)^2-z^2))
y = eval(g)
```

Simulation 6



Simulation 6

Here the treatment effect is both heterogeneous and varies with z, and non-constant in x:

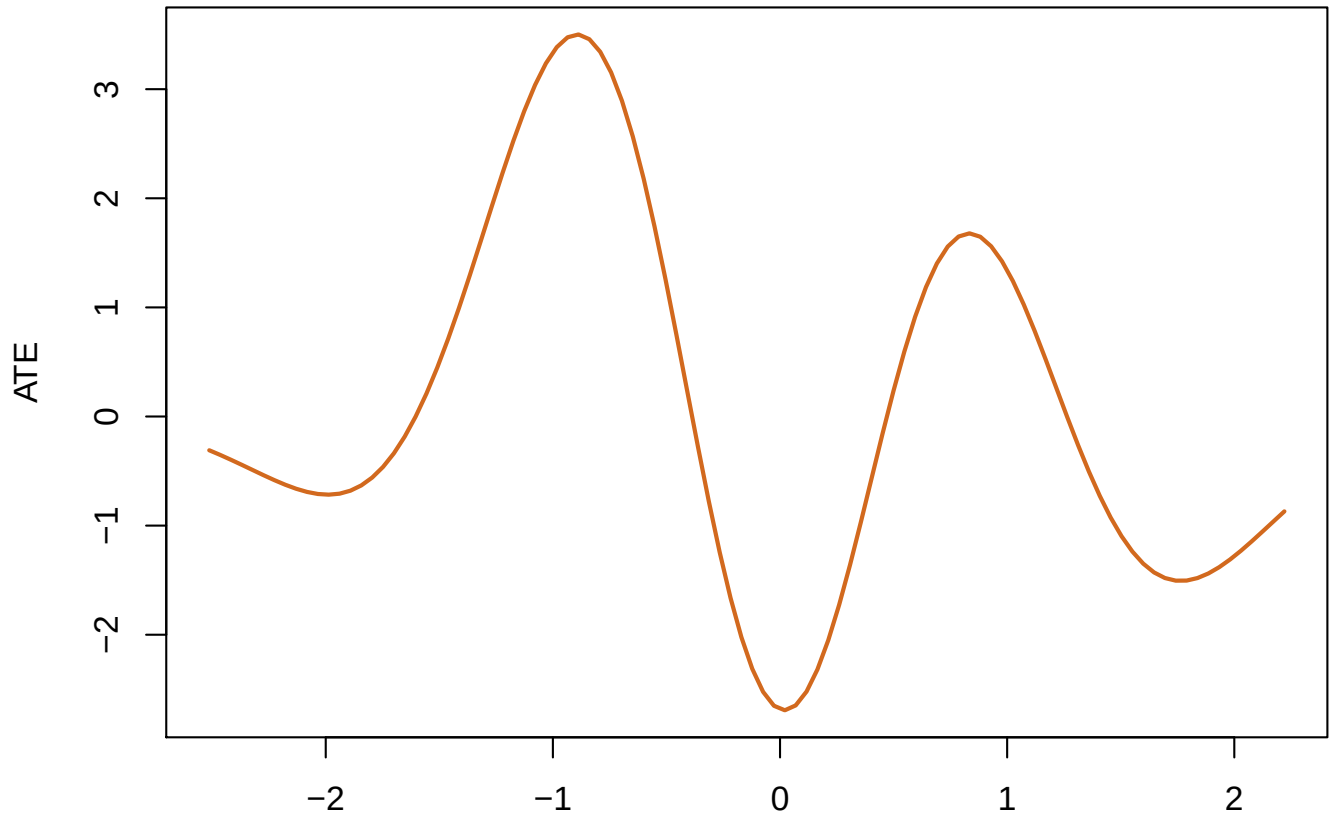
$$\begin{aligned}\text{ATE}(x) &= \int \frac{\partial E[y|x,z]}{\partial x} p(z)\, dz \\ &\approx \frac{1}{N}\sum_i \frac{\partial E[y_i|x,z_i]}{\partial x}\end{aligned}$$

Alternatively,

$$\begin{aligned}\text{ATE}(x) &= \frac{d}{dx}\int E[y|x,z]\, p(z)\, dz \\ &\approx \frac{d}{dx}\left(\frac{1}{N}\sum_i E[y_i|x,z_i]\right)\end{aligned}$$

Simulation 6

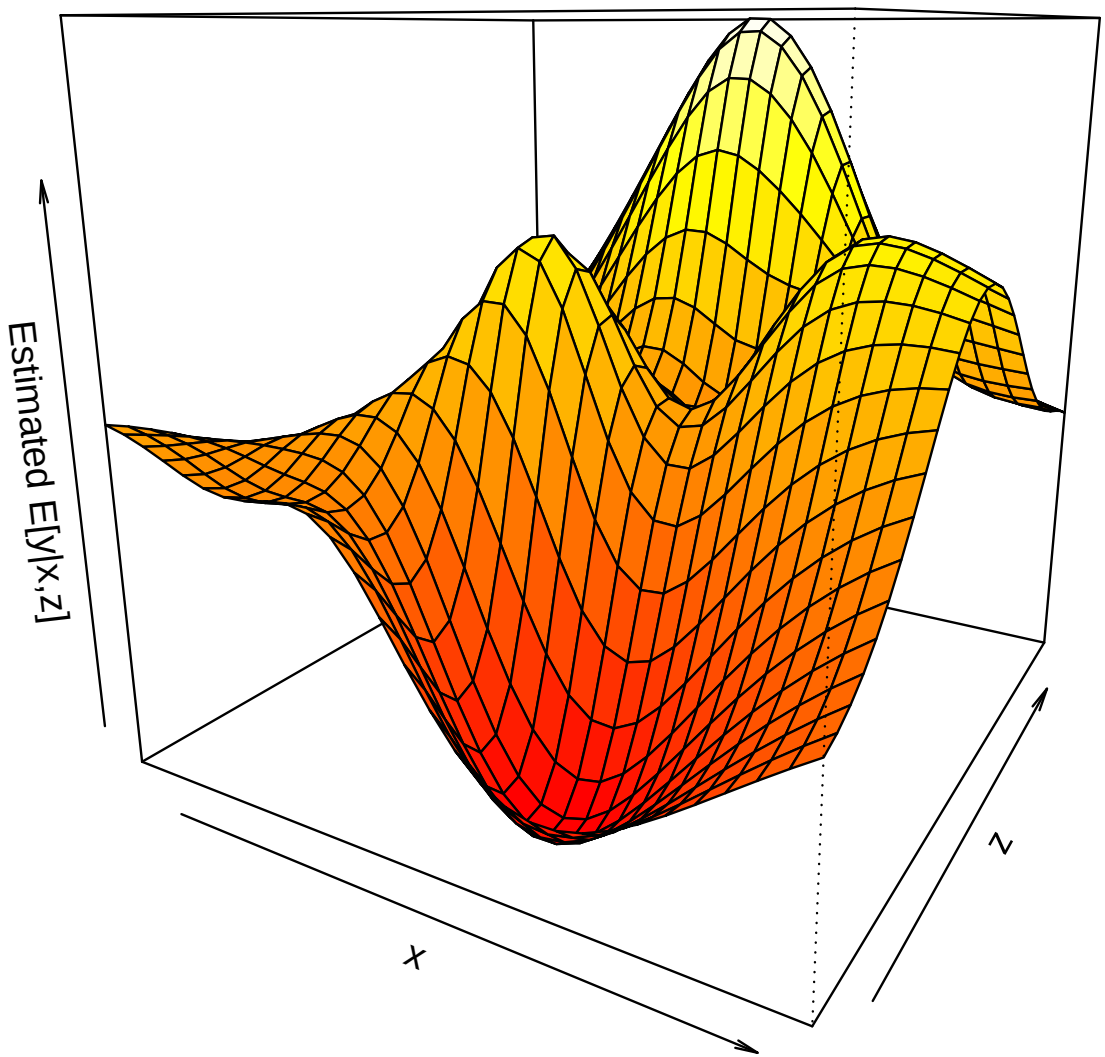
```r
# ATE
dx = D(g,"x") # partial derivative w.r.t. x
xgrid = seq(min(x),max(x),length.out=100) # calculate ate on xgrid
ate = sapply(xgrid, function(a) mean(eval(dx,envir=list(x=a,z=z))))
```



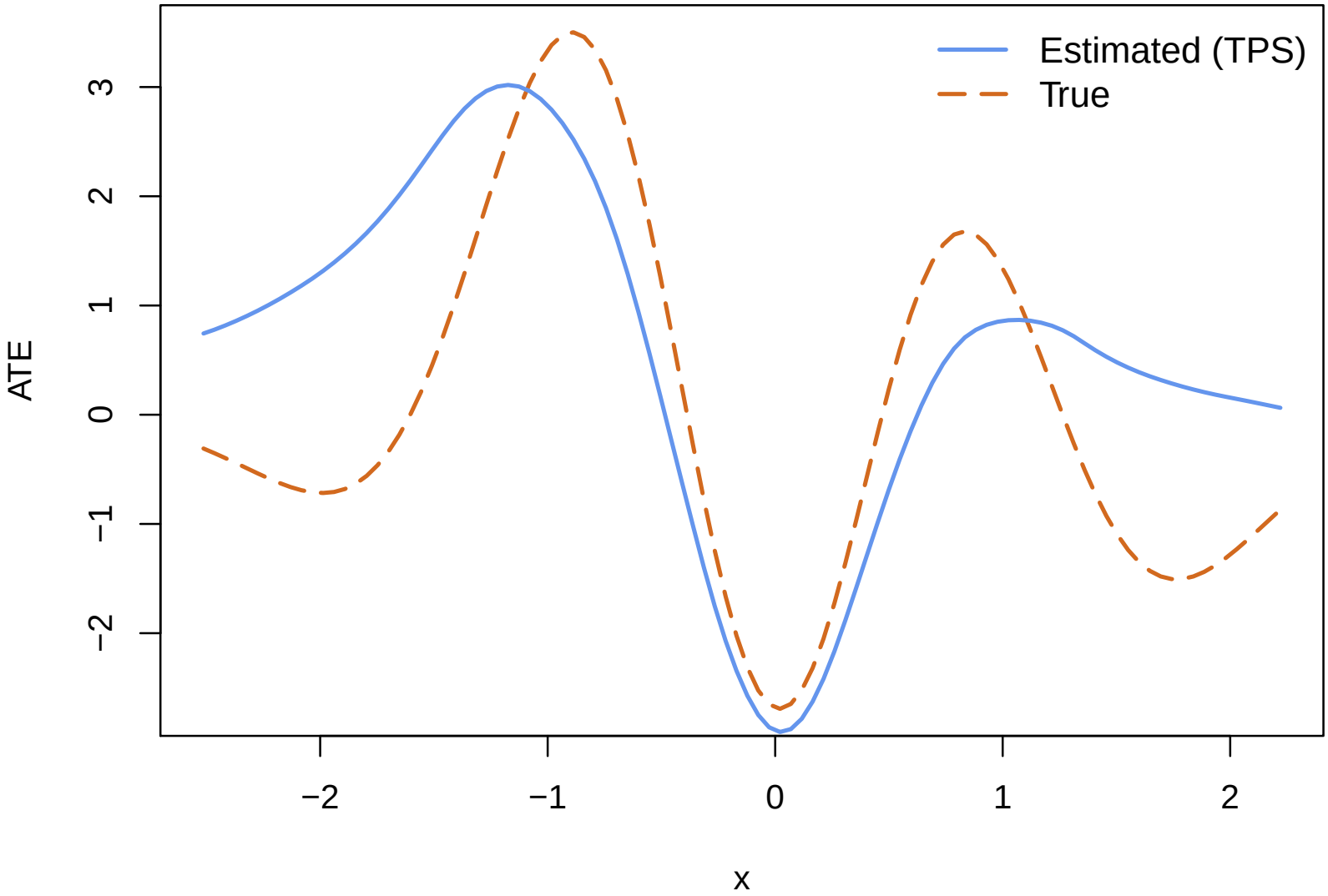
Simulation 6

```
######################
# Thin-plate Spline #
######################
fit = gam(y ~ s(x,z),family=gaussian)

# Estimated ATE
# here we first compute E[y|do(x)] = Integrate(E[y|x,z]) over z
# then we take the derivative of E[y|do(x)] to obtain ATE
require(numDeriv)
cef = function(a) mean(predict(fit,data.frame(x=a,z=z)))
atehat = sapply(xgrid,function(a) grad(cef,a))
```

Simulation 6



Simulation 6



Simulation 6

To choose the best statistical model for $E[y|x,z]$, we can split our data set into training and test sets and perform model selection.

