第 4 章：回帰分析入門

Jonathan Roth

数理計量経済学 I
ブラウン大学

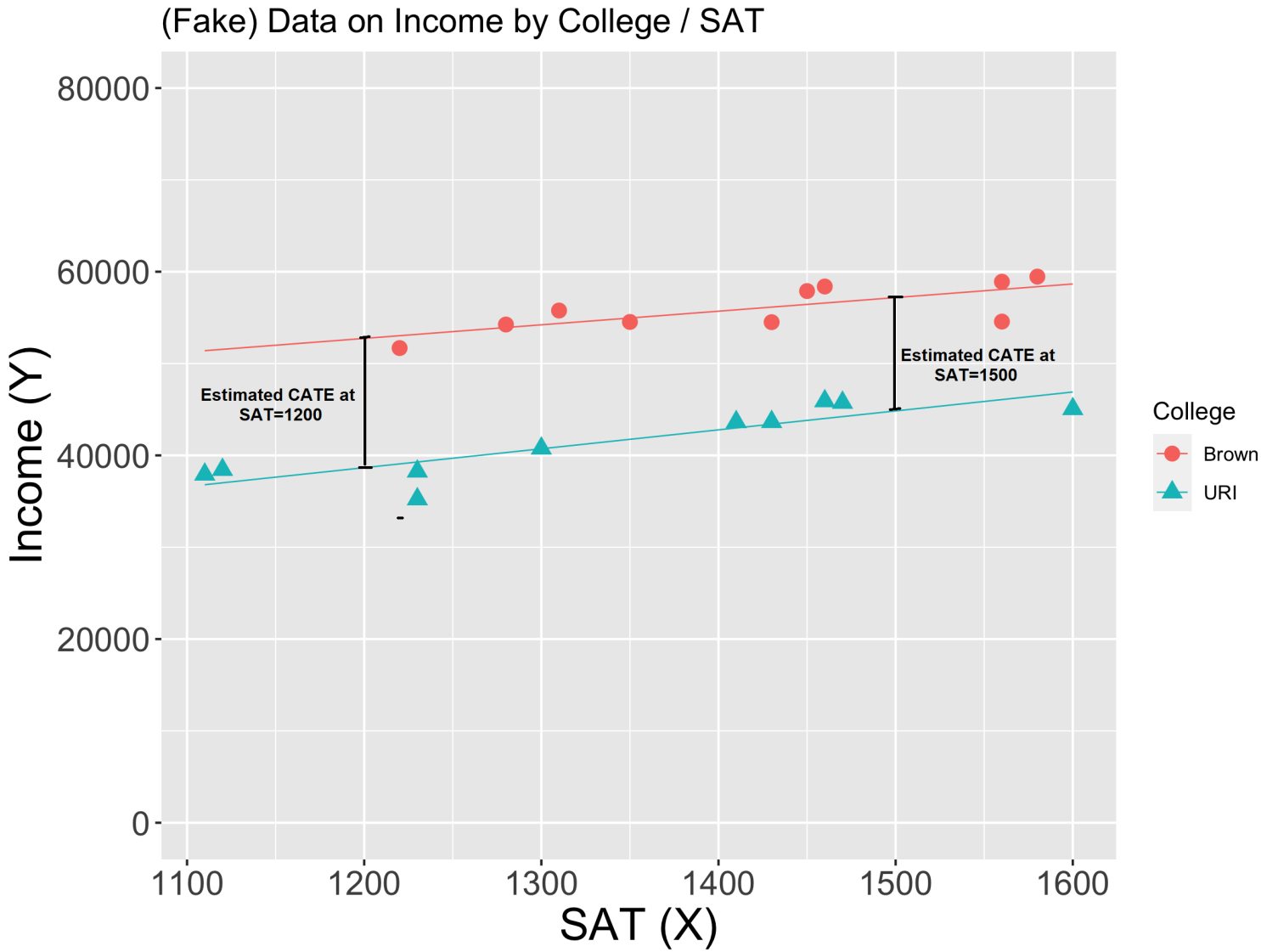
モチベーション

- 条件付き非交絡性の下では、処置群と対照群の結果の平均を X_i で条件付けて比較することで、条件付き平均処置効果（CATE）を学べることを示しました：

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

- X_i が離散的で、各 x の値に対して多くの観測値がある（N_x が大きい）場合、中心極限定理を用いて各条件付き期待値を推定し、「推論」を行う方法を示しました。

- しかし、X_i が連続量である場合はどうでしょうか？



- これらの CEF の推定値があれば、任意の x において $CATE(x)$ を推定できます。

アウトライン

1. 母集団回帰
2. 標本回帰 (OLS)
3. 回帰分析の実践への応用

回帰分析の導入

- 回帰（**regression**）の考え方は、特定の関数形式（例：線形、2 次式など）を用いて、ユニット間で外挿を行うことで条件付き期待値関数（CEF）を推定するプロセスを定式化することです。
- これから答えるべきいくつかの未解決の疑問があります：
- 手元にあるサンプルで、どのように CEF を近似すればよいでしょうか？（すなわち、どのようにデータに線を引くか！）
- CEF の推定値に対して、どのように信頼区間を構築し、仮説検定を行えばよいでしょうか？
- もし実際の CEF が、推定で使用した形式（例：線形）でない場合はどうなるでしょうか？
- これからの数回の講義で、これらの疑問すべてに答えていきます！

ロードマップ

- できること：標本平均を用いて、母平均の推定や仮説検定を行う。
- したいこと：CEF の近似を推定し、それらに関する仮説検定を行う。

「できること」を使って「したいこと」を実現するには？

- 1) CEF が特定の形式（例：線形）であると仮定する：

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

- 2) この仮定の下で、α と β が母集団の期待値の関数として表現できることを示す。
- 3) 母平均を標本平均で推定するツールを用いて α, β を推定し、仮説検定を行う。
- 4) たとえ CEF の形式に関する仮定が間違っていても、パラメータ α, β が「良い」近似を提供することを議論する。

