Topic 2: Regression Toolkit

ECON 5783 — University of Arkansas

Prof. Kyle Butts

Fall 2024

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

- Creates a simple and interpretable model of y
- Has desirable properties for causal inference even if the outcome is not linear in covariates

Roadmap

- Conditional Expectation Function and Linear Model
  - Conditional Expectation Function
  - Linear Model of Conditional Expectation Function
- Omitted Variable Bias (OVB)
  - Reinterpreting selection bias as OVB
- Frisch-Waugh-Lovell Theorem
- More Flexible Approximations (`binscatter`)
- Binary Outcome Variable

Prediction model

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

- 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

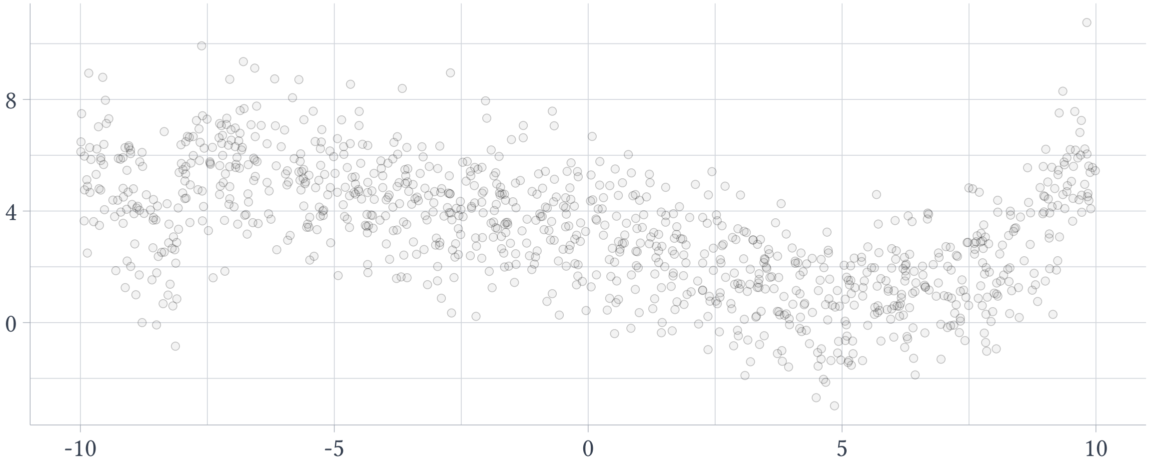
$$y = f(X) + \varepsilon,$$

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

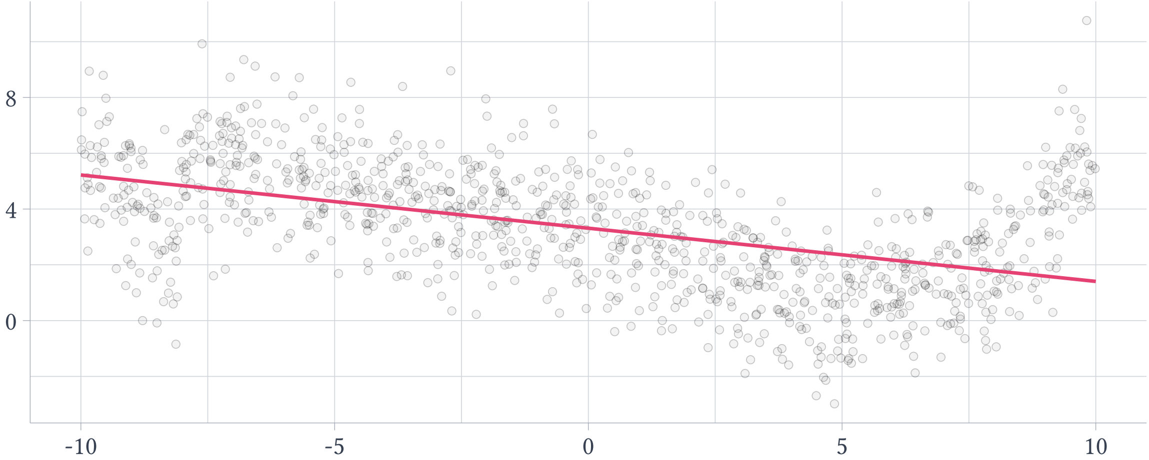
Prediction model

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

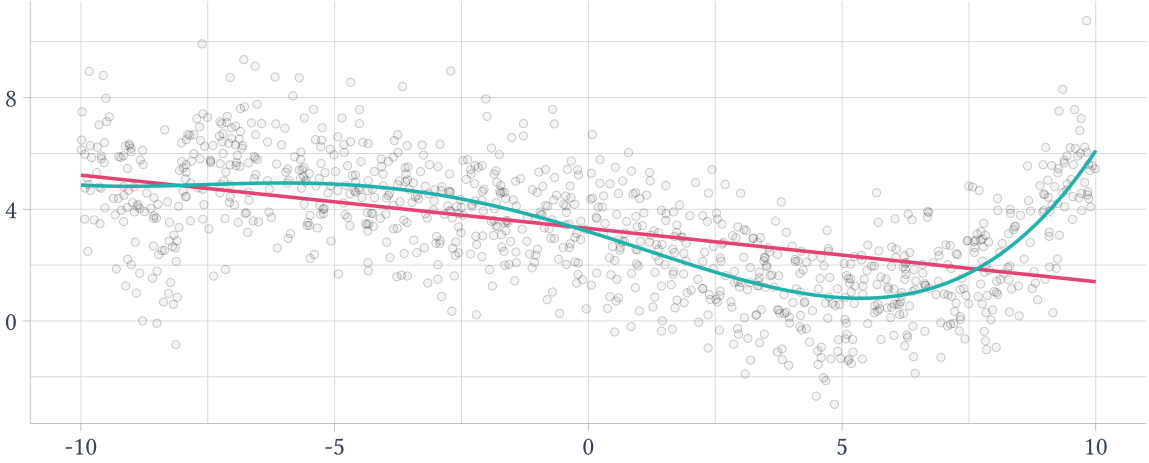
Examples of f:



