12. Conditional Expectation

Spring 2021

Matthew Blackwell

Gov 2002 (Harvard)

Where are we? Where are we going?

- We've learned a lot of probability and the basics of inference.
- Time to move onto to regression! But first: what is regression?
- At its core: how the average of one variable varies with others.

Defining condition expectations

Definition

The **conditional expectation** of Y conditional on $\mathbf{X} = \mathbf{x}$ is:

$$\mu(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}] = \begin{cases} \sum_y y \, \mathbb{P}(Y = y \mid \mathbf{X} = \mathbf{x}) & \text{discrete } Y \\ \int_{-\infty}^{\infty} y \, f_{Y|\mathbf{X}}(y \mid \mathbf{x}) dy & \text{continuous } Y \end{cases}$$

- Expected value of the conditional distribution of Y given $\mathbf{X} = \mathbf{x}$.
  - $\mathbf{X} = (X_1, X_2, \dots, X_k)$ is a random vector ($k = 1$ just an r.v.)
- Viewed as a function of x, it is the **conditional expectation function (CEF)**
  - How does the average value of Y change given different levels of $\mathbf{X}$?

Conditional expectation example

	Support Gay Marriage $Y = 1$	Oppose Gay Marriage $Y = 0$
Female $X = 1$	0.30	0.21
Male $X = 0$	0.22	0.27

- Conditional expectation of gay marriage support Y among men $X = 0$?

$$\begin{aligned}
\mathbb{E}[Y \mid X = 0] &= \sum_y y\, \mathbb{P}(Y = y \mid X = 0) \\
&= 0 \times \mathbb{P}(Y = 0 \mid X = 0) + 1 \times \mathbb{P}(Y = 1 \mid X = 0) \\
&= 1 \times \frac{0.22}{0.22 + 0.27} = 0.45
\end{aligned}$$

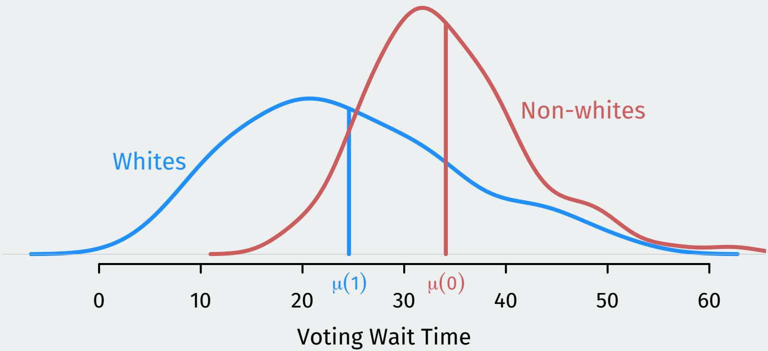
CEF for binary covariates

- 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:

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

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

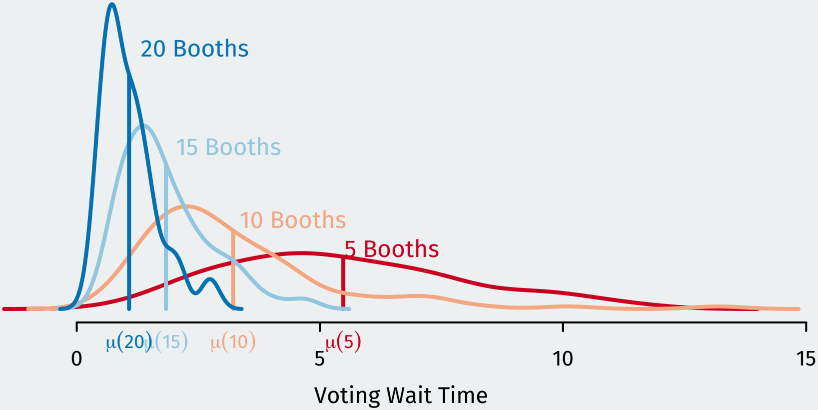
Why is the CEF useful?



- 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 # of polling booths at citizen i's polling station.
- $\mu(x)$ is the mean of Y_i changes as X_i changes:



CEF with multiple covariates

- We can also CEF conditioning on multiple variables $\mu(\mathbf{x})$:

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

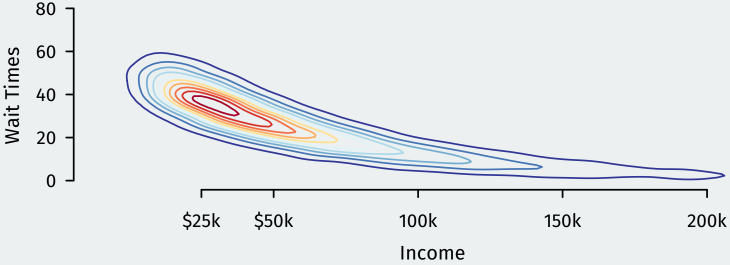
- Why? Allows more credible **all else equal** comparisons (ceteris paribus).
- Ex: average difference in wait times between white and non-white citizens **of the same gender**:

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

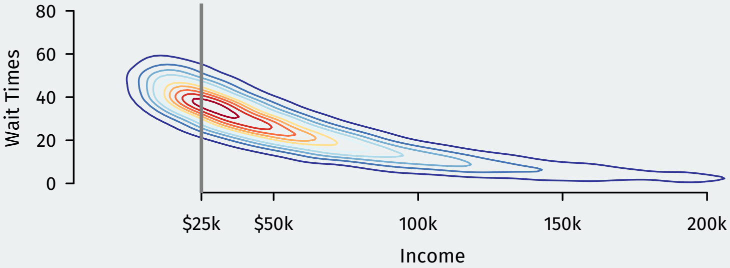
CEF for continuous covariates

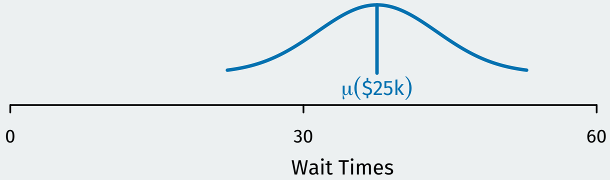
- What if our independent variable, X_i is income?
- Many possible values of $X_i \rightsquigarrow$ many possible values of $\mathbb{E}[Y_i|X_i = x]$.
  - Writing out each value of the CEF no longer feasible.
- Now we will think about $\mu(x) = \mathbb{E}[Y_i|X_i = x]$ as function. What does this function 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

