Lecture 12

Travel cost method

Ivan Rudik
AEM 4510

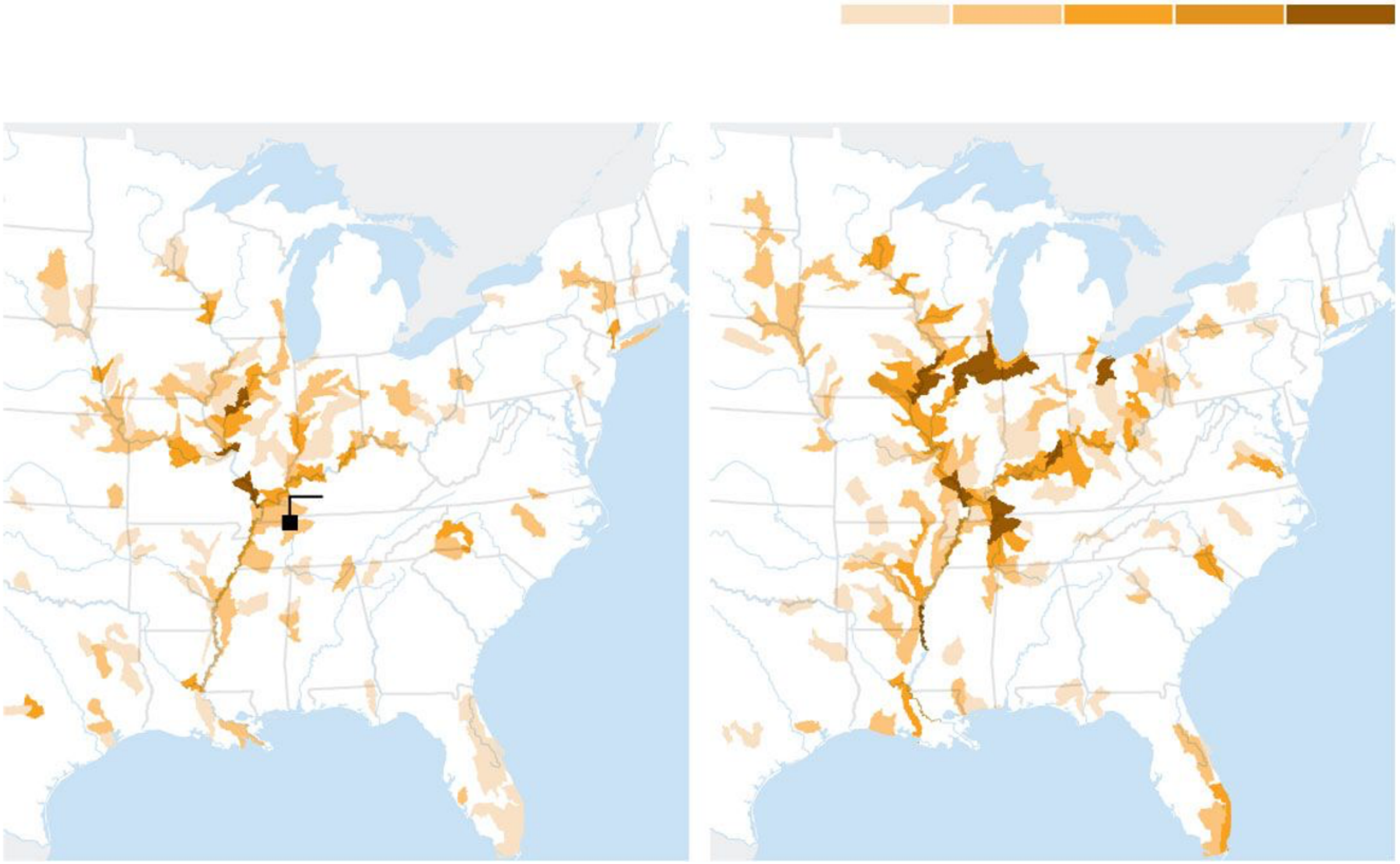
Roadmap

- How do we estimate the value of recreational goods?

Background

Should we separate the Great Lakes and Mississippi?

Should we separate the Great Lakes and Mississippi?



Should we separate the Great Lakes and Mississippi?

The Great Lakes

Carpe diem

Some are worried that Asian carp are poised to invade Lake Michigan

Jul 28th 2012 | From the print edition

Like 21 Tweet 7

WHEN Eric Gittinger, a biologist, goes to work on the Illinois and Mississippi Rivers, he has to look out. The Asian carp that are swimming up from the South, where they escaped from fish farms decades ago, can leap 10 feet in the air or torpedo themselves twice that distance across the water. Larger fish can weigh 40lb (18kg), and Mr Gittinger gets regularly whacked by them.

Yet what most worries people about Asian carp (in fact, several different invasive carp species) is the fact that they are outeating native fish in the rivers, and now seem poised to invade the Great Lakes. This could harm the $7 billion sport-fishing industry, and damage the ecosystem of the largest body of fresh water in the world.

In 2002 the Army Corps of Engineers (ACE) installed a series of electric barriers 37 miles downriver in the Chicago Sanitary and Ship Canal, an artificial channel that links the lakes with the Mississippi and its tributaries. But people fear they may not be working. Recently, multiple traces of Asian-carp DNA have been found in Chicago's Lake Calumet—far beyond the electric fence (see map), and a stone's throw from Lake Michigan.

Should we separate the Great Lakes and Mississippi?

Benefits from barriers accrue to anglers in the Great Lakes, both commercial and recreational

Costs come from cost of building the barriers plus cost of maintaining them, plus costs of reduced shipping (if any), plus any other costs associated with the barriers

How do we figure out the benefits from recreational anglers?

Why do we need travel cost?

Recreational areas **have value**

Why do we need travel cost?

Recreational areas **have value**

Their quality also has value

Why do we need travel cost?

Recreational areas **have value**

Their quality also has value

Not placing a value on recreation is essentially giving it a value of **zero**

Why do we need travel cost?

Recreational areas **have value**

Their quality also has value

Not placing a value on recreation is essentially giving it a value of **zero**

This is likely inappropriate

Why do we need travel cost?

Recreational areas **have value**

Their quality also has value

Not placing a value on recreation is essentially giving it a value of **zero**

This is likely inappropriate

If someone dumped toxic waste in Taughannock does that have zero cost?

What is the travel cost method?