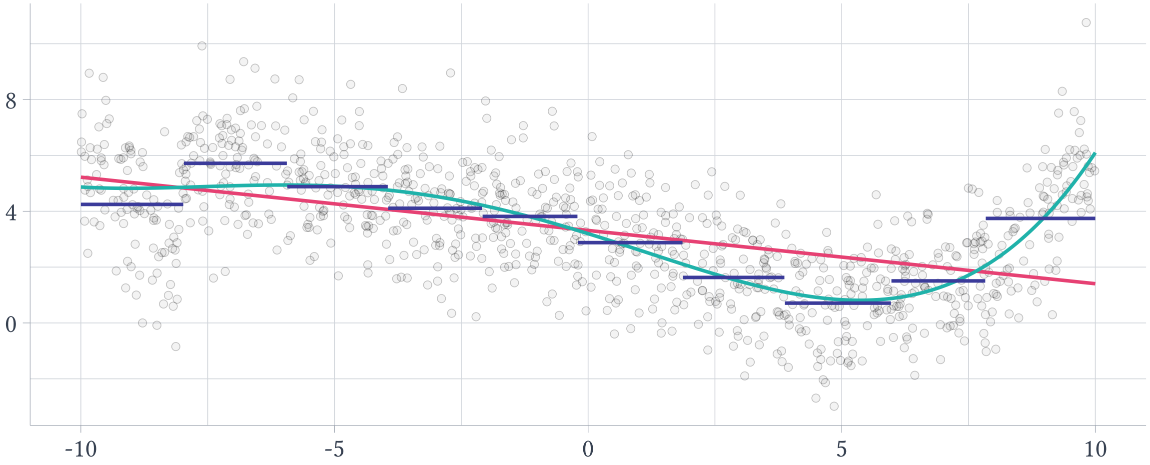
Examples of f: Line



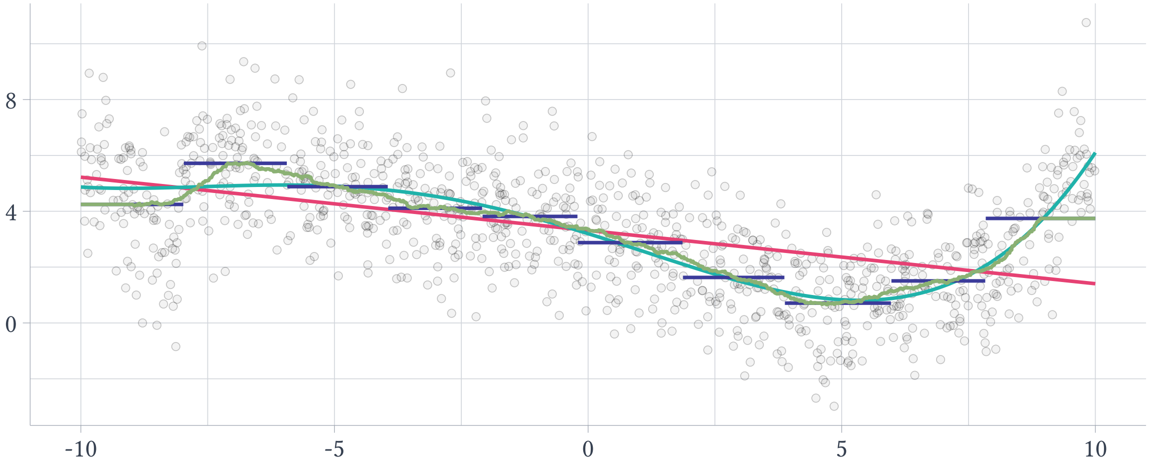
Examples of f: Line, Polynomial (x^4)



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



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



Prediction model

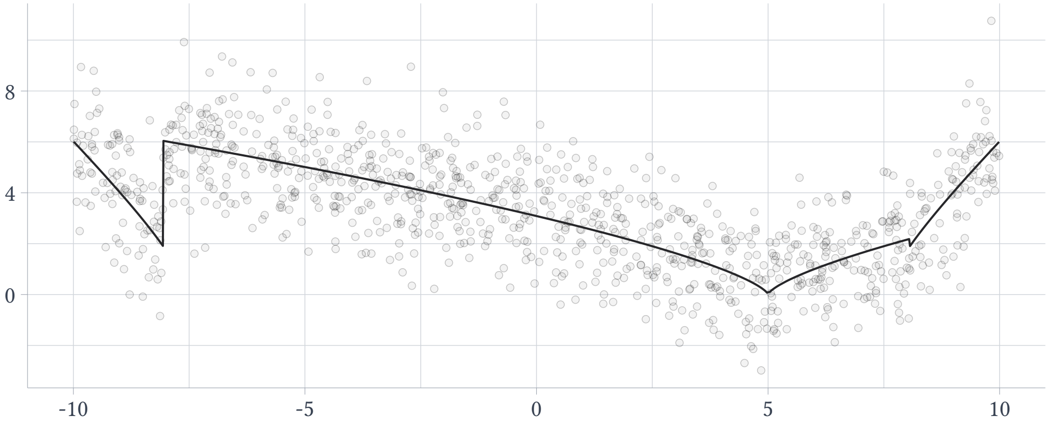
There are many different possible models of f ranging from a linear model; a ‘smooth’ model (polynomial or other); or a fully non-parametric function

The more ‘fancy’ a model:

- The more flexible the relationship between y and X can be
- The larger the risk of overfitting the data
- The less interpretable the model becomes

