Quantitative evaluation

Potential Outcomes Framework and the Assignment Mechanism.

Eduardo Fé

Part I

Introduction.

Some admin...

Eduardo Fé
Senior Lecturer Social Statistics, University of Manchester
eduardo.fe@manchester.ac.uk

Some admin...

We start at **9:30** everyday.

Morning sessions 9:30 to 12:30, coffee break 10:30-11:00

Lunch 12:30-13:30 in ALB, Common Room

Afternoon session: 1:30 to 5, coffee break 3-3:30

Some admin...

Schedule

Tue. Mor.: RCT and Lab 1 (Intro to R);

Tue. Aft.: Randomization Inference and Lab 2 (RCT);

Wed. Mor.: Instrumental Variables and Lab 3 (RI);

Wed. Aft.: DiD and Lab 4 (IV)

Thu. Mor.: RD 1 and Lab 5 (DiD)

Thu. Mor.: RD 2 and Lab 6 (RD)

Fri: Mor.: Capstone project

Introduction.

In the beginning...

- The origin of statistics was tightly intertwined with modern state's efforts to objectively quantify their wealth and health
- The goal was to build a solid foundation for policy making (taxing in particular)
- So, from its origins statistics has thus **implicitly** sought to answer questions about the **causes** and mechanisms of the world.

Introduction.

Paradoxically, **on its own** the traditional *language* of statistics is not designed to provide answers to causal questions

The focus of statistics has been data reduction (estimation) and modelling.

- Good estimation techniques are critical but on their own they lack a causal interpretation
- Similarly, models on their own only reveal only researcher's own view of underlying causal mechanisms.

Introduction.

In recent times, the statistical tool kit has been expanded with a new methodology/language specifically designed to uncover the working causal mechanisms of the world.

- Foundational contributions during the twentieth century, and particularly since the 1970s
- Have introduced and standardised the formal statistical language of causality.
- The **silent 'causal revolution'**: researchers' capabilities to undertake data-based analysis of causal questions and, critically, understand the premises and limitations of these studies, has been substantially magnified.

Introduction

Two (closely related) schools of thought exit when it comes to articulating a framework to develop statistical analysis of causes and effects,

- **Potential outcomes framework** (Rubin, 1974; Holland, 1986)
  - Clearly define the parameter of interest
  - Explicitly spells out the assumptions that will underlie the interpretation of subsequent estimates
  - Estimation methods are, to some extent, of secondary concern
- **Graphical theory of causality** (Balke and Pearl, 1997; Pearl, 2009)
  - Directed acyclic graphs are used to to build a model describing the main causal relationships underlying the problem
  - Estimation is a central problem in this approach (and, although classical approaches are suitable, Bayesian methods are more natural in this setting).

In this course we will work within the potential outcomes framework.

Potential outcomes framework

Origins in, at least, Neyman's 1923 masters dissertation[1] but popularised Rubin (1974), Holland (1986) (see also Rubin, 1990a).

Provides a unifying framework for both randomized experiments AND observational studies.

Throughout, we focus in the standard situation where there are two '*experimental*' conditions[2]: **Treatment** and **Control**.

[1] '*Sur les applications de la theorie des probabilites aux experiences agricoles: Essai des principes*'. Translated in Neyman, Dabrowska, and Speed (1990); see also Neyman (1934) and the commentary in Rubin (1990b).

[2] Here 'experimental' must be broadly understood. We are not circumscribing the study to randomized trials at all. The word 'experimental' could refer to a no-randomized policy or intervention implemented by government, a phenomenon of interest (such as child labour, specific criminal behaviours, market labour statuses, etc)

Potential outcomes framework

Three key ingredients.

Unit of analysis: An entity (person, firm, classroom, cell, etc) with a measurable trait of interest *at a specific point in time*

(Active) Treatment: an action or manipulation that might affect a particular unit. If a unit does not receive the treatment, then the unit must have received the **Control (treatment)**.

Potential outcomes: Associated with each unit there are two potential outcomes at a **future** point in time, which specify the level of some outcome under the treatment/control conditions.

Potential outcomes framework

Three key ingredients. Suppose we are interested in an **outcome** (characteristic) Y (e.g. blood pressure)

Unit of analysis: $i = 1, 2, \dots, N$

(Active) Treatment, T_i such that $T_i = 1$ if unit i received the treatment; otherwise, $T_i = 0$. For instance, $T_i = 1$ if i receives statins[3]; $T_i = 0$ otherwise.

Potential outcomes: $Y_i(1)$ is unit i's potential outcome (blood pressure) under the active treatment (statins); $Y_i(0)$ is unit i's potential outcome under the control treatment (no statins).

It follows that the observed level of Y for unit i is

$$Y_i = Y_i(1) \cdot T_i + Y_i(0) \cdot (1 - T_i) \tag{1}$$

[3] A medication designed to lower cholesterol in blood

Potential outcomes framework: Defining causal effects (I)

The causal effect of the treatment T for unit i results from **comparing** $Y_i(1)$ and $Y_i(0)$.

For example, all these are valid **definitions** of **unit level** causal effects:

- $\tau_{i,1} = Y_i(1) - Y_i(0)$
- $\tau_{i,2} = Y_i(1)/Y_i(0)$
- $\tau_{i,3} = \log Y_i(1) - \log Y_i(0)$

Potential outcomes framework: Defining causal effects (I)

Importantly:

- The **definition** of the unit level causal effect depends **only** on the potential outcomes, but **not** on which potential outcome is observed.
- The definition of a causal effect involves the comparison of i's potential outcomes **at the same point in time** and **after receiving the treatment**.
- In particular, the treatment effect is not defined by comparing potential outcomes at different points in time (e.g. before and after treatment).

Potential outcomes framework: The problem of selection

The potential outcomes framework is ideal to unveil the fundamental problem of causal inference (the problem of selection):

Unit i's potential outcomes cannot be observed simultaneously.

Specifically, a unit can only reveal one potential outcome at a time.

Although the potential outcomes framework helps to clearly defined the unit level causal effect, the latter is not identified or estimable from the data

So, what do we do?!

Potential outcomes framework: The problem of selection

The (compulsory) standard solution is to collect data from units that are subject to the active treatment and units that are subject to the control treatment.

This raises two principal questions

1. How do we define causal effects now?
2. What has determined that some units are subject to the active treatment -and thus reveal $Y_i(1)$- and some units to the control treatment -and thus reveal $Y_i(0)$?

2’ ... or said differently: to what extent are the units in the treatment and control groups as comparable?

Potential outcomes framework: Multiple units.

Consider, first, the problem of defining treatment effects under multiple units.

Suppose we have only 2 units in the sample.

Then there are four treatment *pairs*, (T_1, T_2):

- $(1, 1)$: both units receive the active treatment
- $(0, 0)$: both units receive the control treatment
- $(1, 0)$ or $(0, 1)$ one unit receives the active treatment and the other the control treatment.

...which results in 8 potential outcomes $Y_i(t_1, t_2)$ (four per individual: $Y_i(1, 1)$, $Y_i(1, 0)$, $Y_i(0, 1)$ and $Y_i(0, 0)$ for $i = 1, 2$)...

Potential outcomes framework: Multiple units.