The travel cost method uses observable data on recreation visitation to infer the recreational value of environmental amenities

What is the travel cost method?

The travel cost method uses observable data on recreation visitation to infer the recreational value of environmental amenities

The central idea is that the time and travel cost expenses that people incur to visit a site represent the **price** of access to the site

What is the travel cost method?

The travel cost method uses observable data on recreation visitation to infer the recreational value of environmental amenities

The central idea is that the time and travel cost expenses that people incur to visit a site represent the **price** of access to the site

This means that people's WTP to visit can be estimated based on the number of visits they make to sites of different prices

What is the travel cost method?

The travel cost method uses observable data on recreation visitation to infer the recreational value of environmental amenities

The central idea is that the time and travel cost expenses that people incur to visit a site represent the **price** of access to the site

This means that people's WTP to visit can be estimated based on the number of visits they make to sites of different prices

This gives us a demand curve for sites/amenities, so we can value changes in these environmental amenities

Hotelling

After WWII, the U.S. national park service solicited advice from economists on methods for quantifying the value of specific park properties

Hotelling

After WWII, the U.S. national park service solicited advice from economists on methods for quantifying the value of specific park properties

Would total entrance fee that people pay measure the value?

Hotelling

After WWII, the U.S. national park service solicited advice from economists on methods for quantifying the value of specific park properties

Would total entrance fee that people pay measure the value?

No!

Hotelling

After WWII, the U.S. national park service solicited advice from economists on methods for quantifying the value of specific park properties

Would total entrance fee that people pay measure the value?

No!

Harold Hotelling proposed the first indirect method for measuring the demand of a non-market good in 1947

Hotelling

> Let concentric zones be defined around each park so that the cost of travel to the park from all points in one of these zones is approximately constant. The persons entering the park in a year, or a suitable chosen sample of them, are to be listed according to the zone from which they came. The fact that they come means that the service of the park is at least worth the cost, and this cost can probably be estimated with fair accuracy.

Hotelling

> A comparison of the cost of coming from a zone with the number of people who do come from it, together with a count of the population of the zone, enables us to plot one point for each zone on a demand curve for the service of the park. By a judicious process of fitting, it should be possible to get a good enough approximation to this demand curve to provide, through integration, a measure of consumers' surplus..

Hotelling

> A comparison of the cost of coming from a zone with the number of people who do come from it, together with a count of the population of the zone, enables us to plot one point for each zone on a demand curve for the service of the park. By a judicious process of fitting, it should be possible to get a good enough approximation to this demand curve to provide, through integration, a measure of consumers' surplus..

About twelve years after, Trice and Wood (1958) and Clawson (1959) independently implemented the methodology

Theoretical foundation

Here's the theory for how we can use observed data to tell us something about willingness to pay

Theoretical foundation

Here's the theory for how we can use observed data to tell us something about willingness to pay

Consider a single consumer and a single recreation site

Theoretical foundation

Here's the theory for how we can use observed data to tell us something about willingness to pay

Consider a single consumer and a single recreation site

The consumer has:

- Total number of recreation trips: x, to site of quality: q
- Total budget of time: T
- Working time: H
- Non-recreation, non-work time: l
- Hourly wage: w
- Money cost of reaching the site: c

Theoretical foundation

This lets us write down the consumer's utility maximization problem:

$$\max_{x,z,l} U(x, z, l, q) \quad \text{subject to:} \quad \underbrace{wH = cx + z}_{\text{money budget}}, \ \underbrace{T = H + l + tx}_{\text{time budget}}$$

Theoretical foundation

This lets us write down the consumer's utility maximization problem:

$$\max_{x,z,l} U(x,z,l,q) \quad \text{subject to:} \quad \underbrace{wH = cx + z}_{\text{money budget}}, \ \underbrace{T = H + l + tx}_{\text{time budget}}$$

Combine the two constraints to get:

Theoretical foundation

This lets us write down the consumer's utility maximization problem:

$$\max_{x,z,l} U(x,z,l,q) \quad \text{subject to:} \quad \underbrace{wH = cx + z}_{\text{money budget}},\ \underbrace{T = H + l + tx}_{\text{time budget}}$$

Combine the two constraints to get:

$$\max_{x,z,l} U(x,z,l,q) \quad \text{subject to:} \quad \underbrace{wT = z + (c + wt)x + wl}_{\text{combined money/time budget}}$$

Theoretical foundation

Let $Y = wT$ be the consumer's *full income*, their money value of total time budget

Theoretical foundation

Let $Y = wT$ be the consumer's *full income*, their money value of total time budget

Let $p = c + wt$ be the consumer's *full price*, their total cost to reach the site

Then we can write the problem as:

Theoretical foundation

Let $Y = wT$ be the consumer's *full income*, their money value of total time budget

Let $p = c + wt$ be the consumer's *full price*, their total cost to reach the site

Then we can write the problem as:

$$\max_{x,z,l} U(x, z, l, q) \quad \text{subject to:} \quad \underbrace{Y = z + px + wl}_{\text{combined budget}}$$

Theoretical foundation

Let $Y = wT$ be the consumer's *full income*, their money value of total time budget

Let $p = c + wt$ be the consumer's *full price*, their total cost to reach the site

Then we can write the problem as:

$$\max_{x,z,l} U(x, z, l, q) \quad \text{subject to:} \quad \underbrace{Y = z + px + wl}_{\text{combined budget}}$$

Solve the constraint for z and substitute into the utility function...

Theoretical foundation

$$\max_{x,l} U\left(x, Y - px - wl, l, q\right)$$

Theoretical foundation

$$\max_{x,l} U\left(x, Y - px - wl, l, q\right)$$

This has first-order conditions:

$$[x] \;\; U_x - pU_z = 0 \rightarrow \frac{U_x}{U_z} = p$$

and

$$[l] \;\; -wU_z + U_l = 0 \rightarrow \frac{U_l}{U_z} = w$$

Theoretical foundation

$\frac{U_x}{U_z} = p$ tells us the consumer equates the marginal rate of substitution between recreational trips and consumption to be the full price of the recreational trip

Theoretical foundation

$\frac{U_x}{U_z} = p$ tells us the consumer equates the marginal rate of substitution between recreational trips and consumption to be the full price of the recreational trip

What does this mean?

Theoretical foundation

$\frac{U_x}{U_z} = p$ tells us the consumer equates the marginal rate of substitution between recreational trips and consumption to be the full price of the recreational trip

What does this mean?

