Lecture 2: Causal Inference Using Observational Data

Sheetal Sekhri
University of Virginia

BREAD IGC Summer School, India 2012

July 21, 2012

Using Observational Data

- Many policies and programs are evaluated after their implementation

Using Observational Data

- Many policies and programs are evaluated after their implementation
- Policy design can sometimes provide plausibly exogenous variation

Using Observational Data

- Many policies and programs are evaluated after their implementation
- Policy design can sometimes provide plausibly exogenous variation
- Observational data that can be combined with institutional details for evaluation

Using Observational Data

- Many policies and programs are evaluated after their implementation
- Policy design can sometimes provide plausibly exogenous variation
- Observational data that can be combined with institutional details for evaluation
- Applications may also lend themselves to natural experiments

Using Observational Data

- Many policies and programs are evaluated after their implementation
- Policy design can sometimes provide plausibly exogenous variation
- Observational data that can be combined with institutional details for evaluation
- Applications may also lend themselves to natural experiments
- Observational data can also be used in such contexts

Benefits of Using Observational Data

- Time horizon relative to field experiments

Benefits of Using Observational Data

- Time horizon relative to field experiments
- Not as expensive

Benefits of Using Observational Data

- Time horizon relative to field experiments
- Not as expensive
- Externally valid inference (depending on the data and design)

Benefits of Using Observational Data

- Time horizon relative to field experiments
- Not as expensive
- Externally valid inference (depending on the data and design)
- Less fraught with behavioral concerns

Limitations of Using Observational Data

- Selection concerns

Limitations of Using Observational Data

- Selection concerns
- Data quality

Limitations of Using Observational Data

- Selection concerns
- Data quality
- Data limitations - not all desired data for an application may exist

Limitations of Using Observational Data

- Selection concerns
- Data quality
- Data limitations - not all desired data for an application may exist
- Data restrictions

Working with Observational Data- Methods

- Difference-in-Difference (DID)

Working with Observational Data- Methods

- Difference-in-Difference (DID)
- Regression Discontinuity Design (RDD)

Difference-in-Difference

- Most popular method used in empirical analysis

Difference-in-Difference

- Most popular method used in empirical analysis
- Emulate an experiment with treatment and comparison groups

Difference-in-Difference

- Most popular method used in empirical analysis
- Emulate an experiment with treatment and comparison groups
- Uses panel data and is a two way fixed effects model

DID- Basic Idea

- With panel data on the treated group, can compare pre and post intervention or policy change

DID- Basic Idea

- With panel data on the treated group, can compare pre and post intervention or policy change
- But any discerned effect can arise due to secular changes

DID- Basic Idea

- With panel data on the treated group, can compare pre and post intervention or policy change
- But any discerned effect can arise due to secular changes
- Panel data on comparison group can provide the counterfactual

DID- Basic Idea

- With panel data on the treated group, can compare pre and post intervention or policy change
- But any discerned effect can arise due to secular changes
- Panel data on comparison group can provide the counterfactual
- What would happen to treated group over time in absence of treatment

DID- Implementation

- Isolate the design using tabular or graphic representation

DID- Implementation

- Isolate the design using tabular or graphic representation
- Formalize using regression analysis

Tabular Representation

DID

	Before	After	Difference
Treatment			
Control			
Difference			

Tabular Representation

DID

	Before	After	Difference
Treatment	Y_{T1}	Y_{T2}	
Control			
Difference			

Tabular Representation

DID

	Before	After	Difference
Treatment	Y_{T1}	Y_{T2}	
Control	Y_{C1}	Y_{C2}	
Difference			

Tabular Representation

DID

	Before	After	Difference
Treatment	Y_{T1}	Y_{T2}	$\Delta Y_T = Y_{T2} - Y_{T1}$
Control	Y_{C1}	Y_{C2}	
Difference			

Tabular Representation

DID

	Before	After	Difference
Treatment	Y_{T1}	Y_{T2}	$\Delta Y_T = Y_{T2} - Y_{T1}$
Control	Y_{C1}	Y_{C2}	$\Delta Y_C = Y_{C2} - Y_{C1}$
Difference			

Tabular Representation

DID

	Before	After	Difference
Treatment	Y_{T1}	Y_{T2}	$\Delta Y_T = Y_{T2} - Y_{T1}$
Control	Y_{C1}	Y_{C2}	$\Delta Y_C = Y_{C2} - Y_{C1}$
Difference			$\Delta Y_T - \Delta Y_C$

DID- Identifying Assumption

- Control group shows the time path of the treatment group without the intervention

DID- Identifying Assumption

- Control group shows the time path of the treatment group without the intervention
- Time trends in absence of treatment should be the same

DID- Identifying Assumption

- Control group shows the time path of the treatment group without the intervention
- Time trends in absence of treatment should be the same
- Levels can be different

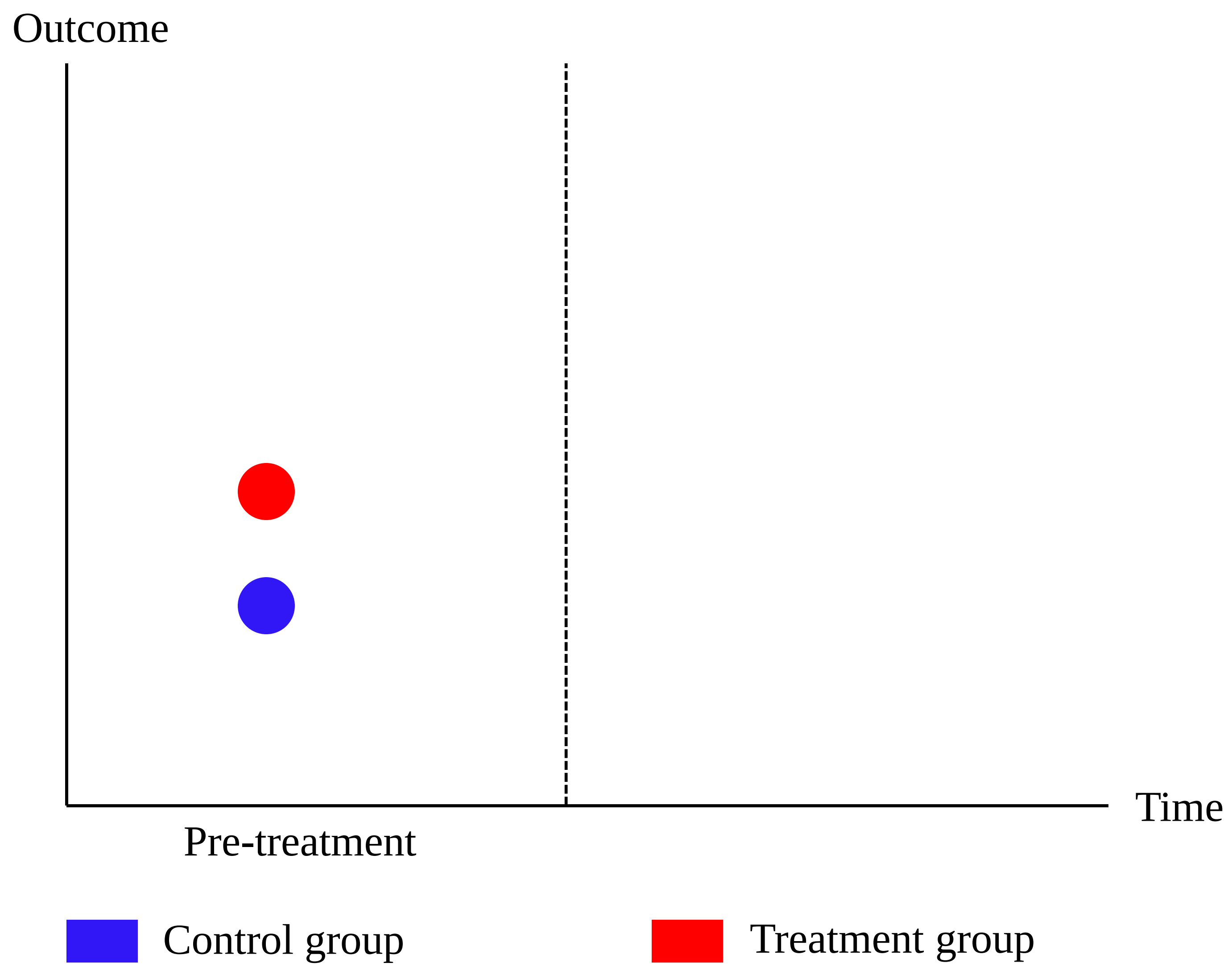
DID- Identifying Assumption

- Control group shows the time path of the treatment group without the intervention
- Time trends in absence of treatment should be the same
- Levels can be different
- If different time trends, effect over or under stated

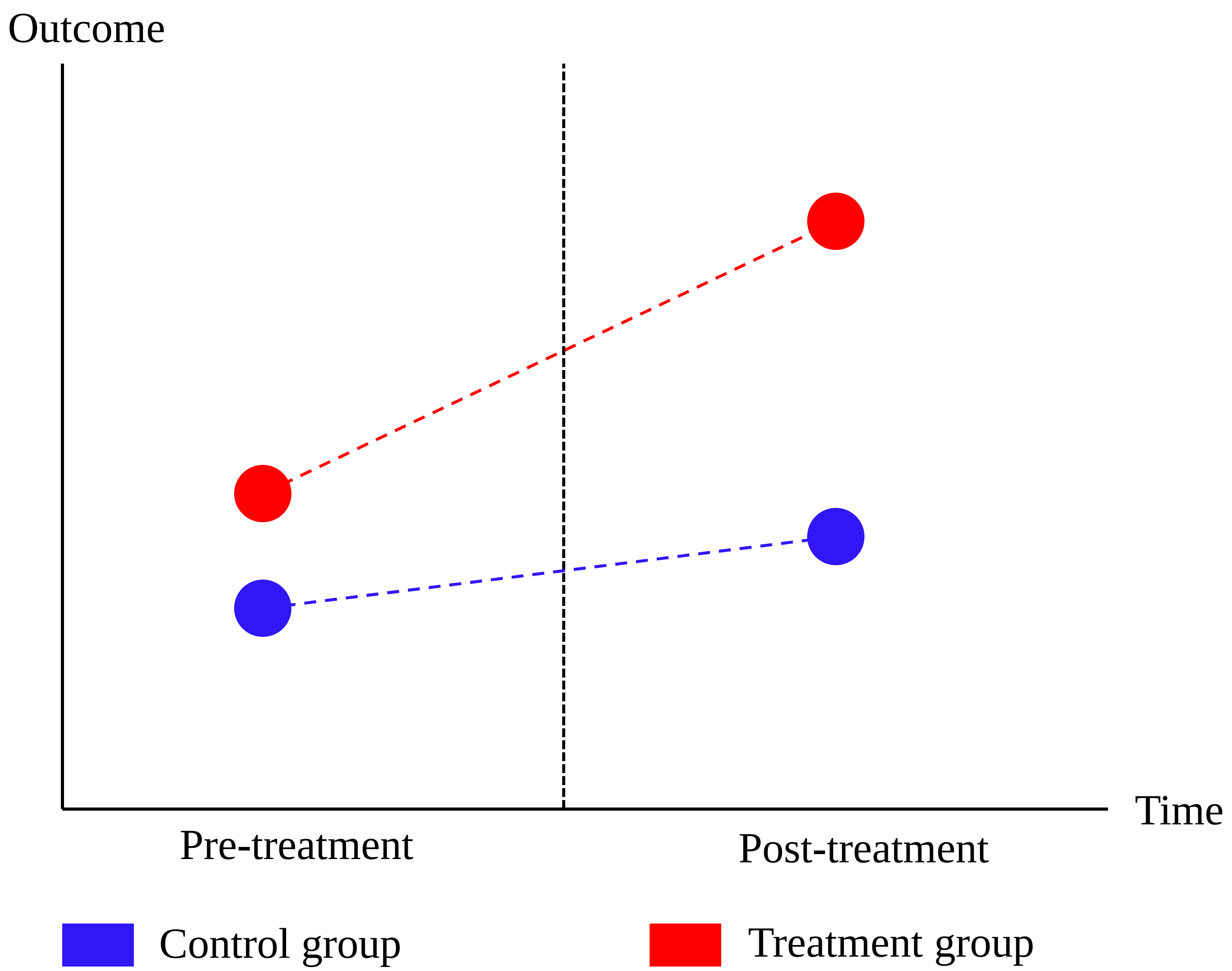
DID- Identifying Assumption

- Control group shows the time path of the treatment group without the intervention
- Time trends in absence of treatment should be the same
- Levels can be different
- If different time trends, effect over or under stated
- Identifying assumption- No differential pre-trends

