

Stataを用いた計量分析入門

麦山 亮太 Ryota MUGIYAMA

学習院大学法学部政治学科

ryota.mugiyama@gakushuin.ac.jp

2024/09/04 2024年度CSRDA計量分析セミナー・夏@Zoom

自己紹介

現所属

2021/04- 學習院大学法学部政治学科

経歴

2019/03 東京大学大学院人文社会系研究科修了、博士（社会学）

2019/04–2021/03 日本学術振興会特別研究員PD・一橋大学経済研究所

専門

社会階層・社会移動、労働市場、家族形成

*より詳しい業績などは[ウェブサイト](#)にて

目次

Stataの基礎とプロジェクト管理

データを加工する

記述統計と基礎的分析

線形回帰分析

重回帰分析を活用する

ロジスティック回帰分析

ロジスティック回帰分析の実質的意味

さらなる学習のために

計量分析を使った論文の標準的な構成

序論 Introduction

先行研究の整理・仮説の提示 Literature review; Hypotheses

方法 Methods

データと変数の説明 Data and variables

変数の記述統計 Descriptive statistics

結果 Results

2変量レベルの分析 Descriptive analysis

多変量解析 Multivariate analysis

議論・結論 Discussions; Conclusion

今日扱う内容

序論 Introduction

先行研究の整理・仮説の提示 Literature review; Hypotheses

方法 Methods

データと変数の説明 Data and variables

変数の記述統計 Descriptive statistics

結果 Results

2変量レベルの分析 Descriptive analysis

多変量解析 Multivariate analysis

議論・結論 Discussions; Conclusion

このセミナーで学ぶこと

適切なデータ分析のワークフロー

- ミスが生まれにくく、共著者や将来の自分にも優しいやり方を学ぶ

分析結果の効率的な（手作業の少ない）出力方法

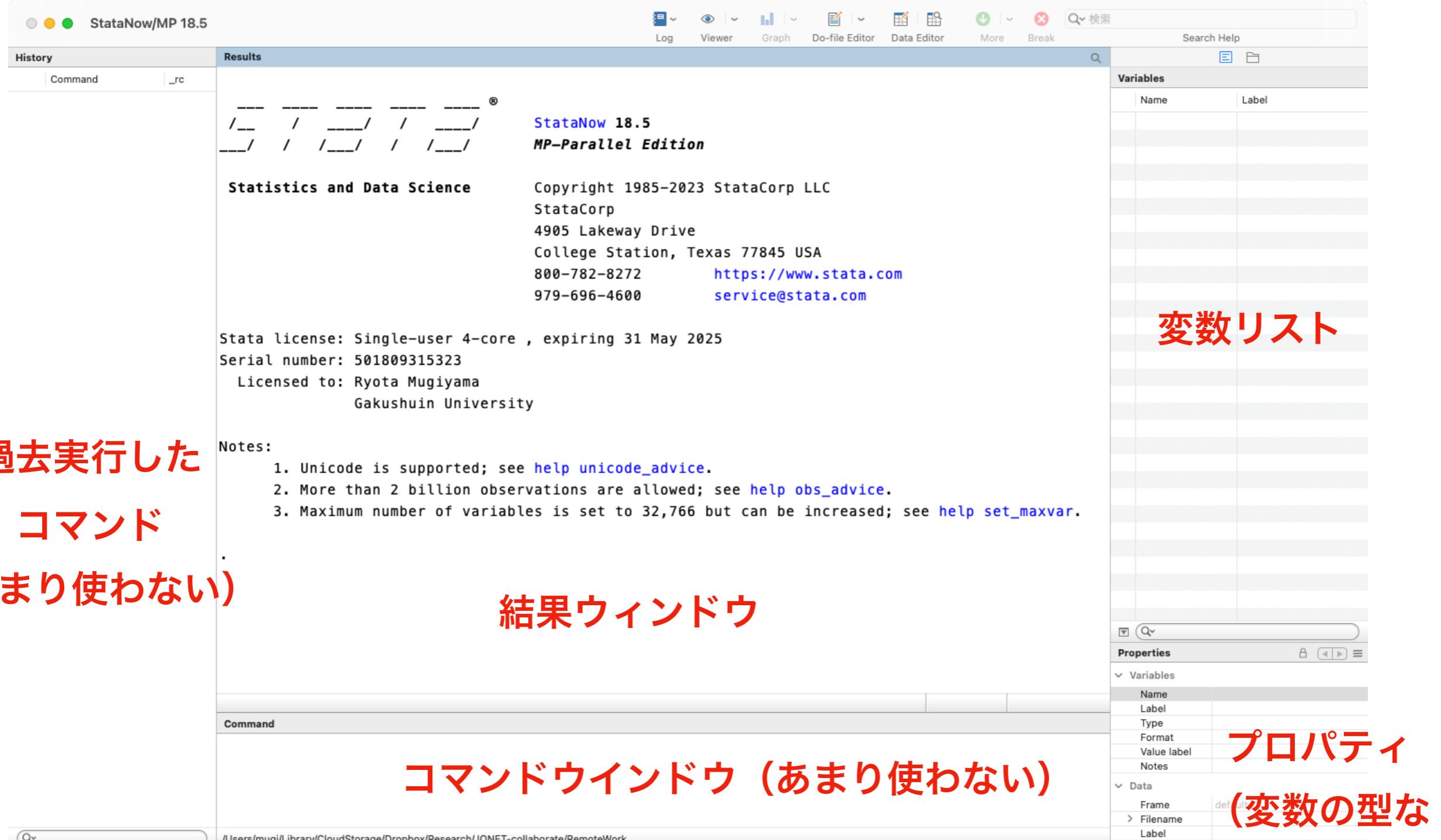
- 無意味な作業の時間を減らすことで、本質的なことを考える時間を取りれる

「意味のわかる」回帰分析をするための方法

- 適切な分析は研究を正しい方向に導く

Stataの基礎とプロジェクトの管理

Stataを開く



