

# Introduction to Forecasting

*ECON 5753 — University of Arkansas*

Prof. Kyle Butts

January 2025

## **Forecasting**

### **Goals of Forecasting**

### **Fitting Models**

Model Selection: Adveristing Example

### **Types of Data**

# Problem of Prediction

The goal of forecasting is to learn the relationship between **input variables** and **outcome variable(s)** so that we can **predict** the outcome variables when we do not observe it.

# Problem of Prediction

The goal of forecasting is to learn the relationship between **input variables** and **outcome variable(s)** so that we can **predict** the outcome variables when we do not observe it.

E.g., Learn about who are potential customers to advertise to based on their observable characteristics

- Input: observable characteristics
- Outcome: whether they purchase a product

# Problem of Prediction

The goal of forecasting is to learn the relationship between **input variables** and **outcome variable(s)** so that we can **predict** the outcome variables when we do not observe it.

E.g., Learn about who are potential customers to advertise to based on their observable characteristics

- Input: observable characteristics
- Outcome: whether they purchase a product

Predict values of a variable in the future, e.g. **time-series** of stock prices

- Input: the time-period
- Output: stock price

# Model

We have an outcome variable  $y$  and a set of  $p$  different predictor variables

$X = (X_1, X_2, \dots, X_p)$ .

- For some observations we observe both  $X$  and  $Y$ 
  - Essential to *fit* the model, i.e. learn the relationship between the two

We can write the model in a general form as

$$y = f_0(X) + \varepsilon,$$

where  $f_0$  is some unknown function of  $X$  and  $\varepsilon$  is the **error term**.

# Model

$$y = f_0(X) + \varepsilon,$$

Here we think of  $f_0(X)$  as the ‘true’ model: this is the best prediction of  $y$  given information on  $X$ . In this case we assume  $\mathbb{E}[y - f_0(X) \mid X] = \mathbb{E}[\varepsilon \mid X] = 0$

- Once we know  $f_0(X)$ , we can not predict any more of the variation in  $y$  using  $X$

## Prediction of $f$

We will use a training sample of data to predict  $f_0$ . We will denote our estimate of  $f_0$  as  $\hat{f}$ . For a given unit, we predict

$$\hat{y} = \hat{f}(X)$$

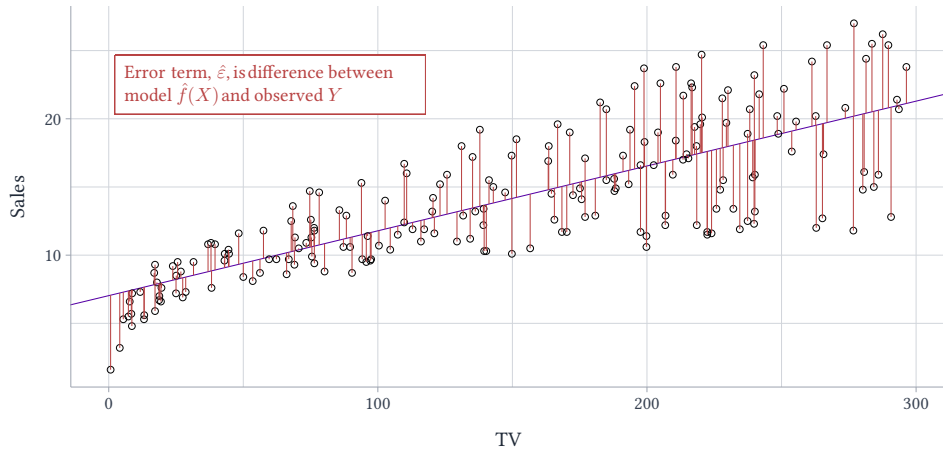


# Error term

The term ‘error term’ is a bit overloaded. It’s worth trying to clarify:

1. If we knew the ‘true’ model,  $f_0(X)$ , then  $\varepsilon \equiv y - f_0(X)$  represents the things that are unpredictable given  $X$   
→ This is the “true” error term
2. If  $\hat{f}(X)$  is an estimated model,  $\hat{\varepsilon} = y - \hat{f}(X)$  is the difference between that unit’s  $y_i$  and their predicted  $y$ ,  $\hat{f}(X_i)$   
→ This is better called the **prediction error**

