R による計量分析：データ解析と可視化

伊藤 岳

富山大学 経済学部 2017 年度後期

Email: gito@eco.u-toyama.ac.jp

January 29, 2018

Agenda

1. まとめ
2. 復習：内生性バイアス
3. 最終課題

講義の概要と目的

達成目標

この講義では，初歩的な計量経済・統計学の知識を前提に，履修者が次のような応用的な経験・知識・スキルを身につけることを目指す．

- データの可視化・解析のスキルを身に付ける —— 基礎的なデータの可視化および計量経済・統計学の推定を，R 言語を用いて行なう方法・スキルを身につける
- 基礎的な推定法，仮定，概念を理解する —— 統計的シミュレーションと経験的な (現実の) データの解析を通して，基礎的な計量経済・統計学の推定法を理解する
- 再現可能な研究を行なう力を身につける —— R 言語による実装・プログラミングを通して，再現可能な研究 (データの取得・整理・可視化・解析の一連の作業を，他者が追試・確認・再現可能な研究) を行なう方法を身につける

講義の概要と目的

達成目標

この講義では，初歩的な計量経済・統計学の知識を前提に，履修者が次のような応用的な経験・知識・スキルを身につけることを目指す．

- データの可視化・解析のスキルを身に付ける —— ヒストグラム，散布図，回帰分析の結果の可視化
- 基礎的な推定法，仮定，概念を理解する —— 中心極限定理，大数の法則，信頼区間，t 検定，回帰分析のシミュレーション
- 再現可能な研究を行なう力を身につける —— 統計的 (モンテカルロ) シミュレーション，コード記述によるデータの読み込み・操作，図表の作成，解析

- 省略した内容：データの自動取得 (演習資料アップロード済み)，GLM，発展的な可視化 (ネットワーク，地図; 一部演習資料アップロード済み)
- 発展的な内容：準実験の手法 (e.g., マッチング，不連続回帰デザイン)

評価

中間課題 2 回と最終課題 (予定)

- 中間課題 2 回 $(20 \times 2 = 40\%)$ と最終課題 (60%) による．課題の詳細は講義時に指示するが，いずれも R のコードと出力 (解析・可視化) 結果の報告・解釈を求める
- 当然ながら，正当な理由 (e.g., インフルエンザ) なく締切を過ぎた提出物は一切受け付けない

$\to$ 中間課題 (2) は最終課題にまとめたので，今日出題する内容で 80%

内生性バイアス：独立変数と誤差項の相関

OLS 推定量の一致性 (再掲)

$$\text{plim}\,\hat{\beta}_1 = \beta_1 + \frac{\text{Cov}(x_i, \epsilon_i)}{\text{Var}(x_i)} = \beta_1$$

- $\text{Cov}(x_i, \epsilon_i) = 0$ (x_i が外生変数) のとき，OLS 推定量は母回帰係数 β_1 の一致推定量になる
- $\text{Cov}(x_i, \epsilon_i) \neq 0$ (x_i が内生変数) のとき，OLS 推定量は母回帰係数 β_1 の一致推定量にならない！

$$\text{plim}\,\hat{\beta}_1 = \beta_1 + \frac{\text{Cov}(x_i, \epsilon_i)}{\text{Var}(x_i)} = \beta_1 + \text{bias}(\hat{\beta}_1) \neq \beta_1 \tag{1}$$

- 外生変数 (exogenous variable)：誤差項と相関しない独立変数
- 内生変数 (endogenous variable)：誤差項と相関する独立変数

なぜ $\mathrm{Cov}(x_i, \epsilon_i) \neq 0$ が生じるのか？

代表的な原因

- 省略変数 (omitted variable)
  - 除外変数／欠落変数でも同じ意味 (訳語の違い)
- 測定誤差 (measurement error)
- 同時性 (simultaneity)

- 選択性バイアス (selectivity bias)
  - この講義では扱わない
  - 処置 (treatment) の状態を個人／ユニットが自ら選択できることに由来するバイアス
- 以下で多用する期待値の演算・記法が分かりにくければ，副読本・鹿野本第 2 章