設定の変更

各種設定変更

EditまたはStata/MP 18.5* → Preferences →

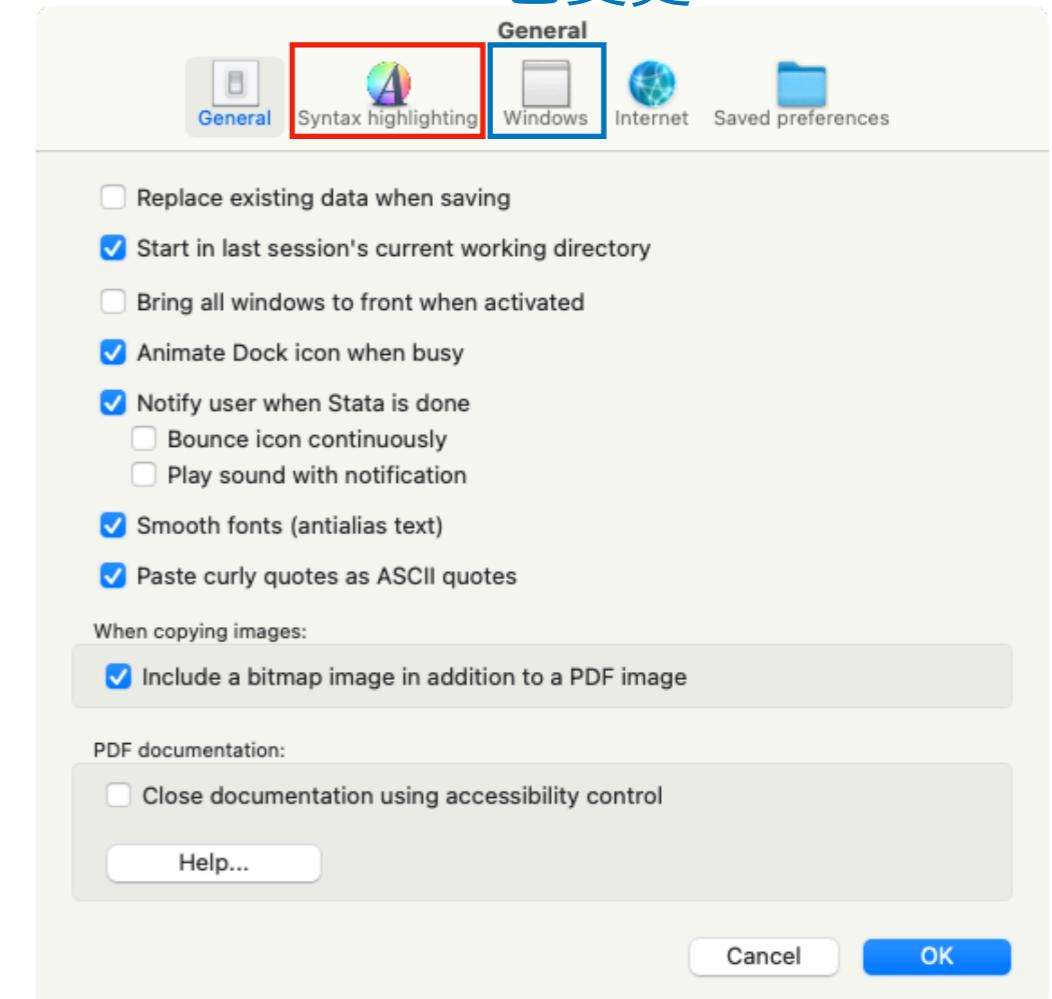
General Preferences

言語変更

EditまたはStata/MP 18.5* → Preferences →

User-interface language

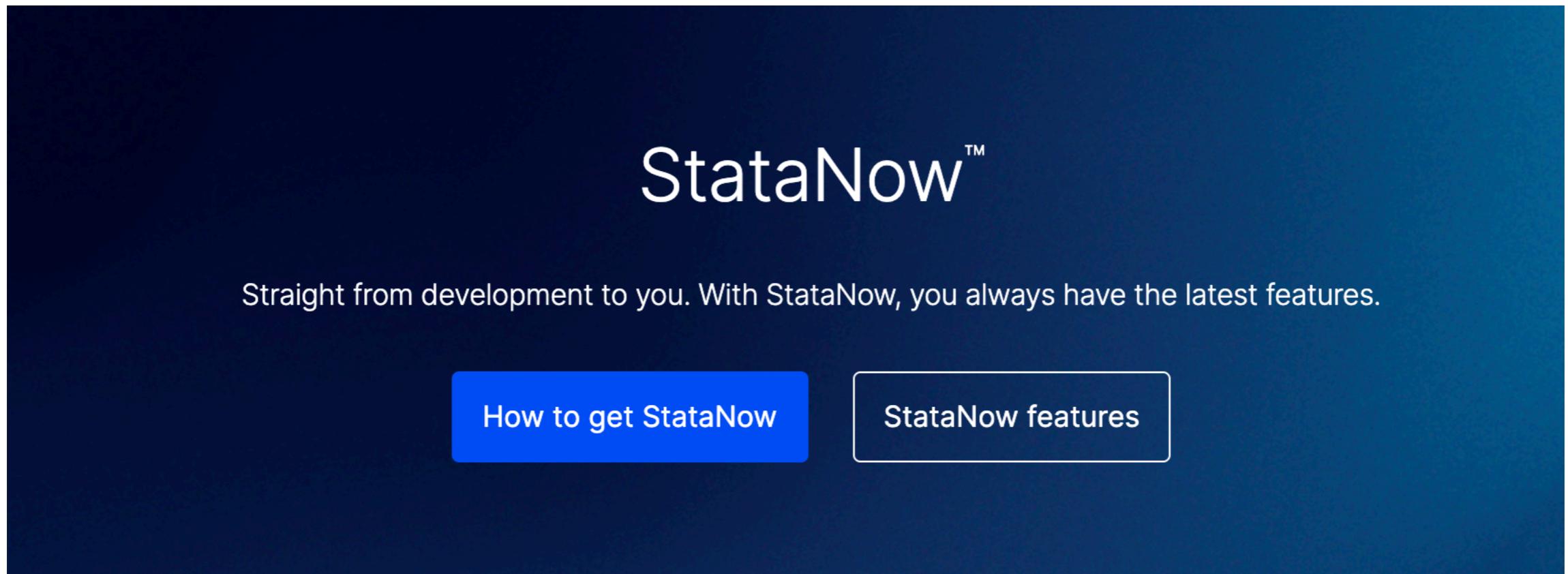
ハイライトの 結果ウインドウ等の
色変更 色変更



*バージョンによってアルファベットや数字が異なります

StataNow

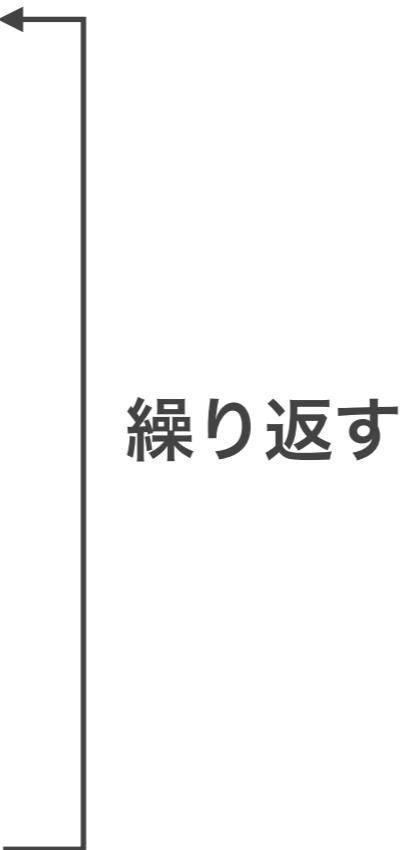
2024年5月からリリースされたStataの新しいバージョン。Stata 18以降にリリースされた新しい変更を、Stata19がリリースされるよりも先に、継続的にアップデートして導入できる



出所) StataNow <https://www.stata.com/statanow/>

計量分析のワークフロー

1. プロジェクトフォルダを作成する
2. 取得したデータをフォルダに入れる
3. データを開く
4. データを加工（変数の作成）
5. データを加工（サンプルの限定）
6. 加工したデータを保存
7. 加工したデータを分析
8. 分析結果の出力
9. 改善点やアイデアを見つける



プロジェクトフォルダの構成の例

- project : あるプロジェクトに関連するファイルをすべて入れる
- code : データの加工・分析に使用するコードを入れる
- codebook : データのコードブックを入れる
- data : 分析に使用するデータを入れる
- manuscript : 論文などの原稿を入れる
- presentation : 学会報告などで使用するスライドを入れる
- results : 分析の出力結果を入れる
- submission : 投稿ファイル、査読コメント・リプライ原稿などを入れる

各フォルダ内はさらに階層化されていてもよい

作業ディレクトリ working directoryの設定

分析のコードを走らせる場所 (=作業ディレクトリ) をPCに教える：

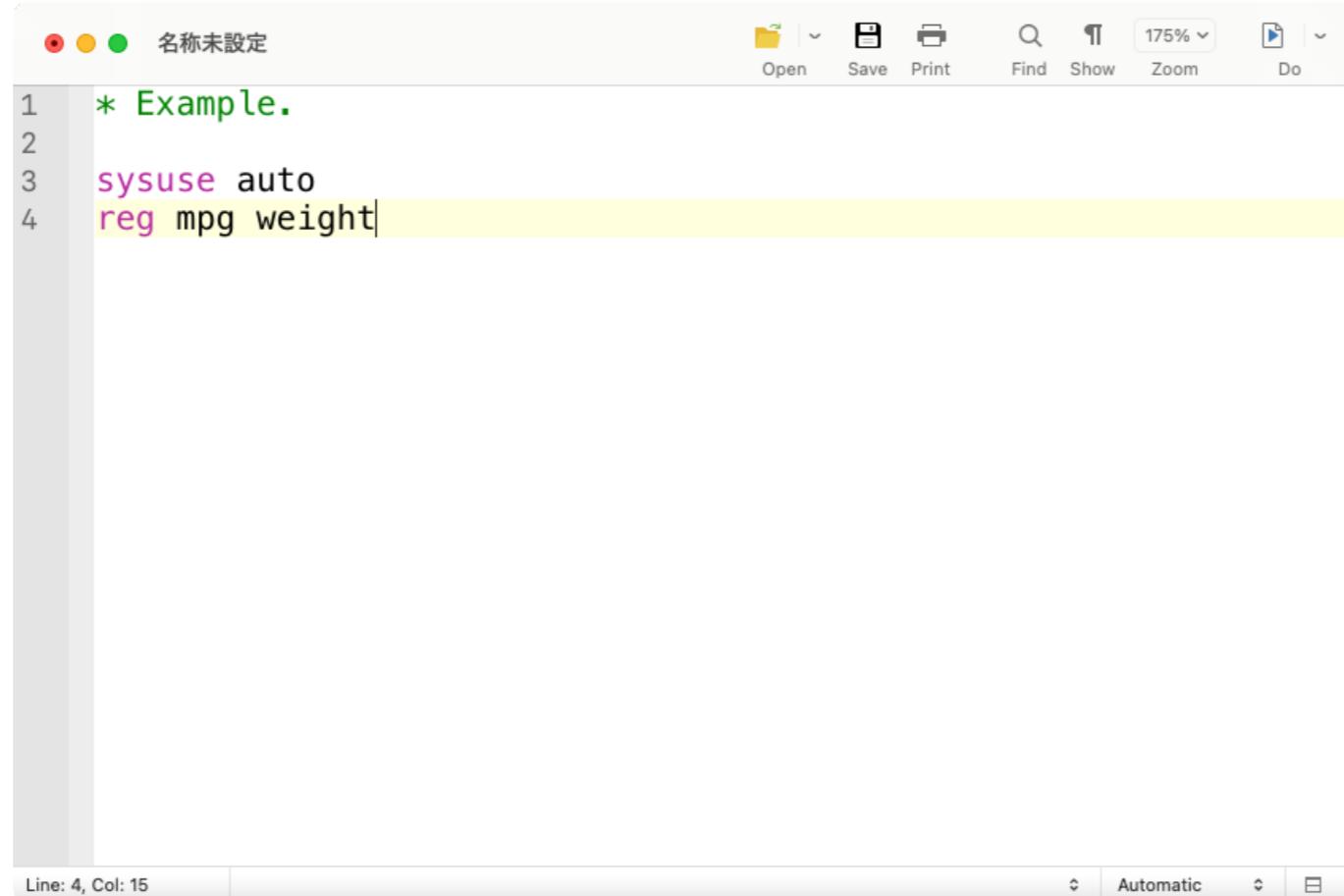
- File → Change working directory
- 作業ディレクトリに設定したいフォルダ内にある**do-fileを開く** (Macの場合はStataのバージョンによってはdirectoryが変わらないかも)

