Conditional CEF: Linear Regression as an Approximation

Le Wang

CEF: Motivation and Running Example

Possible Solutions

Solution (1): Kernel Estimation

Solution 2: KKN Estimator

Solution (3): Parametric Linear Regression

More on Linear Regression

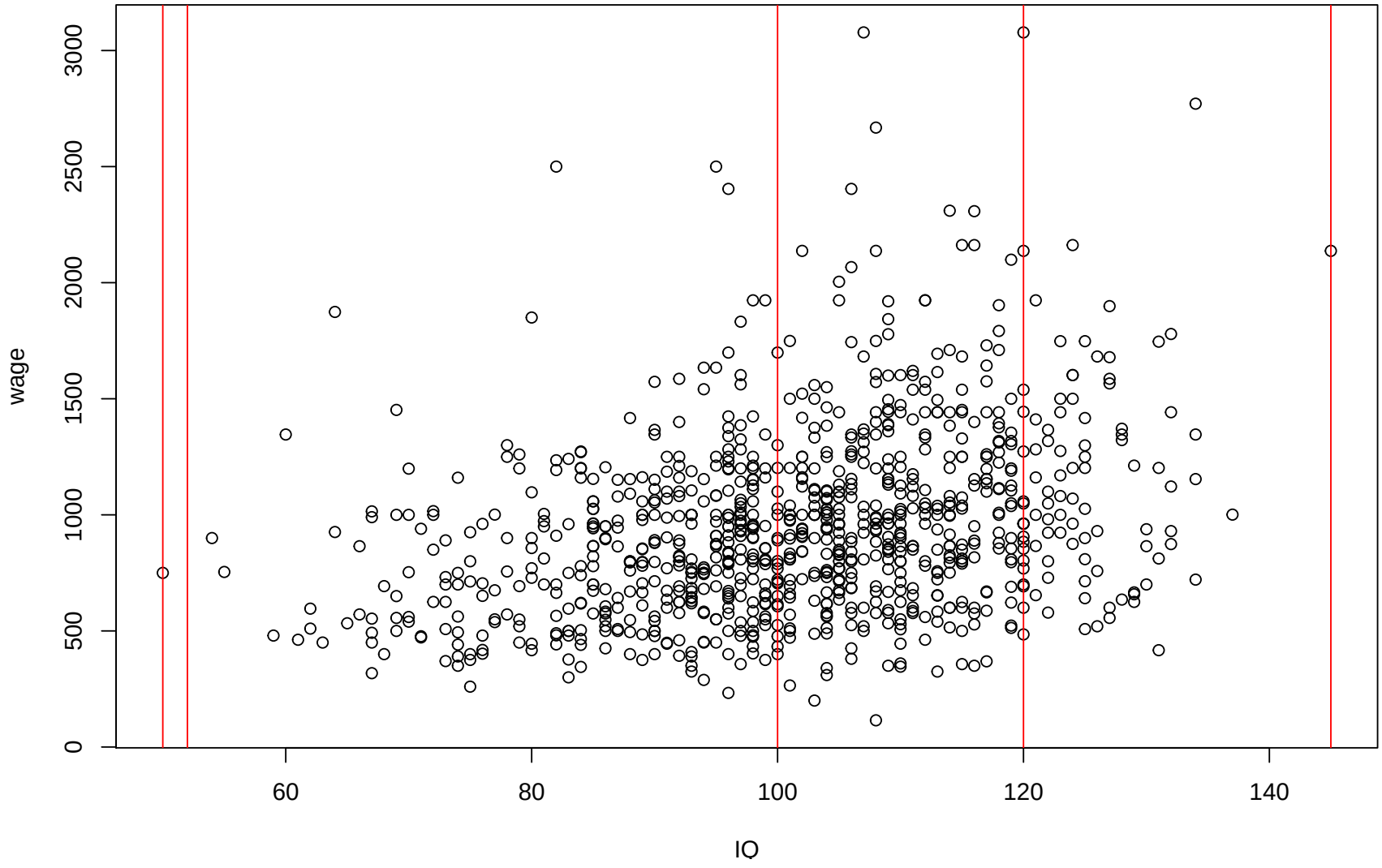
CEF: Motivation and Running Example

Motivation

Question: What if the predictor or explanatory variable, $\boldsymbol{x}$, is continuous?

Let's look at one example

Question: What is the relationship between IQ and Wages?



You already know: not that many low- or high-IQ individuals! Even if it is theoretically well defined, we cannot estimate it!