省略変数

省略変数バイアス

「本来回帰式に加えるべき」独立変数 z_i を，回帰式から無視してしまうことで生じるバイアス

- $\mathrm{Cov}(x_i, z_i) \neq 0$：他の独立変数と相関する
- $\gamma \neq 0$：(母) 偏回帰係数 γ が 0 ではない

教科書第 12 章の例

ゲーム時間・家庭環境・学業成績の関係

- $\mathrm{Cov}(x_i, z_i) \neq 0$：家庭環境 z_i は，ゲーム時間 x_i に影響
- $\gamma \neq 0$：家庭環境 z_i は，学業成績 y に影響

省略変数

- (1) 式 $\operatorname{plim} \hat{\beta}_1 = \beta_1 + \dfrac{\mathrm{Cov}(x_i, \epsilon_i)}{\mathrm{Var}(x_i)} = \beta_1 + \mathrm{bias}(\hat{\beta}_1) \neq \beta_1$ を思い出す
- $y_i = \beta_0 + \beta_1 x_i + \gamma z_i + \epsilon_i$ を考える
- ただし，以下を仮定
  - $\mathrm{E}[z_i] = 0$：z_i の期待値は 0
  - $\mathrm{Cov}(x_i, \epsilon_i) = \mathrm{E}[x_i \epsilon_i] = 0$, $\mathrm{Cov}(z_i, \epsilon_i) = \mathrm{E}[z_i \epsilon_i] = 0$：$\epsilon_i$ は独立変数と相関しない誤差項
  - $\mathrm{Cov}(x_i, z_i) = \mathrm{E}[x_i z_i] \neq 0$：$z_i$ は x_i と相関する
  - z_i は，観察不可能な個体属性 (e.g.,「才能」「コミュニケーション能力」)
  - $\gamma \neq 0$：z_i は y に影響する

省略変数

- z_i が観察不可能な場合，z_i を回帰式に含めることはできず，次の「誤った」回帰式を推定してしまう

$$y_i = \beta_0 + \beta_1 x_i + \eta_i \tag{2}$$

- 「真のモデル」に含まれていた γz_i は，η_i に含まれてしまう

$$\begin{aligned} \eta_i &= \gamma z_i + \epsilon_i & (3) \\ \mathrm{Cov}(x_i, \eta_i) &= \mathrm{E}[x_i \eta_i] \\ &= \mathrm{E}[x_i(\gamma z_i + \epsilon_i)] \\ &= \gamma \mathrm{Cov}(x_i, z_i) \neq 0 & (4) \end{aligned}$$

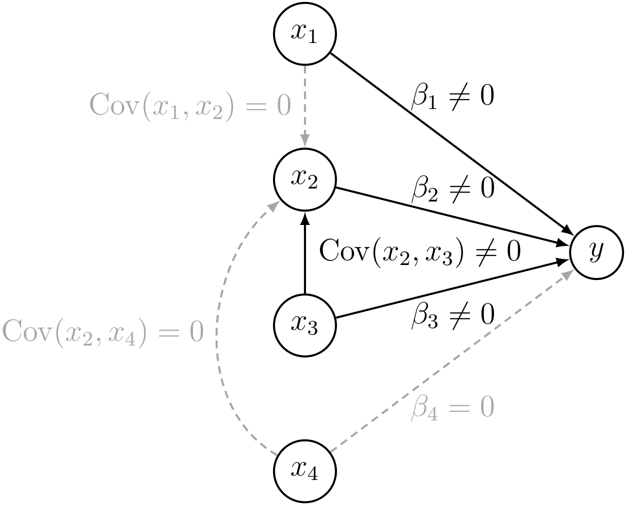
- $\mathrm{Cov}(x_i, z_i) = \mathrm{E}[x_i z_i] \neq 0$ かつ $\gamma \neq 0$ のとき，$\hat{\beta}_1$ は，

$$\mathrm{plim}\, \hat{\beta}_1 = \beta_1 + \frac{\overbrace{\mathrm{Cov}(x_i, \eta_i)}^{=\gamma \mathrm{Cov}(x_i, z_i) \neq 0}}{\mathrm{Var}(x_i)} \neq \beta_1 \tag{5}$$