# Error term



## **Forecasting**

## **Goals of Forecasting**

## **Fitting Models**

Model Selection: Adveristing Example

## **Types of Data**

# Goals of estimation of $f$

There are two related goals when predicting  $f$ :

1. Predict  $y$  as well as possible (**prediction**)

- Think of prediction as a ‘black box’ where the goal is to do as good of a job at predicting  $y$  as possible
- Want to learn the ‘true’  $f_0(X)$ , leaving no information on the table

2. Understand the relationship between  $X$  and  $y$  (**inference**)

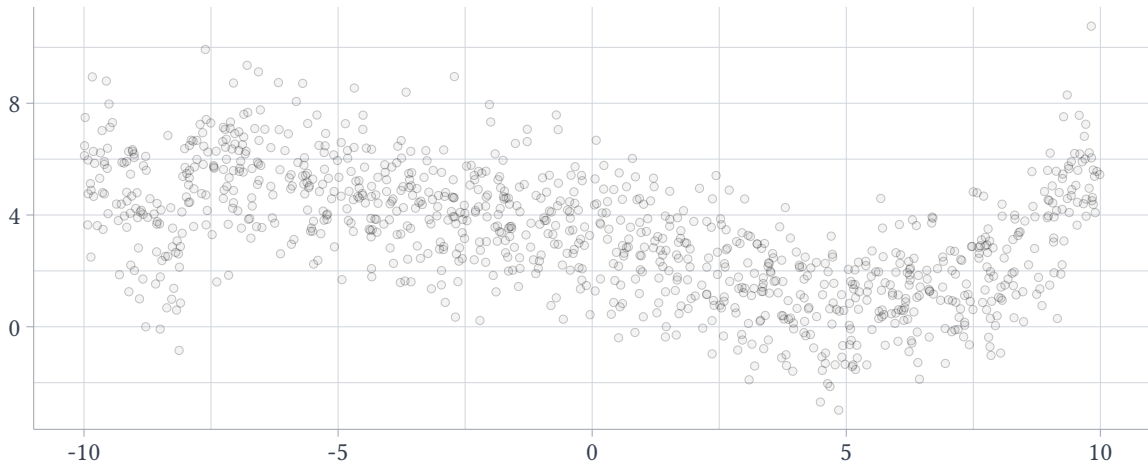
- If our goal is being able to describe the relationship between  $x$  and  $y$ ; i.e. we care about understanding  $f_0$  and not just  $\hat{y}$
- Worth approximating  $f_0$  with a more ‘simple’ functional form so that we can better convey the results to stakeholders.
  - Creates “heuristics” that helps decision-makers

# Examples of Prediction vs. Inference

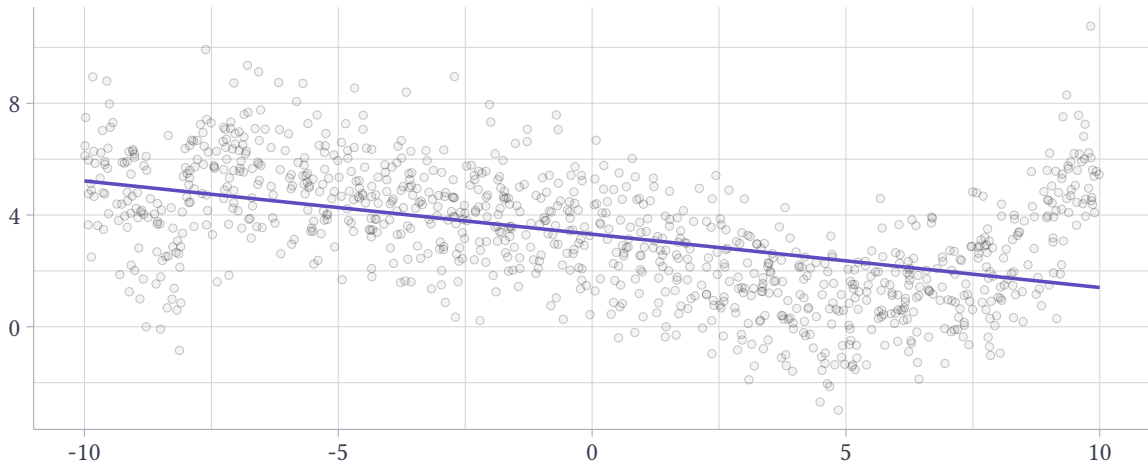
Say we are thinking about housing:

- Example of prediction: “Is the house over/under-priced”  
→ Only care about  $\hat{\text{price}} = f(X)$
- Example of inference: “How much more do homes with a river view sell for?”  
→ Need to know information about  $f$  to answer

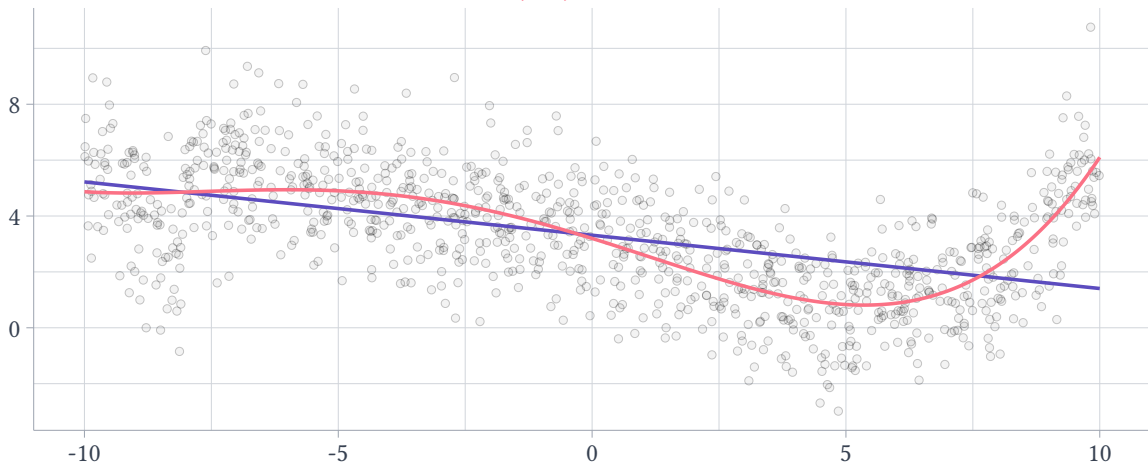
Examples of  $\hat{f}$ :



Examples of  $\hat{f}$ : Line

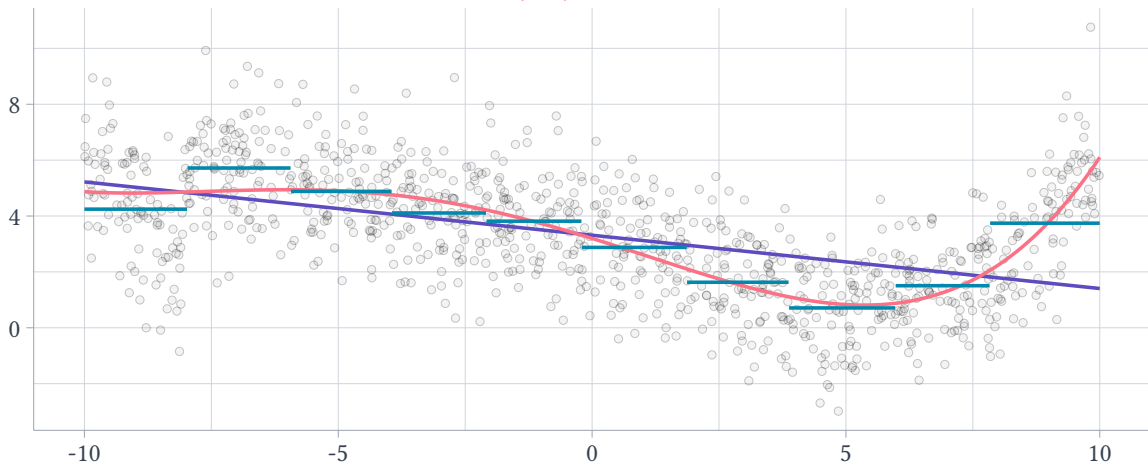


Examples of  $\hat{f}$ : Line, Polynomial ( $x^4$ )