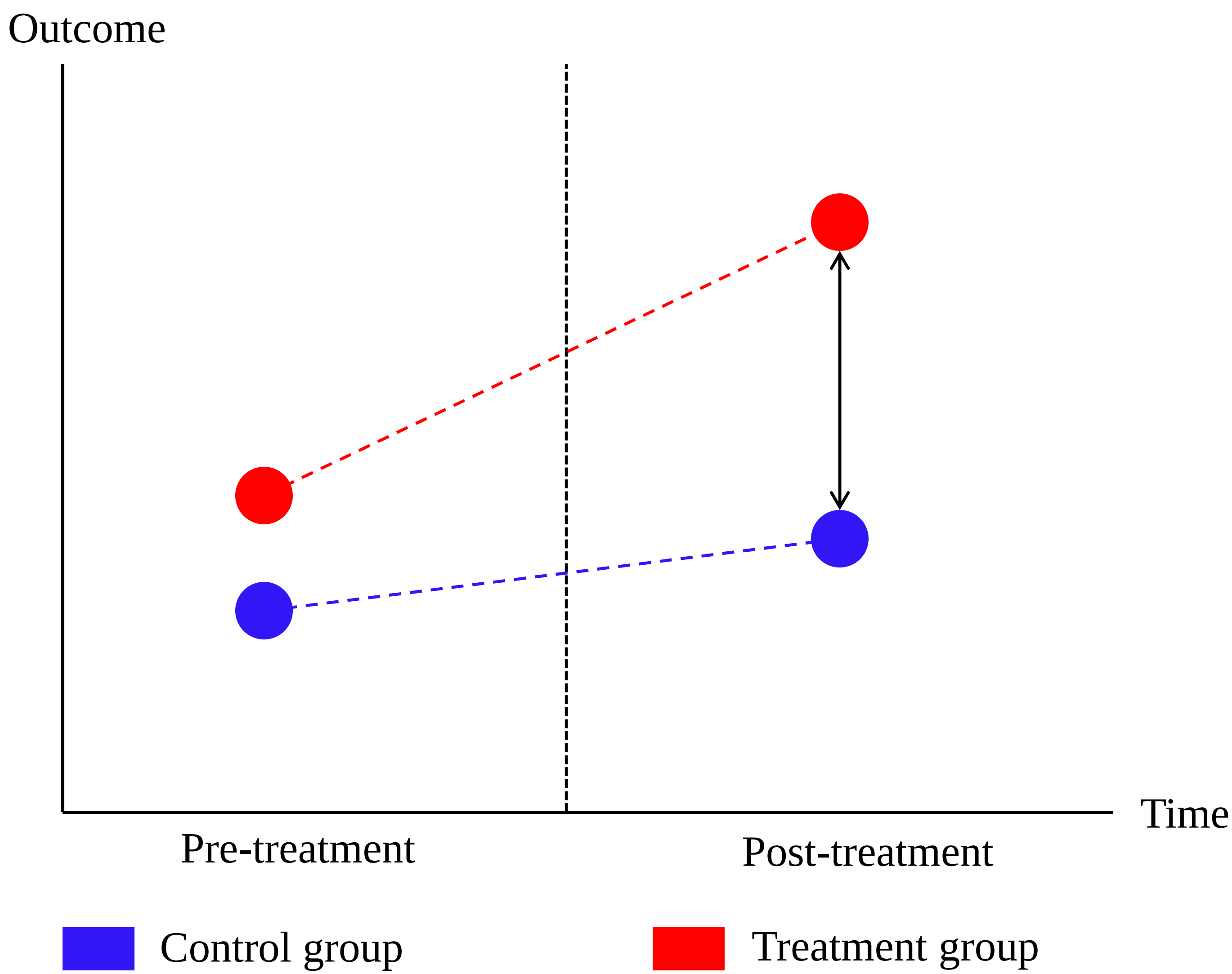
Difference in Difference



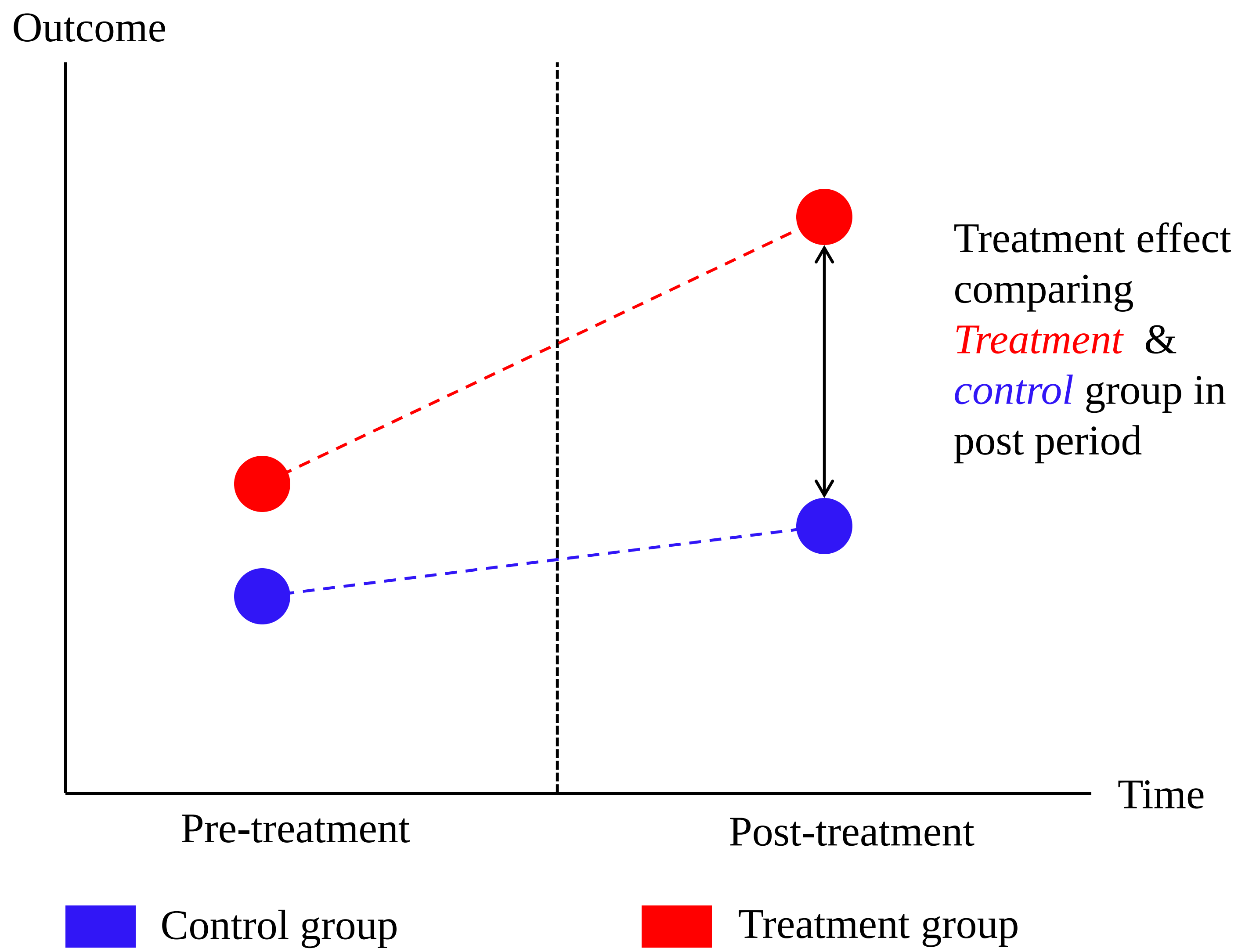
Difference in Difference



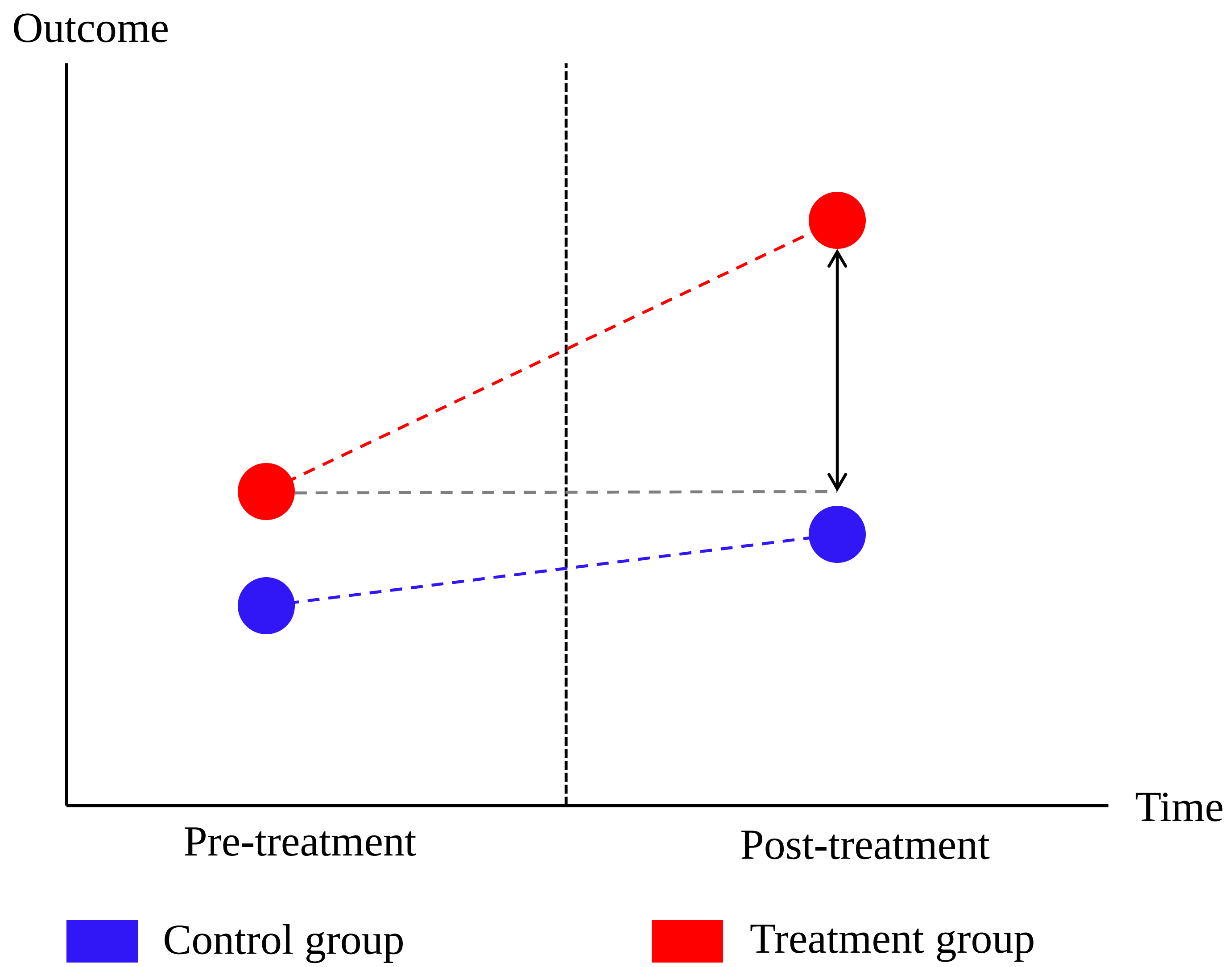
Difference in Difference



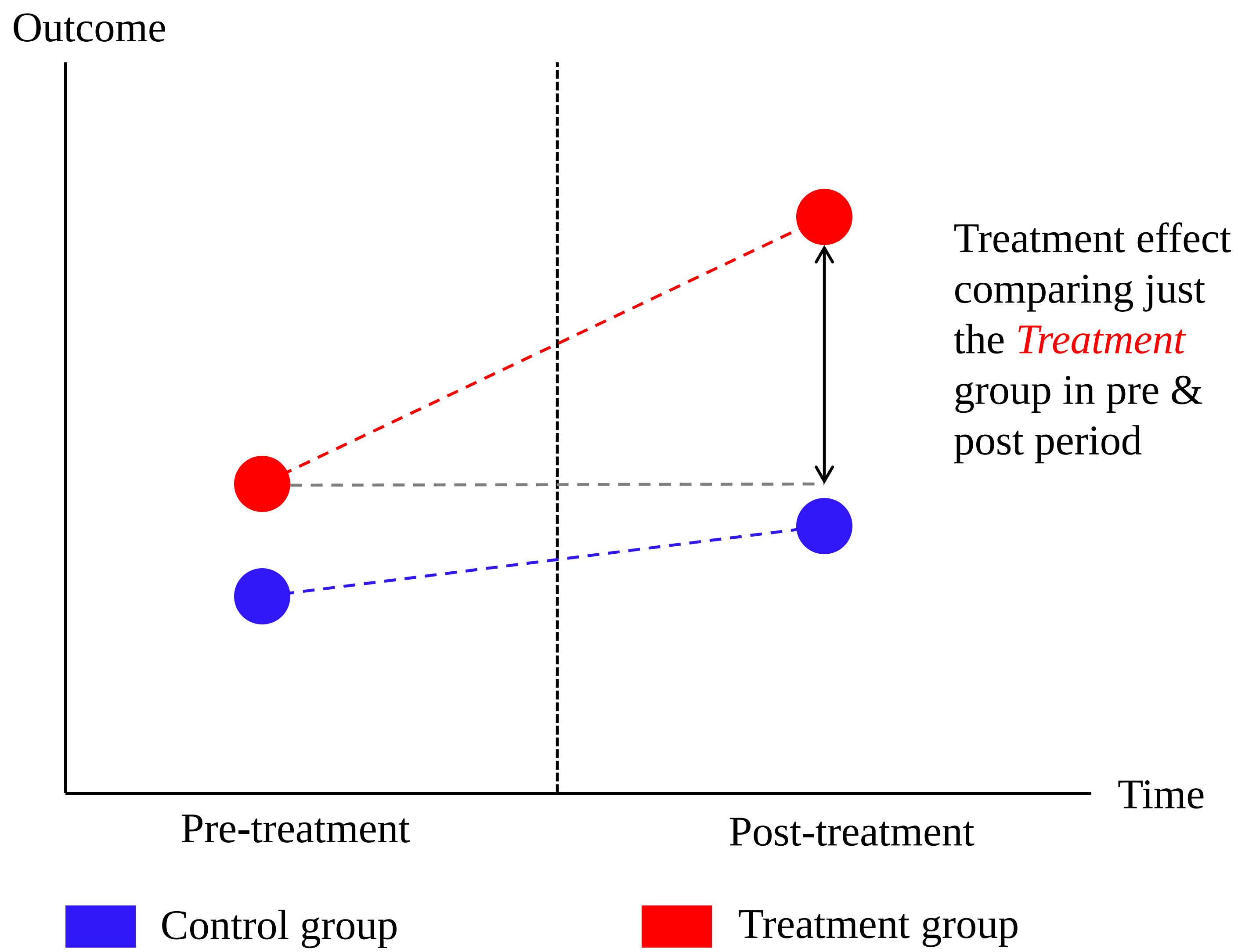
Difference in Difference



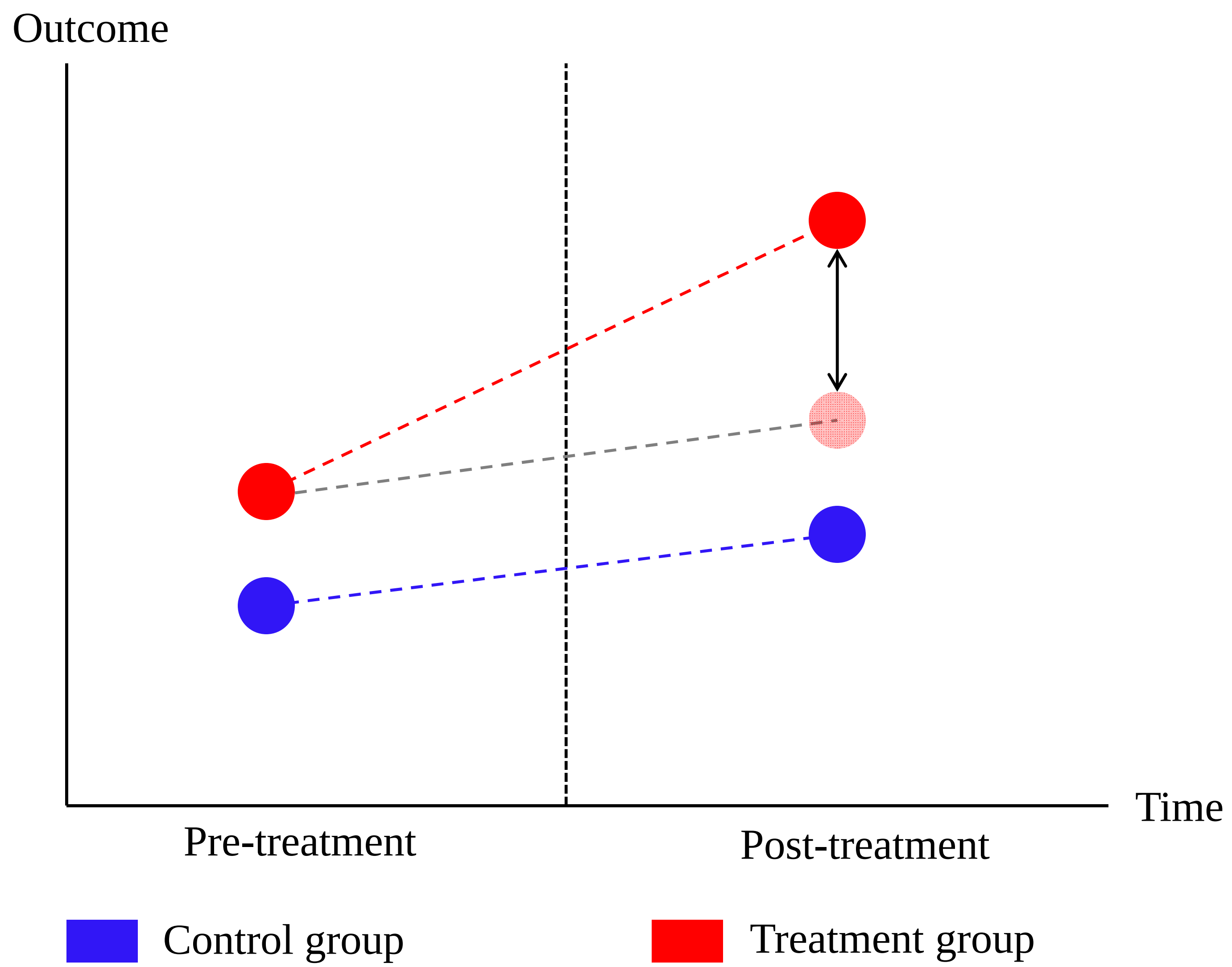