となり，省略変数バイアス (omitted variable bias) が生じる

省略変数：R コード

架空データの構造を確認する：何が問題だったのか？



省略変数：シミュレーション結果とまとめ

- 「独立変数を加え損なう誤り」と「独立変数を加え過ぎる誤り」を比べれば,「独立変数を加え過ぎる誤り」の方が「まだマシ」
  - 「独立変数を加え損なう誤り」は，省略変数バイアスにつながり得る
  - 「独立変数を加え過ぎる誤り」は，省略変数バイアスにつながらない
- 一般的には，理論的・実証的根拠のある独立変数は，回帰式に加える
  - 経済学の理論や先行研究
- 観察不可能な独立変数 (e.g.,「才能」や「能力」) については，何らかの代理変数 (proxy) を検討する
  - IQ の例 (副読本・森田本第 10 章)

測定誤差

- 測定誤差：真のデータと観測されたデータとのズレ
  - 例 記憶違い,「思い出補正」, 年収のような答えにくい質問
- 独立変数に測定誤差ある場合, 内生性が生じる
- $y_i = \beta_0 + \beta_1 x_i + \epsilon_i$ を考える
- 次のように仮定する
  - x_i：独立変数の真の値
  - u_i：測定誤差
  - $w_i = x_i + u_i$：観察できる独立変数の値
  - $\mathrm{E}[u_i] = 0$：測定誤差 u_i の期待値は 0
  - $\mathrm{Cov}(x_i, \epsilon_i) = \mathrm{E}[x_i \epsilon_i] = 0$：$\epsilon_i$ は独立変数と独立
  - $\mathrm{Cov}(x_i, u_i) = \mathrm{E}[x_i u_i] = 0$：測定誤差 u_i と独立変数 x_i は独立

測定誤差

- 実際に観察できる w_i を用いる場合，回帰式は次のように書き換えられる

$$
\begin{aligned}
y_i &= \beta_0 + \beta_1 x_i + \epsilon_i \\
&= \beta_0 + \beta_1 w_i + \eta_i
\end{aligned} \tag{6}
$$

- $w_i = x_i + u_i$ に注意すると，

$$\beta_0 + \beta_1 x_i + \epsilon_i = \beta_0 + \beta_1 x_i + \beta_1 u_i + \eta_i$$

  なので，$\eta_i = \epsilon_i - \beta_1 u_i$ (また，$\mathrm{E}[u_i] = 0$ に注意)
- このとき，$\mathrm{Cov}[w_i, \eta_i] = \mathrm{E}[w_i \eta_i] = 0$ なら，内生性は生じない
- $\mathrm{E}[u_i] = 0$ なので，分散 $\mathrm{Var}(u_i) = \mathrm{E}[u_i^2]$ に注意すると，

$$
\begin{aligned}
\mathrm{E}[w_i \eta_i] &= \mathrm{E}[w_i(\epsilon_i - \beta_1 u_i)] \\
&= -\beta_1 \mathrm{E}[w_i u_i] \\
&= -\beta_1 \mathrm{E}[(x_i + u_i) u_i] \\
&= -\beta_1 \mathrm{E}[u_i^2] \\
&= -\beta_1 \mathrm{Var}(u_i)
\end{aligned} \tag{7}
$$

測定誤差

- 測定誤差によるバイアス：(7) 式において，$\mathrm{Var}(u_i) > 0$ なので，$\beta_1 \neq 0$ である限り，$\mathrm{E}[w_i \eta_i] = -\beta_1 \mathrm{Var}(u_i) \neq 0$
- つまり，w_i と η_i は相関してしまう
- 測定誤差 u_i と独立変数 x_i が独立と仮定したので (したがって $\mathrm{Var}(w_i) = \mathrm{Var}(x_i + u_i) = \mathrm{Var}(x_i) + \mathrm{Var}(u_i)$),

$$
\begin{aligned}
\operatorname{plim} \hat{\beta}_1 &= \beta_1 + \frac{\overbrace{\mathrm{Cov}(w_i, \eta_i)}^{=-\beta_1 \mathrm{Var}(u_i) \neq 0}}{\mathrm{Var}(w_i)} \\
&= \beta_1 - \beta_1 \left(\frac{\mathrm{Var}(u_i)}{\mathrm{Var}(x_i) + \mathrm{Var}(u_i)} \right) \\
&= \beta_1 \left(1 - \frac{\mathrm{Var}(u_i)}{\mathrm{Var}(x_i) + \mathrm{Var}(u_i)} \right) \\
&= \beta_1 \left(\frac{\mathrm{Var}(x_i)}{\mathrm{Var}(x_i) + \mathrm{Var}(u_i)} \right) \neq \beta_1 \qquad (8)
\end{aligned}
$$