How can we obtain

$$\mathbb{E}[\text{wages} \mid \text{IQ}]$$

Of course, we can still calculate the CEF values for those that exist in the data.

```
##    Group.1         x
## 1       50  750.0000
## 2       54  900.0000
## 3       55  754.0000
## 4       59  480.0000
## 5       60 1346.0000
## 6       61  462.0000
## 7       62  553.0000
## 8       63  450.0000
## 9       64 1400.0000
## 10      65  533.0000
## 11      66  718.0000
## 12      67  636.3333
## 13      68  546.5000
## 14      69  831.6000
## 15      70  810.4000
## 16      71  629.6667
## 17      72  872.7500
## 18      73  637.5000
## 19      74  605.8750
```

Possible Solutions

Possible Solutions

There are many ways to define these solutions (or models): **Nonparametric** vs. **Parametric**

In **econometrics:** whether or not functional form is assumed

In **machine learning:** does the model have a fixed number of parameters? Or, does the number of parameters grow twith the amount of data?

Parametric models have the advantage of often being faster to use, but the disadvantage of making stronger (perhaps wrong) assumptions about the nature of the conditional distribution.

Nonparametric Models are more flexible, but often computationally intractable for large datasets, let alone big data!

1. Kernel Estimation
2. K-nearest neighbors (KNN)
3. OLS estimation (Linear Regression)

Solution (1): Kernel Estimation

Solution (1): Kernel Estimation

We now turn this quantity

$$\mathbb{E}[\text{wages}|\text{IQ} = 60]$$

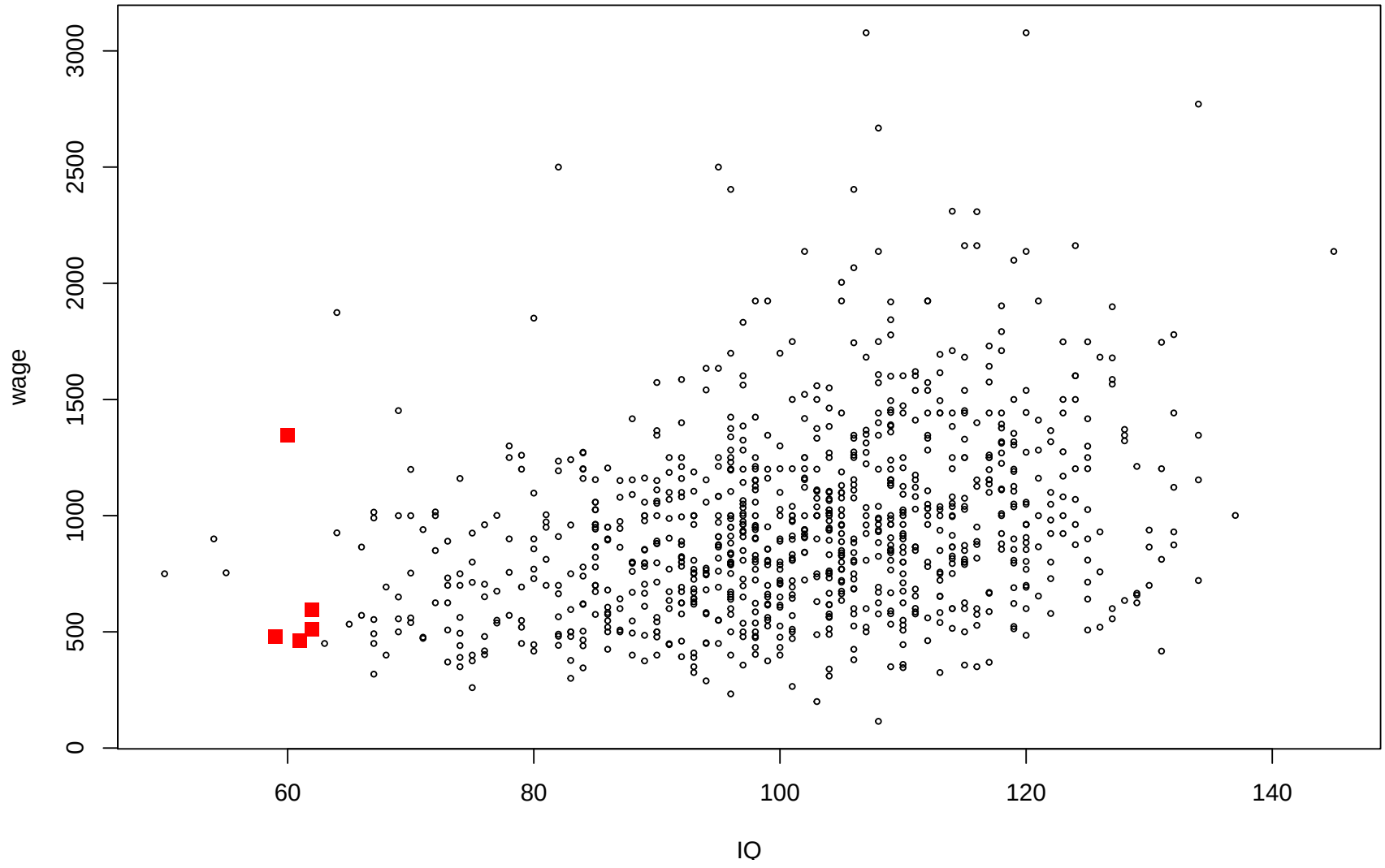
to the wages for those with similar IQ

$$\mathbb{E}[\text{wages}|\text{IQ} \approx 60]$$

Question: What do you mean by similar?

Question: What do you mean by similar?

Similarity: Perhaps within two-unit differences of one's own IQ? If so, it is pretty straightforward to implement in R.


wage
0
500
1000
1500
2000
2500
3000
60
80
100
120
140
IQ

Algorithm:

1. Choose your definition of **closeness** based on a particular distance measure, say **d**. Here we will use Euclidean distance as measure.

Algorithm:

1. Choose your definition of **closeness** based on a particular distance measure, say **d**. Here we will use Euclidean distance as measure.
2. For getting the predictions, iterate from 1 to total number of training data points

(a) Select the subsample of the original sample within the radius, **d**, of the training data point.

Algorithm:

1. Choose your definition of **closeness** based on a particular distance measure, say **d**. Here we will use Euclidean distance as measure.