今回はダウンロードした「code」フォルダを作業ディレクトリとして指定する。
Stataの画面の下部が次のように（末尾が「/code」に）なるはず



do-fileと保存の方法

コマンドウィンドウに doedit と入力して実行 (Enter)



```
* Example.  
sysuse auto  
reg mpg weight
```

do-file上部の「Save」をクリック、またはctrl + s (Macならcommand + s) して保存。はじめて保存するときには名前をつけて保存することになる。

保存先は「code」フォルダとする

パッケージをインストールする

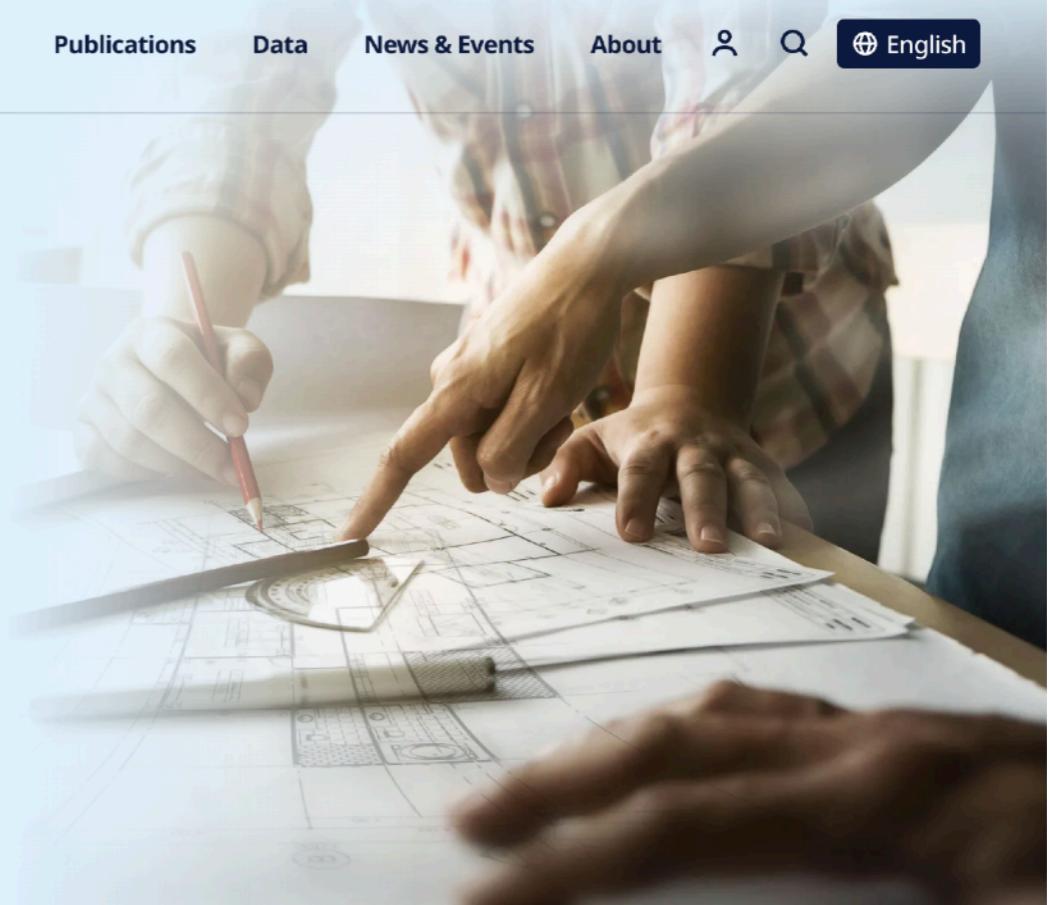
もともと組み込まれている関数のほか、他のユーザーが開発したパッケージをインストールして使うことができる。今回のセミナーで使うものは以下：

```
fre  estout  coefplot  stripplot  mdesc  
desctable  cleanplots
```

一度インストールしてしまえば、その後はほかの普通のコマンドと同じように使うことができる（毎回インストールする必要はない）

0_install2024-09-04.doのコードを実行してパッケージをインストールしよう

サンプルデータ：PIAAC



The screenshot shows the OECD website's header with navigation links: Topics, Countries & regions, Publications, Data, News & Events, About, and English. Below the header, a breadcrumb trail reads "OECD > About > PIAAC". The main title "Survey of Adult Skills (PIAAC)" is displayed prominently in large, bold, dark blue text. A descriptive paragraph below the title states: "The Survey of Adult Skills, a product of the PIAAC, measures adults' proficiency in literacy, numeracy and the ability to solve problems in technology-rich environments." A blue button labeled "Directorate for Education and Skills" is visible at the bottom left.

出所) Survey of Adult Skills (PIAAC) <https://www.oecd.org/en/about/programmes/piaac.html>

参考) 日本語での解説ページ：国立教育政策研究所 https://www.nier.go.jp/04_kenkyu_annai/div03-shogai-piaac-pamph.html

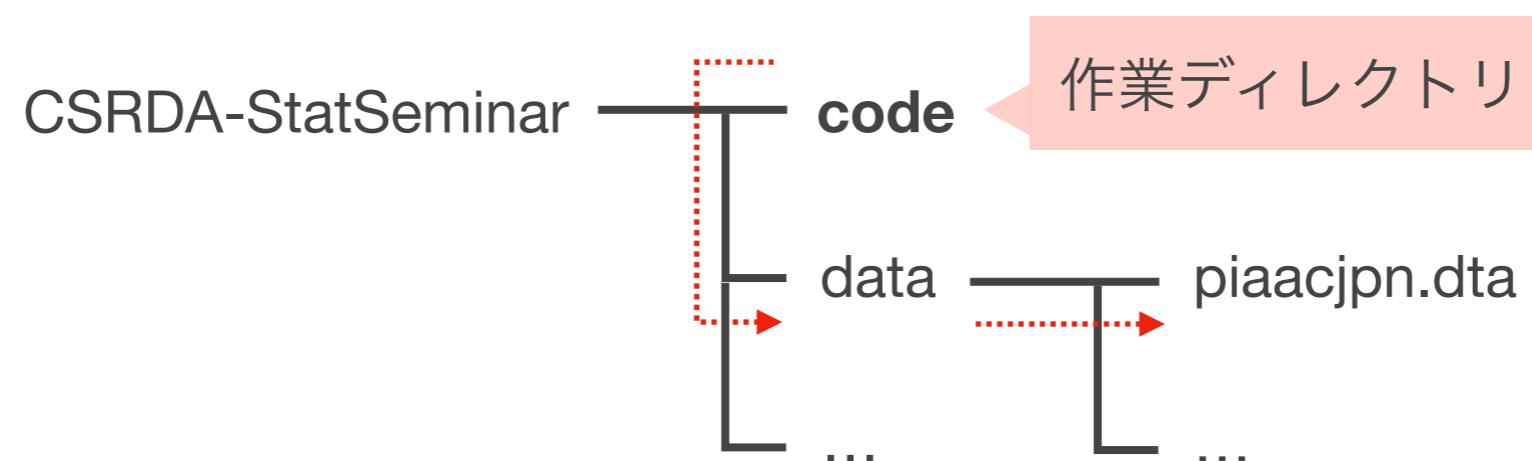
データを開き、中身を確認する

1_variables2024-09-04.doを開き、データを読み込んでみよう (1.0)

作業ディレクトリからの相対的な位置でファイルを参照することができる

../というふうにすると、作業ディレクトリから1つ上の階層に戻ることができる

"../data/piaacjpn.dta"はどこを指すのか：



do-file editorタブ上で複数のdo-fileを開く方法

Macの場合：

PCのフォルダ上にあるdo-fileをダブルクリックして開く、またはWindowsと同様の手順

Windowsの場合：

Stataのメニュー上からFile → Open → 開きたいdo-fileを選択。

PCのフォルダ上にあるdo-fileをダブルクリックすると、2つめのStataウィンドウが開いてしまい、どこで何の操作をしているのかわからなくなってしまう

絶対にやるべきでないコードの書き方の例

```
use "../data/piaacjpn.dta", clear  
  
regress earnhrbonus age i.gender  
  
recode age (25/34 = 1)(35/44 = 2)(45/54 = 3)(55/64 = 4), gen(ageg)  
  
regress earnhrbonus i.ageg i.gender  
  
drop if gender == 2  
  
regress earnhrbonus i.ageg  
  
summarize i.ageg
```

データの加工と分析は混ぜてはいけない

```
use "../data/piaacjpn.dta", clear
```

データの加工

```
regress earnhrbonus age i.gender
```

```
recode age (25/34 = 1)(35/44 = 2)(45/54 = 3)(55/64 = 4), gen(ageg)
```

```
regress earnhrbonus i.ageg i.gender
```

```
drop if gender == 2
```

```
regress earnhrbonus i.ageg
```