Difference in Difference



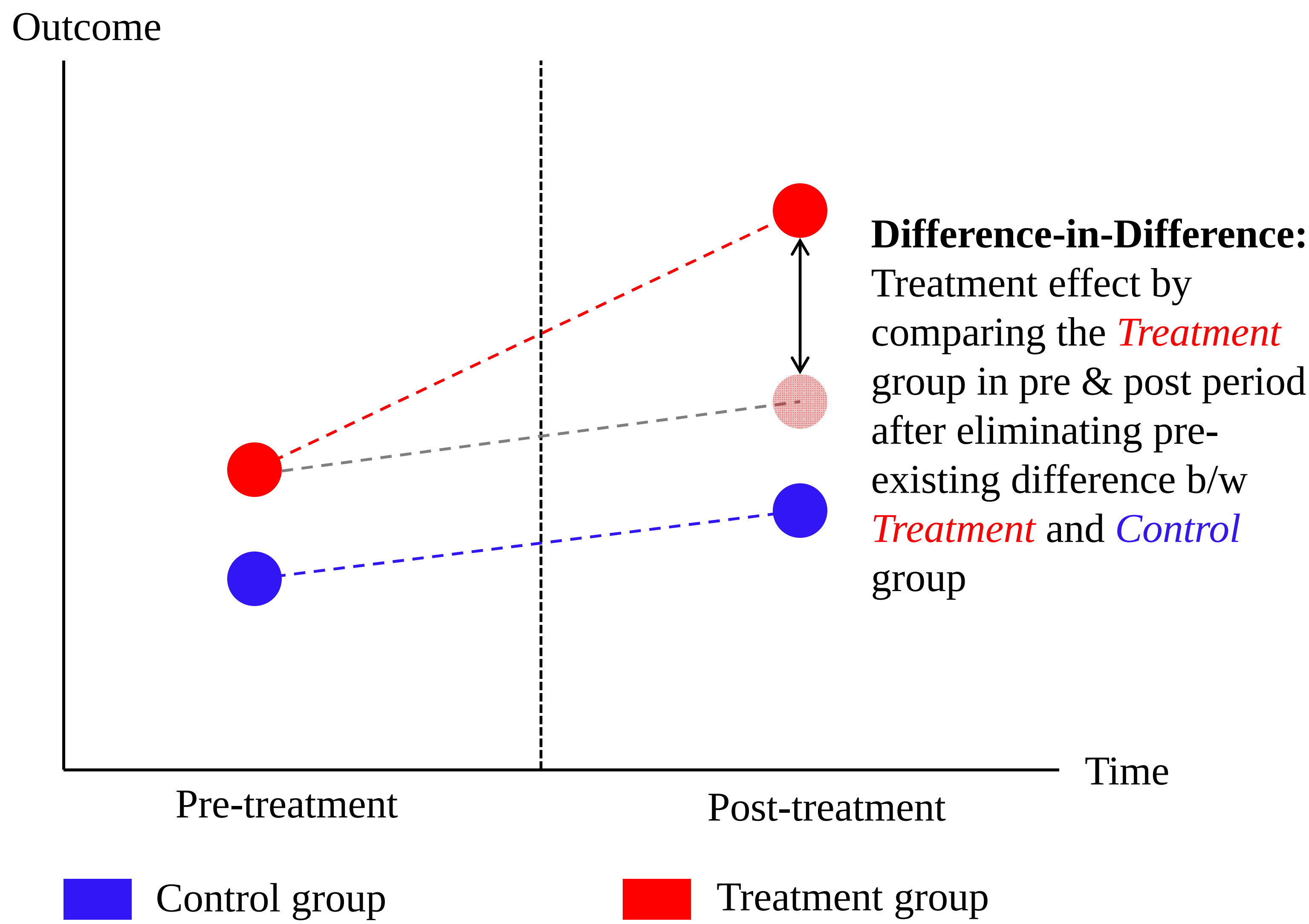
Difference in Difference



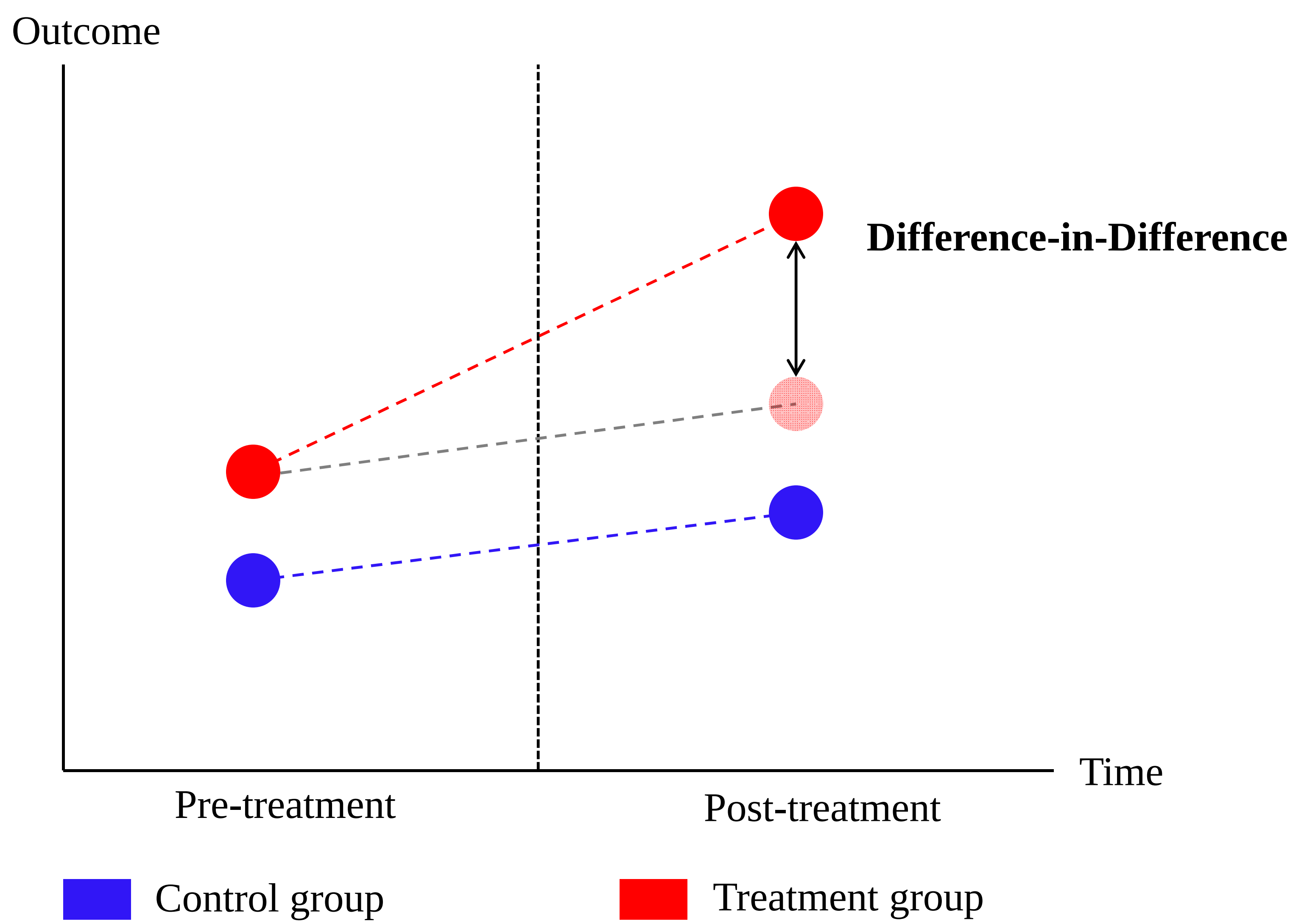
Difference in Difference



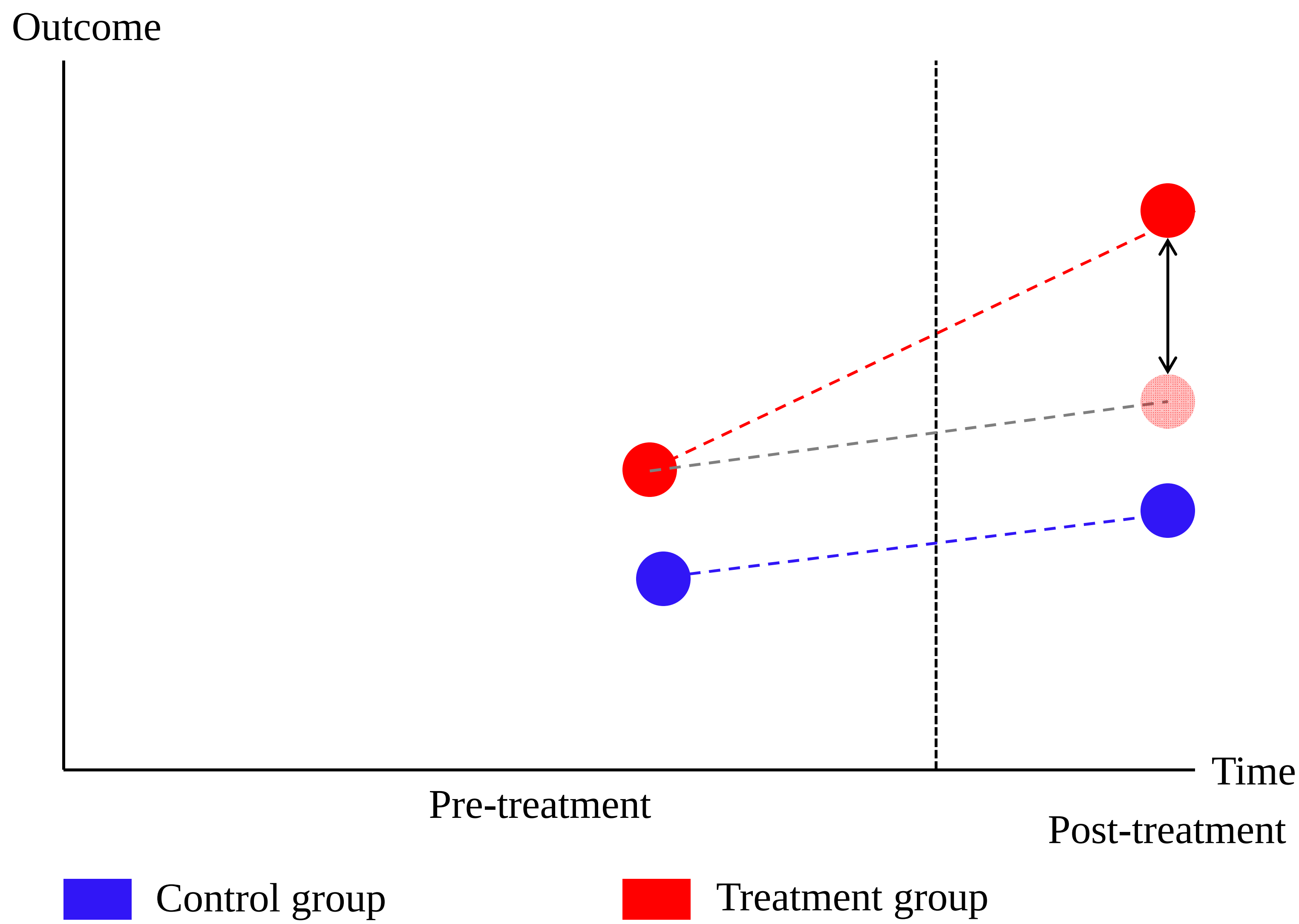
Difference in Difference



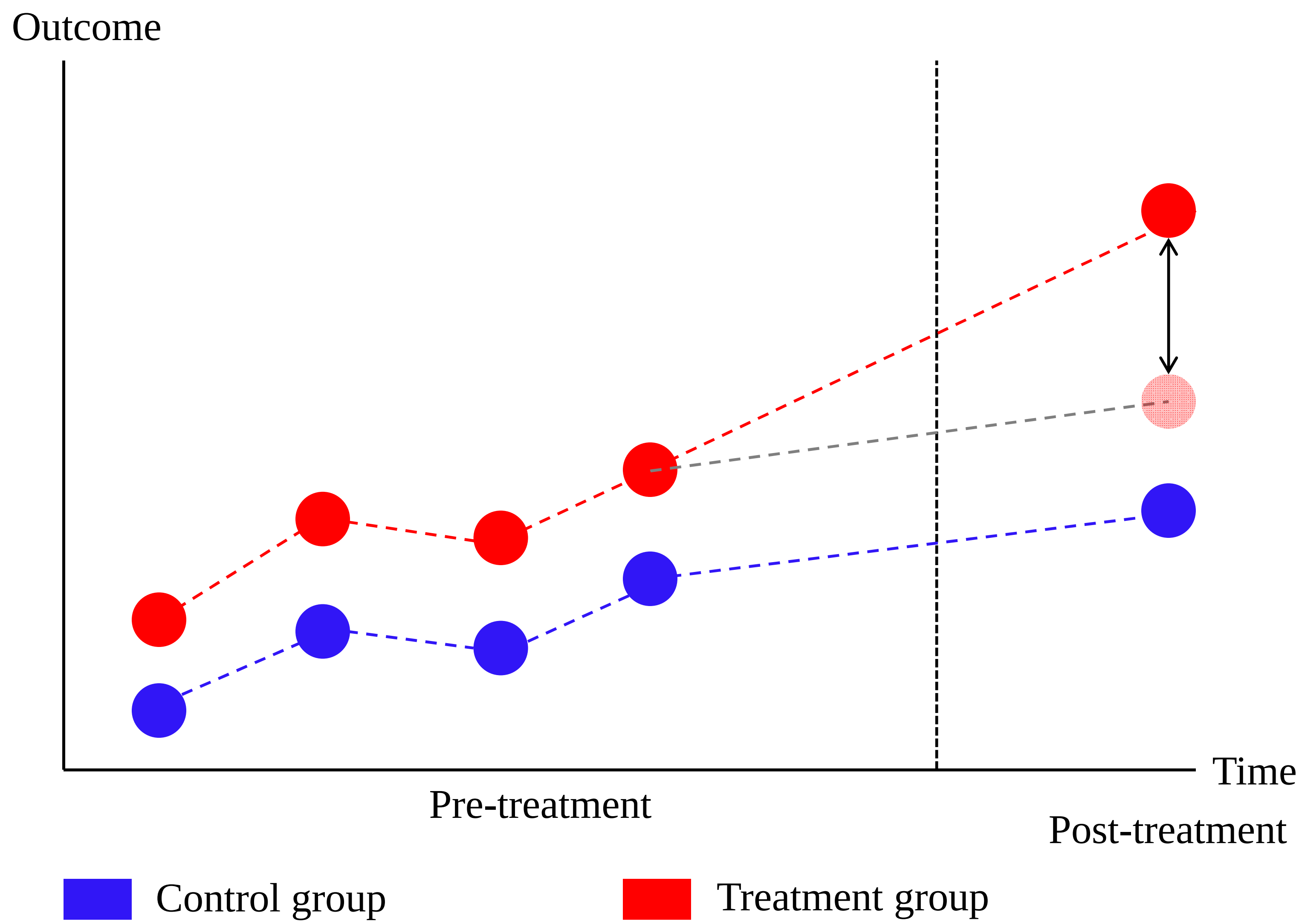
Difference in Difference: Parallel Trend Assumption



