Chapter 4: Introduction to Regression

Jonathan Roth

Mathematical Econometrics I
Brown University

Motivation

- We showed that under conditional unconfoundedness we can learn the conditional average treatment effect (CATE) by comparing outcome means for the treatment/control group conditional on X_i:

$$CATE(x) = E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x]$$

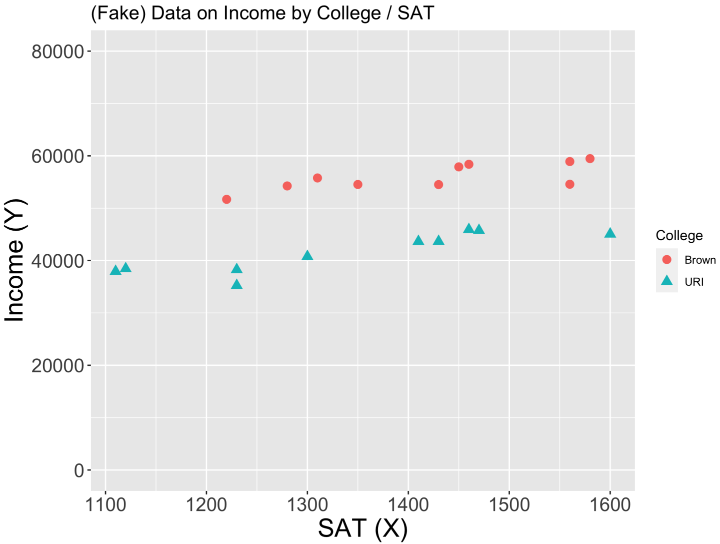
- When X_i is discrete and we have many observations per x-value (N_x is large), we showed how we can use the Central Limit Theorem to estimate each of these conditional means and "do inference"

Motivation

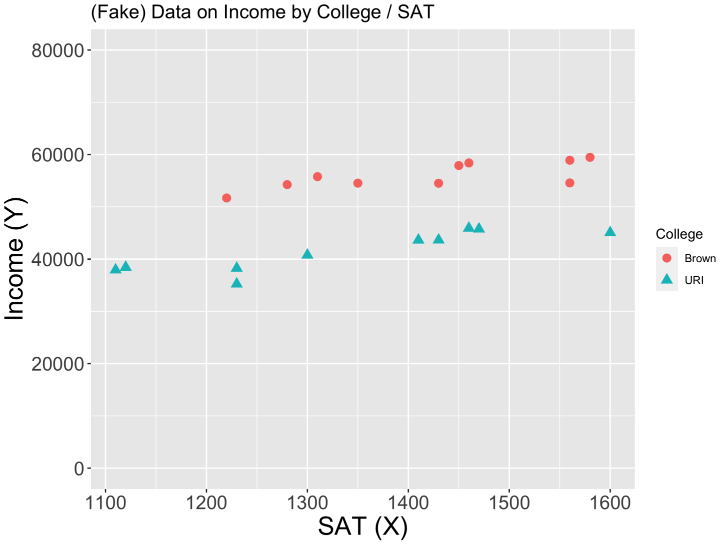
- We showed that under conditional unconfoundedness we can learn the conditional average treatment effect (CATE) by comparing outcome means for the treatment/control group conditional on X_i:

$$CATE(x) = E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x]$$

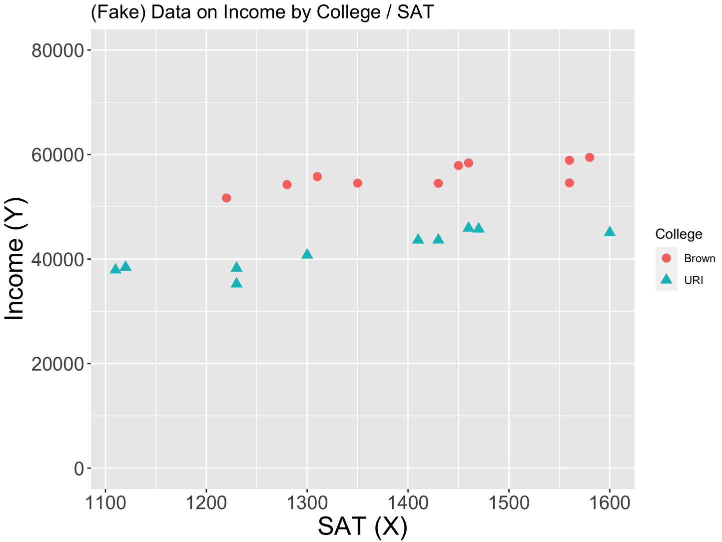
- When X_i is discrete and we have many observations per x-value (N_x is large), we showed how we can use the Central Limit Theorem to estimate each of these conditional means and "do inference"
- But what about when X_i is continuous?



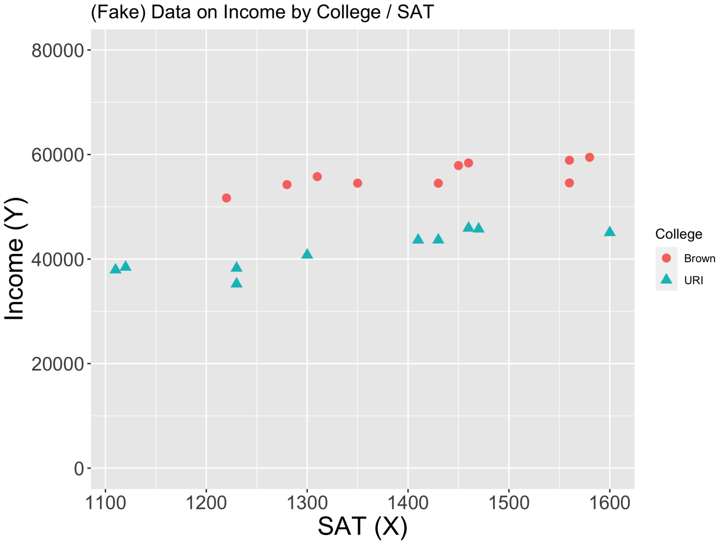
- Suppose this is our data



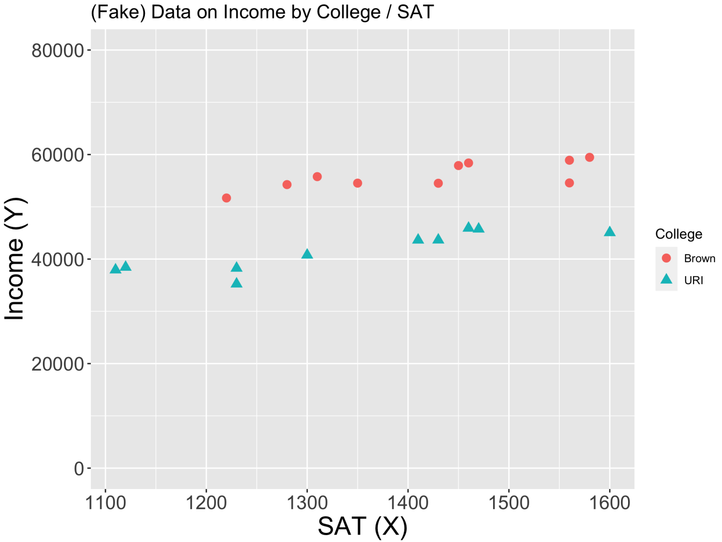
- We might be willing to assume that college attendance is as-good-as-random conditional on SAT score



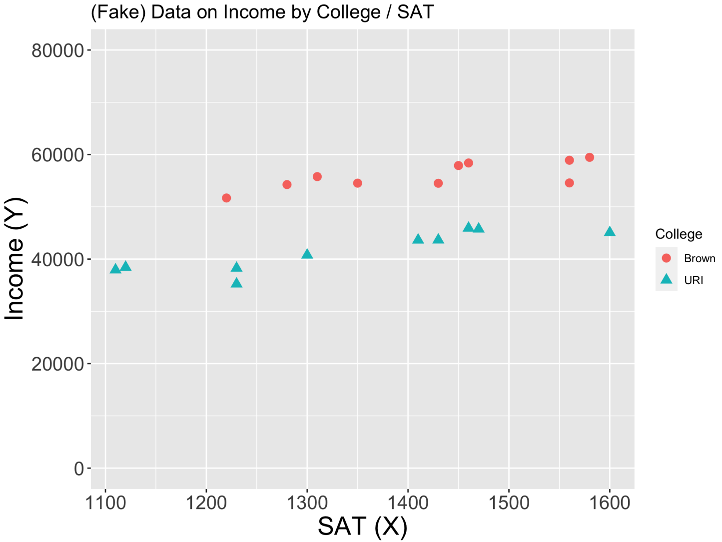
- Say we're interested in the CATE at $X = 1350$. Theory tells us to compare average income for Brown/URI SAT scores with $X = 1350$.