...but this means that we can define many unit causal effects based on several meaningful comparisons of potential outcomes $Y_i(t_1, t_2)$:

- $Y_1(1, 0)$ with $Y_1(0, 0)$
- $Y_1(1, 0)$ with $Y_1(0, 1)$
- $Y_1(1, 1)$ with $Y_1(0, 0)$
- $Y_1(1, 1)$ with $Y_1(0, 1)$
- ...

This creates uncertainty in the definition. The solution is to **assume out** some of these potential outcomes...

Potential outcomes framework: SUTVA

To define a causal parameter when multiple units are considered, we need to:

- Specify if treatments are comparable across units
- Specify if the treatment status of a unit can affect the outcomes of other unit(s)

This information is normally not available, therefore we need to **assume** the terms of these relationships.

The most common assumption in the literature is the Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1980).

Potential outcomes framework: SUTVA

Stable Unit Treatment Value Assumption (SUTVA; Rubin, 1980).

Assumption 1: (SUTVA) There is no interference between units and there is no versions of the treatments leading to 'technical errors'.

Two restrictions:

No interference: Unit i's outcomes do not depend on the treatments that other units received.

No versions of the treatments: Implies no hidden variation in treatments (only one version of each treatment).

Potential outcomes framework: SUTVA

SUTVA No interference.

This was a main concern in early work in agricultural experiments (Neyman et al., 1935, Fisher, 1935, Rubin, 1990b). In those settings, ‘guard rows’ were used to separate experimental plots and avoid contamination (interference).

This can be a strong assumption in observational studies. For instance,

- Immunisation, where inoculation of vaccines to substantial sub-populations can generate positive externalities for the non-vaccinated)
- Evaluation of school policies, where there may be potential spillovers between students in the same year/school.

Potential outcomes framework: SUTVA

SUTVA No versions of the treatments.

Again, this can be problematic in practice.

- For instance, if testing a new pill, the concentration in all the pills must be the same.
- If studying retirement, some individuals might be partially retired, and variation in part time work might be substantial.

This component of SUTVA is very often not explored in research (a mistake).

A note on assumptions.

It is important to bear in mind that

- Assumptions are essential to causal inference
- But these assumptions are generally not testable -thus always open to criticism
- In empirical research you need to be clear about
  - **All** assumptions are problematic
  - But some assumptions are more credible than others
  - Don't try to convince readers about why your assumptions are right (for every assumption you make, there exists at least 1 researcher that will have an issue with it)...
  - Rather discuss the pros and cons of the assumptions underpinning your work.
  - An always try to work with credible assumptions for your specific application.

Potential outcomes framework: SUTVA

We will maintain SUTVA throughout this course.

Suppose that you have $i = 1, 2$ units in your data. Under SUTVA, the potential outcomes $Y_i(T_1, T_2)$ can now be rewritten as

$$Y_i(T_1, T_2) = Y_i(T_i) \text{ for all } i = 1, 2 \tag{2}$$

by virtue of no interference.

More generally, consider $i = 1, 2, \ldots, N$ units. Let $\mathbf{T} = (T_1, T_2, \ldots, T_N)'$ be the $N \times 1$ vector of treatments for the full sample. Then, SUTVA implies that

$$Y_i(\mathbf{T}) = Y_i(T_i) \tag{3}$$

Note that SUTVA was already implicit in equation (1).

Potential outcomes framework: Assignment Mechanism.

Despite its strength, SUTVA does nothing to ameliorate the fundamental problem of selection (missing data on one potential outcome per unit).

Selection cannot be resolved unless we incorporation of carefully chosen assumptions

In this respect, the most critical assumptions will be those contributing to (or shaping our) understanding of the process that determined which units were allocated to the active treatment or control treatment.

That process is known as the **assignment mechanism**.

Potential outcomes framework: Assignment Mechanism.

The assignment mechanism is a characterisation of the likelihood of each feasible assignment of units into treatment and control groups.

Specifically, the assignment mechanism

1. quantifies the probability of each assignment.
2. establishes if that probability depends on the potential outcomes or the values of some other variables.

Potential outcomes framework: Assignment Mechanism.

Running example.

Consider, to begin with, a 2 unit case, $N = 2$. There are four (that is 2^N) assignments to treatment:

$$\mathbf{T}_1 = (1, 1)', \mathbf{T}_2 = (1, 0)', \mathbf{T}_3 = (0, 1)', \mathbf{T}_4 = (0, 0)'.$$

1. The assignment mechanism quantifies the probability of $\mathbf{T}_j$, say π_j, for $j = 1, ..., 4$.

For instance, if we allocate units to treatment/control by flipping a fair coin, then $\pi_1 = \pi_2 = ... = \pi_4 = 1/4$

May be, however, only one unit can be allocated to the active treatment. In this other example $\pi_1 = \pi_4 = 0$ but $\pi_2 = \pi_3 = 1/2$

Potential outcomes framework: Assignment Mechanism.

Running example.

Consider, to begin with, a 2 unit case, $N = 2$. There are four (that is 2^N) assignments to treatment:

$$\mathbf{T}_1 = (1,1)', \mathbf{T}_2 = (1,0)', \mathbf{T}_3 = (0,1)', \mathbf{T}_4 = (0,0)'.$$

2. The assignment mechanism specifies if the likelihood of treatment depends on the potential outcomes or some other variables. Define $\mathbf{Y}(0) = (Y_1(0), Y_2(0))'$, $\mathbf{Y}(1) = (Y_1(1), Y_2(1))'$ and let $\mathbf{X}$ be a set of variables that might determine treatment

Then the assignment mechanism defines, for $j = 1, \dots, 4$

$$\pi_j = P(\mathbf{T}_j | \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) \tag{4}$$

This notation means that the likelihood of assignment $\mathbf{T}_j$ depends on all the potential outcomes of all units, as well as the value of the variables in $\mathbf{X}$ for all units.

Potential outcomes framework: Assignment Mechanism.

Running example.

Consider, to begin with, a 2 unit case, $N = 2$. There are four (that is 2^N) assignments to treatment:

$$\mathbf{T}_1 = (1,1)', \mathbf{T}_2 = (1,0)', \mathbf{T}_3 = (0,1)', \mathbf{T}_4 = (0,0)'.$$

2. The assignment mechanism specifies if the likelihood of treatment depends on the potential outcomes or some other variables. Define $\mathbf{Y}(0) = (Y_1(0), Y_2(0))'$, $\mathbf{Y}(1) = (Y_1(1), Y_2(1))'$ and let $\mathbf{X}$ be a set of variables that might determine treatment

If assignment was defined by, for instance, flipping a fair coin, then the assignment would not depend on the potential outcomes or $\mathbf{X}$, and then

$$\pi_j = P(\mathbf{T}_j | \mathbf{Y}(0), \mathbf{Y}(1), \mathbf{X}) = P(\mathbf{T}_j) \tag{5}$$

Potential outcomes framework: Assumptions on the Assignment Mechanism.

Individualistic assignment. A unit's treatment status does not depend on the outcomes and assignments of other units.

Warning: It rules out spill overs (externalities, peer effects, partner effects) which are common in everyday life.

Potential outcomes framework: Assumptions on the Assignment Mechanism.