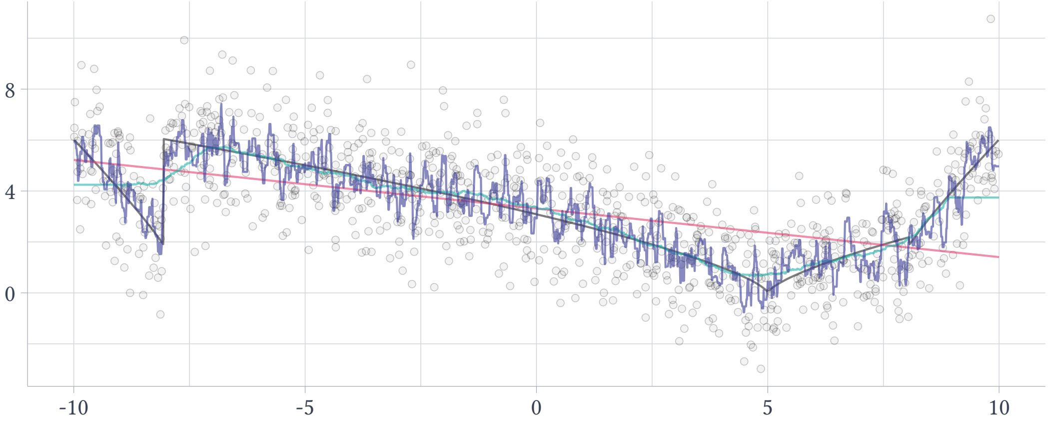
Flexibility vs. Overfitting

True $g(x)$



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"

Roadmap

- Conditional Expectation Function and Linear Model
  - Conditional Expectation Function
  - Linear Model of Conditional Expectation Function
- Omitted Variable Bias (OVB)
  - Reinterpreting selection bias as OVB
- Frisch-Waugh-Lovell Theorem
- More Flexible Approximations (`binscatter`)
- Binary Outcome Variable

The Conditional Expectation Function

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

$$g(x) \equiv \mathbb{E}[y_i \mid X_i = x]$$

- This reads "$g(x)$ is the expected value of y_i conditional on the unit having $X_i = x$"

The Conditional Expectation Function

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

$$g(x) \equiv \mathbb{E}[y_i \mid X_i = x]$$

- This reads "$g(x)$ is the expected value of y_i conditional on the unit having $X_i = x$"

The easiest way to estimate this for a given x is to average y_i for units with $X_i = x$.

The Conditional Expectation Function

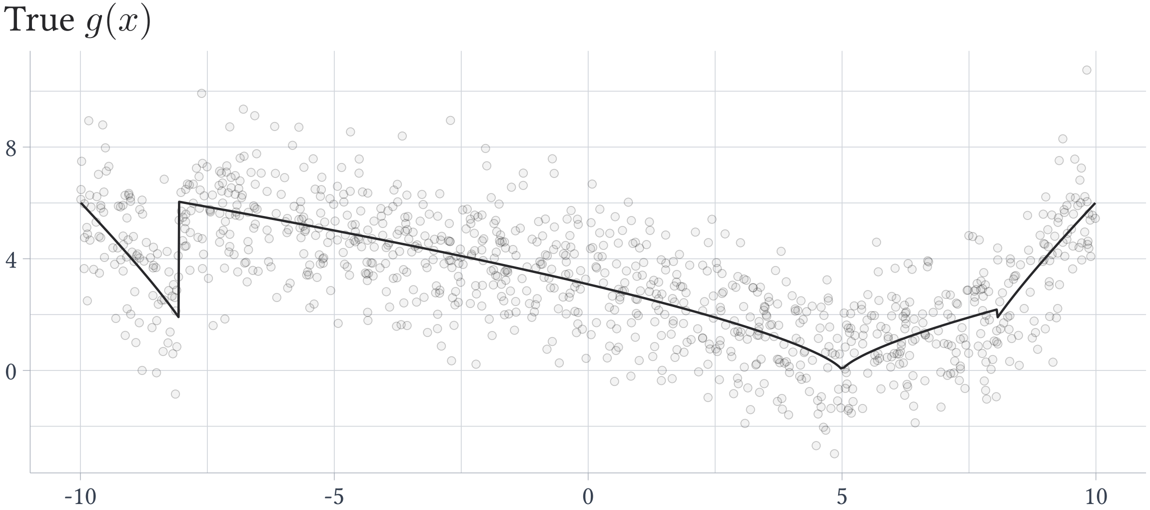
In particular, we will think a lot about the Conditional Expectation Function (CEF) of y_i given $X_i = (X_{i1}, \dots, X_{ip})'$:

$$g(x) \equiv \mathbb{E}[y_i \mid X_i = x]$$

- This reads "$g(x)$ is the expected value of y_i conditional on the unit having $X_i = x$"

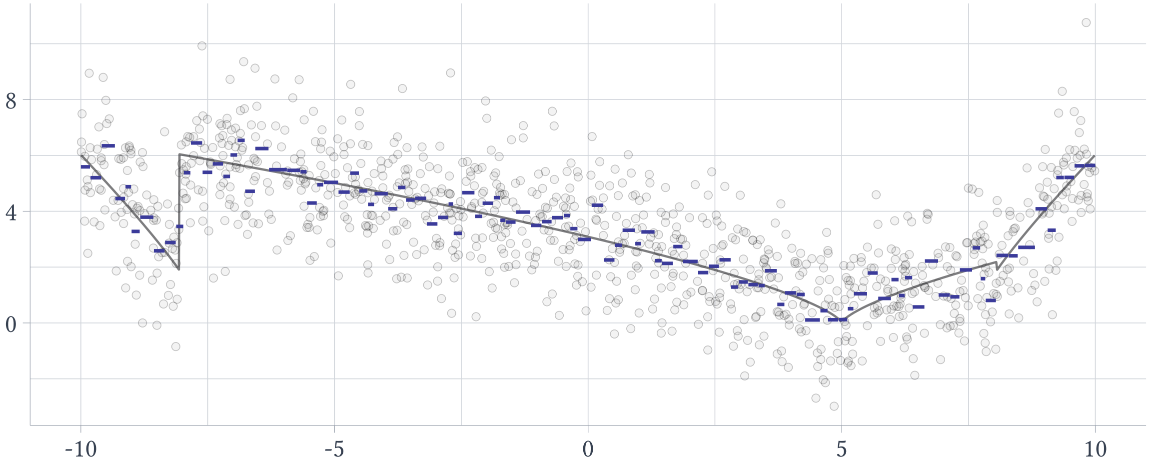
The easiest way to estimate this for a given x is to average y_i for units with $X_i = x$.

- Only uses observations with $X_i = x$ (or $X_i \approx x$ when X_i is continuous), so that is the relavent 'n' when considering sample size



True $g(x)$
8
4
0
-10
-5
0
5
10

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



The Conditional Expectation Function

Uses of the conditional expectation function:

1. Descriptive: how y on average covaries with X
   - $\rightarrow$ By definition compare $g(x_1)$ to $g(x_2)$
2. Prediction: if we know X_i, our best guess for y_i is $g(X_i)$
   - $\rightarrow$ Will prove 'best guess' next
3. Causal inference: what happens to y_i if we manipulate X_i
   - $\rightarrow$ Sometimes

Prediction Error and the CEF

The prediction error of the conditional expectation function is given by $\varepsilon_i = y_i - g(X_i)$. For any x, we have

$$\begin{aligned}
\mathbb{E}[\varepsilon_i \mid X_i = x] &= \mathbb{E}[y_i - \mathbb{E}[y_i \mid X_i = x] \mid X_i = x] \\
&= \mathbb{E}[y_i \mid X_i = x] - \mathbb{E}[y_i \mid X_i = x] \\
&= 0
\end{aligned}$$

