Topic 2: Regression Toolkit

ECON 5783 — University of Arkansas

Prof. Kyle Butts

September 2025

Regression Toolkit

We learn:

→ What is the "Conditional Expectation Function (CEF)"

→ Curse of dimensionality and why we need models

→ Modelling choices

→ What is Omitted Variable Bias and why is it useful for thinking of causal effects

→ Deeper understanding of regression: FWL Theorem

Readings to complement this lecture are:

→ Angrist and Pischke (2009) *Mostly Harmless Econometrics*, Chatper 3: Intro, section 3.1, and section 3.2

Linear Regression Bootcamp

This set of slides will serve as a 'bootcamp' into one of the most popular tools in the applied researcher's toolkit: linear regression

$\rightarrow$ Creates a simple and interpretable model of y

$\rightarrow$ Has desirable properties for causal inference even if the outcome is not linear in covariates

Conditional Expectation Function

- Conditional Expectation Function

Modeling the CEF

- Linear Model of Conditional Expectation Function
- Making models more flexible
- More Flexible Approximations (`binscatter`)

Omitted Variable Bias (OVB)

- Reinterpreting selection bias as OVB

Frisch-Waugh-Lovell Theorem

Conditional Expectation Function

- Conditional Expectation Function

Modeling the CEF

- Linear Model of Conditional Expectation Function
- Making models more flexible
- More Flexible Approximations (`binscatter`)

Omitted Variable Bias (OVB)

- Reinterpreting selection bias as OVB

Frisch-Waugh-Lovell Theorem

The Conditional Expectation Function

In particular, we will think a lot about the Conditional Expectation Function (CEF) of y_i given $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})'$:

$$m(\boldsymbol{x}) \equiv \mathbb{E}[y_i \mid \boldsymbol{X}_i = \boldsymbol{x}]$$

$\rightarrow$ This reads "$m(\boldsymbol{x})$ is the expected value of y_i conditional on the unit having $\boldsymbol{X}_i = \boldsymbol{x}$"

The Conditional Expectation Function

In particular, we will think a lot about the Conditional Expectation Function (CEF) of y_i given $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})'$:

$$m(\boldsymbol{x}) \equiv \mathbb{E}[y_i \mid \boldsymbol{X}_i = \boldsymbol{x}]$$

$\rightarrow$ This reads "$m(\boldsymbol{x})$ is the expected value of y_i conditional on the unit having $\boldsymbol{X}_i = \boldsymbol{x}$"

The easiest way to estimate this for a given $\boldsymbol{x}$ is to average y_i for units with $\boldsymbol{X}_i = \boldsymbol{x}$.

The Conditional Expectation Function

In particular, we will think a lot about the Conditional Expectation Function (CEF) of y_i given $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{ip})'$:

$$m(\boldsymbol{x}) \equiv \mathbb{E}[y_i \mid \boldsymbol{X}_i = \boldsymbol{x}]$$

$\rightarrow$ This reads "$m(\boldsymbol{x})$ is the expected value of y_i conditional on the unit having $\boldsymbol{X}_i = \boldsymbol{x}$"

The easiest way to estimate this for a given $\boldsymbol{x}$ is to average y_i for units with $\boldsymbol{X}_i = \boldsymbol{x}$.

$\rightarrow$ Only uses observations with $\boldsymbol{X}_i = \boldsymbol{x}$ (or $\boldsymbol{X}_i \approx \boldsymbol{x}$ when $\boldsymbol{X}_i$ is continuous), so that is the relavent 'n' when considering sample size

Estimation Procedure for CEF

For each value of $\boldsymbol{x}$,

$\rightarrow$ Subset population to only the units with $\boldsymbol{X}_i = \boldsymbol{x}$

$\rightarrow$ Average value of y_i for those units

This is the value $m(\boldsymbol{x})$ for that given $\boldsymbol{x}$

Example with Discrete Variable

Say, we have w_i as wages and D_i is an indicator for college attendance. Then, we can estimate the CEF of wages conditional on college attendance as

$$g(1) = \mathbb{E}[y_i \mid D_i = 1] \quad \text{and} \quad g(0) = \mathbb{E}[y_i \mid D_i = 0]$$

This is just just the average wage for college attendees and non-attendees

Regression Framework

To estimate these given a sample of workers, we could regress

$$w_i = \alpha + D_i\tau + u_i$$

$\rightarrow$ $\hat{\alpha}$ is our estimate of $g(0)$ and $\hat{\alpha} + \hat{\tau}$ is our estimate of $g(1)$

Example with two discrete variables

Now, we have w_i as wages and $\boldsymbol{X}_i$ consists of an indicator for college attendence, D_i, and F_i is an indicator for being a female.

Now we have four distinct values of $\boldsymbol{X}_i$, $(D_i, F_i) \in \{(0,0), (1,0), (0,1), (0,0)\}$.

$\rightarrow$ Estimating the CEF would consist of just four sub-sample averages (e.g. female college attendees)

Regression Framework

Note, it is *not* enough to regress wages on an indicator for college and an indicator for female. We need to include an interaction term as well!

$$w_i = \alpha + D_i\beta_1 + F_i\beta_2 + (D_i \cdot F_i)\beta_3 + u_i$$

$\rightarrow$ Without the interaction term, we are assuming the effect of college is the same for female and male workers

Example with single continuous variable

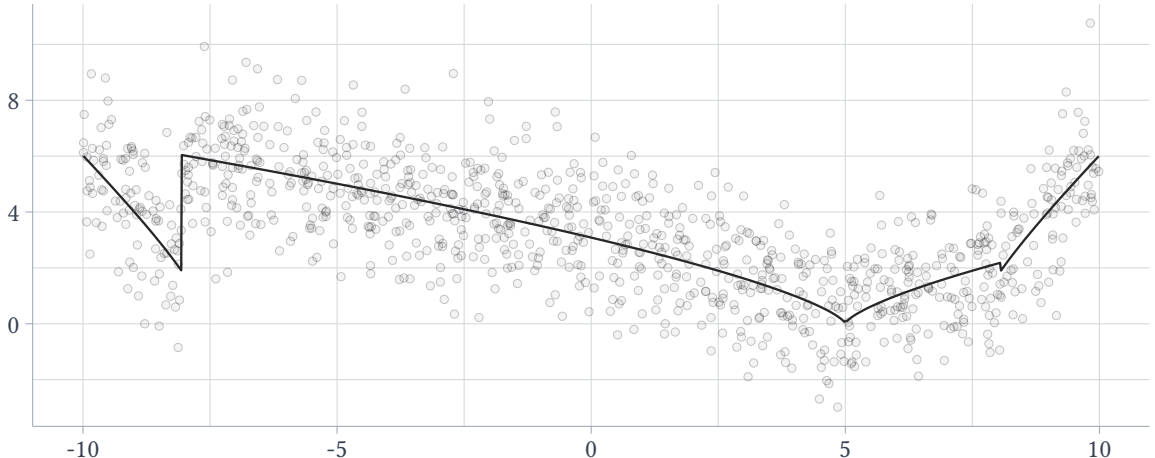
If we move from a discrete to a single continuous X_i, we can still estimate the CEF by averaging y_i for individuals with $X_i \approx x$