データの分析

```
summarize i.ageg
```

整理されていないコードはあとから見て自分が困るだけでなく、誤った結果を出すリスクを高め、結果の再現性も損なう

“dual workflow” (Long, 2009) のすすめ

最低限、データの加工とデータの分析でdo-fileを分ける

分析に関するdo-file内では、図表などを作成するための一時的なものを除いて、原則データの加工をすべきではない

0_install2024-09-04.do

1_variables2024-09-04.do

2_sample2024-09-04.do

データの加工

3_descriptive2024-09-04.do

データの分析

4_regression2024-09-04.do

5_logit2024-09-04.do

do-fileの書き方についてのtips

- 上から順番に実行すれば途中でエラーが出ることなく論文に掲載する図表がすべて出力されるのが原則（100～120行目は飛ばして……みたいなのはダメ）
- do-fileの名前は、何に関する、いつ作成・編集したものなのかがわかるようなものにするのがよい（たとえば「handling2024-09-04.do」など）。大きな変更があったときには、日付部分の名前を更新したdo-fileを作る
- do-fileの冒頭に何のdo-fileなのかわかるようにメモを残すとよい
- 類似する作業に関わるコードはまとめてフォルダに入れて管理する方法もある（ただし相対パスに配慮する必要あり）

Master do-fileから個別のdo-fileを実行する

_master2024-09-04.doを開いて中身を確認してみよう

```
/*-----  
Stataによる計量分析の実践 演習用do-file  
master2022-09-07.do  
-----  
Ryota Mugiyama (Gakushuin University)  
2022-09-07  
-----*/  
  
clear all // 何らかのファイルを開いている場合はこれらをすべて削除する  
macro drop _all // 何らかのmacro変数を使っている場合はこれらをすべて削除する  
set more off // -more-が表示されて推定結果の表示が途中で中断しないようにする。  
set scheme cleanplots // 先にインストールしたcleanplot schemeを使用するよう設定。  
  
capture log close // すでに開いているlogがある場合はこれを閉じます  
log using "log_statSeminar2022-09-07.log",replace // 新しく名前をつけたlogファイルを作成します  
  
*** 0. Install user-developed packages: 授業で使用するパッケージのインストール  
*do "0_install2022-09-07.do" // 自分の場合はインストール不要なのでコメントアウトしておきます  
  
*** Generate variables: 変数の作成  
do "1_variable2022-09-07.do"  
  
*** Select sample: 分析に用いるサンプルの抽出  
do "2_sample2022-09-07.do"  
  
*** Descriptive analysis: 記述的分析  
do "3_descriptive2022-09-07.do"  
  
*** Regression analysis: 回帰分析  
do "4_regression2022-09-07.do"  
  
*** Logit analysis: ロジスティック回帰分析  
do "5_logit2022-09-07.do"  
  
*** Advanced analysis: 時間があればやります  
do "6_advanced2022-09-07.do"  
  
log close // logを閉じます
```

Automatic ◊ Line: 16, Col: 45

データの加工

データの加工

データを手に入れたらすぐ分析.....とはならず、ほとんどの場合はもともとのデータを加工して、分析計画を実行できるようなデータを作成する必要がある
データの加工がずさんだと、とんでもない間違いが起こる

American Sociological Review

ASA American Sociological Association

12.444 Impact Factor
5-Year Impact Factor 13.153
[Journal Indexing & Metrics »](#)

Does Diversity Pay? A Replication of Herring (2009)

Dragana Stojmenovska, Thijs Bol, Thomas Leopold

First Published July 7, 2017 | Research Article |  [Check for updates](#)

<https://doi.org/10.1177/0003122417714422>

[Article information ▾](#)  58 

Abstract

In an influential article published in the *American Sociological Review* in 2009, Herring finds that diverse workforces are beneficial for business. His analysis supports seven out of eight hypotheses on the positive effects of gender and racial diversity on sales revenue, number of customers, perceived relative market share, and perceived relative profitability. This comment points out that Herring's analysis contains two errors. First, missing codes on the outcome variables are treated as substantive codes. Second, two control variables—company size and establishment size—are highly skewed, and this skew obscures their positive associations with the predictor and outcome variables. We replicate Herring's analysis correcting for both errors. The findings support only one of the original eight hypotheses, suggesting that diversity is nonconsequential, rather than beneficial, to business success.

SAGE Recommends ▾

データ加工のフロー

元のデータ : piaacjpn.dta

| | x1 | x2 | x3 |
|-----|----|----|----|
| 1 | | | |
| 2 | | | |
| 3 | | | |
| ... | | | |

変数作成後データ : piaacjpn-variable.dta

| | x4 | x5 | x6 |
|-----|----|----|----|
| 1 | | | |
| 2 | | | |
| 3 | | | |
| ... | | | |

サンプル限定後データ : piaacjpn-sample.dta

| | x4 | x5 | x6 |
|-----|----|----|----|
| 1 | | | |
| 3 | | | |
| 6 | | | |
| ... | | | |

データ加工は以下の操作からなる :

1. データ合併 : 複数のサンプルを合併する（行を加える）操作 = 今回は扱わない
2. 変数作成 : 元々のデータに変数（列）を加える操作
3. サンプル限定 : サンプル（行）を削除する操作

上記のフローはコード上で混同せず、別々に行うとよい

変数作成でよく使うコード

generate : 新たに変数を作成する

replace : 条件節で指定して、既存の変数の値を書き換える

recode : 既存の変数の値を書き換える

label variable : 変数に名前をつける

label value : 変数の値に名前をつける

label define : 変数の値につけるための名前を準備する

fre : 変数の数値とラベル、度数分布をチェックする (パッケージ)

1_variable2024-09-04.doを開き、変数を作成したり、名前をつけたりしてみよう
(1.1 – 1.4)

変数を作成したデータを保存する

分析に使う変数を作成したら、そのデータを保存する

元々のデータの容量が大きい場合には、作成した変数のみを残したデータを保存するとよい

keep：選択した変数のみを残し、他を削除する

drop：選択した変数を削除する

このフローを経ることで、自分が分析しているデータは元データを加工したデータなのだ、ということを明確に意識できる（元データにミスがあるのか、元データの加工過程にミスがあるのかを区別できる）

サンプル限定でよく使うコード

keep if : 条件に合うケースのみを残す

drop if : 条件に合うケースを除外する

2_sample2024-09-04.doを開き、サンプルを限定してみよう (2.1 – 2.2)

*Stataにおける欠損値".."は、無限大という数字で認識されている。たとえば働いているケースだけを分析したいと思って**keep if work >= 1**というコードを実行すると、働いているケースに加えて、workが欠損のケースも削除されずに残ってしまうことに注意。

サンプル限定の2つのステップ (1)

研究対象を絞るためのサンプル限定 (2.1)

- 元データから自分の研究が想定する母集団 (population) に対応するサンプルを抽出するための処理。たとえば、分析を女性に限定する、25–64歳に限定する、など
- サンプル限定に使用する変数に顕著に欠損が多い場合には問題となりうる (今回なら、年齢や働いているか否かの変数が欠損している場合)

サンプル限定の2つのステップ (2)

研究対象のうち、調査や定義の過程で欠損してしまうケースの削除 (2.2)

- あらかじめ、分析に使用する各変数でどれくらい欠損が生じているのかをチェックする
- 最近では欠損値除外前のサンプルサイズと除外後のサンプルサイズ（何の欠損でどれくらいサンプルサイズが減るか）を併記するのが規範となっている
- リストワイズ削除 (Listwise deletion; 分析に使用する変数の少なくとも1つが欠損であるようなケースを除外すること) では、欠損が完全にランダムに生じている (Missing Completely At Random; MCAR) と仮定している。リストワイズ削除をすると通常記述統計量にはバイアスが生じるが、とはいえ回帰分析では条件によっては一致推定量を得られる*

*Little, Roderick J., James R. Carpenter, and Katherine J. Lee. 2022. “A Comparison of Three Popular Methods for Handling Missing Data: Complete-Case Analysis, Inverse Probability Weighting, and Multiple Imputation.” *Sociological Methods & Research* <https://doi.org/10.1177/00491241221113873>.

Stataのプログラムで使う演算子

| | |
|--------|-----------------|
| a + b | aにbを足す |
| a - b | aからbを引く |
| a * b | aにbをかける |
| a / b | aをbで割る |
| a ^ b | aをb乗する |
| a = b | aをbに代入 |
| a == b | aとbは等しい |
| a != b | aとbは等しくない |
| a ~= b | aとbは等しくない |
| a > b | aはbより大きい |
| a < b | aはbより小さい |
| a >= b | aはbより大きいかまたは等しい |
| a <= b | aはbより小さいかまたは等しい |
| & | かつ |
| | または |

Stataのプログラムでよく使う記号

| | |
|------------|---------------------------|
| "a" | aが文字列であることを示す |
| , | , 以下はオプションであることを示す |
| /// | コードの改行 |
| . | 値に含まれている場合、欠損値 (NA) を示す |
| # | 回帰分析における交互作用項（掛け算項）の指定 |
| ## | 回帰分析における下位項目を含む交互作用項の指定 |
| /* aaaa */ | /* */で囲まれた部分はコメントアウト |
| // aaaa | // 以下、同じ行に書かれた部分はコメントアウト |
| *aaaa | * が一番はじめにある場合にはコメントアウトを意味 |

コードを書くための一般的な注意点

- 変数の名前は多少長くてもよいのでわかりやすい名前をつける：たとえば学歴ならedではなくeducation、最低でもedu, educくらいの長さがよい。変数名に全角文字は一応使えるがおすすめしない。なおStataの変数名は最大32文字なので注意
- もともとのデータに入っている質問項目などをそのまま使わない：たとえばq1_1が性別に関する質問項目で、その値をそのまま使うとしても、q1_1のままにせず、gen sex = q1_1、というふうに、必ず新しく変数を作る
- 変数には必ず変数ラベルをつける (label variable)
- カテゴリ変数の値には必ず値ラベルをつける (label define / label value)
- こまめにやっている作業のメモを残す
- 長過ぎるdo-fileは（作業単位で）分割する