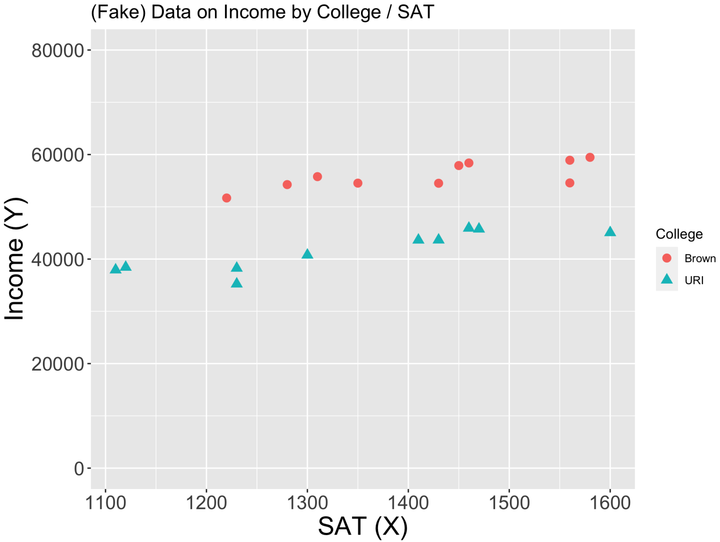
- We could estimate the average at Brown using our 1 student with $X = 1350$. But that estimate is very noisy, & we can't apply the CLT



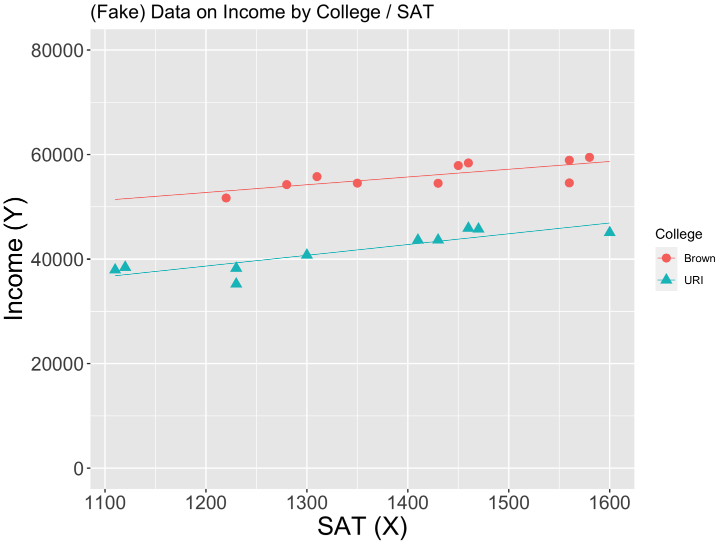
- We could estimate the average at Brown using our 1 student with $X = 1350$. But that estimate is very noisy, & we can't apply the CLT
- Moreover, we don't have any URI students with $X = 1350$!



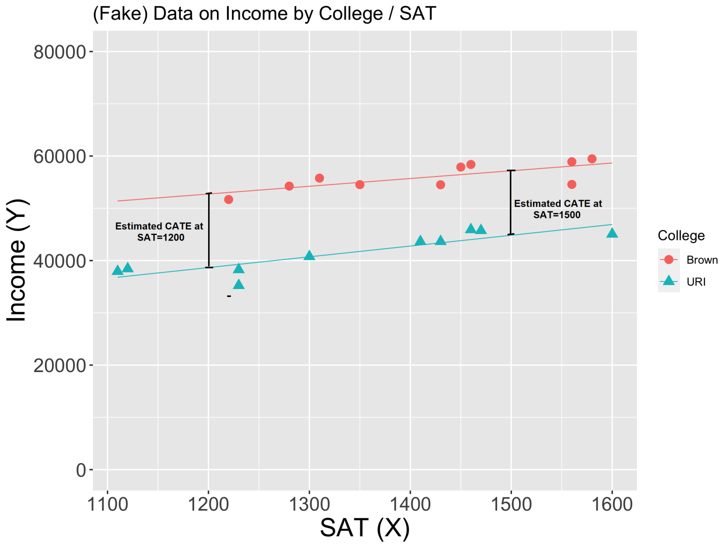
- Clearly, we need to extrapolate from students with other SAT scores.



- Clearly, we need to extrapolate from students with other SAT scores.
- What would you do if you were eyeballing it?



- Clearly, we need to extrapolate from students with other SAT scores.
- What would you do if you were eyeballing it?
- Probably draw a line through the points to estimate the CEFs!



- With these CEF estimates in hand, we can estimate $CATE(x)$ at any x

Outline

1. Population Regression
2. Sample Regressions (OLS)
3. Putting Regression into Practice

Introduction to Regression

- The idea of **regression** is to formalize the process of estimating the conditional expectation function (CEF) by extrapolating across units using a particular functional form (e.g. linear, quadratic, etc.)

Introduction to Regression

- The idea of **regression** is to formalize the process of estimating the conditional expectation function (CEF) by extrapolating across units using a particular functional form (e.g. linear, quadratic, etc.)
- There are a few outstanding questions that we need to answer:

Introduction to Regression

- The idea of **regression** is to formalize the process of estimating the conditional expectation function (CEF) by extrapolating across units using a particular functional form (e.g. linear, quadratic, etc.)

- There are a few outstanding questions that we need to answer:

- How do we approximate the CEF in the sample that we have? (I.e. how to draw the lines through the data!)

Introduction to Regression

- The idea of **regression** is to formalize the process of estimating the conditional expectation function (CEF) by extrapolating across units using a particular functional form (e.g. linear, quadratic, etc.)

- There are a few outstanding questions that we need to answer:

- How do we approximate the CEF in the sample that we have? (I.e. how to draw the lines through the data!)

- How can we construct confidence intervals / do hypothesis tests for the estimates of the CEF?

Introduction to Regression

- The idea of **regression** is to formalize the process of estimating the conditional expectation function (CEF) by extrapolating across units using a particular functional form (e.g. linear, quadratic, etc.)
- There are a few outstanding questions that we need to answer:
- How do we approximate the CEF in the sample that we have? (I.e. how to draw the lines through the data!)
- How can we construct confidence intervals / do hypothesis tests for the estimates of the CEF?
- What happens if the real CEF doesn't take the form we've used for estimation (e.g. isn't linear)?

Introduction to Regression

- The idea of **regression** is to formalize the process of estimating the conditional expectation function (CEF) by extrapolating across units using a particular functional form (e.g. linear, quadratic, etc.)
- There are a few outstanding questions that we need to answer:
- How do we approximate the CEF in the sample that we have? (I.e. how to draw the lines through the data!)
- How can we construct confidence intervals / do hypothesis tests for the estimates of the CEF?
- What happens if the real CEF doesn't take the form we've used for estimation (e.g. isn't linear)?
- We'll try to answer all of these questions over the next several lectures!

Roadmap

- **What we know how to do:** Estimate and test hypotheses about population means using sample means
- **What we want to do:** Estimate approximations to the CEF and test hypotheses about them

Roadmap

- **What we know how to do:** Estimate and test hypotheses about population means using sample means
- **What we want to do:** Estimate approximations to the CEF and test hypotheses about them

How can we use what know to do what we want?

Roadmap

- **What we know how to do:** Estimate and test hypotheses about population means using sample means
- **What we want to do:** Estimate approximations to the CEF and test hypotheses about them

How can we use what know to do what we want?

- 1) Assume the CEF takes a particular form, e.g. linear:

$$E[Y_i|X_i = x] = \alpha + x\beta$$

Roadmap

- **What we know how to do:** Estimate and test hypotheses about population means using sample means
- **What we want to do:** Estimate approximations to the CEF and test hypotheses about them

How can we use what know to do what we want?

- 1) Assume the CEF takes a particular form, e.g. linear:

$$E[Y_i|X_i = x] = \alpha + x\beta$$

- 2) Show that under this assumption, α and β can be represented as functions of population means.

Roadmap

- **What we know how to do:** Estimate and test hypotheses about population means using sample means
- **What we want to do:** Estimate approximations to the CEF and test hypotheses about them

How can we use what know to do what we want?

- 1) Assume the CEF takes a particular form, e.g. linear:

$$E[Y_i|X_i = x] = \alpha + x\beta$$

- 2) Show that under this assumption, α and β can be represented as functions of population means.
- 3) Use our tools for estimating population means using sample means to estimate α, β and test hypotheses about them

Roadmap

- **What we know how to do:** Estimate and test hypotheses about population means using sample means
- **What we want to do:** Estimate approximations to the CEF and test hypotheses about them

How can we use what know to do what we want?

- 1) Assume the CEF takes a particular form, e.g. linear:

$$E[Y_i|X_i = x] = \alpha + x\beta$$

- 2) Show that under this assumption, α and β can be represented as functions of population means.
- 3) Use our tools for estimating population means using sample means to estimate α, β and test hypotheses about them
- 4) Argue that even if our assumption about the form of the CEF is wrong, the parameters α, β may provide a "good" approximation.

The "Least Squares" Problem

- Suppose X_i is scalar and the CEF is linear (we'll relax both later):

$$E[Y_i|X_i = x] = \alpha + x\beta$$

The "Least Squares" Problem

- Suppose X_i is scalar and the CEF is linear (we'll relax both later):

$$E[Y_i|X_i = x] = \alpha + x\beta$$

- A useful fact which we will exploit is that when this is true (α, β) solve the "least squares" problem:

$$(\alpha, \beta) = \arg\min_{a,b} E[(Y_i - (a + bX_i))^2]$$

The "Least Squares" Problem

- Suppose X_i is scalar and the CEF is linear (we'll relax both later):

$$E[Y_i|X_i = x] = \alpha + x\beta$$

- A useful fact which we will exploit is that when this is true (α, β) solve the "least squares" problem:

$$(\alpha, \beta) = \arg\min_{a,b} E[(Y_i - (a + bX_i))^2]$$

- Where does this come from?!

Starting with a Simpler Problem

- To show that (α, β) solve a "least-squares" problem, let's first consider a simpler, related problem:

Starting with a Simpler Problem

- To show that (α, β) solve a "least-squares" problem, let's first consider a simpler, related problem:

- Suppose that we want to find a constant u to minimize

$$\min_{u} E[(Y_i - u)^2]$$

Starting with a Simpler Problem

- To show that (α, β) solve a "least-squares" problem, let's first consider a simpler, related problem:
- Suppose that we want to find a constant u to minimize

$$\min_u E[(Y_i - u)^2]$$

- What constant u should we choose?

Starting with a Simpler Problem

- To show that (α, β) solve a "least-squares" problem, let's first consider a simpler, related problem:
- Suppose that we want to find a constant u to minimize

$$\min_u E[(Y_i - u)^2]$$

- What constant u should we choose? The population mean $\mu = E[Y_i]$!

