Gov 2000: 7. What is Regression?

Matthew Blackwell

Fall 2016

1. Relationships between Variables
2. Conditional Expectation
3. Estimating the CEF
4. Linear CEFs and Linear Projections
5. Least Squares

Where are we? Where are we going?

- What we've been up to: estimating parameters of population distributions. Generally we've been learning about a single variable.
- This week and for the rest of the term, we'll be interested in the relationships between variables. How does one variable change we change the values of another variable? These will be the bread and butter of the class moving forward.

AJR data

- How do we draw this line?

1/ Relationships between Variables

What is a relationship and why do we care?

- Most of what we want to do in the social science is learn about how two variables are related
- Examples:
  - Does turnout vary by types of mailers received?
  - Is the quality of political institutions related to average incomes?
  - Does conflict mediation help reduce civil conflict?

Notation and conventions

- Y_i - the dependent variable or outcome or regressand or left-hand-side variable or response
  - Voter turnout
  - Log GDP per capita
  - Number of battle deaths
- X_i - the independent variable or explanatory variable or regressor or right-hand-side variable or treatment or predictor
  - Social pressure mailer versus Civic Duty Mailer
  - Average Expropriation Risk
  - Presence of conflict mediation

Joint distribution review

- (Y_i, X_i) are draws from an i.i.d. joint distribution $f_{Y,X}$
  - Y_i and X_i are measured on the same unit i
  - WARNING different than our use of Y_i and X_i as r.v.s for different groups.
  - There, Y_i and X_i corresponded to different units.
- Several ways to summarize the joint population distribution:
  - Covariance/correlation
  - Conditional expectation
- Today we'll spend a lot of time thinking about the relevant populations parameters for estimating relationships.
  - Population-first approach.

2/ Conditional Expectation

Conditional expectation function

- Conditional expectation function (CEF): how the mean of Y_i changes as X_i changes.

$$\mu(x) = \mathbb{E}[Y_i|X_i = x]$$

- The CEF is a feature of the joint distribution of Y_i and X_i:

$$\mathbb{E}[Y_i|X_i = x] = \int_{-\infty}^{\infty} y f_{Y|X}(y|x)dy$$

- Goal of regression is to estimate CEF: $\widehat{\mu}(x) = \widehat{\mathbb{E}}[Y_i|X_i = x]$

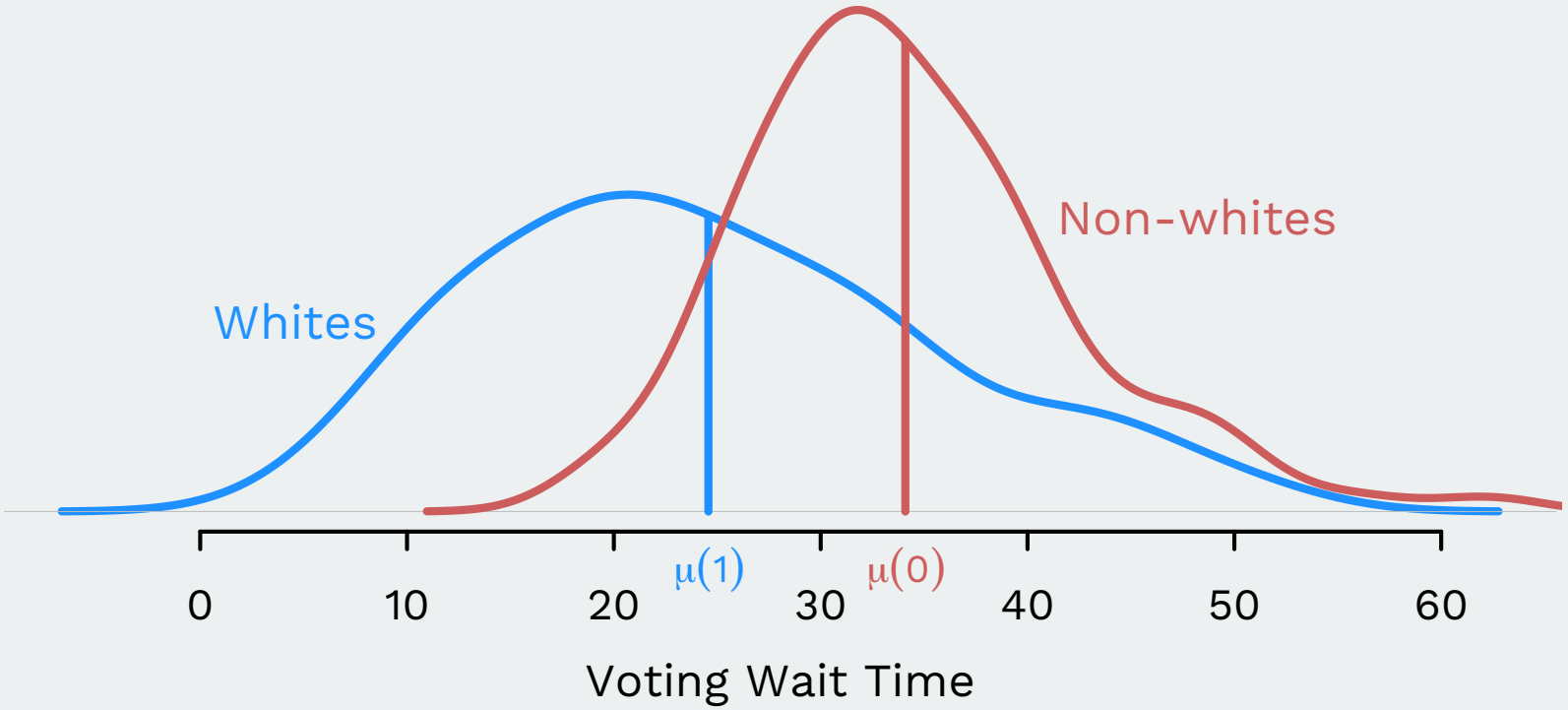
CEF for binary covariates

- Example:
  - Y_i is the time respondent i waited in line to vote.
  - $X_i = 1$ for whites, $X_i = 0$ for non-whites.
- Then the mean in each group is just a conditional expectation:

$$\mu(\text{white}) = E[Y_i|X_i = \text{white}]$$
$$\mu(\text{non-white}) = E[Y_i|X_i = \text{non-white}]$$

- Notice here that since X_i can only take on two values, 0 and 1, then these two conditional means completely summarize the CEF.

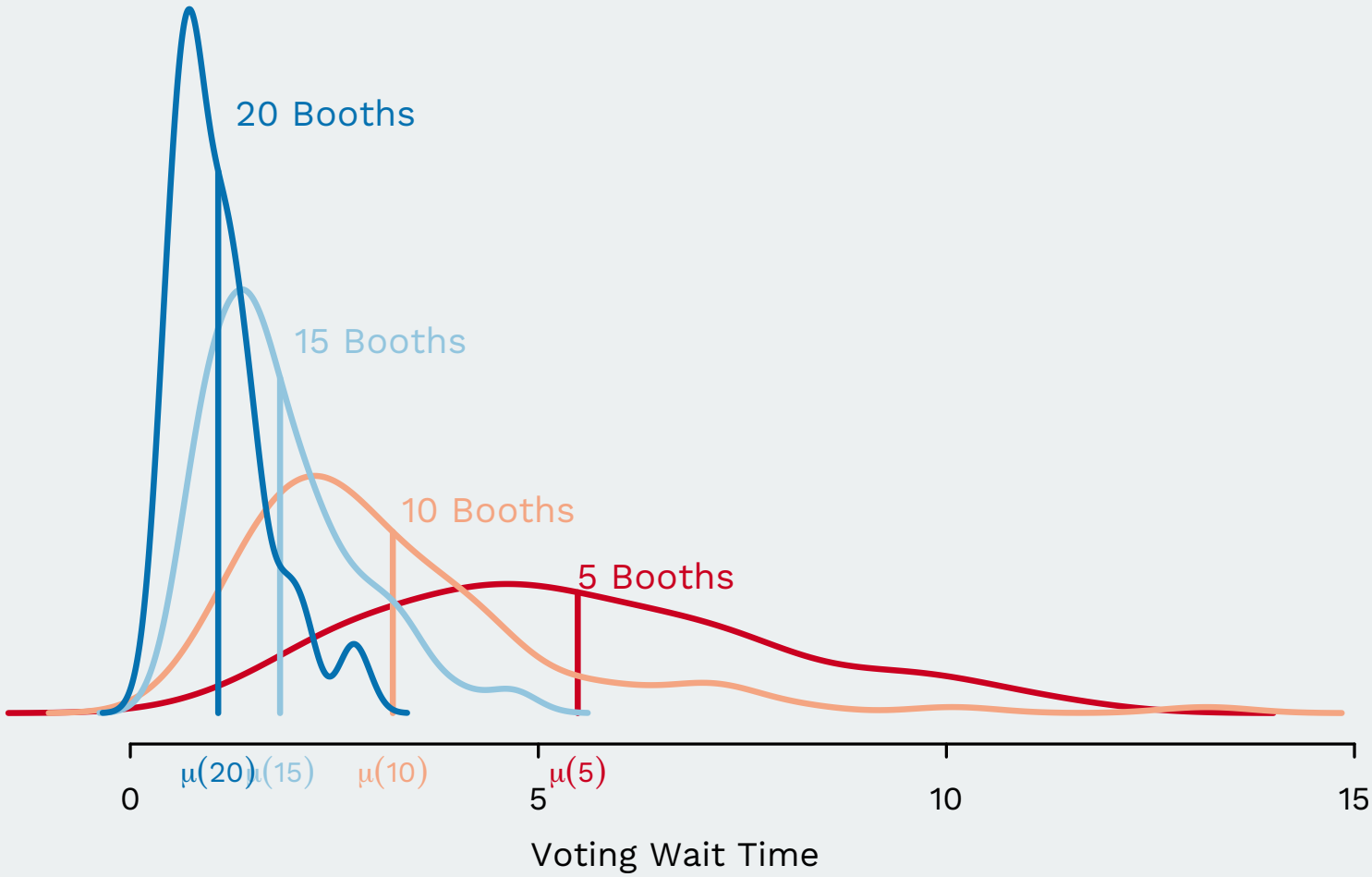
Why is the CEF useful?



- The CEF encodes relationships between variables.
- If $\mu(\text{white}) < \mu(\text{non-white})$, so that waiting times for whites are shorter on average than for non-whites.
- Indicates a relationship in the population between race and wait times.

CEF for discrete covariates

- New covariate: X_i is the number of polling booths at citizen i's polling station.
- The mean of Y_i changes as X_i changes:



CEF with multiple covariates

- We could also be interested in the CEF conditioning on multiple variables:

$$\begin{aligned}
\mu(\text{white, man}) &= \mathbb{E}[Y_i | X_i = \text{white}, Z_i = \text{man}] \\
\mu(\text{white, woman}) &= \mathbb{E}[Y_i | X_i = \text{white}, Z_i = \text{woman}] \\
\mu(\text{non-white, man}) &= \mathbb{E}[Y_i | X_i = \text{non-white}, Z_i = \text{man}] \\
\mu(\text{non-white, woman}) &= \mathbb{E}[Y_i | X_i = \text{non-white}, Z_i = \text{woman}]
\end{aligned}$$

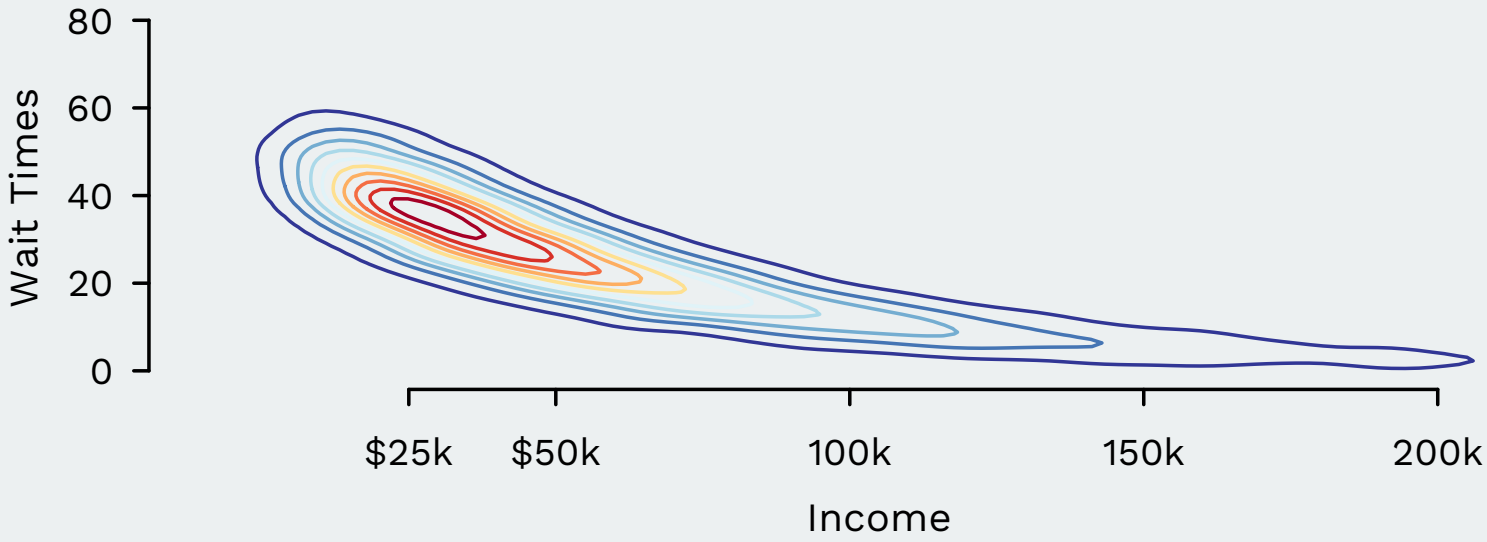
- Why? Allows more credible all else equal comparisons (ceteris paribus).
- Ex: average difference in wait times between white and non-white citizens of the same gender:

$$\mu(\text{white, man}) - \mu(\text{non-white, man})$$

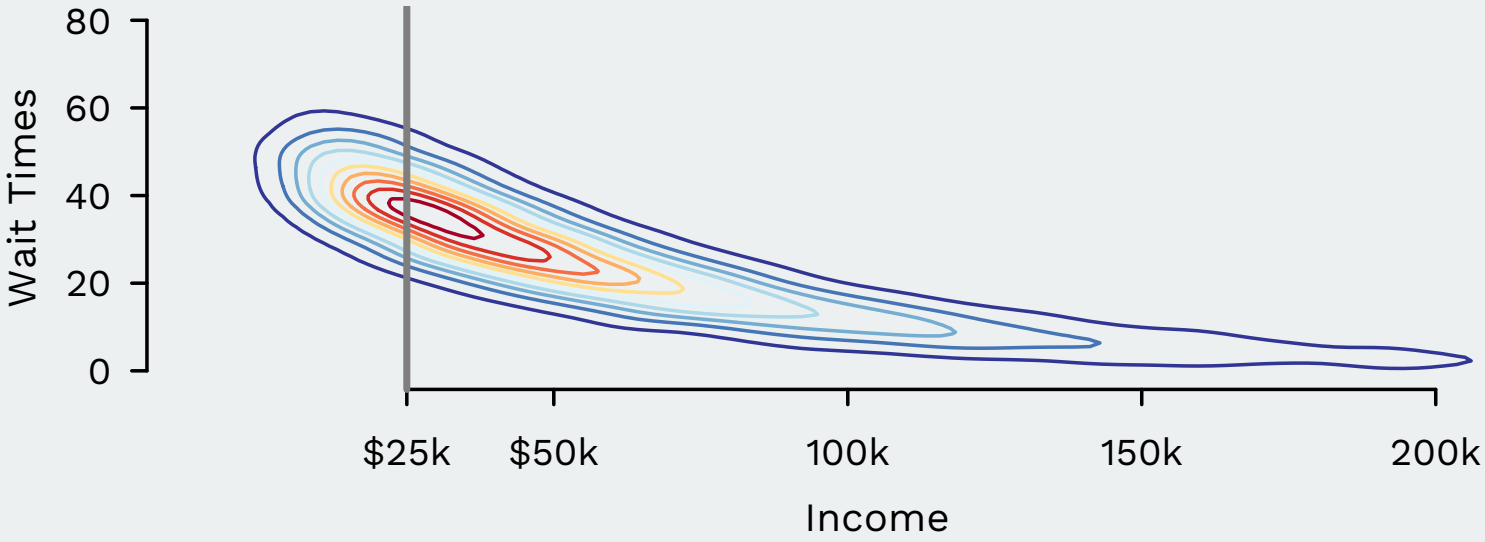
CEF for continuous covariates

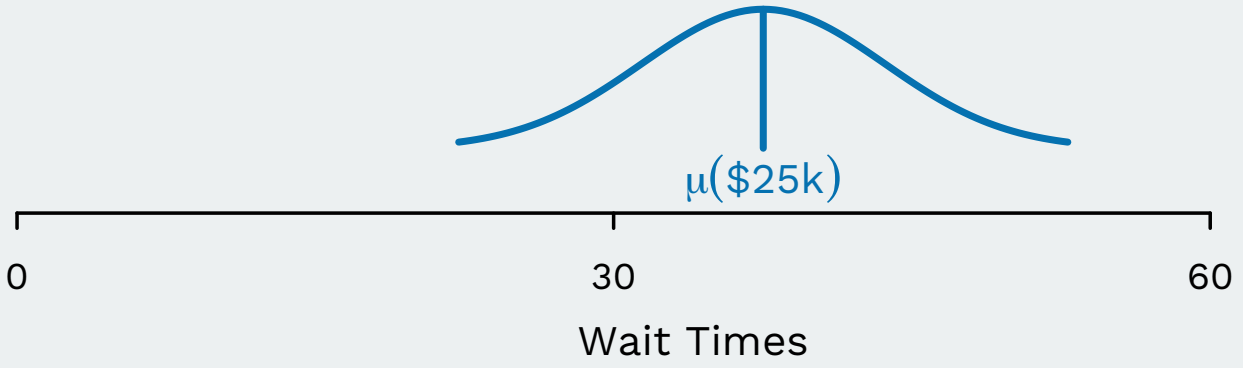
- What if our independent variable, X_i is income?
- Many possible values of $X_i \rightsquigarrow$ many possible values of $\mathbb{E}[Y_i|X_i = x]$.
  - Writing out each value of the CEF no longer feasible.
- Now we will think about $\mu(x) = \mathbb{E}[Y_i|X_i = x]$ as function. What does this function look like:
  - Linear: $\mu(x) = \alpha + \beta x$
  - Quadratic: $\mu(x) = \alpha + \beta x + \gamma x^2$
  - Crazy, nonlinear: $\mu(x) = \alpha/(\beta + x)$
- These are unknown functions in the population! This is going to make producing an estimator $\widehat{\mu}(x)$ very difficult!