測定誤差

- 減衰バイアス (attenuation bias)：(8) 式において ($\mathrm{Var}(u_i) > 0$ のとき)，$\hat{\beta}_1$ は絶対値で小さくなる方向にバイアスが生じる
  - $\beta_1 > 0$ なら，

$$\beta_1 > \beta_1 \left(\frac{\mathrm{Var}(x_i)}{\mathrm{Var}(x_i) + \mathrm{Var}(u_i)} \right) > 0 \tag{9}$$

  - $\beta_1 < 0$ なら，

$$0 > \beta_1 \left(\frac{\mathrm{Var}(x_i)}{\mathrm{Var}(x_i) + \mathrm{Var}(u_i)} \right) > \beta_1 \tag{10}$$

- ただし，「過小評価」のバイアスなので，回帰係数の「方向」は判別できる
- 対処法
  - 「正しい」データを集める
  - 測定誤差に配慮したモデルを構築する
  - 操作変数 (instrument variable)

同時性

- 2 つの回帰式を考える
  - 例 警察官数と犯罪件数
    - 地域 i の犯罪件数を決める式：$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$
    - 地域 i の警察官数を決める式：$x_i = \gamma_0 + \gamma_1 y_i + \eta_i$
    - 警察官数は犯罪件数を減らすが，犯罪が多い地域により多くの警察官が配備される．さらに $\cdots$
    - $x_i \to y_i \to x_i \to y_i \to \cdots$ といった相互依存関係 (フィードバック)
- 次のように仮定する
  - $\mathrm{E}[\epsilon_i] = \mathrm{E}[\eta_i] = 0$：誤差項の期待値は 0
  - $\mathrm{Cov}[\epsilon_i, \eta_i] = \mathrm{E}[\epsilon_i \eta_i] = 0$：誤差項は独立

同時性

- β_1 を正しく (一致) 推定するためには，x_i が外生変数でなくてはならない
- つまり，$\mathrm{Cov}[x_i, \epsilon_i] = \mathrm{E}[x_i\epsilon_i] = 0$ でなくてはならない
- しかし，ϵ_i が大きくなると y_i の値が大きくなり，$\gamma_1 > 0$ なら y_i の値が大きくなると x_i の値も大きくなる (連立方程式を一方について解けば，他方の誤差項を含む；副読本・鹿野本第 12 章も参照)

$$\begin{aligned}\mathrm{E}[x_i\epsilon_i] &= \mathrm{E}[(\gamma_0 + \gamma_1 y_i + \eta_i)\epsilon_i] \\ &= \gamma_0 \overbrace{\mathrm{E}[\epsilon_i]}^{=0} + \gamma_1 \mathrm{E}[y_i\epsilon_i] + \overbrace{\mathrm{E}[\epsilon_i\eta_i]}^{=0} \\ &= \gamma_1 \mathrm{E}[y_i\epsilon_i] \end{aligned} \tag{11}$$

- $\gamma_1 \neq 0$ のとき $\mathrm{E}[x_i\epsilon_i] = \gamma_1 \mathrm{E}[y_i\epsilon_i] \neq 0$ となり，内生性が生じる

$$\operatorname{plim} \hat{\beta}_1 = \beta_1 + \frac{\overbrace{\mathrm{Cov}(x_i, \epsilon_i)}^{=\gamma_1 \mathrm{E}[y_i\epsilon_i] \neq 0}}{\mathrm{Var}(x_i)} \neq \beta_1 \tag{12}$$

- 対処法
  - 何らかの実験を行なう (警察官をランダムに配置する？)
  - 操作変数を使う (e.g, 教科書 11 章の選挙と警察官数の増減の例)