Probabilistic assignment. This means that any unit has a non-zero probability of being allocated to the treatment or control group

Warning: Some subpopulations might have 0/1 likelihood of assignment.

Potential outcomes framework: Assumptions on the Assignment Mechanism.

Unconfounded assignment: The assignment mechanism is unconfounded if it does not depend on units' potential outcomes. In other words, for any unit, $T_i \perp (Y_i(0), Y_i(1))$.

Warning: This is **THE** critical assumption and it is a strong one. It rules out situations where units self-select into treatment on the basis of perceived advantages or disadvantages derived from the treatment.

Potential outcomes framework: Assumptions on the Assignment Mechanism.

The combination of SUTVA, with the assumptions of an individualistic, probabilistic and unconfounded assignment will be enough to resolve the problem of selection. These assumptions will enable us to use simple statistics to estimate causal effects.

However, unconfounded assignment is rarely encountered in natural environments; to relax this assumption (and thus increase the credibility of our causal analysis) we will need to introduce further restrictions on the assignment mechanism by restricting the process through which unconfoundedness is violated.

Classification of assignment mechanisms

We can offer a classification of assignment mechanisms depending on whether or not the researcher has full knowledge and control of the assignment mechanism. For practical purposes, we can distinguish between:

- **A classical randomized experiment**. An individualistic, probabilistic and unconfounded treatment assignment with a known form which is controlled by the experimenter.
- **Observational studies**. The assignment mechanism is not controlled or known by the experimenter
  - **A regular assignment mechanism**. An observational study with a individualistic, probabilistic and unconfounded assignment mechanism
  - **Irregular assignment mechanism**. For our purposes, this will be observational studies where assignment to treatment is unconfounded but the actual reception of treatment is confounded.

Classification of assignment mechanisms

We can offer a classification of assignment mechanisms depending on whether or not the researcher has full knowledge and control of the assignment mechanism. For practical purposes, we can distinguish between:

- **A classical randomized experiment**. An individualistic, probabilistic and unconfounded treatment assignment with a known form which is controlled by the experimenter.
- **Observational studies**. The assignment mechanism is not controlled or known by the experimenter
  - **A regular assignment mechanism**. An observational study with a individualistic, probabilistic and unconfounded assignment mechanism
  - **Irregular assignment mechanism**. For our purposes, this will be observational studies where assignment to treatment is unconfounded but the actual reception of treatment is confounded.

Part II

Tools of the trade.

Random variables and their distributions

In the statistician's mind, there is always a **population** that we want to study, for which we collect a sample.

Sometimes the sample = the population (finite population approach).

Sometimes the sample is seen as a draw from an infinite population (the **superpopulation** approach). This latter approach is dominant in the statistical and econometric literature (and textbooks).

In this course, we will consider both scenarios, but we now introduce some ideas pertaining to the second framework.

Random variables and their distributions

In the superpopulation framework, statisticians imagine the existence of a model that explains/determines the distribution of units' characteristics; this model explains the variation in any dataset we might draw from the superpopulation.

The hope if that data will be sufficiently informative about the model to allow us to estimate specific characteristics (**moments**) of that model.

Clearly, the underlying model is not observable directly.

Although the characteristics of the model and units might be fixed in the population, each observation in our data is treated as random variable, because we don't know, a priori, which units will be observed in the sample.

Random variables and their distributions

The **distribution** of a specific characteristics such as height, is a function that tells us how different levels of the characteristic are distributed in the superpopulation. Distributions characterise the trait in terms of **probability** (proportion).

For instance, if Y is the superpopulation characteristic ‘height’, then Y_i is unit i’s height, which is a random variable before the sample is drawn -as we don’t know who among the infinite number of units in the superpopulation this unit i is going to be

The **probability distribution** of height, describes the proportion of people with height less that a value y in the superpopulation, $F(y) = P(Y_i \leq y)$; this is a superpopulation **moment** and so **unobservable**.

Random variables and their distributions

We can estimate this distribution, once data are available, with the empirical distribution, which describes, for the sample, the proportion of observations with Y_i under a value y,

$$\hat{F}(y) = \frac{1}{N} \sum_{i=1}^{N} Y_i \cdot \mathbf{1}_{Y_i \leq y}. \tag{6}$$

where $\mathbf{1}_{Y_i \leq y}$ is an *binary indicator* or dummy variable which equals 1 if $Y_i \leq y$ and equals 0 if $Y_i > y$

Random variables and their distributions

When Y takes on a handful of values, say $y_1, \ldots, y_r, \ldots y_R$ for R a small number, then it also makes sense to talk about the probability of each of these y_r, $P(Y_i = y_r)$. This probability is, again unobservable, but we can estimate it, once data are available, with the

$$\hat{P}(y) = \frac{1}{N} \sum_{i=1}^{N} Y_i \cdot \mathbf{1}_{Y_i = y}. \tag{7}$$

where $\mathbf{1}_{Y_i=y}$ is an *binary indicator* or dummy variable which equals 1 if $Y_i = y$ and equals 0 if $Y_i > y$

Moments and conditional expected values

We will be mostly concerned about the superpopulation mean of **expected value** of characteristics, $E(Y)$. If Y takes on just a handful of values, this $E(Y)$ is defined as

$$E(Y) = \sum_{r=1}^{R} y_r \cdot P(Y = y_r) \tag{8}$$

Because in the superpopulation units have different values of Y, we will normally need a measure of how much variation in values there is. This is provided by the superpopulation variance, $V(Y)$, which if Y takes on just a handful of values, is defined as

$$V(Y) = \sum_{r=1}^{R} \Big(y_r - E(Y)\Big)^2 \cdot P(Y = y_r) \tag{9}$$

Moments and conditional expected values

When data are available, Y_i, $i = 1, \ldots, N$ the expected value and variance can normally be estimated with sample equivalents, such as

$$\bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i$$
$$s^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2 \tag{10}$$

These estimators are unbiased[4] and consistent[5] for $E(Y)$ and $V(Y)$ respectively.

[4] Their average, upon repeated sampling from the superpopulation, coincides with the corresponding superpopulation moment that they intend to estimate

[5] As the sample size grows (tends to ∞), they approach the value of the corresponding superpopulation moment that they intend to estimate.

Moments and conditional expected values

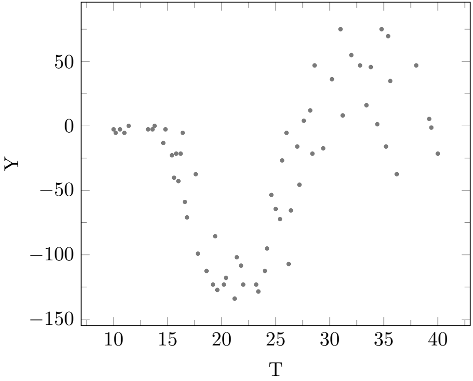
Most often through out the course, we will be interested in the mean of Y when for different values of another characteristic, T. This latter characteristic can take on a few or an ∞ number of values; to simplify the discussion, let's focus on the former case.

This is the **conditional expected value** or conditional mean of Y given $T = t$, $E\left[Y|T = t\right]$

We don't need a formal definition of this; this moment is, essentially, the mean of Y in the superpopulation for those units whose level of T is t.