```
# Create Training and Test Sets
require(caret)
train = createDataPartition(y,p=0.5,list=F)
data = data.frame(x=x,z=z,y=y)
data_train = data[train,]
data_test = data[-train,]
```

Simulation 6

```
######################
# Thin-plate Spline #
######################
fit.tps = gam(y ~ s(x,z),data_train,family=gaussian)

# test err
yhat = predict(fit.tps,data_test)
mean((data_test$y - yhat)^2)

## [1] 0.05288992
```

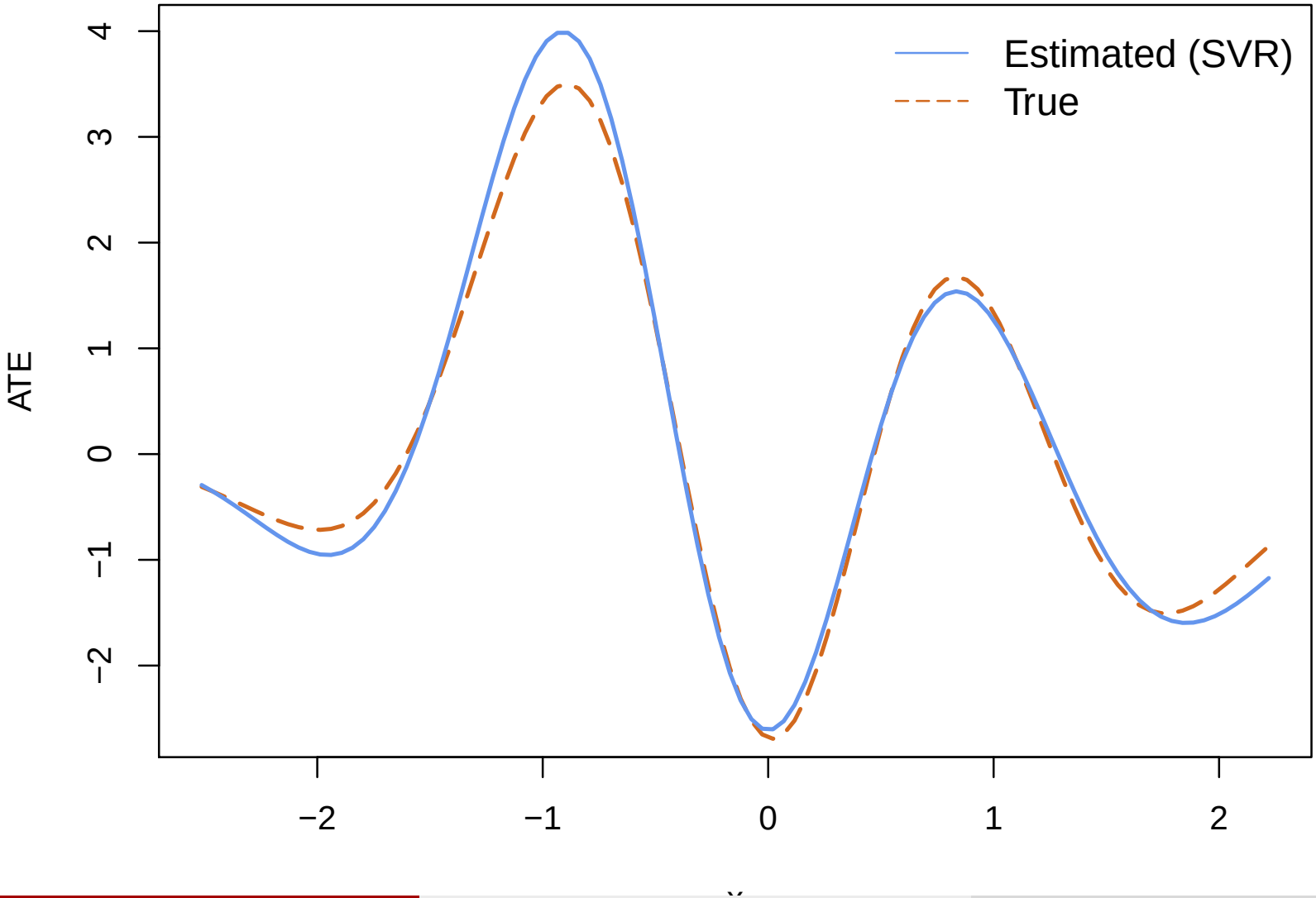
Simulation 6

```
##############################
# Support Vector Regression #
##############################
require(e1071)
fit.svm = svm(y~.,data_train,kernel="radial",
              cost=1e3,gamma=1) # tuning parameters chosen by cv

# test err
yhat = predict(fit.svm,data_test)
mean((data_test$y - yhat)^2)

## [1] 0.02468993
```

Simulation 6

```
###############
# Best Model #
###############
# Fit the best model (here: svr) on the entire data set
# note: the "test set" here is technically still a validation set,
#       since we are using it to select the best model.
#       we can then re-fit the selected model on the combined data
fit = update(fit.svm,data=data)

# Estimated ATE
cef = function(a) mean(predict(fit,data.frame(x=a,z=data$z)))
atehat = sapply(xgrid,function(a) grad(cef,a))
```

Simulation 6

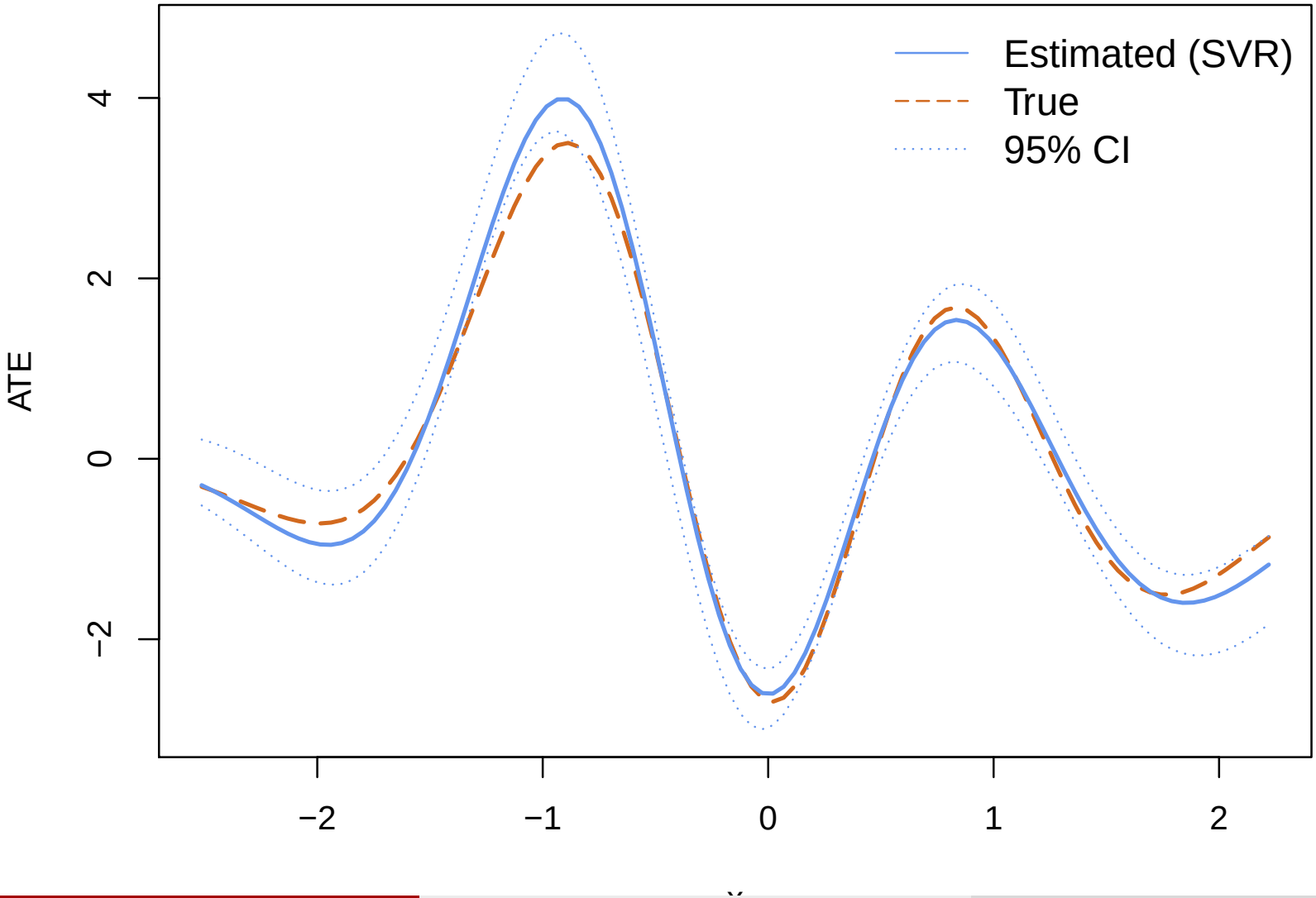


Simulation 6

We can use bootstrap to obtain standard errors and confidence intervals for estimated ATEs.