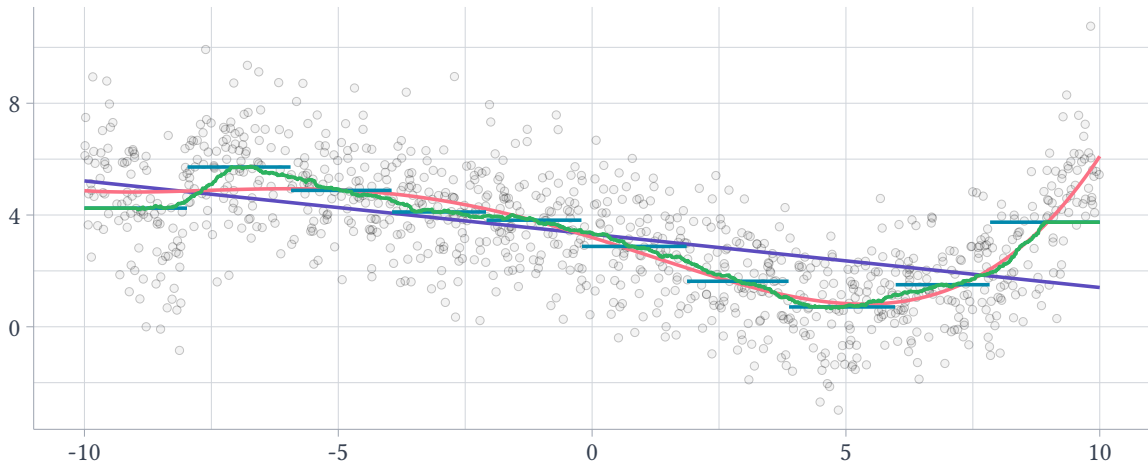




Examples of  $\hat{f}$ : Line, Polynomial ( $x^4$ ), Bins of  $x$



Examples of  $\hat{f}$ : Line, Polynomial ( $x^4$ ), Bins of  $x$ , KNN of  $x$

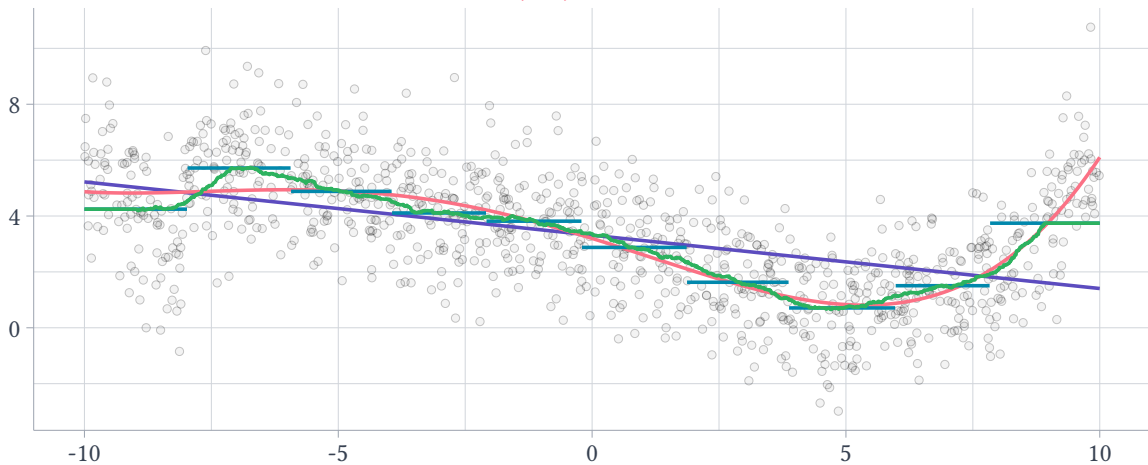


# Model Flexibility

There is a limit to how **flexible** we can make our model

1. If our goal is prediction, we only have a finite amount of data to use to fit the model, so there's a limit on how much we can learn
  - Face the risk of **overfitting** the data (chasing after the random noise  $\varepsilon$ )
2. If our goal is inference, then added flexibility is harder to summarize to stakeholders.

Examples of  $\hat{f}$ : Line, Polynomial ( $x^4$ ), Bins of  $x$ , KNN of  $x$



## **Forecasting**

### **Goals of Forecasting**

### **Fitting Models**

Model Selection: Adveristing Example

### **Types of Data**

# High-level of fitting models

The general framework for forecasting is as follows:

1. Collect a set of data  $\{(y_i, X_i)\}_{i=1}^n$  for a set of observations.
2. Using knowledge of the topic, select a **class** of models  $\mathcal{F}$  that you want to select from  
→ That is,  $\hat{f}$  must be one of the functions in  $\mathcal{F}$ , e.g. linear functions of  $X_i$
3. Using your data, select  $\hat{f} \in \mathcal{F}$  that *predicts*  $y$  “best”  
→ “best” is defined by a **loss function**, e.g. mean-squared prediction error