「最小二乗」問題

- X_i がスカラーであり、CEF が線形であるとします（これらは後で緩和します）：

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

- 私たちが利用する有用な事実は、これが正しいとき、(α, β) は以下の「最小二乗（least squares）」問題の解になるということです：

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

- これはどこから来るのでしょうか？！

より単純な問題から始める

- (α, β) が「最小二乗」問題の解であることを示すために、まず関連するより単純な問題を考えましょう：
- 以下を最小化する定数 u を見つけたいとします：

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

- どのような定数 u を選ぶべきでしょうか？ 母平均 $\mu = E[Y_i]$ です！
- 証明：
  $E[(Y_i - u)^2]$ を u で微分すると $E[-2(Y_i - u)]$ です。
  微分を 0 と置くと：

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

少し難しい問題

- 次に、以下を最小化する関数 $u(x)$ を選びたいとします：

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

- どのような関数 $u(x)$ を選ぶべきでしょうか？ 条件付き期待値 $u(x) = E[Y|X = x]$ です。

- 証明：
  反復期待値の法則より：

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

  したがって、各 x の値について、以下を最小化する $u(x)$ を選びたいわけです：

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

  しかし、前スライドの議論から、その解は $u(x) = E[Y_i|X_i = x]$ となります。

話を戻すと...

- 関数 $u(x) = E[Y_i|X_i = x]$ が以下を最小化することを示しました：

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

- したがって、もし $E[Y_i|X_i = x] = \alpha + \beta x$ ならば、$u(x) = \alpha + \beta x$ は以下を最小化します：

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

- この最小化はすべての関数 $u(\cdot)$ を対象としており、その中には $a + bx$ という形式の線形関数も含まれます。ゆえに：

$E[(Y_i - (\alpha + \beta X_i))^2] \leq E[(Y_i - (a + bX_i))^2]$（すべての a, b に対して）

- これは、(α, β) が以下を解決することを意味します：

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

これで示したかったことが言えました。

なぜこれが有用なのか

- α, β が以下の解であることを示しました：

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

- これがどう役立つのでしょうか？ 最小化問題を解くことで、α, β を母集団の期待値の関数として表現できるからです。

- a と b で微分し、(α, β) において 0 と置きます：

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

- これで 2 つの未知数に対する 2 つの方程式が得られました。これを使って CEF のパラメータ (α, β) を解くことができます。

最小二乗法の解

- 連立方程式の解は以下の通りです：

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

- これらは母平均の連続関数です！

- したがって、以前の講義で学んだツールを使ってこれらを推定し、CEF に関する仮説検定を行うことができます！

ロードマップ

- できること：標本平均を用いて、母平均の推定や仮説検定を行う。
- したいこと：CEF の近似を推定し、それらに関する仮説検定を行う。

「できること」を使って「したいこと」を実現するには？

- 1) CEF が特定の形式（例：線形）であると仮定する：

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

- 2) この仮定の下で、α と β が母集団の期待値の関数として表現できることを示す。✓
- 3) 母平均を標本平均で推定するツールを用いて α, β を推定し、仮説検定を行う。
- 4) たとえ CEF の形式に関する仮定が間違っていても、パラメータ α, β が「良い」近似を提供することを議論する。

アウトライン

1. 母集団回帰 ✓
2. **標本回帰** (OLS)
3. 回帰分析の実践への応用

回帰係数の推定

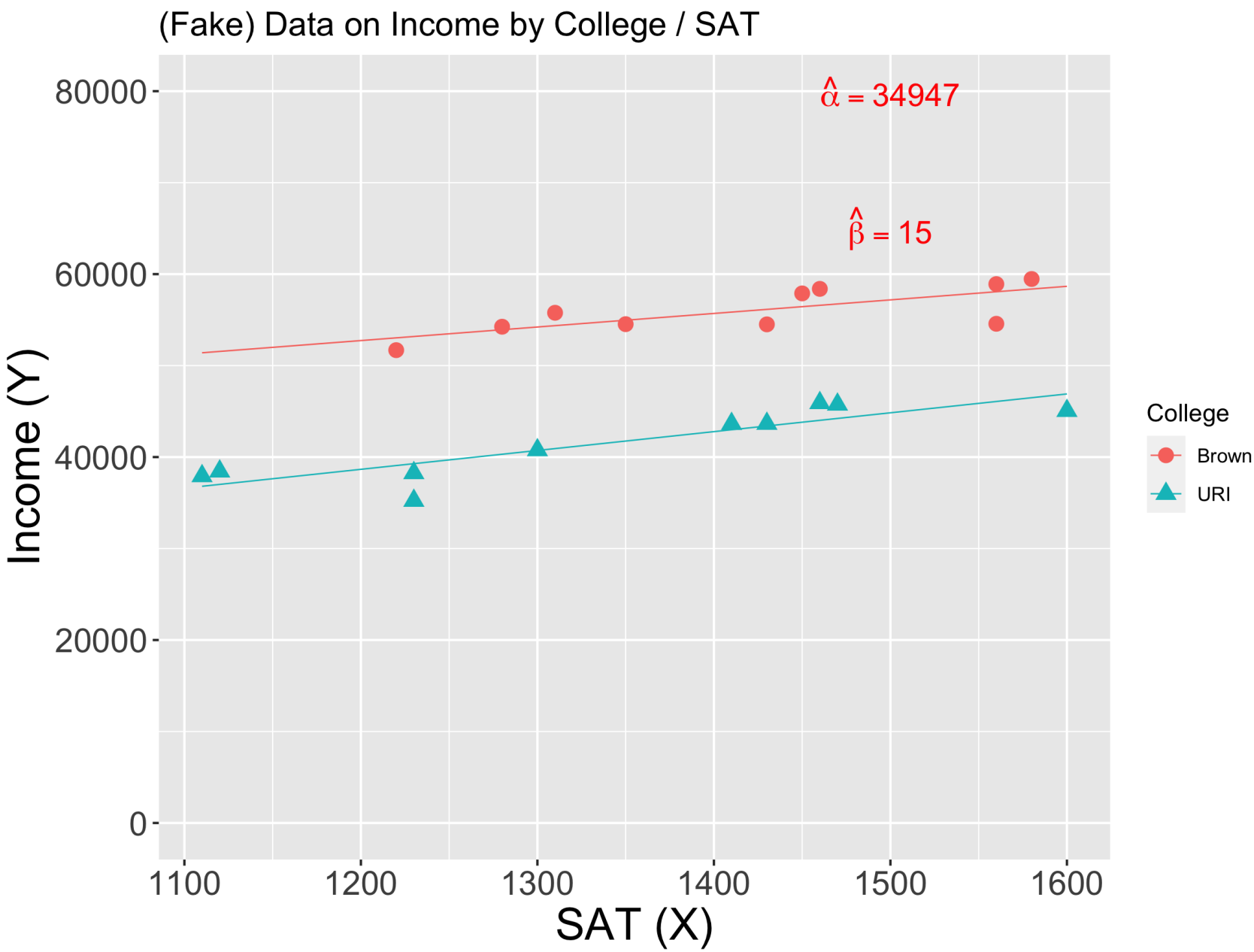
- $E[Y_i \mid X_i = x] = \alpha + \beta x$ のとき、以下が成り立つことを示しました：