$\rightarrow$ Only now, we have a problem where there are infinitely many values of x and we have to decide how close is 'close enough' to x

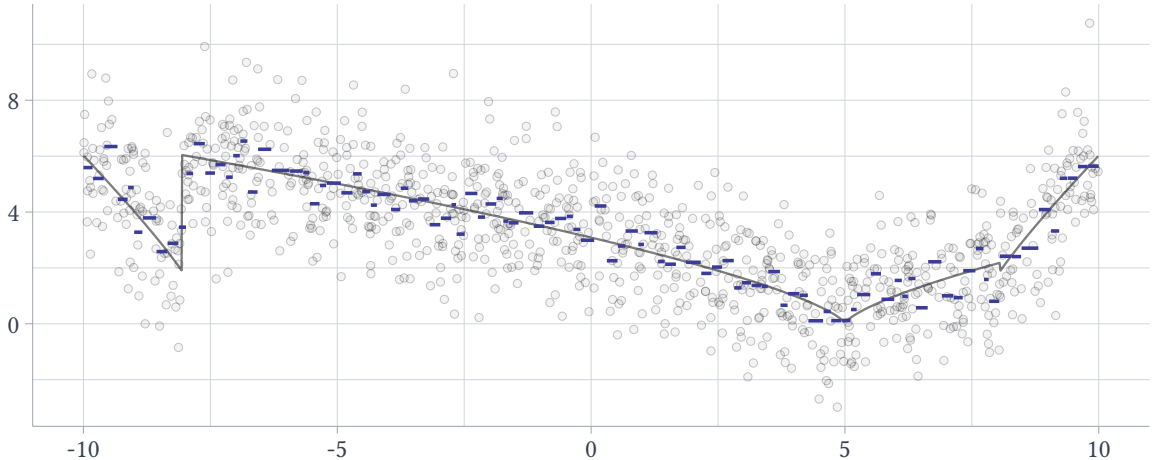
We will estimate by making many small bins of x along the full range of X.

$\rightarrow$ This is called 'non-parametric' estimation in the statistics literature

True $g(x)$



True $g(x)$; Approximate Conditional Expectation Function



Single continuous and single discrete variable

If we have a continuous X_{1i} and a discrete D_i variable, we can do:

$\rightarrow$ Average value of y for bins of X_i *separately* for $D_i = 0$ and $D_i = 1$.

Estimation of the CEF

As we discussed before, we could estimate $g(x) \equiv \mathbb{E}[y_i \mid \boldsymbol{X}_i = x]$ by averaging over individuals with $\boldsymbol{X}_i = x$

$\rightarrow$ In the case where $\boldsymbol{X}_i$ is a discrete variable taking values $x_1, \ldots, x_L$, this is just sub-sample averages for $\boldsymbol{X}_i = x_\ell$

$\rightarrow$ With $\boldsymbol{X}_i$ as a continuous variable, we do a bunch of 'binned' averages.

As we start adding variables, we have to start 'interacting' the variables, creating many many different sub-samples

The “curse of dimensionality”

When $\boldsymbol{X}_i$ is a multi-dimensional vector with many continuous variables, we end up with *a lot* of subsamples we want to take averages of.

The density around any particular value $\boldsymbol{x}$ is typically going to be small or near-zero

$\rightarrow$ In samples, our estimates will be very noisy or even impossible to calculate for given values of $\boldsymbol{x}$

The Conditional Expectation Function

Uses of the conditional expectation function:

1. Descriptive: how y on average changes as X changes
   - By definition, compare $g(x_1)$ to $g(x_2)$

2. Prediction: if we know $\boldsymbol{X}_i$, our best guess for y_i is $m(\boldsymbol{X}_i)$

3. Causal inference: what happens to y_i if we *manipulate* $\boldsymbol{X}_i$
   - Sometimes, the point of this class is to describe when!

Preview of Conditional Expectation Function usages

One main reason why we care about modeling Y is because causal inference is a missing data problem

$\rightarrow$ For the treated units, we do not observe what their outcomes would be in the absence of treatment, $Y(0)$

$\rightarrow$ For the control units, we do not observe their $Y(1)$

Preview of Conditional Expectation Function usages

One main reason why we care about modeling Y is because causal inference is a missing data problem

$\rightarrow$ For the treated units, we do not observe what their outcomes would be in the absence of treatment, $Y(0)$

$\rightarrow$ For the control units, we do not observe their $Y(1)$

If we fit a model for $\mathbb{E}[Y_i(0) \mid \boldsymbol{X}_i = x]$, we can use this to make predictions for the treated units

Preview of Conditional Expectation Function usages

One main reason why we care about modeling Y is because causal inference is a missing data problem

$\rightarrow$ For the treated units, we do not observe what their outcomes would be in the absence of treatment, $Y(0)$

$\rightarrow$ For the control units, we do not observe their $Y(1)$

If we fit a model for $\mathbb{E}[Y_i(0) \mid \boldsymbol{X}_i = x]$, we can use this to make predictions for the treated units

$\rightarrow$ The model predicting out of sample for our treated group requires certain conditions discussed in topic 3

Modelling the CEF

We have an outcome variable y and a set of p different predictor variables $X = (X_1, X_2, \ldots, X_p)$.

$\rightarrow$ For some observations we observe both X and y; this is essential to fit a model

We can write our model in a general form as

$$y = f(X) + \varepsilon,$$

where f is some unknown (but fixed) function of X.

Estimation of the CEF

As we discussed before, we could estimate $m(\boldsymbol{x}) \equiv \mathbb{E}[y_i \mid \boldsymbol{X}_i = \boldsymbol{x}]$ by averaging over individuals with $\boldsymbol{X}_i = \boldsymbol{x}$

But as $\boldsymbol{X}_i$ has more variables, "curse of dimensionality" makes this procedure infeasible

$\rightarrow$ We need to make assumptions on the shape of $g(x)$ to make estimation feasible

Linear Model

It is common to propose a *parametric* model of the conditional expectation function:

$$y_i = \boldsymbol{X}_i'\beta + \text{error} = \sum_{k=1}^{p} X_{i,k}\beta_k + \text{error}$$

$\rightarrow$ We model y as a linear function of the covariates

$\rightarrow$ Assume that one of the variables in $\boldsymbol{X}_i$ is an intercept term

Fitting via Ordinary Least Squares

"Fitting" the linear model involves selecting β to make out model the "best" linear predictor of y:

$$\hat{\beta}_{\texttt{OLS}} \equiv \underset{\beta}{\text{argmin}}\ \mathbb{E}\Big[\big(y_i - \boldsymbol{X}_i'\beta\big)^2\Big]$$

We can optimize this by taking first-order conditions and set equal to zero:

$$\begin{aligned}
&\mathbb{E}\big[\boldsymbol{X}_i\left(y_i - \boldsymbol{X}_i'\beta_{\texttt{OLS}}\right)\big] = 0 \\
\implies\ &\mathbb{E}[\boldsymbol{X}_i y_i] - \mathbb{E}\big[\boldsymbol{X}_i\boldsymbol{X}_i'\big]\beta_{\texttt{OLS}} = 0 \\
\implies\ &\beta_{\texttt{OLS}} = \left(\mathbb{E}\big[\boldsymbol{X}_i\boldsymbol{X}_i'\big]\right)^{-1}\mathbb{E}[\boldsymbol{X}_i y_i]
\end{aligned}$$

Ordinary Least Squares Estimator

We can estimate using a sample of observations:

$$\hat{\beta}_{\mathsf{OLS}} = \left(\sum_{i=1}^{n} \boldsymbol{X}_i \boldsymbol{X}_i' \right)^{-1} \sum_{i=1}^{n} \boldsymbol{X}_i y_i$$

Or in matrix notation

$$\hat{\beta}_{\mathsf{OLS}} = (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{X}'\boldsymbol{y}$$

$\rightarrow$ $\boldsymbol{X}$ is the $n \times k$ matrix with row given by $\boldsymbol{X}_i'$ and $\boldsymbol{y}$ is the column vector of outcome variables

Linearity

When is a linear model of $g(x) \equiv \mathbb{E}[y_i \mid \boldsymbol{X}_i = \boldsymbol{x}]$ a good assumption?