バイアス：まとめ

- 独立変数と誤差項が相関するとき，回帰係数を正しく推定できない
- (原因と結果の「あて」がついていたとしても) 因果効果・因果関係を正しく推定できない
- 主な要因：省略変数，測定誤差，同時性
- では，どうするか？
  - 「取りこぼした」変数がないか，理論的・実証的知見を踏まえ検討する
  - 操作変数を探す
  - 実験データを使う／実験する
  - (自然に生じた) 実験的状況を探す／ (統計的に) 作り出す

最終課題

内容

1. 測定誤差のシミュレーション
2. 架空データの回帰分析と内生性バイアスの検討
3. 現実のデータの回帰分析

- 提出物
  1. R コード (複数の.R ファイルに分けても，1 つにまとめても可)
  2. 結果と解釈を記述した PDF ファイル (1 つにまとめること)
- 提出方法：`gito@eco.u-toyama.ac.jp` にメール添付
- 締切：2/13 (Tue) 23:59
- PDF ファイルは文章と図表で解釈を示すこと．分量の目安は A4 で 3–5 枚
  - R コードのどの部分と文章・図表が対応するかを明記すること (行数等で)
  - 上記 2 と 3 の回帰分析の結果を報告・解釈する上では，教科書第 12 章 (特に pp.243–245) を参照すること
- 日本語のルールや，メールの内容，ファイル名については常識や中間課題時に指定した内容を厳守

最終課題

1. 測定誤差のシミュレーション

- 内容：測定誤差を変化させたシミュレーション結果と，その解釈を示す
- 次の 3 つの場合のシミュレーション結果を示すこと
  - 測定誤差なし (0)，(相対的に) 小さな測定誤差，(相対的に) 大きな測定誤差 (値は任意)
- 下記のコードを使い，β_1 の推定値 $\hat{\beta}_1$ が，測定誤差を指定する `"measurement_error"` の値によってどのように変化するか，ヒストグラムを用いて解釈する
- PDF では，それぞれの場合のヒストグラム (シミュレーション結果) を示す
- その上で，測定誤差をどのような値に指定した結果，どのような $\hat{\beta}_1$ が得られたか，$\hat{\beta}_1$ と `"measurement_error"` の値の関係を明らかにする
  - `"measurement_error"` の値を変化させることは，$\beta_1 \left(\frac{\mathrm{Var}(x_i)}{\mathrm{Var}(x_i) + \mathrm{Var}(u_i)} \right)$ (8 式) の $\mathrm{Var}(u_i)$ を変化させることに等しいことに注意

測定誤差のシミュレーション：R コード

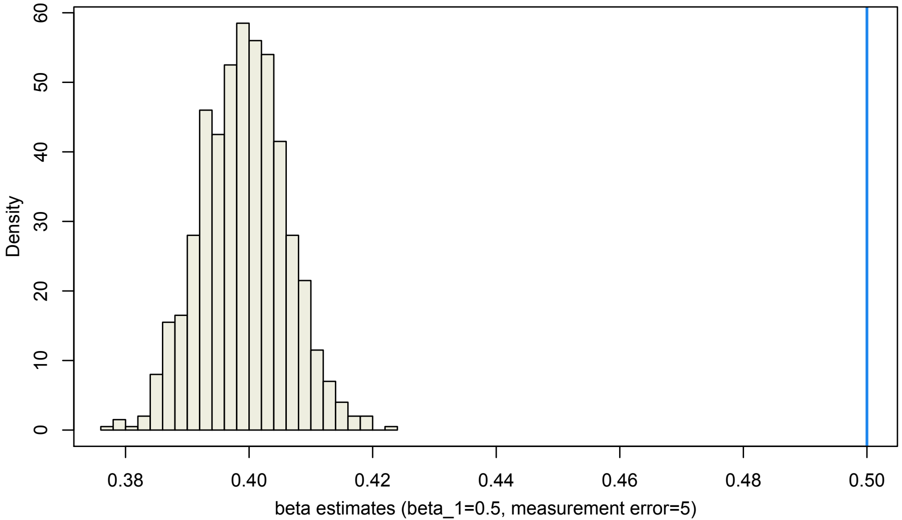
R コードのテンプレート

- R コード：「演習資料」ページの「測定誤差／減衰バイアスのシミュレーション (R コード)：simulating_me.r」
  - URL: `http://cfes-project.eco.u-toyama.ac.jp/wp-content/uploads/simulating_me.r`
- この R コードを用いて，測定誤差のシミュレーションを行なう
- コード前半部分のパッケージの呼び出しや関数定義も実行しておくこと (しなければ動作しない)