```
# Function to calculate ATE
calcATE = function(data,index){
  data_i = data[index,]
  fit_i = update(fit,data=data_i)
  cef_i = function(a) mean(predict(fit_i,data.frame(x=a,z=data_i$z)))
  atehat_i = sapply(xgrid,function(a) grad(cef_i,a))
  return(atehat_i)
}

# Bootstrap
require(boot)
B = 1000 # number of bootstrap samples
bootATE = boot(data,calcATE,R=B,parallel="multicore") # use parallel

# Confidence Intervals
ateci = sapply(1:length(xgrid),
       function(i) boot.ci(bootATE,type="norm",index=i)$normal[2:3])
```

Simulation 6



Discussion: Model Selection in Causal Inference

- The examples show that estimates of the causal effect of x on y using the back-door criterion can suffer from significant inaccuracy if (a) we do not sufficiently control for confounders z, or (b) we do not accurately model $E[y|x,z]$.

- Traditional econometrics focus on (a) and pay little attention to (b). In addition, by relying on models that are typically linear in x, traditional econometrics can be thought of focusing only on first-order effects.

Discussion: Bias-Variance Tradeoff in Causal Inference

- Should our goal be obtaining the most accurate causal effect estimates or unbiased causal effect estimates?
- There exists a popular narrative that the goal of machine learning is prediction, hence machine learning methods are not applicable to causal inference. However, as we have discussed, causal effect estimation can be thought of as a type of prediction as well, which we have called *causal prediction* – predicting the effect of $\text{do}(x)$. Obtaining the best causal effect estimate *should* mean obtaining the most *accurate* causal prediction.

Discussion: Bias-Variance Tradeoff in Causal Inference

- Modern machine learning methods explore the bias-variance tradeoff to improve the predictive performance of statical models. In causal inference – in particular, in the field of econometrics today, whether we should explore the same bias-variance tradeoff in causal effect estimation remains a topic of discussion.

- Part of the reason that the bias-variance tradeoff has not been emphasized in econometrics is that traditionally, statistical methods are evaluated based on their asymptotic properties, while the bias-variance tradeoff exists only in finite samples. Focusing on finite-sample performance and using the training-validation-test approach for model selection is one of the main innovations of modern machine learning.

Structual Estimation

- Econometrics began as a discipline that seeks to link economic theory to data.
- Today in the econometrics literature, causal models based on economic theory are referred to as **structural models**. Their estimation is called **structural estimation**.

The Birth of Econometrics

- The Econometric Society was founded on 29th December 1930 at a gathering during the annual joint meeting of the American Economic Association and the American Statistical Association.
- Ragnar Frisch, a founding member of the Econometric Society and the first editor-in-chief of *Econometrica*, wrote in 1923[15]:

  Intermediate between mathematics, statistics, and economics, we find a new discipline which for lack of a better name, may be called econometrics. Econometrics has as its aim to subject abstract laws of theoretical political economy or 'pure' economics to experimental and numerical verification, and thus to turn pure economics, as far as possible, into a science in the strict sense of the word.

[15] Frisch (1926). Quoted is an English translation of the French original. Underlining mine.

The Birth of Econometrics

- Regarding the name "Econometrics", Frisch later wrote[16]:

  So far, we have been unable to find any better word than "econometrics". We are aware of the fact that in the beginning somebody might misinterpret this word to mean economic statistics only. But ... we believe that it will soon become clear to everybody that the society is interested in economic theory just as much as in anything else.

[16] Bjerkholt (1998).

The Birth of Econometrics

- The Cowles Commission for Research in Economics was founded in 1932 and moved to the University of Chicago from 1939 to 1955[17].
  - Motto: "Theory and Measurement"[18]
- The Cowles Commission made foundational contributions to the early development of Econometrics during its Chicago years.
  1. Introducing the probabilistic framework as well as the methods of modern statistical inference into Econometrics
  2. Estimation of structural simultaneous equations models (SEMs).

[17] The commission moved to Yale in 1955 and has been renamed the Cowles Foundation.

[18] According to Cowles' own statement: "This motto replaced the original Cowles Commission motto 'Science is Measurement,' reflecting the importance of theory that became clear early in the history of Cowles."

The Birth of Econometrics

- Thus at its inception, Econometrics was conceived as "a branch of economics in which *economic theory* and *statistical method* are fused in the analysis of numerical and institutional data" (Hood and Koopmans 1953)[19].

[19] Emphasis mine

Structual Estimation

- A complete structural model may specify preferences, technology, the information available to agents, the constraints under which they operate, and the rules of interaction among agents in market and social settings.
- More generally, we refer to any causal models that use economic theory to specify the functional form of causal relationships as structural models.

Structual Estimation

- Because structural estimation learns the entire causal model, once a model is learned, we can use it to derive $p\left(x_j | \operatorname{do}\left(x_i\right)\right)$ for any $\left\{x_i, x_j\right\}$ in the model.
- In contrast, the two-stage causal effect learning procedure introduced on page 128, whose goal is to learn a single causal effect, has been called **reduced-form analysis** in the econometrics literature[20].
- Difference between the two: structural estimation is a generative approach to causal inference, while reduced-form is a discriminative approach.

[20] Historically, given a structural model $g\left(x, y\right) = 0$ that specifies the relationship governing exogenous variable x and endogenous variable y, if y is solved as a function of x, i.e. $y = f\left(x\right)$, then f is referred to as the **reduced form** of g.

Structual Estimation

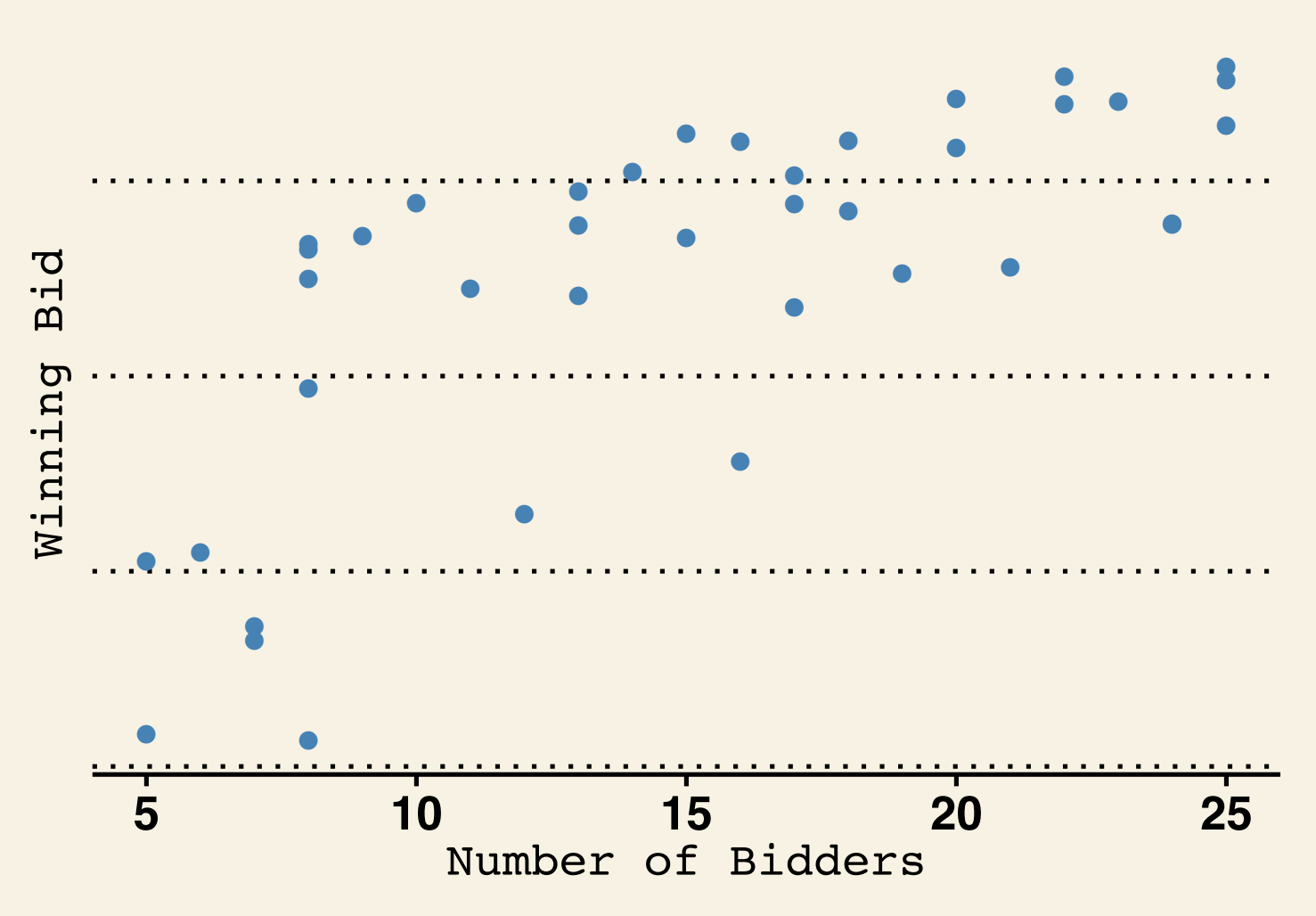
Discriminative Model

Discriminative statistical learning	$p\left(x_j \middle\vert x_i\right)$
Causal effect learning	$p\left(x_j \middle\vert \operatorname{do}\left(x_i\right)\right)$

Generative Model

Generative statistical learning	$p\left(x_1, \ldots x_N; \mathcal{M}\right)$; $\mathcal{M}$: statistical model
Structural estimation	$p\left(x_1, \ldots x_N; \mathcal{M}\right)$; $\mathcal{M}$: causal model

Auction



First-price Sealed-bid Auctions for Identical Goods

Auction

Model

- N risk-neutral bidders
- Independent private value $v_i \sim^{i.i.d.} F(.)$
- Each bidder knows her own v_i and the distribution F, but not the v_i of others
- Observed bids are the Bayesian Nash equilibrium outcome of the game

$\Rightarrow$ Equilibrium bidding strategy:

$$b_i = v_i - \frac{1}{F(v_i)^{N-1}} \int_0^{v_i} F(x)^{N-1} dx$$
$$= v_i - \frac{1}{N-1} \frac{G_N(b_i)}{g_N(b_i)}$$

, where $G_N(.)$ and $g_N(.)$ are the c.d.f. and p.d.f. of the bid distribution.