$\rightarrow$ In some cases, the data might look to grow linearly in each $X_{i,k}$, in which case, it is a reasonable assumption

But many times, the linear model will fall short. For example:

$\rightarrow$ wages seem to grow quadratically in age

$\rightarrow$ Square footage of homes seems to have diminishing returns to price

$\rightarrow$ Returns to work experience depend on whether you have a college degree (interaction)

“Extending” linear models

We can, of course, increase the performance of our model by doing things like polynomials of variables and interaction between terms.

It becomes necessary to distinguish between the variables you are using, $\boldsymbol{X}_i$ and the terms you include in your model:

$$\boldsymbol{W}_i = (g_1(\boldsymbol{X}_i), \ldots, g_K(\boldsymbol{X}_i))'$$

$\rightarrow$ When $g_1(\boldsymbol{X}_i) = x_{1,i}, \ldots, g_p(\boldsymbol{X}_i) = X_{i,p}$ we end up with the original model

$\rightarrow$ But, of course, we can include polynomials and/or interactions as well

A ‘correctly specified’ model

We say a *linear model* is correctly specified if the CEF is exactly equal to the model we are estimating:

$$m(\boldsymbol{x}) \equiv \mathbb{E}[Y_i \mid \boldsymbol{X}_i = \boldsymbol{x}] = \boldsymbol{w}\gamma_0$$

That is, the true conditional expectation is linear our set of terms $\boldsymbol{W}_i$

$\rightarrow$ We have ‘fully used’ the information in $\boldsymbol{X}_i$ with our terms $\boldsymbol{W}_i$

Example: Discrete variables

When $\boldsymbol{X}_i$ is a discrete variable taking values $x_1, \ldots, x_L$, consider a linear model consisting of a set of indicator variables for each value of x_ℓ:

$$y_i = \sum_{\ell=1}^{L} \mathbb{1}[\boldsymbol{X}_i = x_\ell]\beta_\ell + u_i \tag{1}$$

Example: Discrete variables

When $\boldsymbol{X}_i$ is a discrete variable taking values $x_1, \ldots, x_L$, consider a linear model consisting of a set of indicator variables for each value of x_ℓ:

$$y_i = \sum_{\ell=1}^{L} \mathbb{1}[\boldsymbol{X}_i = x_\ell]\beta_\ell + u_i \tag{1}$$

$\rightarrow$ The ordinary least-squares estimator estimates $\hat{\beta}_\ell = \hat{\mathbb{E}}[y_i \mid \boldsymbol{X}_i = x_\ell]$

Example: Discrete variables

When $\boldsymbol{X}_i$ is a discrete variable taking values $x_1, \ldots, x_L$, consider a linear model consisting of a set of indicator variables for each value of x_ℓ:

$$y_i = \sum_{\ell=1}^{L} \mathbb{1}[\boldsymbol{X}_i = x_\ell]\beta_\ell + u_i \tag{1}$$

$\rightarrow$ The ordinary least-squares estimator estimates $\hat{\beta}_\ell = \hat{\mathbb{E}}[y_i \mid \boldsymbol{X}_i = x_\ell]$

$\rightarrow$ In this case, the CEF is *correctly specified* as the linear model (1).

Example: Discrete variables

As mentioned before, if we have multiple discrete variables, we need to include interaction terms as well to ensure we have a correctly specified model

$\rightarrow$ If we do not include interaction terms, we need the true coefficients on the interactions to be 0!

Ommitted Categories

When we include a constant in the regression (or have multiple sets of indicator variables) we have issues of multi-collinearity:

$$y_i = \alpha + \sum_{\ell=2}^{L} \mathbb{1}[\boldsymbol{X}_i = x_\ell]\beta_\ell + u_i$$

We need to drop (at least) one of the indicator variables (say $\mathbb{1}[\boldsymbol{X}_i = x_1]$). This serves as the "reference category"

$$\hat{\beta}_\ell = \hat{\mathbb{E}}[y_i \mid \boldsymbol{X}_i = x_\ell] - \hat{\mathbb{E}}[y_i \mid \boldsymbol{X}_i = x_1]$$

$\hat{\beta}_\ell$ is the mean of group ℓ relative to the omitted group

Prediction model

We have an outcome variable y and a set of p different predictor variables $X = (X_1, X_2, \ldots, X_p)$.

$\rightarrow$ For some observations we observe both X and y; this is essential to fit the model

We can write the model in a general form as

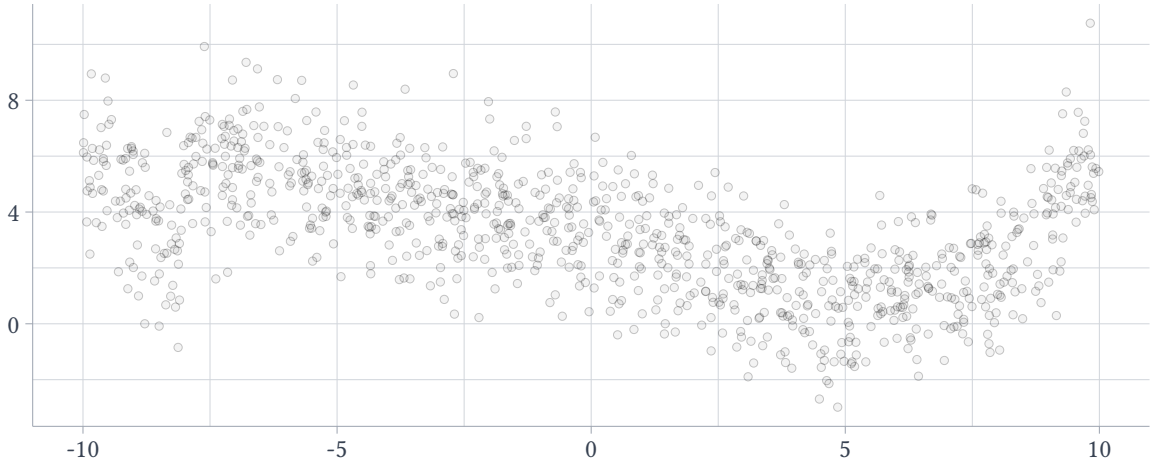
$$y = f(X) + \varepsilon,$$

where f is some unknown (but fixed) function of X. By definition $\varepsilon \equiv y - f(X)$ is the error term that is needed to fit the data perfectly