Moments and conditional expected values

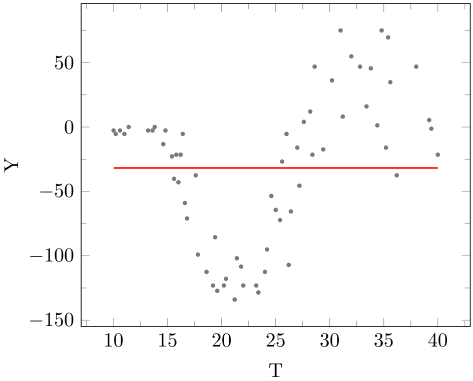
Suppose that this was a superpopulation that we could observe in its entirety



A wavy superpopulation.

Moments and conditional expected values

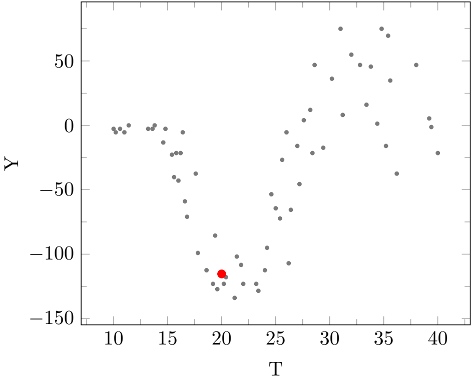
Suppose that this was a superpopulation that we could observe in its entirety



The population expected value is $E(Y) = -31.812$; it doesn't vary with X.

Moments and conditional expected values

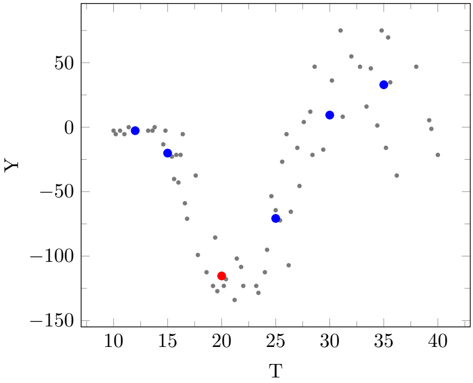
Suppose that this was a superpopulation that we could observe in its entirety



The population conditional expected value at $T = 20$ is $E(Y|T = 20) = -115.38$;

Moments and conditional expected values

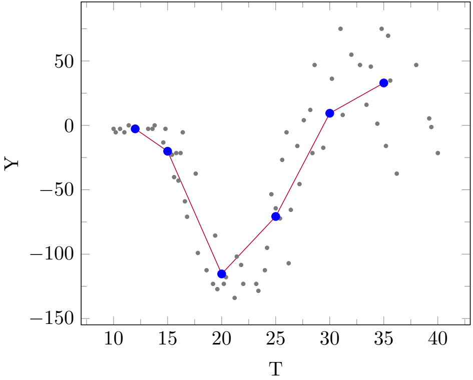
Suppose that this was a superpopulation that we could observe in its entirety



The population conditional expected value varies with T; it is function of T ...

Moments and conditional expected values

Suppose that this was a superpopulation that we could observe in its entirety



... so can write $E(Y|T = t) = g(t)$ where $g(.)$ is unknown; the purple line approximates $g(t)$

Regression and Nonparametric Regression

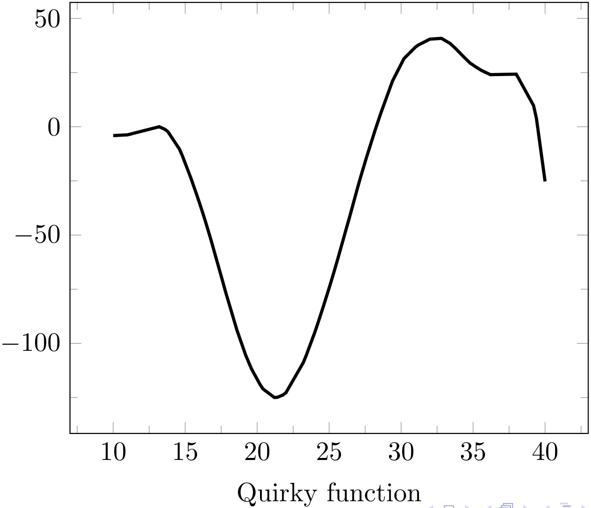
Most often, in this course, T will be binary, taking on just to values: 0 and 1. In this cases, we can estimate the unknown superpopulation $E\big[Y_i|T = t\big]$ with the sample mean for units with $T_i = t$.

However, we will sometimes need either a model characterising $E\big[Y_i|T = t\big]$ for all t or estimating $E\big[Y_i|X = x\big]$ for a continuous X. In the latter case, there might not be observations in the sample with $X = x$, in which case we need to use the data in some way to infer what $E\big[Y_i|X = x\big]$ might be

This is were **parametric** and **nonparametric** regression is helpful.

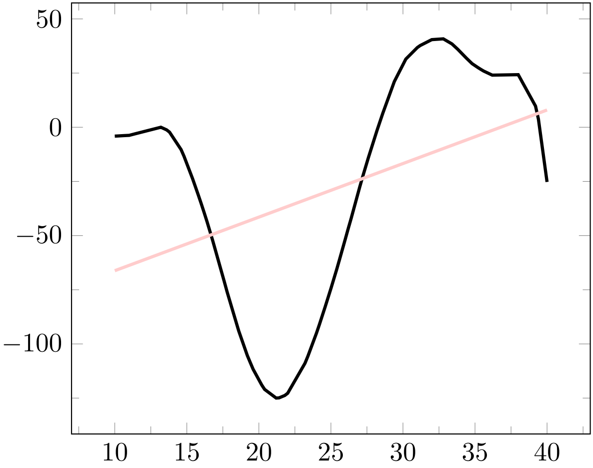
Regression and Nonparametric Regression.

Any sufficiently smooth function can be well approximated by a polynomial function globally...



Regression and Nonparametric Regression.

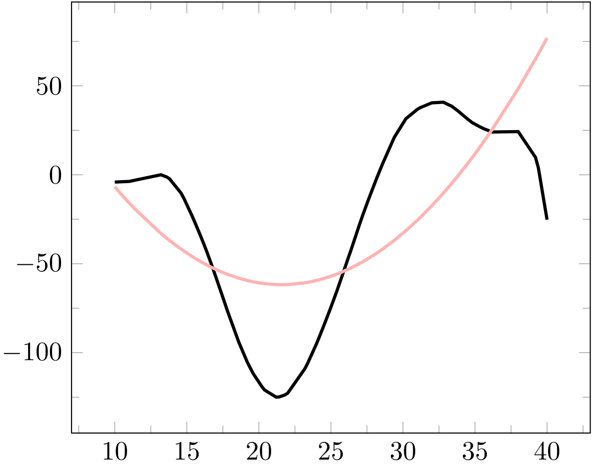
Any sufficiently smooth function can be well approximated by a polynomial function globally...



Approximation by polynomial of order 1

Regression and Nonparametric Regression.