# Selecting class of models

Let's break this down. First, we have to select a class of models  $\mathcal{F}$

- This involves selecting variables we want to include in the model
- Specifying a functional form for the model

# Selecting class of models

Let's break this down. First, we have to select a class of models  $\mathcal{F}$

- This involves selecting variables we want to include in the model
- Specifying a functional form for the model

For example, we might think about a simple *linear model* of our  $X$  variables:

$$f(X_i) = \alpha + X_{i1}\beta_1 + X_{i2}\beta_2$$

$\mathcal{F}$  would consist of all the models of this form, i.e. we are selecting over values of  $(\alpha, \beta_1, \beta_2)$



# title

For example, we might think about a simple *linear model* of our  $X$  variables:

$$f(X_i) = \alpha + X_{i1}\beta_1 + X_{i2}\beta_2$$

This is a restrictive model:

- no polynomials of  $X_1$  (e.g. wages are non-linear in age)
- no interaction between  $X_1$  and  $X_2$  (e.g. college degree changes return to experience)

## Selecting class of models

Perhaps, we want to look within a wider class of  $\mathcal{F}$ :

$$f(X_i) = \alpha + X_{i1}\beta_1 + X_{i1}^2\beta_2 + X_{i2}\beta_3 + X_{i2}^2\beta_4 + X_{i1}X_{i2}\beta_5 + u_i$$

$\mathcal{F}$  consists of all functions of this form

- some of the coefficients can be 0, so this class is *strictly* more general than the last.

## Selecting class of models

Perhaps, we want to look within a wider class of  $\mathcal{F}$ :

$$f(X_i) = \alpha + X_{i1}\beta_1 + X_{i1}^2\beta_2 + X_{i2}\beta_3 + X_{i2}^2\beta_4 + X_{i1}X_{i2}\beta_5 + u_i$$

$\mathcal{F}$  consists of all functions of this form

- some of the coefficients can be 0, so this class is *strictly* more general than the last.

Can imagine creating a bunch more terms or having things other than quadratics

- e.g.  $\mathbb{1}[20 < \text{Age}_i \leq 30]$

# Prediction Error

We want to be able to evaluate which model in our selected class,  $\mathcal{F}$ , does the “best” job at predicting  $y$ .

Given a model  $f \in \mathcal{F}$ , we want to evaluate how good our model does at predicting observations  $y$ . For this, define the **prediction error** as

$$\hat{\varepsilon} = \underbrace{y}_{\text{true value}} - \underbrace{f(X)}_{\text{predicted value}}$$

→ The prediction error depends on the choice of model  $f \in \mathcal{F}$

# Prediction Error

$$\hat{\varepsilon} = \underbrace{y}_{\text{true value}} - \underbrace{f(X)}_{\text{predicted value}}$$

Large  $\hat{\varepsilon}$  means you did a poor job of predicting that observation. That could be because of

1. **Reducible errors:** The model is bad at predicting  $y$ , i.e.  $f(X) \neq f_0(X)$
2. **Irreducible errors:** Or, the true noise  $\varepsilon$  is making  $y$  far away from the systematic component  $f(X)$  for this observation

# Prediction Error

We can rewrite our prediction error as

$$\begin{aligned}\hat{\varepsilon} &= y - f(X) \\ &= f_0(X) + \varepsilon - f(X) \\ &= \underbrace{f_0(X) - f(X)}_{\text{reducible}} + \underbrace{\varepsilon}_{\text{irreducible}}\end{aligned}$$

**Remember:** we do not know  $f_0$ , so we can not separate the two.

# Loss functions

To provide a summary measure of fit, we want to *average* prediction error over many observations. This will find a 'average' prediction error

If we took the simple mean of prediction error, positive and negative prediction errors would cancel out

- An error of -1 and 1 would be just as bad as -4 and 4.