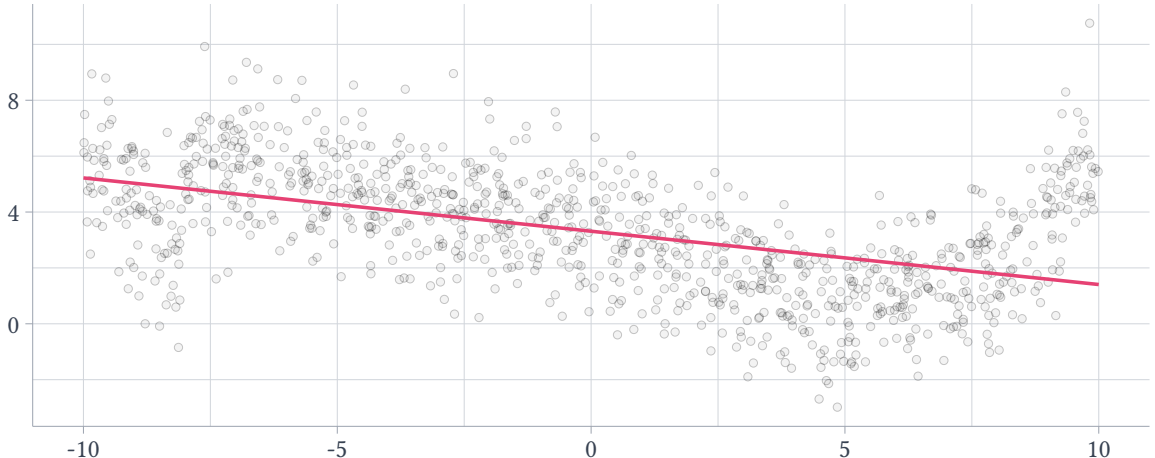
Prediction *model*

There are many different possible models of f ranging from a linear model; a 'smooth' model (polynomial or other); or a fully non-parametric function

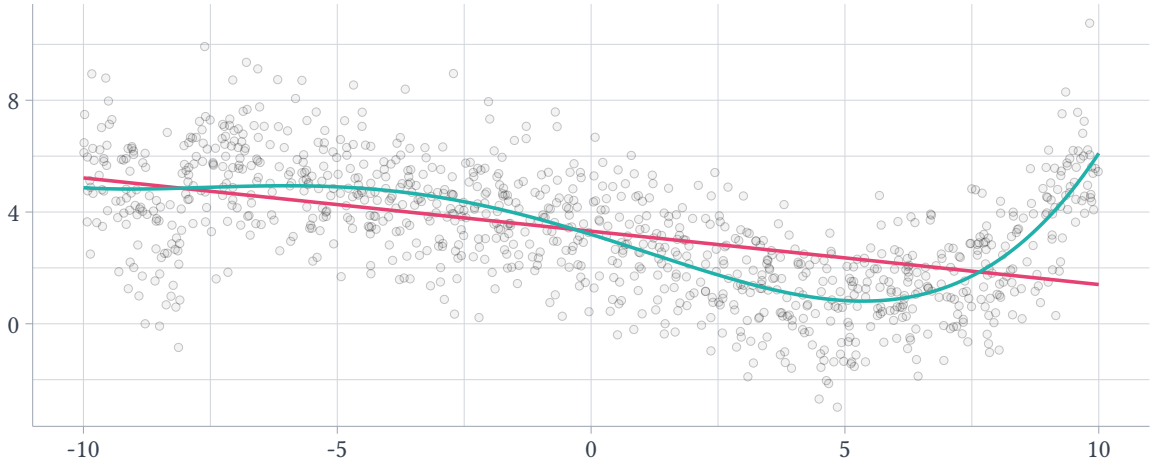
Examples of f:



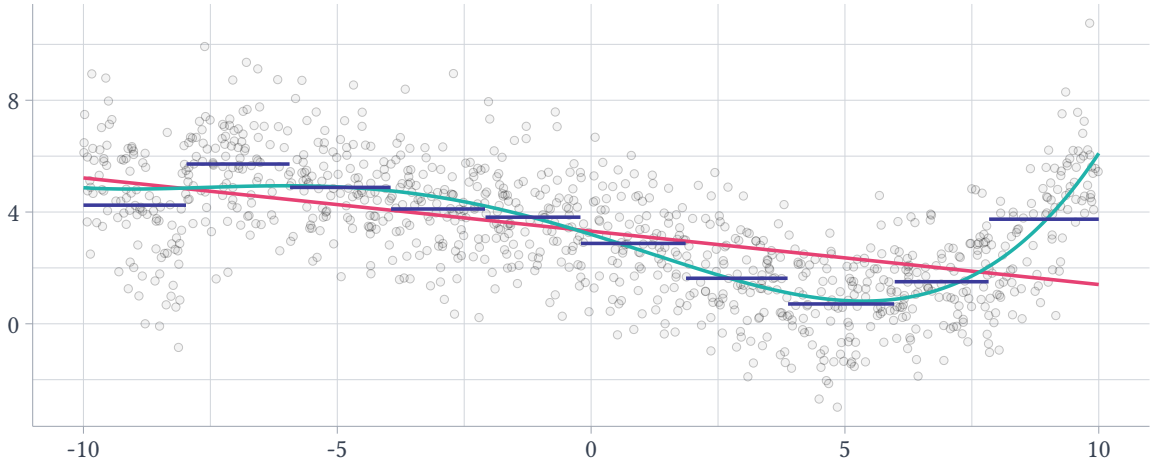
Examples of f: Line



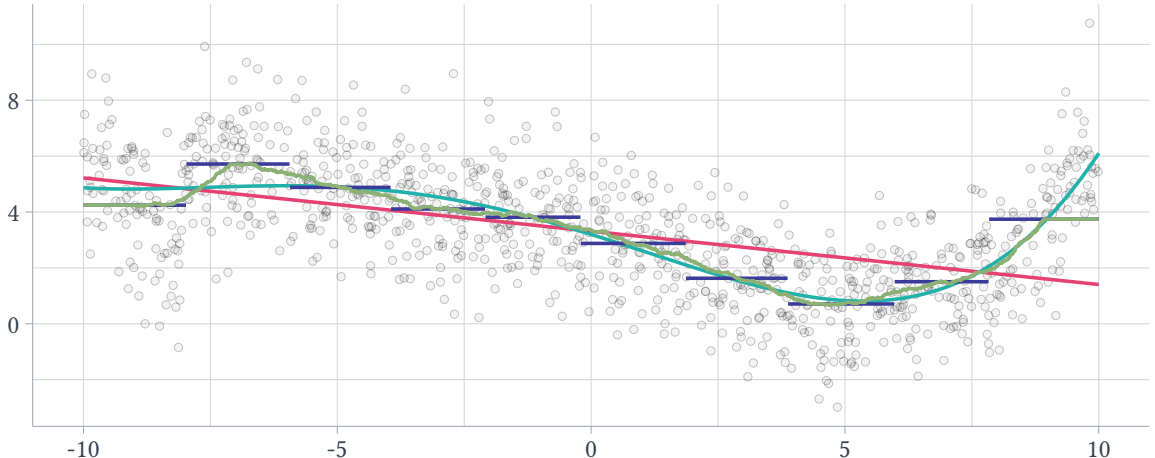
Examples of f: Line, Polynomial (x^4)



