# 11. (Linear) Regression

Fall 2023

Matthew Blackwell

Gov 2002 (Harvard)

# Where are we? Where are we going?

- Learned about estimation and inference in general.
- Now: building to a specific estimator, least squares regression.
- First we need to understand what a "linear model" is and when/why we need it.
    - No estimators quite yet. First, let's understand what we are estimating.
- Linear model is ubiquitous but poorly understood. Lots of subtlety here.

# Regression derivatives and partial effects

- Goal of regression: how mean of $Y$ changes with $X$.

$$\mu(\mathbf{x}) = \mathbb{E}[Y \mid \mathbf{X} = \mathbf{x}]$$

- For continuous regressors, we can use the partial derivative:

$$\frac{\partial \mu(x_1, \ldots, x_k)}{\partial x_1}$$

- For binary $X_1$, we can use the difference in conditional expectations:

$$\mu(1, x_2, \ldots, x_k) - \mu(0, x_2, \ldots, x_k)$$

- "Partial effect" of $X_1$ holding other included variables constant
- Exact form will depend on the functional form of $\mu(\mathbf{x})$.
    - How do we decide what form $\mu(\mathbf{x})$ should take?

# Estimating the CEF for discrete covariates

- To motivate function form, useful to think about estimation.
- How do we estimate $\mu(x) = \mathbb{E}[Y|X = x]$ for binary $X$?
- **Subclassification**: calculate sample averages with levels of $X_i$:

$$\hat{\mu}(1) = \frac{1}{n_1} \sum_{i=1}^{n} Y_i X_i$$

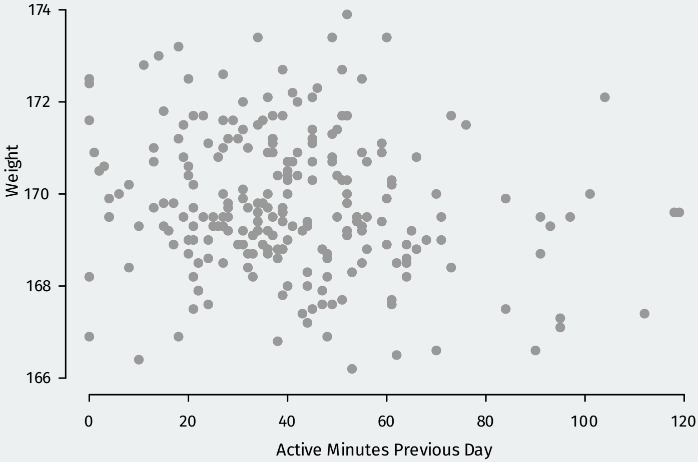
  - $n_1 = \sum_{i=1}^{n} X_i$ is the number of units with $X_i = 1$ in the sample.
- More generally for any discrete $X_i$:

$$\hat{\mu}(x) = \frac{\sum_{i=1}^{N} Y_i \mathbb{I}(X_i = x)}{\sum_{i=1}^{N} \mathbb{I}(X_i = x)}$$