## Loss functions: Mean-squared prediction error

The most common loss-function is the **mean-square (prediction) error** (MSE):

$$\text{MSE} \equiv \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 \quad (1)$$



## Loss functions: Mean-squared prediction error

The most common loss-function is the **mean-square (prediction) error** (MSE):

$$\text{MSE} \equiv \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \sum_{i=1}^n \hat{\varepsilon}_i^2 \quad (1)$$

If we collect  $\hat{\varepsilon}_i$  as a vector, we can use linear algebra to more simply write it as

$$\text{MSE} = \frac{1}{n} \hat{\varepsilon}' \hat{\varepsilon}$$

# Mean-square prediction error

$y_i$	$\hat{y}_i$	$\hat{\varepsilon}_i$
3.7	4.20	
4.1	4.18	
5.6	5.48	
2.9	3.29	
8.8	8.81	

Calculate mean-square prediction error:

## Mean-square prediction error

$y_i$	$\hat{y}_i$	$\hat{\varepsilon}_i$
3.7	4.20	0.5
4.1	4.18	0.08
5.6	5.48	-0.12
2.9	3.29	0.39
8.8	8.81	0.01

Calculate mean-square prediction error:

$$\begin{aligned}\text{MSPE} &= \frac{1}{5} (0.5^2 + 0.08^2 + (-0.12)^2 + 0.39^2 + 0.01^2) \\ &= 0.0846\end{aligned}$$

# Loss functions

The mean-squared prediction error is not the only loss-function:

- The mean-absolute prediction error  $\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$   
→ Will estimate the *median* of  $y$  given  $X$

# Loss functions

The mean-squared prediction error is not the only loss-function:

- The mean-absolute prediction error  $\frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$   
→ Will estimate the *median* of  $y$  given  $X$
- Imagine a setting where you're predicting whether someone has a disease; you would want to penalize false-negatives more than false-positives

## In-sample vs. Out-of-sample prediction error

As a forecaster, you will **fit** a model using a set of observations  $\{(x_1, y_1), \dots, (x_n, y_n)\}$ . This is called the **training data**.

We can calculate the **in-sample MSE** by formula (1) averaging over all observations in the training data.

- This tells us how good we do at predicting the data *we trained the model on*.

## In-sample vs. Out-of-sample prediction error

If our goal is prediction, we really want to know how the model would predict on *new* observations that we *have not seen before*

- It is common to hold out a set of **test data** that is NOT used for training the model, but just for evaluating it's performance

## Why use 'test data'?

It is common to try and 'pick' from a set of models based on how they do at in-sample prediction:

- That is, select the model with the smallest *in-sample MSE*.



# Why use 'test data'?

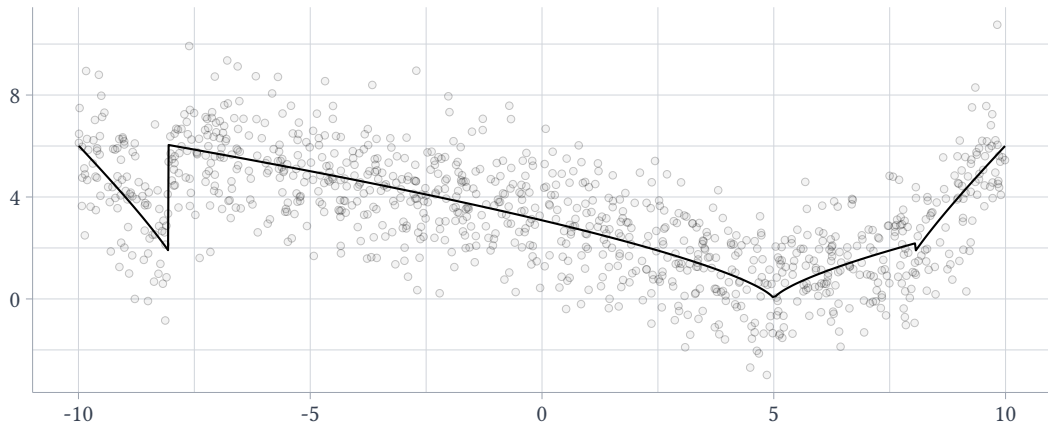
It is common to try and 'pick' from a set of models based on how they do at in-sample prediction:

- That is, select the model with the smallest *in-sample MSE*.