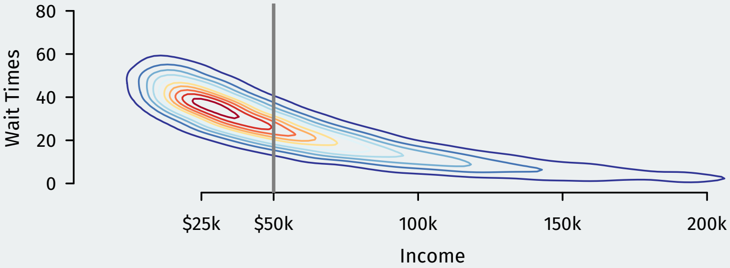


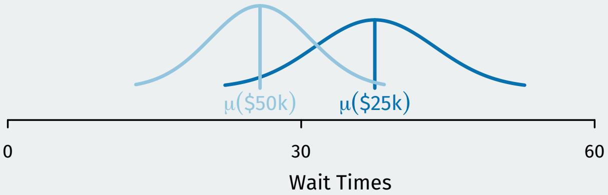
Wait times and income



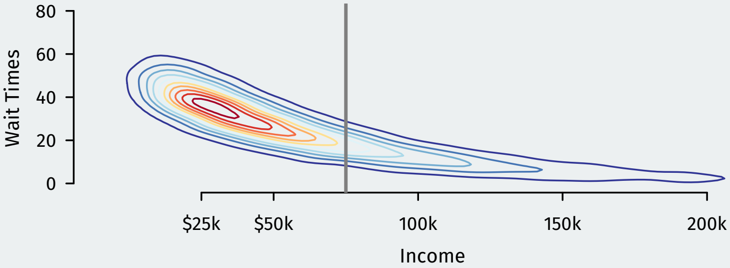


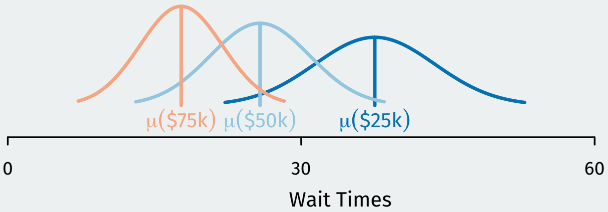
Wait times and income



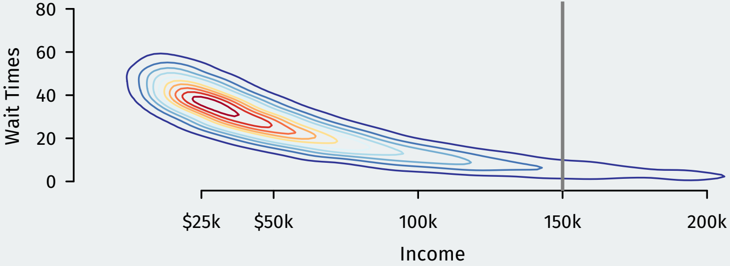


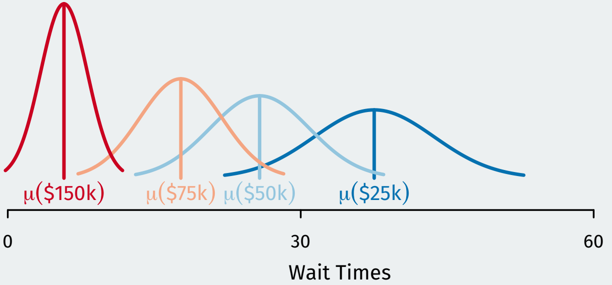
Wait times and income





Wait times and income





Conditional expectations as random variables

- The conditional expectation is a function of $\mathbf{x}$: $\mu(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$.
    - Not random: for a particular $\mathbf{x}$, $\mu(\mathbf{x})$ is a number.
    - Conditional expectation given an event.
- What about the conditional expectation given an r.v., $\mathbb{E}[Y \mid \mathbf{X}]$?
    - Why? Best prediction about Y given we get to know $\mathbf{X}$.
- Obtained by plugging r.v. into the CEF: $\mathbb{E}[Y \mid X] = \mu(X)$
- This is itself a random variable! For binary X:

$$\mathbb{E}[Y \mid X] = \begin{cases} \mu(0) & \text{with prob. } \mathbb{P}(X = 0) \\ \mu(1) & \text{with prob. } \mathbb{P}(X = 1) \end{cases}$$

- Has an expectation, $\mathbb{E}[\mathbb{E}[Y \mid X]]$, and a variance, $\mathbb{V}[\mathbb{E}[Y \mid X]]$.

Law of iterated expectations

Simple Law of Iterated Expectations

If $\mathbb{E}|Y| < \infty$, for any random vector $\mathbf{X}$, $\mathbb{E}\left\{\mathbb{E}[Y \mid \mathbf{X}]\right\} = E[Y]$.

- Expectation of the conditional expectation is the marginal expectation.
  - Discrete version: $\mathbb{E}\left[\mathbb{E}[Y \mid X]\right] = \sum_x \mathbb{E}[Y \mid X = x]\mathbb{P}(X = x) = \mathbb{E}[Y]$
  - Continuous version: $\mathbb{E}\left[\mathbb{E}[Y \mid X]\right] = \int_x \mathbb{E}[Y \mid X = x] f_X(x) dx = \mathbb{E}[Y]$
- General version allows for two conditioning sets:

Law of Iterated Expectations

If $\mathbb{E}|Y| < \infty$, for any random vectors $\mathbf{X}_1$ and $\mathbf{X}_2$,

$$\mathbb{E}\left\{\mathbb{E}[Y \mid \mathbf{X}_1, \mathbf{X}_2] \mid \mathbf{X}_1\right\} = E[Y \mid \mathbf{X}_1].$$

- “Averaging” over what is not constant ($\mathbf{X}_2$).

Example: law of iterated expectations

	Support Gay Marriage $Y = 1$	Oppose Gay Marriage $Y = 0$	Marginal
Female $X = 1$	0.30	0.21	0.51
Male $X = 0$	0.22	0.27	0.49
Marginal	0.52	0.48	

- $\mathbb{E}[Y \mid X = 1] = 0.59$ and $\mathbb{E}[Y \mid X = 0] = 0.45$.
- $\mathbb{P}(X = 1) = 0.51$ (females) and $\mathbb{P}(X = 0) = 0.49$ (males).
- Plug into the iterated expectations:

$$\begin{aligned}\mathbb{E}[\mathbb{E}[Y \mid X]] &= \mathbb{E}[Y \mid X = 0]\mathbb{P}(X = 0) + \mathbb{E}[Y \mid X = 1]\mathbb{P}(X = 1) \\ &= 0.45 \times 0.49 + 0.59 \times 0.51 = 0.52 = \mathbb{E}[Y]\end{aligned}$$

Properties of conditional expectations

1. $\mathbb{E}[c(X)Y \mid X] = c(X)\mathbb{E}[Y \mid X]$ for any function $c(X)$.
    - Example: $\mathbb{E}[X^2 Y \mid X] = X^2\mathbb{E}[Y \mid X]$ (If we know X, then we also know X^2)
2. If X and Y are independent r.v.s, then

$$\mathbb{E}[Y \mid X = x] = \mathbb{E}[Y].$$

3. If $X \perp\!\!\!\perp Y \mid Z$, then

$$\mathbb{E}[Y \mid X = x, Z = z] = \mathbb{E}[Y \mid Z = z]$$

4. Linearity: $\mathbb{E}[Y + X \mid Z] = \mathbb{E}[Y \mid Z] + E[X \mid Z]$

CEF errors and projection

- CEF error: $e = Y - \mathbb{E}[Y \mid \mathbf{X}]$
- Properties of the CEF error:

  1. $\mathbb{E}[e \mid \mathbf{X}] = 0$
  2. $\mathbb{E}[e] = 0$
  3. If $\mathbb{E}[|Y|^r] < \infty$ for $r \geq 1$, then $\mathbb{E}[|e|^r] < \infty$
  4. For any function $h(\mathbf{X})$, $h(\mathbf{X})$ is uncorrelated with e: $\mathbb{E}[h(\mathbf{X})e] = 0$

- Last property: CEF errors are **orthogonal** to the space of functions of $\mathbf{X}$.
  - $\mathbb{E}[Y \mid \mathbf{X}]$ is the **projection** of Y on the space of all functions of $\mathbf{X}$.
  - Closest point in that space to Y.
- These properties are definitional, not assumptions.

Conditional Expectation as Best Predictor

- Suppose we want to predict Y based on random vector $\mathbf{X}$.
  - We can use any function $g(\mathbf{X})$ as our predictor.
- Mean squared error of our predictions:

$$\mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$$

- What function will minimize this error? The CEF, $\mu(\mathbf{x})$!
- If $E[Y^2] < \infty$, then for any predictor $g(\mathbf{X})$,

$$\mathbb{E}\left[(Y - g(\mathbf{X}))^2\right] \geq \mathbb{E}\left[(Y - \mu(\mathbf{X}))^2\right]$$

Conditional Variance

Definition

The **conditional variance** of a Y given $\mathbf{X} =$ is defined as:

$$\sigma^2(\mathbf{x}) = \mathbb{V}[Y \mid \mathbf{X} = \mathbf{x}] = \mathbb{E}\left[(Y - \mu(\mathbf{x}))^2 \mid \mathbf{X} = \mathbf{x}\right]$$

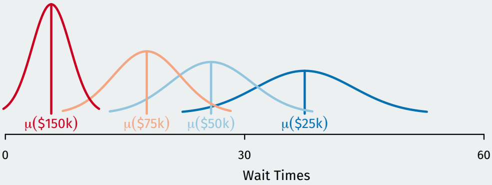
- Spread of the conditional distribution around its expectation.
- By definition, same as the variance of the CEF errors:

$$\mathbb{V}[Y \mid \mathbf{X} = \mathbf{x}] = \mathbb{V}[e \mid \mathbf{X} = \mathbf{x}] = \mathbb{E}[e^2 \mid \mathbf{X} = \mathbf{x}]$$

- Can re-express in the usual way:

$$\mathbb{V}[Y \mid \mathbf{X} = \mathbf{x}] = \mathbb{E}\left[Y^2 \mid \mathbf{X} = \mathbf{x}\right] - \left(\mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]\right)^2$$

Skedasticity



- The error is **homoskedastic** if $\sigma^2(\mathbf{x}) = \sigma^2$ does not depend on **x**.
    - Homoskedasticity greatly simplifies math, but often strong and implausible.
- The error is **heteroskedastic** if $\sigma^2(\mathbf{x})$ does depend on **x**
    - Hetero = different, skedastic = scatter
- Default assumption should be the less restrictive one: heteroskedastic

Conditional variance as a random variable

- Conditional variance is just a function of $\mathbf{x}$: $\sigma^2(\mathbf{x}) = \mathbb{V}[Y \mid \mathbf{X} = \mathbf{x}]$
- $\sigma^2(\mathbf{X}) = \mathbb{V}[Y \mid \mathbf{X}]$ is an r.v. and a function of $\mathbf{X}$, just like $\mathbb{E}[Y \mid \mathbf{X}]$.
- With a binary X:

$$\mathbb{V}[Y \mid X] = \begin{cases} \sigma^2(0) & \text{with prob. } \mathbb{P}(X = 0) \\ \sigma^2(1) & \text{with prob. } \mathbb{P}(X = 1) \end{cases}$$

- **Theorem** (Law of Total Variance/EVE's law):

$$\mathbb{V}[Y] = \mathbb{E}[\mathbb{V}[Y \mid \mathbf{X}]] + \mathbb{V}[\mathbb{E}[Y \mid \mathbf{X}]]$$

- The total variance can be decomposed into:
    1. the average of the within group variance ($\mathbb{E}[\mathbb{V}[Y \mid \mathbf{X}]]$) and
    2. how much the average varies between groups ($\mathbb{V}[\mathbb{E}[Y \mid \mathbf{X}]]$).

Res