Auction

Structural Estimation

1. For each auction[a], nonparametrically estimate $G_N(.)$ and $g_N(.)$ from observed bids $\{b_1, \ldots, b_N\}$.
2. For each bidder, calculate

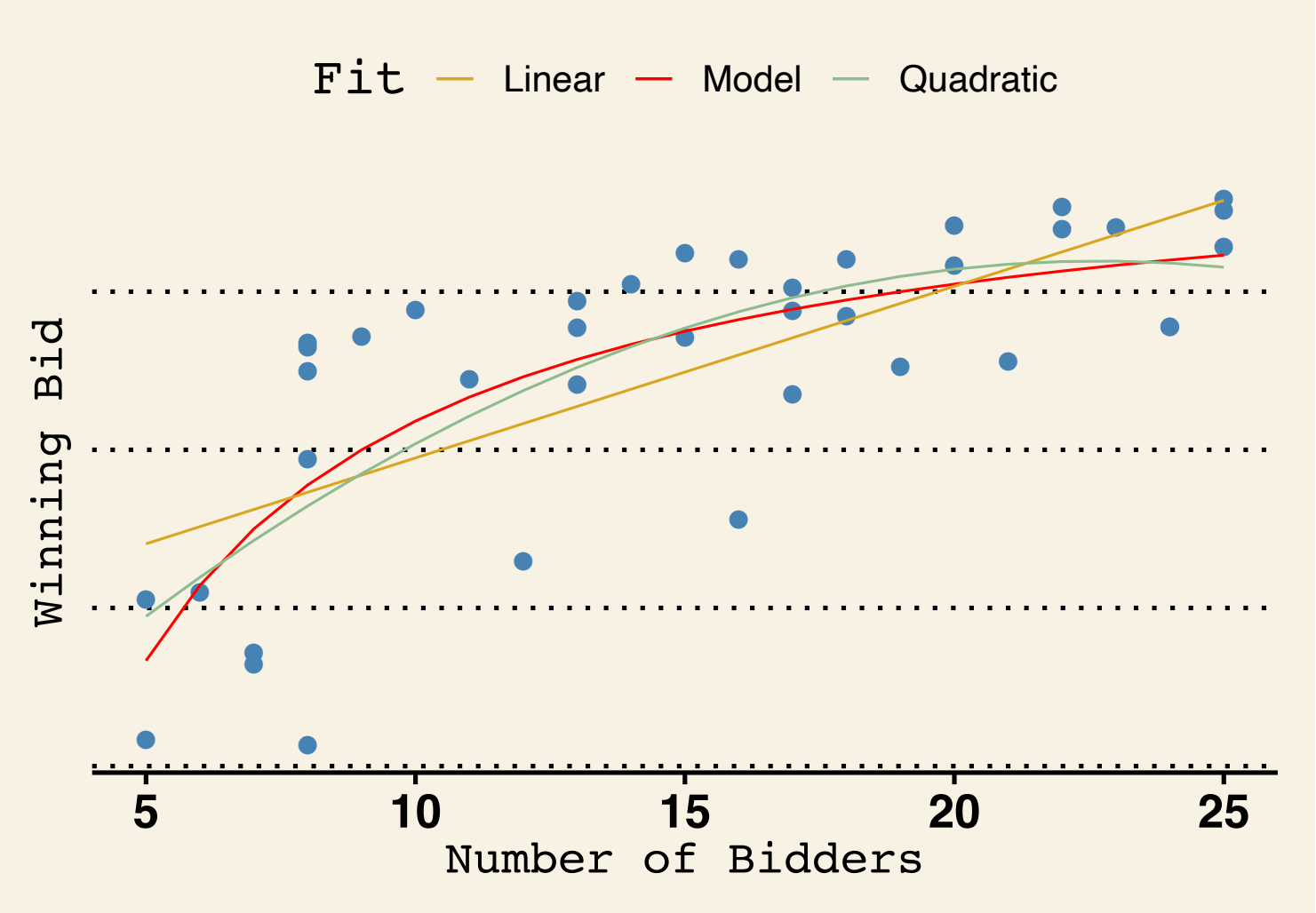
$$\widehat{v}_i = b_i + \frac{1}{N-1} \frac{\widehat{G}_N(b_i)}{\widehat{g}_N(b_i)} \tag{11}$$

3. Use $\widehat{v}_i$ to nonparametrically estimate $F(.)$
4. $\widehat{F}(.)$ can be used to predict the winning bid in an $N-$bidder auction:

$$E[\max\{b_i\}] = E\left[\max\left\{v_i - \frac{1}{\widehat{F}(v_i)^{N-1}} \int_0^{v_i} \widehat{F}(x)^{N-1} dx\right\}\right]$$

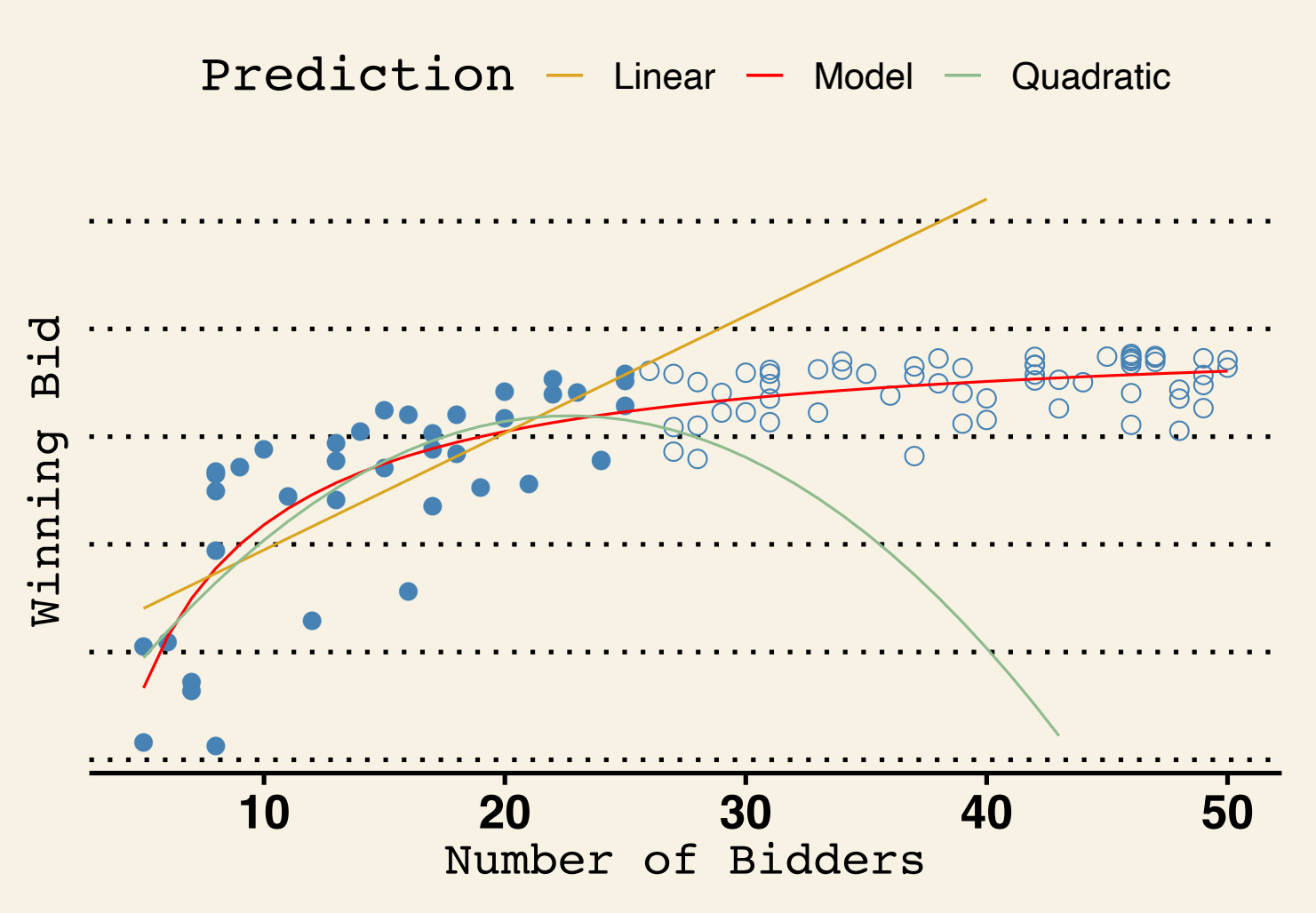
[a] See Guerre et al. (2000).