Stataの欠損値の仕様

Stataでは欠損値「.」の後にアルファベットをつけて欠損値を区別することができる。たとえば、「.a」「.b」など

異なる理由で欠損になったケースを区別したい場合に使うことがある。たとえば「.a」は無回答による欠損、「.b」は「わからない」を選択したことによる欠損、など。

詳しい説明は以下：<https://www.stata.com/manuals/dmissingvalues.pdf>

データ合併でよく使うコード

append：元々のデータに新しいデータを結合して、新しい行を追加する

merge：元々のデータの変数を参照して、新しい列を追加する

append : データを下に結合する

append using "xxx.dta"

開いているデータ

| | country | x1 | x2 | x3 |
|-----|---------|----|----|----|
| 1 | Japan | 1 | 3 | 4 |
| 2 | Japan | 2 | 3 | 6 |
| ... | | | | |

using "結合したいデータ"

| | country | x1 | x2 | x3 |
|-----|---------|----|----|----|
| 1 | Korea | 2 | 5 | 3 |
| 2 | Korea | 2 | 1 | 5 |
| ... | | | | |

| | country | x1 | x2 | x3 |
|-----|---------|----|----|----|
| 1 | Japan | 1 | 3 | 4 |
| 2 | Japan | 2 | 3 | 6 |
| ... | | | | |

| | country | x1 | x2 | x3 |
|-----|---------|----|----|----|
| 1 | Korea | 2 | 5 | 3 |
| 2 | Korea | 2 | 1 | 5 |
| ... | | | | |

append : データを下に結合する

対応する変数がない場合にも結合され、その場合変数の値はmissing (.) になる

開いているデータ

| | country | x1 | x2 | x3 |
|-----|---------|----|----|----|
| 1 | Japan | 1 | 3 | 4 |
| 2 | Japan | 2 | 3 | 6 |
| ... | | | | |

using "結合したいデータ"

| | country | x1 | x2 | x4 |
|-----|---------|----|----|----|
| 1 | Korea | 2 | 5 | 10 |
| 2 | Korea | 2 | 1 | 15 |
| ... | | | | |

| | country | x1 | x2 | x3 | x4 |
|-----|---------|----|----|----|----|
| 1 | Japan | 1 | 3 | 4 | . |
| 2 | Japan | 2 | 3 | 6 | . |
| ... | | | | | |
| 1 | Korea | 2 | 5 | . | 10 |
| 2 | Korea | 2 | 1 | . | 15 |
| ... | | | | | |

記述統計と基礎的分析

1変数の集計：要約統計量

計量分析でまずははじめにやるべきは、用いるサンプルの要約統計量を集計して、データの特徴をつかむこと

- 平均はどれくらい？
- ばらつき（標準偏差）はどれくらい？
- 最大値は？最小値は？

3_descriptive2024-09-04.do を開き、summarizeコマンドを使って要約統計量を算出してみよう（3.1.1）

要約統計量を算出する

このようなときには`dtable` (Stata 18から実装) を使うと、ラベルを保持したまま、かつ論文に即座に掲載できるレベルの要約統計量の一覧を出力できる

Stata 17以前の場合は、`descstable` が代替的なコマンドとして使える

| | Summary |
|--------------------|-----------------------|
| N | 2,805 |
| Hourly wage | 1,775.167 (1,149.641) |
| Age | 43.928 (10.884) |
| Gender | |
| Men | 1,483 (52.9%) |
| Women | 1,322 (47.1%) |
| Level of education | |
| Junior high | 240 (8.6%) |
| Senior high | 994 (35.4%) |
| Junior college | 692 (24.7%) |
| University | 879 (31.3%) |
| Numeracy score | 294.185 (43.362) |

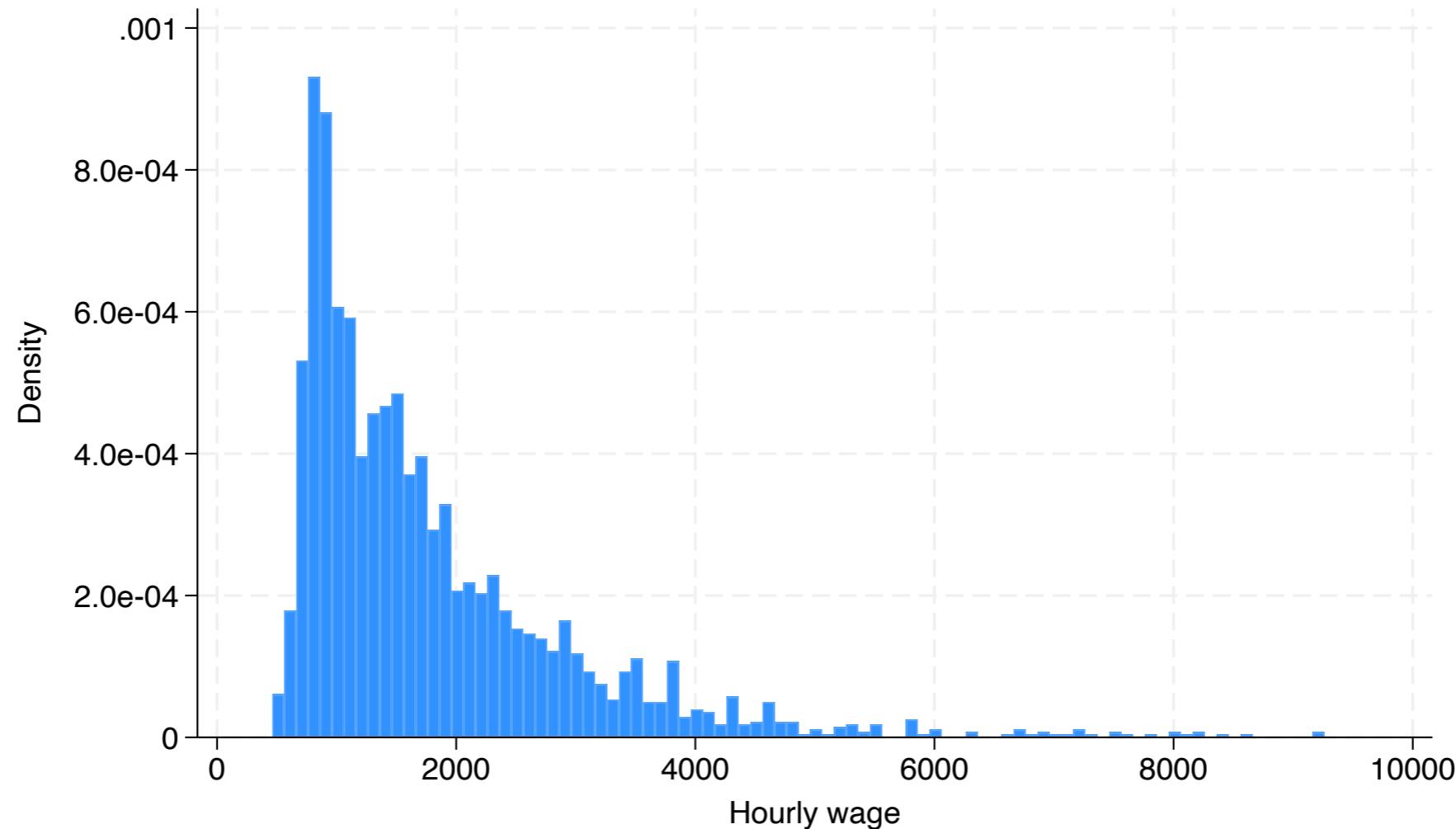
Table #: Descriptive Statistics (N = 2805)

| | Mean/Prop. | SD |
|---------------------------|------------|----------|
| Hourly wage | 1775.167 | 1149.641 |
| Age | 43.928 | 10.884 |
| Gender | .471 | |
| <i>Level of education</i> | | |
| Junior high | .086 | |
| Senior high | .354 | |
| Junior college | .247 | |
| University | .313 | |
| Numeracy score | 294.185 | 43.362 |

`dtable/descstable`コマンドを使って要約統計量を書き出してみよう (3.1.2)

