Topic 1: Introduction to Forecasting

ECON 4753 — University of Arkansas

Prof. Kyle Butts

Fall 2024

Roadmap

Forecasting

Goals of Forecasting

Evaluating Models

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

Prediction model

We have an outcome variable Y and a set of p different predictor variables $X = (X_1, X_2, \ldots, X_p)$.

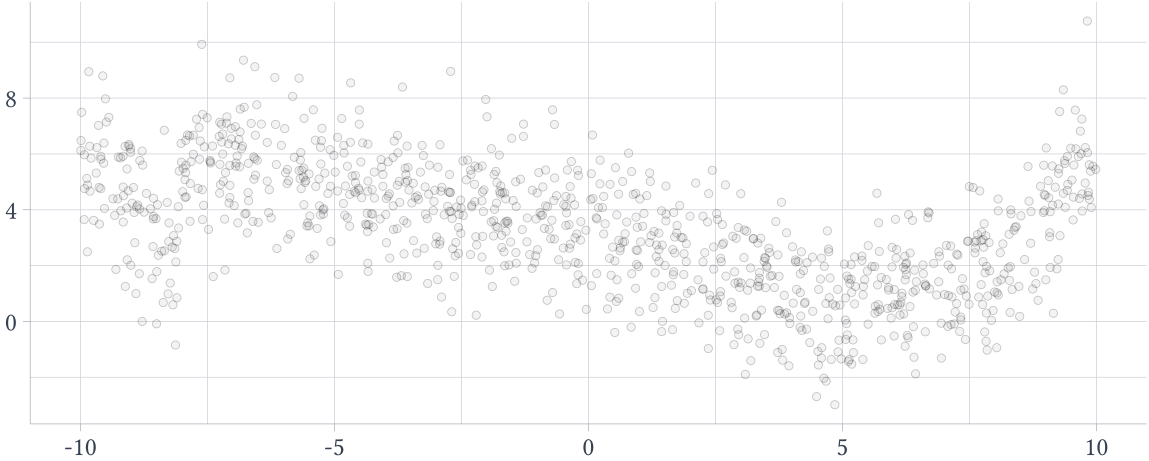
- For some observations we observe both X and Y; this is essential to fit the model

We can write the model in a general form as

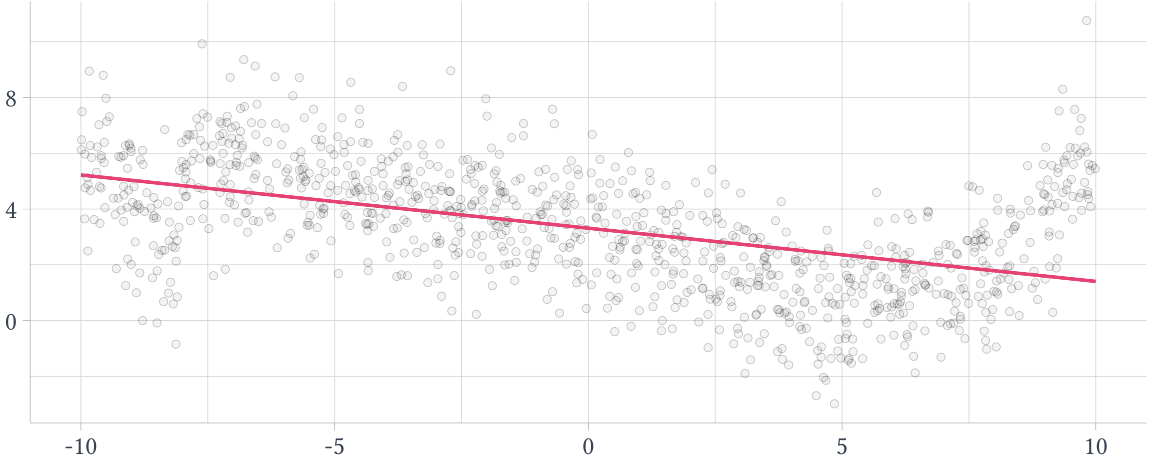
$$Y = f(X) + \varepsilon,$$

where f is some unknown function of X and ε is the error term which we assume is unrelated to X and mean zero.