Auction



First-price Sealed-bid Auctions for Identical Goods

Auction



First-price Sealed-bid Auctions for Identical Goods

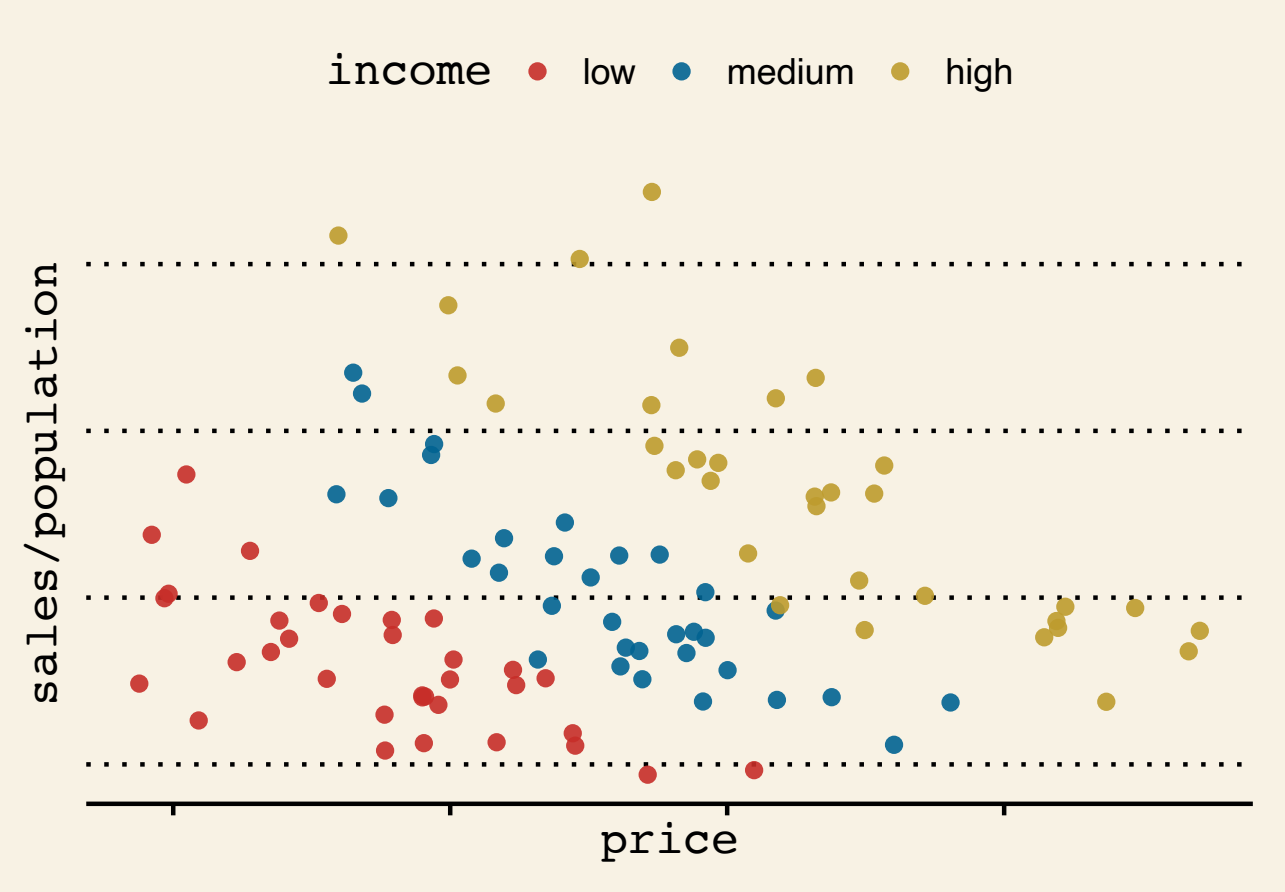
Auction

- Here, there is no confounding between N (the number of bidders) and $b_{\max}$ (the winning bid). Hence $f(N) \equiv E[b_{\max} | N]$ *nonparametrically identifies* the effect of N on $b_{\max}$.
- The estimation problem is to learn $f(N)$ from data. Here, theory helps specify the functional form of $f(N)$ and therefore serves as a model selection mechanism.
- Theory also helps us to learn the values of the bidders (equation (11)) – which cannot be identified nonparametrically – by specifying the functional form of the mapping from $\{v_i\}$ to $\{b_i\}$.

Monopoly

A monopoly firm's pricing and sales in different geographical markets

Data: price, sales, average income, population for each market



Monopoly

Model: Demand

In each market m with population N_m and mean income I_m, consumers choose between the monopoly product and an outside good. Individual utilities are given by:

$$\begin{aligned} U_{i0}^m &= \epsilon_{i0}^m \\ U_{i1}^m &= \beta_0 + \beta_1 I_m - \beta_2 p_m + \epsilon_{i1}^m \end{aligned} \tag{12}$$

, where (U_{i0}^m, U_{i1}^m) are respectively the indirect utilities of the outside good and the monopoly product, and $\epsilon_{ij}^m \sim \text{Gumbel}\,(0,1)$.

(12) $\Rightarrow q_m \sim \text{Binomial}\,(N_m, \pi_m)$, where

$$\pi_m = \frac{\exp\left(\beta_0 + \beta_1 I_m - \beta_2 p_m\right)}{1 + \exp\left(\beta_0 + \beta_1 I_m - \beta_2 p_m\right)}$$

Monopoly

Model: Supply

For each market m, given demand $q_m(p)$, the monopoly firm chooses p to maximize:

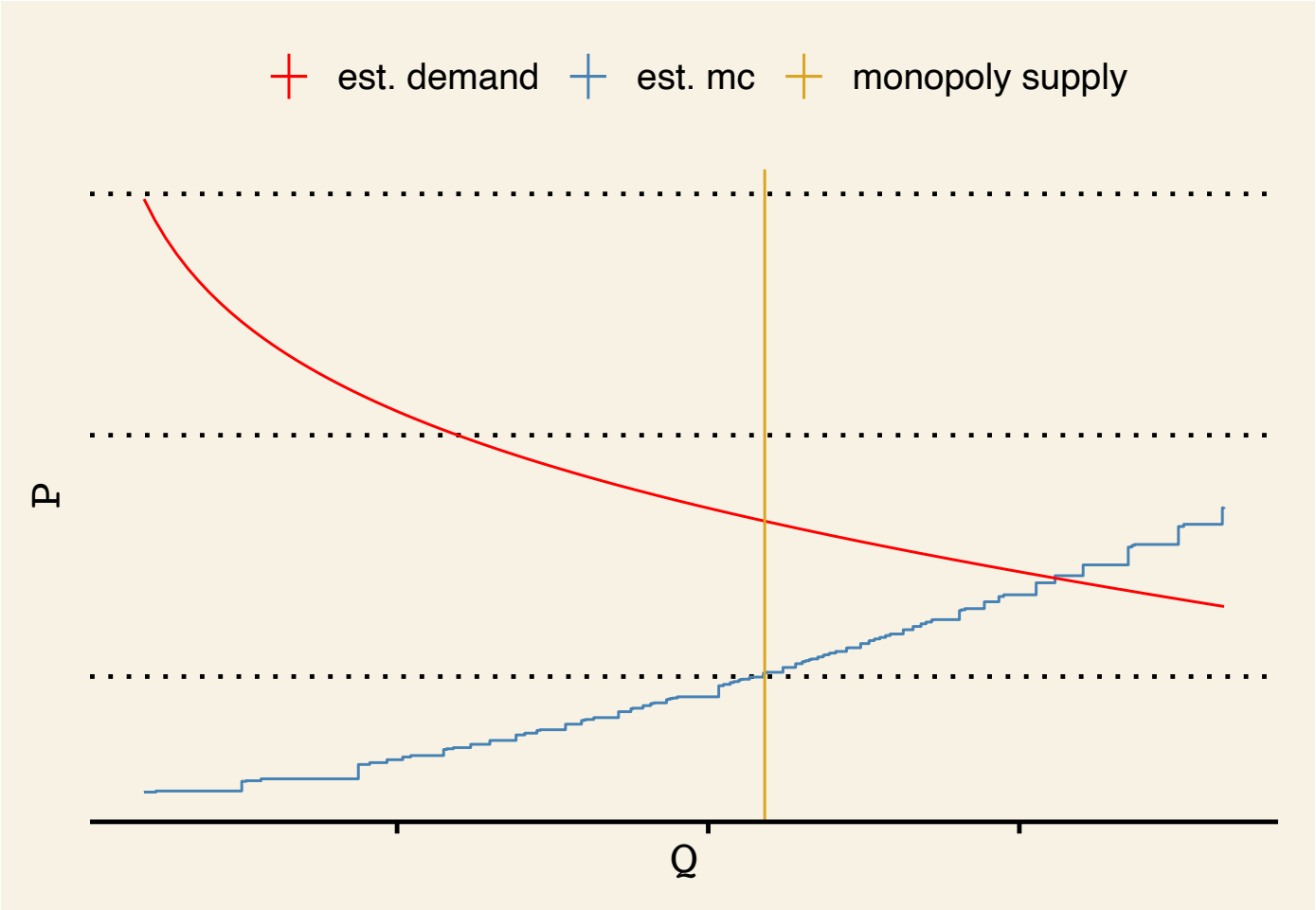
$$\max_p \{p \times q_m(p) - c(q_m(p))\} \tag{13}$$

, where $c(q)$ is the firm's cost function.