一変数の分布：ヒストグラム

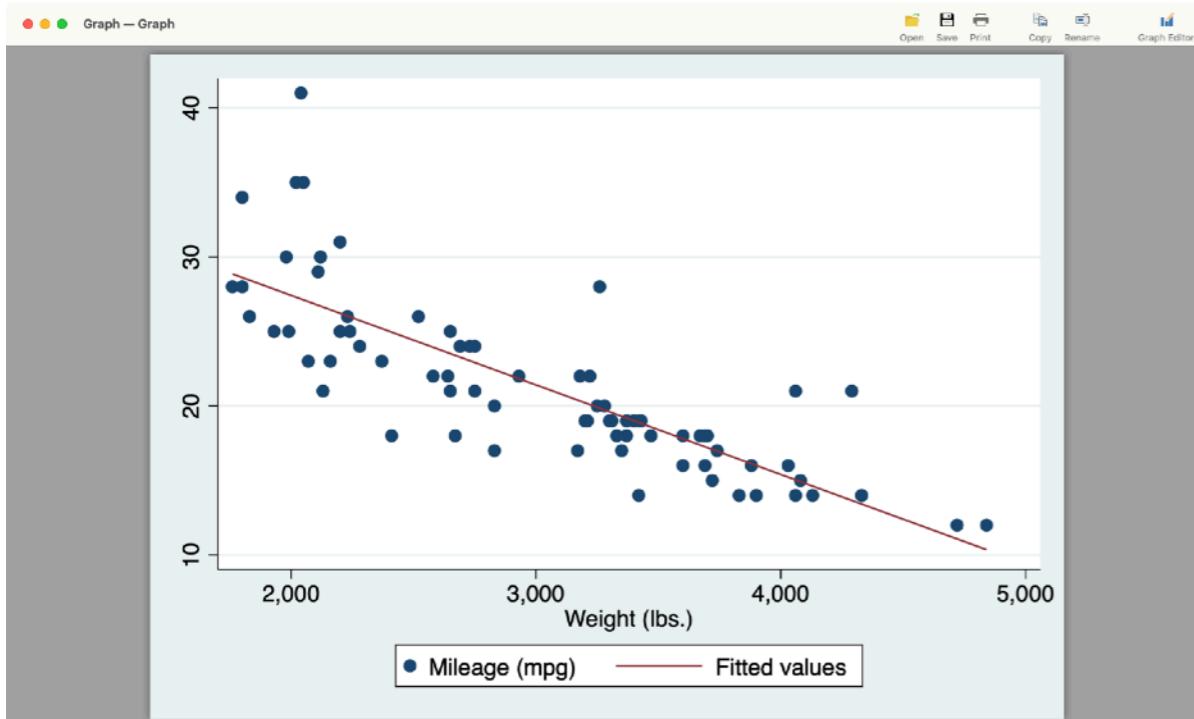
連続変数は要約統計量だけでなく分布を確認することも大事



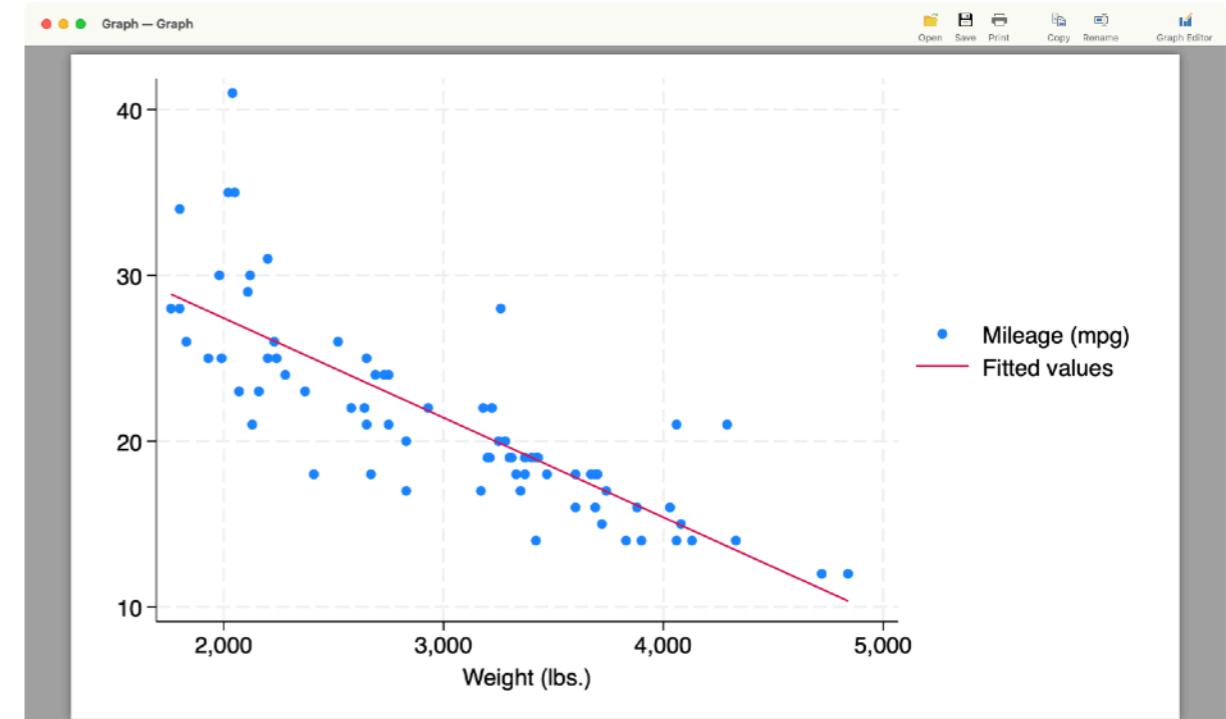
ヒストグラム、カーネル密度のグラフを作成してみよう (3.1.3)

Stata 18のグラフデザイン変更

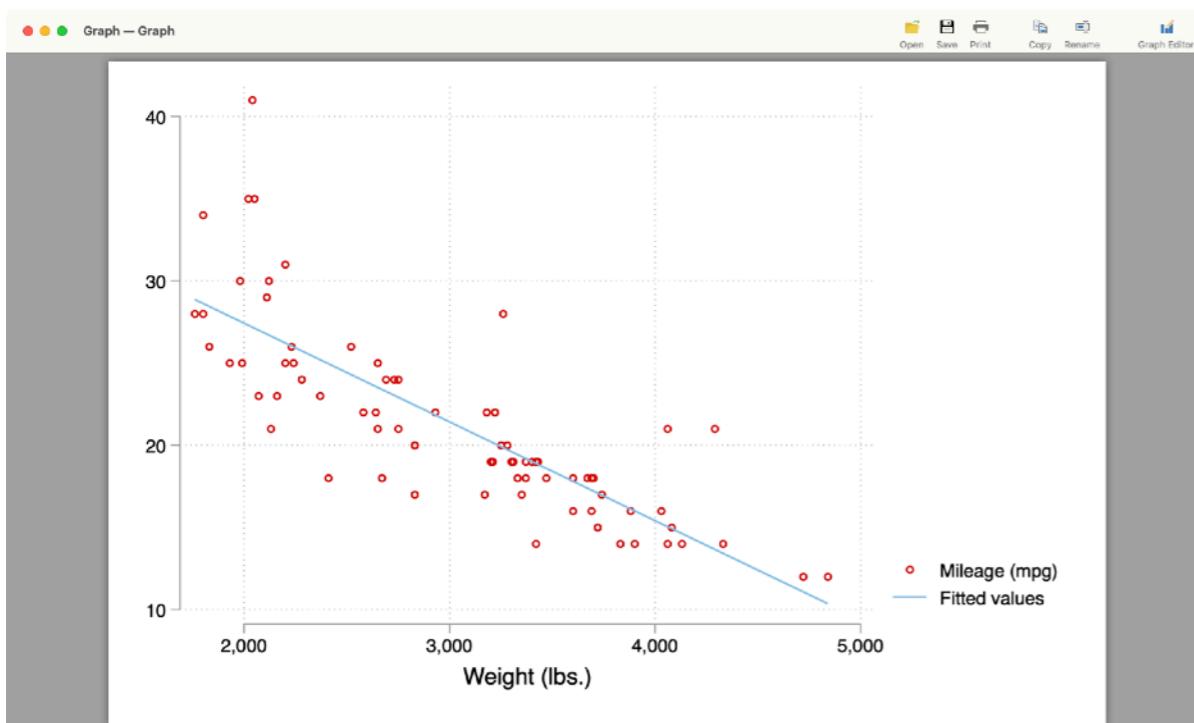
Stata 17まで



Stata 18以降



scheme(cleanplots)