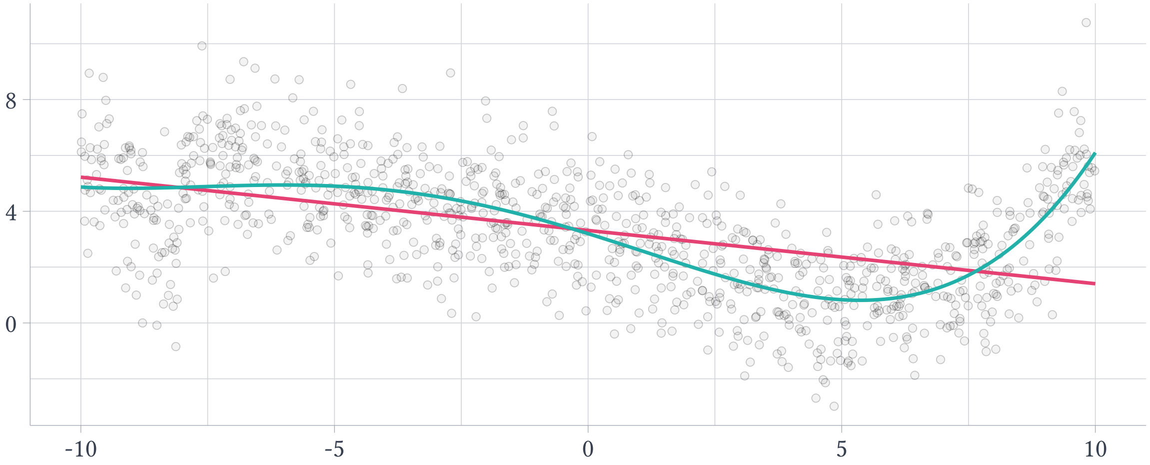
Examples of f:



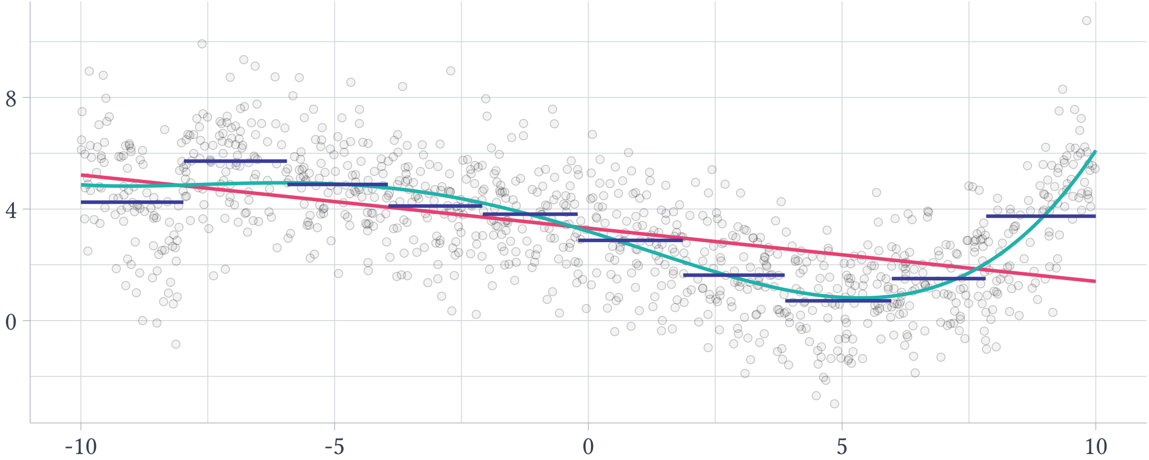
Examples of f: Line



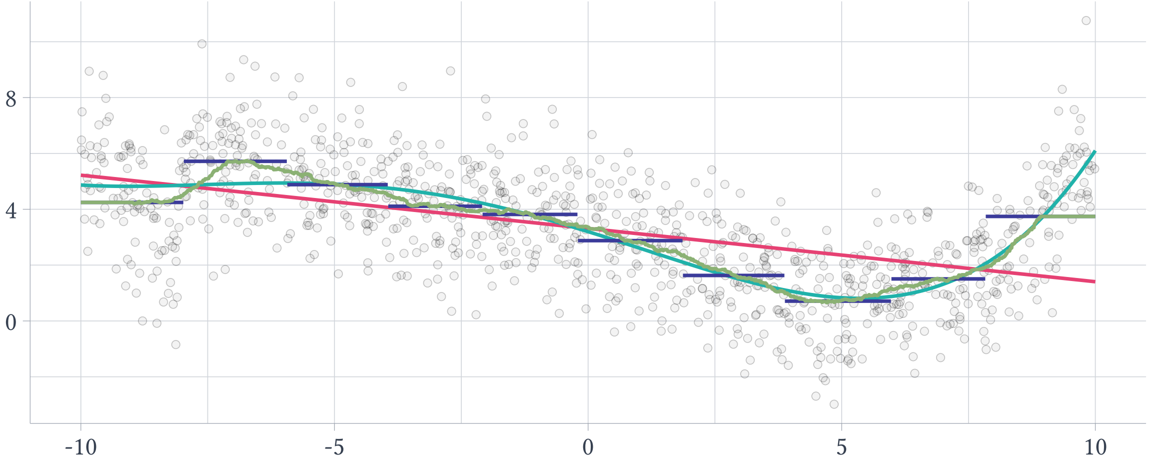
Examples of f: Line, Polynomial (x^4)



Examples of f: Line, Polynomial $\left(x^4\right)$, Bins of x



Examples of f: Line, Polynomial $\left(x^4\right)$, Bins of x, KNN of x

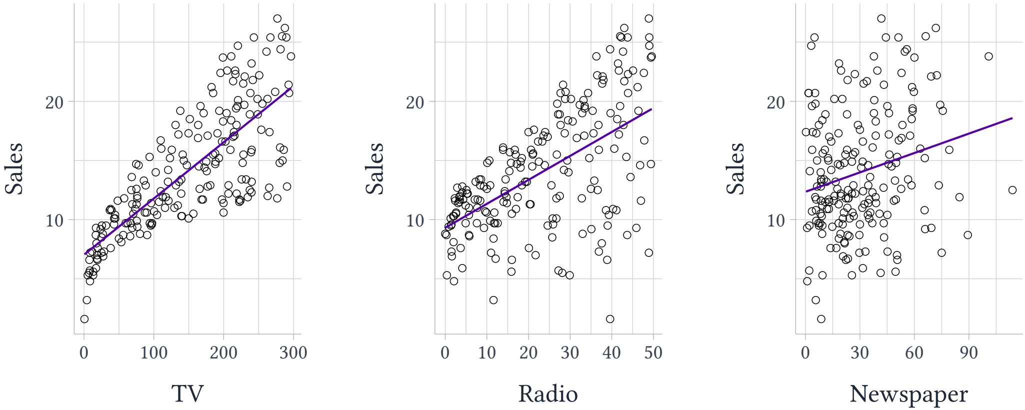


Advertising Example

Let's given an 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

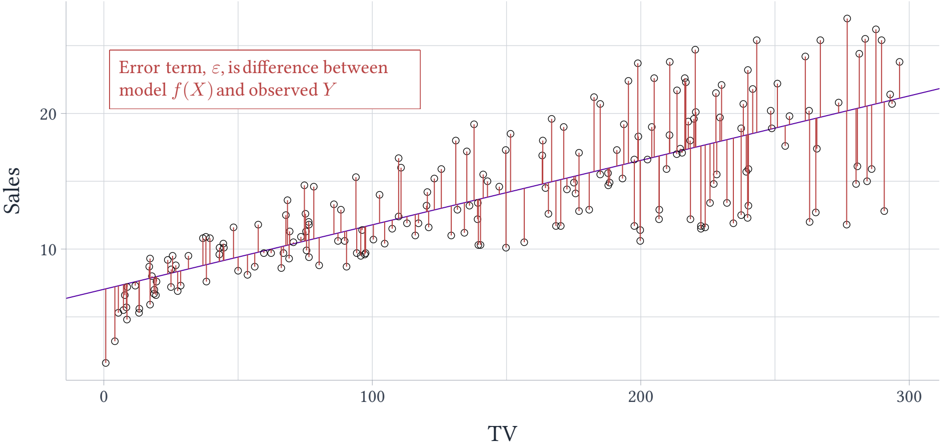
We see that sales are higher in markets that have spending on TV, Radio, and Newspaper ads separately.