2. For getting the predictions, iterate from 1 to total number of training data points

(a) Select the subsample of the original sample within the radius, **d**, of the training data point.

(b) Calculate the average of the outcome variable (continuous variable) or the conditional distribution of the outcome variable (discrete variable)

Algorithm:

1. Choose your definition of **closeness** based on a particular distance measure, say **d**. Here we will use Euclidean distance as measure.

2. For getting the predictions, iterate from 1 to total number of training data points

(a) Select the subsample of the original sample within the radius, **d**, of the training data point.

(b) Calculate the average of the outcome variable (continuous variable) or the conditional distribution of the outcome variable (discrete variable)

(c) Report the average (continuous variable) as your prediction, or the most likely outcome (discrete variable) as your prediction.

Alternative Expression of NP: Weighted Averages

Another way to express what we just did:

$$\mathbb{E}[\text{wages} \mid \text{IQ} = 50] = \mathbb{E}[\text{wages} \mid \text{IQ} \approx 50]$$

Alternative Expression of NP: Weighted Averages

Another way to express what we just did:

$$\begin{aligned}\mathbb{E}[\text{wages} \mid \text{IQ} = 50] &= \mathbb{E}[\text{wages} \mid \text{IQ} \approx 50] \\ &= \frac{\sum \mathbb{I}[50 + 2 \geq \text{IQ}_i \geq 50 - 2] \cdot \text{wages}_i}{\sum \mathbb{I}[50 + 2 \geq \text{IQ}_i \geq 50 - 2]}\end{aligned}$$

Alternative Expression of NP: Weighted Averages

Another way to express what we just did:

$$\begin{aligned}\mathbb{E}[\text{wages} \mid \text{IQ} = 50] &= \mathbb{E}[\text{wages} \mid \text{IQ} \approx 50] \\ &= \frac{\sum \mathbb{I}[50 + 2 \geq \text{IQ}_i \geq 50 - 2] \cdot \text{wages}_i}{\sum \mathbb{I}[50 + 2 \geq \text{IQ}_i \geq 50 - 2]} \\ &= \sum w_i \cdot \text{wages}\end{aligned}$$

where $w_i = \frac{\mathbb{I}[50+2 \geq \text{IQ}_i \geq 50-2]}{\sum \mathbb{I}[50+2 \geq \text{IQ}_i \geq 50-2]}$

Example

IQ	*wages*	$\mathbb{I}[50+2 \geq \mathsf{IQ}_i \geq 50-2]$
51	100	1
49	50	1
53	110	0
53	115	0
44	45	0

Approach 1: $\frac{100+50}{2} = 75$

Example

IQ	*wages*	$\mathbb{I}[50+2 \geq \mathsf{IQ}_i \geq 50-2]$
51	100	1
49	50	1
53	110	0
53	115	0
44	45	0

Approach 1: $\frac{100+50}{2} = 75$

Approach 2:

$$\frac{1 \cdot 100 + 1 \cdot 50 + 0 \cdot 110 + 0 \cdot 115 + 0 \cdot 45}{1 + 1 + 0 + 0 + 0}$$