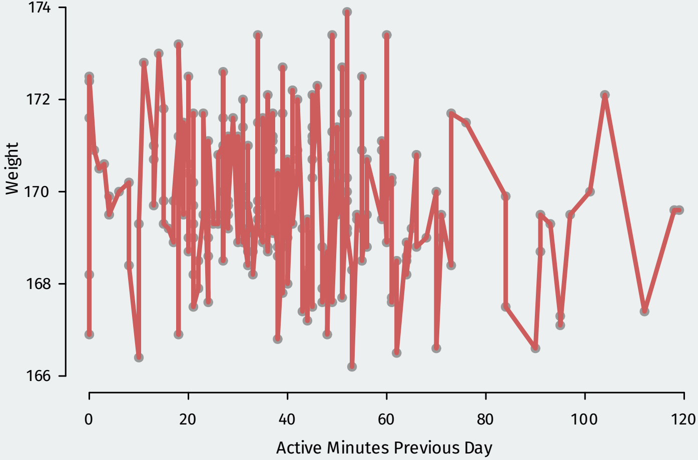
# Continuous covariates

- What if $X$ is continuous? Subclassification fall apart.
  - Each $i$ has a unique value: $\sum_{i=1}^{N} \mathbb{I}(X_i = x) = 1$
  - Very noisy estimates
  - What about any $x$ not in the sample?
- **Stratification**: bin $X_i$ into categories and treat like as discrete.
  - Every $x$ in the same bin gets the same conditional expectation.
  - Depends on arbitrary bin cutoffs/sizes.
- Example:
  - Personal data science: I wear an activity tracker and have a smart scale.
  - Relationship between my weight and active minutes in the previous day.

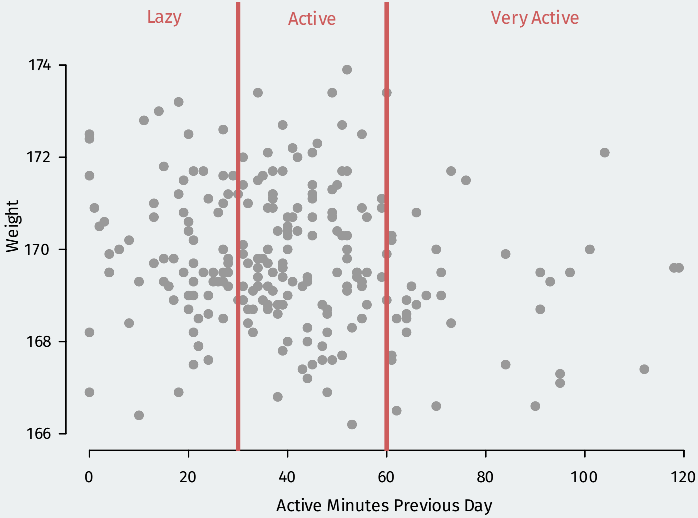
# Continuous covariate example



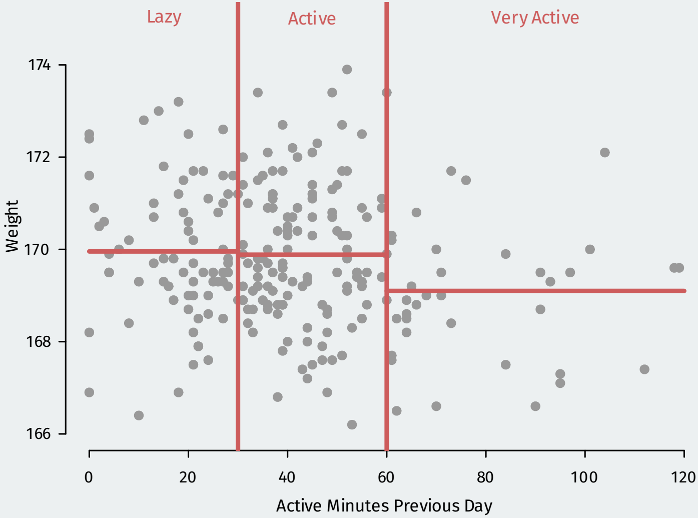
# Continuous covariate CEF: interpolation



# Continuous covariate CEF: stratification



# Continuous covariate CEF: stratification



# Linear CEFs

- Statification requires lots of choices/hidden assumptions.
    - Number of categories, cutoffs for the categories, constant means within strata, etc.
- Alternative: assuming that the CEF is **linear**:

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

- **Intercept**, $\beta_0$: the condition expectation of $Y_i$ when $X_i = 0$
- **Slope**, $\beta_1$: change in the CEF 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.
  - $\beta_0$: average income among people with 0 years of education.
  - $\beta_1$: expected difference in income between two adults that differ by 1 year of education.
- Why is linearity an assumption?