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

Error term



Error term, ε, is difference between model $f(X)$ and observed Y
20
10
Sales
0
100
200
300
TV

Error term

In the previous figure, we were able to determine ε because we assumed we *knew* the model f and therefore could observe $f(X)$ for each market.

In reality, we do not know f and can never observe ε. But, we can try and estimate it ...

Roadmap

Forecasting

Goals of Forecasting

Evaluating Models

Types of Data

Why estimate f?

There are two related reasons to try and predict f:

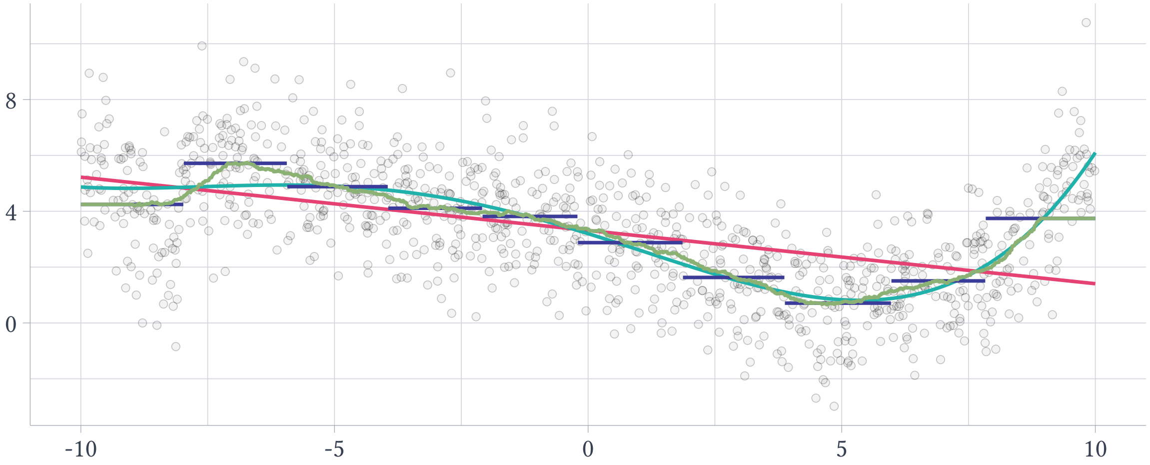
1. Predict y as good as possible (prediction)
   - $\rightarrow$ Think of prediction as a 'black box' where the goal is to do as good of a job at predicting y as possible
2. Understand the relationship between x and y (inference)
   - $\rightarrow$ If our goal is being able to describe the relationship between x and y.

Model Flexibility

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

1. If our goal is prediciton, 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
   $\rightarrow$ Face the risk of overfitting the data (chasing after the random noise ε)
2. If our goal is inference, then added flexibility is harder to summarize to stakeholders.

Examples of f: Line, Polynomial $\left(x^4\right)$, Bins of x, KNN of x



Roadmap

Forecasting

Goals of Forecasting

Evaluating Models

Types of Data

Prediction Error

Given our model, we will want to be able to evaluate how good our model does at predicting observations y

Define the prediction error as

$$\hat{\varepsilon} = \underbrace{y}_{\text{true value}} - \underbrace{\hat{y}}_{\text{predicted value}}$$

Prediction Error

$$\hat{\varepsilon} = \underbrace{y}_{\text{true value}} - \underbrace{\hat{y}}_{\text{predicted value}}$$

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

1. The linear model is bad at predicting y
2. Or, the true noise ε is making y far away from the systematic component $f(X)$ for this observation

Mean-square prediction error

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

Mean-square prediction error

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

The mean-square (prediction) error (MSE) is calculated as:

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

- Average of squared prediction error

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:

$$\text{MSPE} = \frac{1}{5}\left(0.5^2 + 0.08^2 + -0.12^2 + 0.39^2 + 0.01^2\right)$$
$$= 0.0846$$

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

As a forecaster, you will fit a model using a set of observations $\{(x_1, y_1), \ldots, (x_n, y_n)\}$. 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*.