The value of the recreational trip to the consumer, in dollar terms, is revealed by the full price p

Theoretical foundation

$$U_x - pU_z = 0 \qquad -wU_z + U_l = 0$$

The above FOCs are two equations, the consumer had two choices (x,l) so we had two unknowns

We can thus solve for x (and l) as a function of the parameters (p,Y,q):

$$x = f(p, Y, q)$$

This is simply the consumer's **demand curves** for recreation as a function of the full price p, full budget Y, and quality q

Theoretical foundation

$$x = f(p, Y, q)$$

If we observe consumers going to sites of different full prices $p_1, p_2, \ldots, p_n$, we are moving up and down their recreation demand curve

Theoretical foundation

$$x = f(p, Y, q)$$

If we observe consumers going to sites of different full prices $p_1, p_2, \ldots, p_n$, we are moving up and down their recreation demand curve

This lets us trace out the demand curve

Theoretical foundation

$$x = f(p, Y, q)$$

If we observe consumers going to sites of different full prices $p_1, p_2, \ldots, p_n$, we are moving up and down their recreation demand curve

This lets us trace out the demand curve

Changing Y or q shifts the demand curve in or out: these are income and quasi-price effects

Theoretical foundation

$$x = f(p, Y, q)$$

If we observe consumers going to sites of different full prices $p_1, p_2, \ldots, p_n$, we are moving up and down their recreation demand curve

This lets us trace out the demand curve

Changing Y or q shifts the demand curve in or out: these are income and quasi-price effects

Once we have it, we can compute surplus!

Zonal (single-site) model

Here's the most basic travel cost model to start

Zonal (single-site) model

Here's the most basic travel cost model to start

- Construct distance zones (i) as concentric circles emanating from the recreation site
    - Travel costs from all points within each zone to the site are sufficiently close in magnitude to justify neglecting the differences

Zonal (single-site) model

Here's the most basic travel cost model to start

- Construct distance zones (i) as concentric circles emanating from the recreation site
  - Travel costs from all points within each zone to the site are sufficiently close in magnitude to justify neglecting the differences
- From a sample of visitors (v_i) at the recreation site, determine zones of origin and their populations (n_i)

Zonal (single-site) model

Here's the most basic travel cost model to start

- Construct distance zones (i) as concentric circles emanating from the recreation site
  - Travel costs from all points within each zone to the site are sufficiently close in magnitude to justify neglecting the differences
- From a sample of visitors (v_i) at the recreation site, determine zones of origin and their populations (n_i)
- Calculate the per capita visitation rates for each zone of origin $(t_i = (v_i/n_i))$

Zonal (single-site) model

- Construct a travel cost measure (tc_i) that reflects the round-trip costs of travel from the zone of origin to the recreation site (time and gas), + an entry fee (fee) which may be zero and does not vary across zones

Zonal (single-site) model

- Construct a travel cost measure (tc_i) that reflects the round-trip costs of travel from the zone of origin to the recreation site (time and gas), + an entry fee (fee) which may be zero and does not vary across zones

- Collect relevant socioeconomic data (s_i) such as income and education for each distance zone

Zonal (single-site) model

- Construct a travel cost measure (tc_i) that reflects the round-trip costs of travel from the zone of origin to the recreation site (time and gas), + an entry fee (fee) which may be zero and does not vary across zones

- Collect relevant socioeconomic data (s_i) such as income and education for each distance zone

- Use statistical methods to estimate the trip demand curve: the relationship between per-capita visitation rates, cost per visit, [and travel costs to other sites (tc_{si})] controlling for socioeconomic differences

Zonal (single-site) model

- Construct a travel cost measure (tc_i) that reflects the round-trip costs of travel from the zone of origin to the recreation site (time and gas), + an entry fee (fee) which may be zero and does not vary across zones

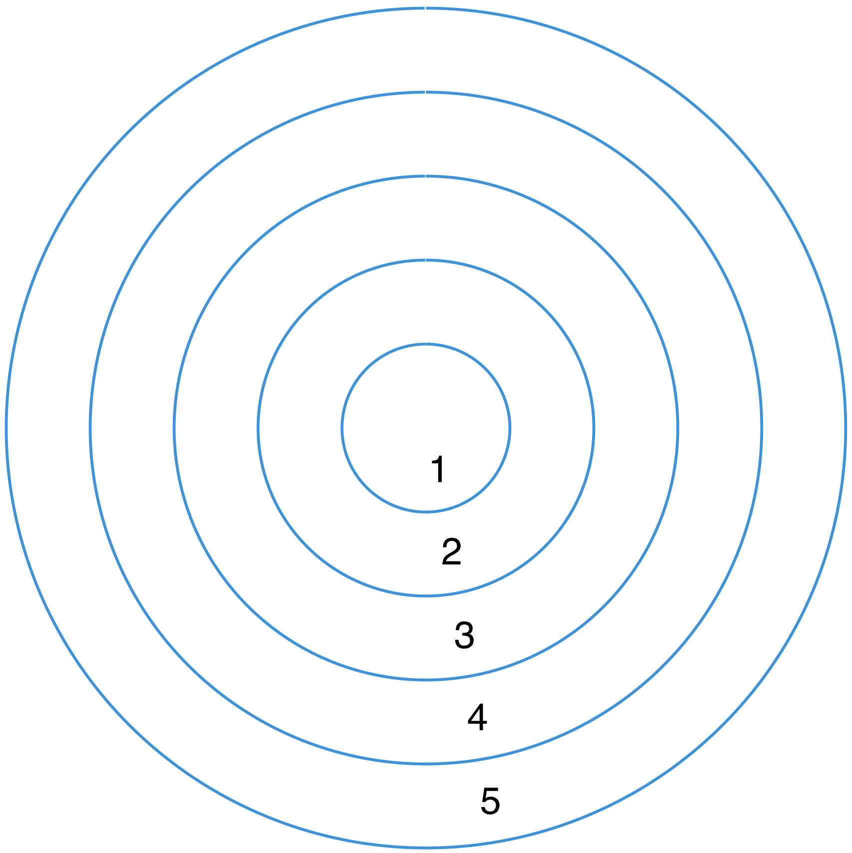
- Collect relevant socioeconomic data (s_i) such as income and education for each distance zone

- Use statistical methods to estimate the trip demand curve: the relationship between per-capita visitation rates, cost per visit, [and travel costs to other sites (tc_{si})] controlling for socioeconomic differences