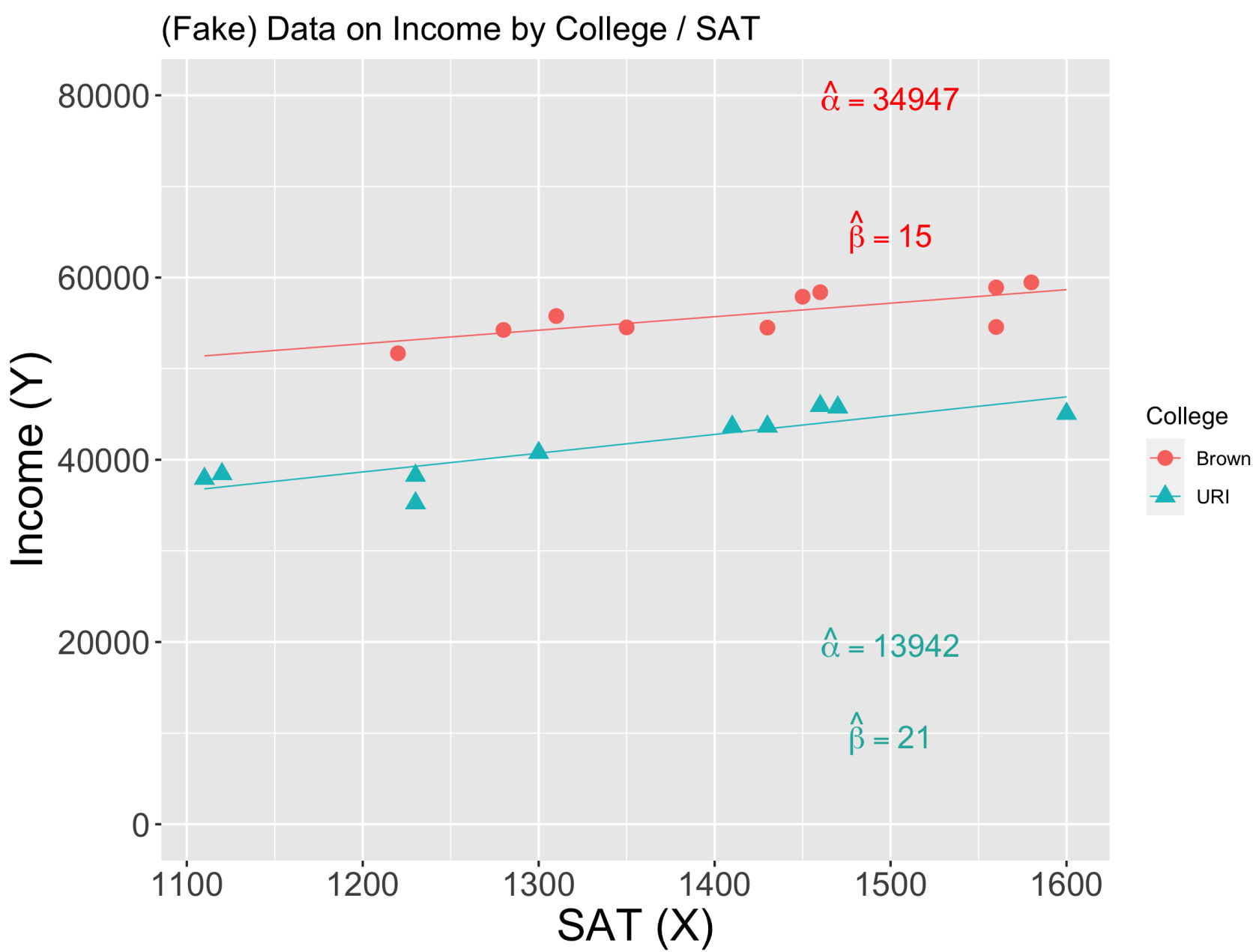
$$\beta = \frac{E[(X_i - E[X_i])(Y_i - E[Y_i])]}{E[(X_i - E[X_i])^2]} = \frac{Cov(X_i, Y_i)}{Var(X_i)}$$
$$\alpha = E[Y_i] - E[X_i]\beta$$

- どのように α, β を推定すればよいでしょうか？
  母平均を標本平均に置き換えればよいのです！

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

- これらの $\hat{\alpha}, \hat{\beta}$ は最小二乗法（Ordinary Least Squares, OLS）係数と呼ばれます。
  - これらは標本の最小化問題 $\min_{a,b} \frac{1}{N}\sum_i (Y_i - (a + bX_i))^2$ の解となります。


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



- $E[Y_i|D_i = 1, X_i = 1350]$ の推定値はいくらでしょうか？
  $\hat{\alpha} + \hat{\beta} \cdot 1350 = 34947 + 15 \cdot 1350 = 55197$ です。

OLS の一致性

- 標本平均（の関数）に関する結果を用いて、$\hat{\beta}$ が β に対して一致性を持つこと、すなわち $\hat{\beta} \to_p \beta$ を示すことができます。

- 以下が成り立ちます：

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

- 同様に、$\hat{\alpha} \to_p \alpha$ も示せます。

OLS の漸近分布

- OLS 推定量 $\hat{\alpha}, \hat{\beta}$ は標本平均の連続関数です。
- したがって、中心極限定理と連続写像定理を用いて、これらが漸近的に正規分布に従うことを示すことができます。
- 具体的には、以下を示します：