Prediction Error and the CEF

The prediction error of the conditional expectation function is given by $\varepsilon_i = y_i - g(X_i)$. For any x, we have

$$\begin{aligned}
\mathbb{E}[\varepsilon_i \mid X_i = x] &= \mathbb{E}[y_i - \mathbb{E}[y_i \mid X_i = x] \mid X_i = x] \\
&= \mathbb{E}[y_i \mid X_i = x] - \mathbb{E}[y_i \mid X_i = x] \\
&= 0
\end{aligned}$$

The prediction error is unpredictable given $X_i = x$

- We have *used up* all the information that X_i can give us.
- This is not true for general $f(X)$

Mean-square prediction error

To provide a summary measure of fit, we want a ‘average’ prediction error over the population

- If we took the average of prediction error, positive and negative prediction errors would cancel out

Mean-square prediction error

To provide a summary measure of fit, we want a 'average' prediction error over the population

- If we took the average of prediction error, positive and negative prediction errors would cancel out

The mean-square (prediction) error (MSE) for some model f is calculated as:

$$\text{MSE}(f) \equiv \mathbb{E}\Big[\left(y_i - f(X_i)\right)^2\Big] \tag{1}$$

- Average over the population

Optimal model for y

The model f that minimizes the mean-square prediction error is the conditional expectation function.

$$\mathbb{E}\left[(y_i - f(X_i))^2\right] = \mathbb{E}\left[(y_i - g(X_i) + g(X_i) - f(X_i))^2\right]$$

Optimal model for y

The model f that minimizes the mean-square prediction error is the conditional expectation function.

$$\mathbb{E}\left[(y_i - f(X_i))^2\right] = \mathbb{E}\left[(y_i - g(X_i) + g(X_i) - f(X_i))^2\right]$$

$$= \mathbb{E}\left[(y_i - g(X_i))^2\right] + \mathbb{E}\left[(g(X_i) - f(X_i))^2\right] + 2\,\mathbb{E}[(y_i - g(X_i))\,(f(X_i) - g(X_i))]$$

- The first term does not depend on f

Optimal model for y

The last term equals 0:

$$
\begin{aligned}
&\mathbb{E}[(y_i - g(X_i))\,(f(X_i) - g(X_i))] \\
&\quad = \mathbb{E}[\,\mathbb{E}[(y_i - g(X_i))\,(f(X_i) - g(X_i)) \mid X_i]] \\
&\quad = \mathbb{E}[(\,\mathbb{E}[y_i \mid X_i] - g(X_i))\,(f(X_i) - g(X_i))] \\
&\quad = \mathbb{E}[(g(X_i) - g(X_i))\,(f(X_i) - g(X_i))] \\
&\quad = 0
\end{aligned}
$$

Optimal model for y

The model f that minimizes the mean-square prediction error is the conditional expectation function.

$$\underset{f}{\operatorname{argmin}}\ \mathbb{E}\Big[(y_i - f(X_i))^2\Big] = \mathbb{E}\Big[(y_i - g(X_i) + g(X_i) - f(X_i))^2\Big]$$

$$= \underset{f}{\operatorname{argmin}}\ \mathbb{E}\Big[(y_i - g(X_i))^2\Big] + \mathbb{E}\Big[(g(X_i) - f(X_i))^2\Big] + 0$$

- Minimizing this with respect to f only involves the second term so we set $f(X_i) = g(X_i)$

Optimal model for y

Therefore, in terms of mean-square prediction error, the conditional expectation function is the best predictor of y

Roadmap

- Conditional Expectation Function and Linear Model
  - Conditional Expectation Function
  - Linear Model of Conditional Expectation Function
- Omitted Variable Bias (OVB)
  - Reinterpreting selection bias as OVB
- Frisch-Waugh-Lovell Theorem
- More Flexible Approximations (`binscatter`)
- Binary Outcome Variable

Estimation of the CEF

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

- In the case where X_i is a discrete variable taking values $x_1, \ldots, x_L$, this is just sub-sample averages for $X_i = x_\ell$

Estimation of the CEF

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

- In the case where X_i is a discrete variable taking values $x_1, \ldots, x_L$, this is just sub-sample averages for $X_i = x_\ell$