- $t_i = g(tc_i + fee; tc_{si}, s_i) + \varepsilon_i$ where g can be linear

Zonal (single-site) model

Here's a simple example of a set of zones 1-5:



Zonal (single-site) model

Suppose we have the following data:

```
## # A tibble: 5 × 5
##   zone   dist   pop  cost   vpp
##   <chr> <dbl> <dbl> <dbl> <dbl>
## 1 A         2 10000    20    15
## 2 B        30 10000    30    13
## 3 C        90 20000    65     6
## 4 D       140 10000    80     3
## 5 E       150 10000    90     1
```

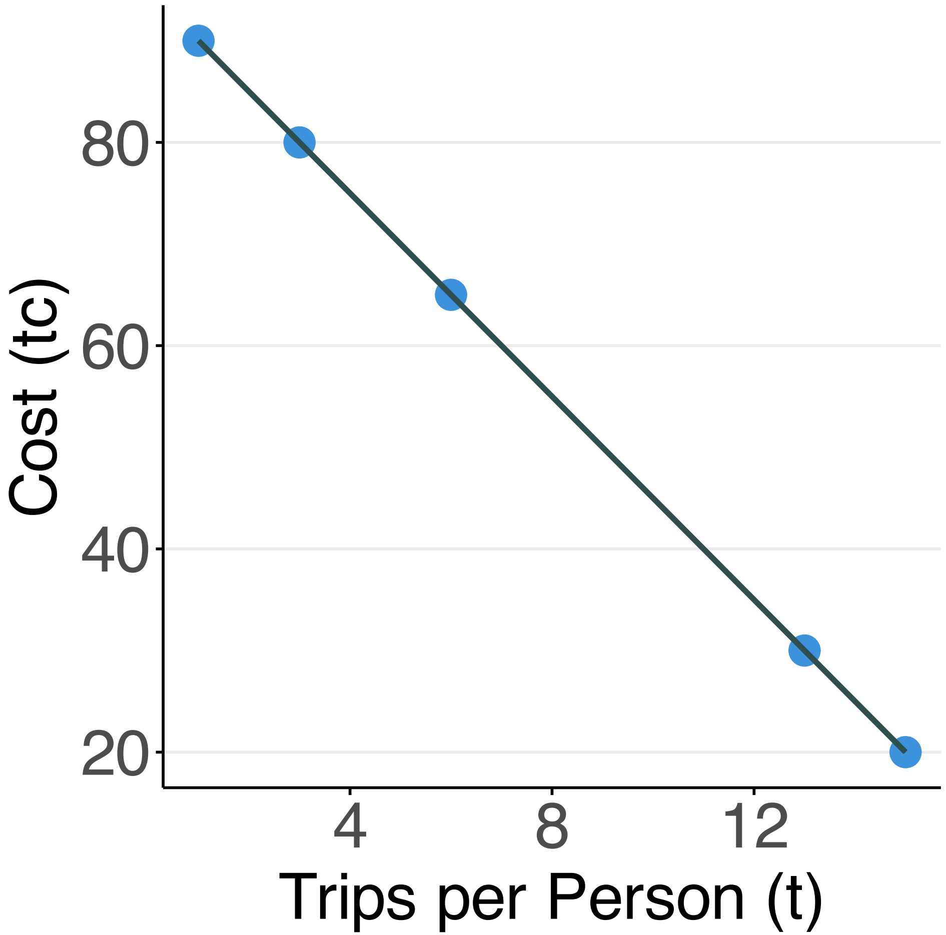
If we plot cost by visits per person, we have a measure of the demand curve...

Zonal (single-site) model

This is a very simple example where it happens to be an exactly straight line, most likely the data won't be this perfect

The line is simply from estimating:

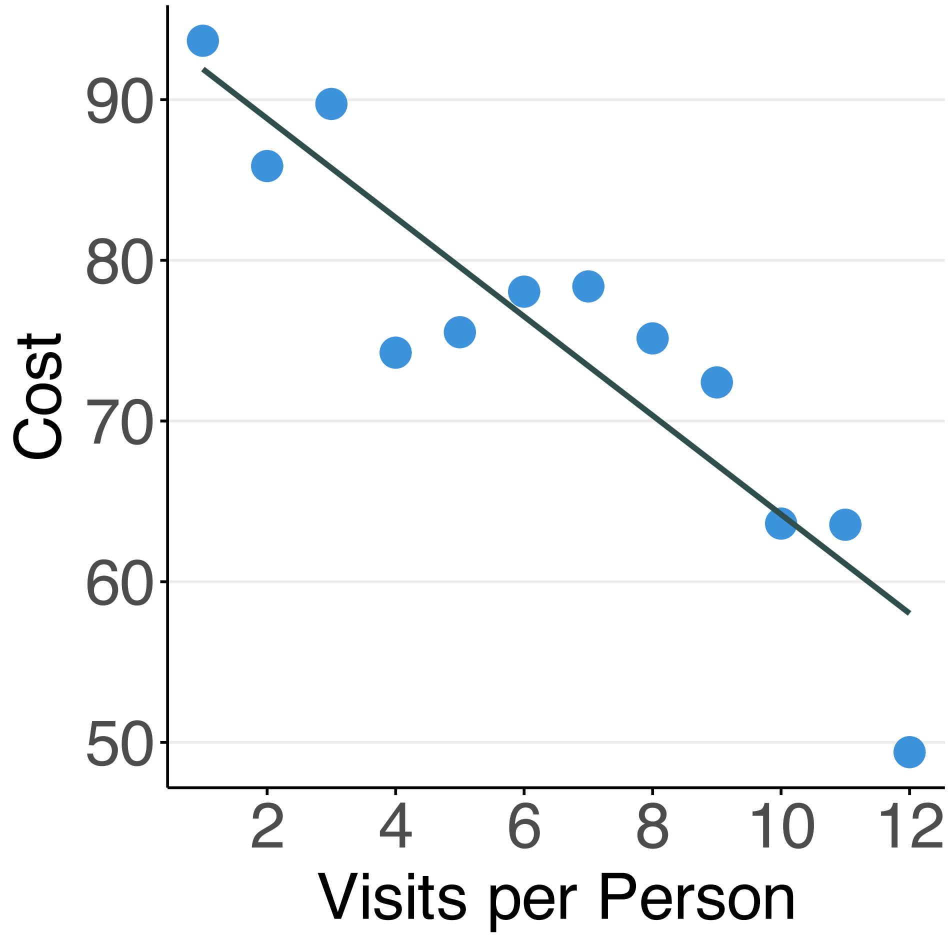
$$t_i = \beta_0 + \beta_1 tc_i + \varepsilon_i$$