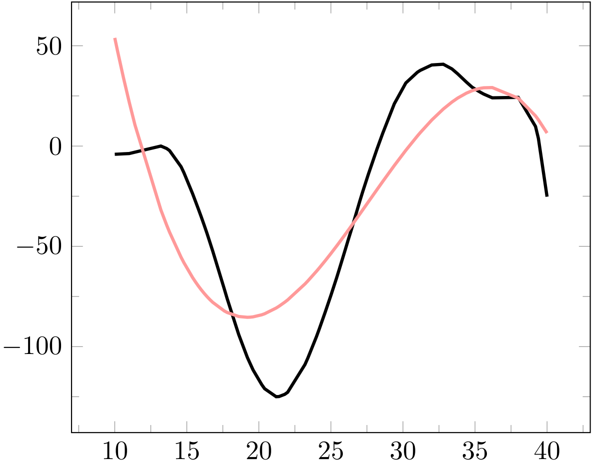
Any sufficiently smooth function can be well approximated by a polynomial function globally...



Approximation by polynomial of order 2

Regression and Nonparametric Regression.

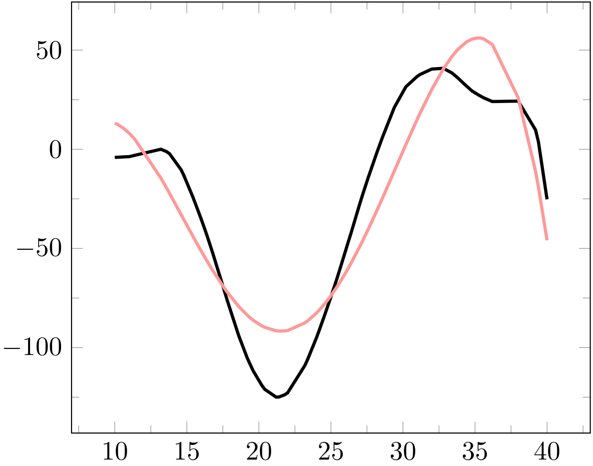
Any sufficiently smooth function can be well approximated by a polynomial function globally...



Approximation by polynomial of order 3

Regression and Nonparametric Regression.

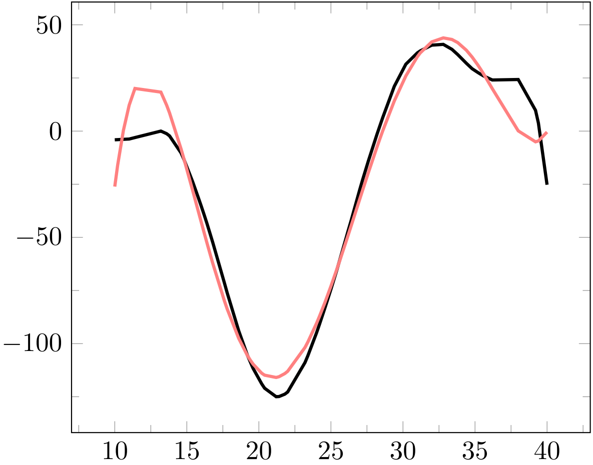
Any sufficiently smooth function can be well approximated by a polynomial function globally...



Approximation by polynomial of order 4

Regression and Nonparametric Regression.

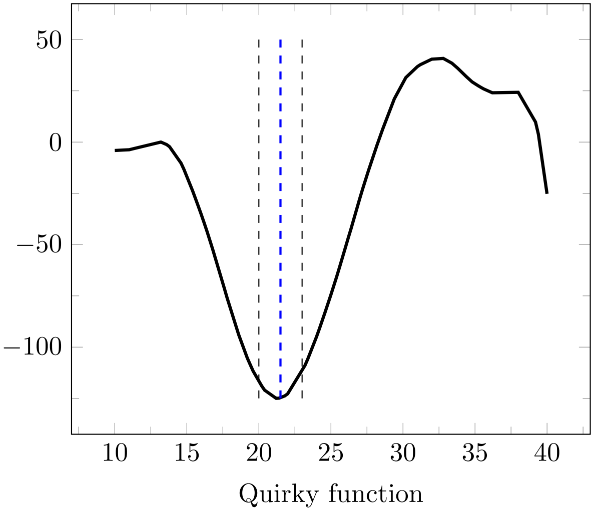
Any sufficiently smooth function can be well approximated by a polynomial function globally...



Approximation by polynomial of order 5

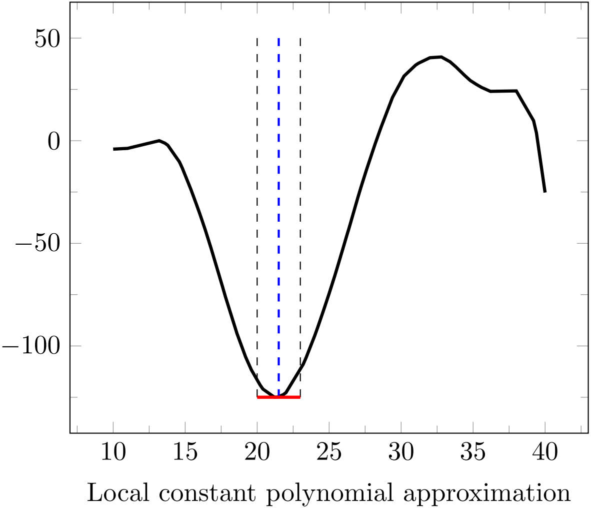
Regression and Nonparametric Regression

but the same principle applies locally, at a point, *c*.



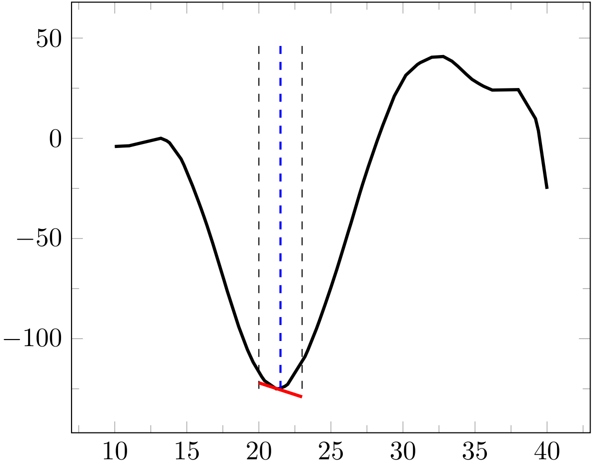
Regression and Nonparametric Regression

but the same principle applies locally, at a point, c.



Regression and Nonparametric Regression

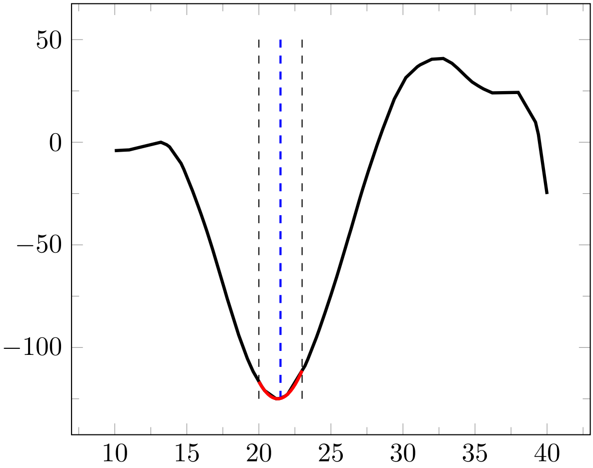
but the same principle applies locally, at a point, c.