Example

IQ	*wages*	$\mathbb{I}[50+2 \geq \mathsf{IQ}_i \geq 50-2]$
51	100	1
49	50	1
53	110	0
53	115	0
44	45	0

Approach 1: $\frac{100+50}{2} = 75$

Approach 2:

$$\frac{1 \cdot 100 + 1 \cdot 50 + 0 \cdot 110 + 0 \cdot 115 + 0 \cdot 45}{1+1+0+0+0}$$

$$= \frac{1}{2} \cdot 100 + \frac{1}{2} \cdot 50 + \frac{0}{2} \cdot 110 + \frac{0}{2} \cdot 115 + \frac{0}{2} \cdot 45 = 75$$

Limitations of Solution 1

Our proposed solution: discretization (as we have seen in one of the previous examples)Shortcoming:

1. Arbitrary bundling of the points.
2. Use of only limited points (sometimes may not be any observations)!

Question: What to do?

Solution 1 (Extension): Nonparametric Kernel Estimation:

$$\mathbb{E}[\text{wages} \mid \text{IQ} = 50] = \mathbb{E}[\text{wages} \mid \text{IQ} \approx 50]$$

Solution 1 (Extension): Nonparametric Kernel Estimation:

$$\begin{aligned}\mathbb{E}[\text{wages} \mid \text{IQ} = 50] &= \mathbb{E}[\text{wages} \mid \text{IQ} \approx 50] \\ &= \frac{\sum \mathbb{K}[50, IQ_i] \cdot \text{wages}_i}{\sum \mathbb{K}[50, \text{IQ}_i]}\end{aligned}$$

Solution 1 (Extension): Nonparametric Kernel Estimation:

$$\begin{aligned}\mathbb{E}[\text{wages} \mid \text{IQ} = 50] &= \mathbb{E}[\text{wages} \mid \text{IQ} \approx 50] \\ &= \frac{\sum \mathbb{K}[50, IQ_i] \cdot \text{wages}_i}{\sum \mathbb{K}[50, \text{IQ}_i]} \\ &= \sum w_i \cdot \text{wages}\end{aligned}$$

where $w_i = \frac{\mathbb{K}[50,\text{IQ}_i]}{\sum \mathbb{K}[50,\text{IQ}_i]}$

Find a kernel function that assigns declining weights with respect to the distance to the data point, x.

$$\mathbb{K}[50, \mathrm{IQ}_i] = \mathbb{K}[\frac{IQ_i - 50}{h}]$$

where h is called the **bandwidth** (you can think of it as window)

Examples of Kernel

1. **Gaussian** or **Normal**

$$\mathbb{K}(z) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{u^2}{2})$$

2. **Uniform**

$$\mathbb{K}(z) = \frac{1}{2}\mathbb{I}(|u| \leq 1)$$

3. **Epanechnikov**

$$\mathbb{K}(z) = \frac{3}{4}(1 - u^2)\mathbb{I}(|u| \leq 1)$$

Solution 2: KKN Estimator

Solution 2: KKN Estimator

The book also discusses an alternative approach, which, to some extent, solve the issue where no observations exist for the discretization approach above:

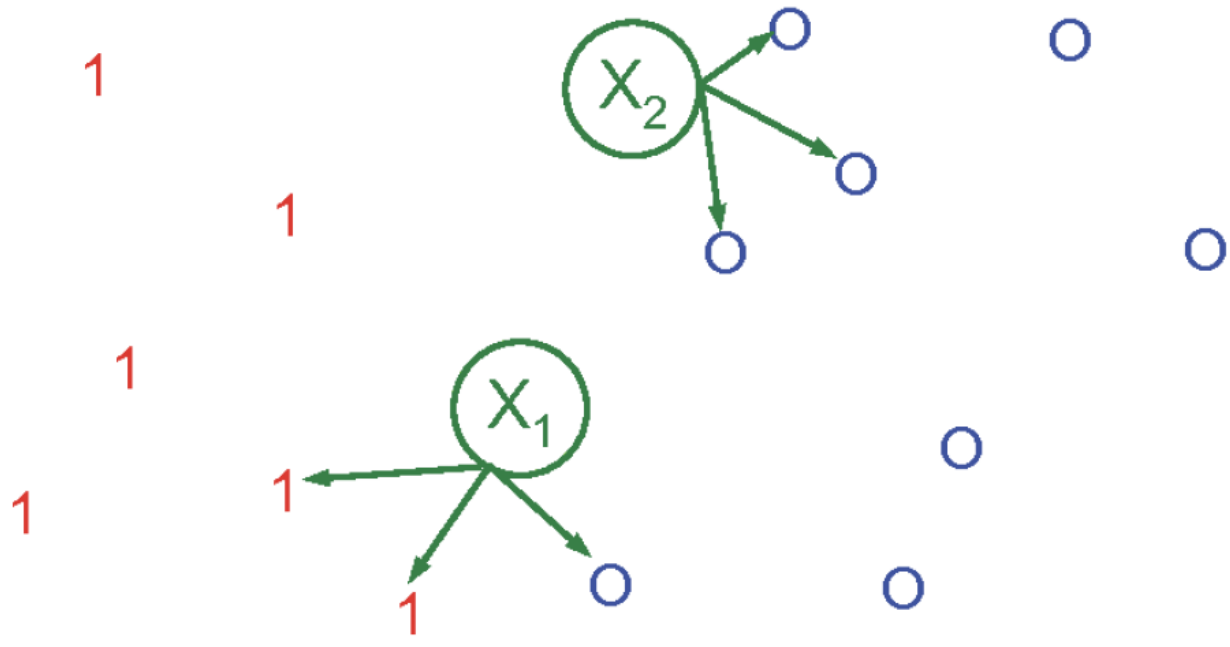
K-nearest Neighbors

As the name suggests, you use only K nearest neighbors around the value, x, to calculate the mean and use it to obtain the prediction.

This method is an example of memory-based learning or instance-based learning.

The 3 nearest neighbors of point x_{1}have labels 1,1,0, so we

$$\mathbb{E}[y|x = x_1] = \frac{1+1+0}{3} = \frac{2}{3}$$



Algorithm

1. Calculate the distance between your value of x and all x in your data, sort the data, and pick the first k observations

Algorithm

1. Calculate the distance between your value of x and all x in your data, sort the data, and pick the first k observations
2. For getting the predictions, iterate from 1 to total number of training data points

(a) Report the average as your prediction for these k observations, or the most likely outcome as your prediction.

Alternative Expression of KKN: Weighted Averages VersionAnother way to express what we just did:

$$\begin{aligned}\mathbb{E}[\text{wages}|\text{IQ} = 50] &= \mathbb{E}[\text{wages}|\text{IQ} \in D] \\ &= \frac{\sum \mathbb{I}[\text{IQ} \in D] \cdot \text{wages}_i}{\sum \mathbb{I}[\text{IQ} \in D]} \\ &= \sum w_i \cdot \text{wages}\end{aligned}$$

IQ	*wages*	$d = IQ - 50$	$\mathbb{I}[D]\cdot$
51	100	1	1
49	50	1	1
53	110	3	0
53	115	3	0
44	45	6	0

Approach 1: $\frac{100+50}{2} = 75$

IQ	*wages*	$d = IQ - 50$	$\mathbb{I}[D]\cdot$
51	100	1	1
49	50	1	1
53	110	3	0
53	115	3	0
44	45	6	0

Approach 1: $\frac{100+50}{2} = 75$

Approach 2:

$$\frac{1 \cdot 100 + 1 \cdot 50 + 0 \cdot 110 + 0 \cdot 115 + 0 \cdot 45}{1 + 1 + 0 + 0 + 0}$$
$$= \frac{1}{2} \cdot 100 + \frac{1}{2} \cdot 50 + \frac{0}{2} \cdot 110 + \frac{0}{2} \cdot 115 + \frac{0}{2} \cdot 45 = 75$$

Solution (3): Parametric Linear Regression

Solution (3): Parametric Linear Regression

Parametric Model (Only two parameters!): No matter what x values, I will estimate the following model

$$\text{wage} = \beta_0 + \beta_1 \cdot \text{IQ} + \epsilon$$

My prediction from the previous model

$$\text{wage} = 116.992 + 8.303 \cdot \text{IQ}$$

One time for all values!

$$\text{IQ} = 50 \implies \mathbb{E}[\text{wage}|\text{IQ} = 50] = 116.992 + 8.303 \cdot 50$$
$$\text{IQ} = 60 \implies \mathbb{E}[\text{wage}|\text{IQ} = 60] = 116.992 + 8.303 \cdot 60$$

Alternative Expression of OLS: Weighted Averages

Remember that it is an estimator of the underlying conditional mean. We are somehow weighting some of the observations to obtain an estimate of the conditional mean. But what does this weighting function look like?

$$\mathbb{E}[\text{wage}|\text{IQ} = 50] = \beta_0 + \beta_1 \cdot 50$$

$$\beta_0 = \bar{y} - \beta_1 \cdot \bar{x}$$
$$\beta_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2}$$