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

- Effect of HS degree is the same as the effect of college degree.
- Put another way: average partial effects are constant $\frac{\partial\mu(x)}{\partial x} = \beta_1$

# Linear CEF with nonlinear effects

- What if we think the effect is nonlinear?
- We can include nonlinear transformations:

$$\mu(x) = \beta_0 + x\beta_1 + x^2\beta_2$$

  - Partial effect now varies: $\partial\mu(x)/\partial x = \beta_1 + 2x\beta_2$
- **Linear** means linear in the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_k)$, not $\mathbf{X}$.
- We can also include **interactions** between covariates:

$$\mu(x_1, x_2) = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3$$

  - Average partial effect of $X_1$ depends on $X_2$: $\partial\mu(x_1, x_2)/\partial x_1 = \beta_1 + x_2\beta_3$

# Linear CEF with a binary covariate

- Wait-times ($Y_i$) and race ($X_i = 1$ for white, $X_i = 0$ for POC)
  - Two possible values of the CEF: $\mu_1$ for whites and $\mu_0$ for POC.
- Can write the CEF as follows:

$$\mu(x) = x\mu_1 + (1 - x)\mu_0 = \mu_0 + x\left(\mu_1 - \mu_0\right) = \beta_0 + x\beta_1$$

- No assumptions, just rewriting! Interpretations:
  - $\beta_0 = \mu_0$: expected wait-time for POC
  - $\beta_1 = \mu_1 - \mu_0$: diff. in avg. wait times between whites and POC.
- $>$ 2 categories: dummies for all but category and everything is linear.

# Linear CEF with multiple binary covariates

- What if we have two binary covariates, $X_1$ (race) and $X_2$ (1 urban/0 rural):

$$\mu(x_1, x_2) = \begin{cases} \mu_{00} & \text{if } x_1 = 0 \text{ and } x_2 = 0 \text{ (POC, rural)} \\ \mu_{10} & \text{if } x_1 = 1 \text{ and } x_2 = 0 \text{ (white, rural)} \\ \mu_{01} & \text{if } x_1 = 0 \text{ and } x_2 = 1 \text{ (POC, urban)} \\ \mu_{11} & \text{if } x_1 = 1 \text{ and } x_2 = 1 \text{ (white, urban)} \end{cases}$$

- Can rewrite this without assumptions as a linear CEF with interaction:

$$\mu(x_1, x_2) = \beta_0 + x_1\beta_1 + x_2\beta_2 + x_1x_2\beta_3$$

- Interpretations:
    - $\beta_0 = \mu_{00}$: average wait times for rural POC.
    - $\beta_1 = \mu_{10} - \mu_{00}$: diff. in means for rural whites vs rural POC.
    - $\beta_2 = \mu_{01} - \mu_{00}$: diff. in means for urban POC vs rural POC.
    - $\beta_3 = (\mu_{11} - \mu_{01}) - (\mu_{10} - \mu_{00})$: diff. in urban racial diff. vs rural racial diff.
- Generalizes to $p$ binary variables if **all interactions included (saturated)**