This is *a bad thing to do*; by focusing on fitting the current sample very well, you are risking **overfitting** the data

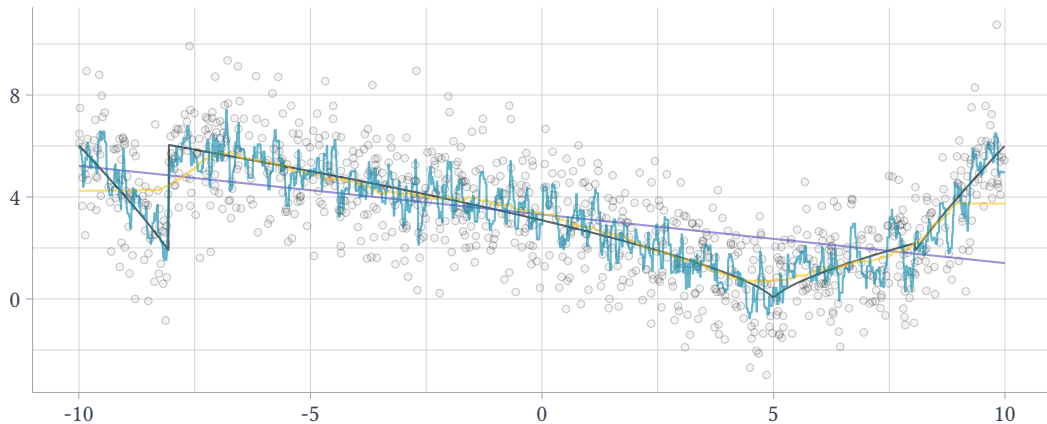
# Flexibility vs. Overfitting

True  $f(x)$



# Flexibility vs. Overfitting

True  $f(x)$ , Line, Somewhat flexible, Highly flexible



# Flexibility vs. Overfitting

By making the model more and more *flexible*, you risk overfitting more and more

- Your model tries to improve the *in-sample* mean-squared error, but it worsens your *out-of-sample* MSE.

A solution is to evaluate your model fit using outside ‘testing data’

# Sample-splitting

We will not spend much time in this course discussing sample-splitting/cross-validation and model selection, but I want to give just one example so you're aware of it

- A lot of you might recognize this from Hyunseok's class

# Sample-splitting

We will not spend much time in this course discussing sample-splitting/cross-validation and model selection, but I want to give just one example so you're aware of it

- A lot of you might recognize this from Hyunseok's class

Say you want to fit a polynomial, but are concerned with over-fitting. We can tackle this with sample-splitting:

- Take a random half your data and fit a polynomial of order  $k$
- Evaluate MSE on the other half your data (test set)

Do this for  $k = 1, \dots, K$  and pick the polynomial degree that minimizes test-data MSE

- See ILSR section 5.1 for cross-validation

# Sample-splitting

This technique is not as common when your model is more simple (e.g. regression model with a few terms)

- In some sense, you are preventing yourself from overfitting by making the model simple

# High-level of fitting models

The general framework for forecasting is as follows:

1. Collect a set of data  $(y, X)$  for a set of observations.
2. Using knowledge of the topic, select a **class** of models  $\mathcal{F}$  that you want to select from  
→ That is,  $\hat{f}$  must be one of the functions in  $\mathcal{F}$ .
3. Using your data (*perhaps on a training sample*), select  $\hat{f} \in \mathcal{F}$  that minimizes the **loss function**  
→ E.g. mean-squared prediction error (*perhaps on testing sample*)



## Bias-variance trade-off

This discussion of increasing flexibility leading to increasing the noise of the model fit is a well-known problem. It is called the **Bias-Variance Tradeoff**:

1. **Bias**: When the model we fit,  $\hat{f}(x)$ , does a poor job fitting the true model  $f_0(x)$
2. **Variance**: The variability of the model we fit,  $\hat{f}(x)$ , across samples
  - Repeated sampling: the model we estimate varies from estimate to estimate

# Bias-variance trade-off

This discussion of increasing flexibility leading to increasing the noise of the model fit is a well-known problem. It is called the **Bias-Variance Tradeoff**:

1. **Bias**: When the model we fit,  $\hat{f}(x)$ , does a poor job fitting the true model  $f_0(x)$
2. **Variance**: The variability of the model we fit,  $\hat{f}(x)$ , across samples
  - Repeated sampling: the model we estimate varies from estimate to estimate

This is a 'trade-off'. To lower bias by adding flexibility, you're adding variance (noisiness) to the estimate

# Model Selection

Our general approach seems to follow:

- Select class of models to choose from,  $\mathcal{F}$ .
- Find  $\hat{f} \in \mathcal{F}$  that minimizes the in- or out-of- sample MSPE