Examples of f: Line, Polynomial $\left(x^4\right)$, Bins of x



Examples of f: Line, Polynomial $\left(x^4\right)$, Bins of x, KNN of x



Prediction model

There are many different possible models of f ranging from a linear model; a 'smooth' model (polynomial or other); or a fully non-parametric function

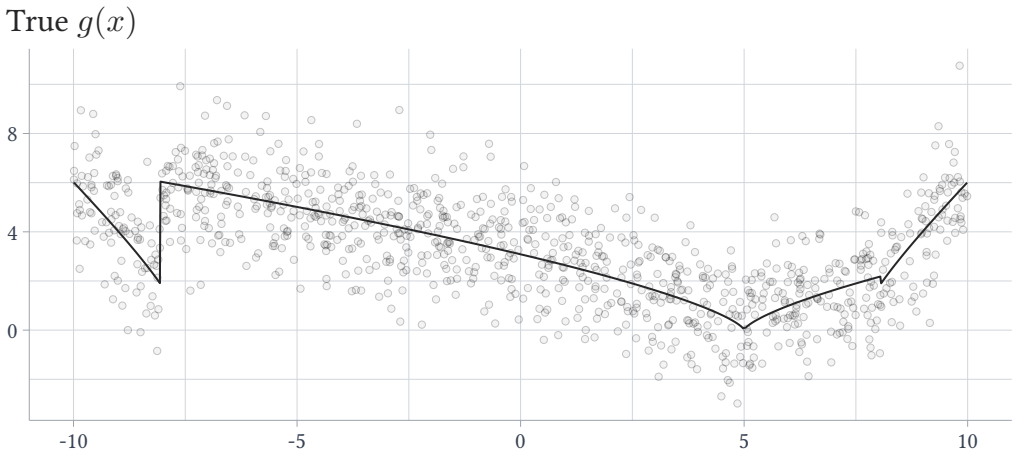
The more 'fancy' a model:

$\rightarrow$ The more flexible the relationship between y and X can be

$\rightarrow$ The larger the risk of overfitting the data

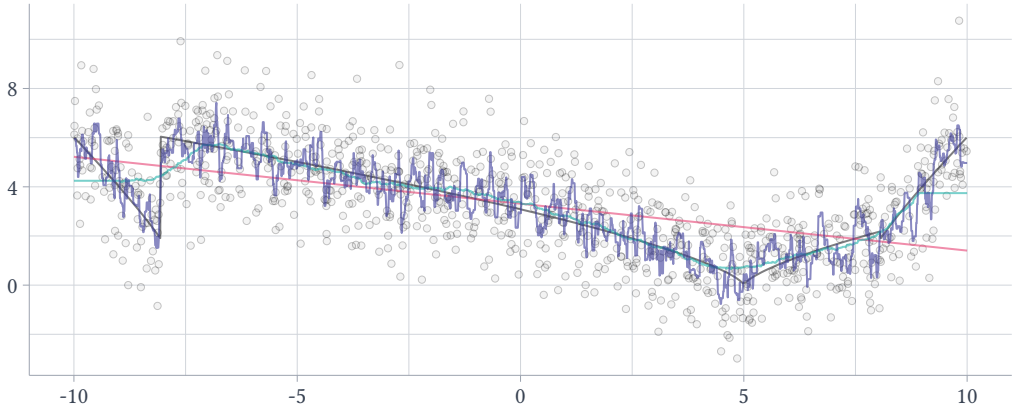
$\rightarrow$ The less interpretable the model becomes

Flexibility vs. Overfitting



Flexibility vs. Overfitting

True $f(x)$, Line, Somewhat flexible, Highly flexible



Flexibility vs. Overfitting

By making the model more and more *flexible*, you risk overfitting more and more

→ A solution is to evaluate your model fit using outside 'testing data' (hold out some observations from fitting the model)

Flexibility vs. Overfitting

By making the model more and more *flexible*, you risk overfitting more and more

→ A solution is to evaluate your model fit using outside 'testing data' (hold out some observations from fitting the model)

This technique is not as common when you care more about the associations between variables (interpreting the model)

→ Not really a good reason other than "that is more complicated"

Partially linear model

The Partially linear model mixes high model flexibility in a key variable we care about and linear model for the rest of the covariates:

$$y_i = \mu(X_i) + \boldsymbol{Z}_i'\beta + u_i$$

$\rightarrow$ $\mu(X_i)$ is a highly flexible function

$\rightarrow$ $\boldsymbol{Z}_i'$ is a set of *linear* control variables

This allows you to prevent the curse of dimensionality by linearly controlling for most of the variables. Allows a flexible model for the key variable of interest, $\boldsymbol{X}_i$, that is good for graphing

Partially linear model

$$y_i = \mu(X_i) + \boldsymbol{Z}_i'\beta + u_i$$

One recent way of estimating this is using a 'binscatter' regression

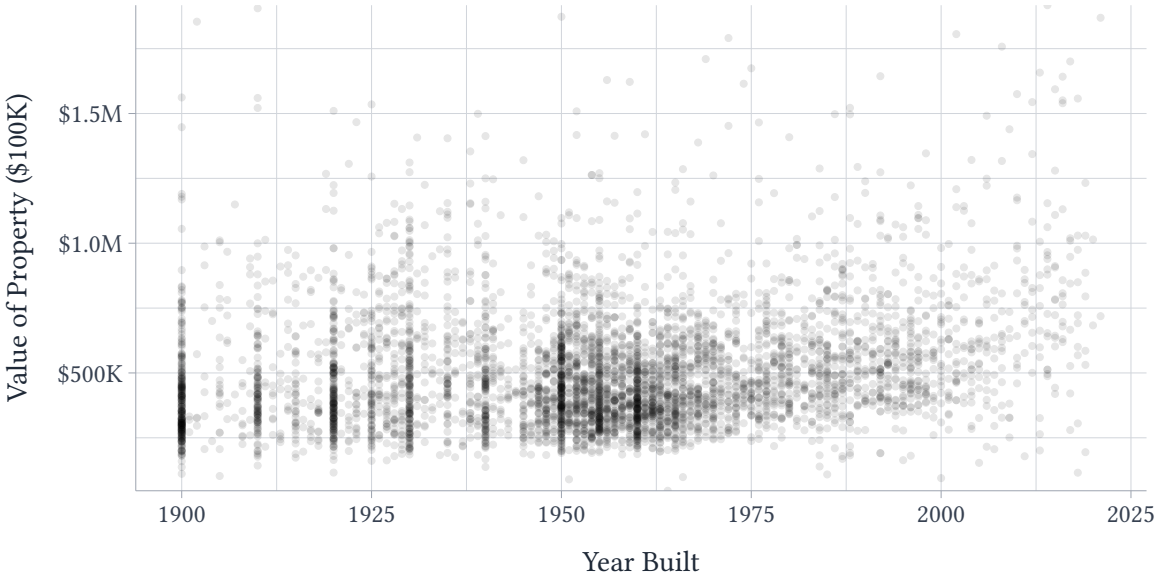
→ Popularized by Raj Chetty and coauthors since they had millions of observations (too many for normal scatterplots)