(13) $\Rightarrow$

$$c'(q_m) = p_m + \left[q'_m(p_m)\right]^{-1} q_m \tag{14}$$

Monopoly



Estimated marginal cost and demand curves
for a market with median income and population

Monopoly

- Here, theory helps us to learn the marginal cost function of the monopoly firm as well as the consumer utility function – neither of which is observed and neither can be nonparametrically identified.
- Using the estimation results, we can conduct welfare analysis and make normative statements.
  - For example, calculating the total deadweight loss due to monopoly.

Counterfactual Simulation

- One of the benefits of learning a structural model is that it allows us to predict the effect of a completely new treatment – a treatment that has never been observed before.
  - If in the observed data, x is always equal to 0, what would be the effect of $\text{do}(x = 1)$?
- Because structural estimation learns an entire structural model, once we have learned a model with variables $\{x_1, \ldots, x_n\}$, we can use it to generate data from the distribution $p(x_1, \ldots, x_n | \text{do}(x_j = a))$ for any hypothetical manipulation $\text{do}(x_j = a)$. This is called **counterfactual simulation**.

Counterfactual Simulation

Counterfactual Simulation

What if Caesar never crossed the Rubicon?

Monopoly

What happens if the government imposes a 20% sales tax on the company?

After tax:

$$\begin{aligned} \Delta \texttt{ Consumer Surplus:} &\quad -27.83\% \\ \Delta \texttt{ Total Surplus:} &\quad -27.95\% \end{aligned}$$

Tax incidence:

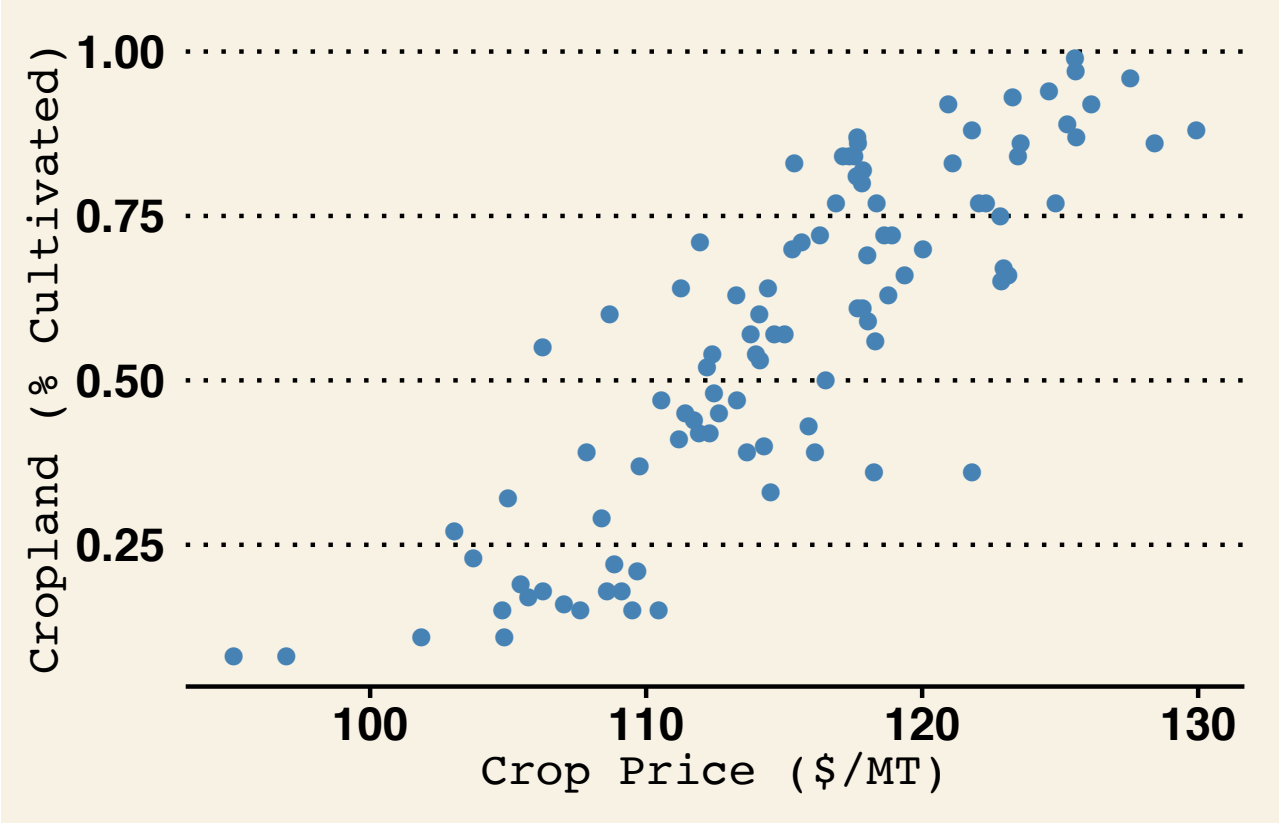
$$\texttt{Consumer:} \quad 26.65\%$$

Dynamic Structural Model

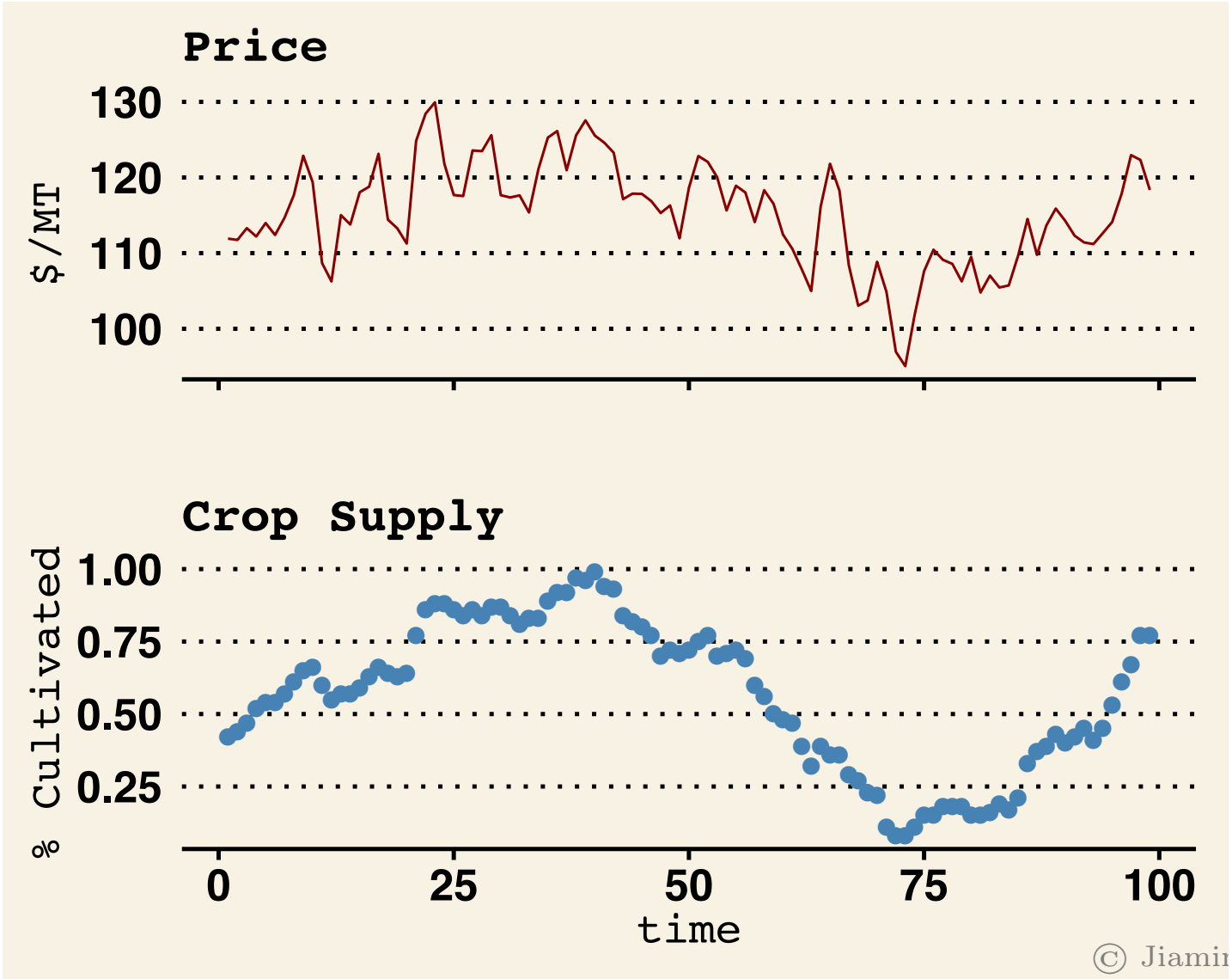
- In a changing environment, with new information arriving each period, individual are forward-looking when making decisions: choices are made partly based on expectations of the future.
- Decisions are also often influenced by the past. Since it can be costly to transition from one state to another, payoffs to different choices are often history-dependent: our past partly shapes our future.
- In dynamic models, treatment effects can be time-varying and it's often useful to distinguish between short-run and long-run effects.