連続変数をグループ間で比較する

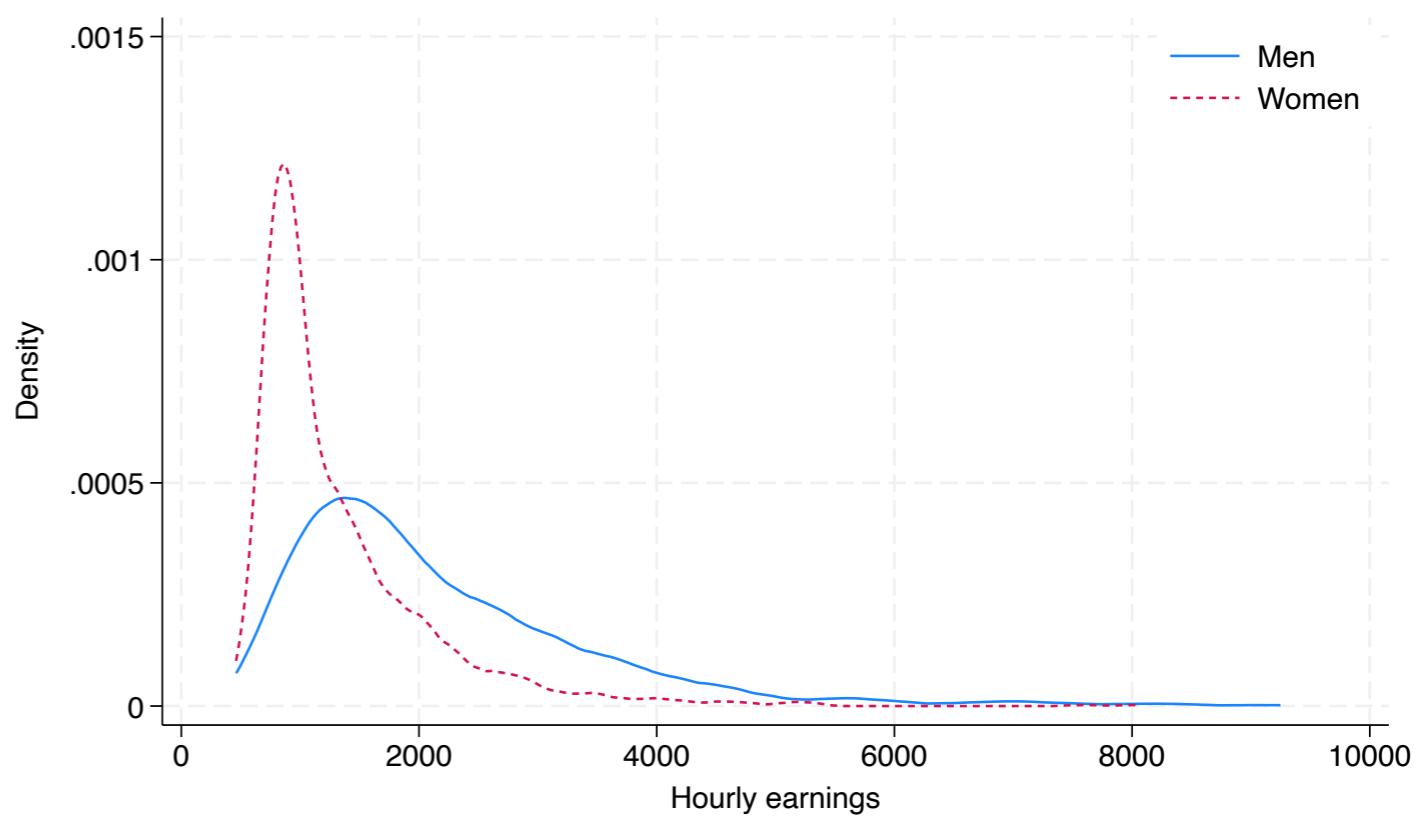
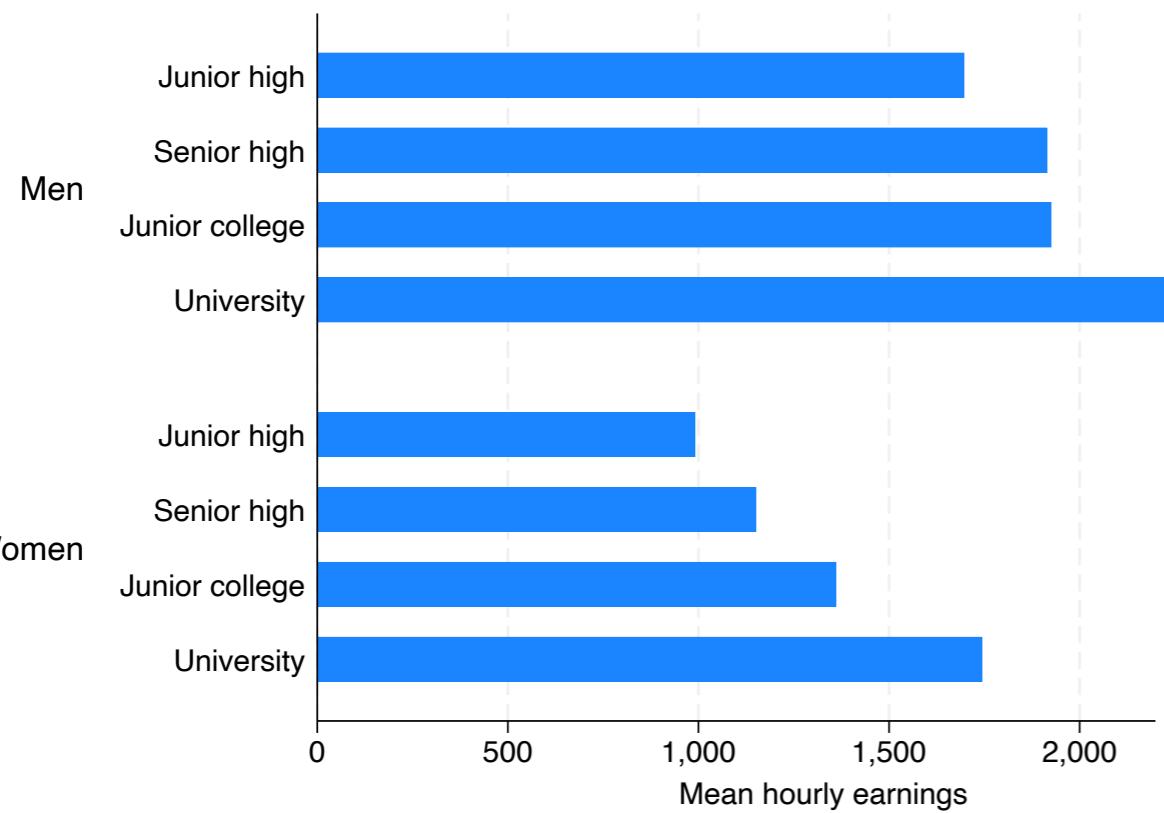
性別によって、変数の平均値や標準偏差にはどの程度違いがあるだろうか？

dtable/desctableコマンドを使ってグループ別の集計をしてみよう（3.2.1）

| | Gender | | Test |
|--------------------|-----------------------|---------------------|--------|
| | Men | Women | |
| N | 1,483 (52.9%) | 1,322 (47.1%) | |
| Hourly wage | 2,172.005 (1,277.289) | 1,330.001 (774.602) | <0.001 |
| Age | 43.704 (11.042) | 44.179 (10.702) | 0.249 |
| Numeracy score | 300.933 (44.878) | 286.616 (40.289) | <0.001 |
| Level of education | | | |
| Junior high | 139 (9.4%) | 101 (7.6%) | <0.001 |
| Senior high | 506 (34.1%) | 488 (36.9%) | |
| Junior college | 219 (14.8%) | 473 (35.8%) | |
| University | 619 (41.7%) | 260 (19.7%) | |

より効果的なプレゼンテーション

グループ別平均値の棒グラフ、複数グループ別の棒グラフ、カーネル密度グラフを作成してみよう (3.2.2)



カテゴリ変数の分布をグループ間で比較する

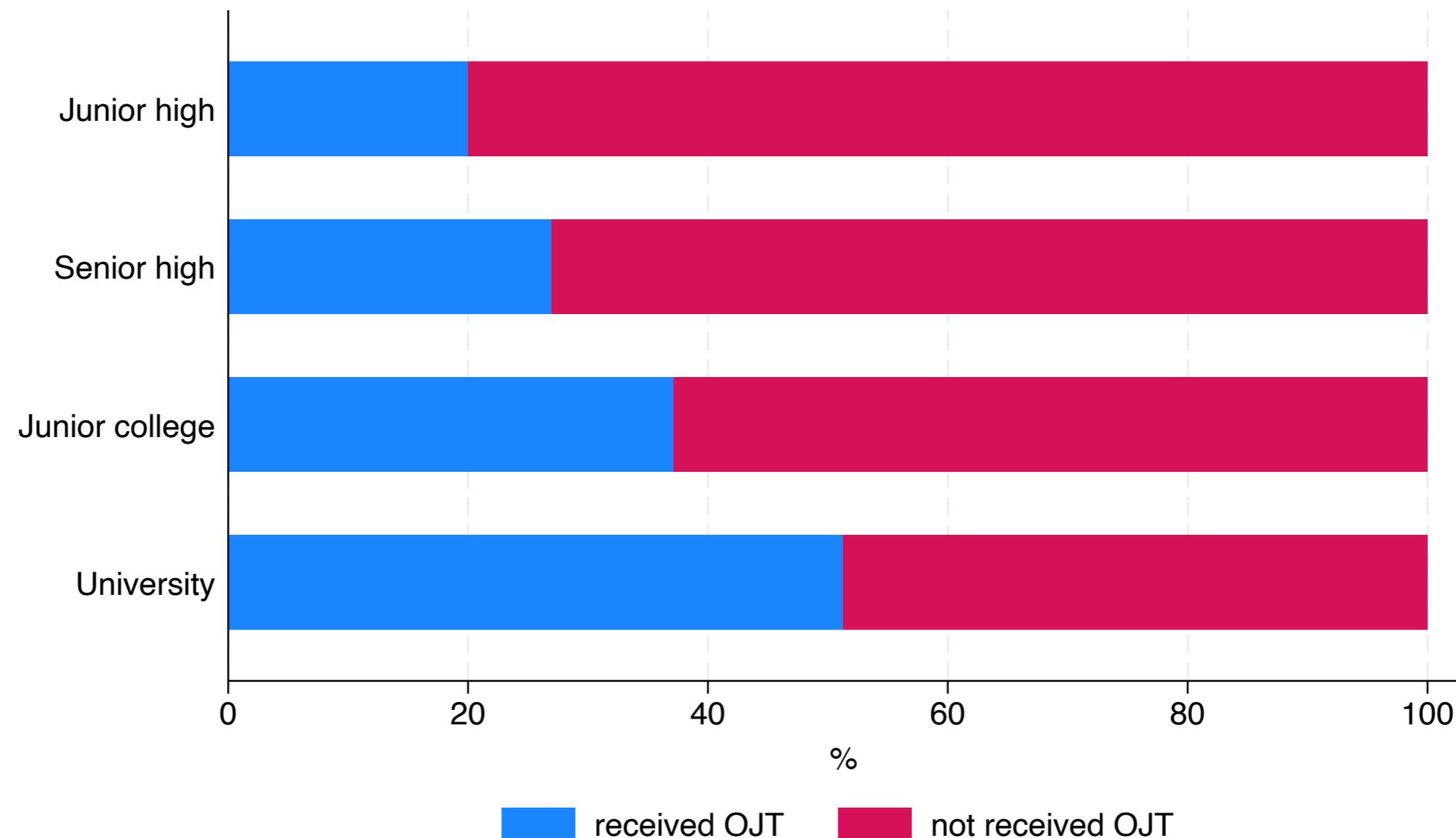
学歴によって、1年の間に職場での教育訓練（OJT）を受ける割合はどれくらい違うだろうか？