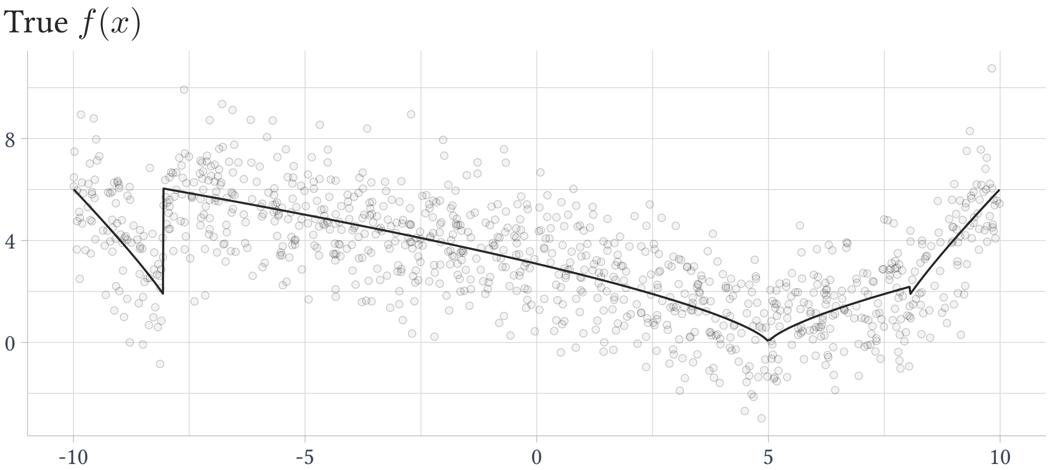
Why use 'test data'?

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*.

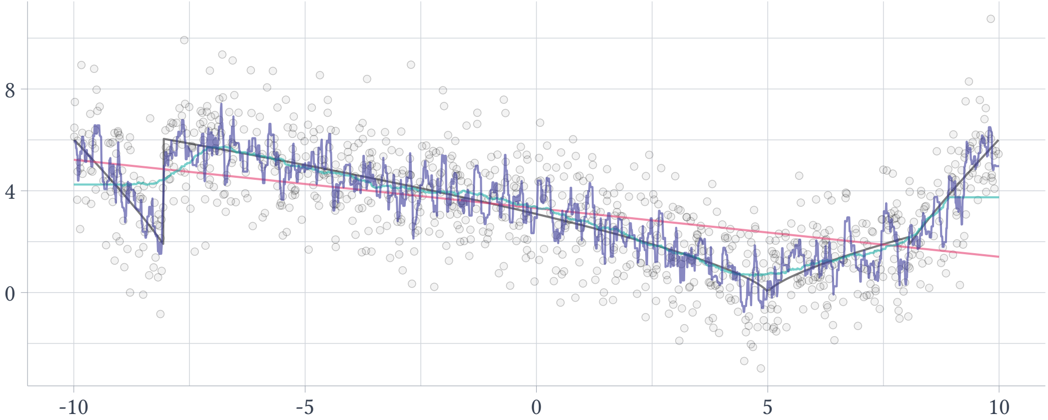
This is *a bad thing to do*; by focusing on fitting the current sample very well, you are risking overfitting the data

Flexibility vs. Overfitting



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

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

Flexibility vs. Overfitting

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

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

This technique is not as common when you care more about the associations between variables (interpreting the model)

- Not really a good reason other than "that is more complicated"

Bias-variance trade-off

This discussing of increasing flexiblity 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 $f(x)$
2. Variance: When the model we fit, $\hat{f}(x)$, is very variable across samples

    $\rightarrow$ In repeated sampling, the model we estimate varies from estimate to estimate

Bias-variance trade-off

This discussing of increasing flexiblity 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 $f(x)$
2. Variance: When the model we fit, $\hat{f}(x)$, is very variable across samples
   $\rightarrow$ In 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

Roadmap

Forecasting

Goals of Forecasting

Evaluating Models

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_b
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
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
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:

hedge_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%