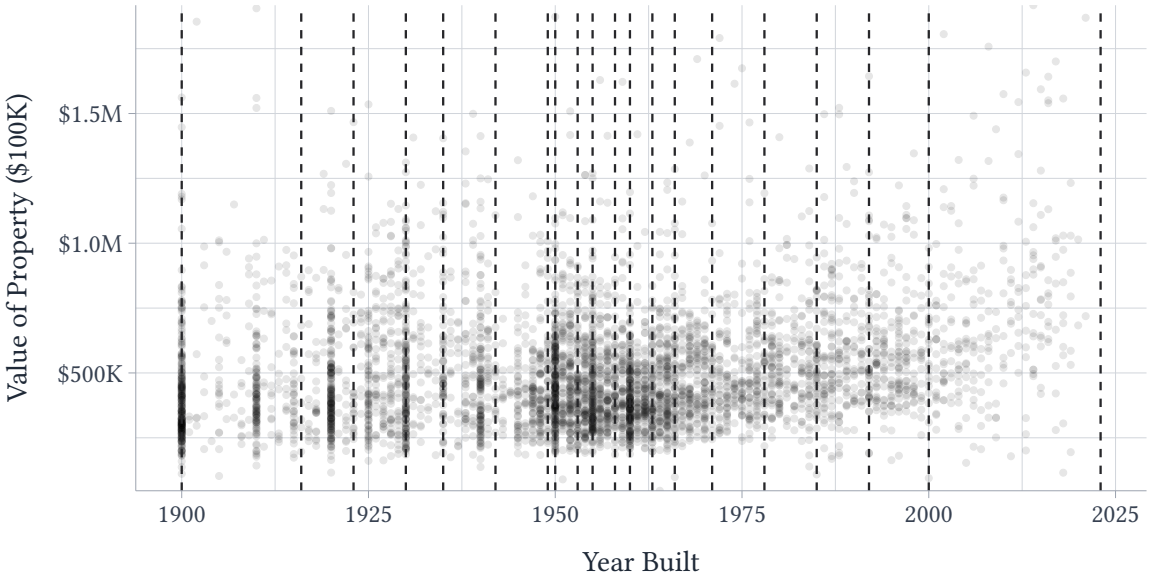
Binscatter Regression

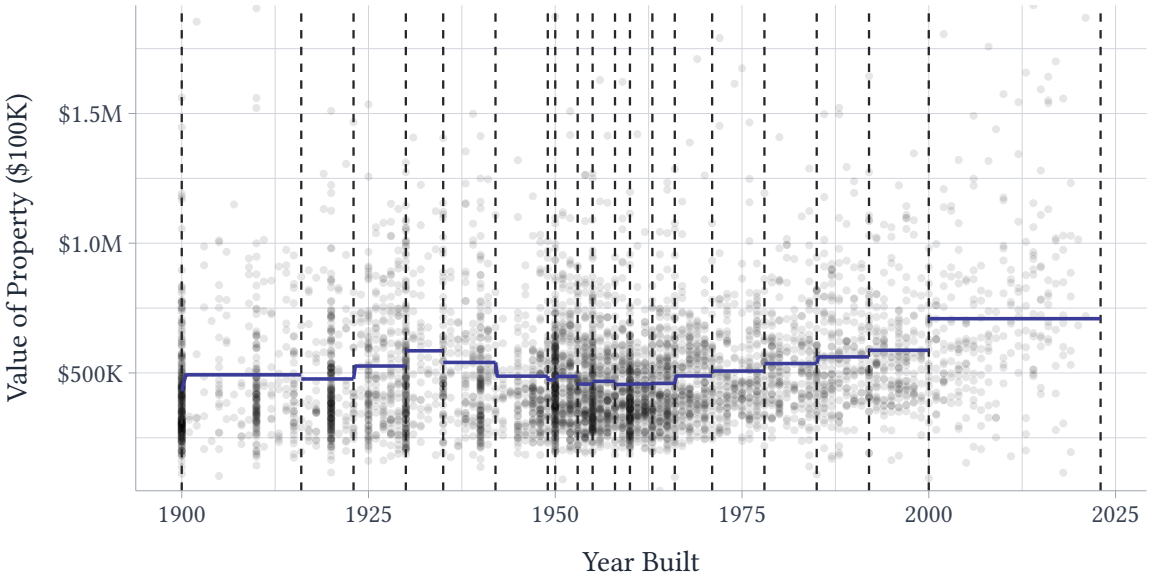
$$y_i = \mu(X_i) + \boldsymbol{Z}_i'\beta + u_i$$

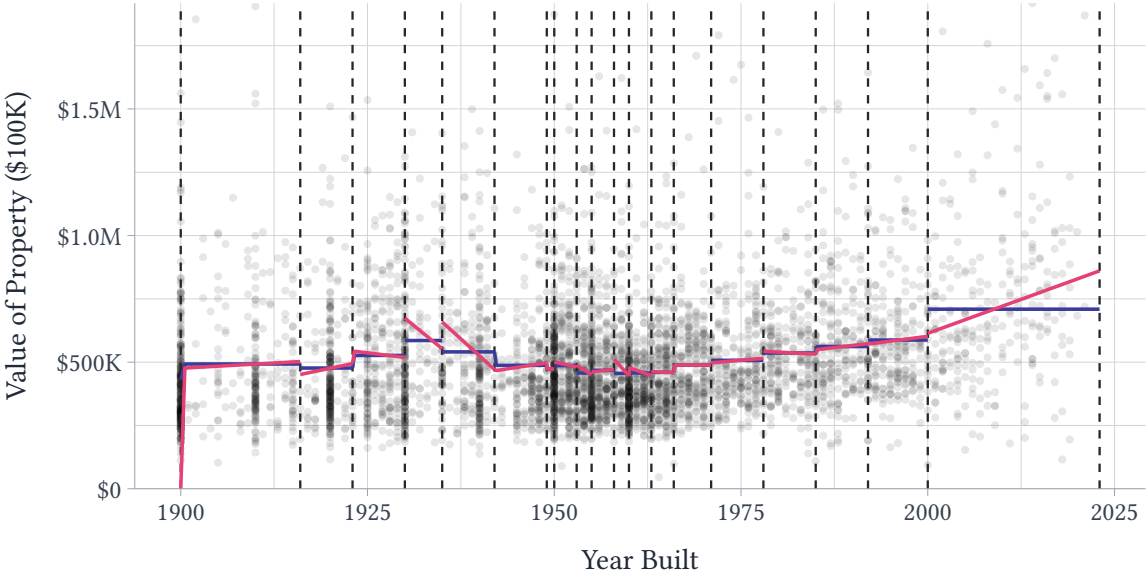
One recent way of estimating this is using a 'binscatter' regression:

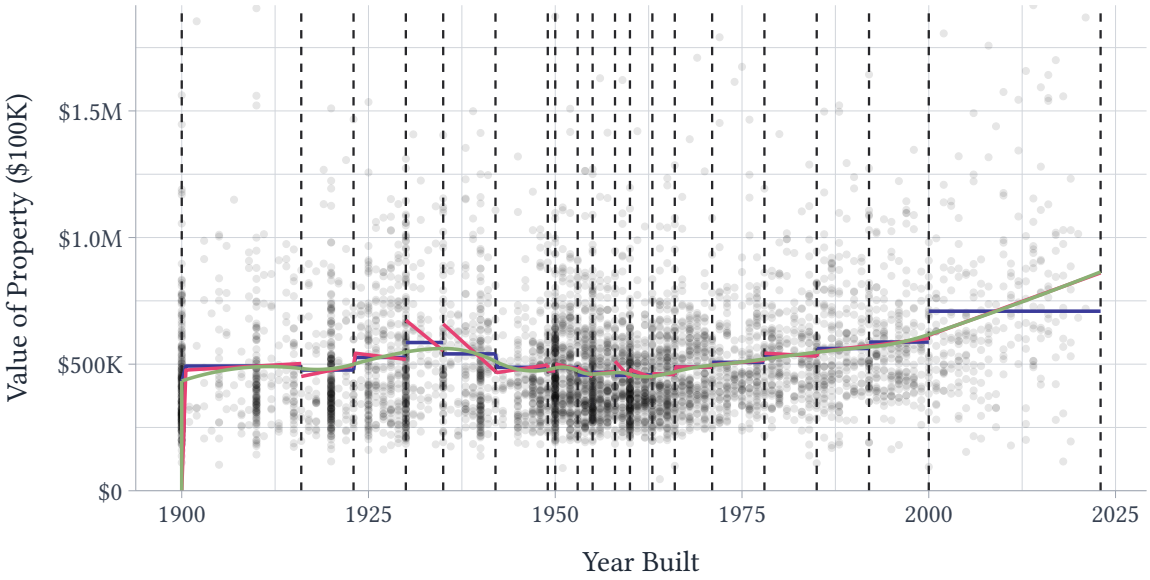
1. Chop X_i variable into J bins with an equal number of observations into each bin
2. Fit some polynomial of X_i just within each bin (interact X_i polynomial with bin indicators)


Value of Property ($100K)
$1.5M
$1.0M
$500K
1900
1925
1950
1975
2000
2025
Year Built


Value of Property ($100K)
$1.5M
$1.0M
$500K
1900
1925
1950
1975
2000
2025
Year Built


Value of Property ($100K)
$1.5M
$1.0M
$500K
1900
1925
1950
1975
2000
2025
Year Built


Value of Property ($100K)
$1.5M
$1.0M
$500K
$0
1900
1925
1950
1975
2000
2025
Year Built


Value of Property ($100K)
$1.5M
$1.0M
$500K
$0
1900
1925
1950
1975
2000
2025
Year Built

Conditional Expectation Function

- Conditional Expectation Function

Modeling the CEF

- Linear Model of Conditional Expectation Function
- Making models more flexible
- More Flexible Approximations (`binscatter`)

Omitted Variable Bias (OVB)

- Reinterpreting selection bias as OVB

Frisch-Waugh-Lovell Theorem

Difference between true model and model we estimate

Say there is a true causal model for y

$$y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i$$

$\rightarrow$ Assume $\mathbb{E}[\varepsilon_i \mid \boldsymbol{X}_i] = 0$ so that β_1 is the true causal effect

Difference between true model and model we estimate

Say there is a true causal model for y

$$y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i$$

$\rightarrow$ Assume $\mathbb{E}[\varepsilon_i \mid \boldsymbol{X}_i] = 0$ so that β_1 is the true causal effect

But we only estimate a 'short' regression specification

$$y_i = \delta_0 + X_{i1}\delta_1 + error$$

What is the relationship between β_1 the true causal effect and the coefficient δ_1?

Omitted Variable Bias

$$\underbrace{y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i}_{\text{“long regression”}} \quad \text{and} \quad \underbrace{y_i = \delta_0 + X_{i1}\delta_1 + \text{error}_i}_{\text{“short regression”}}$$

We have the following relationship:

$$\begin{aligned}
\delta_1 &= \frac{\text{Cov}(X_1, y)}{\text{Var}\,(X_1)} \\
&= \frac{\text{Cov}(X_1, \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon)}{\text{Var}\,(X_1)}
\end{aligned}$$

Omitted Variable Bias

$$\underbrace{y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i}_{\text{“long regression”}} \quad \text{and} \quad \underbrace{y_i = \delta_0 + X_{i1}\delta_1 + \text{error}_i}_{\text{“short regression”}}$$

We have the following relationship:

$$\begin{aligned}
\delta_1 &= \frac{\text{Cov}(X_1, y)}{\text{Var}\,(X_1)} \\
&= \frac{\text{Cov}(X_1, \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon)}{\text{Var}\,(X_1)} \\
&= \beta_1 + \beta_2 \frac{\text{Cov}\,(X_1, X_2)}{\text{Var}\,(X_1)}
\end{aligned}$$

Omitted Variable Bias

$$\hat{\delta}_1 = \beta_1 + \beta_2 \frac{\text{Cov}\,(X_1, X_2)}{\text{Var}\,(X_1)}$$

The reason this is true is due to regression being a prediction model!

$\rightarrow$ If X_1 and X_2 are correlated, then knowing about X_1 tells me information on X_2

$\rightarrow$ I would want to use that implicit information on X_2 to predict y as well!

$\Longrightarrow$ take the effect of β_2 times how I think X_1 tells me about X_2

Omitted Variable Bias

$$\hat{\delta}_1 = \beta_1 + \beta_2 \frac{\text{Cov}\left(X_1, X_2\right)}{\text{Var}\left(X_1\right)}$$

We can often times 'sign' the bias:

$\rightarrow$ The sign of β_2 is what we think the effect of X_2 is on y

$\rightarrow$ $\text{Cov}\left(X_1, X_2\right)$ is how X_1 and X_2 are related in the population

Signing the Bias

	$\text{Cov}(X_1, X_2) > 0$	$\text{Cov}(X_1, X_2) < 0$	$\text{Cov}(X_1, X_2) = 0$
$\beta_2 > 0$	positive bias	negative bias	no bias
$\beta_2 < 0$	negative bias	positive bias	no bias
$\beta_2 = 0$	no bias	no bias	no bias

If X_2 is unrelated to X_1 *or* X_2 has no effect on y, then we have no problem

Omitted Variable Bias

Let X_1 is an indicator variable, call it D.

$$\begin{aligned}\text{Cov}\,(D, X_2) &= \mathbb{E}[(D - \mathbb{E}[D])(X_2 - \mathbb{E}[X_2])] \\ &= \mathbb{E}[D(X_2 - \mathbb{E}[X_2])] \\ &= \pi\,\mathbb{E}[X_2 \mid D = 1] - \pi\,\mathbb{E}[X_2]\end{aligned}$$

Let $\pi = \mathbb{P}(D = 1)$ and note from definition, $\text{Var}\,(D) = \pi(1 - \pi)$. Then,

$$\delta_1 = \beta_1 + \frac{\beta_2}{(1 - \pi)}\,(\,\mathbb{E}[X_2 \mid D = 1] - \mathbb{E}[X_2])$$

Selection Bias

$$\delta_1 = \beta_1 + \frac{\beta_2}{(1-\pi)} \left(\mathbb{E}[X_2 \mid D = 1] - \mathbb{E}[X_2] \right)$$

In our context of D being a treatment indicator, δ_1 is our treatment effect estimate and β_1 is the true ATT.