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

ここで

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

- これが有用なのは、推定された $\hat{\sigma}$ を用いて、$\hat{\beta} \pm 1.96\hat{\sigma}/\sqrt{N}$ という形式で β の信頼区間を形成できるからです。

OLS の漸近分布の導出

- 回帰残差（regression residual）を $\varepsilon_i = Y_i - (\alpha + X_i\beta)$ と定義します。これは次を意味します：

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

- (α, β) について導出した 1 階の条件（FOC）は、この残差の期待値が 0 であり、説明変数（regressor）と直交することを意味します：
$E[\varepsilon_i] = E[X_i\varepsilon_i] = 0$

- 平均を取ると、$\bar{Y} = \alpha + \bar{X}\beta + \bar{\varepsilon}$ となります。よって
$Y_i - \bar{Y} = (X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})$ です。

OLS の漸近分布（続き）

- $Y_i - \bar{Y} = (X_i - \bar{X})\beta + (\varepsilon_i - \bar{\varepsilon})$ であることが分かりました。
- したがって：

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

OLS の漸近分布（続き）

- ゆえに：

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

- LLN と CMT より、$\left(\frac{1}{N}\sum_{i=1}^{N}(X_i-\bar{X})^2\right)^{-1} \to_p Var(X_i)^{-1}$ です。
- CLT より、$\sqrt{N}\frac{1}{N}\sum_{i=1}^{N}(X_i-E[X_i])\varepsilon_i \to_d \mathrm{N}(0, Var((X_i-E[X_i])\varepsilon_i))$ です。
- LLN, CLT および スルツキーの定理より、$\bar{\varepsilon}\sqrt{N}(\bar{X}-E[X_i]) \to_d 0 \times \mathrm{N}(0, Var(X_i)) = 0$ です。

漸近理論の仕上げ (!)

- これらをまとめると、以下が導かれます：

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

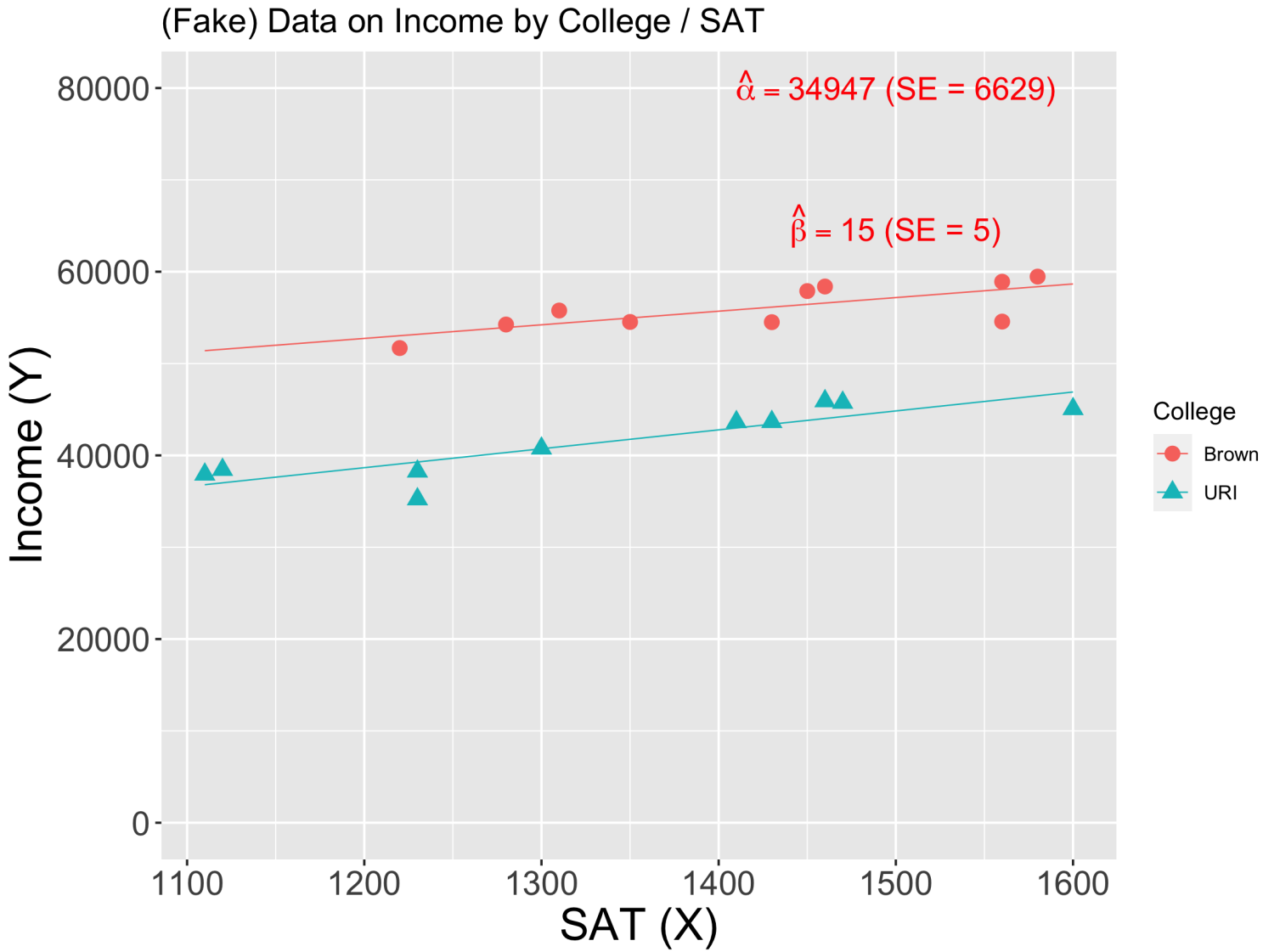
ここで

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

- 以前と同様に、標本平均を用いて分散 σ^2 を推定できます：

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

- 同様の手順で、$\hat{\alpha}$ も漸近的に正規分布に従うことを示せます。（公式は後ほど紹介します！）



- β の信頼区間は $\hat{\beta} \pm 1.96 \times SE \approx [5, 25]$ です。

表記と用語についての補足

- しばしば、「以下の（母集団）回帰を考える」と言われることがあります：

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

- これが意味するのは、「$(\alpha, \beta) = \arg\min_{a,b} E[(Y_i - (a + bX_i))^2]$ と定義する」ということです。