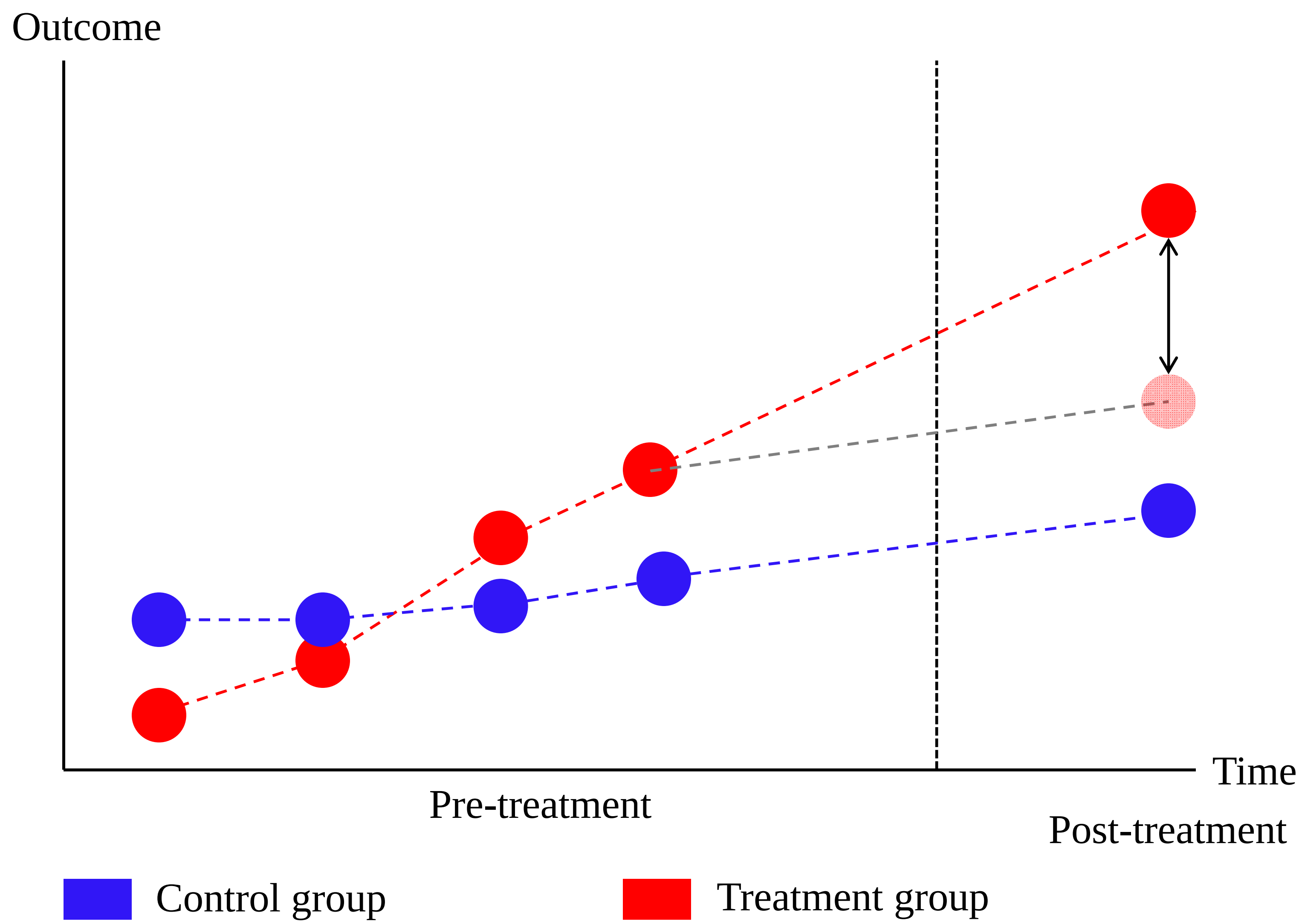
Difference in Difference: Parallel Trend Assumption



Difference in Difference: Parallel Trend Assumption



Difference in Difference: Parallel Trend Assumption Violated



DID- Regression Analysis

- Suppose our policy change effects villages such that a set of villages are treated

DID- Regression Analysis

- Suppose our policy change effects villages such that a set of villages are treated
- We have data over time for all villages

DID- Regression Analysis

- Suppose our policy change effects villages such that a set of villages are treated
- We have data over time for all villages
- The panel has only 2 time periods

DID- Regression Analysis

- Suppose our policy change effects villages such that a set of villages are treated
- We have data over time for all villages
- The panel has only 2 time periods
- Post is an indicator that switches to 1 after the intervention

DID- Regression Analysis

- Suppose our policy change effects villages such that a set of villages are treated
- We have data over time for all villages
- The panel has only 2 time periods
- Post is an indicator that switches to 1 after the intervention
- T is an indicator that takes value 1 for the villages to be treated

DID- Regression Analysis

- Outcome variable Y varies by villages and time

DID- Regression Analysis

- Outcome variable Y varies by villages and time
- $Y_{it} = \alpha_0 + \alpha_1 \; Post + \alpha_2 \; T + \alpha_3 \; Post * T + \varepsilon_{it}$

DID- Regression Analysis

- Outcome variable Y varies by villages and time
- $Y_{it} = \alpha_0 + \alpha_1 \ Post + \alpha_2 \ T + \alpha_3 \ Post * T + \varepsilon_{it}$
- The panel has only 2 time periods

DID- Regression Analysis

- Outcome variable Y varies by villages and time
- $Y_{it} = \alpha_0 + \alpha_1 \ \mathit{Post} + \alpha_2 \ T + \alpha_3 \ \mathit{Post} * T + \varepsilon_{it}$
- The panel has only 2 time periods
- Post is an indicator that switches to 1 after the intervention

DID- Regression Analysis

- Outcome variable Y varies by villages and time
- $Y_{it} = \alpha_0 + \alpha_1 \; Post + \alpha_2 \; T + \alpha_3 \; Post * T + \varepsilon_{it}$
- The panel has only 2 time periods
- Post is an indicator that switches to 1 after the intervention
- T is an indicator that takes value 1 for the villages to be treated

Tabular Representation

DID

	Before	After	Difference
Treatment	$\alpha_0 + \alpha_2$		
Control			
Difference			

Tabular Representation

DID

	Before	After	Difference
Treatment	$\alpha_0 + \alpha_2$	$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3$	
Control			
Difference			

Tabular Representation

DID

	Before	**After**	**Difference**
Treatment	$\alpha_0 + \alpha_2$	$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3$	$\alpha_1 + \alpha_3$
Control			
Difference			

Tabular Representation

DID

	Before	After	Difference
Treatment	$\alpha_0 + \alpha_2$	$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3$	$\alpha_1 + \alpha_3$
Control	α_0		
Difference			

Tabular Representation

DID

	Before	After	Difference
Treatment	$\alpha_0 + \alpha_2$	$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3$	$\alpha_1 + \alpha_3$
Control	α_0	$\alpha_0 + \alpha_1$	
Difference			

Tabular Representation

DID

	Before	After	Difference
Treatment	$\alpha_0 + \alpha_2$	$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3$	$\alpha_1 + \alpha_3$
Control	α_0	$\alpha_0 + \alpha_1$	α_1
Difference			

Tabular Representation

DID

	Before	After	Difference
Treatment	$\alpha_0 + \alpha_2$	$\alpha_0 + \alpha_1 + \alpha_2 + \alpha_3$	$\alpha_1 + \alpha_3$
Control	α_0	$\alpha_0 + \alpha_1$	α_1
Difference			α_3

DID- Robustness and Extension

- With many years data before the intervention, possible to check for pre-trends to make estimation more credible

DID- Robustness and Extension

- With many years data before the intervention, possible to check for pre-trends to make estimation more credible

- Common support required - check for significant overlap in distributions of T and C

DID- Robustness and Extension

- With many years data before the intervention, possible to check for pre-trends to make estimation more credible
- Common support required - check for significant overlap in distributions of T and C
- Placebo test- No effect should be discerned if treatment is randomly considered to occur in any year prior to actual date

DID- Robustness and Extension

- With many years data before the intervention, possible to check for pre-trends to make estimation more credible
- Common support required - check for significant overlap in distributions of T and C
- Placebo test- No effect should be discerned if treatment is randomly considered to occur in any year prior to actual date
- Balance across treatment and control- selection model to show determinants of treatment not time varying

DID- Extension

- $Y_{it} = \beta_0 + T_t + V_i + \beta_2 \; T_t * V_i + \varepsilon_{it}$

DID- Extension

- $Y_{it} = \beta_0 + T_t + V_i + \beta_2 \ T_t * V_i + \varepsilon_{it}$
- T_t full set of year fixed effects

DID- Extension

- $Y_{it} = \beta_0 + T_t + V_i + \beta_2 \; T_t * V_i + \varepsilon_{it}$
- T_t full set of year fixed effects
- V_i full set of village fixed effects

DID- Extension

- $Y_{it} = \beta_0 + T_t + V_i + \beta_2 \; T_t * V_i + \varepsilon_{it}$
- T_t full set of year fixed effects
- V_i full set of village fixed effects
- Allows for covariance between T_t and $T_t * V_i$ and V_i and $T_t * V_i$

DID- Extension

- $Y_{it} = \beta_0 + T_t + V_i + \beta_2 \; T_t * V_i + \varepsilon_{it}$
- T_t full set of year fixed effects
- V_i full set of village fixed effects
- Allows for covariance between T_t and $T_t * V_i$ and V_i and $T_t * V_i$
- Systematic differences between villages allowed

DID- Extension

- $Y_{it} = \beta_0 + T_t + V_i + \beta_2 \; T_t * V_i + \varepsilon_{it}$
- T_t full set of year fixed effects
- V_i full set of village fixed effects
- Allows for covariance between T_t and $T_t * V_i$ and V_i and $T_t * V_i$
- Systematic differences between villages allowed
- Allow for intervention to occur in years with different outcome variable

Regression Discontinuity Design- RDD

- Resource allocation based on a cutoff- scores, date of birth, rationing cutoffs

Regression Discontinuity Design- RDD

- Resource allocation based on a cutoff- scores, date of birth, rationing cutoffs
- Can use an RD design in such settings

Regression Discontinuity Design- RDD

- Resource allocation based on a cutoff- scores, date of birth, rationing cutoffs
- Can use an RD design in such settings
- Powerful way of addressing selection

Regression Discontinuity Design- RDD

- Resource allocation based on a cutoff- scores, date of birth, rationing cutoffs
- Can use an RD design in such settings
- Powerful way of addressing selection
- Observable characteristics in T and C can be different

Regression Discontinuity Design- RDD

- Resource allocation based on a cutoff- scores, date of birth, rationing cutoffs
- Can use an RD design in such settings
- Powerful way of addressing selection
- Observable characteristics in T and C can be different
- Common support not needed

RDD- Basic Idea

- The control and treated observations are very similar around the cutoff

RDD- Basic Idea

- The control and treated observations are very similar around the cutoff
- Scoring barely above the cutoff matter of chance

RDD- Basic Idea

- The control and treated observations are very similar around the cutoff
- Scoring barely above the cutoff matter of chance
- Unobservable characteristics like ability very similar but one group gets treatment and other does not

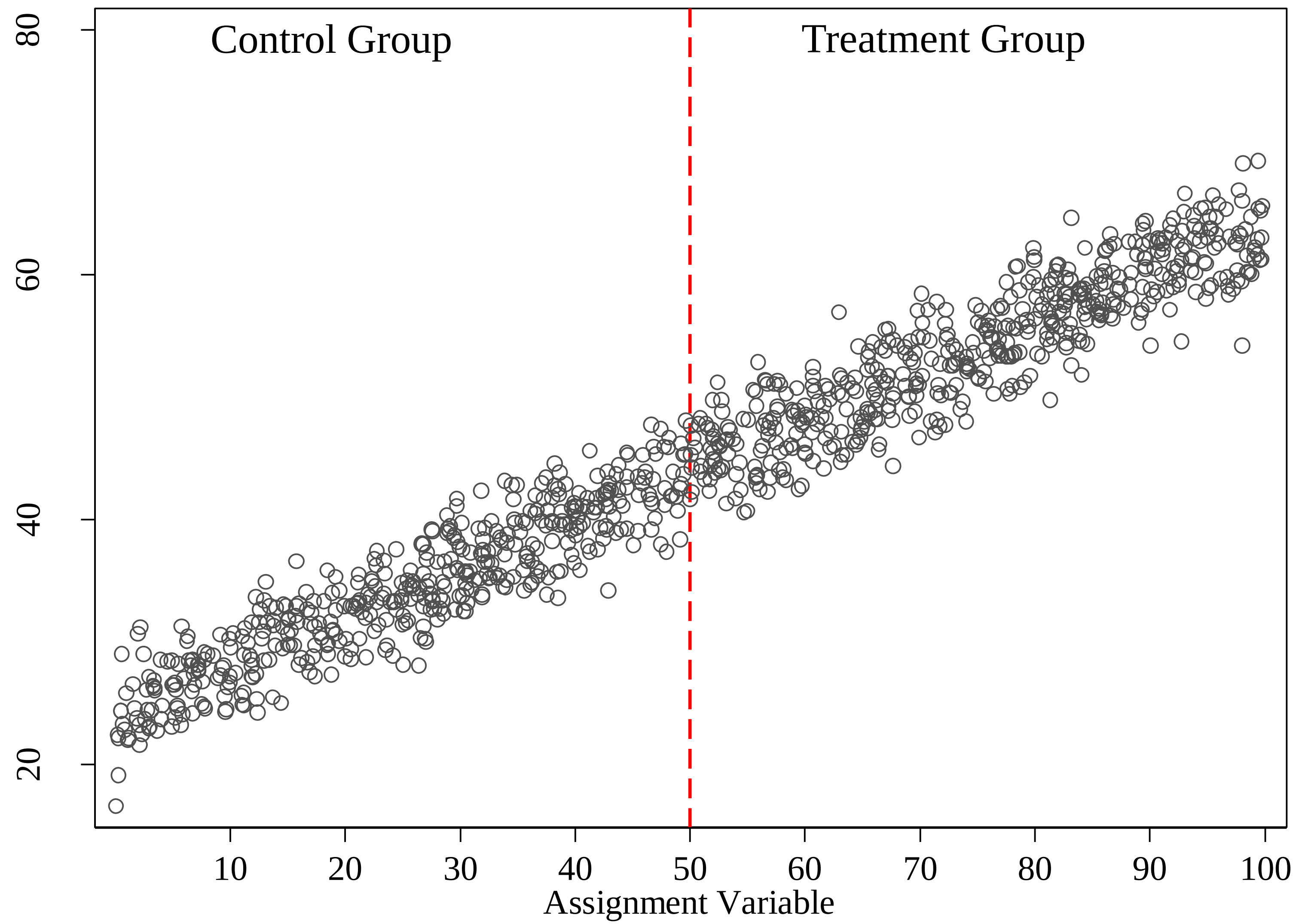
RDD- Basic Idea

- The control and treated observations are very similar around the cutoff
- Scoring barely above the cutoff matter of chance
- Unobservable characteristics like ability very similar but one group gets treatment and other does not
- Selection process completely known and can be modeled