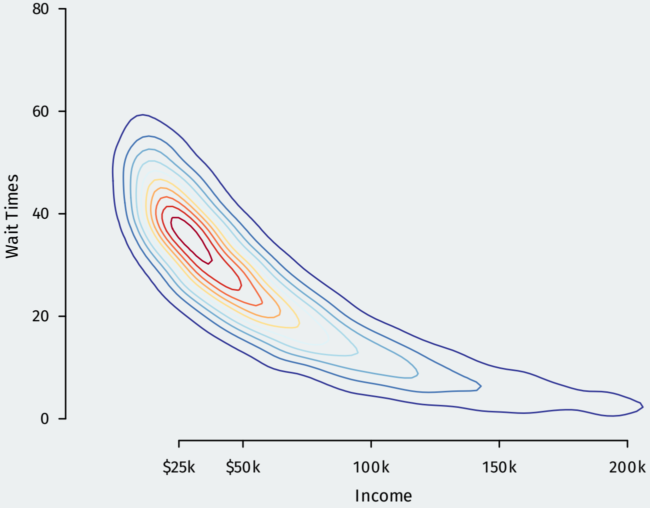
# Linear approximation

- Outside of saturated discrete settings, CEF almost never truly linear.
- Alternative goal: find **best linear predictor** of $Y$ given $X$.
- Formally, linear function of $X$ that **minimizes squared prediction errors**:

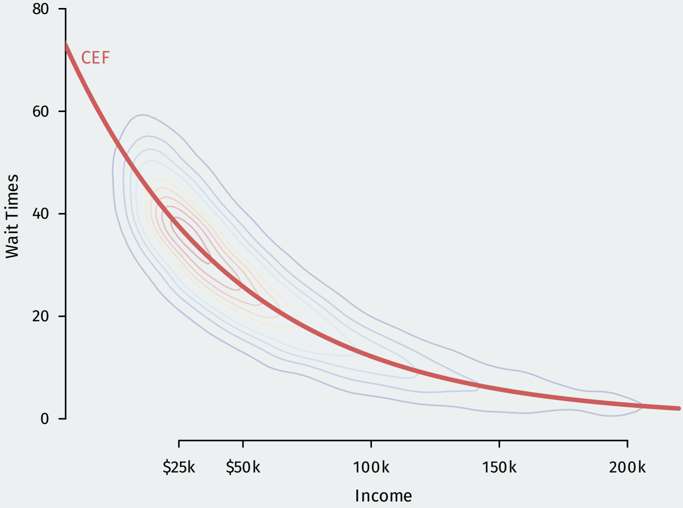
$$(\beta_0, \beta_1) = \underset{(b_0, b_1)}{\arg\min}\, \mathbb{E}[(Y - (b_0 + b_1 X))^2]$$

- $m(x) = \beta_0 + \beta_1 X$ is called the **linear projection** of $Y$ onto $X$.
  - $\beta_1 = \text{Cov}(X, Y)/\mathbb{V}[X]$
  - $\beta_0 = \mu_Y - \mu_X \beta_1$, where $\mu_Y = \mathbb{E}[Y]$ and $\mu_X = \mathbb{E}[X]$
- In general, $m(x)$ 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.

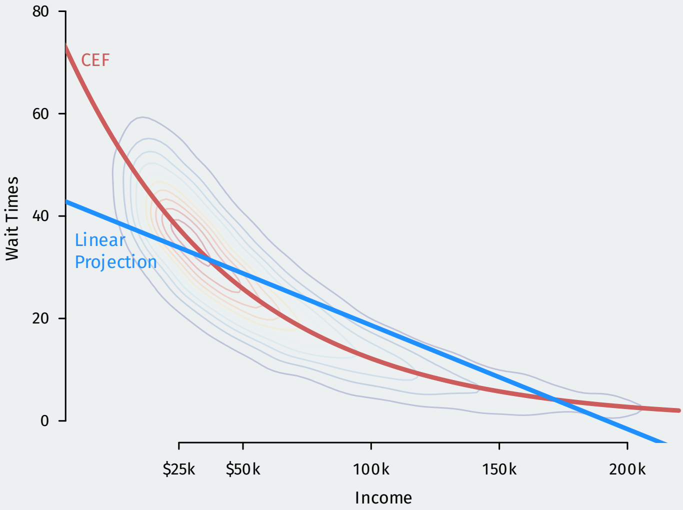
# Linear approximation



# Linear approximation



# Linear approximation



# Best linear predictor

- We'll almost always condition on a vector $\mathbf{X} = (X_1, \dots, X_k)'$:

$$m(\mathbf{x}) = m(x_1, \dots, x_k) = x_1\beta_1 + \cdots + x_k\beta_k = \mathbf{x}'\boldsymbol{\beta}$$