測定誤差のシミュレーション：R コード

```r
## Regression model
reg_mod = as.formula(y ~ w_1 + x_2 + x_3)
## ----------------------------------------------------
## Simulation: Measurement error
## ----------------------------------------------------
## Simulation parameters
set.seed(123456) ## random seed
n_sim = 10^3 ## N simulations
sample_size = 10^3 ## sample size
## ----------------------------------------------------
## Beta and measurement error
true_beta_1 = 0.5 ## true beta_1
measurement_error = 5 ## measurement error
## ----------------------------------------------------
## Simulation
est_mat_baseline = ols_sim(
    reg_mod, n_sim = n_sim, sample_size = sample_size, lambda = 0,
    beta_1 = true_beta_1, measurement = measurement_error
    )
## ----------------------------------------------------
## Skim results
est_mat_baseline
## ----------------------------------------------------
## Plot result
plot_mat = est_mat_baseline %>% filter(variable == "w_1")
coef_hist(plot_mat, ref_line = true_beta_1)
## ----------------------------------------------------
```

測定誤差のシミュレーション：R コード



最終課題

2. 架空データの回帰分析と内生性バイアスの検討

- 内容：架空データの回帰分析の結果から，内生性バイアスの原因を推定する
- 架空データ：演習資料ページ「サンプル・データ」内の “fake_data.csv”
- 架空データについて R で回帰分析 (OLS) した結果を示し，解釈する
- 架空データは，下記の変数を含んでいる
  - dv: 従属変数 (e.g., 子供の成績)
  - idv_1: 独立変数 1 (e.g., 世帯年収)
  - idv_2: 独立変数 2 (e.g., ゲーム時間)
  - idv_3: 独立変数 3 (e.g., 家庭環境)
  - idv_4: 独立変数 4 (e.g., 生まれた日の天候)
- ただし，この架空データについては「idv_2 と idv_3 は相関し，かつ idv_1 には何らかの理由によって測定誤差が生じ，近似的な値しか測定できないこと」が分かっている
- いくつかの回帰分析を行ない，その結果を比較することで，何が原因でどのようなバイアスが生じているかを明らかにする

最終課題

2. 架空データの回帰分析と内生性バイアス

- なぜ，どのような回帰式を推定したかを説明すること
- ある回帰分析と他の回帰分析の結果を比較することで，なぜ，どのようなバイアスが生じていると考えられるかを説明すること
- 演習資料「省略変数バイアスのシミュレーション (RMarkdown)：regression_fall2017」などを参照し，複数の回帰分析の結果を図表で示す
- PDF ファイルの文章による結果の提示と原因の考察は，数式を含めても含めなくてもよい
- R コードについては，演習資料ページの「読み込み用 R コード：read_fakedata.r」を修正・追記した.R ファイルを提出すること
  - “fake_data.csv” へのパスを指定すれば，データの読み込みはこのコードで完了
  - 変数・回帰式を指定して，回帰分析の実行・図表の出力を行なうコードを追記すればよい

架空データの回帰分析と内生性バイアス：R コード

架空データと R コードのテンプレート

- 架空データ：「演習資料」ページ内「サンプル・データ」下の "fake_data.csv"
  - URL: `http://cfes-project.eco.u-toyama.ac.jp/wp-content/uploads/fake_data.csv`
- R コード：「演習資料」ページ内「サンプル・データ」下の「読み込み用 R コード：read_fakedata.r」
  - URL: `http://cfes-project.eco.u-toyama.ac.jp/wp-content/uploads/read_fakedata.r`
- csv の読み込み用のサンプルコード
- 自分で 1 から書ければ，必ずしも使用しなくてもよい
- いずれの場合も使用したコード全体を含む.R ファイルを提出すること