Crop Supply

Data: crop price and percentage of cropland cultivated in a county for $t = 1, \ldots, T$



Crop Supply



Crop Supply

Model

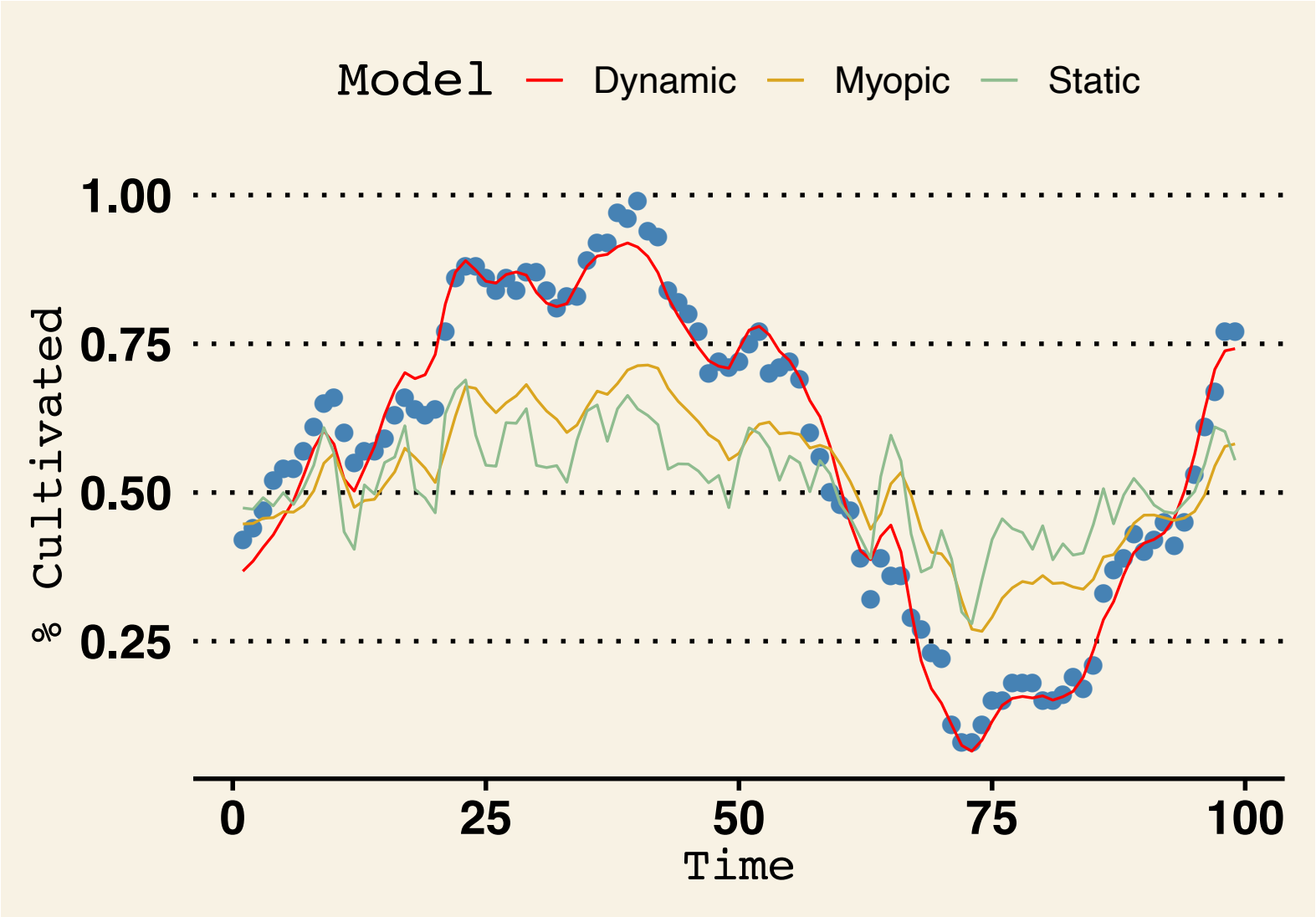
- At the beginning of each period t, each field owner decides whether or not to plant the crop in the current period.
- The decision is based on observed period$-t$ price as well as expectations of future prices.
- If a field has not been cultivated for k periods, then in order to (re)-cultivate it, the farmer needs to pay a one time cost $c(k)$.
- Farmers have **rational expectation**: their expectations of future prices are unbiased conditional on the information they have.
    - Here we assume that crop prices follow an AR(1) process, which is known to the farmers.

Crop Supply

Counterfactual Simulation:

- How would crop supply change in response to changes in crop prices if farmers are myoptic: if they are not forward-looking?
- How would crop supply change in response to changes in crop prices if farmers are static: if they are neither forward-looking, nor subject to any re-cultivation costs, so that planting decisions are made entirely based on current prices?

Crop Supply


Model
Dynamic
Myopic
Static
1.00
0.75
0.50
0.25
% Cultivated
0
25
50
75
100
Time

Crop Supply

- In general, if we are interested in the effect of x on y, but x is self-selected based on expectations of y, then without any measures of such expectations, the causal effect cannot be nonparametrically identified and we need to rely on theory to specify how expectations are formed.
- Here, farmers take crop prices as given and decide whether or not to plant[21]. Models like this are called **dynamic discrete choice models**.
- If prices are endogenous – if farmers' planting decisions affect equilibrium prices[22] – then we need to model both crop supply and crop demand. Such models are called **dynamic general equilibrium models**.

[21] This is a reasonable assumption since we are looking at a single county.
[22] For example, if we look at all the farmers in a country or in the world.

Why Structural Estimation

S-0. Structural models, by *explicitly* modeling the data-generating causal mechanisms, make clear what prior knowledge (assumptions) are relied upon to draw causal inference[23].

[23] There is a mistaken belief among some practitioners that structural estimation *is* causal mechanism learning. This is incorrect: structural estimation is learning based on an assumed causal mechanism.

Why Structural Estimation

By using theory to specify the functional forms of causal relationships, structural models can be used to:

S-1. identify causal effects or the values of unobserved variables that *cannot* be nonparametrically identified[24].

S-2. serve as a model selection mechanism[25] for causal effects that *can* be identified nonparametrically.

[24] For this reason, structural estimation is sometimes described as *identification by functional form*.

[25] i.e., determine the functional form of $E[y|\text{do}(x)]$.

Why Structural Estimation

S-3.1. Using structural models, what we learn from one set of data $\mathcal{D} \sim p(x, y)$ can be potentially used to explain and predict data drawn from another distribution, say $p(u, v)$, if $\{x, y\}$ and $\{u, v\}$ are generated from a similar causal mechanism.

- In other words, what we learn from one observed phenomenon can be used to explain and predict other related phenomena.
- For example, we can learn individuals' risk aversion from their investment behavior, which in turn, can help explain and predict their career choices.

Why Structural Estimation

S-3.2. Structural models make it possible to predict the effects of existing treatments in a new population/environment, or the effects of completely new treatments.

- To do so, a structural model must be "deep" enough so that its parameters remain invariant in the new population/environment, or when new treatments are applied.
- The concept of invariance is closely related to the concept of stability for causal relationships. The need for invariant parameters is key to causal analysis and policy evaluation.

S-4. Once we have learned a structural model, we can use it to generate synthetic data and perform counterfactual simulations.

Three Types of Program Evaluation Problems

P-1 Evaluating the impacts of historical programs on outcomes.

P-2 Forecasting the impacts of programs implemented in one population/environment in other populations/environments.

P-3 Forecasting the impacts of programs never historically experienced.

- P-1 is the problem of internal validity.
- P-2 is the problem of external validity.
- P-2 and P-3 often require the use of structural models.
- For all three types of problems, if we want to evaluate welfare impact, we need a structural model.

Why Structural Estimation

S-5. Structural models, by linking economic theories based on individual preferences to data, allow the economist to make welfare calculations and normative statements.

- Individual choices reveal information about their preferences and the potential outcomes they face[26].

[26] Heckman and Vytlacil (2007): "Incorporating choice into the analysis of treatment effects is an essential and distinctive ingredient of the econometric approach .. An assignment in [the RCM] is an assignment to treatment, not an assignment of *incentives* and *eligibility* for treatment with the agent making treatment choices. [The statistical treatment effect approach] has only one assignment mechanism and treats noncompliance with it as a problem rather than as a source of information on agent preferences, as in the econometric approach ... Accounting for uncertainty and subjective valuations of outcomes ... is a major contribution of the econometric approach."

Why Structural Estimation

S-6. Like all scientific models, structural models can potentially deliver better predictive performance than statistical models trained on single data sets, because their parameters can be learned from a combination of data from various sources that share the same underlying causal mechanism[27].

[27] This point is related to S-3.1 and S-3.2: different observed phenomena can all help inform the values of the same set of "deep" parameters.

Appendix: Causal Reasoning in Econometrics

The reduced-form econometric approach to causal reasoning starts with assuming a "true model":

$$y = \alpha_0 + \alpha_1 x + u \tag{15}$$

, where x is observed, u is unobserved, and α_1 represents the average causal effect of x on y.

To estimate α, we can run a linear regression:

$$y = \beta_0 + \beta_1 x + e \tag{16}$$

If u and x are linearly uncorrelated, then we say x is exogenous and OLS estimation of (16) produces $\widehat{\beta}_1$ that is an unbiased estimate of α_1. Otherwise, x is endogenous and $\widehat{\beta}_1$ will be a biased estimate of α_1.

Appendix: Causal Reasoning in Econometrics

Limitations:

1. The econometric approach does not clearly distinguish between causal reasoning and statistical modeling, between what assumptions are causal and what are statistical.

Appendix: Causal Reasoning in Econometrics

- (15) is often confused with (16). (16) is a statistical model, where the linear relation between x and y is purely a statistical assumption. OLS estimation of (16) will always produce $\widehat{\beta}$ that are *unbiased* estimates of

$$\beta^* = \underset{\beta=(\beta_0,\beta_1)}{\arg\min}\, E\left[(y-\beta_0-\beta_1 x)^2\right]$$

, in the sense that $E\left(\widehat{\beta}\right) = \beta^*$.

Appendix: Causal Reasoning in Econometrics

- (15), on the other hand, is (part of) a causal model that makes the causal assumption that y is *determined* by x and variables in u, in addition to making a statistical assumption on the functional form of their relationship. In other words, (15) should be considered a structural equation.

- β_1 is a statistical parameter. α_1 is a causal parameter.

Appendix: Causal Reasoning in Econometrics

- When the econometrics literature states that: *"when the error term is correlated with the regressor, the estimation result is biased,"* it really has in mind *two* models: (15) and (16).
  - By "the error term," it is referring to u not to e.
  - By "the estimation result is biased," it is saying that $E\left(\widehat{\beta}\right) \neq \alpha$, i.e., the statistical model (16) produces a biased estimate of the causal model (15).
    - $\widehat{\beta}_1$ is an *unbiased* estimate of the statistical parameter β_1^*, but a *biased* estimate of the causal parameter α_1.