Local linear polynomial approximation

Regression and Nonparametric Regression

but the same principle applies locally, at a point, c.



Local quadratic polynomial approximation

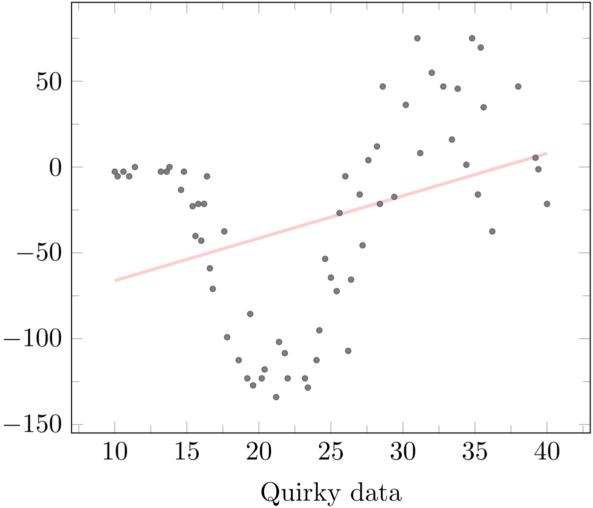
Regression and Nonparametric Regression

The least squares method provides a way of finding the best *global* polynomial approximation to a data set. More precisely, if $E\left[Y_i|X_i = x\right] = \mu(x)$ for some unknown function $\mu(x)$, least squares can be used to find the first order polynomial in X

$$E\Big[Y_i|X_i = x\Big] = \beta_0 + \beta_1 \cdot X_i \tag{11}$$

that is closest to the unknown $\mu(x)$ that produced the data. Here β_0, β_1 are unknown parameters that least squares finds by minimising the global error made by the linear model when approximating the observed Y_i,

Regression and Nonparametric Regression



Regression and Nonparametric Regression

but least squares can also find the best second order polynomial in X

$$E\big[Y_i|X_i = x\big] = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2$$

Regression and Nonparametric Regression

but least squares can also find the best second order polynomial in X

$$E\big[Y_i|X_i = x\big] = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2$$

- or third order $E\big[Y_i|X_i = x\big] = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + \beta_3 \cdot X_i^3$

Regression and Nonparametric Regression

but least squares can also find the best second order polynomial in X

$$E\left[Y_i|X_i = x\right] = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2$$

- or third order $E\left[Y_i|X_i = x\right] = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + \beta_3 \cdot X_i^3$
- or fourth order $E\left[Y_i|X_i = x\right] = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + \beta_3 \cdot X_i^3 + \beta_4 \cdot X_i^4$

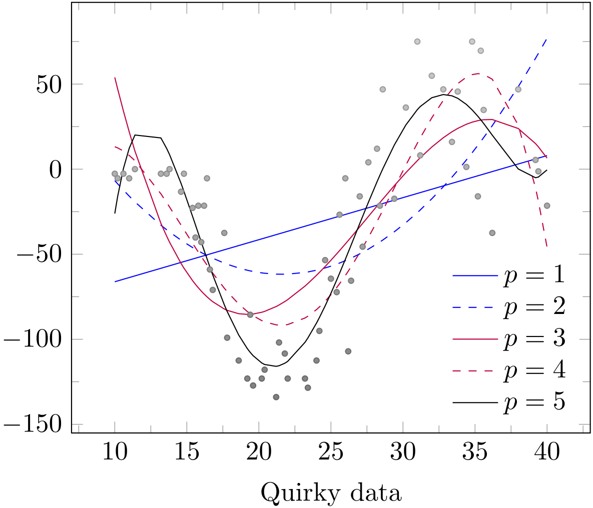
Regression and Nonparametric Regression

but least squares can also find the best second order polynomial in X

$$E\big[Y_i|X_i = x\big] = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2$$

- or third order $E\big[Y_i|X_i = x\big] = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + \beta_3 \cdot X_i^3$
- or fourth order $E\big[Y_i|X_i = x\big] = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + \beta_3 \cdot X_i^3 + \beta_4 \cdot X_i^4$
- or fifth order $E\big[Y_i|X_i = x\big] = \beta_0 + \beta_1 \cdot X_i + \beta_2 \cdot X_i^2 + \beta_3 \cdot X_i^3 + \beta_4 \cdot X_i^4 + \beta_5 \cdot X_i^5$

Regression and Nonparametric Regression



Regression and Nonparametric Regression

It turns out that to apply the same principle *locally*. Rather than trying to approximate $E[Y_i|X_i = x] = \mu(x)$ with a single model, we can approximate $\mu(x)$ at one specific location, say $x = c$ only. To do this, you only need to apply least squares with a couple of modifications:

- First, tell least squares on which point, $x = c$, you want to focus estimation. In practice this translates in using $(X_i - c)$, $(X_i - c)^2, \ldots (X_i - c)^p$ instead of $X_i, X_i^2, \ldots, X_i^p$
- Second, tell least squares which observations are important for estimation. This can be done using a **kernel** (weight) function, $K(c; h)$ that penalises observations away from c for some **bandwidth parameter** h that needs to be selected.

Regression and Nonparametric Regression

Then, for a give p, the best estimator of $E[Y_i|X_i = x] = \mu(x)$ at $x = c$ is the least squares estimator of β_0 in the following least squares problem:

$$\min_{\beta_0,\ldots,\beta_p} \sum_{i=1}^{N} e_i^2 \cdot K(c; h)$$

where

$$e_i = Y_i - \beta_0 - \beta_1 \cdot (X_i - c) - \beta_2 \cdot (X_i - c)^2 - \ldots - \beta_p \cdot (X_i - c)^p$$

The least squares estimator of β_0 is a valid estimator of $\mu(c) = E\left[Y_i|X_i = c\right]$; the **nonparametric regression estimator of** $E\left[Y_i|X_i = c\right]$

Nonparametric Regression: choice of kernel

The choice of kernel is discussed in several sources and is not a critical decision, although specific choices of kernel have more desirable theoretical properties[6]

A common choice in practice is the uniform kernel. When estimating $\hat{\mu}^-$ this kernel is defined as,

$$K(c;h) = \begin{cases} 1 & \text{if } c - h \leq X_i < c \\ 0 & \text{otherwise} \end{cases} \tag{12}$$

Essentially, this choice of kernel results in a least squares regression using only observations with X falling within $(c-h, c+h)$

[6] See Hardle, 1990; Li and Racine, 2006

Nonparametric Regression: choice of p

The choice of the order of the polynomial is more consequential.