架空データの回帰分析と内生性バイアス：R コード

```
library(stargazer)
library(tidyverse)

path2yourData = "" ## assign path2yourData the path to your fake_data.csv
fake_data = read_csv(path2yourData)

## Now ready to execute OLS estimation using fake_data
```

最終課題

3. 現実のデータの回帰分析

- 内容：CEPII Gravity dataset を用いて，貿易の重力モデル (gravity model) を推定し，結果を解釈する
- CEPII Gravity dataset は，下記から取得できる
  - URL: `https://sites.google.com/site/hiegravity/data-sources`
  - 表示されるウェブサイト内の，“lighter version” というリンク (zip ファイル)
- 取得・解凍した.dta ファイル (Stata ファイル) へのパスを取得し，演習資料内の R コードを用いてデータを整形する
  - データ整形用の.R ファイル：
  - .dta ファイルへのパスを各自の環境に合わせて変更しなければ動作しない
- 整形したデータセットを用いて，重力モデルを推定する
  - データ整形用の.R ファイルに回帰分析と結果の図表出力用のコードを追記し，提出すること
  - 当然ながら，上記の.dta ファイルへのパスの修正も加えておくこと

最終課題

3. 現実のデータの回帰分析

- 貿易の重力モデル：次スライド
- 推定する回帰式
  1. 最小限の重力モデル (テンプレートの R コードに記載済み)
  2. Head et al. (2010) Table 2 の “OLS” に準じるモデル
- PDF には 2 つの重力モデルの推定結果 (回帰分析) を図表で示すこと
- 従属変数・独立変数を対数変換しているか否かに注意しつつ，回帰係数の意味を明示的に解釈すること

  例 弾力性，半弾力性
  - 変数名や内容については，Head et al. (2010) を参照
- 上の 2 つのモデルの推定結果で係数に変化が生じれば，その原因を明示的に考察する

貿易の重力モデル

Tinbergen (1962)

- A workhorse, "one of the most successful empirical models in economics" (Anderson, 2011; Head & Mayer, 2015)
- Based on the analogy with Newtonian physics, the gravity model sees the bilateral trade flow as a multipliable function of size and distance effects:

$$Trade_{ijt} = \frac{GDP_{it} GDP_{jt}}{Distance_{ijt}}, \tag{13}$$

$$\ln Trade_{ijt} = \ln GDP_{it} + \ln GDP_{jt} - \ln Distance_{ijt}, \tag{14}$$

$$y_{ijt} = \alpha + \boldsymbol{x}_{it}^{\top} \boldsymbol{\beta} + \boldsymbol{x}_{jt}^{\top} \boldsymbol{\gamma} + \boldsymbol{z}_{ijt}^{\top} \boldsymbol{\phi} + \epsilon_{ijt}. \tag{15}$$

- Applications include: flows of service trade (e.g., Head et al., 2009), foreign direct investment (e.g., Head & Ries, 2008), migration (e.g., Anderson, 2011; Lewer & Van den Berg, 2008; Ravenstein, 1885; Zipf, 1946, 1949), and refugees (e.g., Echevarria & Gardeazabal, 2016), and likelihood of international armed conflict (e.g., Hegre, 2008)

現実のデータの回帰分析：R コード

R コードのテンプレート

- R コード：「演習資料」ページ内「7. 課題用テンプレート」下の「最終課題 (3) 用データ操作コード：prepare_gravity.r」
  - URL: `http://cfes-project.eco.u-toyama.ac.jp/wp-content/uploads/prepare_gravity.r`
- やや発展的なデータ操作を含むので，このコードを追記・修正した.R ファイルを提出すること

現実のデータの回帰分析：R コード

```r
## Install additional packages. Execute only once
# library(devtools)
# install_cran("haven")
# install_cran("sjPlot")

## Load packages
library(haven)
library(sjPlot)
library(stargazer)
library(stringr)
library(tidyverse)
```