Starting with a Simpler Problem

- To show that (α, β) solve a "least-squares" problem, let's first consider a simpler, related problem:

- Suppose that we want to find a constant u to minimize

$$\min_u E[(Y_i - u)^2]$$

- What constant u should we choose? The population mean $\mu = E[Y_i]$!

- Proof:
  The derivative of $E[(Y_i - u)^2]$ with respect to u is $E[2(Y_i - u)]$.

Starting with a Simpler Problem

- To show that (α, β) solve a "least-squares" problem, let's first consider a simpler, related problem:

- Suppose that we want to find a constant u to minimize

$$\min_u E[(Y_i - u)^2]$$

- What constant u should we choose? The population mean $\mu = E[Y_i]$!

- Proof:
  The derivative of $E[(Y_i - u)^2]$ with respect to u is $E[2(Y_i - u)]$.
  Setting the derivative to 0, we obtain

$$E[2(Y_i - \mu)] = 0 \Rightarrow 2E[Y_i] = 2u \Rightarrow u = E[Y_i].$$

Now A Slightly Harder Problem

- Now suppose we want to choose the function $u(x)$ to minimize

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

Now A Slightly Harder Problem

- Now suppose we want to choose the function $u(x)$ to minimize

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- What function $u(x)$ should we choose?

Now A Slightly Harder Problem

- Now suppose we want to choose the function $u(x)$ to minimize

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- What function $u(x)$ should we choose? The conditional expectation $u(x) = E[Y|X = x]$.

Now A Slightly Harder Problem

- Now suppose we want to choose the function $u(x)$ to minimize

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- What function $u(x)$ should we choose? The conditional expectation $u(x) = E[Y|X = x]$.

- Proof:
  By the law of iterated expectations,

$$E[(Y_i - u(X_i))^2] = E[E[(Y_i - u(X_i))^2|X_i]].$$

Now A Slightly Harder Problem

- Now suppose we want to choose the function $u(x)$ to minimize

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- What function $u(x)$ should we choose? The conditional expectation $u(x) = E[Y|X = x]$.

- Proof:
  By the law of iterated expectations,

$$E[(Y_i - u(X_i))^2] = E[E[(Y_i - u(X_i))^2 | X_i]].$$

  Thus, for each value of x, we want to choose $u(x)$ to minimize

$$E[(Y_i - u(x))^2 | X_i = x].$$

Now A Slightly Harder Problem

- Now suppose we want to choose the function $u(x)$ to minimize

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- What function $u(x)$ should we choose? The conditional expectation $u(x) = E[Y|X = x]$.

- Proof:
  By the law of iterated expectations,

$$E[(Y_i - u(X_i))^2] = E[E[(Y_i - u(X_i))^2|X_i]].$$

  Thus, for each value of x, we want to choose $u(x)$ to minimize

$$E[(Y_i - u(x))^2|X_i = x].$$

  However, our argument on the previous slide implies that the solution is $u(x) = E[Y_i|X_i = x]$.

Looping Back...

- We've shown that the function $u(x) = E[Y_i|X_i = x]$ solves

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

Looping Back...

- We've shown that the function $u(x) = E[Y_i|X_i = x]$ solves

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- Thus, if $E[Y_i|X_i = x] = \alpha + \beta x$, then $u(x) = \alpha + \beta x$ minimizes

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

Looping Back...

- We've shown that the function $u(x) = E[Y_i|X_i = x]$ solves

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- Thus, if $E[Y_i|X_i = x] = \alpha + \beta x$, then $u(x) = \alpha + \beta x$ minimizes

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- The minimization above was over *all* functions $u(\cdot)$, including linear ones of the form $a + bx$. Hence,

$$E[(Y_i - (\alpha + \beta X_i))^2] \leq E[(Y_i - (a + bX_i))^2] \text{ for all } a, b.$$

Looping Back...

- We've shown that the function $u(x) = E[Y_i|X_i = x]$ solves

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- Thus, if $E[Y_i|X_i = x] = \alpha + \beta x$, then $u(x) = \alpha + \beta x$ minimizes

$$\min_{u(\cdot)} E[(Y_i - u(X_i))^2]$$

- The minimization above was over *all* functions $u(\cdot)$, including linear ones of the form $a + bx$. Hence,

$$E[(Y_i - (\alpha + \beta X_i))^2] \leq E[(Y_i - (a + bX_i))^2] \text{ for all } a, b.$$

- This implies that (α, β) solve

$$\min_{a,b} E[(Y_i - (a + bX_i))^2],$$

as we wanted to show

Why This is Useful

- So we've shown that α, β are the solutions to

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- How does this help us?

Why This is Useful

- So we've shown that α, β are the solutions to

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- How does this help us? By solving the minimization problem, we can express α, β as functions of population expectations.

Why This is Useful

- So we've shown that α, β are the solutions to

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- How does this help us? By solving the minimization problem, we can express α, β as functions of population expectations.

- Let's take the derivative w.r.t. a and b and set them to zero at (α, β):

Why This is Useful

- So we've shown that α, β are the solutions to

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- How does this help us? By solving the minimization problem, we can express α, β as functions of population expectations.

- Let's take the derivative w.r.t. a and b and set them to zero at (α, β):

$$E[-2(Y_i - (\alpha + \beta X_i))] = 0$$

Why This is Useful

- So we've shown that α, β are the solutions to

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- How does this help us? By solving the minimization problem, we can express α, β as functions of population expectations.

- Let's take the derivative w.r.t. a and b and set them to zero at (α, β):

$$E[-2(Y_i - (\alpha + \beta X_i))] = 0$$
$$E[-2X_i(Y_i - (\alpha + \beta X_i))] = 0$$

Why This is Useful

- So we've shown that α, β are the solutions to

$$\min_{a,b} E[(Y_i - (a + bX_i))^2].$$

- How does this help us? By solving the minimization problem, we can express α, β as functions of population expectations.

- Let's take the derivative w.r.t. a and b and set them to zero at (α, β):

$$E[-2(Y_i - (\alpha + \beta X_i))] = 0$$
$$E[-2X_i(Y_i - (\alpha + \beta X_i))] = 0$$

- We now have 2 equations with 2 unknowns, which we can use to solve for the CEF parameters (α, β)

The Least Squares Solution

- The solution to the system of equations is as follows:

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

$$\alpha = E[Y_i] - E[X_i]\beta$$

The Least Squares Solution

- The solution to the system of equations is as follows:

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$
$$\alpha = E[Y_i] - E[X_i]\beta$$

- These are continuous functions of population means!

The Least Squares Solution

- The solution to the system of equations is as follows:

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$
$$\alpha = E[Y_i] - E[X_i]\beta$$

- These are continuous functions of population means!

- We can therefore use the tools from previous lectures to estimate them and test hypotheses about the CEF!

Roadmap

- **What we know how to do:** Estimate and test hypotheses about population means using sample means
- **What we want to do:** Estimate approximations to the CEF and test hypotheses about them

How can we use what know to do what we want?

- 1) Assume the CEF takes a particular form, e.g. linear:

$$E[Y_i|X_i = x] = \alpha + x\beta \qquad \checkmark$$

- 2) Show that under this assumption, α and β can be represented as functions of population means. ✓
- 3) Use our tools for estimating population means using sample means to estimate α, β and test hypotheses about them
- 4) Argue that even if our assumption about the form of the CEF is wrong, the parameters α, β may provide a "good" approximation.

Outline

1. Population Regression ✓
2. Sample Regressions (OLS)
3. Putting Regression into Practice

Estimating Regression Coefficients

- We showed that when $E[Y_i \mid X_i = x] = \alpha + \beta x$

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

$$\alpha = E[Y_i] - E[X_i]\beta$$

- How can we estimate α, β?

Estimating Regression Coefficients

- We showed that when $E[Y_i \mid X_i = x] = \alpha + \beta x$

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

$$\alpha = E[Y_i] - E[X_i]\beta$$

- How can we estimate α, β?
  Replace population means with sample averages!

Estimating Regression Coefficients

- We showed that when $E[Y_i \mid X_i = x] = \alpha + \beta x$

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

$$\alpha = E[Y_i] - E[X_i]\beta$$

- How can we estimate α, β?
  Replace population means with sample averages!

$$\hat{\beta} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{\widehat{Cov}(X_i, Y_i)}{\widehat{Var}(X_i)}$$

Estimating Regression Coefficients

- We showed that when $E[Y_i \mid X_i = x] = \alpha + \beta x$

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$

$$\alpha = E[Y_i] - E[X_i]\beta$$

- How can we estimate α, β?
  Replace population means with sample averages!

$$\hat{\beta} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{\widehat{Cov}(X_i, Y_i)}{\widehat{Var}(X_i)}$$

$$\hat{\alpha} = \bar{Y} - \bar{X}\hat{\beta}$$

Estimating Regression Coefficients

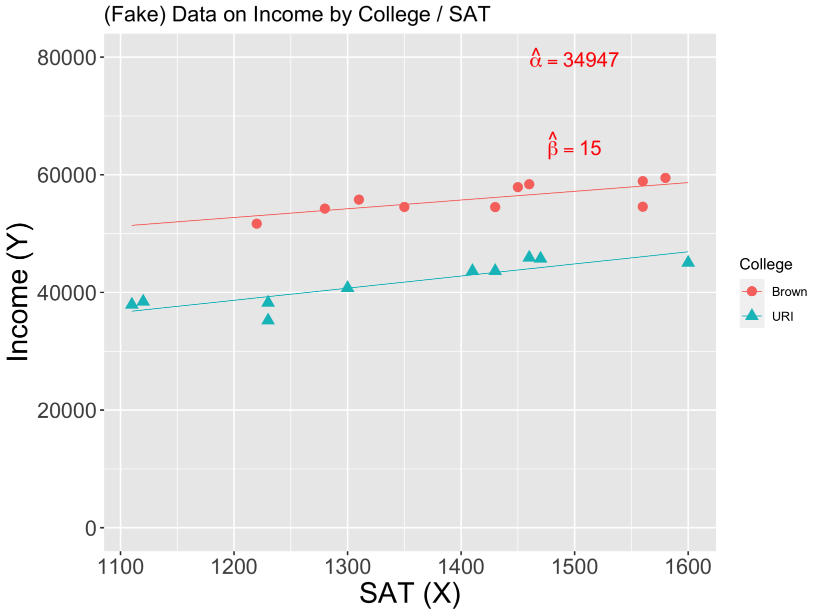
- We showed that when $E[Y_i \mid X_i = x] = \alpha + \beta x$