  - Linear predictor when $\mathbf{X} = \mathbf{x}$
- $\mathbf{X}$ is now a $k \times 1$ random vector of covariates:
  - May contain nonlinear transformations/interactions of "real" variables.
  - Typically, $X_1 = 1$ and is the intercept/constant.
- Assumptions ("Regularity conditions"):
  1. $\mathbb{E}[Y^2] < \infty$ (outcome has finite mean/variance)
  2. $\mathbb{E}\|\mathbf{X}\|^2 < \infty$ ($\mathbf{X}$ has finite means/variances/covariances)
  3. $\mathbf{Q}_{\mathbf{XX}} = \mathbb{E}[\mathbf{XX}']$ is positive definite (columns of $\mathbf{X}$ are linearly independent)

# Linear Projection

- How to find $\boldsymbol{\beta}$? Minimize squared prediction error!

$$\boldsymbol{\beta} = \arg\min_{\mathbf{b} \in \mathbb{R}^k} \mathbb{E}\left[(Y - \mathbf{X}'\mathbf{b})^2\right]$$

- After some calculus:

$$\boldsymbol{\beta} = \mathbf{Q}_{\mathbf{XX}}^{-1}\mathbf{Q}_{\mathbf{X}Y} = \left(\mathbb{E}[\mathbf{XX}']\right)^{-1}\mathbb{E}[\mathbf{X}Y]$$

  - $\mathbb{E}[\mathbf{XX}']$ is $k \times k$ and $\mathbb{E}[\mathbf{X}Y]$ is $k \times 1$

- Notes about the $m(\mathbf{x}) = \mathbf{x}'\boldsymbol{\beta}$:

  - $\boldsymbol{\beta}$ is a population quantity and possible quantity of interest.
  - Well-defined under very mild assumptions!
  - Not necessarily a conditional mean nor a causal effect!

# Projection errors

- Projection error: $e = Y - \mathbf{X}'\boldsymbol{\beta}$
- Decomposition of $Y$ into the linear projection and error: $Y = \mathbf{X}'\boldsymbol{\beta} + e$
- Properties of the projection error:
  - $\mathbb{E}[\mathbf{X}e] = 0$
  - $\mathbb{E}[e] = 0$ when $\mathbf{X}$ contains a constant.
  - Together, implies $\text{Cov}(X_j, e) = 0$ for all $j = 1, \dots, k$
- Distinct from CEF errors: $u = Y - \mu(\mathbf{X})$ which had the additional property: $\mathbb{E}[u \mid \mathbf{X}] = 0$
  - Zero conditional mean is stronger: CEF errors are 0 at every value of $\mathbf{X}$
  - $\mathbb{E}[\mathbf{X}e] = 0$ just says they are uncorrelated.

# Regression coefficients

- Sometimes useful to separate the constant:

$$Y = \beta_0 + \mathbf{X}'\boldsymbol{\beta} + e$$

  where $\mathbf{X}$ doesn't have a constant.

- Solution for $\boldsymbol{\beta}$ more interpretable here:

$$\boldsymbol{\beta} = \mathbb{V}[\mathbf{X}]^{-1}\text{Cov}(\mathbf{X}, Y), \qquad \beta_0 = \mu_Y - \boldsymbol{\mu}'_{\mathbf{X}}\boldsymbol{\beta}$$

# Interpretation of the coefficients

- Interpretation of $\beta_j$ depends on what nonlinearities are included.
- Simplest case: no polynomials or interactions.
- $\beta_j$ is the average change in predicted outcome for a one-unit change in $X_j$ holding other variables fixed.
- Let's compare:

$$m(x_1 + 1, x_2) = \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2$$
$$m(x_1, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

- Then:

$$m(x_1 + 1, x_2) - m(x_1, x_2) = \beta_1$$

- Holds for all values of $x_2$ and even if we add more variables.