Zonal (single-site) model

The data will most likely look like this, but even this is probably too clean

It ignores things like income, other sites, other household characteristics



Zonal (single-site) model

Based on the estimate model coefficients, construct the (inverse) demand curve

Zonal (single-site) model

Based on the estimate model coefficients, construct the (inverse) demand curve

For each zone: predict total visitation given various fees

Zonal (single-site) model

Based on the estimate model coefficients, construct the (inverse) demand curve

For each zone: predict total visitation given various fees

Entry fee on the y-axis (price), and the number of predicted total visits on the x-axis (quantity)

Zonal (single-site) model

Based on the estimate model coefficients, construct the (inverse) demand curve

For each zone: predict total visitation given various fees

Entry fee on the y-axis (price), and the number of predicted total visits on the x-axis (quantity)

The demand curve is different for different zone because different social economic variables

Zonal (single-site) model

Based on the estimate model coefficients, construct the (inverse) demand curve

For each zone: predict total visitation given various fees

Entry fee on the y-axis (price), and the number of predicted total visits on the x-axis (quantity)

The demand curve is different for different zone because different social economic variables

The (use) value of the park/site to each zone is given by the area underneath the corresponding demand curve

Issues with the single-site model

What are some potential issues and concerns with this approach?

Issues with the single-site model

What are some potential issues and concerns with this approach?

It ignores non-use value (automatically zero for non-users)

Issues with the single-site model

What are some potential issues and concerns with this approach?

It ignores non-use value (automatically zero for non-users)

What are the right zones to choose?

Issues with the single-site model

What are some potential issues and concerns with this approach?

It ignores non-use value (automatically zero for non-users)

What are the right zones to choose?

What is the right functional form for demand?

Issues with the single-site model

What are some potential issues and concerns with this approach?

It ignores non-use value (automatically zero for non-users)

What are the right zones to choose?

What is the right functional form for demand?

How do we measure the opportunity cost of time?

Issues with the single-site model

What are some potential issues and concerns with this approach?

It ignores non-use value (automatically zero for non-users)

What are the right zones to choose?

What is the right functional form for demand?

How do we measure the opportunity cost of time?

How do we treat multi-purpose trips?

Issues with the single-site model

What are some potential issues and concerns with this approach?

It ignores non-use value (automatically zero for non-users)

What are the right zones to choose?

What is the right functional form for demand?

How do we measure the opportunity cost of time?

How do we treat multi-purpose trips?

How do we value particular site attributes? Can't disentangle them at a single site

Multi-site model

To value particular site attributes we need to have multiple sites (with different attributes!)

Multi-site model

To value particular site attributes we need to have multiple sites (with different attributes!)

We can answer questions like:

Multi-site model

To value particular site attributes we need to have multiple sites (with different attributes!)

We can answer questions like:

What is the benefit of a fish restocking program?

- Need to know the value of fish catch rate for visitors

Multi-site model

To value particular site attributes we need to have multiple sites (with different attributes!)

We can answer questions like:

What is the benefit of a fish restocking program?

- Need to know the value of fish catch rate for visitors

What is the benefit of water clarity?

Multi-site model

To value particular site attributes we need to have multiple sites (with different attributes!)

We can answer questions like:

What is the benefit of a fish restocking program?

- Need to know the value of fish catch rate for visitors

What is the benefit of water clarity?

What is the benefit of tree replanting?

Multi-site model

Suppose we have a dataset with a large number of individuals and sites

Multi-site model

Suppose we have a dataset with a large number of individuals and sites

Individuals are given by $i = 1, \ldots, N$ and sites are given by $j = 1, \ldots, J$

Multi-site model

Suppose we have a dataset with a large number of individuals and sites

Individuals are given by $i = 1, \ldots, N$ and sites are given by $j = 1, \ldots, J$

We observe the number of times each individual visited each site

Multi-site model

Suppose we have a dataset with a large number of individuals and sites

Individuals are given by $i = 1, \ldots, N$ and sites are given by $j = 1, \ldots, J$

We observe the number of times each individual visited each site

The multi-site model works as follows

Multi-site model

Step 1: Do the single-site estimation for each site:

$$T_{ij} = \beta_{0j} + \beta_{1j} tc_{ij} + \beta_{2j} tc_{sij} + \beta_{3j} s_i + \varepsilon_{ij}$$

Multi-site model

Step 1: Do the single-site estimation for each site:

$$T_{ij} = \beta_{0j} + \beta_{1j} tc_{ij} + \beta_{2j} tc_{sij} + \beta_{3j} s_i + \varepsilon_{ij}$$

Step 2: Recover all the βs from each step 1 regression so that we have a set of J β_{0j}s for $j = 1 \ldots, J$, β_{1j}s for $j = 1 \ldots, J$, etc

Multi-site model

Step 1: Do the single-site estimation for each site:

$$T_{ij} = \beta_{0j} + \beta_{1j}tc_{ij} + \beta_{2j}tc_{sij} + \beta_{3j}s_i + \varepsilon_{ij}$$

Step 2: Recover all the βs from each step 1 regression so that we have a set of J β_{0j}s for $j = 1 \ldots, J$, β_{1j}s for $j = 1 \ldots, J$, etc

These βs tell us the slope (β_{1j}) and intercept $(\beta_{0j}, \beta_{2j}, \beta_{3j})$

Multi-site model

Step 1: Do the single-site estimation for each site:

$$T_{ij} = \beta_{0j} + \beta_{1j} tc_{ij} + \beta_{2j} tc_{sij} + \beta_{3j} s_i + \varepsilon_{ij}$$

Step 2: Recover all the βs from each step 1 regression so that we have a set of J β_{0j}s for $j = 1 \ldots, J$, β_{1j}s for $j = 1 \ldots, J$, etc

These βs tell us the slope (β_{1j}) and intercept ($\beta_{0j}, \beta_{2j}, \beta_{3j}$)

β_{2j}, β_{3j} capture how the cost of substitute sites and household characteristics: they shift demand up and down