$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$
$$\alpha = E[Y_i] - E[X_i]\beta$$

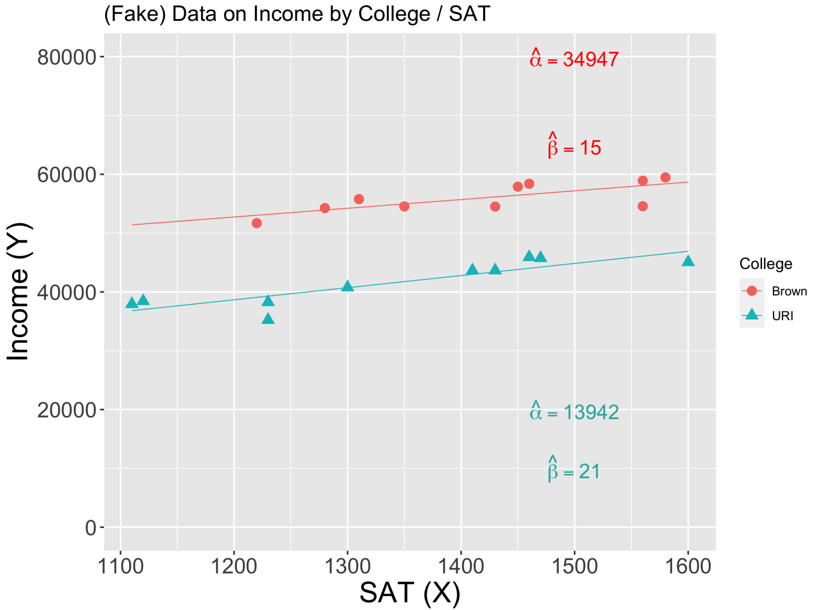
- How can we estimate α, β?
  Replace population means with sample averages!

$$\hat{\beta} = \frac{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2} = \frac{\widehat{Cov}(X_i, Y_i)}{\widehat{Var}(X_i)}$$
$$\hat{\alpha} = \bar{Y} - \bar{X}\hat{\beta}$$

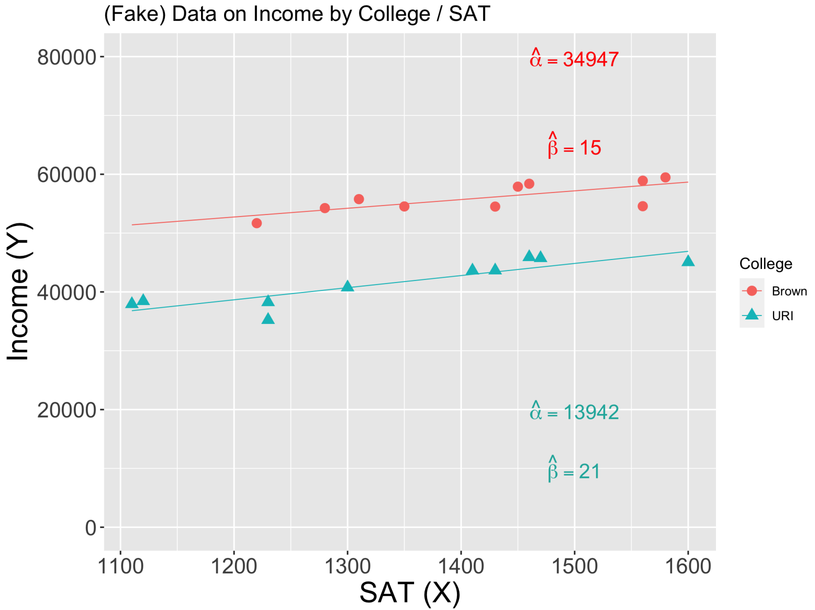
- These $\hat{\alpha}, \hat{\beta}$ are called *ordinary least squares* (OLS) coefficients
  - They solve the "sample analog" problem, $\min_{a,b} \frac{1}{N}\sum_i (Y_i - (a + bX_i))^2$



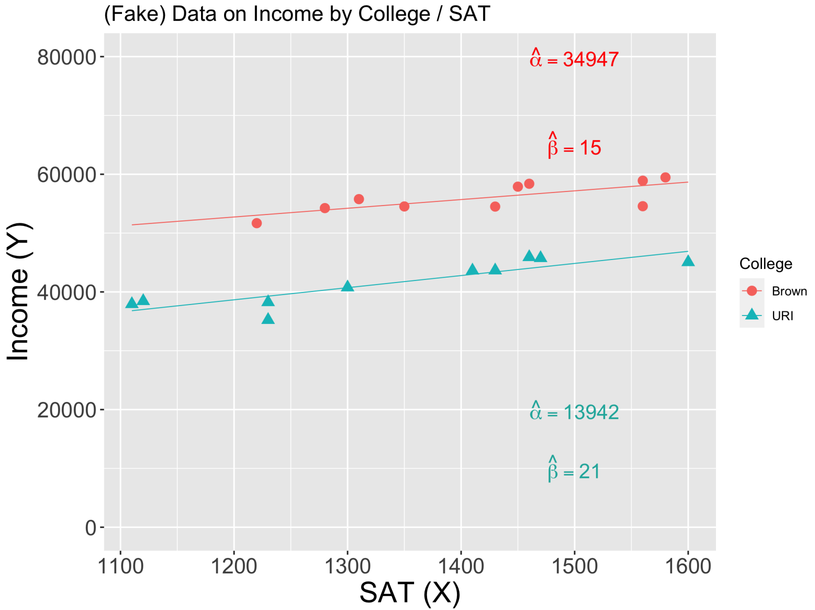
(Fake) Data on Income by College / SAT
$\hat{\alpha}$ = 34947
$\hat{\beta}$ = 15
Income (Y)
SAT (X)
80000
60000
40000
20000
0
1100
1200
1300
1400
1500
1600
College
Brown
URI



(Fake) Data on Income by College / SAT
$\hat{\alpha} = 34947$
$\hat{\beta} = 15$
$\hat{\alpha} = 13942$
$\hat{\beta} = 21$
Income (Y)
SAT (X)
80000
60000
40000
20000
0
1100
1200
1300
1400
1500
1600
College
Brown
URI



- What is the estimated value of $E[Y_i | D_i = 1, X_i = 1350]$?



- What is the estimated value of $E[Y_i|D_i = 1, X_i = 1350]$?
  $\hat{\alpha} + \hat{\beta} \cdot 1350 = 34947 + 15 \cdot 1350 = 55197$.

Consistency of OLS

- We can now use our results for (functions of) sample averages to show that $\hat{\beta}$ is consistent for β, i.e. $\hat{\beta} \to_p \beta$.

Consistency of OLS

- We can now use our results for (functions of) sample averages to show that $\hat{\beta}$ is consistent for β, i.e. $\hat{\beta} \rightarrow_p \beta$.

- We have that

$$\hat{\beta} = \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$

Consistency of OLS

- We can now use our results for (functions of) sample averages to show that $\hat{\beta}$ is consistent for β, i.e. $\hat{\beta} \to_p \beta$.

- We have that

$$\begin{aligned}\hat{\beta} &= \left(\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2\right)^{-1}\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) \\ &= \left(\frac{1}{N}\sum_{i=1}^{N}X_i^2 - \bar{X}^2\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}X_iY_i - \bar{X}\bar{Y}\right)\end{aligned}$$

Consistency of OLS

- We can now use our results for (functions of) sample averages to show that $\hat{\beta}$ is consistent for β, i.e. $\hat{\beta} \rightarrow_p \beta$.

- We have that

$$\begin{aligned}
\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) \\
&= \left(\frac{1}{N} \sum_{i=1}^{N} X_i^2 - \bar{X}^2 \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y} \right) \\
&\rightarrow_p \left(E[X_i^2] - E[X_i]^2 \right)^{-1} \left(E[X_i Y_i] - E[X_i]E[Y_i] \right)
\end{aligned}$$

Consistency of OLS

- We can now use our results for (functions of) sample averages to show that $\hat{\beta}$ is consistent for β, i.e. $\hat{\beta} \to_p \beta$.

- We have that

$$\begin{aligned}
\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) \\
&= \left(\frac{1}{N} \sum_{i=1}^{N} X_i^2 - \bar{X}^2 \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y} \right) \\
&\to_p \left(E[X_i^2] - E[X_i]^2 \right)^{-1} (E[X_i Y_i] - E[X_i]E[Y_i]) \\
&= Var(X_i)^{-1} Cov(X_i, Y_i)
\end{aligned}$$

Consistency of OLS

- We can now use our results for (functions of) sample averages to show that $\hat{\beta}$ is consistent for β, i.e. $\hat{\beta} \to_p \beta$.

- We have that

$$\begin{aligned}
\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) \\
&= \left(\frac{1}{N} \sum_{i=1}^{N} X_i^2 - \bar{X}^2 \right)^{-1} \left(\frac{1}{N} \sum_{i=1}^{N} X_i Y_i - \bar{X}\bar{Y} \right) \\
&\to_p \left(E[X_i^2] - E[X_i]^2 \right)^{-1} \left(E[X_i Y_i] - E[X_i]E[Y_i] \right) \\
&= Var(X_i)^{-1} Cov(X_i, Y_i) = \beta
\end{aligned}$$

Consistency of OLS

- We can now use our results for (functions of) sample averages to show that $\hat{\beta}$ is consistent for β, i.e. $\hat{\beta} \to_p \beta$.