Selection Bias

$$\delta_1 = \beta_1 + \frac{\beta_2}{(1-\pi)} \left(\mathbb{E}[X_2 \mid D = 1] - \mathbb{E}[X_2] \right)$$

In our context of D being a treatment indicator, δ_1 is our treatment effect estimate and β_1 is the true ATT.

We see that if the mean of X_2 differs for the treatment group, then our estimate is biased

$\rightarrow$ E.g. if D is college attendance and X_2 is parental income, then our treatment effect is biased if college attendees have difference average parental income

OVB In Practice

A lot of research will run regressions that look like

$$y_i = D_i\tau + \boldsymbol{X}_i'\beta + \varepsilon_i$$

The *key things* you will want to do is think through what might show up in the error term

1. If those omitted variables are correlated with D_i (after controlling for $\boldsymbol{X}_i$) and have an effect on y_i, then you have problems interpreting the effect as causal

Conditional Expectation Function

- Conditional Expectation Function

Modeling the CEF

- Linear Model of Conditional Expectation Function
- Making models more flexible
- More Flexible Approximations (`binscatter`)

Omitted Variable Bias (OVB)

- Reinterpreting selection bias as OVB

Frisch-Waugh-Lovell Theorem

Projection Matrix

Before we describe the Frisch-Waugh-Lovell theorem, let's define a few terms. Consider our regression estimator

$$\hat{\beta} = \left(X'X\right)^{-1} X'y$$

We could then create fitted values by multiplying X by our coefficient of interest:

$$X\hat{\beta} = X\left(X'X\right)^{-1} X'y \equiv P_X y$$

→ We define the Projection Matrix as P_X to be the fitted values from a regression of a variable on the variables X.

Residuals

The residuals from the regression are given by $\hat{\varepsilon} = y - \hat{y} = y - P_X y$

In matrix notation, we can write this as $\hat{\varepsilon} = (I - P_X)y$. We define M_X to be the annihilator matrix with $M_X \equiv I - P_X$

Residuals

The residuals from the regression are given by $\hat{\varepsilon} = y - \hat{y} = y - P_X y$

In matrix notation, we can write this as $\hat{\varepsilon} = (I - P_X)y$. We define M_X to be the annihilator matrix with $M_X \equiv I - P_X$

$\rightarrow$ The annihilator matrix first predicts y using a linear model of X and then subtracts off the prediction

Residuals

From regression algebra we have the residuals are (linearly) uncorrelated with $\boldsymbol{X}_i$

$$\mathbb{E}[\boldsymbol{X}_i \hat{\varepsilon}_i] = 0$$

Residuals

From regression algebra we have the residuals are (linearly) uncorrelated with $\boldsymbol{X}_i$

$$\mathbb{E}[\boldsymbol{X}_i \hat{\varepsilon}_i] = 0$$

If we assume that the CEF $\mathbb{E}[y_i \mid \boldsymbol{X}_i] = \boldsymbol{X}_i' \beta$, then we can go further and say

$$\mathbb{E}[\hat{\varepsilon}_i \mid \boldsymbol{X}_i = x] = 0$$

$\rightarrow$ the remaining variation in y_i, given by $\hat{\varepsilon}_i$, is unpredictable given $\boldsymbol{X}_i$

Frisch-Waugh-Lovell Theorem

Consider the regression

$$y_i = \tau D_i + W_i'\beta + u_i$$

$\rightarrow$ D_i is a scalar variable of interest and W_i is a $k \times 1$ vector of covariates

We can of course estimate the regression coefficients $\hat{\tau}$ and $\hat{\beta}$ jointly in a single regression

Frisch-Waugh-Lovell Theorem

The FWL theorem shows that instead of doing one regression, we could estimate $\hat{\tau}_{\texttt{OLS}}$ by the series of steps:

1. Regress y_i on W_i and grab the residuals, $M_W y$
2. Regress D_i on W_i and grab the residuals, $M_W D$
3. Regress $M_W y$ on $M_W D$ to estimate $\hat{\tau}_{\texttt{FWL}}$

Frisch-Waugh-Lovell Theorem

The FWL theorem shows that instead of doing one regression, we could estimate $\hat{\tau}_{\texttt{OLS}}$ by the series of steps:

1. Regress y_i on W_i and grab the residuals, $M_W y$
2. Regress D_i on W_i and grab the residuals, $M_W D$
3. Regress $M_W y$ on $M_W D$ to estimate $\hat{\tau}_{\texttt{FWL}}$

The estimate $\hat{\tau}_{\texttt{FWL}}$ is going to be *numerically identical* to $\hat{\tau}_{\texttt{OLS}}$.

$\rightarrow$ Up to degree-of-freedom correction, the standard errors will be identical as well (including robust and clustered standard errors)

- The final regression pretends we didn't estimate the K coefficients on W_i

Frisch-Waugh-Lovell Theorem

The FWL Theorem shows us how to think about the regression coefficient in a multivariate regression:

$\rightarrow$ We are predicting D_i and y_i using covariates W_i

$\rightarrow$ We are removing that predictable variation and seeing if the "remaining variation" in y_i and D_i are linearly correlated

Frisch-Waugh-Lovell Theorem

The FWL Theorem shows us how to think about the regression coefficient in a multivariate regression:

$\rightarrow$ We are predicting D_i and y_i using covariates W_i

$\rightarrow$ We are removing that predictable variation and seeing if the "remaining variation" in y_i and D_i are linearly correlated

To be clear, we do not have to run these regression; we can interpret our regression results as if we had run it using this procedure

Example of Frisch-Waugh-Lovell Thinking

We want to know the causal effect of college on earnings

→ D_i is an indicator for a person going to college

→ y_i is the worker's earnings at age 25

→ W_i is a vector of covariates we think are important determinents of college attendance and/or earnings

Run this regression:

$$y_i = \tau D_i + W_i'\beta + u_i$$

Example of Frisch-Waugh-Lovell Thinking

The regression estimate will do the following:

→ Predict whether a worker would go to college given the covariates W_i. The difference between D_i and the prediction $\hat{D}_i$ is *hopefully* due to random reasons

→ Predict how those covariates W_i would affect future earnings and remove that prediction. The remaining variation in wages is hopefully driven by (i) either college attendance, or (ii) other reasons that are uncorrelated with going to college

It is important therefore to know a lot about your subject and know what causes treatment uptake D_i

Example of Frisch-Waugh-Lovell Thinking

College attendance

Like with omitted variable bias, this is a story of what variables did we not include. In our college attendance example, say W_i is parental income and GPA.

$\rightarrow$ Both are important drivers of college attendance, but not the only ones

What are examples of other variables that can drive attendance?