Alternative Expression of OLS: Weighted Averages

Let me show you one results before proceeding

$$
\begin{aligned}
\beta_1 &= \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} - \frac{\sum(x_i - \bar{x})\bar{y}}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} - \frac{\bar{y} \cdot \sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} - \frac{\bar{y} \cdot N \cdot \frac{1}{N}\sum(x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} - \frac{\bar{y} \cdot N \cdot \left[\frac{1}{N}\sum x_i - \frac{1}{N}\sum \bar{x}\right]}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} - \frac{\bar{y} \cdot N \cdot \left[\bar{x} - \frac{1}{N} \cdot N \cdot \bar{x}\right]}{\sum(x_i - \bar{x})^2} \\
&= \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2}
\end{aligned}
$$

$$
\begin{aligned}
\beta_0 + \beta_1 \cdot 50 &= (\bar{y} - \beta_1 \cdot \bar{x}) + \beta_1 \cdot 50 \\
&= \bar{y} + \beta_1(50 - \bar{x}) \\
&= \bar{y} + \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} \cdot (50 - \bar{x}) \\
&= \frac{1}{N}\sum y_i + \frac{\sum(x_i - \bar{x})y_i}{\sum(x_i - \bar{x})^2} \cdot (50 - \bar{x}) \\
&= \sum \frac{1}{N} \cdot y_i + \sum \frac{(50 - \bar{x}) \cdot (x_i - \bar{x})}{\sum(x_i - \bar{x})^2} y_i \\
&= \sum \left[\frac{1}{N} + \frac{(50 - \bar{x}) \cdot (x_i - \bar{x})}{\sum(x_i - \bar{x})^2} \right] y_i \\
&= \sum w_i \cdot y_i
\end{aligned}
$$

More on Linear Regression

Linear Regression: Prediction

Now, we know the underlying mechanism behind how we calculate the conditional mean. We turn to a more convenient way to calculate our conditional mean function and the prediction.

Suppose that the model that you run is given by

$$y = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k + \epsilon$$

The implied conditional mean is given by

$$\mathbb{E}[y|x] = \beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

You can just substitute in your x values to this equation.

Linear Regression: Partial Effects

Remember that the second use of our conditional mean function is to figure out what the partial effects of a variable is

$$\mathbb{E}[\text{wage}|\text{IQ} = x_1] - \mathbb{E}[\text{wage}|\text{IQ} = x_0]$$

Examples:

$$\begin{aligned}
&\mathbb{E}[\text{wage}|\text{IQ} = 51] - \mathbb{E}[\text{wage}|\text{IQ} = 50] \\
&= (\beta_0 + \beta_1 \cdot 51) - (\beta_0 + \beta_1 \cdot 50) \\
&= \beta_1
\end{aligned}$$

Linear Regression: Partial Effects

Remember that the second use of our conditional mean function is to figure out what the partial effects of a variable is

$$\mathbb{E}[\text{wage}|\text{IQ} = x_1] - \mathbb{E}[\text{wage}|\text{IQ} = x_0]$$

Examples:

$$\begin{aligned}
&\mathbb{E}[\text{wage}|\text{IQ} = 51] - \mathbb{E}[\text{wage}|\text{IQ} = 50] \\
&= (\beta_0 + \beta_1 \cdot 51) - (\beta_0 + \beta_1 \cdot 50) \\
&= \beta_1
\end{aligned}$$

$$\begin{aligned}
&\mathbb{E}[\text{wage}|\text{IQ} = 52] - \mathbb{E}[\text{wage}|\text{IQ} = 51] \\
&= (\beta_0 + \beta_1 \cdot 52) - (\beta_0 + \beta_1 \cdot 51) \\
&= \beta_1
\end{aligned}$$

Partial Effects: More General Form

$$\mathbb{E}[y|x_1, x_2, x_3 \ldots, x_k] = m(x) = m(x_1, x_2, \ldots, x_k)$$

Partial Effects:

$$\frac{\partial}{\partial x_1}\mathbb{E}[y|x_1, x_2, x_3 \ldots, x_k] = \frac{\partial}{\partial x_1} m(x)$$

Interpretation:

When x_1 increases by one unit, by how much will y (or the averages of y) will increase?

Partial Effects: More General Form

$$\mathbb{E}[y|x_1, x_2, x_3 \ldots, x_k] = m(x) = m(x_1, x_2, \ldots, x_k)$$

Partial Effects:

$$\frac{\partial}{\partial x_1}\mathbb{E}[y|x_1, x_2, x_3 \ldots, x_k] = \frac{\partial}{\partial x_1} m(x)$$

Interpretation:

When x_1 increases by one unit, by how much will y (or the averages of y) will increase?