Multi-site model

Step 3: Take each set of J coefficient estimates and use them as the dependent variable in a regression on site attributes z:

$$\hat{\beta}_{0j} = \alpha_{00} + \alpha_{01} z_j + \epsilon_{0j}$$

$$\hat{\beta}_{1j} = \alpha_{10} + \alpha_{11} z_j + \epsilon_{1j}$$

$$\hat{\beta}_{2j} = \alpha_{20} + \alpha_{21} z_j + \epsilon_{2j}$$

$$\hat{\beta}_{3j} = \alpha_{30} + \alpha_{31} z_j + \epsilon_{3j}$$

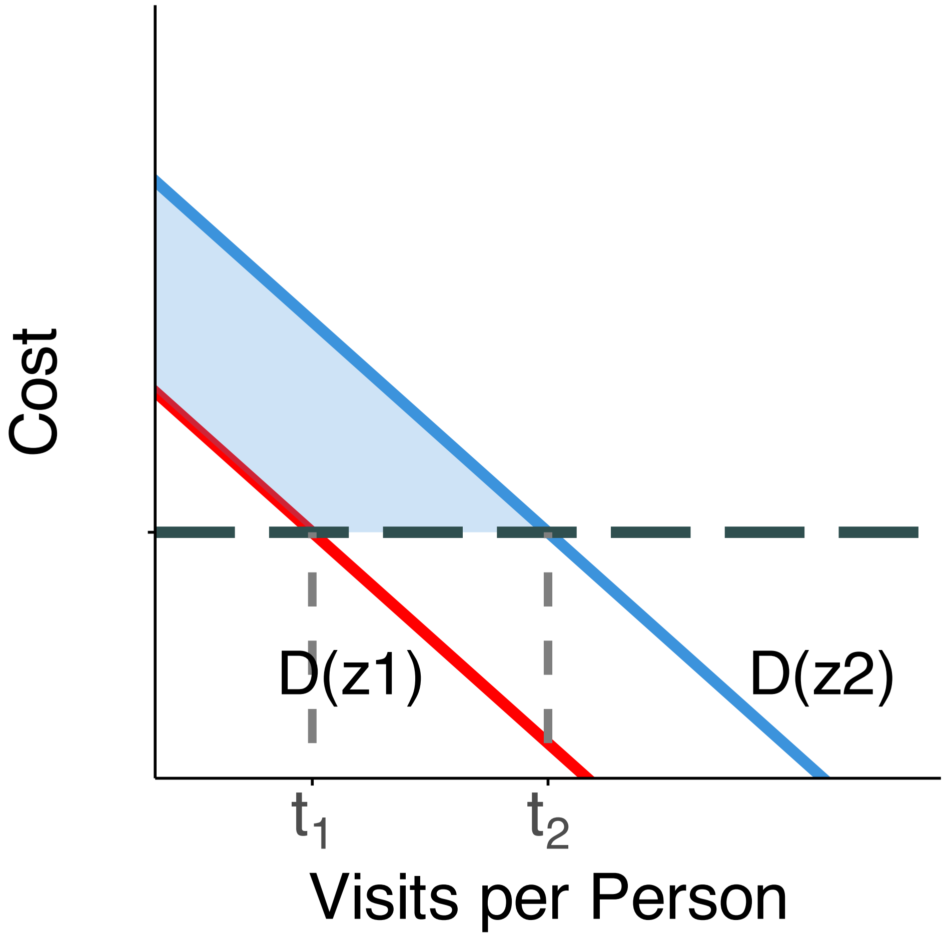
The $\alpha_{\times 1}$ coefficients tell us how the demand curve shifts $(\alpha_{00}, \alpha_{02}, \alpha_{03})$ or rotates (α_{01}) as we change z

Valuing attributes with a multi-site model

If we improve the quality of a site from z_1 to z_2, demand for that site shifts up

The gain in CS, holding the cost fixed, is given by the blue area

Once we estimate demand curves, we can see how welfare changes when we alter quality characteristics!



Multi-site example

```
trip_data
```

```
## # A tibble: 2,600 × 7
##    house_num  site trips income travel_cost travel_cost_other water_clarity
##        <int> <int> <dbl>  <dbl>       <dbl>             <dbl>         <dbl>
##  1         1     1     4 40450.       38.9              16.4          0.506
##  2         2     1     5 60304.       29.8              37.5          0.506
##  3         3     1     5 66681.       42.2              67.2          0.506
##  4         4     1     5 52886.       11.0              51.3          0.506
##  5         5     1     5 69282.       15.7               7.72         0.506
##  6         6     1     5 36948.        4.30             48.0          0.506
##  7         7     1     6 60866.        5.31             91.0          0.506
##  8         8     1     5 35557.       65.0             161.           0.506
##  9         9     1     5 64880.       14.5              24.3          0.506
## 10        10     1     4 38491.       13.6              26.5          0.506
## # … with 2,590 more rows
```

First stage estimation

```r
# first stage of multi-site
site_estimates <- map_dfr(unique(trip_data$site), function(site_in){
  lm(trips ~ travel_cost + travel_cost_other + income,
     trip_data %>% filter(site == site_in)) %>%
    broom::tidy() %>%
    select(estimate) %>%
    mutate(site = site_in) %>%
    list() %>%
    tibble_row() %>%
    unlist()
}) %>%
  select(1:5) %>%
  magrittr::set_colnames(c("intercept", "own_price", "cross_price", "income", "site"))
```

First stage estimation

```
site_estimates
```

```
## # A tibble: 26 × 5
##    intercept own_price cross_price    income  site
##        <dbl>     <dbl>       <dbl>     <dbl> <dbl>
##  1     2.99   -0.0161      0.0106  0.0000321     1
##  2     2.45   -0.0117      0.0101  0.0000397     2
##  3     2.37   -0.0197      0.0111  0.0000450     3
##  4     2.33   -0.0187      0.0119  0.0000438     4
##  5     2.05   -0.0143      0.0139  0.0000450     5
##  6    -0.236  -0.00668     0.00972 0.0000321     6
##  7     2.67   -0.0210      0.0118  0.0000395     7
##  8    -0.346  -0.00395     0.00987 0.0000324     8
##  9     2.98   -0.0133      0.0107  0.0000315     9
## 10    -0.103  -0.00943     0.0105  0.0000302    10
## # … with 16 more rows
```