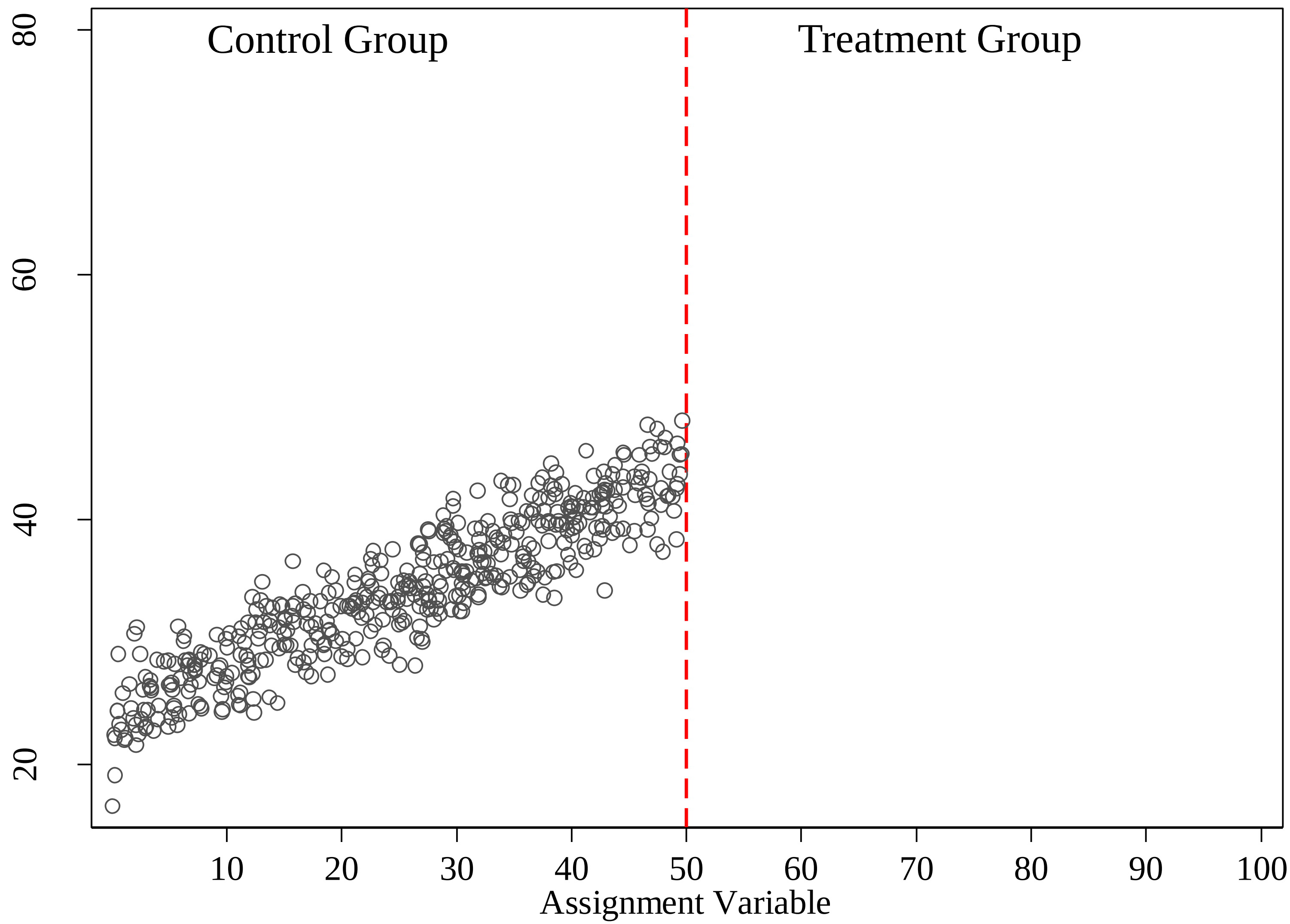
RDD- Basic Idea

- The control and treated observations are very similar around the cutoff
- Scoring barely above the cutoff matter of chance
- Unobservable characteristics like ability very similar but one group gets treatment and other does not
- Selection process completely known and can be modeled
- Regression function between assignment and outcome variable determined

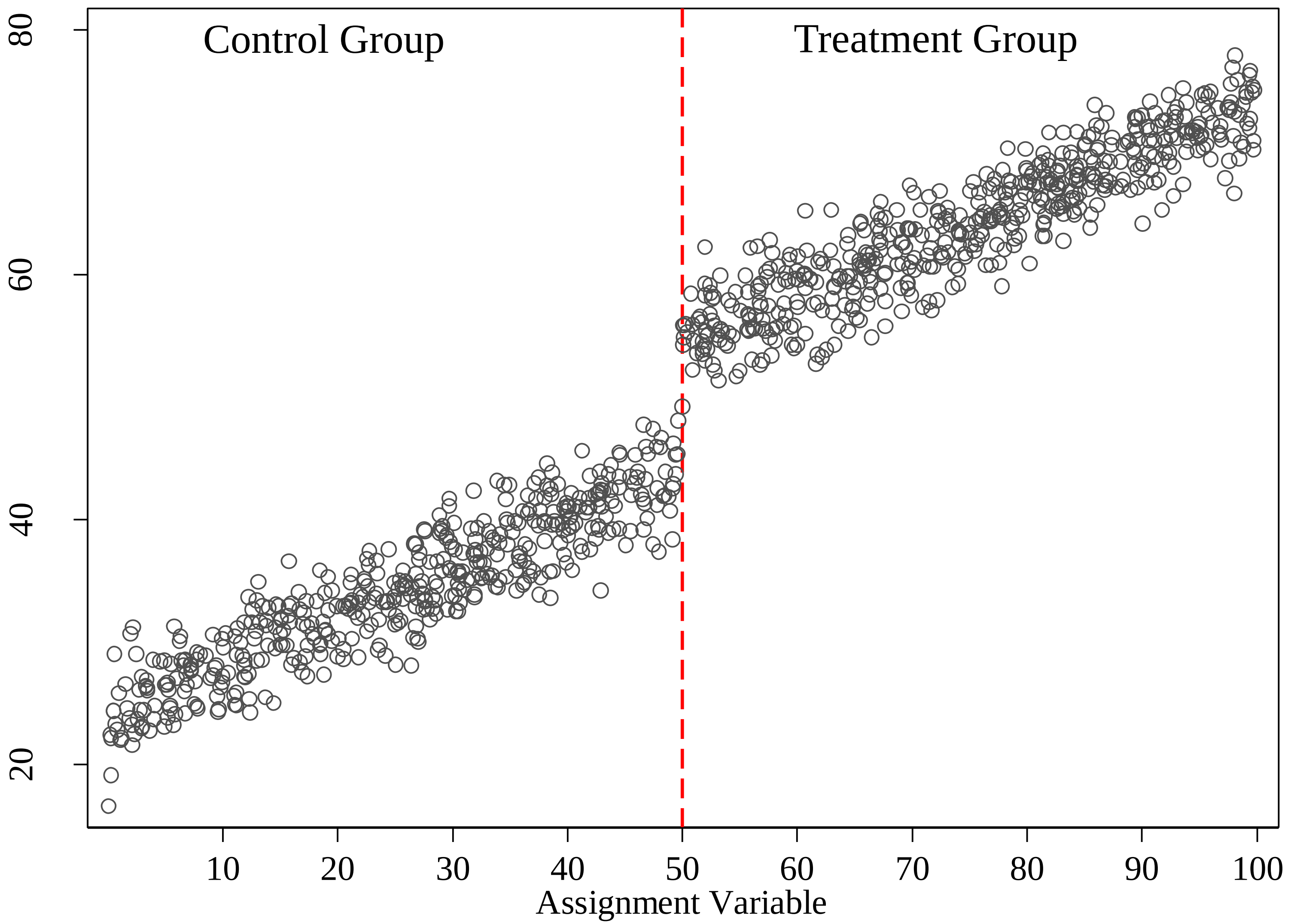
Regression Discontinuity: No Effect



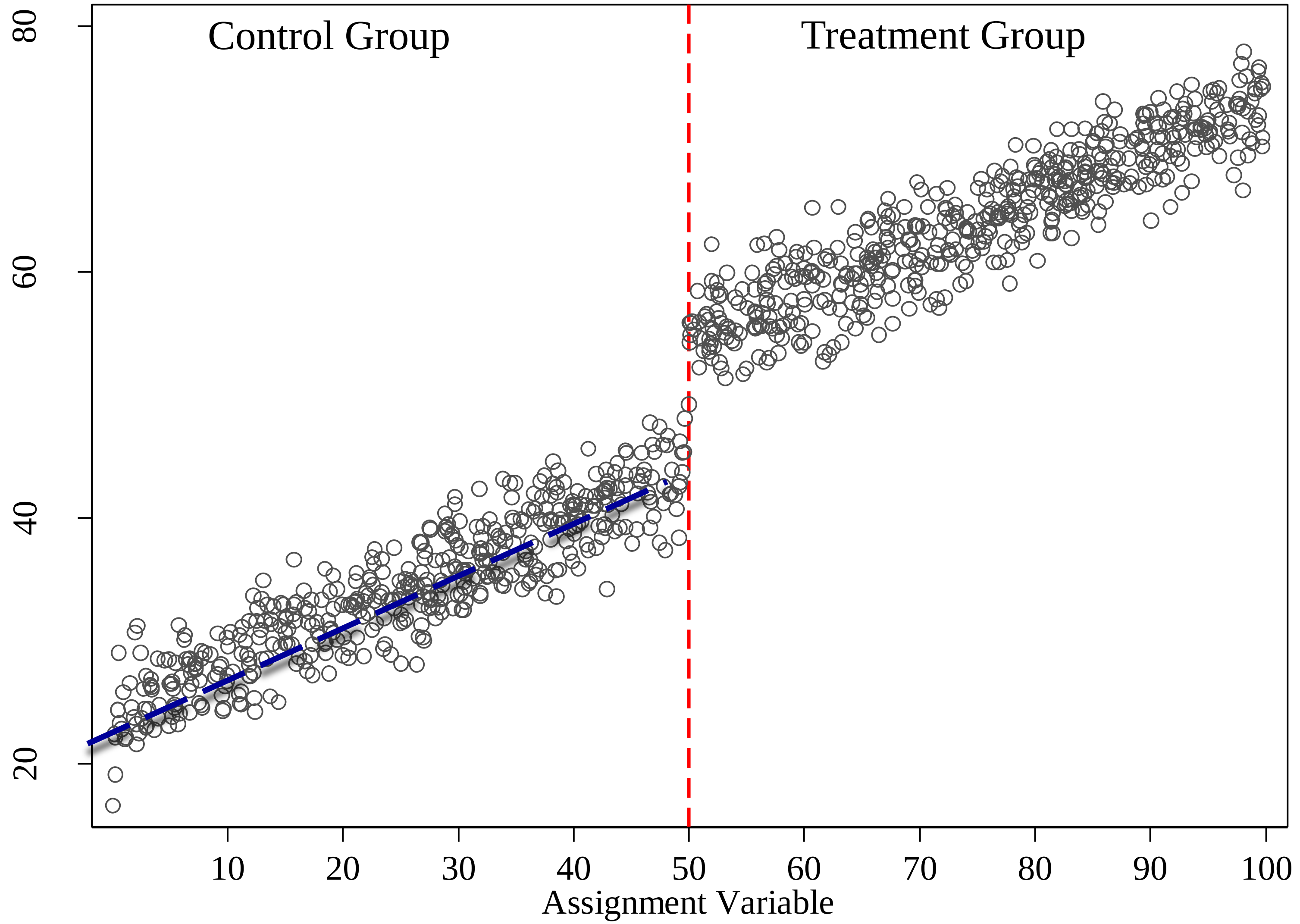
Regression Discontinuity: Significant Effect



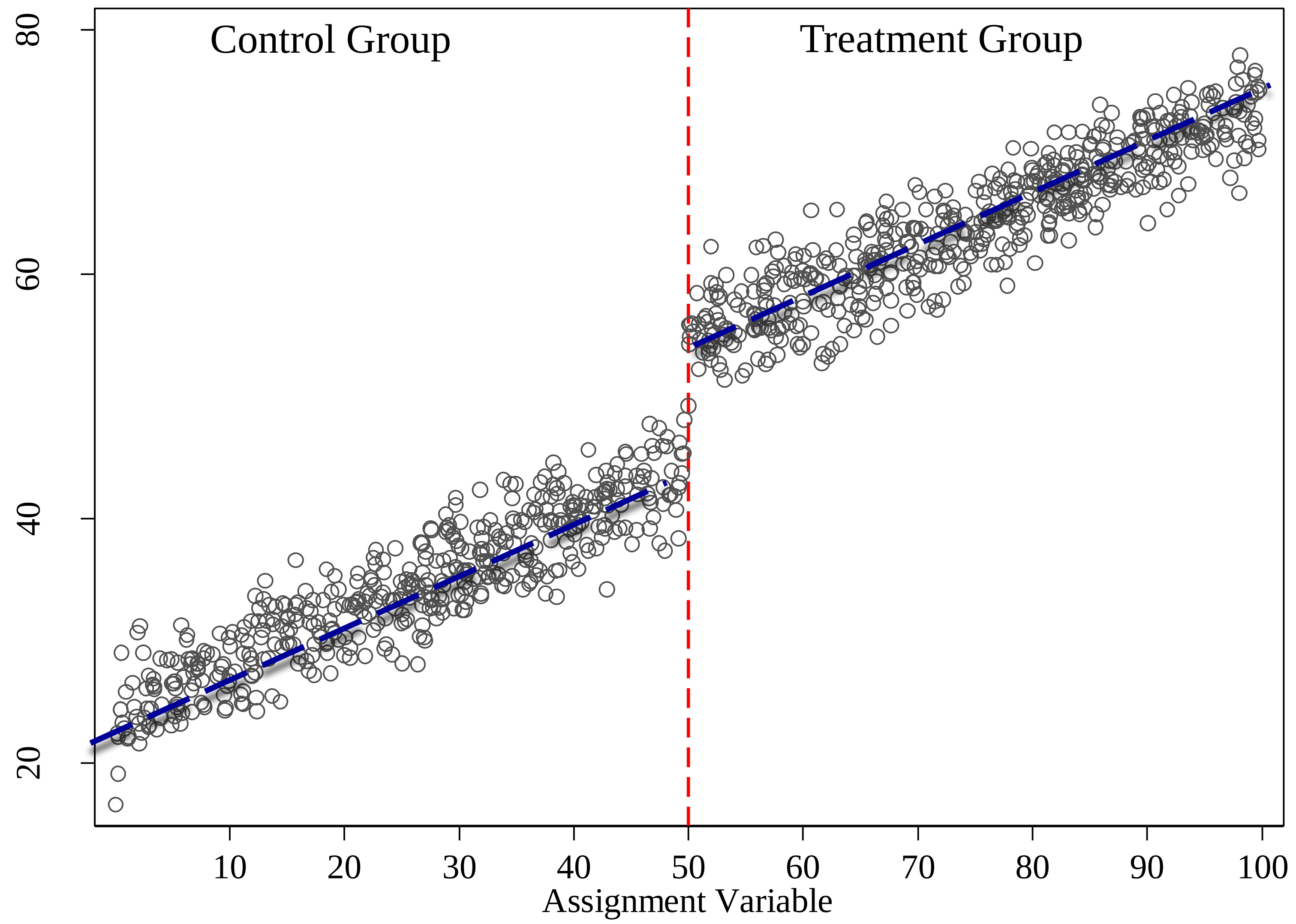
Regression Discontinuity: Significant Effect