1. Linear Models

$$\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k$$

Partial Effects:

$$\frac{\partial}{\partial x_k}[\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \cdots + \beta_k \cdot x_k] = \beta_k$$

Example

$$\begin{aligned}&\mathbb{E}[\text{wage}|\text{married},\text{black},\text{IQ}]\\&= 71.734355 + 177.020344 \cdot \text{married} + 7.338611 \cdot \text{IQ} - 118.030901 \cdot \text{black}\end{aligned}$$

1. One-unit increase in IQ will lead to

Example

$$\begin{aligned}&\mathbb{E}[\text{wage}|\text{married},\text{black},\text{IQ}]\\&= 71.734355 + 177.020344 \cdot \text{married} + 7.338611 \cdot \text{IQ} - 118.030901 \cdot \text{black}\end{aligned}$$

1. One-unit increase in IQ will lead to

$$\begin{aligned}&\frac{\partial}{\partial \text{IQ}} [71.734355 + 177.020344 \cdot \text{married} + 7.338611 \cdot \text{IQ} - 118.030901 \cdot \text{black}]\\&= 7.338611\end{aligned}$$

Example

$$\begin{aligned}&\mathbb{E}[\text{wage}|\text{married},\text{black},\text{IQ}]\\&= 71.734355 + 177.020344 \cdot \text{married} + 7.338611 \cdot \text{IQ} - 118.030901 \cdot \text{black}\end{aligned}$$

1. One-unit increase in IQ will lead to

$$\begin{aligned}&\frac{\partial}{\partial \text{IQ}} [71.734355 + 177.020344 \cdot \text{married} + 7.338611 \cdot \text{IQ} - 118.030901 \cdot \text{black}]\\&= 7.338611\end{aligned}$$

2. One unit increase in marital status (change from 0 to 1): (it is not really derivative, but for consistency)

Example

$$\begin{aligned}&\mathbb{E}[\text{wage}|\text{married},\text{black},\text{IQ}]\\&= 71.734355 + 177.020344 \cdot \text{married} + 7.338611 \cdot \text{IQ} - 118.030901 \cdot \text{black}\end{aligned}$$

1. One-unit increase in IQ will lead to

$$\begin{aligned}&\frac{\partial}{\partial \text{IQ}} [71.734355 + 177.020344 \cdot \text{married} + 7.338611 \cdot \text{IQ} - 118.030901 \cdot \text{black}]\\&= 7.338611\end{aligned}$$

2. One unit increase in marital status (change from 0 to 1): (it is not really derivative, but for consistency)

$$\begin{aligned}&\frac{\partial}{\partial \text{married}} [71.734355 + 177.020344 \cdot 1 + 7.338611 \cdot \text{IQ} - 118.030901 \cdot \text{black}]\\&= 177.020344\end{aligned}$$

Again, it does not depend on any other things such as x. It is a constant!

Why do we still use OLS?

We know that the linear regression or the OLS approach gives us

1. A weird form of weighted averages
2. A constant partial effect

Question: Why is it still useful?!

Answer: It is the best linear approximation!

$$\min \mathbb{E}\left[(m(x) - x'\beta)^2\right] \implies \beta = \mathbb{E}[xx']^{-1}\mathbb{E}[x'y]$$

You can think of $m(x)$ as your dependent variable

If your problem is to minimize the following

$$\mathbb{E}[(y - x'\beta)^2]$$

Your solution is

$$\mathbb{E}[xx']^{-1}\mathbb{E}[xy]$$

You can think of $m(x)$ as your dependent variable

If your problem is to minimize the following

$$\mathbb{E}[(y - x'\beta)^2]$$

Your solution is

$$\mathbb{E}[xx']^{-1}\mathbb{E}[xy]$$

When your problem is to minimize the following

$$\mathbb{E}[(m(x) - x'\beta)^2]$$

You can think of $m(x)$ as your dependent variable

If your problem is to minimize the following

$$\mathbb{E}[(y - x'\beta)^2]$$

Your solution is

$$\mathbb{E}[xx']^{-1}\mathbb{E}[xy]$$

When your problem is to minimize the following

$$\mathbb{E}[(m(x) - x'\beta)^2]$$

Your solution is

$$\mathbb{E}[xx']^{-1}\mathbb{E}[xm(x)]$$

We can show that

$$\mathbb{E}[xx']^{-1}\mathbb{E}[xm(x)] = \mathbb{E}[xx']^{-1}\mathbb{E}[xy]$$

In your homework, you have already showed a simpler version of this

$$cov(x, y) = cov(x, \mathbb{E}[y \mid x])$$

$$\mathbb{E}[xx']^{-1}\mathbb{E}[xm(x)] = \mathbb{E}[xx']^{-1}\mathbb{E}[x\mathbb{E}[y \mid x]]$$