Take estimates, join with water clarity

```
# merge in water clarity
estimation_df <- site_estimates %>%
  left_join(trip_data %>% distinct(site, water_clarity))
```

```
## Joining, by = "site"
```

```
estimation_df
```

```
## # A tibble: 26 × 6
##    intercept own_price cross_price    income  site water_clarity
##        <dbl>     <dbl>       <dbl>     <dbl> <dbl>         <dbl>
##  1     2.99   -0.0161      0.0106  0.0000321     1         0.506
##  2     2.45   -0.0117      0.0101  0.0000397     2         0.503
##  3     2.37   -0.0197      0.0111  0.0000450     3         0.515
##  4     2.33   -0.0187      0.0119  0.0000438     4         0.515
##  5     2.05   -0.0143      0.0139  0.0000450     5         0.515
##  6    -0.236  -0.00668     0.00972 0.0000321     6         0.481
##  7     2.67   -0.0210      0.0118  0.0000395     7         0.539
##  8    -0.346  -0.00395     0.00987 0.0000324     8         0.482
```

Second stage

```
# second stage of multi-site
demand_shifts <- map_dfr(names(estimation_df)[1:4],
        function(coefficient) {
          reg_formula <- as.formula(paste0(coefficient, " ~ water_clarity"))
          lm(reg_formula, estimation_df) %>%
            broom::tidy() %>%
            mutate(coeff = coefficient) %>%
            slice(2)
        }
) |>
  select(term, estimate, coeff)
```

Second stage

```r
demand_shifts
```

```
## # A tibble: 4 × 3
##   term          estimate coeff
##   <chr>            <dbl> <chr>
## 1 water_clarity 48.0      intercept
## 2 water_clarity -0.171    own_price
## 3 water_clarity  0.0241   cross_price
## 4 water_clarity  0.000165 income
```

The estimates column tells us how a change in water clarity (from 0 to 100%), shifts or rotates our demand curve

Real world data: central park

Standard travel cost method is costly

Real world data: central park

Standard travel cost method is costly

Need to survey households

Real world data: central park

Standard travel cost method is costly

Need to survey households

This takes time and money

Real world data: central park

Standard travel cost method is costly

Need to survey households

This takes time and money

What alternatives do we have?

Mobility data from cell phones

Cell phones track where people live, go, etc

Mobility data from cell phones

Cell phones track where people live, go, etc

We can use these data to do the travel cost method

Mobility data from cell phones

Cell phones track where people live, go, etc

We can use these data to do the travel cost method

Same data used by NYT, WaPo, etc for COVID analysis of restaurants, etc

Mobility data from cell phones

Cell phones track where people live, go, etc

We can use these data to do the travel cost method

Same data used by NYT, WaPo, etc for COVID analysis of restaurants, etc

Here we will be looking at visits to central park

Mobility data from cell phones

```
central_park_data = read_csv("data/12-central-park-phone-data.csv")
central_park_data
```

```
## # A tibble: 22,972 × 13
##    visitor_cbgs  year month location_name       latitude longitude scaled_visits
##           <dbl> <dbl> <dbl> <chr>                  <dbl>     <dbl>         <dbl>
##  1 340030032003  2018     8 Harlem Meer             40.8     -74.0          34.8
##  2 340030032003  2018     8 Harlem Meer             40.8     -74.0          69.5
##  3 340030032003  2018     8 Harlem Meer             40.8     -74.0          34.8
##  4 340030034011  2018    11 Diana Ross Playgro…     40.8     -74.0          59.8
##  5 340030034011  2019     8 Diana Ross Playgro…     40.8     -74.0          46
##  6 340030034011  2019    11 Central Park            40.8     -74.0          92.9
##  7 340030034023  2018     9 East 72nd Street P…     40.8     -74.0         257.
##  8 340030035002  2018     3 East 72nd Street P…     40.8     -74.0         184.
##  9 340030035002  2019     5 Cherry Hill Founta…     40.8     -74.0          38.4
## 10 340030040022  2018     1 Central Park            40.8     -74.0         110.
## # … with 22,962 more rows, and 6 more variables: visits <dbl>,
## #   travel_distance_km <dbl>, travel_time_minutes <dbl>,
## #   visits_per_person <dbl>, median_age <dbl>, median_income <dbl>
```

Real world data: central park

The dataframe tells us for each **census block group (CBG)** (600-3000 person locations):

- visits per month to a particular location in central park by all cell phones in the CBG
- how far the CBG is from the central park location (time and distance)
- The median income of the CBG
- The median age of the CBG

```
## # A tibble: 22,972 × 13
##    visitor_cbgs  year month location_name        latitude longitude scaled_visits
##           <dbl> <dbl> <dbl> <chr>                   <dbl>     <dbl>         <dbl>
##  1 340030032003  2018     8 Harlem Meer              40.8     -74.0          34.8
##  2 340030032003  2018     8 Harlem Meer              40.8     -74.0          69.5
##  3 340030032003  2018     8 Harlem Meer              40.8     -74.0          34.8
```

Visits by where people live

```
##
    |
    |                                                                        |    0%
    |
    |                                                                        |    1%
    |
    |=                                                                       |    1%
    |
    |=                                                                       |    2%
    |
    |==                                                                      |    2%
    |
    |==                                                                      |    3%
    |
    |===                                                                     |    4%
    |
    |===                                                                     |    5%
    |
    |====                                                                    |    5%
    |
    |====                                                                    |    6%
```

Travel cost estimation with cell data

We don't have the exact cost of households going to central park, but we have variables that are a good proxy

Travel cost estimation with cell data

We don't have the exact cost of households going to central park, but we have variables that are a good proxy

Estimate a simple travel cost model, what does it tell you (tip: use `feols` instead of `lm`)

Travel cost estimation with cell data

We don't have the exact cost of households going to central park, but we have variables that are a good proxy

Estimate a simple travel cost model, what does it tell you (tip: use `feols` instead of `lm`)

```
central_park_demand = feols(
  log(visits) ~ log(travel_distance_km),
  central_park_data
  ) |>
  tidy() |>
  select(term, estimate)
```

```
## NOTE: 237 observations removed because of infinite values (RHS: 237).
```