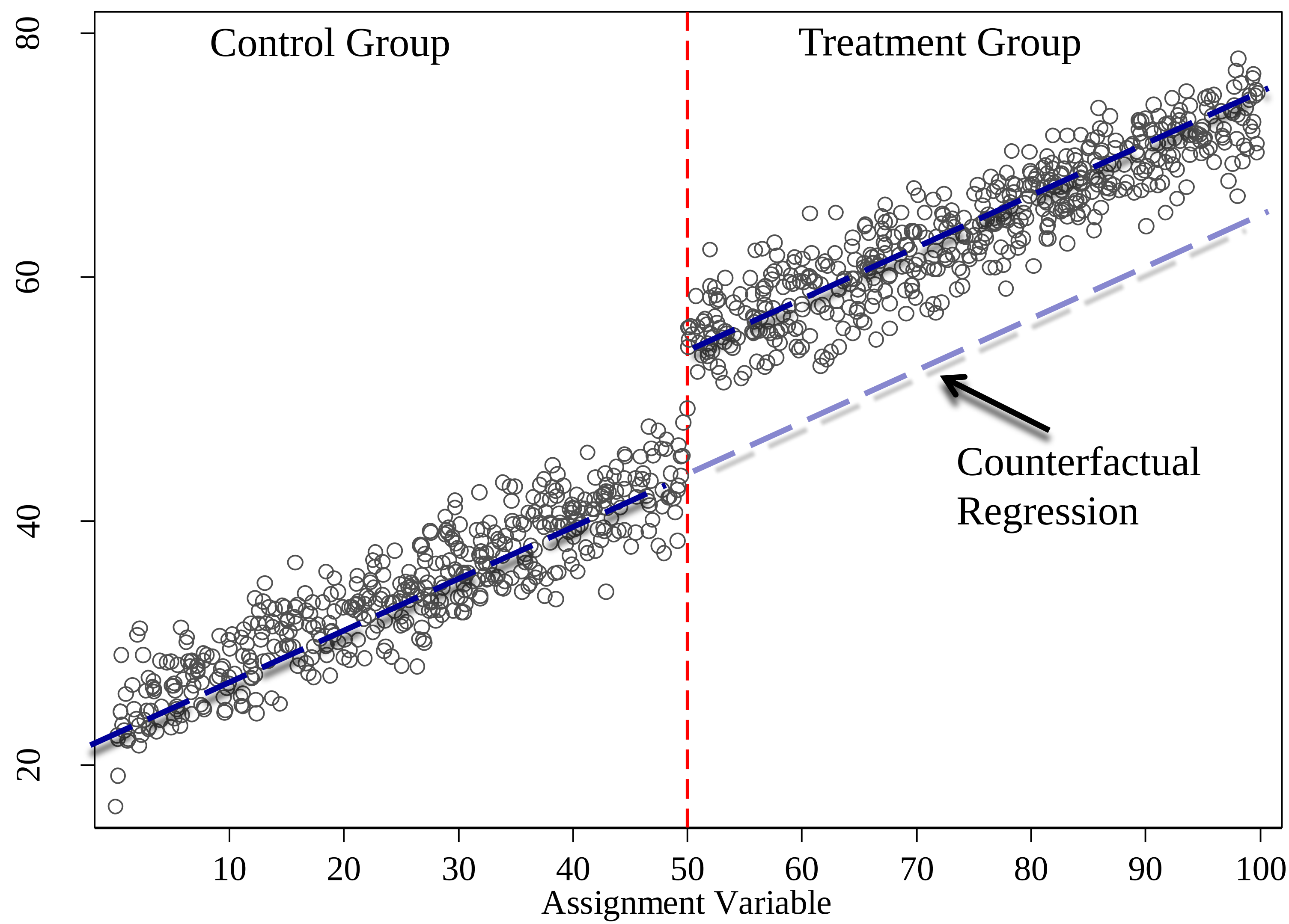
Regression Discontinuity: Significant Effect



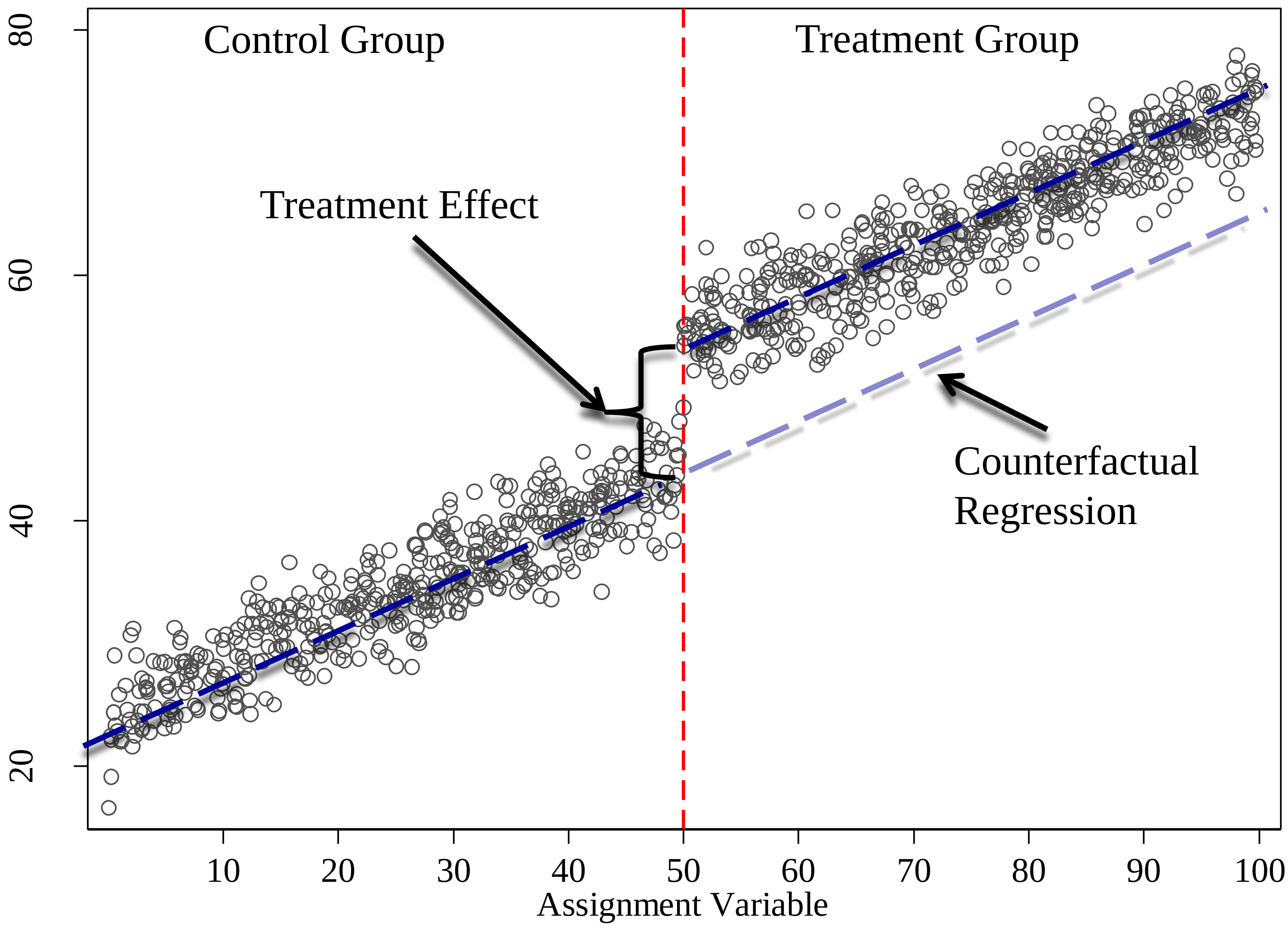
Regression Discontinuity: Significant Effect



Regression Discontinuity: Significant Effect



Regression Discontinuity: Significant Effect



RDD- Implementation

- Probability of treatment should be discontinuous at the cutoff- T sample on one side

RDD- Implementation

- Probability of treatment should be discontinuous at the cutoff- T sample on one side
- Those offered T should take it up and control groups should not be able to get treated

RDD- Implementation

- Probability of treatment should be discontinuous at the cutoff- T sample on one side
- Those offered T should take it up and control groups should not be able to get treated
- Sharp versus fuzzy design require different approaches

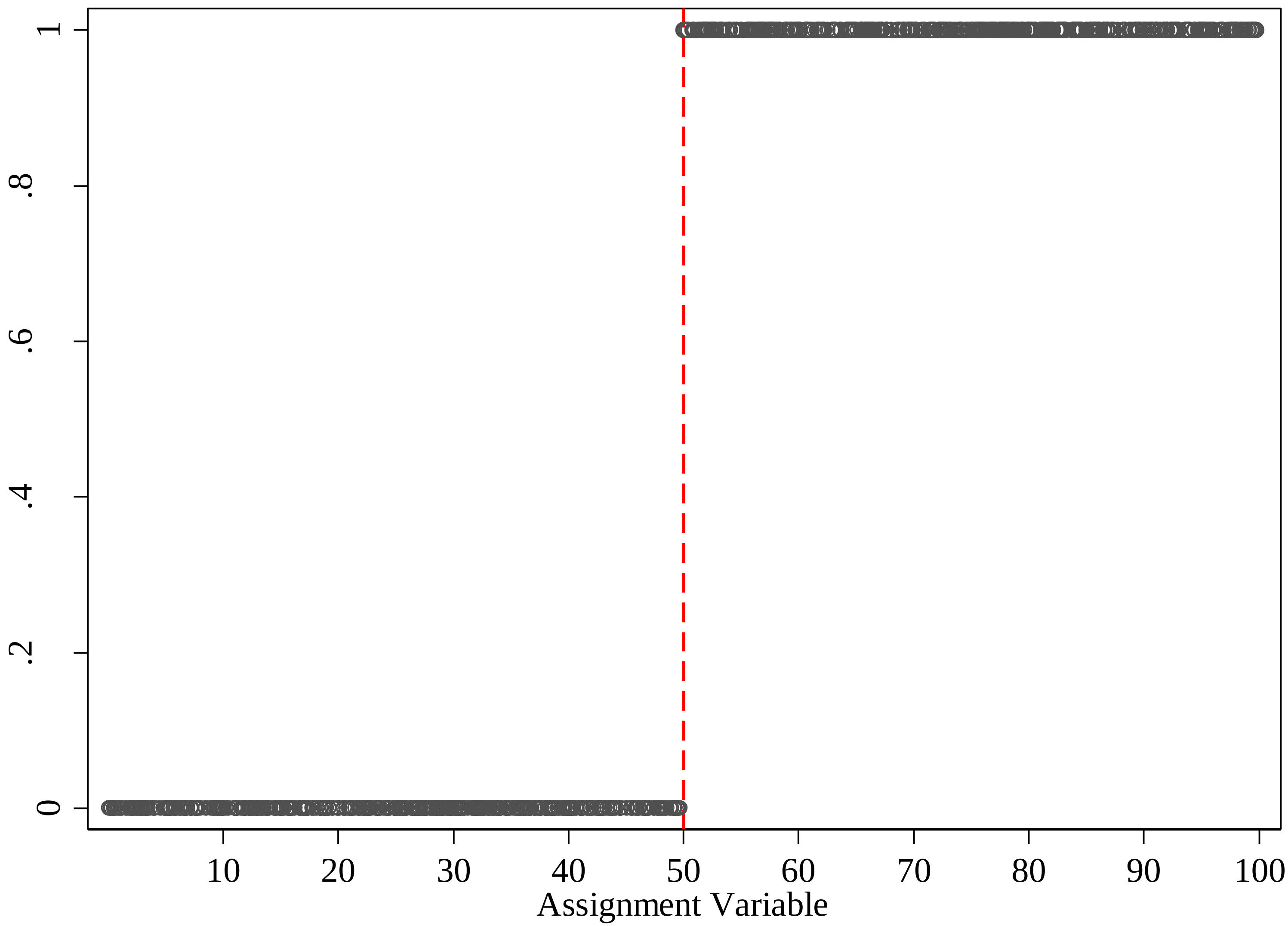
RDD- Implementation

- Probability of treatment should be discontinuous at the cutoff- T sample on one side
- Those offered T should take it up and control groups should not be able to get treated
- Sharp versus fuzzy design require different approaches
- The pr of T changes from 0 to 1 at the cutoff in sharp design