When X_i is a multi-dimensional vector with many continuous variables, the density around any particular value x is typically going to be small or near-zero

- The so-called “curse of dimensionality”

Linear Model

Instead, it is common to propose a *parametric* model of the conditional expectation function:

$$y_i = X_i'\beta + \text{error}$$

- We model y as a linear function of the covariates
- Most of the time, we assume X_i contains a constant for an intercept

Linear Model

Instead, it is common to propose a *parametric* model of the conditional expectation function:

$$y_i = X_i' \beta + \text{error}$$

- We model y as a linear function of the covariates
- Most of the time, we assume X_i contains a constant for an intercept

Similar to the conditional expectation function, we can find the "best" linear predictor of y:

$$\hat{\beta}_{\texttt{OLS}} \equiv \operatorname*{argmin}_{\beta} \mathbb{E}\left[\left(y_i - X_i' \beta \right)^2 \right]$$

- Same as before but searching over only linear functions of X

Ordinary Least Squares

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

$$\mathbb{E}\big[X_i\,\big(y_i - X_i'\beta_{\mathsf{OLS}}\big)\big] = 0$$

$$\implies \mathbb{E}[X_i y_i] - \mathbb{E}\big[X_i X_i'\big]\beta_{\mathsf{OLS}} = 0$$

$$\implies \beta_{\mathsf{OLS}} = \left(\mathbb{E}\big[X_i X_i'\big]\right)^{-1}\mathbb{E}[X_i y_i]$$

- The best linear predictor of y is the ordinary-least squares estimate
- Similar math shows β_{OLS} is the best linear predictor of the CEF function $\mathbb{E}[y_i \mid X_i = x]$

Ordinary Least Squares Estimator

We can estimate using a sample of observations:

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

Or in matrix notation

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

- X is the $n \times k$ matrix with row given by X_i' and y is the column vector of outcome variables

Indicator Variables

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

- In some cases, the data might look to grow linearly in X_i, in which case, it is a reasonable assumption

Indicator Variables

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

- In some cases, the data might look to grow linearly in X_i, in which case, it is a reasonable assumption

A linear model means *linear in parameters*; we can include polynomial terms to allow for non-linear (but smooth) model of y_i given X_i

Discrete variables

When 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:

$$y_i = \sum_{\ell=1}^{L} \mathbb{1}[X_i = x_\ell]\beta_\ell + u_i \tag{2}$$

Discrete variables

When 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:

$$y_i = \sum_{\ell=1}^{L} \mathbb{1}[X_i = x_\ell]\beta_\ell + u_i \tag{2}$$

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

Discrete variables

When 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:

$$y_i = \sum_{\ell=1}^{L} \mathbb{1}[X_i = x_\ell]\beta_\ell + u_i \tag{2}$$

- The ordinary least-squares estimator estimates $\hat{\beta}_\ell = \hat{\mathbb{E}}[y_i \mid X_i = x_\ell]$
- In this case, the CEF is *correctly specified* as the linear model (2).

Ommitted Categories

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

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

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

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

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

Roadmap

Conditional Expectation Function and Linear Model

- Conditional Expectation Function
- Linear Model of Conditional Expectation Function

Omitted Variable Bias (OVB)

- Reinterpreting selection bias as OVB

Frisch-Waugh-Lovell Theorem

More Flexible Approximations (`binscatter`)

Binary Outcome Variable

Difference between true model and model we estimate

Say there is a true causal model for y

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

- Assume $\mathbb{E}[\varepsilon_i \mid X_i] = 0$ 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

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

- Assume $\mathbb{E}[\varepsilon_i \mid X_i] = 0$ 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

$$y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i \quad \text{and} \quad y_i = \delta_0 + X_{i1}\delta_1 + error$$

We have the following relationship:

$$\begin{aligned}
\delta_1 &= \frac{\operatorname{cov}(X_1, y)}{\operatorname{var}(X_1)} \\
&= \frac{\operatorname{cov}(X_1, \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon)}{\operatorname{var}(X_1)}
\end{aligned}$$