- (α, β) は「母集団回帰係数」と呼ばれます。

- 同様に、「方程式 (1) を OLS で推定する」という表現は、α, β の標本対応物、すなわち $\hat{\alpha}, \hat{\beta}$ を OLS で計算することを意味します。

アウトライン

1. 母集団回帰 ✓
2. 標本回帰 (OLS)✓
3. **回帰分析の実践への応用**

実験データの分析に回帰を用いる

- 実験（RCT）があるとき、平均処置効果は平均の差によって識別されることを思い出してください：

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

- ここで、次のように書けることに注目してください：

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

- したがって、CEF $E[Y_i|D_i = d]$ は d に関して線形であり、傾き係数 β はまさに実験において ATE を識別する推定対象そのものになります！

- 同様に、OLS の傾き係数 $\hat{\beta}$ は ATE を推定する標本平均の差になります：$\hat{\beta} = \bar{Y}_1 - \bar{Y}_0 = \hat{\tau}$。

- ゆえに、OLS を ATE の推定や標準誤差の算出のための便利なツールとして使うことができます。

例：WorkAdvance

- 背景：大卒と非大卒の労働者の格差は、時間の経過とともに拡大しています。
- しかし、誰もが伝統的な大学の環境で成功するわけではありません。
- **WorkAdvance** は、高賃金の業界（IT、医療、製造など）で通用するスキルを身につけさせるための職業訓練プログラムです。

- MDRC は、初期スクリーニングを通過した人々を対象に、訓練プログラムへの参加権をランダムに割り当てるランダム化比較試験を実施しました。

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

- 以下の OLS 回帰を推定します：

$$\underbrace{Y_i}_{\text{2-3 年後の所得}} = \alpha + \beta \underbrace{D_i}_{\text{処置ダミー}} + \varepsilon_i$$

係数	推定値	標準誤差 (SE)
$\hat{\alpha}$	14636	425
$\hat{\beta}$	1965	609

- 推定された処置効果は？ $\hat{\beta} = 1965$
- 処置効果の信頼区間は？
  $\hat{\beta} \pm 1.96 \times SE_\beta = 1965 \pm 1.96 \times 609 = [771, 3159]$
- 推定された対照群の平均は？ $\hat{\alpha} = 14636$

ロードマップ

- できること：標本平均を用いて、母平均の推定や仮説検定を行う。
- したいこと：CEF の近似を推定し、それらに関する仮説検定を行う。

「できること」を使って「したいこと」を実現するには？

- 1) CEF が特定の形式（例：線形）であると仮定する：

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

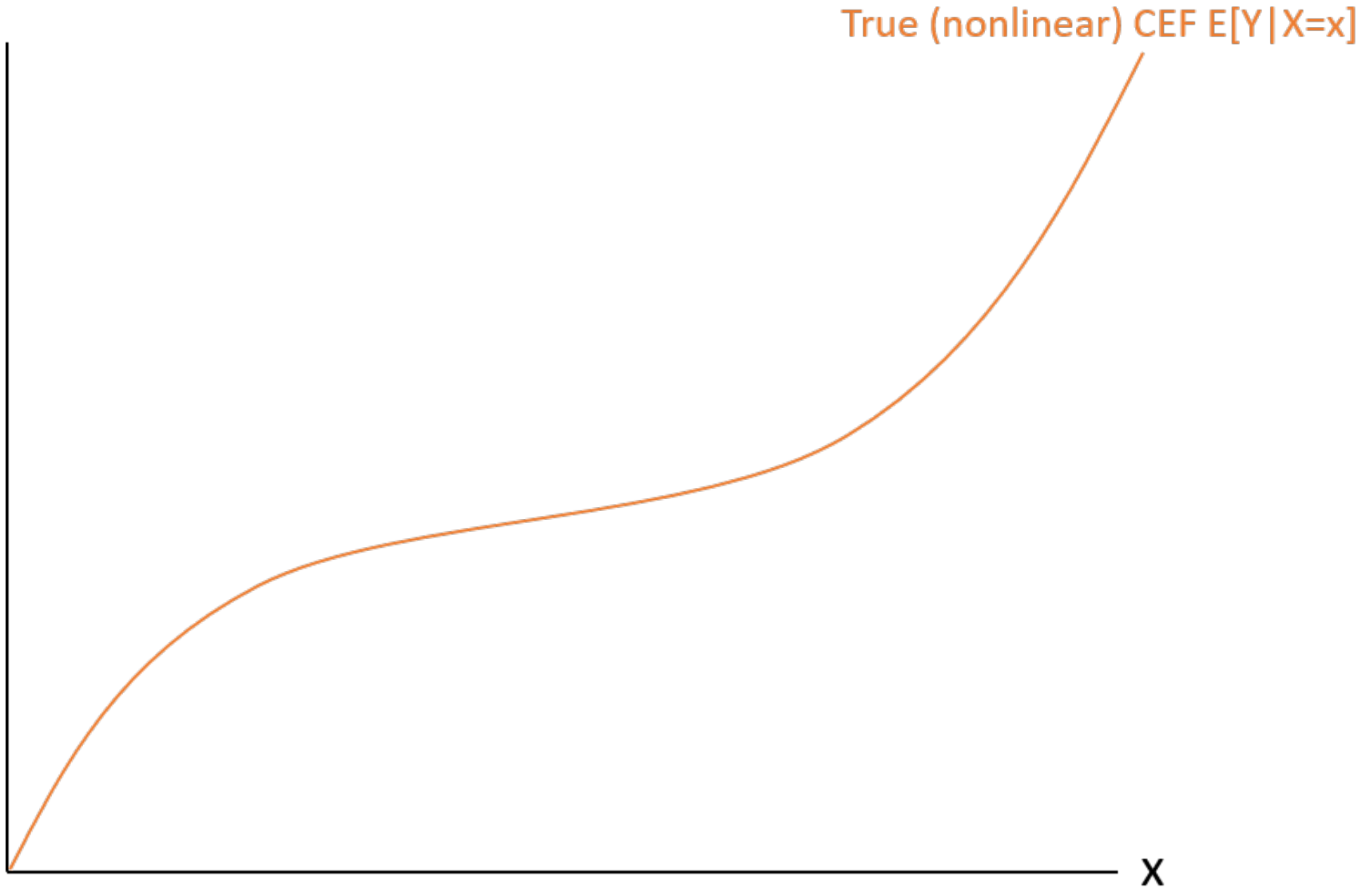
- 2) この仮定の下で、α と β が母集団の期待値の関数として表現できることを示す。✓
- 3) 母平均を標本平均で推定するツールを用いて α, β を推定し、仮説検定を行う。✓
- 4) たとえ CEF の形式に関する仮定が間違っていても、パラメータ α, β が「良い」近似を提供することを議論する。