The secret sauce of forecasting is in selecting  $\mathcal{F}$

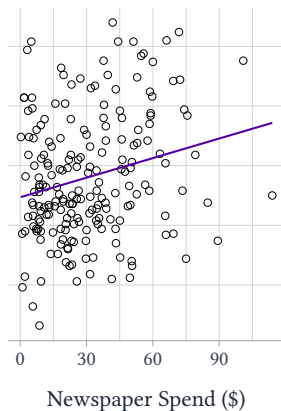
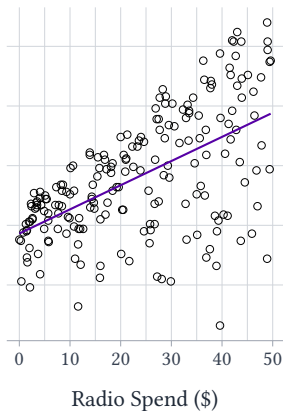
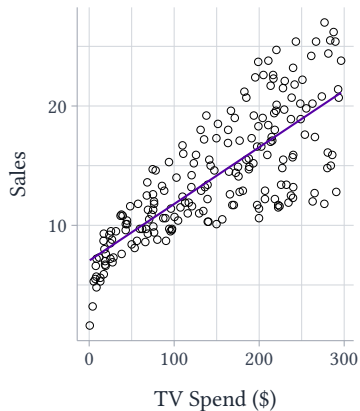
- E.g. sports teams all have a ton of data to help them draft the best players; all will minimize MSPE to fit their model; all will do sample-splitting; etc.
  - The advantage is who can pick the best **variables** to include in their model
  - Often called “**feature selection**”

## Model Selection: Advertising Example

Say you're a business and you want to use advertising to boost sales. You have a bunch of different markets (e.g. cities) and you have data on how you've spent your advertising budget in those markets and the sales in that market

# Advertising Example

*Single-variable predictors*



## Advertising Example

We see that sales are higher in markets that have spending on TV, Radio, and Newspaper ads separately.

# Advertising Example

We see that sales are higher in markets that have spending on TV, Radio, and Newspaper ads separately.

These single scatter plots with line of best-fits are a somewhat poor model:

- Are there synergies between different advertising strategies (are they substitutes or complements to one another)?
- Do places with more TV ads also have more radio ads? Then how can we tell if it is TV ads that are helping or if it is really radio ads

# Advertising Example

We see that sales are higher in markets that have spending on TV, Radio, and Newspaper ads separately.

These single scatter plots with line of best-fits are a somewhat poor model:

- Are there synergies between different advertising strategies (are they substitutes or complements to one another)?
- Do places with more TV ads also have more radio ads? Then how can we tell if it is TV ads that are helping or if it is really radio ads

**Key takeaway:** Forecasting models get better the more carefully you think about the context you are in



# Interactions Matter

Over the next few weeks, we will learn a lot about regression methodology. We will do so for a set of covariates,  $X_i$ .

- These could be a set of variables like age, height, batting average, etc.
- But, these could also be functions of variables, e.g.  $\mathbb{1}[\text{Age}_i = 21]$  or  $\text{height} \times \text{weight}$

It is important to remember that the world can feature a lot of non-linearities and interactive effects

- Your model should reflect those too!

## **Forecasting**

### **Goals of Forecasting**

### **Fitting Models**

Model Selection: Adveristing Example

### **Types of Data**

## Cross-sectional Data

**Cross-sectional data** consists of many different *units* viewed at a point in time:

school_id	avg_sat_math	pct_white	pct_black
01M539	657	28.6%	13.3%
02M294	395	11.7%	38.5%
02M308	418	3.1%	28.2%
02M545	613	1.7%	3.1%
01M292	410	3.9%	24.4%
01M696	634	45.3%	17.2%

# Time-series Data

**Time-series data** consists of a single observational unit viewed over multiple points in time:

month	day	hour	bikers	temp
1	1	0	16	0.24
1	1	1	40	0.22
1	1	2	32	0.22
⋮	⋮	⋮	⋮	⋮
12	31	21	52	0.40
12	31	22	38	0.38
12	31	23	31	0.36

# Panel Data

**Panel data** is like time-series data,  
but for many different observational  
units:

fund_manager	month	return
1	1	-3.34%
1	2	3.76%
1	3	12.97%
⋮	⋮	⋮
2000	48	-3.76%
2000	49	2.25%
2000	50	6.68%