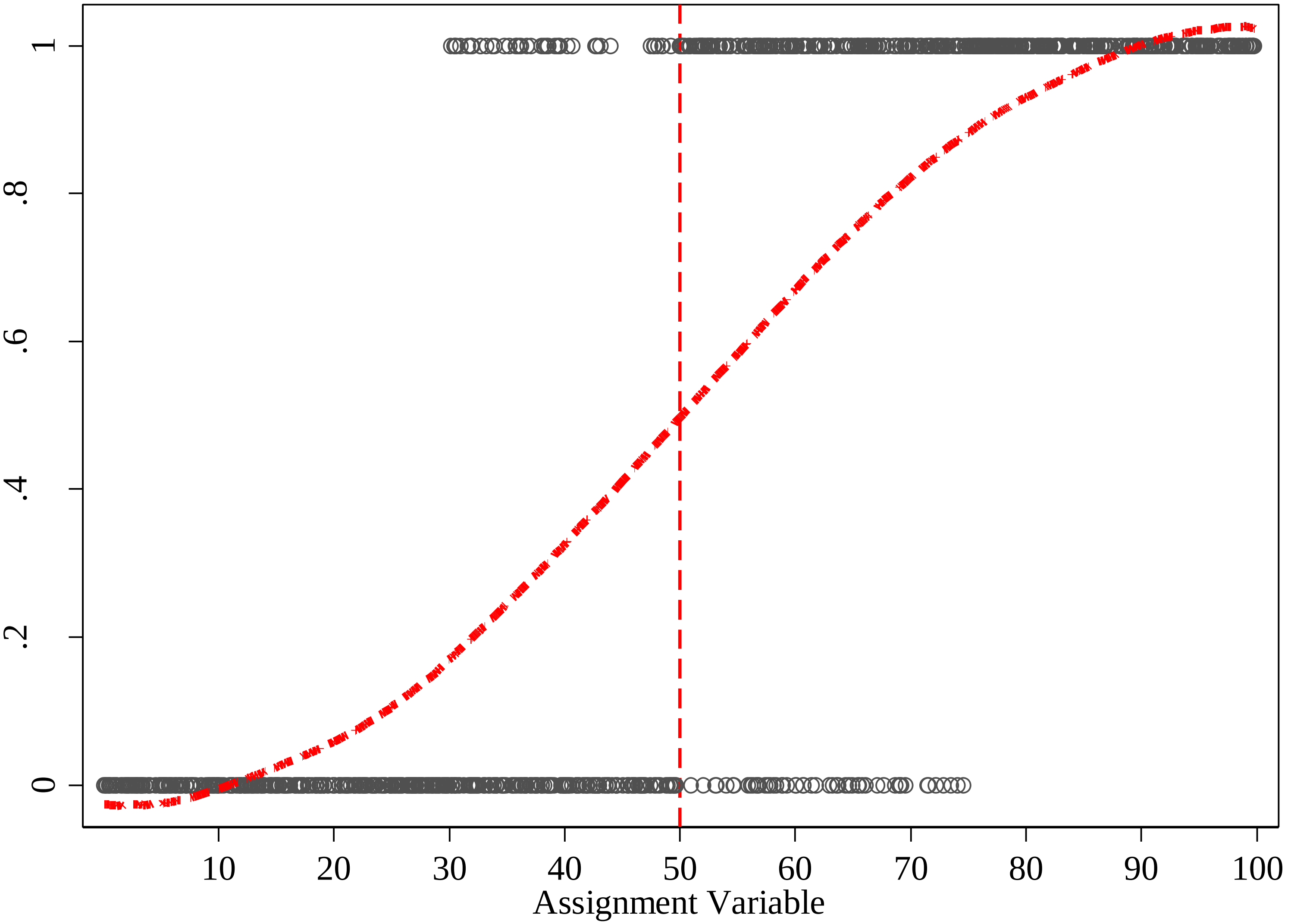
RDD- Implementation

- Probability of treatment should be discontinuous at the cutoff- T sample on one side
- Those offered T should take it up and control groups should not be able to get treated
- Sharp versus fuzzy design require different approaches
- The pr of T changes from 0 to 1 at the cutoff in sharp design
- If the pr does not change very sharply or the over rides are high, use assignment as IV for treatment

Regression Discontinuity: Sharp Design



Regression Discontinuity: Fuzzy Design



RDD- Implementation

- Parametric , semi-parametric or non parametric methods can be used for estimation

RDD- Implementation

- Parametric , semi-parametric or non parametric methods can be used for estimation
- Mis-specified functional form can be a problem

RDD- Implementation

- Parametric , semi-parametric or non parametric methods can be used for estimation
- Mis-specified functional form can be a problem
- Discontinuity in regression functions at the cutoff is the treatment effect

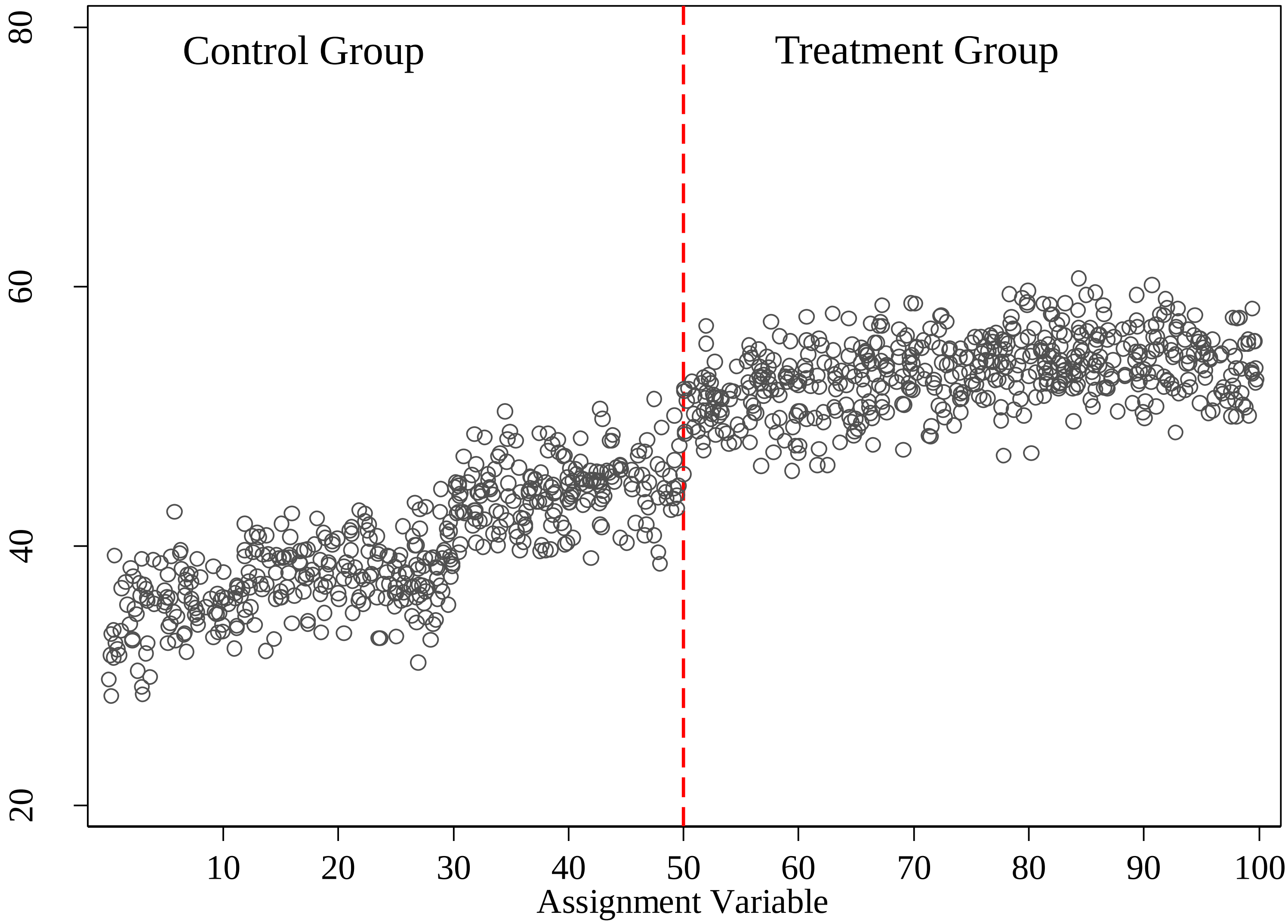
RDD- Implementation

- Parametric , semi-parametric or non parametric methods can be used for estimation
- Mis-specified functional form can be a problem
- Discontinuity in regression functions at the cutoff is the treatment effect
- Functional forms can generate spurious effects or biased effects

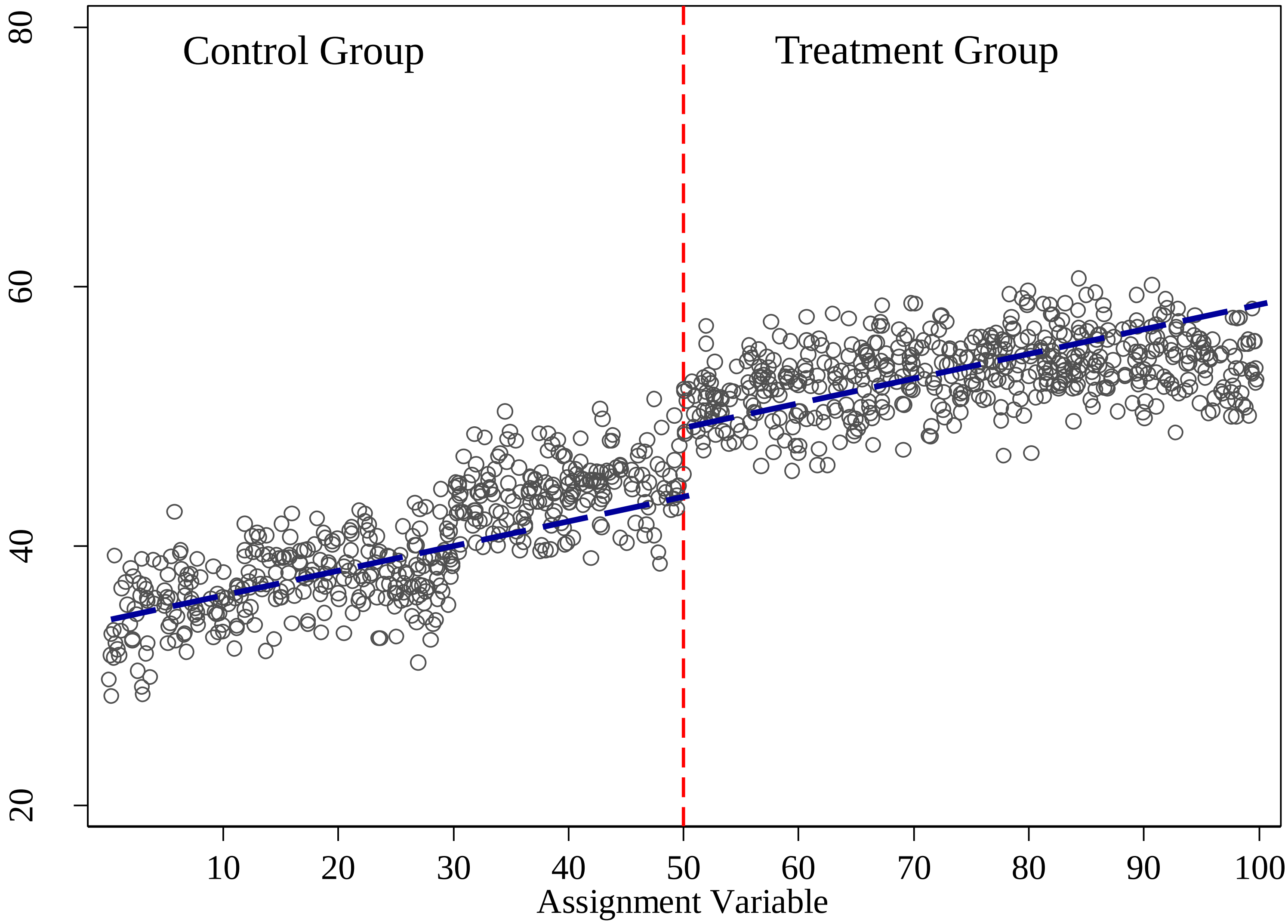
RDD- Implementation

- Parametric , semi-parametric or non parametric methods can be used for estimation
- Mis-specified functional form can be a problem
- Discontinuity in regression functions at the cutoff is the treatment effect
- Functional forms can generate spurious effects or biased effects
- Non linear functional forms estimated as linear regression functions is an example

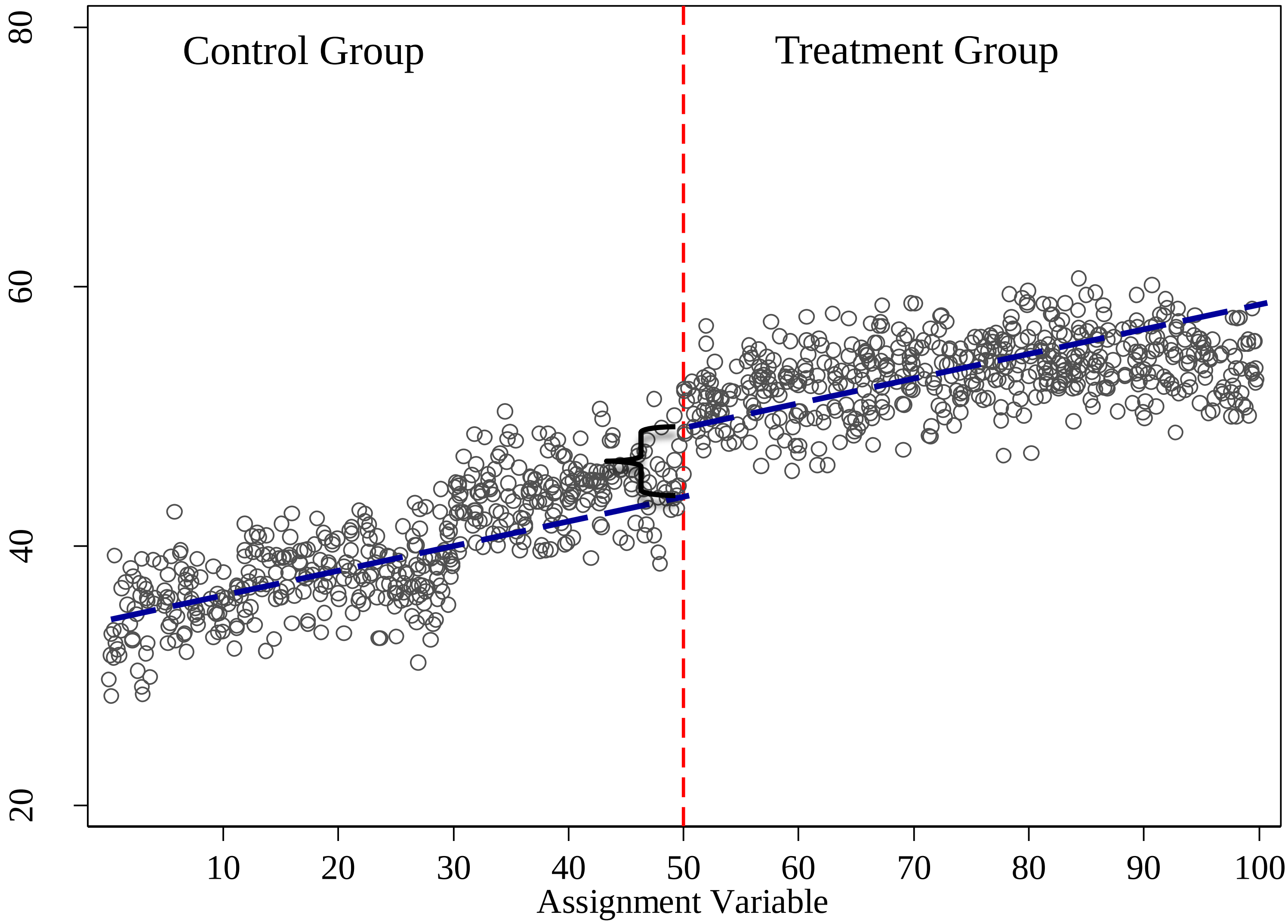
Threats to RD: Nonlinear Functional Form