Wait times and income

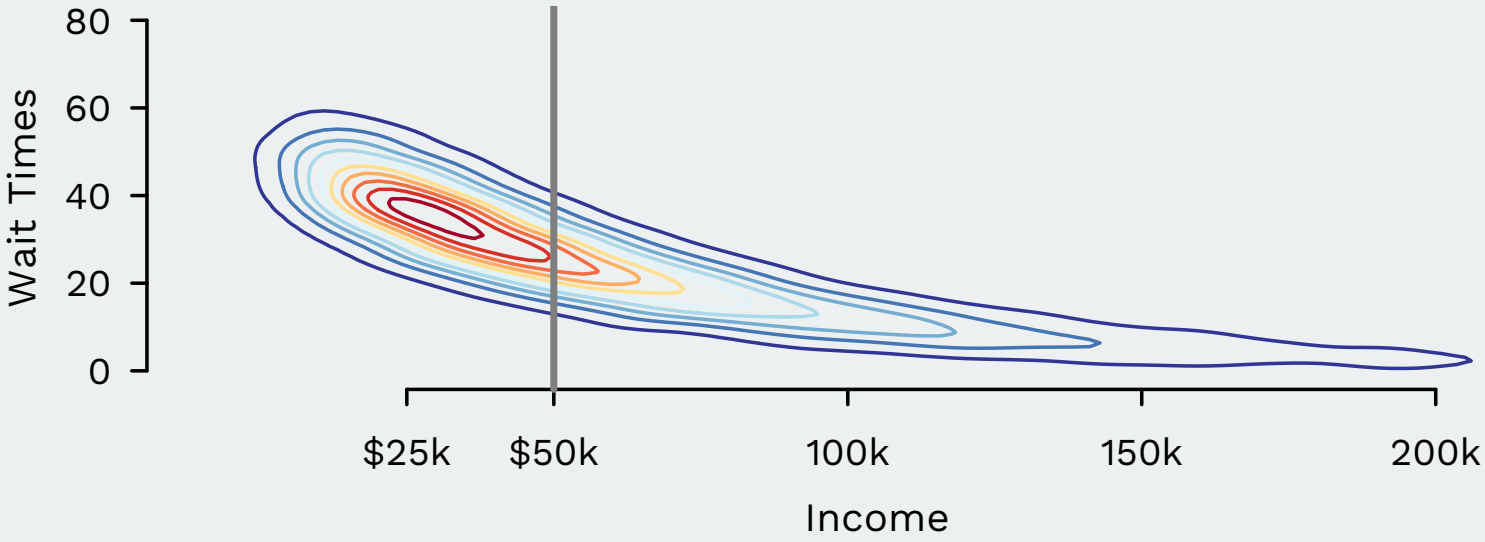


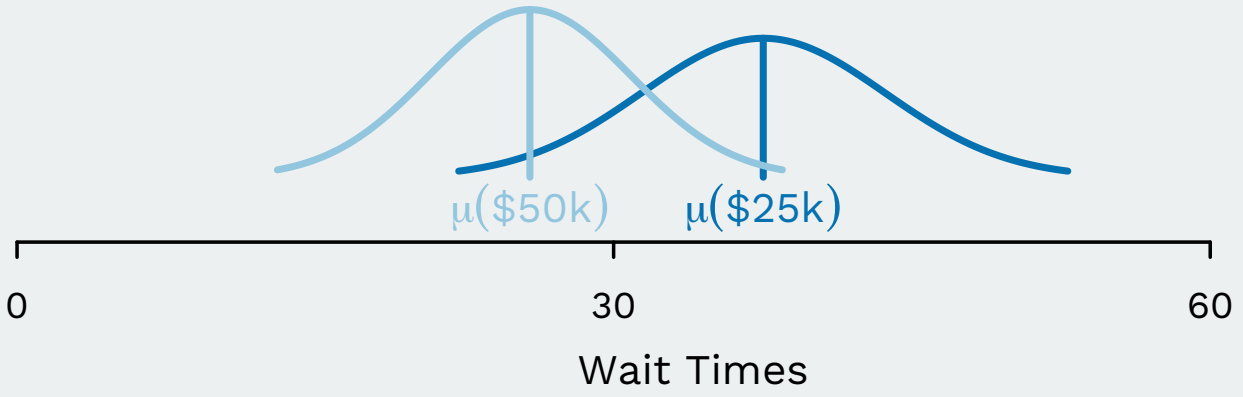
Wait times and income



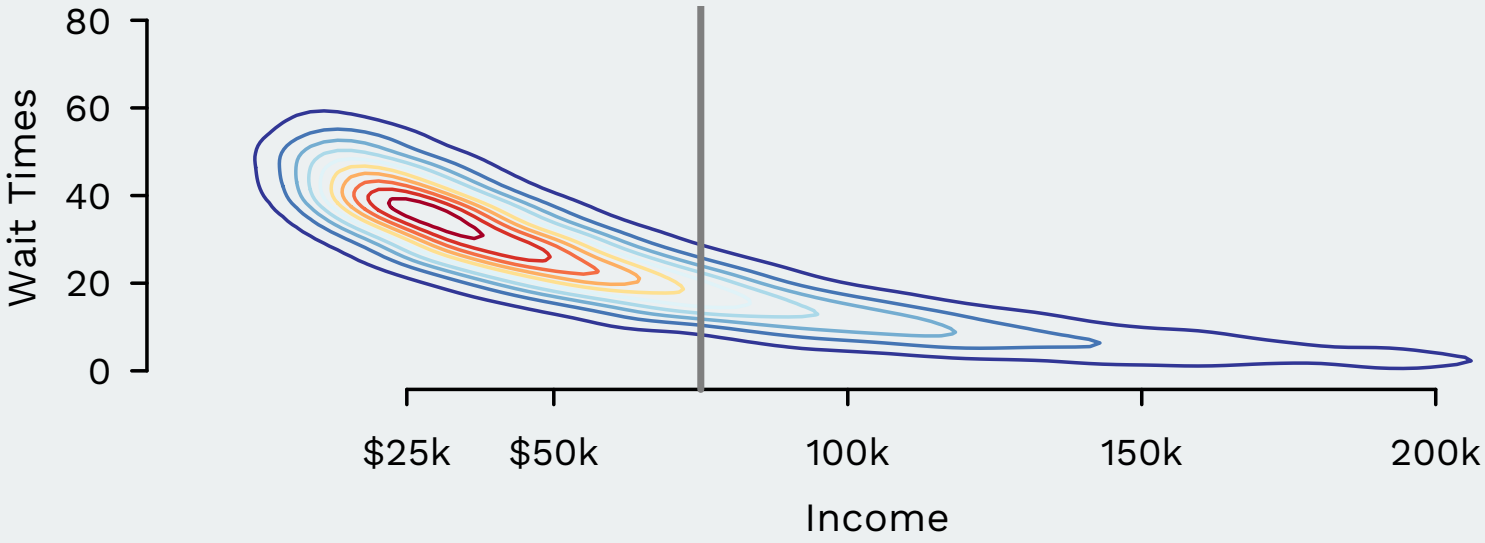


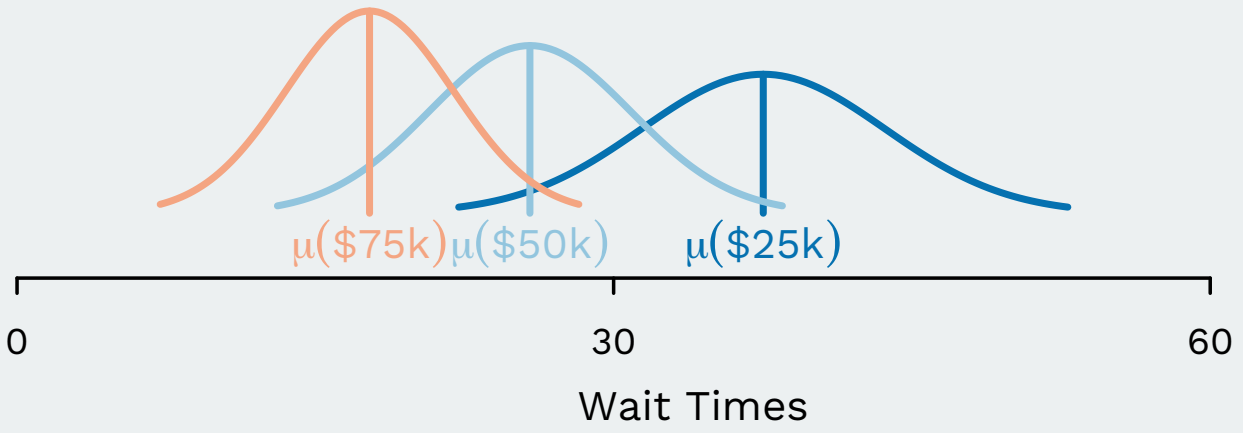
Wait times and income



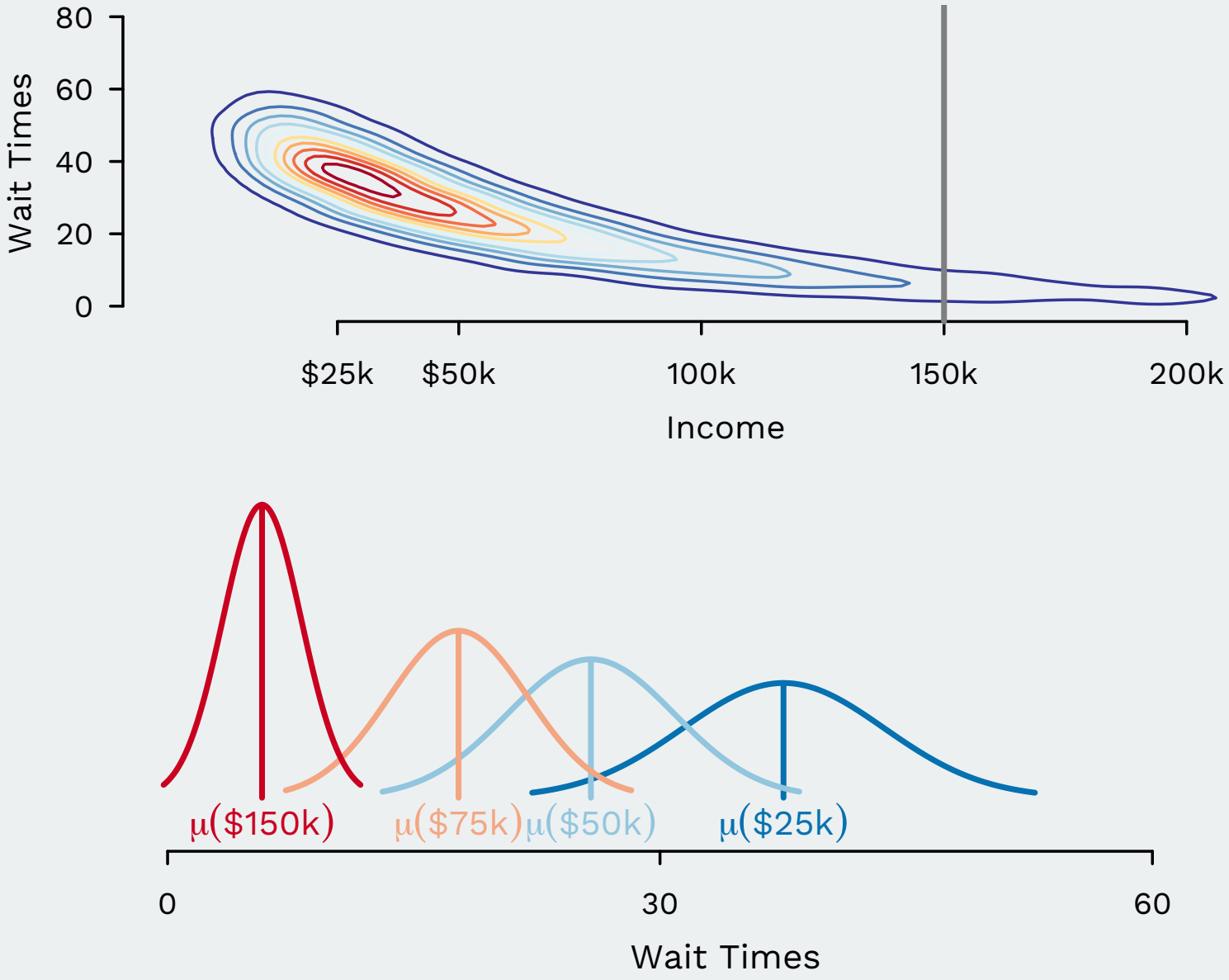


Wait times and income





Wait times and income



The CEF decomposition

- We can always decompose Y_i into the CEF and an error:

$$Y_i = \mathbb{E}[Y_i|X_i] + u_i$$

- Here, the CEF error has two definitional properties:
  - The mean of the error doesn't depend on X_i: $\mathbb{E}[u_i|X_i] = \mathbb{E}[u_i] = 0$
  - The error is uncorrelated with any function of X_i.
- Y_i can be decomposed into the part "explained by X_i" and a part that is uncorrelated with X_i.

Best predictor

- Another reason to focus on the CEF: it generates best predictions about Y_i using X_i.
- Let $g(X_i)$ be some function that generates prediction and define the mean squared error (MSE) of the prediction as:

$$\mathbb{E}[(Y_i - g(X_i))^2]$$

- What function should you pick? The CEF minimizes this prediction error:

$$\mathbb{E}[(Y_i - g(X_i))^2] \geq \mathbb{E}[(Y_i - \mu(X_i))^2]$$

- We say the CEF is the best predictor of Y_i among functions of X_i.
  - ...in terms of squared error.

3/ Estimating the CEF

Estimating the CEF for binary covariates

- How do we estimate $\widehat{\mathbb{E}}[Y_i|X_i = x]$?
- Sample means within each group:

$$\widehat{\mathbb{E}}[Y_i|X_i = 1] = \frac{1}{n_1} \sum_{i:X_i=1} Y_i$$

$$\widehat{\mathbb{E}}[Y_i|X_i = 0] = \frac{1}{n_0} \sum_{i:X_i=0} Y_i$$

- $n_1 = \sum_{i=1}^{n} X_i$ is the number of women in the sample.
- $n_0 = n - n_1$ is the number of men.
- $\sum_{i:X_i=1}$ sum only over the i that have $X_i = 1$, meaning that i is a woman.
- $\rightsquigarrow$ estimate the mean of Y_i conditional on X_i by just estimating the means within each group of X_i.

Binary covariate example

```
## mean of log GDP among non-African countries
mean(ajr$logpgp95[ajr$africa == 0], na.rm = TRUE)
```

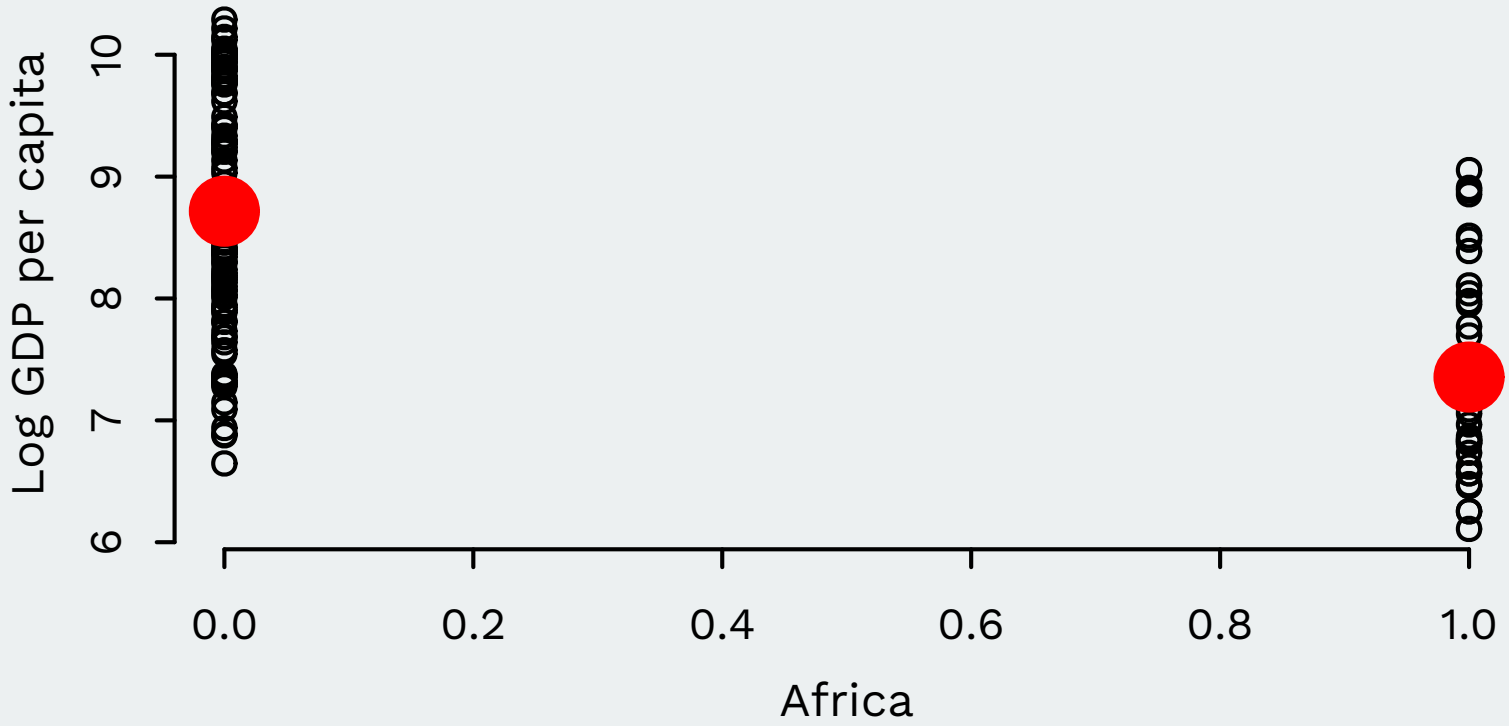
```
## [1] 8.716
```

```
## mean of log GDP among African countries
mean(ajr$logpgp95[ajr$africa == 1], na.rm = TRUE)
```

```
## [1] 7.355
```

Binary covariate CEF plot

```
plot(ajr$africa, ajr$logpgp95, ylab = "Log GDP per capita", xlab = "Africa",
    bty = "n")
points(x = 0, y = mean(ajr$logpgp95[ajr$africa == 0], na.rm = TRUE),
    pch = 19, col = "red", cex = 3)
points(x = 1, y = mean(ajr$logpgp95[ajr$africa == 1], na.rm = TRUE),
    pch = 19, col = "red", cex = 3)
```



Discrete covariate: estimating the CEF

- What if X_i isn't binary, but takes on > 2 discrete values?
- The same logic applies, we can still estimate $\mathbb{E}[Y_i|X_i = x]$ with the sample mean among those who have $X_i = x$:

$$\widehat{\mathbb{E}}[Y_i|X_i = x] = \frac{1}{n_x} \sum_{i:X_i=x} Y_i$$

Discrete covariate example

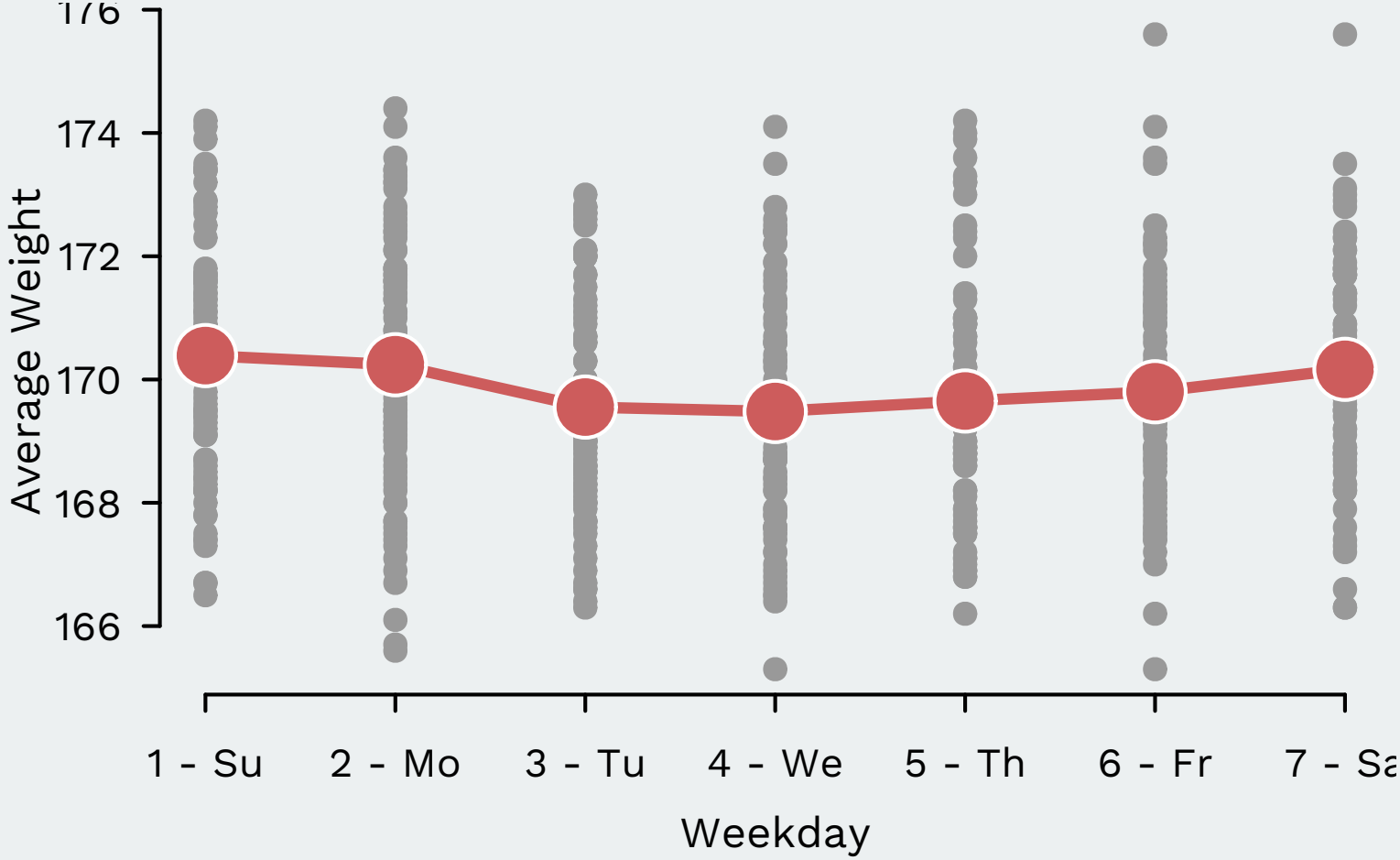
- I've been collecting data on my own weight for a while.
- How does my weight (Y_i) varied by the day of the week (X_i)?
- Calculate the mean weight for each day of the week:

```
weight <- read.csv("../data/weight.csv", stringsAsFactors = FALSE)
weight$weekday <- as.numeric(format(as.Date(weight$date, format = "%m/%d/%y%n%H:%M"),
    "%w")) + 1
weight$date <- as.Date(weight$date, format = "%m/%d/%y%n%H:%M")
day.means <- rep(NA, times = 7)
names(day.means) <- c("1 - Su", "2 - Mo", "3 - Tu", "4 - We", "5 - Th",
    "6 - Fr", "7 - Sa")
for (i in 1:7) {
    day.means[i] <- mean(weight$weight[weight$weekday == i])
}
day.means
```

```
## 1 - Su 2 - Mo 3 - Tu 4 - We 5 - Th 6 - Fr 7 - Sa
##  170.4  170.2  169.6  169.5  169.7  169.8  170.2
```

Discrete covariate CEF plot

```
plot(x = weight$weekday, y = weight$weight, xaxt = "n", xlab = "Weekday",
    ylab = "Average Weight", pch = 19, col = "grey60")
lines(x = 1:7, y = day.means, pch = 19, col = "indianred", lwd = 3)
points(x = 1:7, y = day.means, pch = 21, col = "white", cex = 3, bg = "indianred")
axis(side = 1, at = 1:7, labels = names(day.means))
```



Continuous covariate (I): each unique value gets a mean

- What if X_i is continuous? Can we calculate a mean for every value of X_i?
- Not really, because remember the probability that two values will be the same in a continuous variable is 0.
- Thus, we'll end up with a very "jumpy" function, $\widehat{\mathbb{E}}[Y_i|X_i = x]$, since n_x will be at most 1 for any value of x.

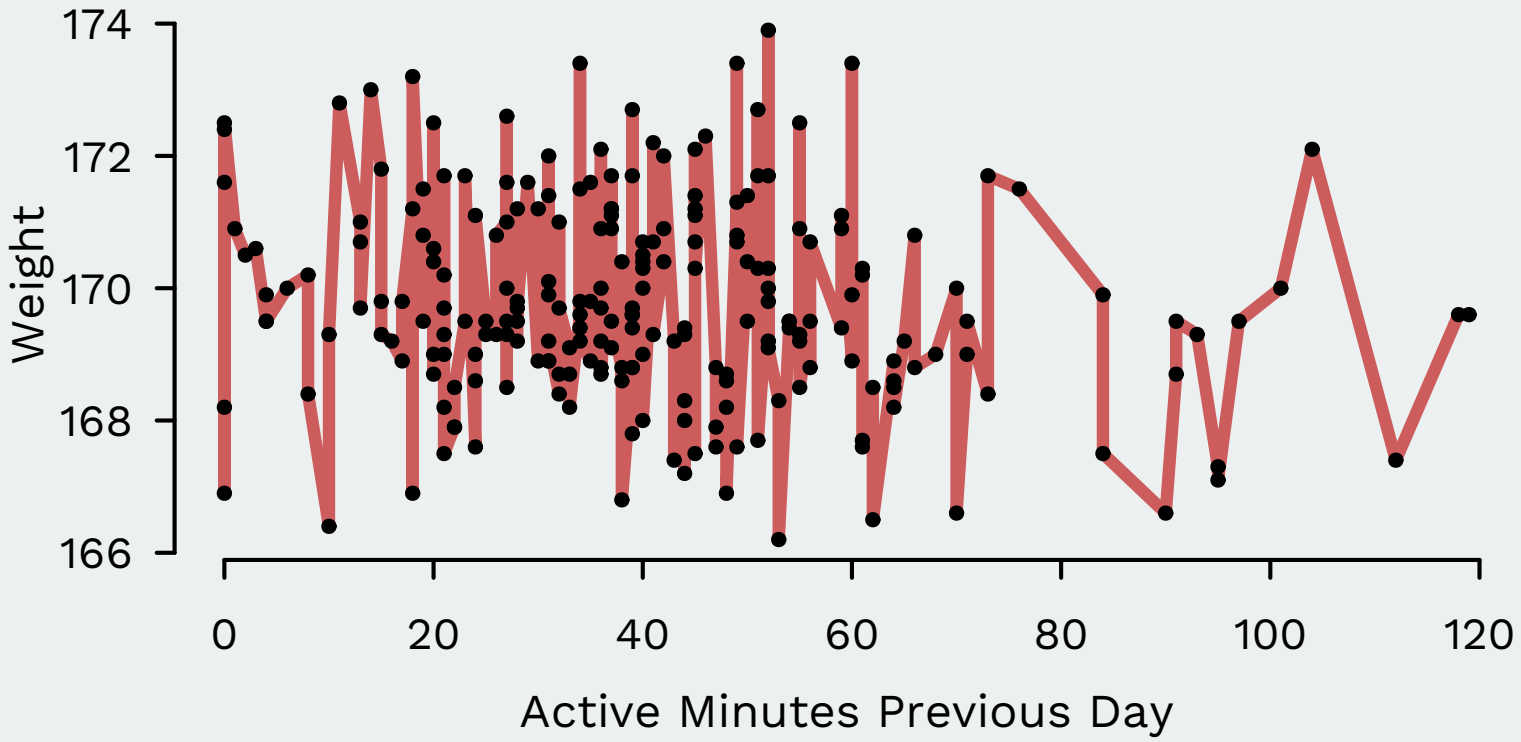
Continuous covariate (I) example

- I also wear an activity tracker and that collects how active I am during the day
- Let's look at the relationship between my weight and my active minutes in the previous day using this approach.

```
fitbit <- read.csv("../data/fitbit.csv", stringsAsFactors = FALSE)
fitbit$date <- as.Date(fitbit$date, format = "%m/%d/%y")
## lag fitbit by one day
fitbit$date <- fitbit$date + 1
## merge fitbit and weight data
weight <- merge(weight, fitbit, by = "date")
```

Continuous covariate (I) CEF plot

```
plot(weight$active.mins[order(weight$active.mins)],
     weight$weight[order(weight$active.mins)], type = "l", lwd = 3, pch = 19,
     col = "indianred",xlab = "Active Minutes Previous Day", ylab = "Weight")
points(weight$active.mins, weight$weight, pch = 19, cex = 0.5)
```



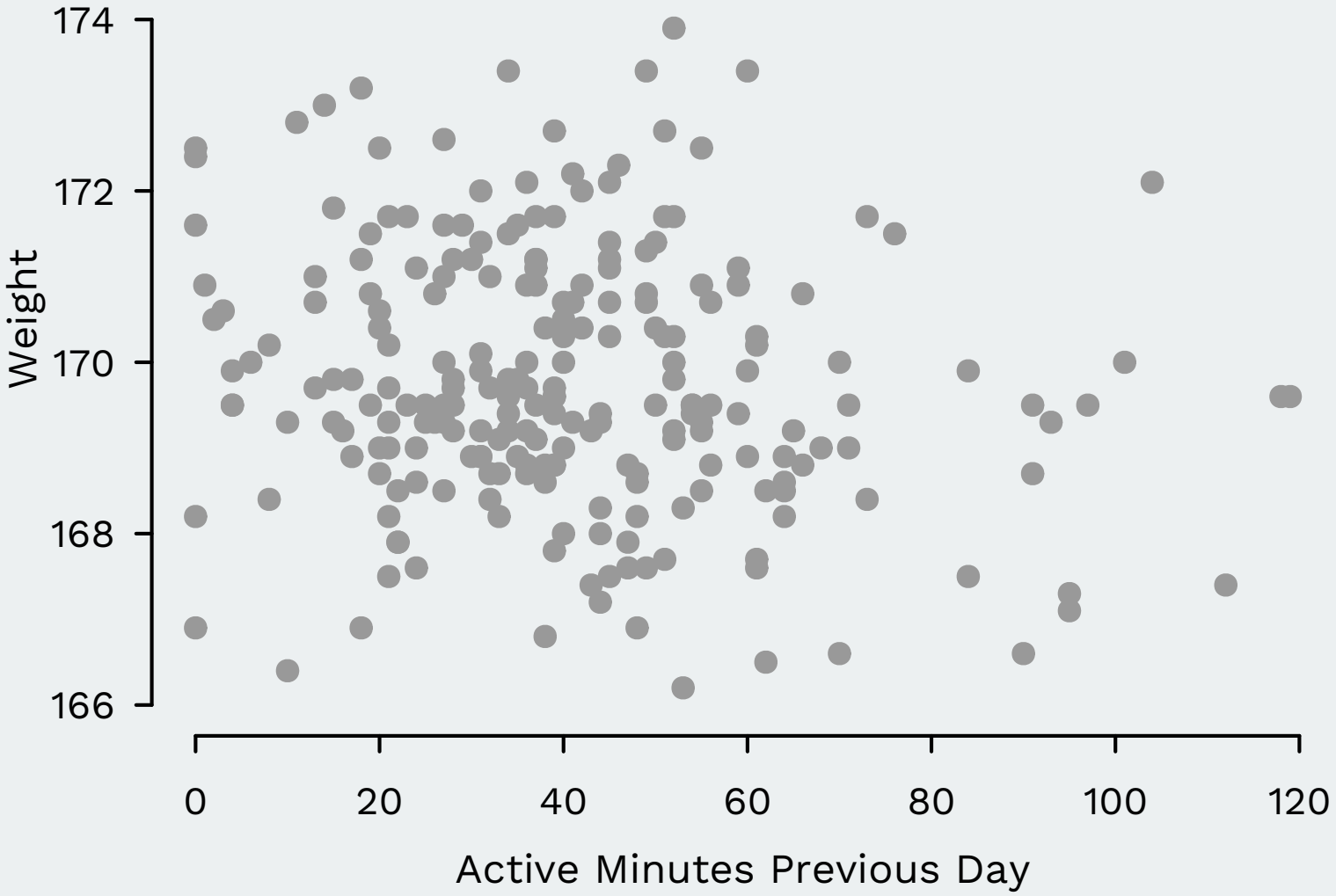
- Not a useful summary of the relationship between X_i and Y_i.