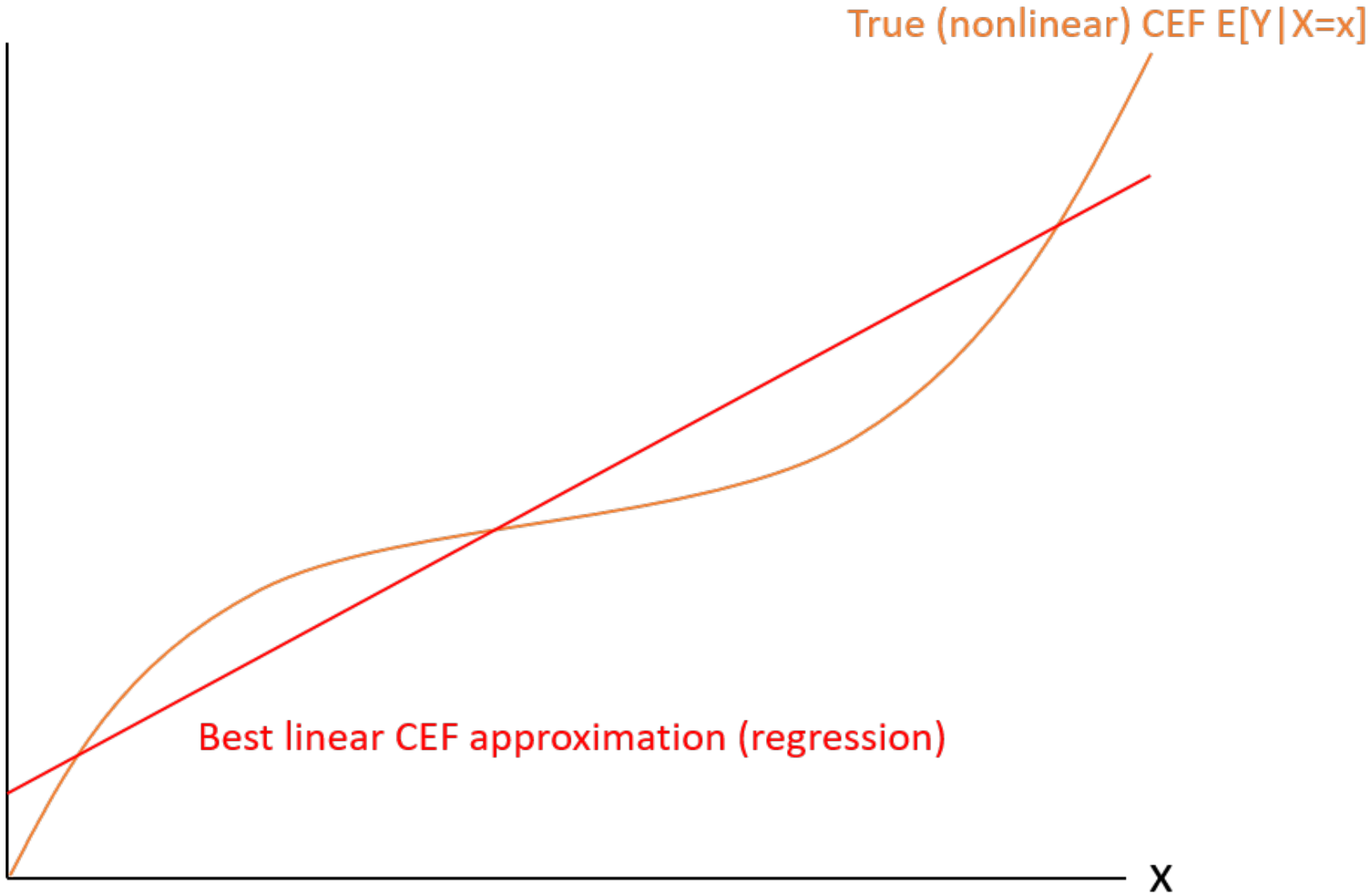
近似としての回帰分析

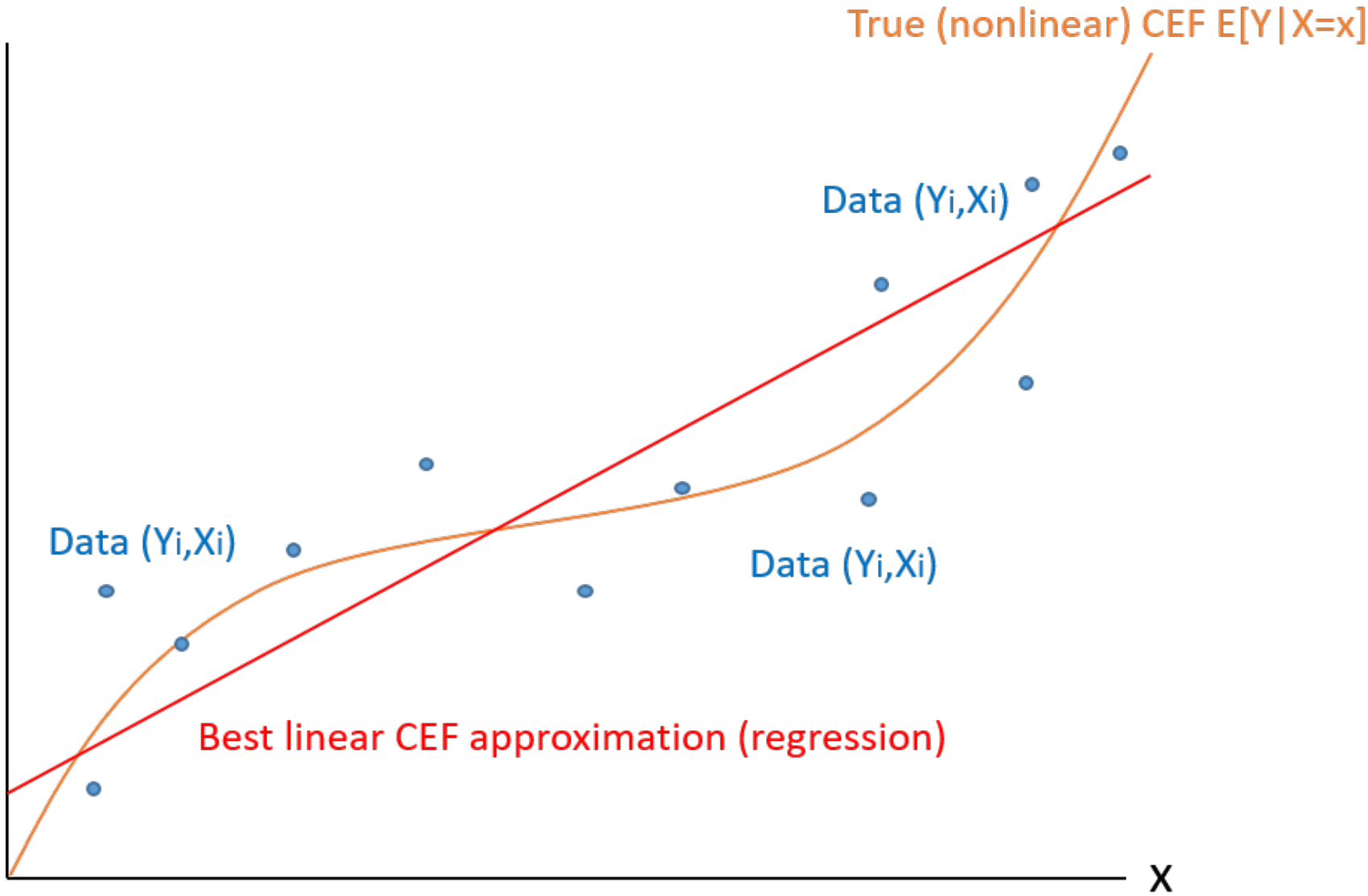
- これまでは、条件付き期待値が線形であると仮定してきました：$E[Y_i|X_i = x] = \alpha + \beta x$
- もし真の CEF が線形でなかったらどうなるでしょうか？！