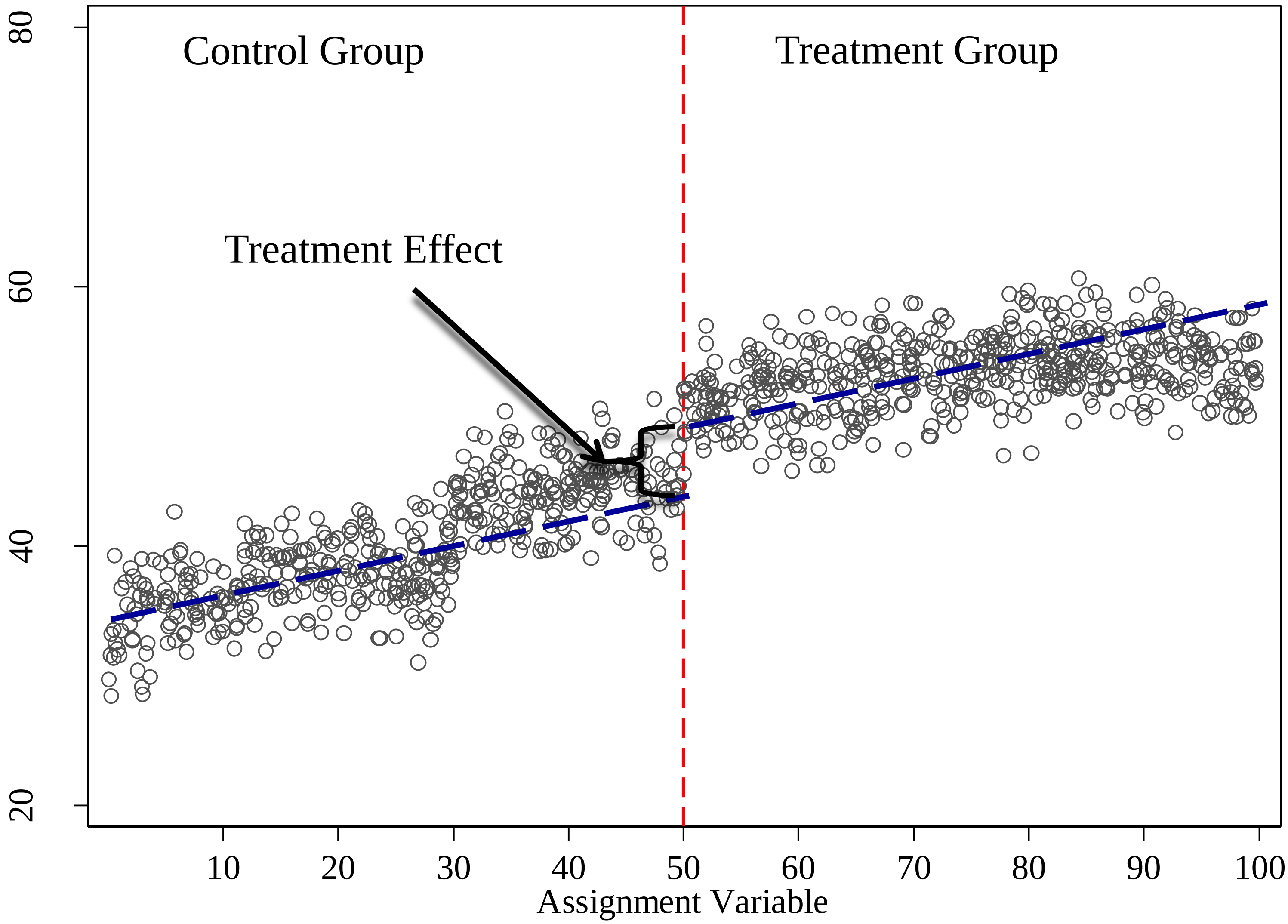
Threats to RD: Nonlinear Functional Form



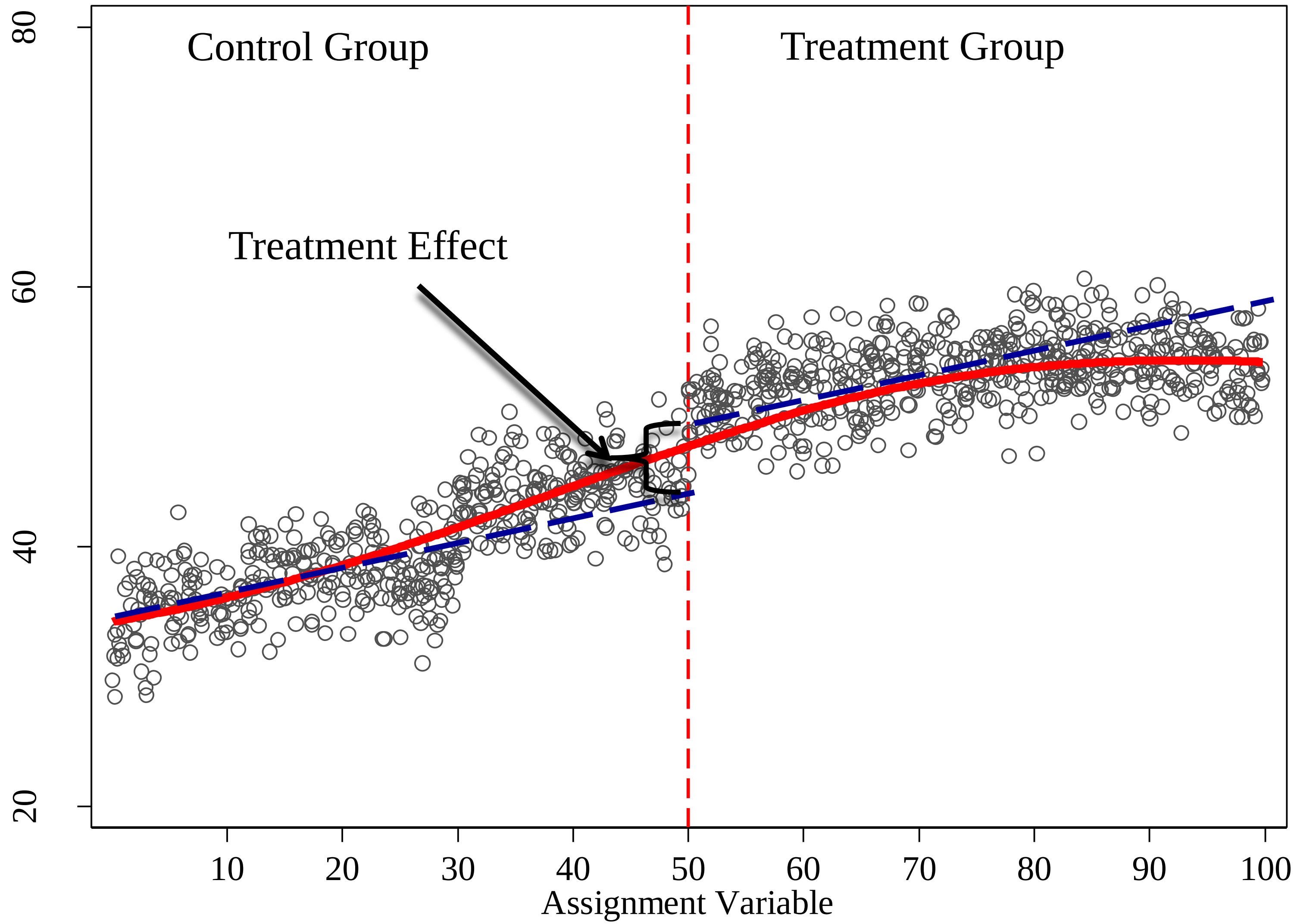
Threats to RD: Nonlinear Functional Form



Threats to RD: Nonlinear Functional Form



Threats to RD: Nonlinear Functional Form



RDD- Functional Form

- Visual inspection of the data - normalize the AV by subtracting the cutoff from the observation score

RDD- Functional Form

- Visual inspection of the data - normalize the AV by subtracting the cutoff from the observation score
- Over fitting the model allowing for interaction terms as well

RDD- Functional Form

- Visual inspection of the data - normalize the AV by subtracting the cutoff from the observation score
- Over fitting the model allowing for interaction terms as well
- Will reduce power and need a lot of data around he cutoff

RDD- Functional Form

- Visual inspection of the data - normalize the AV by subtracting the cutoff from the observation score
- Over fitting the model allowing for interaction terms as well
- Will reduce power and need a lot of data around he cutoff
- Sensitivity analysis to different functional forms

RDD- Functional Form

- Non parametric approaches - local linear regressions

RDD- Functional Form

- Non parametric approaches - local linear regressions
- Sensitivity to bandwidth and kernel choice

RDD- Functional Form

- Non parametric approaches - local linear regressions
- Sensitivity to bandwidth and kernel choice
- In semi parametric approaches, smooth function estimated with splines and covariates can be controlled

RDD- Threats to the Validity

- The cutoffs should be unknown to the population

RDD- Threats to the Validity

- The cutoffs should be unknown to the population
- Cutoffs should not be manipulated

RDD- Threats to the Validity

- The cutoffs should be unknown to the population
- Cutoffs should not be manipulated
- Testing for manipulation, can perform Mcrary's test

RDD- Threats to the Validity

- The cutoffs should be unknown to the population
- Cutoffs should not be manipulated
- Testing for manipulation, can perform Mcrary's test
- Other potential outcomes should be continuous to avoid alternative confounding interpretations

RDD- Threats to the Validity

- The cutoffs should be unknown to the population
- Cutoffs should not be manipulated
- Testing for manipulation, can perform Mcrary's test
- Other potential outcomes should be continuous to avoid alternative confounding interpretations
- Test for continuity of several available control variables

Threats to RD: Manipulation of the Assignment Variable

Threats to RD: Manipulation of the Assignment Variable

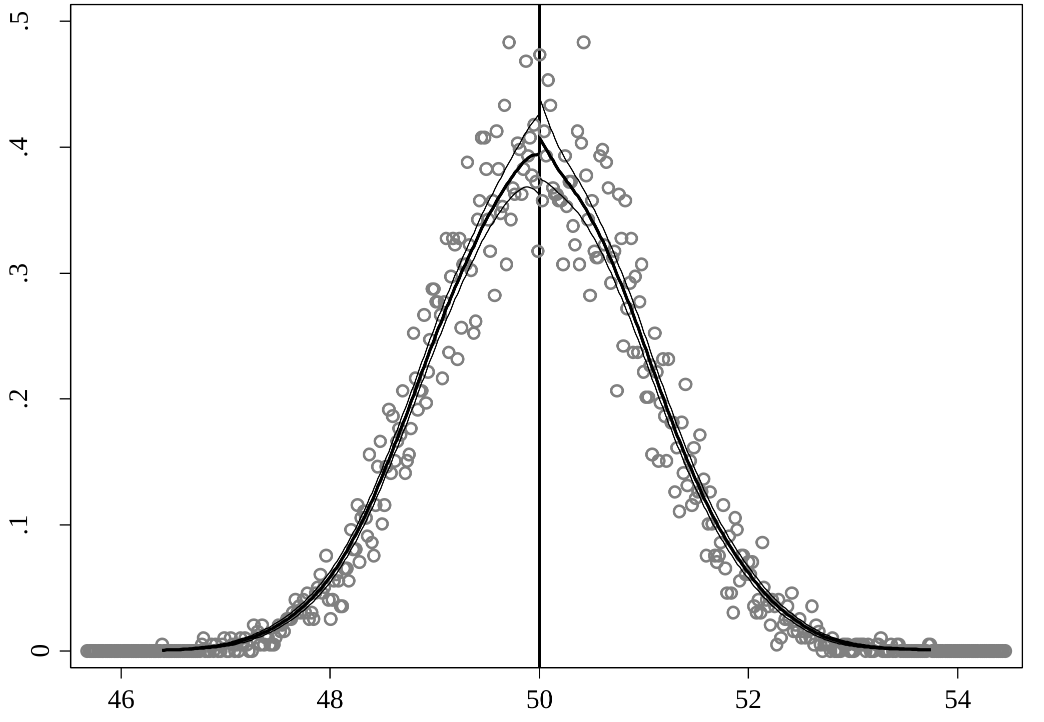
McCrary Test (2008)

- Statistical test for testing discontinuity of the assignment variable at the cutoff point

Threats to RD: Manipulation of the Assignment Variable

McCrary Test (2008)

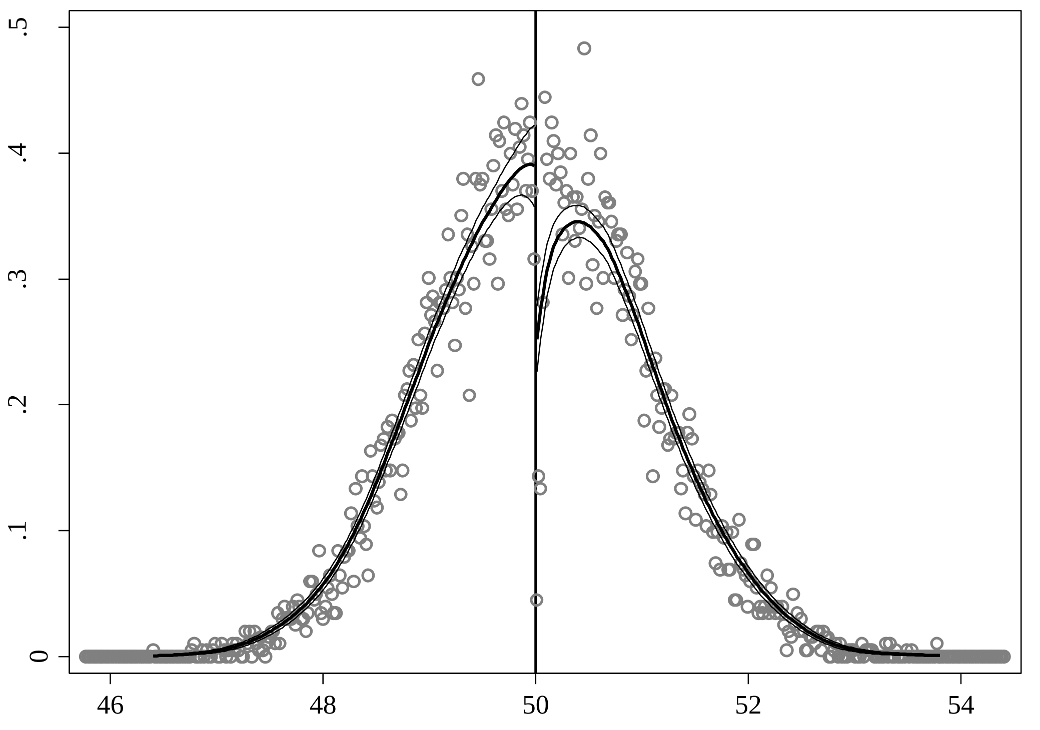
- Statistical test for testing discontinuity of the assignment variable at the cutoff point
- Assignment variable satisfying “McCrary” Test



Threats to RD: Manipulation of the Assignment Variable

McCrary Test (2008)

- Statistical test for testing discontinuity of the assignment variable at the cutoff point
- Assignment variable violating "McCrary" Test



RDD- LATE Estimator

- The limitation of RDD - effect isolated at cutoff
- Cutoff may not be policy relevant or results may not be externally valid
- RD frontier can arise if cutoff varies by years or sites
- Can pool different cutoff to get a more general estimate for the range over which cutoff varies

RDD- LATE Estimator

- The limitation of RDD - effect isolated at cutoff
- Cutoff may not be policy relevant or results may not be externally valid
- RD frontier can arise if cutoff varies by years or sites
- Can pool different cutoff to get a more general estimate for the range over which cutoff varies
- More generalizable but masks heterogeneity