Continuous covariate (II): stratify and take means

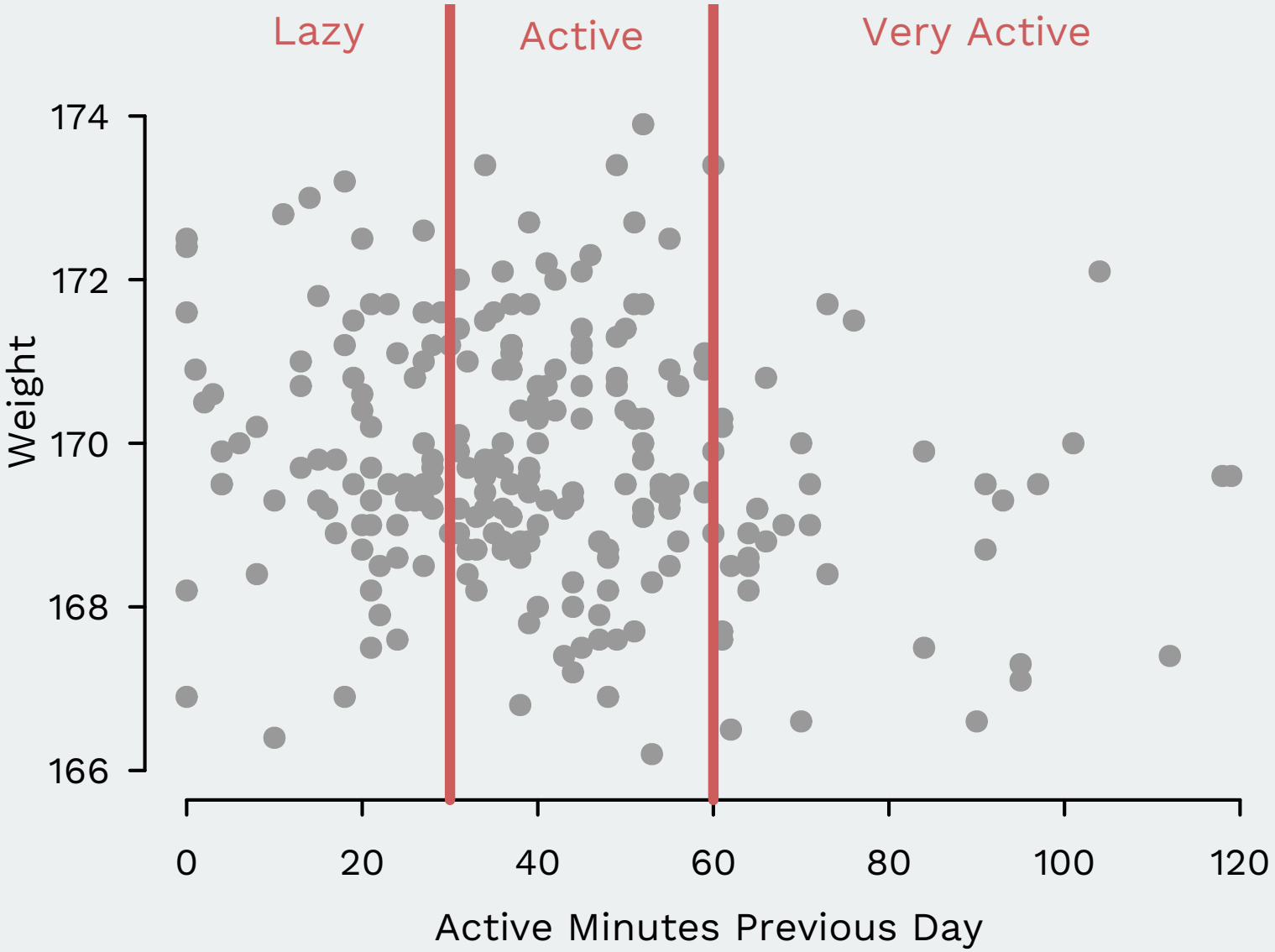
- So, that seems like each value of X_i won't work, but maybe we can take the continuous variable and turn it into a discrete variable. We call this stratification.
- Once it's discrete, we can just calculate the means within each strata.
- For instance, we could break up the "Active Minutes" variable into 3 categories: lazy ($<$ 30mins), active (30-60mins), and very active ($>$60min).

```
lowactivity.mean <- mean(weight$weight[weight$active.mins < 30])
medactivity.mean <- mean(weight$weight[weight$active.mins >= 30 & weight$active.mins
    60])
hiactivity.mean <- mean(weight$weight[weight$active.mins >= 60])
```

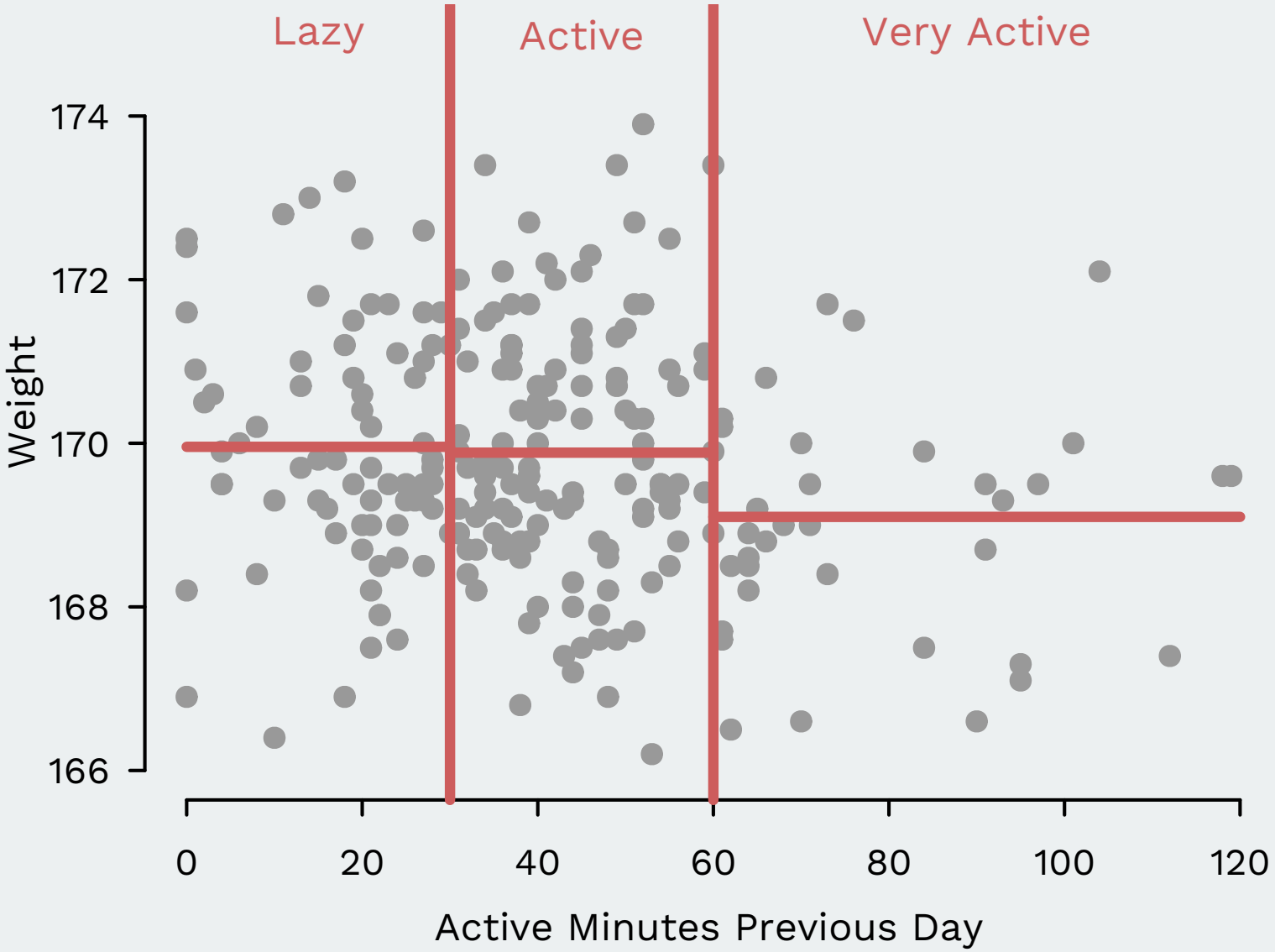
Continuous covariate (II) stratified CEF



Continuous covariate (II) stratified CEF



Continuous covariate (II) stratified CEF



4/ Linear CEFs and Linear Projections

Linear CEFs

- Obviously, estimation is going to be difficult with continuous covariates.
  - Even stratification had many hidden assumptions: number of categories, cutoffs for the categories, constant means within strata, etc.
- We can side-step some of these issues by assuming that the CEF is linear:

$$\mu(x) = \mathbb{E}[Y_i|X_i = x] = \beta_0 + \beta_1 x$$

- Intercept, β_0: the condition expectation of Y_i when $X_i = 0$
- Slope, β_1: average change in the mean of Y_i given a one-unit change in X_i

Why is linearity an assumption?

- Example: Y_i is income, X_i is years of education.
  - β_0: average income among people with 0 years of education.
  - β_1: expected difference in income between two adults that differ by 1 year of education.
- Why is linearity an assumption?

$$\mathbb{E}[Y_i|X_i = 12] - \mathbb{E}[Y_i|X_i = 11] = \mathbb{E}[Y_i|X_i = 16] - \mathbb{E}[Y_i|X_i = 15] = \beta_1$$

- Effect of getting HS degree is the same as the effect of getting college degree.

Linear CEF with a binary covariate

- Return to wait-times and race example, with $X_i = 1$ being white and $X_i = 0$ being non-white.
  - Two possible values of the CEF: $\mu(1)$ for whites and $\mu(0)$ for non-whites.
- Can write the CEF as follows:

$$\mathbb{E}[Y_i|X_i = x] = \mu(x) = \mu(0) + \big(\mu(1) - \mu(0)\big)\, x$$

- Rewriting with $\beta_0 = \mu(0)$ and $\beta_1 = \mu(1) - \mu(0)$:

$$\mu(x) = \beta_0 + \beta_1 x$$

- No assumptions, just rewriting!
  - β_0: expected wait-time for non-whites
  - β_1: difference in expected wait times between whites and non-whites.

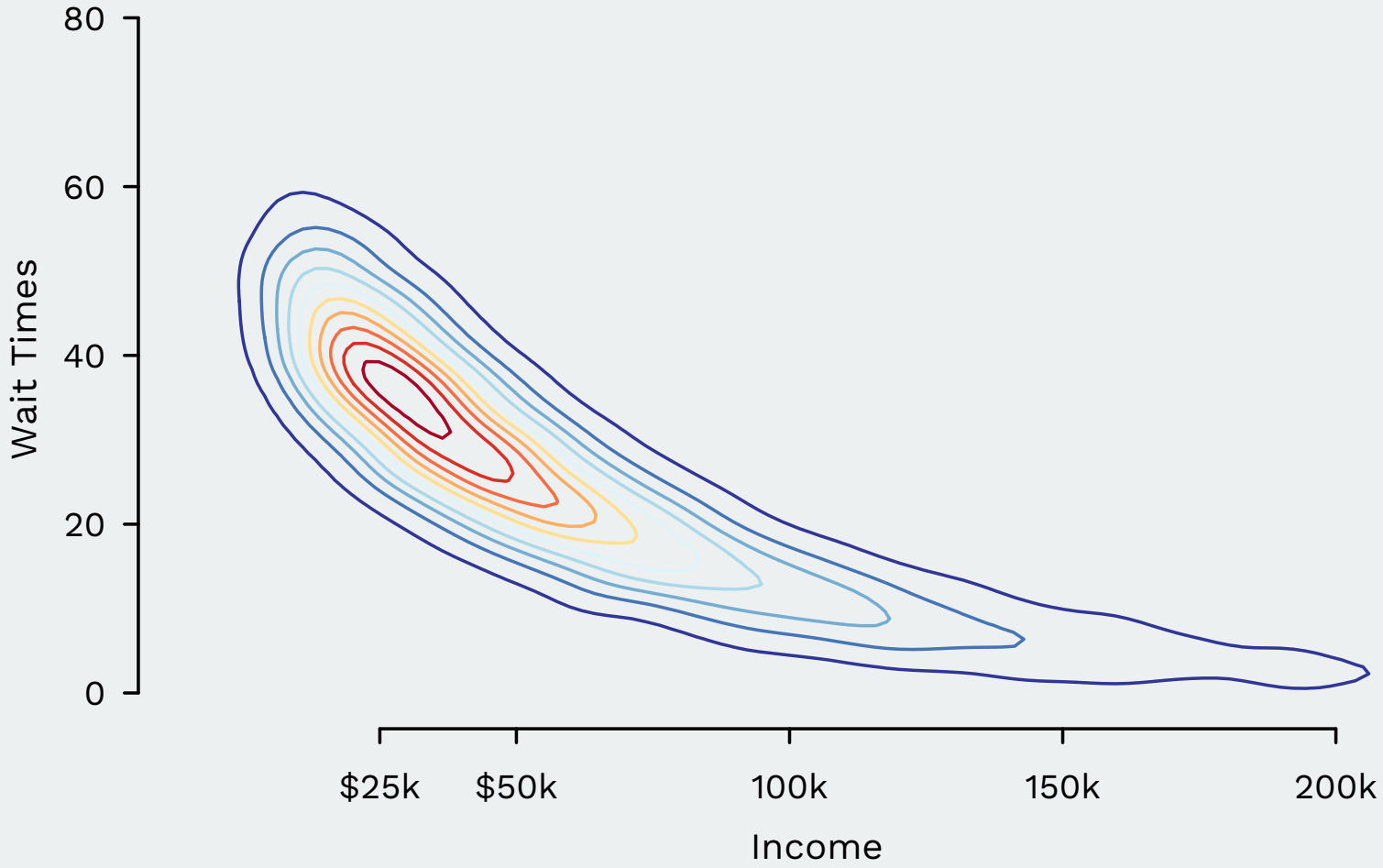
Linear approximation

- Ugh, what if the CEF isn't linear but we assume it is?
- Better to think of there being a population line of best fit that is the best linear approximation to Y_i.
- Mathematically, find the linear function of X_i that minimizes the squared prediction errors:

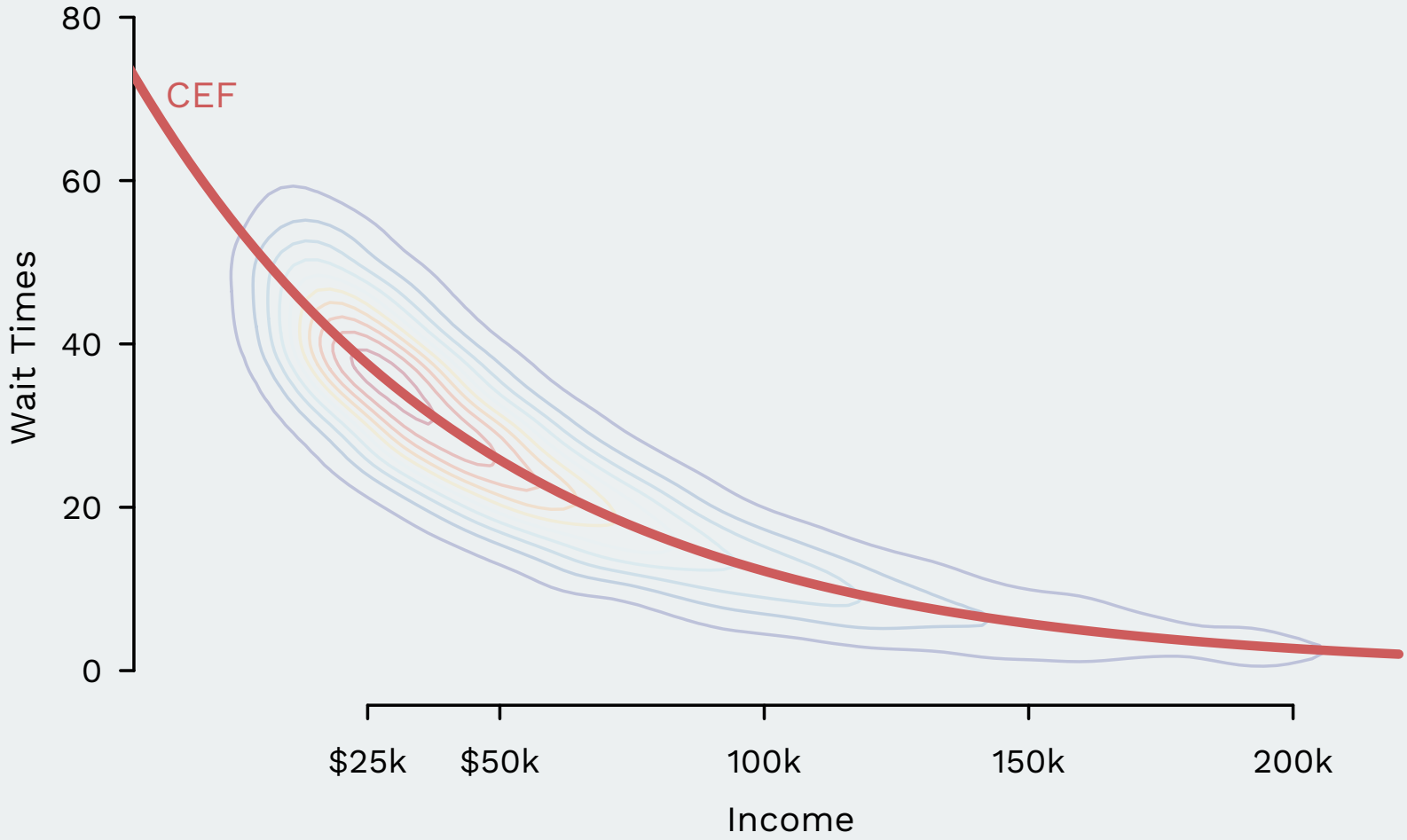
$$(\beta_0, \beta_1) = \underset{(b_0, b_1)}{\arg\min}\, \mathbb{E}[(Y_i - (b_0 + b_1 X_i))^2]$$

- Resulting function $\beta_0 + \beta_1 X_i$ is called the linear projection or the population linear regression of Y_i onto X_i.
- In general, distinct from the CEF:
  - CEF, $\mu(x)$ is the best predictor of Y_i among all functions.
  - Linear projection is best predictor among linear functions.

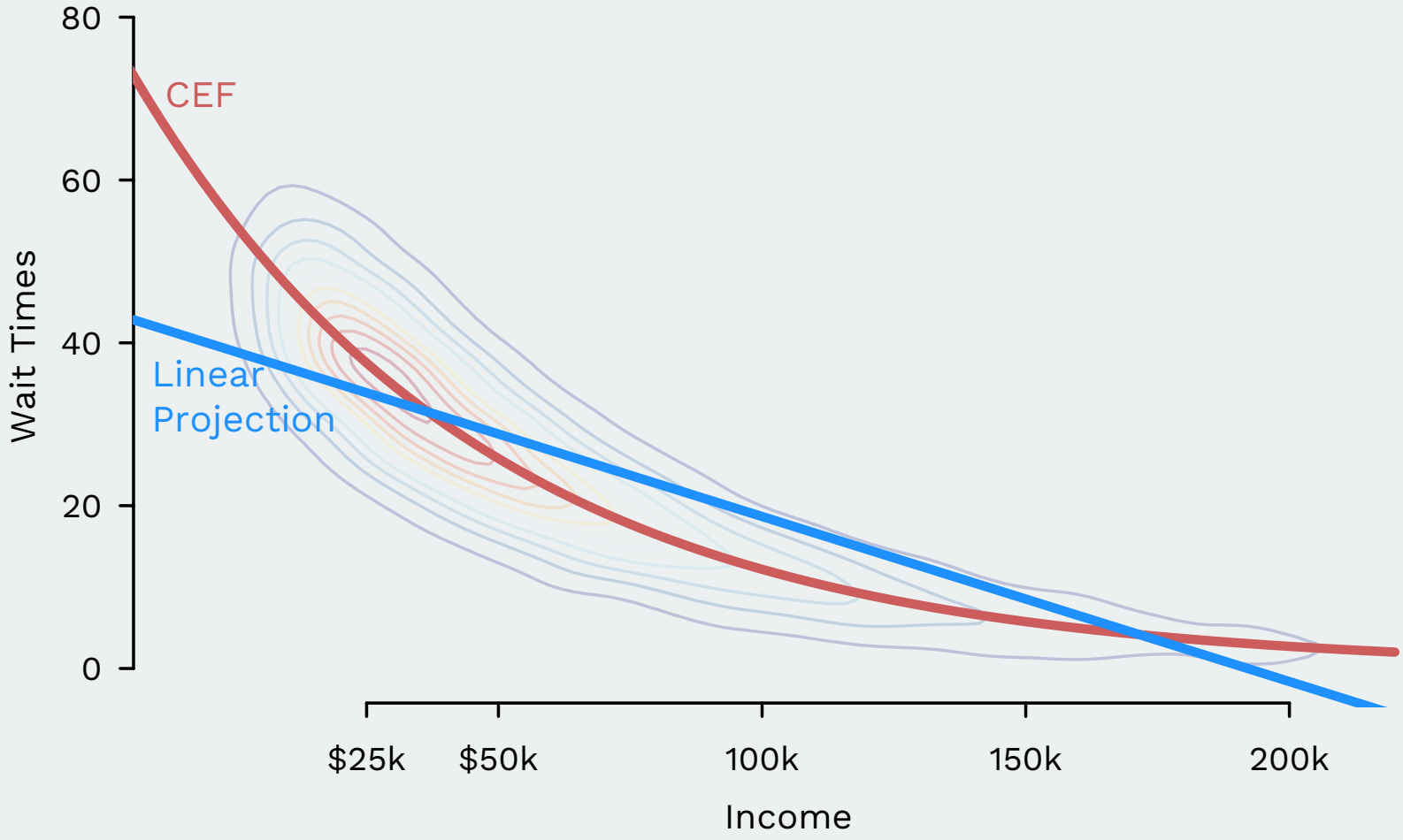
Linear approximation



Linear approximation



Linear approximation



Population linear projection

- Can we relate the intercept and slope of the population line of best fit to the joint distribution of Y_i and X_i?
- Yes, using some multivariate calculus, can show:

$$\beta_0 = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i]$$
$$\beta_1 = \frac{\text{Cov}[Y_i, X_i]}{\mathbb{V}[X_i]}$$

- What's awesome about the linear projection is that it exists and is well-defined even if the CEF is nonlinear.

Why the linear projection?

- Two handy results about the linear projection:

CEF is linear

If the CEF is a linear function, $\mathbb{E}[Y_i|X_i] = b_0 + b_1 X_i$, then it will be equal to the linear projection: $\mathbb{E}[Y_i|X_i] = \beta_0 + \beta_1 X_i$.

Linear projection approximates CEF

The linear projection is the best linear approximation to the CEF, so that:

$$(\beta_0, \beta_1) = \underset{(b_0, b_1)}{\arg\min}\, \mathbb{E}[(\mu(X_i) - (b_0 + b_1 X_i))^2]$$

5/ Least Squares

Back up and review

- To review our approach:
  - Defined a population line of best fit, $\beta_0 + \beta_1 X_i$.
  - If CEF is linear, it is equal to this line.
- Either way, β_0 and β_1 are valid population parameters just like μ or σ^2!
- Sample: $\{(Y_1, X_1), \ldots, (Y_n, X_n)\}$ are i.i.d. draws from a population joint distribution, $f_{(Y,X)}(y, x)$
- How can we use this sample to estimate β_0, β_1?

Sample line of best fit

- To get the linear projection, we found the population line of best fit:

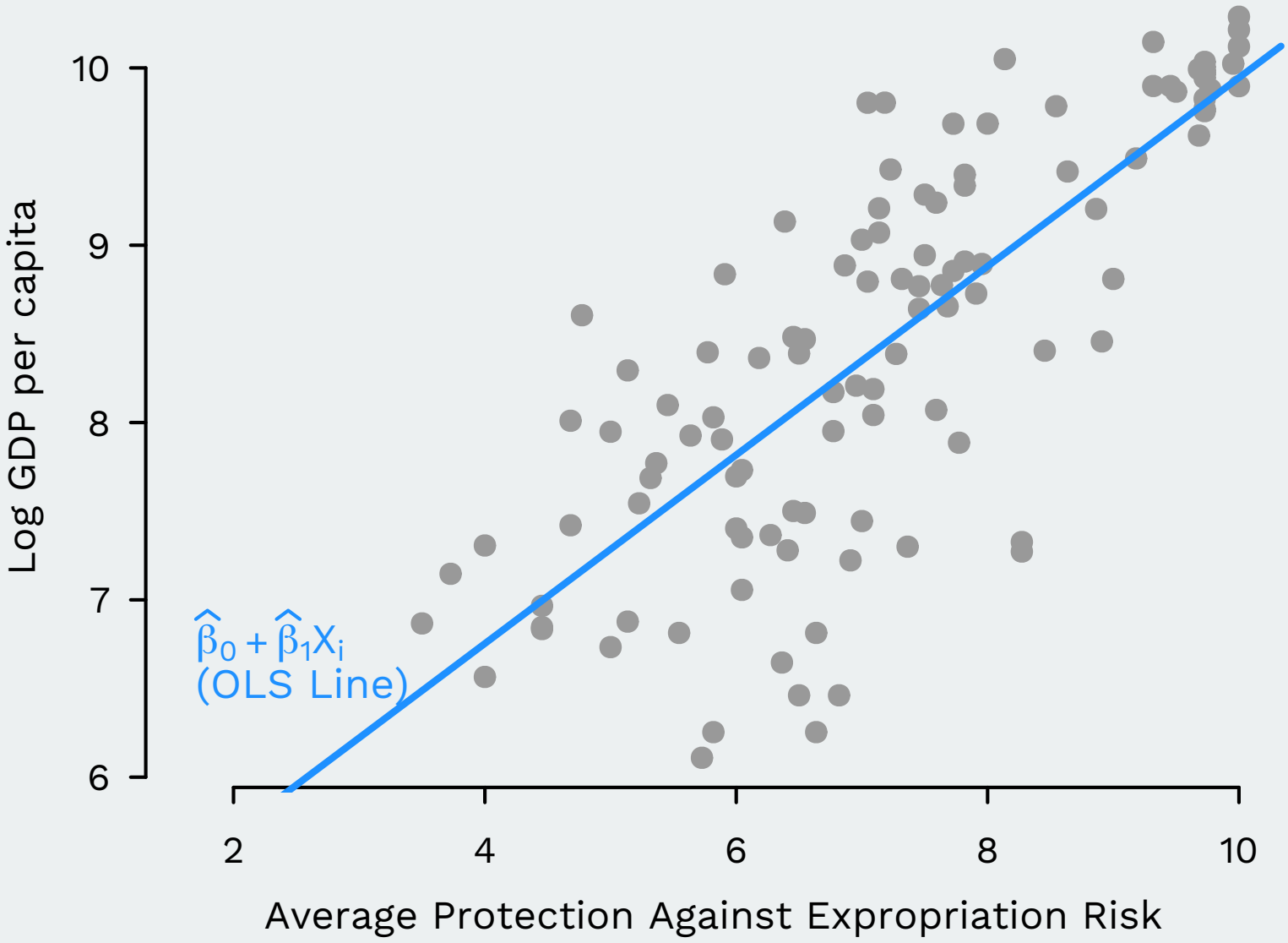
$$(\beta_0, \beta_1) = \arg\min_{(b_0,b_1)} \mathbb{E}[(Y_i - b_0 - b_1 X_i)^2]$$

- To get the sample line of best fit, we replace the population expectation with a sample mean:

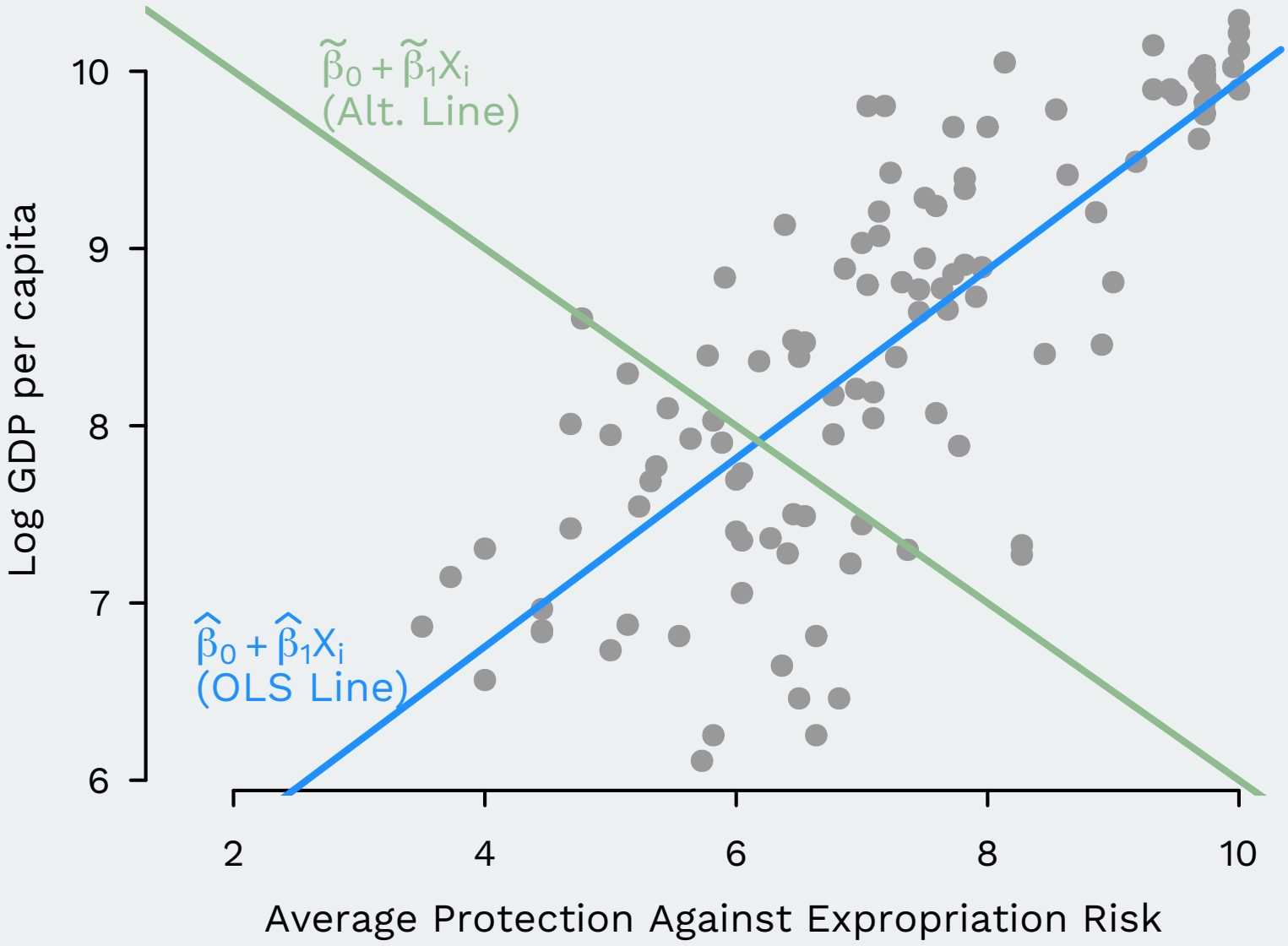
$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{b_0,b_1} \frac{1}{n} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

- This estimator is called least squares (LS) or ordinary least squares (OLS).

Fitted OLS lines



Fitted OLS line



Fitted values and residuals

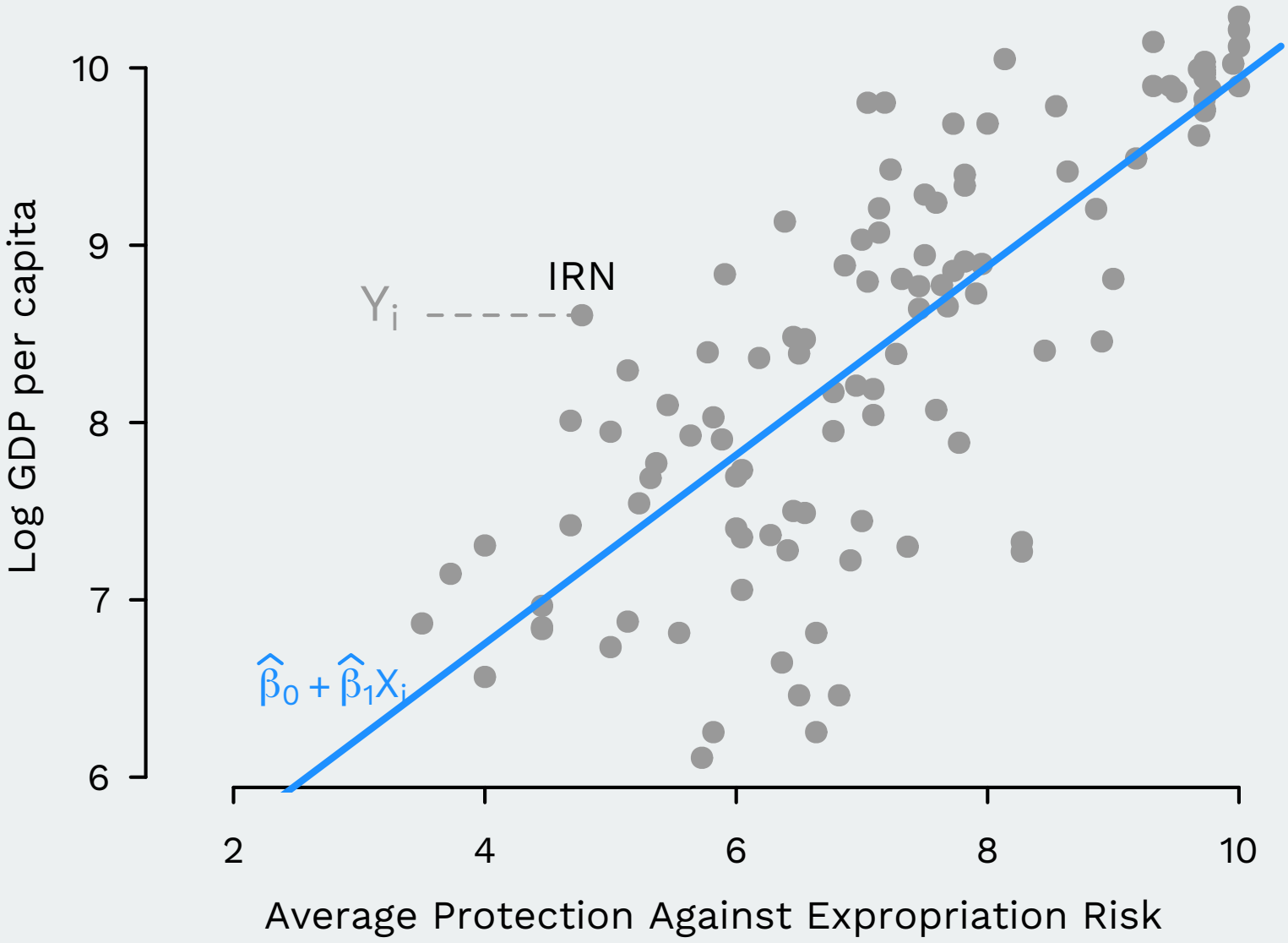
- Definition A fitted value is the estimated conditional mean of Y_i for a particular observation with independent variable X_i:

$$\widehat{Y}_i = \widehat{\mathbb{E}}[Y_i|X_i] = \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

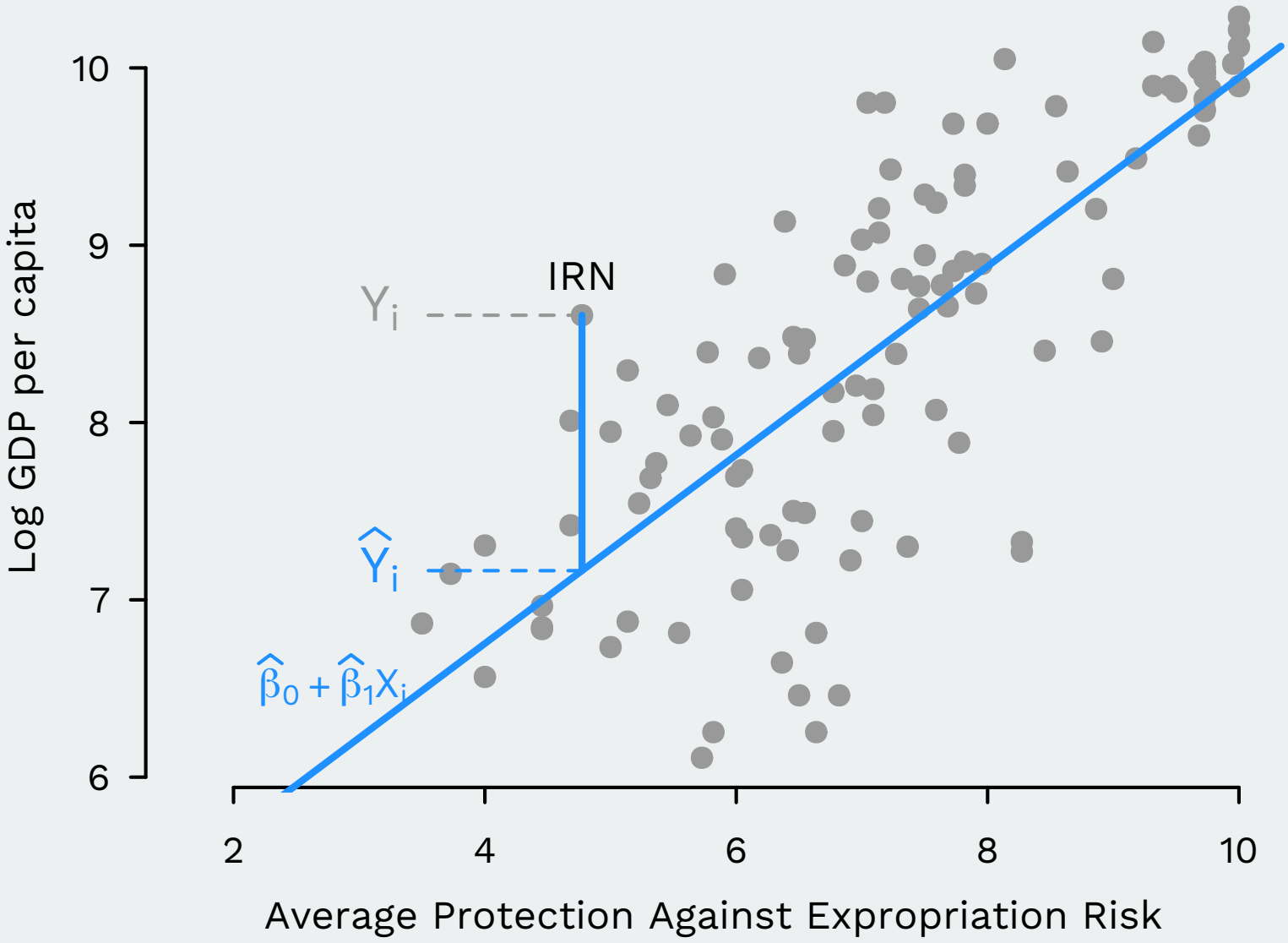
- Definition The residual is the difference between the actual value of Y_i and the fitted value, $\widehat{Y}_i$:

$$\hat{u}_i = Y_i - \widehat{Y}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$$