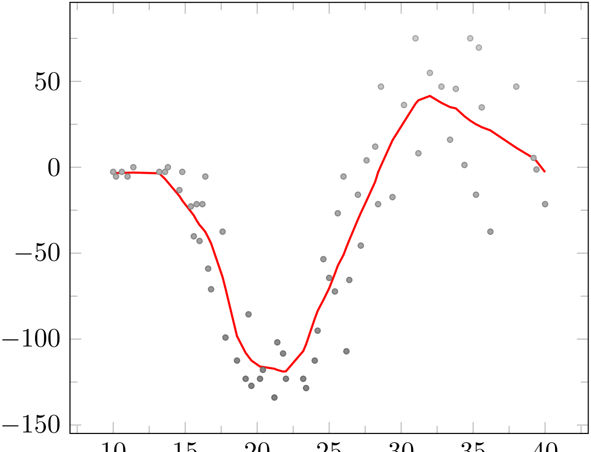
In practice it is recommended to select $p = 1$ (local linear regression).

Estimates based on $p = 0$ have poor performance in the boundaries, (were some of the estimators we will discuss in this course need to be computed).

For a given choice of bandwidth, increasing p tends to increase the accuracy of the approximation, but also increases the variability of the estimator.

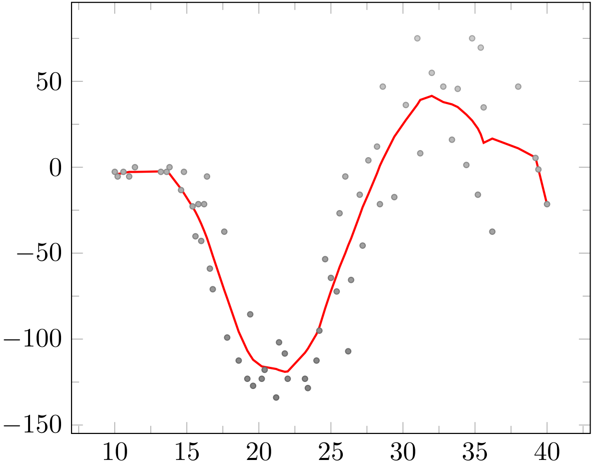
Large choices of p may incur in over-fitting, which increases the variability of the estimates.

Nonparametric Regression: choice of p



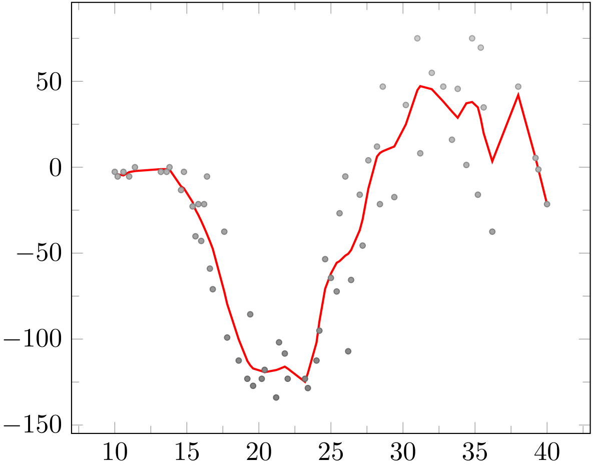
Polynomial order 0 and bandwidth 1

Nonparametric Regression: choice of p



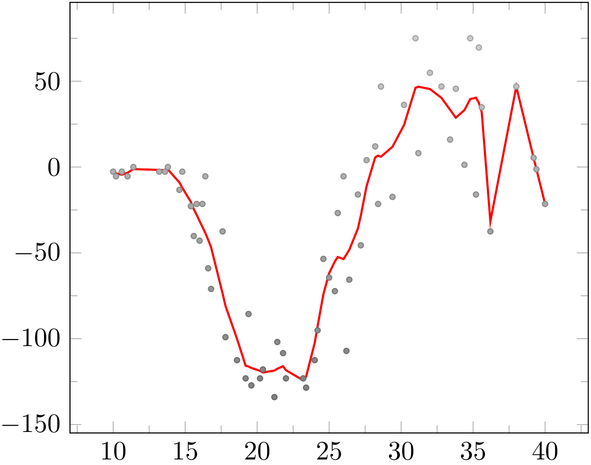
Polynomial order 1 and bandwidth 1

Nonparametric Regression: choice of p



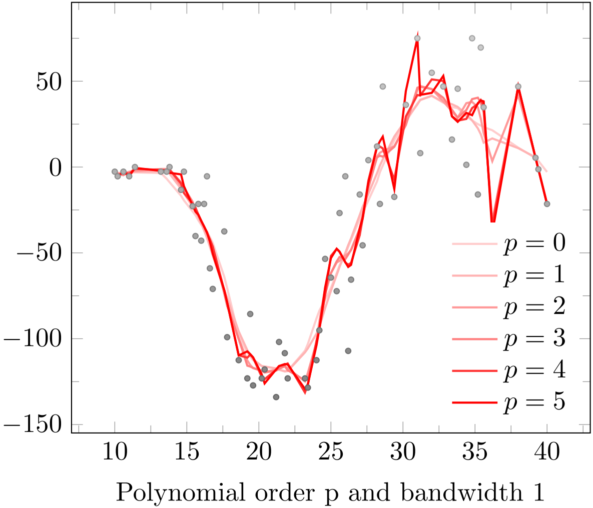
Polynomial order 2 and bandwidth 1

Nonparametric Regression: choice of p



Polynomial order 3 and bandwidth 1

Nonparametric Regression: choice of p



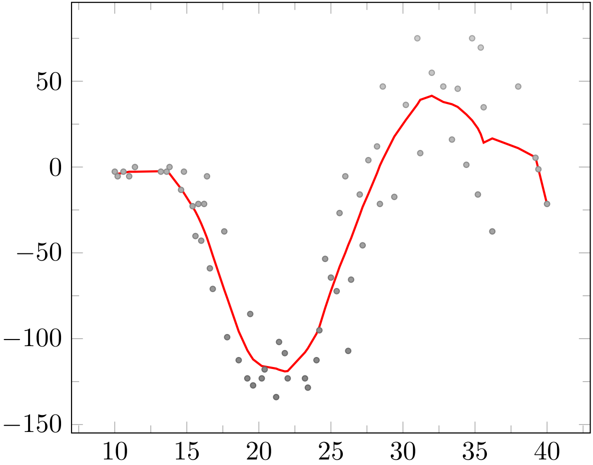
Nonparametric Regression: Choice of bandwidth.

The choice of bandwidth involves at **bias-variance trade-off**

Small bandwidths reduce the bias or approximation error, but they tend to increase the variability (variance) of the estimated coefficients -because fewer observations will be avaiable for estimation

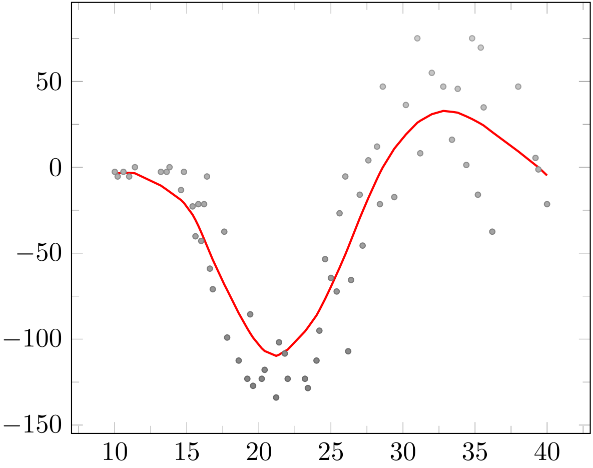
Big bandwidths, produce estimates that more stable (smaller variance) but they will result in more smoothing bias if the unknown regression function differs a lot from the polynomial model use for approximation.