$$\begin{aligned}\mathbb{E}[xx']^{-1}\mathbb{E}[xm(x)] &= \mathbb{E}[xx']^{-1}\mathbb{E}[x\mathbb{E}[y \mid x]] \\ &= \mathbb{E}[xx']^{-1}\mathbb{E}[\mathbb{E}[xy \mid x]\end{aligned}$$

$$\begin{aligned}\mathbb{E}[xx']^{-1}\mathbb{E}[xm(x)] &= \mathbb{E}[xx']^{-1}\mathbb{E}[x\mathbb{E}[y \mid x]] \\ &= \mathbb{E}[xx']^{-1}\mathbb{E}[\mathbb{E}[xy \mid x] \\ &= \mathbb{E}[xx']^{-1}\mathbb{E}[xy]\end{aligned}$$

This actually suggests a group-level regression when micro data do not exist.

Alternative Proof:

$\beta^{\text{linear regression}}$ solves the following problem

$$\mathbb{E}[(y - x'\beta)^2]$$

Alternative Proof:

$\beta^{\text{linear regression}}$ solves the following problem

$$\mathbb{E}[(y - x'\beta)^2]$$

We can also show that this problem is equivalent to

$$\mathbb{E}[(m(x) - x'\beta)^2]$$

Why?

$$\mathbb{E}[(y - x'\beta)^2] = \mathbb{E}[(y - m(x) + m(x) - x'\beta)^2]$$

Why?

$$\begin{aligned}\mathbb{E}[(y - x'\beta)^2] &= \mathbb{E}[(y - m(x) + m(x) - x'\beta)^2] \\ &= \mathbb{E}[(y - m(x))^2] - 2\mathbb{E}[(y - m(x))(y - x'\beta)] \\ &\quad + \mathbb{E}[(m(x) - x'\beta)^2]\end{aligned}$$

Why?

$$\begin{aligned}
\mathbb{E}[(y - x'\beta)^2] &= \mathbb{E}[(y - m(x) + m(x) - x'\beta)^2] \\
&= \mathbb{E}[(y - m(x))^2] - 2\mathbb{E}[(y - m(x))(y - x'\beta)] \\
&+ \mathbb{E}[(m(x) - x'\beta)^2] \\
&= \mathbb{E}[(y - m(x))^2] - 2\mathbb{E}[\epsilon(y - x'\beta)] \\
&+ \mathbb{E}[(m(x) - x'\beta)^2]
\end{aligned}$$

1. $\mathbb{E}[(y - m(x))^2]$ does not involve β
2. equal to zero

Hence, the problems are exactly the same, and so are the solutions!

Improvements of OLS

The fatct that it is the best linear approximation does NOT mean that we can further improve the approximation using the linear form of nonlinear equations

1. Interaction Terms

(a) Interaction terms between the same variables: Polynomial terms (partial effects depend on itself)
(b) Interaction terms bretween two different variables: Partial effects depend on the other variable

1. Quadratic Models:

$$\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2 + \cdots + \beta_k \cdot x_k$$

Partial Effects:

$$\frac{\partial}{\partial x_1}[\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2 + \cdots + \beta_k \cdot x_k] = \beta_1 + 2 \cdot \beta_2 \cdot x_1$$

1. Quadratic Models:

$$\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2 + \cdots + \beta_k \cdot x_k$$

Partial Effects:

$$\frac{\partial}{\partial x_1}[\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2 + \cdots + \beta_k \cdot x_k] = \beta_1 + 2 \cdot \beta_2 \cdot x_1$$

2. Interaction Models

$$\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2 + \ldots$$

Partial Effects:

$$\frac{\partial}{\partial x_1}[\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2 + \ldots] = \beta_1 + \beta_3 \cdot x_2$$

1. Quadratic Models:

$$\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2 + \cdots + \beta_k \cdot x_k$$

Partial Effects:

$$\frac{\partial}{\partial x_1}[\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_1^2 + \cdots + \beta_k \cdot x_k] = \beta_1 + 2 \cdot \beta_2 \cdot x_1$$

2. Interaction Models

$$\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2 + \ldots$$

Partial Effects:

$$\frac{\partial}{\partial x_1}[\beta_0 + \beta_1 \cdot x_1 + \beta_2 \cdot x_2 + \beta_3 \cdot x_1 \cdot x_2 + \ldots] = \beta_1 + \beta_3 \cdot x_2$$

3. Infinite number of possible models

Model Selection

Two questions naturally follow:

1. Can I test the model specification?
2. Which model is preferred? Model selection criteria.