- We have that

$$\begin{aligned}
\hat{\beta} &= \left(\frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})^2\right)^{-1} \frac{1}{N}\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y}) \\
&= \left(\frac{1}{N}\sum_{i=1}^{N}X_i^2 - \bar{X}^2\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N}X_iY_i - \bar{X}\bar{Y}\right) \\
&\to_p \left(E[X_i^2] - E[X_i]^2\right)^{-1}(E[X_iY_i] - E[X_i]E[Y_i]) \\
&= Var(X_i)^{-1}Cov(X_i, Y_i) = \beta
\end{aligned}$$

- Analogously, we can show that $\hat{\alpha} \to_p \alpha$.

Asymptotic Distribution for OLS

- Our OLS estimates $\hat{\alpha}, \hat{\beta}$ are continuous functions of sample means.

Asymptotic Distribution for OLS

- Our OLS estimates $\hat{\alpha}, \hat{\beta}$ are continuous functions of sample means.

- We can therefore use the Central Limit Theorem and Continuous Mapping Theorem to show that they are asymptotically normally distributed

Asymptotic Distribution for OLS

- Our OLS estimates $\hat{\alpha}, \hat{\beta}$ are continuous functions of sample means.
- We can therefore use the Central Limit Theorem and Continuous Mapping Theorem to show that they are asymptotically normally distributed
- In particular, we will show that

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow_d N(0, \sigma^2),$$

where

$$\sigma^2 = \frac{Var((X_i - E[X_i])\varepsilon_i)}{Var(X_i)^2}$$

Asymptotic Distribution for OLS

- Our OLS estimates $\hat{\alpha}, \hat{\beta}$ are continuous functions of sample means.

- We can therefore use the Central Limit Theorem and Continuous Mapping Theorem to show that they are asymptotically normally distributed

- In particular, we will show that

$$\sqrt{N}(\hat{\beta} - \beta) \to_d N(0, \sigma^2),$$

  where

$$\sigma^2 = \frac{Var((X_i - E[X_i])\varepsilon_i)}{Var(X_i)^2}$$

- This is useful because we can then form CIs for β of the form $\hat{\beta} \pm 1.96\hat{\sigma}/\sqrt{N}$, where $\hat{\sigma}$ is our estimate of σ.

Deriving the Asymptotic Distribution for OLS

- Define the **regression residual** $\varepsilon_i = Y_i - (\alpha + X_i\beta)$, implying

$$Y_i = \alpha + X_i\beta + \varepsilon_i$$

Deriving the Asymptotic Distribution for OLS

- Define the **regression residual** $\varepsilon_i = Y_i - (\alpha + X_i\beta)$, implying

$$Y_i = \alpha + X_i\beta + \varepsilon_i$$

- The first-order conditions we derived for (α, β) imply this residual is mean-zero and **orthogonal** to the **regressor**: $E[\varepsilon_i] = E[X_i\varepsilon_i] = 0$

Deriving the Asymptotic Distribution for OLS

- Define the **regression residual** $\varepsilon_i = Y_i - (\alpha + X_i\beta)$, implying

$$Y_i = \alpha + X_i\beta + \varepsilon_i$$

- The first-order conditions we derived for (α, β) imply this residual is mean-zero and **orthogonal** to the **regressor**: $E[\varepsilon_i] = E[X_i\varepsilon_i] = 0$

- Taking means, $\bar{Y} = \alpha + \bar{X}\beta + \bar{\varepsilon}$.

Deriving the Asymptotic Distribution for OLS

- Define the **regression residual** $\varepsilon_i = Y_i - (\alpha + X_i\beta)$, implying

$$Y_i = \alpha + X_i\beta + \varepsilon_i$$

- The first-order conditions we derived for (α, β) imply this residual is mean-zero and **orthogonal** to the **regressor**: $E[\varepsilon_i] = E[X_i\varepsilon_i] = 0$

- Taking means, $\bar{Y} = \alpha + \bar{X}\beta + \bar{\varepsilon}$. So $Y_i - \bar{Y} = (X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})$

Asymptotic Distribution for OLS (cont.)

- We just derived that $Y_i - \bar{Y} = (X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})$.
- Thus,

$$\hat{\beta} = \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$

$$=$$

Asymptotic Distribution for OLS (cont.)

- We just derived that $Y_i - \bar{Y} = (X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})$.
- Thus,

$$\hat{\beta} = \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y})$$

$$= \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})((X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon}))$$

Asymptotic Distribution for OLS (cont.)

- We just derived that $Y_i - \bar{Y} = (X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})$.
- Thus,

$$\begin{aligned}
\hat{\beta} &= \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(Y_i - \bar{Y}) \\
&= \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})((X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})) \\
&= \beta + \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon})
\end{aligned}$$

Asymptotic Distribution for OLS (cont.)

- Hence,

$$\sqrt{N}(\hat{\beta} - \beta) = \left(\frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})^2 \right)^{-1} \sqrt{N} \frac{1}{N} \sum_{i=1}^{N} (X_i - \bar{X})(\varepsilon_i - \bar{\varepsilon})$$

Asymptotic Distribution for OLS (cont.)

- Hence,

$$\begin{aligned}\sqrt{N}(\hat{\beta}-\beta) &= \left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1}\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})(\varepsilon_i-\bar{\varepsilon}) \\ &= \left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1}\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(X_i-E[X_i])\varepsilon_i \\ &- \left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1}\left(\bar{\varepsilon}\sqrt{N}(\bar{X}-E[X_i])\right)\end{aligned}$$

- By LLN and CMT, $\left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1} \to_p Var(X_i)^{-1}$

Asymptotic Distribution for OLS (cont.)

- Hence,

$$\sqrt{N}(\hat{\beta}-\beta) = \left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1}\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})(\varepsilon_i-\bar{\varepsilon})$$
$$= \left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1}\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(X_i-E[X_i])\varepsilon_i$$
$$- \left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1}\left(\bar{\varepsilon}\sqrt{N}(\bar{X}-E[X_i])\right)$$

- By LLN and CMT, $\left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1} \to_p Var(X_i)^{-1}$

- By CLT, $\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(X_i-E[X_i])\varepsilon_i \to_d \mathrm{N}(0, Var((X_i-E[X_i])\varepsilon_i))$.

Asymptotic Distribution for OLS (cont.)

- Hence,

$$\sqrt{N}(\hat{\beta}-\beta) = \left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1}\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})(\varepsilon_i-\bar{\varepsilon})$$
$$= \left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1}\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(X_i-E[X_i])\varepsilon_i$$
$$- \left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1}\left(\bar{\varepsilon}\sqrt{N}(\bar{X}-E[X_i])\right)$$

- By LLN and CMT, $\left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1} \to_p Var(X_i)^{-1}$
- By CLT, $\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(X_i-E[X_i])\varepsilon_i \to_d \mathrm{N}(0, Var((X_i-E[X_i])\varepsilon_i))$.
- By LLN, CLT, and Slutsky, $\bar{\varepsilon}\sqrt{N}(\bar{X}-E[X_i]) \to_d 0\times\mathrm{N}(0, Var(X_i)) = 0$

Finishing the Asymptotics (!)

- Putting all the pieces together, we see that

$$\sqrt{N}(\hat{\beta} - \beta) \rightarrow_d N(0, \sigma^2),$$

where

$$\sigma^2 = \frac{Var((X_i - E[X_i])\varepsilon_i)}{Var(X_i)^2}$$

- As before, we can estimate the variance σ^2 using sample averages,

$$\hat{\sigma}^2 = \frac{\frac{1}{N}\sum_i((X_i - \bar{X})\hat{\varepsilon}_i)^2}{\left(\frac{1}{N}\sum_i(X_i - \bar{X})^2\right)^2}, \text{ where } \hat{\varepsilon}_i = Y_i - (\hat{\alpha} + X_i\hat{\beta})$$

Finishing the Asymptotics (!)

- Putting all the pieces together, we see that

$$\sqrt{N}(\hat{\beta} - \beta) \to_d N(0, \sigma^2),$$

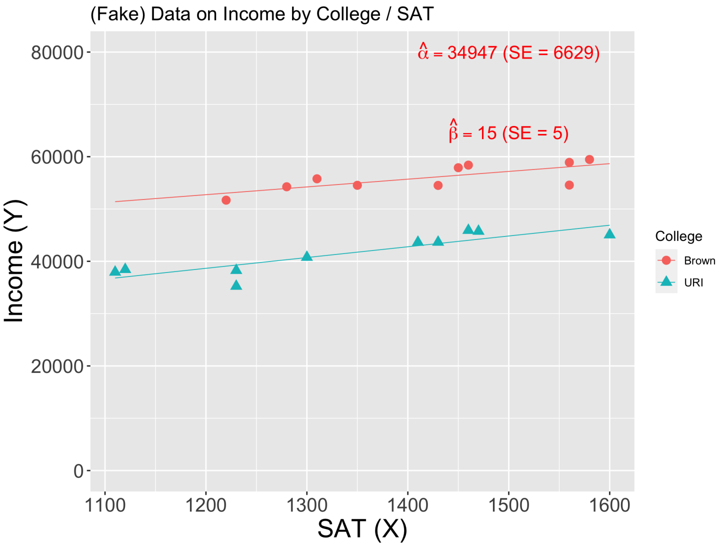
where

$$\sigma^2 = \frac{Var((X_i - E[X_i])\varepsilon_i)}{Var(X_i)^2}$$

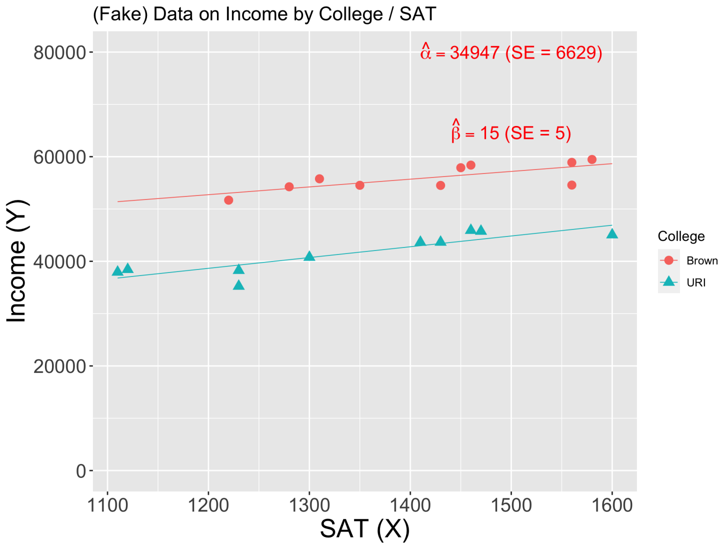
- As before, we can estimate the variance σ^2 using sample averages,

$$\hat{\sigma}^2 = \frac{\frac{1}{N}\sum_i ((X_i - \bar{X})\hat{\varepsilon}_i)^2}{\left(\frac{1}{N}\sum_i (X_i - \bar{X})^2\right)^2}, \text{ where } \hat{\varepsilon}_i = Y_i - (\hat{\alpha} + X_i\hat{\beta})$$