現実のデータの回帰分析：R コード

```r
## ----------------------------------------------------
## Gravity dataset
## ----------------------------------------------------
## Path to dta file
## ----------------------------------------------------
## UPDATE THE FILEPATH BELOW BEFORE EXECUTING THE FOLLOWING CODE BLOCKS
path2yourData = "/Users/Gaku/Dropbox/040-ToyamaNIHU/090-ToyamaTeaching/030-
     DataVisualization/020-R_code_fall2017/090-Regression/gravity/col_regfile09.dta"
## ----------------------------------------------------
grav_data_light_raw = read_stata(path2yourData)
```

現実のデータの回帰分析：R コード

```r
## ---------------------------------------------------
## Modify data
## ---------------------------------------------------
gravity_data = grav_data_light_raw %>%
    filter(!is.na(flow), flow > 0) %>%
    mutate_at(
        vars(flow, gdp_d, gdp_o, distw, pop_d, pop_o),
        funs("ln" = log(.))
        ) %>%
    mutate(
        lnGDPpcDest = gdp_d_ln - pop_d_ln, lnGDPpcOrig = gdp_o_ln - pop_o_ln,
        BothGATT = gatt_o*gatt_d,
        ColdWar = ifelse(year <= 1991, 1, 0),
        ColonyAlways = is.na(indepdate)*col_hist,
        ACP_EU = ifelse(acp_to_eu == 1, 1, 0),
        ACP_EU = ifelse(eu_to_acp == 1, 1, 0)
        ) %>%
    select(
        Year = year, iso_origin = iso_o, iso_destination = iso_d,
        lnTradeFlow = flow_ln, lnGDPpcDest, lnGDPpcOrig,
        lnPopOrig = pop_o_ln, lnPopDest = pop_d_ln, lnDistance = distw_ln,
        RTA = rta, ColonialHistory = col_hist, ColonyAlways, ACP_EU, ColdWar,
        SharedLanguage = comlang_off, SharedBorder = contig,
        SharedCurrency = comcur, SharedLegal = comleg
        )
```

現実のデータの回帰分析：R コード

```
## Simple gravity equation
simple_mod = as.formula(lnTradeFlow ~ lnGDPpcDest + lnGDPpcOrig + lnDistance)
simple_gravity = lm(simple_mod, data = gravity_data)

## Print results
summary(simple_gravity)
stargazer(simple_gravity, type = "text")
```

- 複数の回帰分析の結果を 1 枚の表で示すには，`stargazer()` 関数を用いればよい
- 演習資料「省略変数バイアスのシミュレーション (RMarkdown)：regression_fall2017」に例がある

Anderson, James E (2011) The gravity model. *Annual Review of Economics* 31(1): 133–160.

Echevarria, Jon & Javier Gardeazabal (2016) Refugee gravitation. *Public Choice* 169(3-4): 269–292.

Head, Keith & Thierry Mayer (2015) *Gravity Equations: Workhorse,Toolkit, and Cookbook* volume 4. Oxford: Elsevier BV.

Head, Keith; Thierry Mayer & John Ries (2009) How remote is the offshoring threat? *European Economic Review* 53(4): 429–444.

Head, Keith; Thierry Mayer & John Ries (2010) The erosion of colonial trade linkages after independence. *Journal of International Economics* 81(1): 1–14.

Head, Keith & John Ries (2008) FDI as an outcome of the market for corporate control: Theory and evidence. *Journal of International Economics* 74(1): 2–20.

Hegre, Håvard (2008) Gravitating toward War: Preponderance May Pacify, but Power Kills. *Journal of Conflict Resolution* 52(4): 566–589.

Lewer, Joshua J & Hendrik Van den Berg (2008) A gravity model of immigration. *Economics Letters* 99(1): 164–167.

Ravenstein, E. G (1885) The Laws of Migration. *Journal of the Statistical Society of London* 48(2): 167–235.

Tinbergen, Jan (1962) *Shaping the World Economy: Suggestions for an International Economic Policy*. New York: Twentieth Century Fund.

Zipf, George K (1946) The $\frac{P_1 P_2}{D}$ Hypothesis: On the Intercity Movement of Persons. *American Sociological Review* 11(6): 677–686.

Zipf, George K (1949) *Human Behavior and the Principle of Least Effort*. Cambridge, MA: Addison-Wesley.