Appendix: Causal Reasoning in Econometrics

- However, this is not always made clear. Most econometrics textbooks use one equation to represent both models, confusing what is causal and what is statistical.

- The requirement that u be (linearly) uncorrelated with x is often stated as an essential assumption on the linear regression model itself, under which the OLS estimator is unbiased[28,29] – again confusing what is causal and what is statistical.

[28] For example, in Stock and Watson (2010), this is listed as one of the "Least Squares Assumptions."

[29] The linear regression model as a statistical model, of course, places no such assumptions on the error term.

Appendix: Causal Reasoning in Econometrics

Structural vs. Statistical Equation

Failure to understand (15) as a structural equation can lead to confusion and unhappiness in life. Take, for example, the following causal model:

$$x \sim U(0,1), y \sim U(0,1)$$
$$u \leftarrow y - x$$

, where "$\leftarrow$" means that the variable on the left is determined by the variables on the right.

Appendix: Causal Reasoning in Econometrics

Structural vs. Statistical Equation (cont.)

The causal effect of x on y is 0 – no effect. However, the causal model implies the following statistical equation:

$$y = \alpha x + u \tag{17}$$

, where $\alpha = 1$ and u is correlated with both x and y.

Failure to understand (17) as a statistical rather than structural equation will lead us to wrongly conclude that a regression of y on x will produce biased causal estimates – it won't. It will only produce a biased estimate of α, but α is not the average causal effect of x on y.

Appendix: Causal Reasoning in Econometrics

Limitations:

2. The econometric approach does not clearly distinguish between nonparametric and parametric identification, and between identification and estimation.

Appendix: Causal Reasoning in Econometrics

- Identification in the econometric approach refers to whether the parameters in a "true model" can be uniquely determined given infinite data on the observed variables. A "true model" like (15), however, already makes parametric assumptions on the underlying causal structure.
- Therefore, while causal inference based on causal graphical models recognizes causal effect learning as a two-stage process – identification and estimation – and uses the back-door and front-door criteria to establish clear rules for nonparametric identification, the econometric approach fails to do so[30].

[30] This failure leads to confusion over how to apply statistical and machine learning models to causal inference.

Appendix: Causal Reasoning in Econometrics

Limitations:

3. The econometric approach does not provide an easily operational way of choosing control variables for identifying a desired causal effect.

Appendix: Causal Reasoning in Econometrics

- If x is endogenous, we can try to find control variables z such that conditional on z, u is no longer correlated with x in the following model:

$$y = \alpha_0 + \alpha_1 x + \alpha_2 z + u \tag{18}$$

Then we can run the following regression:

$$y = \beta_0 + \beta_1 x + \beta_2 z + e \tag{19}$$

and $\widehat{\beta}_1$ will be an unbiased estimate of α_1 – our causal parameter of interest.

Appendix: Causal Reasoning in Econometrics

- The requirement that u be (linearly) uncorrelated with x conditional on z can be thought of as implied by the back-door criterion.
- However, unlike the back-door criterion, the econometric approach to causal reasoning does not offer a clear guidance on the choice of z, because it is not based on a clear thinking of the underlying causal mechanism (such as a causal diagram representation)[31].

[31] Clear thinking on causal mechanism forms part of the appeal of structural estimation over the reduced-form approach.

Appendix: Causal Reasoning in Econometrics

- The statistics and econometrics literature often state that in order for the strategy outlined on page 245 to be successful – in order for causal effect to be identifiable by conditioning – there must be no unmeasured confounding (statistics) or selection on observables (econometrics).

- However, the back-door criterion shows that we do not need to observe and condition on all confounders, only a *sufficient* set of variables that renders all back-door paths blocked.

    - In Figure 1, W is the confounder, but we do not need to observe it: the causal effect of X on Y is identifiable by conditioning on C.

Appendix: Causal Reasoning in Econometrics

- There are now more than one "true models", because for most problems, multiple sets of variables exist that can render all back-door paths blocked ($\{C\}, \{W\}, \{C, W\}$ in Figure 1).
- The econometrics literature defines **omitted variable bias** as the bias that arises when a variable is omitted from the "true model." But with more than one true model – with multiple sets of z that can sufficiently control for confounding – what is an omitted variable?

Appendix: Causal Reasoning in Econometrics

- Not only is (18) no longer unique, its meaning is now no longer clear. Once we include control variables z, (18) can no longer be interpreted as a structural equation in a causal model. The back-door criterion makes it clear that z can include not only direct causes of y (W in Figure 1), but also variables that are not causes of y (C in Figure 1). Therefore, (18) is no longer a structural equation specifying the relation between y and its determinants. The meaning of the equation itself has become unclear, not to mention its ability to offer meaningful guidance on the choice of z.

- The back-door criterion also makes it clear that z should not include descendants of x (C in Figure 2) and should not include colliders that may open a back-door path (E in Figure 2). The econometric approach offers none of these guidances.

Appendix: Causal Reasoning in Econometrics

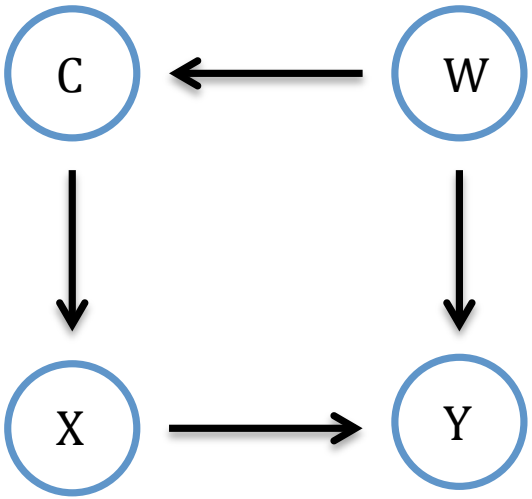


Figure: 1

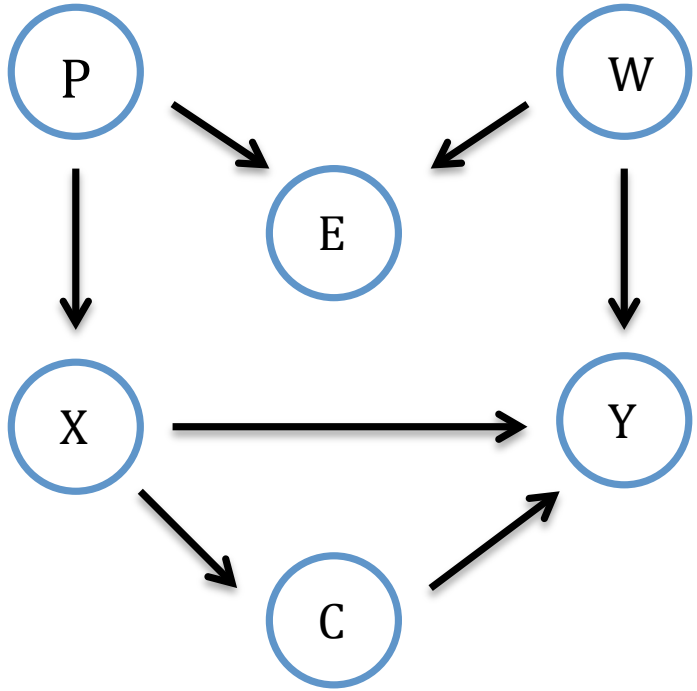


Figure: 2

Acknowledgement

Some slides are adapted from the following sources:

- Abu-Mostafa, Y. S., M. Magdon-Ismail, and H. Lin. 2012. *Learning from Data*. AMLBook.
- Antoine de Saint-Exupéry. 2001. *Le Petit Prince*. Harcourt, Inc. (Original work published 1943.)
- Angrist, J. D. and J. Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.
- Doré, G. *The Dore Illustrations for Dante's Divine Comedy*. Dover Publications, 1st edition (1976).
- Hernán, M. A. and J. M. Robins. 2019. *Causal Inference*. CRC Press.
- James, G., D. Witten, T. Hastie, and R. Tibshirani. 2013. *An Introduction to Statistical Learning: with Applications in R*. Springer.
- Silva, R. *Causal Inference in Machine Learning*. Talk at Imperial College London, retrieved on 2017.01.01. [link]

Reference I


Bajari, P., D. Nekipelov, S. P. Ryan, and M. Yang. 2015. "Machine Learning Methods for Demand Estimation," *American Economic Review*, 105(5).


Bjerkholt, O. 1998. "Ragnar Frisch and the Foundation of the Econometric Society," In Steinar Strøm (ed.), *Econometrics and Economic Theory in the 20th Century: The Ragnar Frisch Centennial Symposium*, Cambridge University Press.


Deaton, A. and N. Cartwright. 2018. "Understanding and misunderstanding randomized controlled trials," *Social Science & Medicine*, 210.


Frisch, R. 1926. "Sur un problème d'économie pure," *Norsk Matematisk Forenings Skrifter*, Series I, No. 16.


Guerre, E., I. Perrigne, and Q. Vuong. 2000. "Optimal Nonparametric Estimation of First-Price Auctions," *Econometrica,* 68(3).


Heckman, J. J. and E. J. Vytlacil. 2007. "Econometric Evaluation of Social Programs, Part I: Causal Models, Structural Models and Econometric Policy Evaluation." In J. Heckman and E. Leamer (eds.), *Handbook of Econometrics*, Vol. 6, Elsevier.

Reference II


Hood, W. C. and T. C. Koopmans (eds.) 1953. *Studies in Econometric Method*, Cowles Commission Monograph 14, Wiley.


Russell, B., 1912. *The Problems of Philosophy*. Arc Manor, Rockville, MD (2008).


Scott, P. T. 2013. "Dynamic Discrete Choice Estimation of Agricultural Land Use," Working Paper.


Shalizi, C. R. 2016. *Advanced Data Analysis from an Elementary Point of View*. Manuscript.


Stock, J. H. and M. W. Watson. 2010. *Introduction to Econometrics* (3rd ed.). Pearson Education.

Thank you!!

For more, check out: https://jiamingmao.github.io/data-analysis