# Interpretation with nonlinear terms

- What if we include a nonlinear function of one covariate?

$$m(x_1, x_1^2, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2,$$

- One-unit change in $x_1$ is more complicated:

$$m(x_1 + 1, (x_1 + 1)^2, x_2) = \beta_0 + \beta_1(x_1 + 1) + \beta_2(x_1 + 1)^2 + \beta_3 x_2$$
$$m(x_1, x_1^2, x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2,$$

- Better to think of the **marginal effect** of $X_{i1}$:

$$\frac{\partial m(x_1, x_1^2, x_2)}{\partial x_1} = \beta_1 + 2\beta_2 x_1$$

- Interpretations:
  - $\beta_1$: "effect" of $X_{i1}$ on predicted $Y_i$ when $X_{i1} = 0$ (holding $X_{i2}$ fixed)
  - $\beta_2/2$: how that "effect" changes as $X_{i1}$ changes
  - Maybe better to visualize than to interpret

# Interpretation with interactions

- What if we include an interaction between two covariates?

$$m(x_1, x_2, x_1x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- Two different **marginal effects** of interest:

$$\frac{\partial m(x_1, x_2, x_1x_2)}{\partial x_1} = \beta_1 + \beta_3 x_2,$$
$$\frac{\partial m(x_1, x_2, x_1x_2)}{\partial x_2} = \beta_2 + \beta_3 x_1$$

- Interpretations:
  - $\beta_1$: the marginal effect of $X_{i1}$ on predicted $Y_i$ when $X_{i2} = 0$.
  - $\beta_2$: the marginal effect of $X_{i2}$ on predicted $Y_i$ when $X_{i1} = 0$.
  - $\beta_3$: the change in the marginal effect of $X_{i1}$ due to a one-unit change in $X_{i2}$ **OR** the change in the marginal effect of $X_{i2}$ due to a one-unit change in $X_{i1}$.

# Partitioned Regression

$$(\alpha, \beta, \gamma) = \underset{(a,b,c)\in\mathbb{R}^3}{\arg\min}\ \mathbb{E}[(Y_i - (a + bX_i + cZ_i))^2]$$

- Can we get an expression for just $\beta$? With some tricks, yes!
- Population residuals from projection of $X_i$ on $Z_i$:

$$\widetilde{X}_i = X_i - (\delta_0 + \delta_1 Z_i) \quad \text{where} \quad (\delta_0, \delta_1) = \underset{(d_0,d_1)\in\mathbb{R}^2}{\arg\min}\ \mathbb{E}[(X_i - (d_0 + d_1 Z_i))^2]$$

  - $\widetilde{X}_i$ is now **orthogonal** to $Z_i$ so that $\text{cov}(\widetilde{X}_i, Z_i) = \mathbb{E}[\widetilde{X}_i Z_i] = 0$
- Project $Y$ onto these residuals gives $\beta$ as coefficient:

$$\beta = \frac{\text{cov}(Y_i, \widetilde{X}_i)}{\mathbb{V}[\widetilde{X}_i]}$$

  - Helps with interpretation: connects multivariate regression coefficients to simple regression coefficients.
  - The relationship captured by $\beta$ is between the outcome and the variation in $X_i$ not linearly explained by $Z_i$

# Partition regression more generally

- More general linear projection coefficients:

$$\boldsymbol{\beta} = \left(\mathbb{E}[\mathbf{X}\mathbf{X}']\right)^{-1} \mathbb{E}[\mathbf{X}Y]$$

- Let $\mathbf{X}_{i,-k}$ be the set of covariates without entry $k$.
- Now define $\widetilde{X}_{ik} = X_{ik} - m_k(\mathbf{X}_{i,-k})$
    - $m_k(\mathbf{X}_{i,-k})$ is the BLP of $X_{ik}$ on $\mathbf{X}_{i,-k}$
- Generic coefficient $\beta_k$ is:

$$\beta_k = \frac{\text{cov}(Y_i, \widetilde{X}_{ik})}{\mathbb{V}[\widetilde{X}_{ik}]}$$