Travel cost estimation with cell data

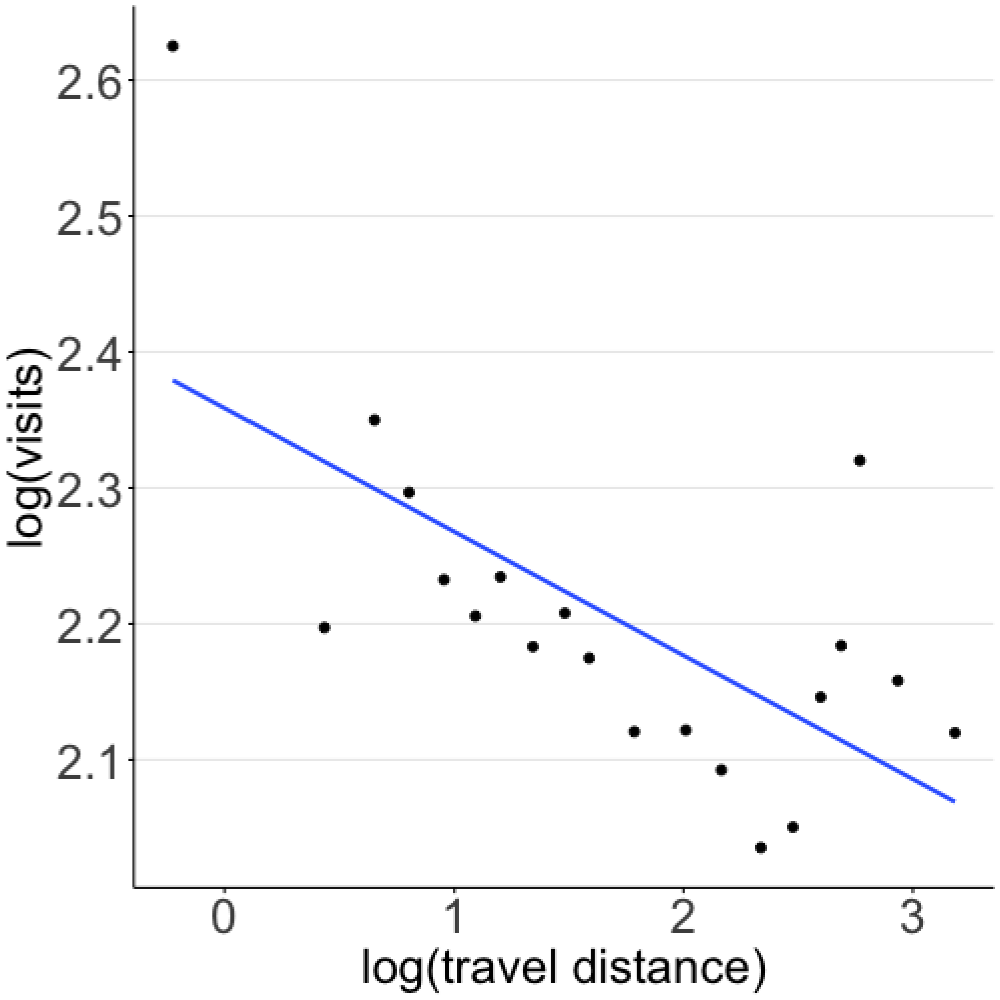
Regression: `log(visits) ~ log(travel_distance_km)`

```
central_park_demand
```

```
## # A tibble: 2 × 2
##   term                   estimate
##   <chr>                     <dbl>
## 1 (Intercept)              2.10
## 2 log(travel_distance_km)  -0.0593
```

What do the estimates mean?

Visualizing the relationship



The number of visits decreases in distance

The slope is the elasticity (-0.0593)

A 1 percent increase in distance decreases visits by 0.0593 percent

The elasticity and omitted variables

Other things probably affect how far someone lives from central park and how often they visit central park

The elasticity and omitted variables

Other things probably affect how far someone lives from central park and how often they visit central park

Ideas?

The elasticity and omitted variables

Other things probably affect how far someone lives from central park and how often they visit central park

Ideas?

New regression controlling for these factors:

```
central_park_demand = feols(
  log(visits) ~ log(travel_distance_km) + log(median_income) + log(median_age),
  central_park_data
  ) |>
  tidy() |>
  select(term, estimate)
```

```
## NOTE: 2,036 observations removed because of NA and infinite values (RHS: 2,036).
```

The elasticity and omitted variables

```
central_park_demand
```

```
## # A tibble: 4 × 2
##   term                    estimate
##   <chr>                      <dbl>
## 1 (Intercept)               0.578
## 2 log(travel_distance_km)  -0.0252 versus -0.593
## 3 log(median_income)        0.0858
## 4 log(median_age)           0.134
```

The elasticity dropped by two-thirds!

The elasticity and omitted variables

```
central_park_demand
```

```
## # A tibble: 4 × 2
##   term                        estimate
##   <chr>                          <dbl>
## 1 (Intercept)                   0.578
## 2 log(travel_distance_km)     -0.0252 versus -0.593
## 3 log(median_income)            0.0858
## 4 log(median_age)               0.134
```

The elasticity dropped by two-thirds!

Why?

The elasticity and omitted variables

Rich people go to central park more than poorer people

Older people go to central park more than younger people

Where do richer older people tend to live?

The elasticity and omitted variables

```
feols(log(travel_distance_km) ~ log(median_income), central_park_data) |> tidy()
```

```
## # A tibble: 2 × 5
##   term               estimate std.error statistic p.value
##   <chr>                 <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)           7.65    0.0942       81.2       0
## 2 log(median_income)   -0.520   0.00831     -62.6       0
```

```
feols(log(travel_distance_km) ~ log(median_age), central_park_data) |> tidy()
```

```
## # A tibble: 2 × 5
##   term            estimate std.error statistic p.value
##   <chr>              <dbl>     <dbl>     <dbl>   <dbl>
## 1 (Intercept)         5.95    0.0913      65.1       0
## 2 log(median_age)    -1.15    0.0250     -46.2       0
```

Richer and older people live closer to central park

The elasticity and omitted variables

Why does this matter?

The elasticity and omitted variables

Why does this matter?

Rich people can afford to live in Manhattan and they also like parks a lot

The elasticity and omitted variables

Why does this matter?

Rich people can afford to live in Manhattan and they also like parks a lot

Ignoring this makes it seem like the average person visits a lot less if they live further away

The elasticity and omitted variables

Why does this matter?

Rich people can afford to live in Manhattan and they also like parks a lot

Ignoring this makes it seem like the average person visits a lot less if they live further away

But it is just the fact that poorer households tend to live in the outer boroughs of New York and likely cannot afford as many trips as richer households