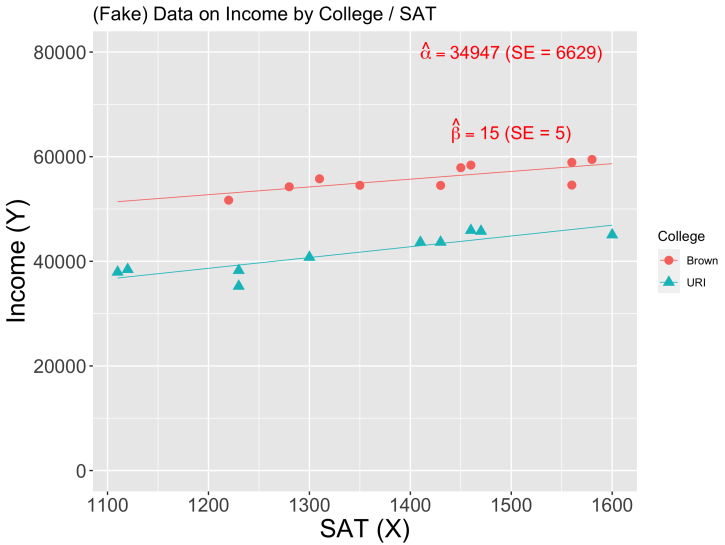
- Can do similar steps to show $\hat{\alpha}$ is asymptotically normally distributed as well. (We'll show formulas later!)



(Fake) Data on Income by College / SAT
α̂ = 34947 (SE = 6629)
β̂ = 15 (SE = 5)
80000
60000
40000
20000
0
Income (Y)
1100
1200
1300
1400
1500
1600
SAT (X)
College
Brown
URI



- A CI for β is $\hat{\beta} \pm 1.96 \times SE$



- A CI for β is $\hat{\beta} \pm 1.96 \times SE \approx [5, 25]$

Aside on notation/terminology

- Oftentimes people will say: consider the (population) regression

$$Y_i = \alpha + \beta D_i + \varepsilon_i \tag{1}$$

- What they mean is: "define $(\alpha, \beta) = \arg\min_{a,b} E[(Y_i - (a + bX_i))^2]$"

- (α, β) are referred to as the "population regression coefficients"

Aside on notation/terminology

- Oftentimes people will say: consider the (population) regression

$$Y_i = \alpha + \beta D_i + \varepsilon_i \tag{1}$$

- What they mean is: "define $(\alpha, \beta) = \arg\min_{a,b} E[(Y_i - (a + bX_i))^2]$"
- (α, β) are referred to as the "population regression coefficients"
- Likewise, people will say "We estimate equation (1) by OLS" to mean that they compute the sample analogs to α, β via OLS, i.e. $\hat{\alpha}, \hat{\beta}$.

Outline

1. Population Regression✓
2. Sample Regressions (OLS)✓
3. Putting Regression into Practice

Using Regressions to Analyze RCTs

- Recall that when we have an experiment, the average treatment effect is identified by a different in means:

$$\tau = E[Y_i(1) - Y_i(0)] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

Using Regressions to Analyze RCTs

- Recall that when we have an experiment, the average treatment effect is identified by a different in means:

$$\tau = E[Y_i(1) - Y_i(0)] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

- Observe that we can write:

$$E[Y_i|D_i = d] = E[Y_i|D_i = 0] + (E[Y_i|D_i = 1] - E[Y_i|D_i = 0]) \cdot d$$

Using Regressions to Analyze RCTs

- Recall that when we have an experiment, the average treatment effect is identified by a different in means:

$$\tau = E[Y_i(1) - Y_i(0)] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

- Observe that we can write:

$$\begin{aligned} E[Y_i|D_i = d] &= E[Y_i|D_i = 0] + (E[Y_i|D_i = 1] - E[Y_i|D_i = 0]) \cdot d \\ &= \alpha + \beta d \end{aligned}$$

Using Regressions to Analyze RCTs

- Recall that when we have an experiment, the average treatment effect is identified by a different in means:

$$\tau = E[Y_i(1) - Y_i(0)] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

- Observe that we can write:

$$\begin{aligned} E[Y_i|D_i = d] &= E[Y_i|D_i = 0] + (E[Y_i|D_i = 1] - E[Y_i|D_i = 0]) \cdot d \\ &= \alpha + \beta d \end{aligned}$$

- Thus, the CEF $E[Y_i|D_i = d]$ is linear in d, and the slope coefficient β is exactly the estimand which identifies the ATE in an experiment!

Using Regressions to Analyze RCTs

- Recall that when we have an experiment, the average treatment effect is identified by a different in means:

$$\tau = E[Y_i(1) - Y_i(0)] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

- Observe that we can write:

$$\begin{aligned} E[Y_i|D_i = d] &= E[Y_i|D_i = 0] + (E[Y_i|D_i = 1] - E[Y_i|D_i = 0]) \cdot d \\ &= \alpha + \beta d \end{aligned}$$

- Thus, the CEF $E[Y_i|D_i = d]$ is linear in d, and the slope coefficient β is exactly the estimand which identifies the ATE in an experiment!

- Analogously, the OLS slope coefficient $\hat{\beta}$ is the difference in sample means which estimates the ATE: $\hat{\beta} = \bar{Y}_1 - \bar{Y}_0 = \hat{\tau}$.

Using Regressions to Analyze RCTs

- Recall that when we have an experiment, the average treatment effect is identified by a different in means:

$$\tau = E[Y_i(1) - Y_i(0)] = E[Y_i|D_i = 1] - E[Y_i|D_i = 0]$$

- Observe that we can write:

$$\begin{aligned} E[Y_i|D_i = d] &= E[Y_i|D_i = 0] + (E[Y_i|D_i = 1] - E[Y_i|D_i = 0]) \cdot d \\ &= \alpha + \beta d \end{aligned}$$

- Thus, the CEF $E[Y_i|D_i = d]$ is linear in d, and the slope coefficient β is exactly the estimand which identifies the ATE in an experiment!

- Analogously, the OLS slope coefficient $\hat{\beta}$ is the difference in sample means which estimates the ATE: $\hat{\beta} = \bar{Y}_1 - \bar{Y}_0 = \hat{\tau}$.

- We can thus use OLS as a convenient tool for estimating the ATE and getting standard errors

Example - WorkAdvance

- Background: gaps between college-educated and non-college educated workers have widened over time

- Yet not everyone thrives in a traditional college background

- **WorkAdvance** is a job-training program intended to provide people with certifiable skills in high-wage industries (e.g. IT, healthcare manufacturing)

- MDRC conducted a randomized trial that randomized access to the training program among people who passed the initial screening

TABLE 1

WORKADVANCE PROVIDERS AND SAMPLE COMPOSITION AT BASELINE

	PER SCHOLAS	ST. NICKS ALLIANCE	MADISON STRATEGIES GROUP	TOWARDS EMPLOYMENT
Provider characteristics				
Location	Bronx, NY	Brooklyn, NY	Tulsa, OK	Northeast Ohio
Target sector(s)	Information technology	Environmental remediation	Transportation, manufacturing	Health care, manufacturing
Approach	Training first	Training first	Training and placement first until fall 2012; then mostly training first	Training and placement first until fall 2012; then mostly training first
Sample composition				
Average age	31	35	35	35
Female (%)	13	15	16	59
Some college or more (%)	63	44	58	57
Currently/ever employed (%)	13/96	11/98	27/99	27/97

- Estimate OLS regression:

$$\underbrace{Y_i}_{\text{Earnings 2-3 years later}} = \alpha + \beta \underbrace{D_i}_{\text{Treatment indicator}} + \varepsilon_i$$

- Estimate OLS regression:

$$\underbrace{Y_i}_{\text{Earnings 2-3 years later}} = \alpha + \beta \underbrace{D_i}_{\text{Treatment indicator}} + \varepsilon_i$$

Coefficient	Estimate	SE
$\hat{\alpha}$	14636	425
$\hat{\beta}$	1965	609

- Estimate OLS regression:

$$\underbrace{Y_i}_{\text{Earnings 2-3 years later}} = \alpha + \beta \underbrace{D_i}_{\text{Treatment indicator}} + \varepsilon_i$$

Coefficient	Estimate	SE
$\hat{\alpha}$	14636	425
$\hat{\beta}$	1965	609

- What is the estimated treatment effect?

- Estimate OLS regression:

$$\underbrace{Y_i}_{\text{Earnings 2-3 years later}} = \alpha + \beta \underbrace{D_i}_{\text{Treatment indicator}} + \varepsilon_i$$

Coefficient	Estimate	SE
$\hat{\alpha}$	14636	425
$\hat{\beta}$	1965	609

- What is the estimated treatment effect? $\hat{\beta} = 1965$

- What is a CI for the treatment effects?