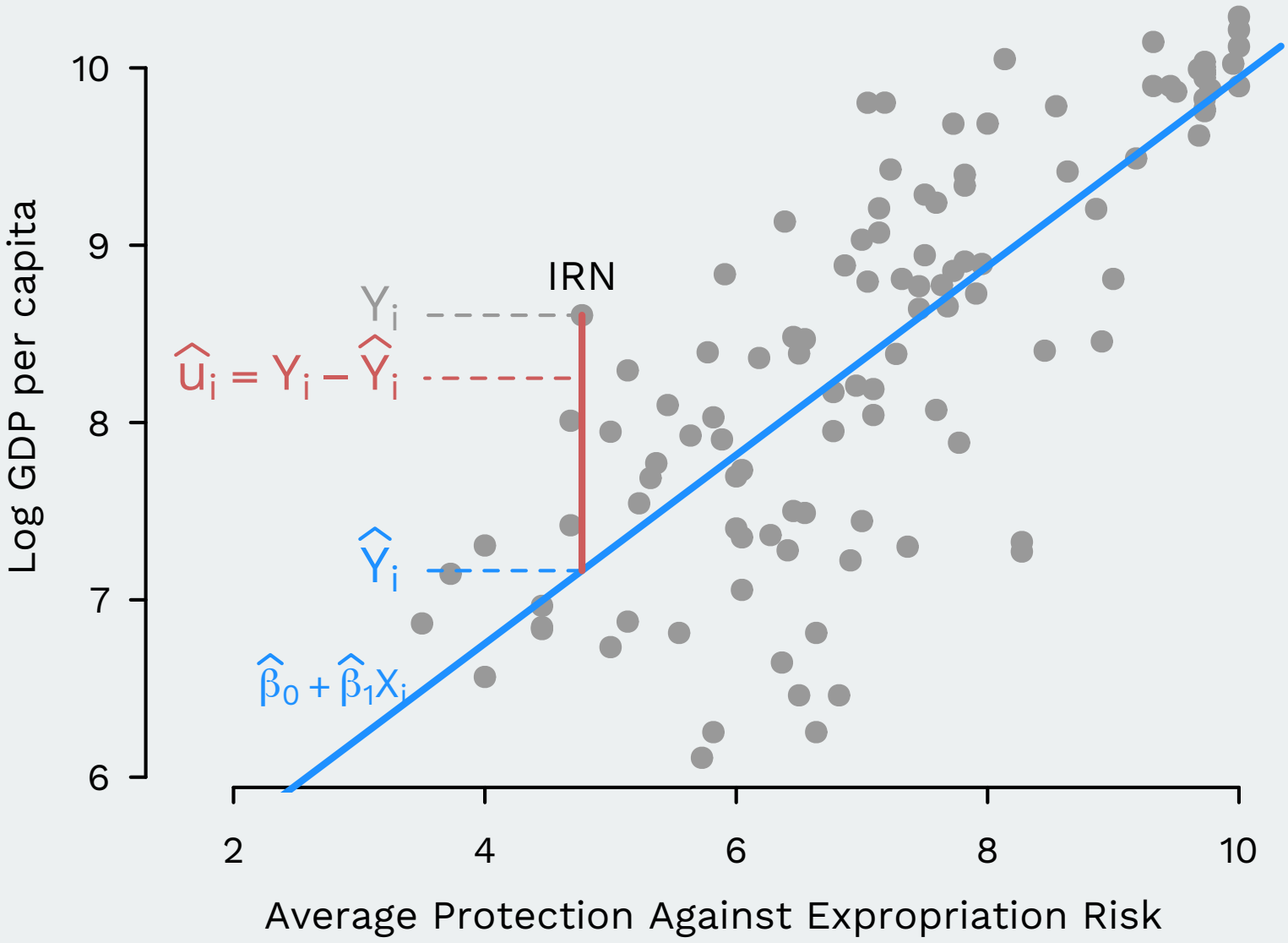
Fitted OLS line



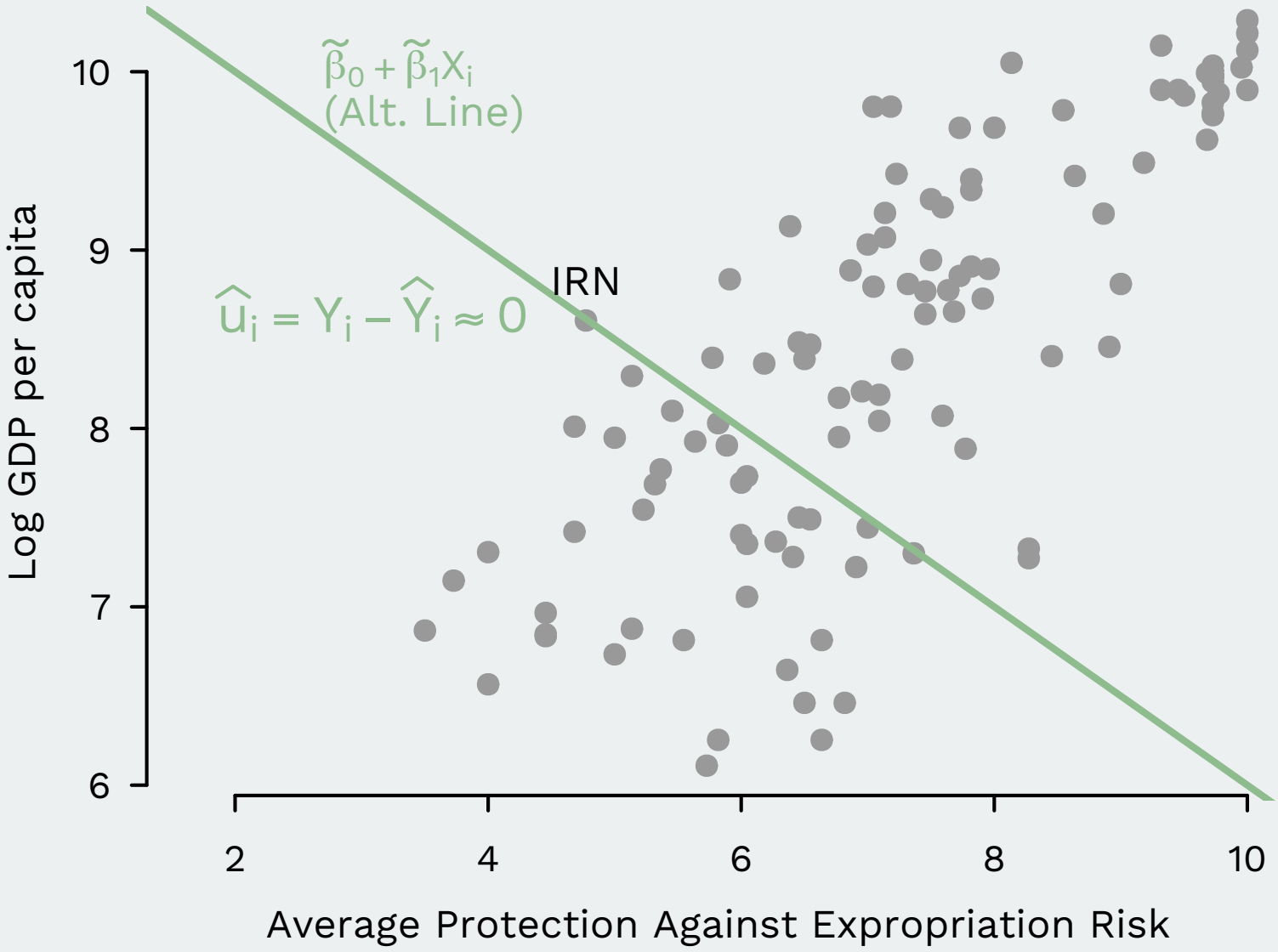
Fitted OLS line



Fitted OLS line



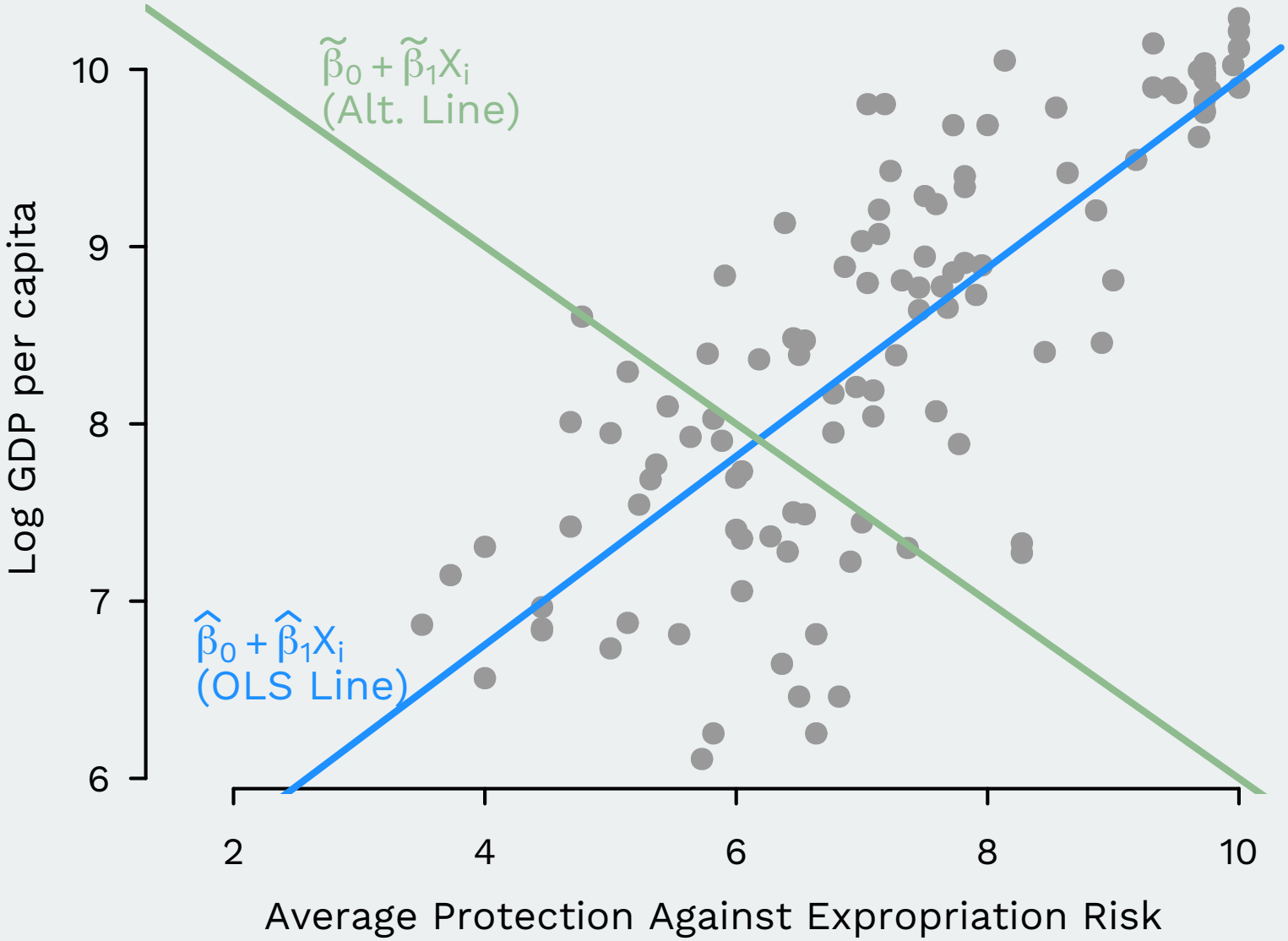
Why not this line?



Minimize the residuals

- The residuals, $\hat{u}_i = Y_i - \widehat{\beta}_0 - \widehat{\beta}_1 X_i$, tell us how well the line fits the data.
  - Larger magnitude residuals means that points are very far from the line
  - Residuals close to 0 mean points very close to the line
- The smaller the magnitude of the residuals, the better we are doing at predicting Y_i
- Choose the line that minimizes the residuals

Which is better at minimizing residuals?



OLS estimator

- OLS estimator defined by minimized squared residuals:

$$(\widehat{\beta}_0, \widehat{\beta}_1) = \arg\min_{b_0, b_1} \frac{1}{n} \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)^2$$

- Can we write the OLS intercept ($\widehat{\beta}_0$) and slope ($\widehat{\beta}_1$) in terms of quantities that we know? Yes!

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

Sample (co)variance

- Sample covariance:

$$\widehat{\text{Cov}}[X, Y] = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X}_n)(Y_i - \overline{Y}_n)$$

- Sample variance:

$$\widehat{\mathbb{V}}[X_i] = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$$

- Thus, we can rewrite the OLS slope as:

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2} = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\mathbb{V}}[X_i]}$$

Linear projection vs OLS

- Compare the linear projection intercept/slope in the population:

$$\beta_0 = \mathbb{E}[Y_i] - \beta_1 \mathbb{E}[X_i]$$
$$\beta_1 = \frac{\text{Cov}[Y_i, X_i]}{\mathbb{V}[X_i]}$$

- With the OLS intercept/slope in the sample:

$$\widehat{\beta}_0 = \overline{Y} - \widehat{\beta}_1 \overline{X}$$
$$\widehat{\beta}_1 = \frac{\widehat{\text{Cov}}(X, Y)}{\widehat{\mathbb{V}}[X_i]}$$

- OLS is just replaces all the population expectations with sample versions!

AJR Example in R

- Let's use those simple formulas we just learned:

```
ajr <- na.omit(ajr[, c("avexpr", "logpgp95")])
cov.xy <- cov(ajr$avexpr, ajr$logpgp95)
var.x <- var(ajr$avexpr)
cov.xy/var.x
```

```
## [1] 0.5319
```

```
mean(ajr$logpgp95) - cov.xy/var.x * mean(ajr$avexpr)
```

```
## [1] 4.626
```

- Compare it to what `lm()`, the OLS function in R produces:

```
coef(lm(logpgp95 ~ avexpr, data = ajr))
```

```
## (Intercept)      avexpr
##      4.6261      0.5319
```

Mechanical properties of least squares

- The residuals will be 0 on average:

$$\sum_{i=1}^{n} \hat{u}_i = 0$$

- The residuals will be uncorrelated with the predictor:

$$\sum_{i=1}^{n} X_i \hat{u}_i = 0 \rightsquigarrow \widehat{\text{Cov}}(X_i, \hat{u}_i) = 0$$

- The residuals will be uncorrelated with the fitted values:

$$\sum_{i=1}^{n} \widehat{Y}_i \hat{u}_i = 0 \rightsquigarrow \widehat{\text{Cov}}(\widehat{Y}_i, \hat{u}_i) = 0$$

Mechanical properties of least squares in R

```
mod <- lm(logpgp95 ~ avexpr, data = ajr)
mean(residuals(mod))
```

```
## [1] -2.006e-17
```

```
cor(ajr$logem4, residuals(mod))
```

```
## [1] -3.185e-17
```

```
cor(fitted(mod), residuals(mod))
```

```
## [1] -1.16e-16
```

Wrap up

- What is regression: estimating the CEF of Y_i given X_i
- Easy to do with sample means with discrete X_i
- Need parametric assumptions when X_i is continuous
- Derived an estimator for linear projection of Y_i on X_i