- 主張：CEF が線形でない場合でも、OLS は CEF に対する「最良線形近似（best linear approximation）」を与えます。
- これが意味するのは、OLS の α, β は以下を最小化するということです：

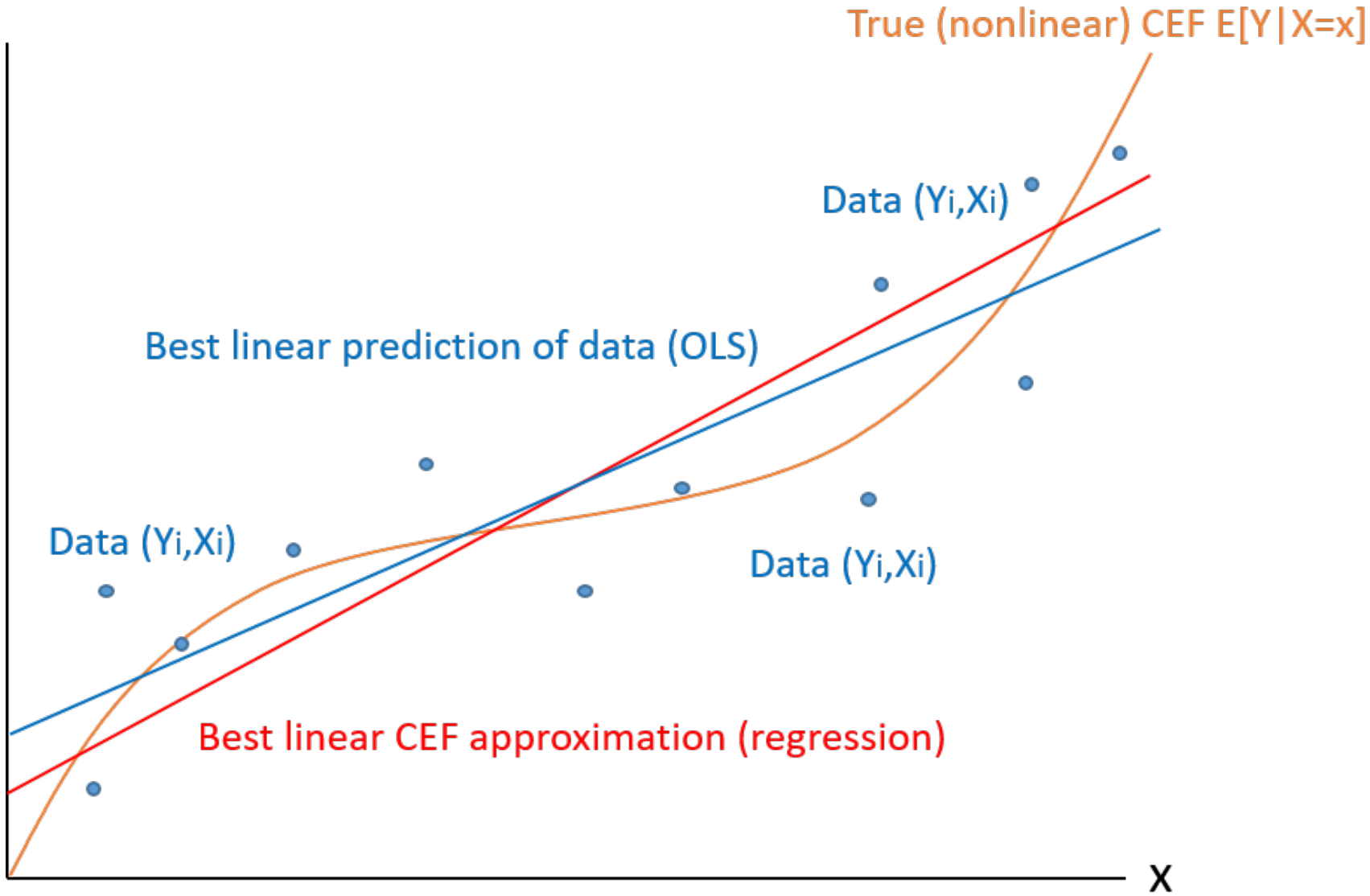
$$min_{\alpha,\beta} E[(E[Y|X] - (\alpha + \beta X))^2]$$