Nonparametric Regression: Choice of bandwidth.



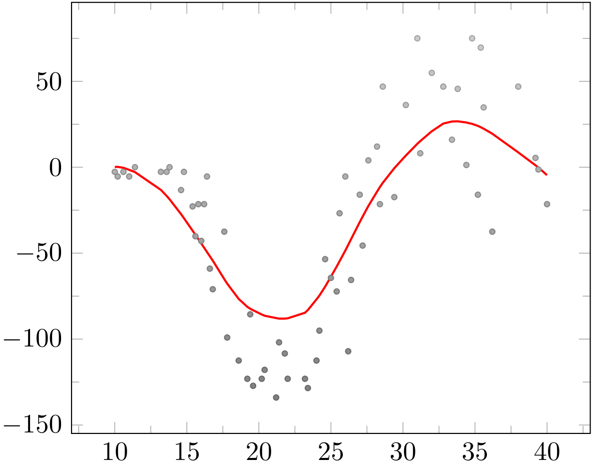
Polynomial order 1 and bandwidth 1

Nonparametric Regression: Choice of bandwidth.



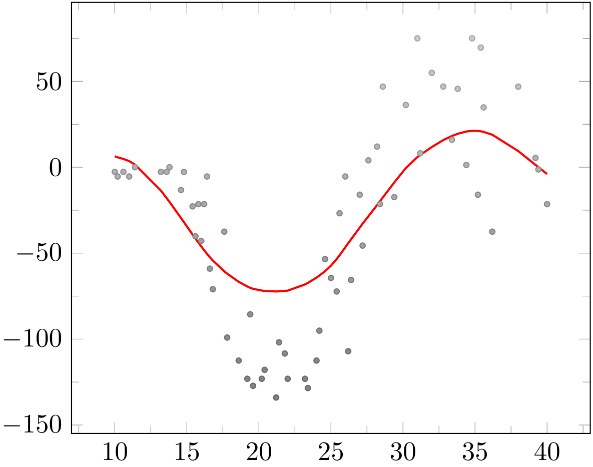
Polynomial order 1 and bandwidth 2

Nonparametric Regression: Choice of bandwidth.



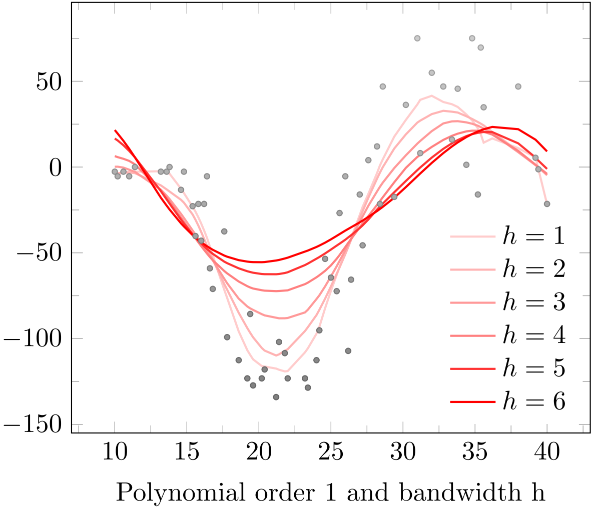
Polynomial order 1 and bandwidth 3

Nonparametric Regression: Choice of bandwidth.



Polynomial order 1 and bandwidth 4

Nonparametric Regression: Choice of bandwidth.



Nonparametric Regression: Choice of bandwidth.

Procedures to select a bandwidth try to balance bias and variance by minimising the Mean Square Error of the SRD estimator,

$$\text{MSE} = \text{Bias}(\hat{\tau}_{SRD})^2 + \text{Variance}(\hat{\tau}_{SRD}) = h^{2(p+1)}B + \frac{1}{nh}V$$

where the constants B and V depend on the curvature and variability of the unknown regression at $x = c$ and the choice of kernels and order of polynomial.

The MSE shows the bias-variance trade-off in a more formal way.

- Increasing h increases the contribution of the bias to the MSE, but decreases the contribution of the variance
- As the sample size increases, the contribution of the variance decreases; but, since h is small

Nonparametric Regression: Choice of bandwidth.

If the constants B and V were known, one can show that the optimal choice of bandwidth is

$$h^* = \left(\frac{V}{2(p+1)B^2} \right)^{1/(2p+3)} \frac{1}{n^{1/(2p+3)}} = \frac{\rho}{n^{1/(2p+3)}} \tag{13}$$

So as N increase h^* decreases, but at a slower rate than N. This has implications for estimation and testing.

Nonparametric Regression: Choice of bandwidth.

When the optimal bandwidth is plugged into the MSE,

$$\text{MSE} = \frac{\rho^{2p+2)}}{n^{2p+2/(2p+3)}} B + \frac{1}{\rho n^{1-1/(2p+3)}} V = \frac{\mathbb{C}}{n^{(2p+2)/(2p+3)}}$$

Since $2p+2/2p+3<1$, this means that increasing the sample size decreases the MSE less than proportionally.
Overall, this means that as the sample size increases, we can reduce the error in approximation by reducing the bandwidth, without paying a penalty in added variability. However, sample size reduction must occur at a slower than the pace at which the sample size increases.

Nonparametric Regression: Choice of bandwidth.

Various bandwidth selection methods exist, but the majority try to find an h such that balance the MSE. Specific methods are discussed in Hardle (1990), Li and Racine (2006), among other sources. The specific of methods are beyond the scope of this course, but we will use them following a black-box approach.

Summary

References I

Balke, A. and **Pearl, J.** (1997). Bounds on treatment effects from studies with imperfect compliance. *Journal of the American Statistical Association*, 92(439): 1171–1176

Fisher, R. (1935). *The Design of Experiments*. Oliver Boyd

Hardle, W. (1990). *Applied Nonparametric Regression*. Cambridge University Press

Holland, P.W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396): 945–960

Li, Q. and **Racine, J.** (2006). *Nonparametric Econometrics: Theory and Practice*. Princeton University Press

Neyman, J., **Iwaszkiewicz, K.**, and **Kolodziejczyk, S.** (1935). Statistical problems in agricultural experimentation. *Supplement to the Journal of the Royal Statistical Society*, 2(2): 107–180. ISSN 14666162

References II

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97(4): 558–625

Neyman, J., **Dabrowska, D.M.**, and **Speed, T.P.** (1990). On the application of probability theory to agricultural experiments. essay on principles. section 9. *Statist. Sci.*, 5(4): 465–472

Pearl, J. (2009). *Causality*. Cambridge University Press

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66: 688–701

Rubin, D.B. (1980). Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371): 591–593. ISSN 01621459

References III

Rubin, D.B. (1990a). Formal mode of statistical inference for causal effects. *Journal of Statistical Planning and Inference*, 25(3): 279 – 292

Rubin, D.B. (1990b). [on the application of probability theory to agricultural experiments. essay on principles. section 9.] comment: Neyman (1923) and causal inference in experiments and observational studies. *Statist. Sci.*, 5(4): 472–480