Omitted Variable Bias

$$y_i = \beta_0 + X_{i1}\beta_1 + X_{i2}\beta_2 + \varepsilon_i \quad \text{and} \quad y_i = \delta_0 + X_{i1}\delta_1 + error$$

We have the following relationship:

$$\begin{aligned}
\delta_1 &= \frac{\operatorname{cov}(X_1, y)}{\operatorname{var}(X_1)} \\
&= \frac{\operatorname{cov}(X_1, \beta_0 + X_1\beta_1 + X_2\beta_2 + \varepsilon)}{\operatorname{var}(X_1)} \\
&= \beta_1 + \beta_2 \frac{\operatorname{cov}(X_1, X_2)}{\operatorname{var}(X_1)}
\end{aligned}$$

Omitted Variable Bias

$$\delta_1 = \beta_1 + \beta_2 \frac{\text{cov}\left(X_1, X_2\right)}{\text{var}\left(X_1\right)}$$

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

- If X_1 and X_2 are correlated, then knowing about X_1 tells me information on X_2.
- I would want to use that implicit information on X_2 to predict y as well!

Omitted Variable Bias

$$\delta_1 = \beta_1 + \beta_2 \frac{\operatorname{cov}(X_1, X_2)}{\operatorname{var}(X_1)}$$

We can often times ‘sign’ the bias:

- The sign of β_2 is what we think the effect of X_2 is on y
- $\operatorname{cov}(X_1, X_2)$ 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

Roadmap

Omitted Variable Bias

When X_1 is an indicator variable, call it D.

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

Let $\pi = \mathbb{P}(D = 1)$. Then,

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

Selection Bias

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

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

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

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

- 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

$$y_i = D_i \tau + X_i' \beta + \varepsilon_i$$

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 X_i) and have an effect on y_i, then you have problems interpreting the effect as causal

Roadmap

- Conditional Expectation Function and Linear Model
  - Conditional Expectation Function
  - Linear Model of Conditional Expectation Function
- Omitted Variable Bias (OVB)
  - Reinterpreting selection bias as OVB
- Frisch-Waugh-Lovell Theorem
- More Flexible Approximations (`binscatter`)
- Binary Outcome Variable

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 $M_X \equiv I - P_X$

- 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 X_i

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

Residuals

From regression algebra we have the residuals are (linearly) uncorrelated with X_i

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

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

$$\mathbb{E}[\hat{\varepsilon}_i \mid X_i] = 0$$

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

Frisch-Waugh-Lovell Theorem

Consider the regression

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

- 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 $\hat{\tau}_{\texttt{OLS}}$.

- Up to degree-of-freedom correction, the standard errors will be identical as well (the FWL 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:

- We are predicting D_i and y_i using covariates W_i
- 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:

- We are predicting D_i and y_i using covariates W_i
- 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

Say D_i is an indicator for a person going to college and 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

- We want to know the causal effect of college on earnings

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

Roadmap

Conditional Expectation Function and Linear Model

- Conditional Expectation Function
- Linear Model of Conditional Expectation Function

Omitted Variable Bias (OVB)

- Reinterpreting selection bias as OVB

Frisch-Waugh-Lovell Theorem

More Flexible Approximations (`binscatter`)

Binary Outcome Variable

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:

$$y_i = f(X_i) + W_i'\beta + u_i$$

- $f(X_i)$ is a highly flexible function
- $W_i'\beta$ is a set of linear control variables

The key advantage is you can control for important factors W_i and still present a bivariate scatter plot that summarizes the (conditional) relationship between X and y

Visual Display of Variation

Roadmap

Conditional Expectation Function and Linear Model

- Conditional Expectation Function
- Linear Model of Conditional Expectation Function

Omitted Variable Bias (OVB)

- Reinterpreting selection bias as OVB

Frisch-Waugh-Lovell Theorem

More Flexible Approximations (`binscatter`)

Binary Outcome Variable