- Estimate OLS regression:

$$\underbrace{Y_i}_{\text{Earnings 2-3 years later}} = \alpha + \beta \underbrace{D_i}_{\text{Treatment indicator}} + \varepsilon_i$$

Coefficient	Estimate	SE
$\hat{\alpha}$	14636	425
$\hat{\beta}$	1965	609

- What is the estimated treatment effect? $\hat{\beta} = 1965$

- What is a CI for the treatment effects?
  $\hat{\beta} \pm 1.96 \times SE_{\beta} =$

- Estimate OLS regression:

$$\underbrace{Y_i}_{\text{Earnings 2-3 years later}} = \alpha + \beta \underbrace{D_i}_{\text{Treatment indicator}} + \varepsilon_i$$

Coefficient	Estimate	SE
$\hat{\alpha}$	14636	425
$\hat{\beta}$	1965	609

- What is the estimated treatment effect? $\hat{\beta} = 1965$

- What is a CI for the treatment effects?
  $\hat{\beta} \pm 1.96 \times SE_{\beta} = 1965 \pm 1.96 \times 609$

- Estimate OLS regression:

$$\underbrace{Y_i}_{\text{Earnings 2-3 years later}} = \alpha + \beta \underbrace{D_i}_{\text{Treatment indicator}} + \varepsilon_i$$

Coefficient	Estimate	SE
$\hat{\alpha}$	14636	425
$\hat{\beta}$	1965	609

- What is the estimated treatment effect? $\hat{\beta} = 1965$

- What is a CI for the treatment effects?
  $\hat{\beta} \pm 1.96 \times SE_{\beta} = 1965 \pm 1.96 \times 609 = [771, 3159]$

- What is the estimated control mean?

- Estimate OLS regression:

$$\underbrace{Y_i}_{\text{Earnings 2-3 years later}} = \alpha + \beta \underbrace{D_i}_{\text{Treatment indicator}} + \varepsilon_i$$

Coefficient	Estimate	SE
$\hat{\alpha}$	14636	425
$\hat{\beta}$	1965	609

- What is the estimated treatment effect? $\hat{\beta} = 1965$

- What is a CI for the treatment effects?
  $\hat{\beta} \pm 1.96 \times SE_{\beta} = 1965 \pm 1.96 \times 609 = [771, 3159]$

- What is the estimated control mean? $\hat{\alpha} = 14636$

Roadmap

- **What we know how to do:** Estimate and test hypotheses about population means using sample means
- **What we want to do:** Estimate approximations to the CEF and test hypotheses about them

How can we use what know to do what we want?

- 1) Assume the CEF takes a particular form, e.g. linear:

$$E[Y_i|X_i = x] = \alpha + x\beta \qquad \checkmark$$

- 2) Show that under this assumption, α and β can be represented as functions of population means. ✓
- 3) Use our tools for estimating population means using sample means to estimate α, β and test hypotheses about them. ✓
- 4) Argue that even if our assumption about the form of the CEF is wrong, the parameters α, β may provide a “good” approximation.

Regressions as Approximations

- So far we've assumed that the conditional expectation is linear: $E[Y_i|X_i = x] = \alpha + \beta x$
- What if the true CEF is not linear?!

Regressions as Approximations

- So far we've assumed that the conditional expectation is linear: $E[Y_i|X_i = x] = \alpha + \beta x$
- What if the true CEF is not linear?!
- Claim: if CEF is not linear, then OLS still gives us the "best linear approximation" to the CEF

Regressions as Approximations

- So far we've assumed that the conditional expectation is linear: $E[Y_i|X_i = x] = \alpha + \beta x$
- What if the true CEF is not linear?!
- Claim: if CEF is not linear, then OLS still gives us the "best linear approximation" to the CEF
- What we mean by this is that the α, β of OLS minimize

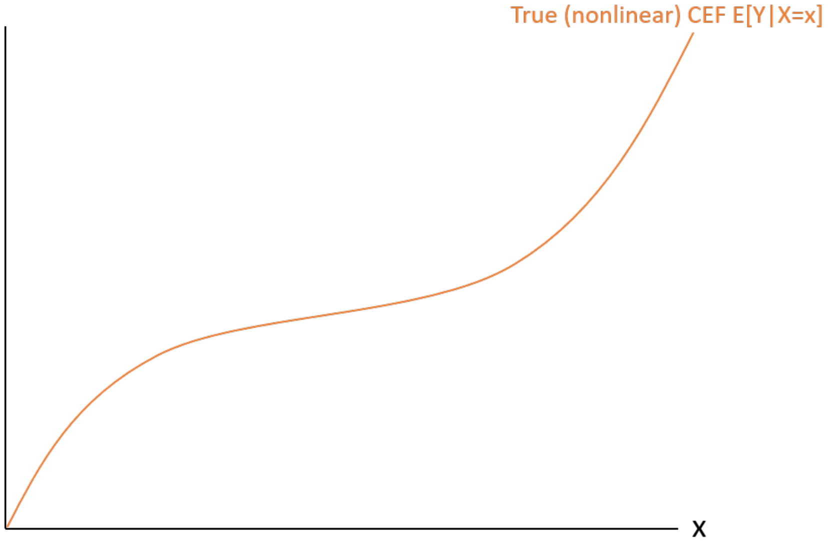
$$min_{\alpha,\beta} E[(E[Y|X] - (\alpha + \beta X))^2]$$

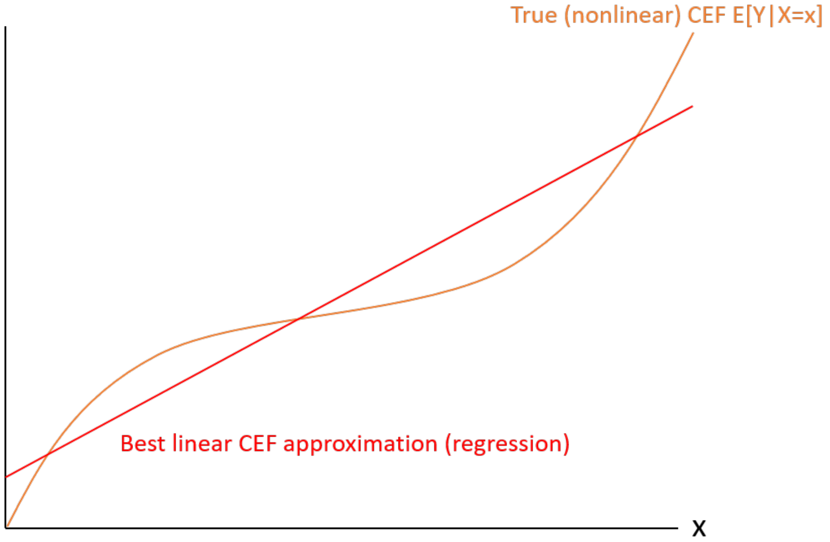
Regressions as Approximations

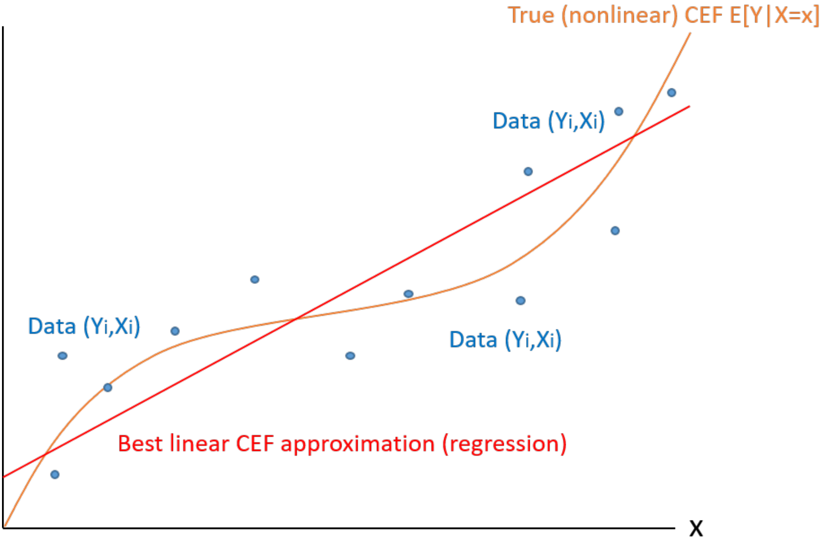
- So far we've assumed that the conditional expectation is linear: $E[Y_i|X_i = x] = \alpha + \beta x$
- What if the true CEF is not linear?!
- Claim: if CEF is not linear, then OLS still gives us the "best linear approximation" to the CEF
- What we mean by this is that the α, β of OLS minimize

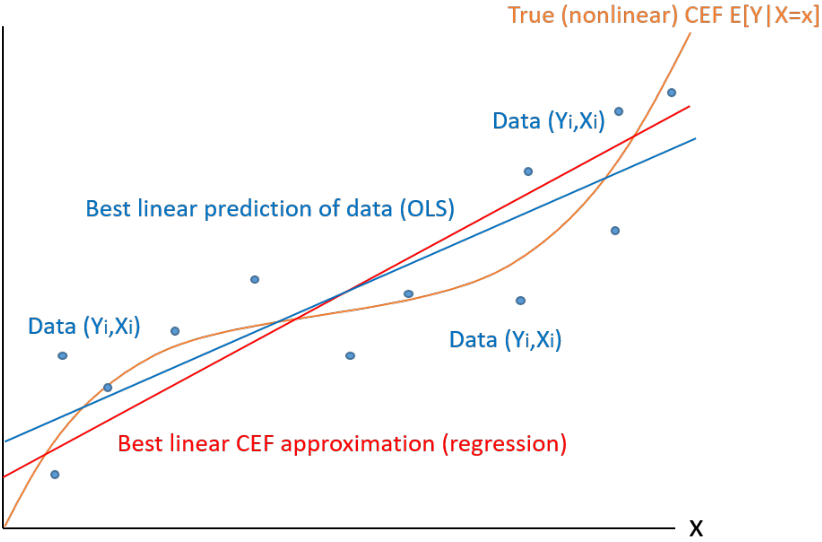
$$min_{\alpha,\beta} E[(E[Y|X] - (\alpha + \beta X))^2]$$

  That is, we get the linear function that's "closest" to the CEF in terms of mean-squared error


True (nonlinear) CEF E[Y|X=x]
X


True (nonlinear) CEF E[Y|X=x]
Best linear CEF approximation (regression)
X


True (nonlinear) CEF E[Y|X=x]
Data (Yi,Xi)
Data (Yi,Xi)
Data (Yi,Xi)
Best linear CEF approximation (regression)
X


True (nonlinear) CEF E[Y|X=x]
Data (Yi,Xi)
Best linear prediction of data (OLS)
Data (Yi,Xi)
Data (Yi,Xi)
Best linear CEF approximation (regression)
X

Proof of OLS as Best Approximation

- We solved for the α, β to minimize $E[(Y-(\alpha+\beta X))^2]$.