# Omitted variable bias

- Consider two projections/regressions with and without some $Z$:

$$m(\mathbf{X}_i, Z_i) = \mathbf{X}_i'\boldsymbol{\beta} + Z_i\gamma, \qquad m_{-z}(\mathbf{X}_i) = \mathbf{X}_i'\boldsymbol{\delta},$$

- How do $\boldsymbol{\beta}$ and $\boldsymbol{\delta}$ relate? Use law of iterated projections:

$$\begin{aligned}
\boldsymbol{\delta} &= \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[\mathbf{X}_iY_i] \\
&= \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[\mathbf{X}_i(\mathbf{X}_i'\boldsymbol{\beta} + Z_i\gamma + e_i)] \\
&= \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\boldsymbol{\beta} + \mathbb{E}[\mathbf{X}_iZ_i]\gamma + \mathbb{E}[\mathbf{X}_ie_i]\right) \\
&= \boldsymbol{\beta} + \underbrace{\left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[\mathbf{X}_iZ_i]}_{\text{coefs from} Z\sim\mathbf{X}}\gamma
\end{aligned}$$

- Leads to the **“omitted variable bias” formula**:

$$\boldsymbol{\delta} = \boldsymbol{\beta} + \boldsymbol{\pi}\gamma, \qquad \boldsymbol{\pi} = \left(\mathbb{E}[\mathbf{X}_i\mathbf{X}_i']\right)^{-1}\mathbb{E}[\mathbf{X}_iZ_i]$$

- $\boldsymbol{\delta} - \boldsymbol{\beta} = \boldsymbol{\pi}\gamma$ is the “bias” but this is misleading.
    - $\boldsymbol{\beta}$ not necessarily “correct”, we’re just relating two projections

# Best linear approximation

- What is the relationship between $m(\mathbf{X})$ and $\mu(\mathbf{X}) = \mathbb{E}[Y \mid \mathbf{X}]$?
  - If $\mu(\mathbf{X})$ is linear, then $\mu(\mathbf{X}) = m(\mathbf{X}) = \mathbf{X}'\boldsymbol{\beta}$.
  - But $\mu(\mathbf{X})$ could be nonlinear, what then?
- Linear projection justification: best linear approximation to $\mu(\mathbf{X})$:

$$\boldsymbol{\beta} = \underset{\mathbf{b}\in\mathbb{R}^K}{\arg\min}\, \mathbb{E}\left[(\mu(\mathbf{X}) - \mathbf{X}'\boldsymbol{\beta})^2\right]$$

- Linear projection is best linear approximation to $Y$ and $\mathbb{E}[Y \mid X]$.
- Limitations:
  - If nonlinearity of $\mu(\mathbf{X})$ is severe, $m(\mathbf{X})$ can only be so good.
  - $m(\mathbf{X})$ can be sensitive to the marginal distribution of $\mathbf{X}$.

# Recap

$$Y = \mathbf{X}'\boldsymbol{\beta} + e$$

- "The Linear Model": is this an assumption?
- Depends on what we assume about the error, $e$
  - If $\mathbb{E}[e \mid \mathbf{X}] = 0$, then we are assuming the CEF is linear, $\mathbb{E}[Y \mid X] = \mathbf{X}'\boldsymbol{\beta}$
  - If just $\mathbb{E}[\mathbf{X}e] = 0$, then this is just a linear projection.
  - First is very strong, second is very mild.
- Why do we care? Affects the properties of OLS.
  - Some finite-sample properties of OLS (unbiasedness) require linear CEF
  - Asymptotic results (consistency, asymptotic normality) apply to both.
  - OLS will consitently estimate something, but maybe not what you want.