つまり、平均二乗誤差の意味で、CEF に最も「近い」線形関数が得られるということです。


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

最良近似としての OLS の証明

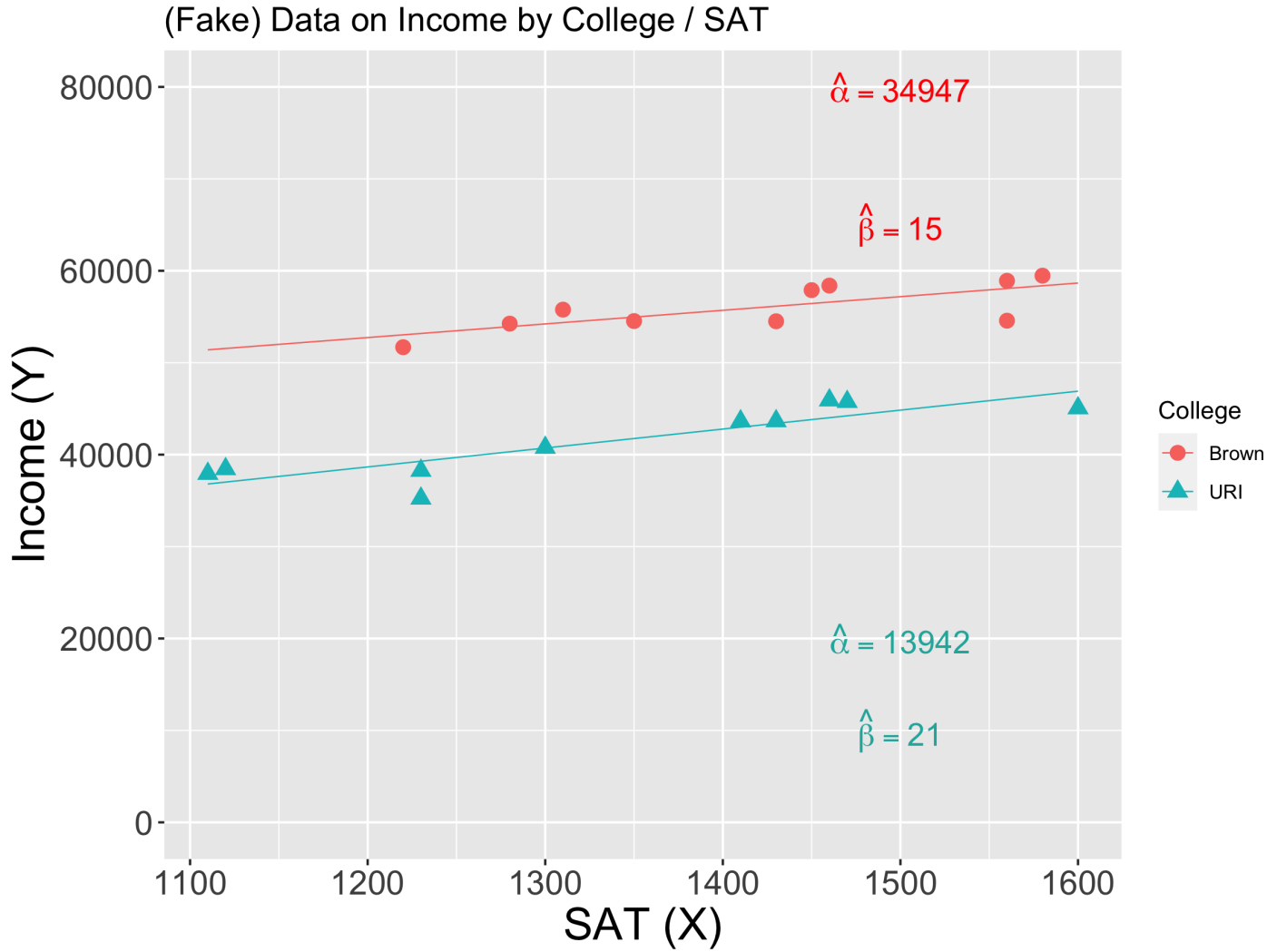
- $E[(Y-(\alpha+\beta X))^2]$ を最小化する α, β を解きました。
- $\mu(x) = E[Y|X=x]$ とします。すると：

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

- 反復期待値の法則（LIE）より：

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

- ゆえに
  $E[(Y-(\alpha+\beta X))^2] = E[(Y-\mu(X))^2] + E[(\mu(X)-(\alpha+\beta X))^2]$。
  第 1 項は β に依存しません。したがって $E[(Y-(\alpha+\beta X))^2]$ を最小化することは、$E[(\mu(X)-(\alpha+\beta X))^2]$ を最小化することと同じです。



- $E[Y_i|D_i = 1, X_i = x] - E[Y_i|D_i = 0, X_i = x] \approx \alpha_1 + \beta_1 x - (\alpha_0 + \beta_0 x)$;
  $\hat{\alpha}_1 + \hat{\beta}_1 x - (\hat{\alpha}_0 + \hat{\beta}_0 x) = (34,947 + 15x) - (13,942 + 21x) = 21,005 - 6x$
- **したがって条件付き無視可能性の下で、**
  $ATE = E[CATE(X_i)] \approx 21,005 - 6E[X_i]$