Proof of OLS as Best Approximation

- We solved for the α, β to minimize $E[(Y - (\alpha + \beta X))^2]$.
- Let $\mu(x) = E[Y|X = x]$. Then we have

Proof of OLS as Best Approximation

- We solved for the α, β to minimize $E[(Y-(\alpha+\beta X))^2]$.
- Let $\mu(x) = E[Y|X=x]$. Then we have

$$E[(Y-(\alpha+\beta X))^2] = E[(Y-\mu(X)+\mu(X)-(\alpha+\beta X))^2]$$

Proof of OLS as Best Approximation

- We solved for the α, β to minimize $E[(Y-(\alpha+\beta X))^2]$.
- Let $\mu(x) = E[Y|X=x]$. Then we have

$$\begin{aligned} E[(Y-(\alpha+\beta X))^2] &= E[(Y-\mu(X)+\mu(X)-(\alpha+\beta X))^2] \\ &= E[(Y-\mu(X))^2]+E[(\mu(X)-(\alpha+\beta X))^2] \\ &\quad +2E[(Y-\mu(X))(\mu(X)-(\alpha+\beta X))] \end{aligned}$$

Proof of OLS as Best Approximation

- We solved for the α, β to minimize $E[(Y-(\alpha+\beta X))^2]$.
- Let $\mu(x) = E[Y|X=x]$. Then we have

$$\begin{aligned} E[(Y-(\alpha+\beta X))^2] &= E[(Y-\mu(X)+\mu(X)-(\alpha+\beta X))^2] \\ &= E[(Y-\mu(X))^2] + E[(\mu(X)-(\alpha+\beta X))^2] \\ &\quad + 2E[(Y-\mu(X))(\mu(X)-(\alpha+\beta X))] \end{aligned}$$

- By the LIE,

$$E[(Y-\mu(X))(\mu(X)-(\alpha+\beta X))] =$$

Proof of OLS as Best Approximation

- We solved for the α, β to minimize $E[(Y-(\alpha+\beta X))^2]$.
- Let $\mu(x) = E[Y|X=x]$. Then we have

$$\begin{aligned} E[(Y-(\alpha+\beta X))^2] &= E[(Y-\mu(X)+\mu(X)-(\alpha+\beta X))^2] \\ &= E[(Y-\mu(X))^2] + E[(\mu(X)-(\alpha+\beta X))^2] \\ &+ 2E[(Y-\mu(X))(\mu(X)-(\alpha+\beta X))] \end{aligned}$$

- By the LIE,

$$E[(Y-\mu(X))(\mu(X)-(\alpha+\beta X))] =$$
$$E[E[(Y-\mu(X))(\mu(X)-(\alpha+\beta X))|X]] =$$

Proof of OLS as Best Approximation

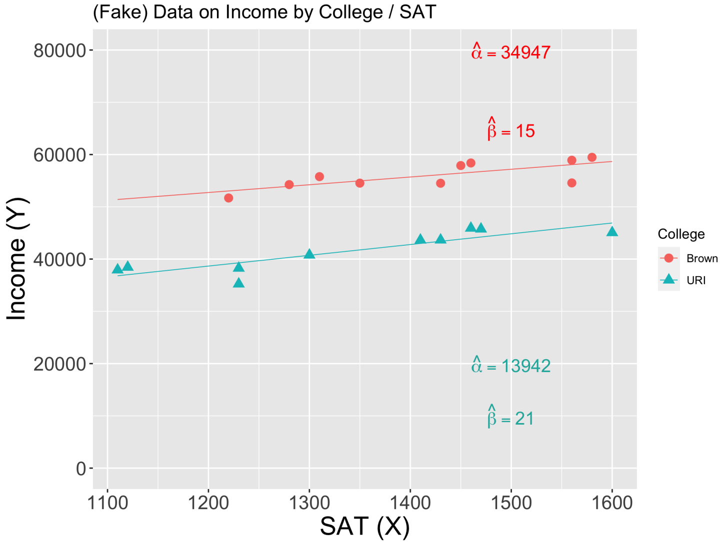
- We solved for the α, β to minimize $E[(Y-(\alpha+\beta X))^2]$.
- Let $\mu(x) = E[Y|X=x]$. Then we have

$$\begin{aligned} E[(Y-(\alpha+\beta X))^2] &= E[(Y-\mu(X)+\mu(X)-(\alpha+\beta X))^2] \\ &= E[(Y-\mu(X))^2]+E[(\mu(X)-(\alpha+\beta X))^2] \\ &\quad +2E[(Y-\mu(X))(\mu(X)-(\alpha+\beta X))] \end{aligned}$$

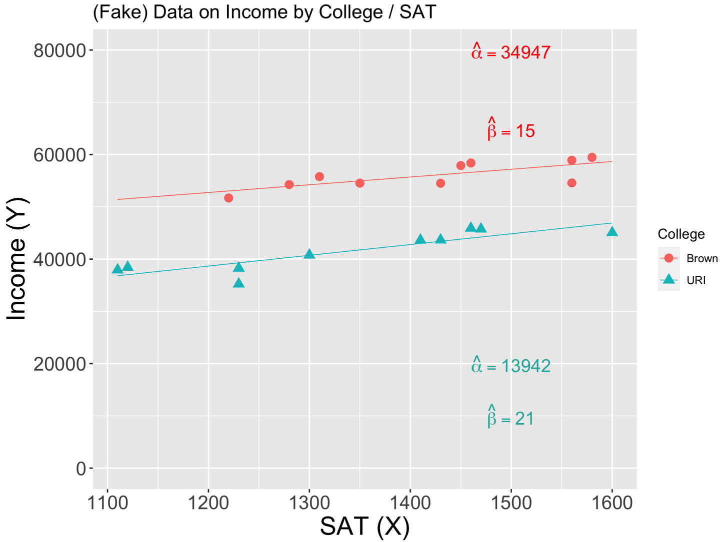
- By the LIE,

$$\begin{aligned} &E[(Y-\mu(X))(\mu(X)-(\alpha+\beta X))] = \\ &E[E[(Y-\mu(X))(\mu(X)-(\alpha+\beta X))|X]] = \\ &E[(\mu(X)-(\alpha+\beta X))\underbrace{E[Y-\mu(X)|X]}_{=0}] = 0 \end{aligned}$$

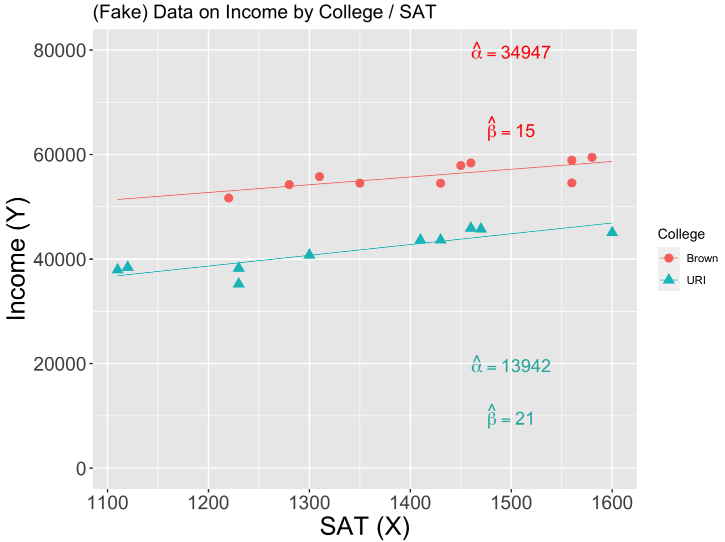
- Hence, $E[(Y-(\alpha+\beta X))^2] = [(Y-\mu(X))^2]+E[\mu(X)-(\alpha+\beta X))^2]$. But the first term doesn't depend on β. So minimizing $E[(Y-(\alpha+\beta X))^2]$ is the same as minimizing $E[\mu(X)-(\alpha+\beta X))^2]$



- $E[Y_i|D_i=1,X_i=x]-E[Y_i|D_i=0,X_i=x]\approx\alpha_1+\beta_1x-(\alpha_0+\beta_0x);$



- $E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x] \approx \alpha_1 + \beta_1 x - (\alpha_0 + \beta_0 x)$; $\hat{\alpha}_1 + \hat{\beta}_1 x - (\hat{\alpha}_0 + \hat{\beta}_0 x) = (34,947 + 15x) - (13,942 + 21x) = 21,005 - 6x$



- $E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x] \approx \alpha_1 + \beta_1 x - (\alpha_0 + \beta_0 x)$; $\hat{\alpha}_1 + \hat{\beta}_1 x - (\hat{\alpha}_0 + \hat{\beta}_0 x) = (34,947 + 15x) - (13,942 + 21x) = 21,005 - 6x$
- So w/conditional ignorability, $ATE = E[CATE(X_i)] \approx 21,005 - 6E[X_i]$