クロス集計表：カテゴリ変数（Y）の度数およびその分布をグループ（X）別に集計した表を指す

tabulateコマンドを使ってグループ別に度数とその分布（割合）を集計してみよう
(3.3.1)

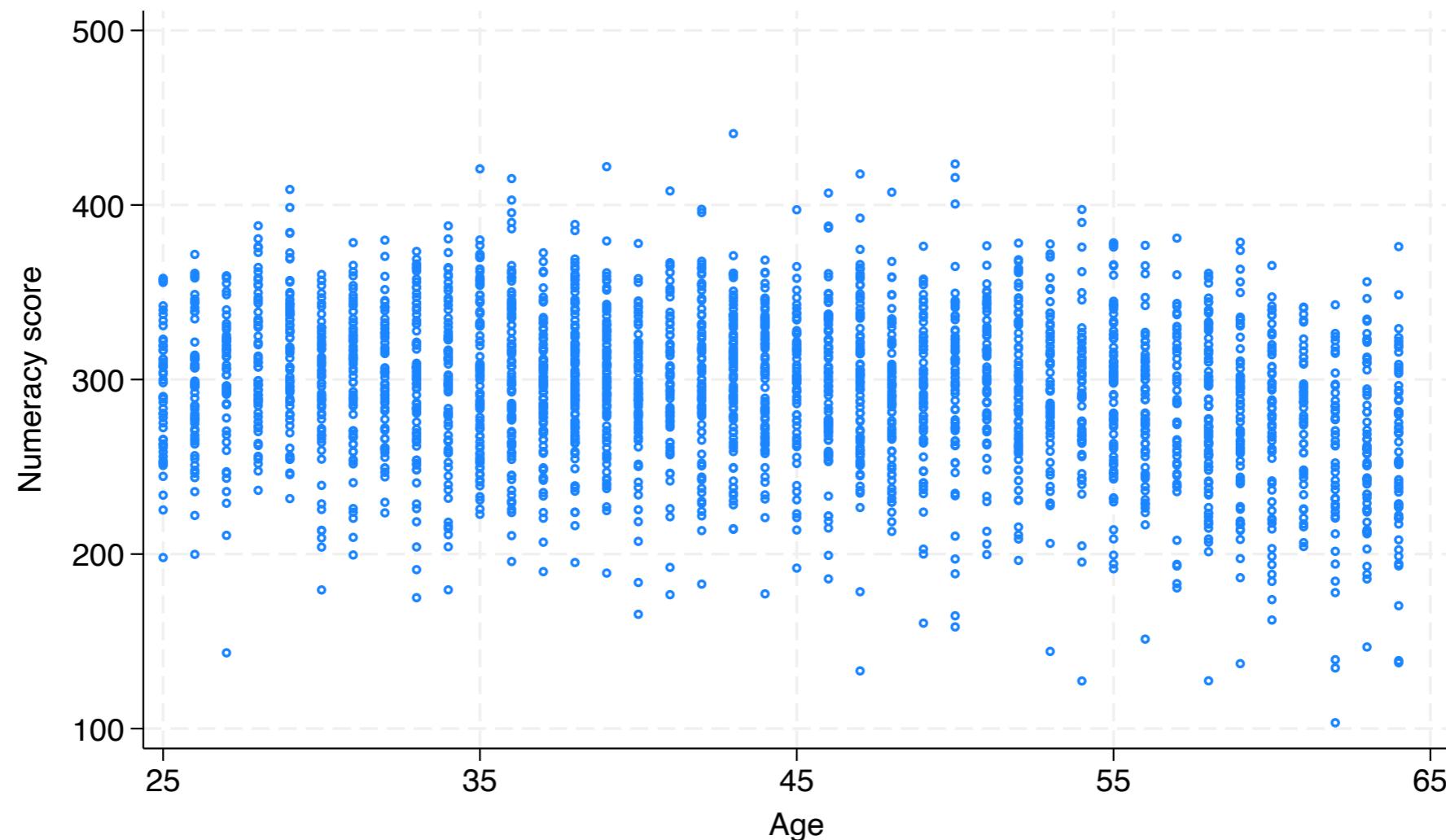
クロス集計表を図で表す

列変数が2値のときのクロス集計表を棒グラフで表してみよう (3.3.2)



散布図と相関係数

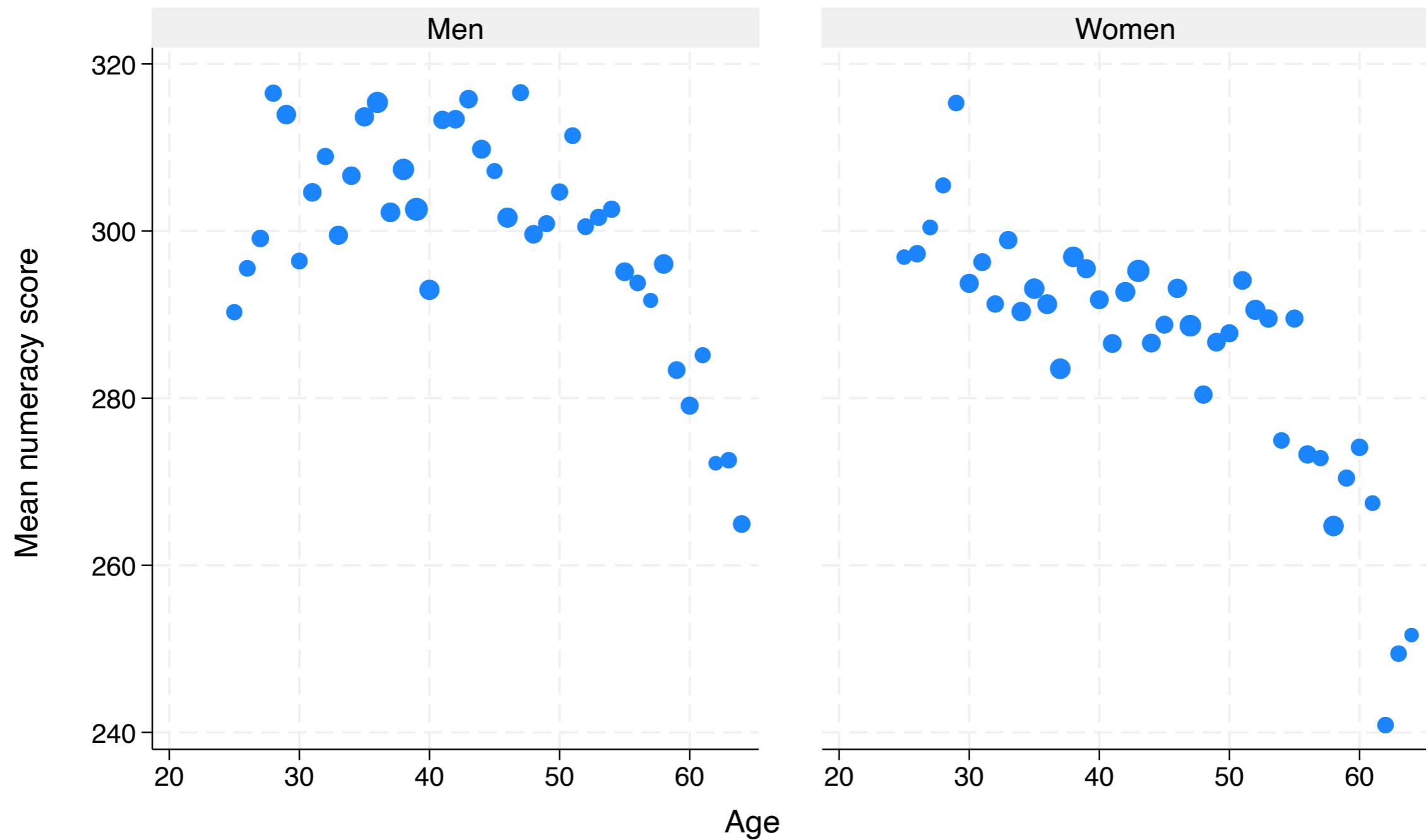
連続変数の値ごとに連続変数の値を比較する場合には、相関係数を計算したり、散布図を作成するのがよい。



相関係数を計算し、散布図を作成してみよう (3.4.1)

散布図から変数間の関係性を探索する

いろいろな散布図を作成してみよう (3.4.2)



線形回帰分析

高いスキルを持つ者は高い賃金を得られるか？

労働者のもつ技能（スキル）を資本と捉える人的資本理論によれば、技能の高い労働者はより高い収益を得ると予想できる

テストで数的思考力を測定したPIAACのデータを用いて、数的思考力と賃金の関係を検討してみよう

【参考文献】

人的資本理論について：

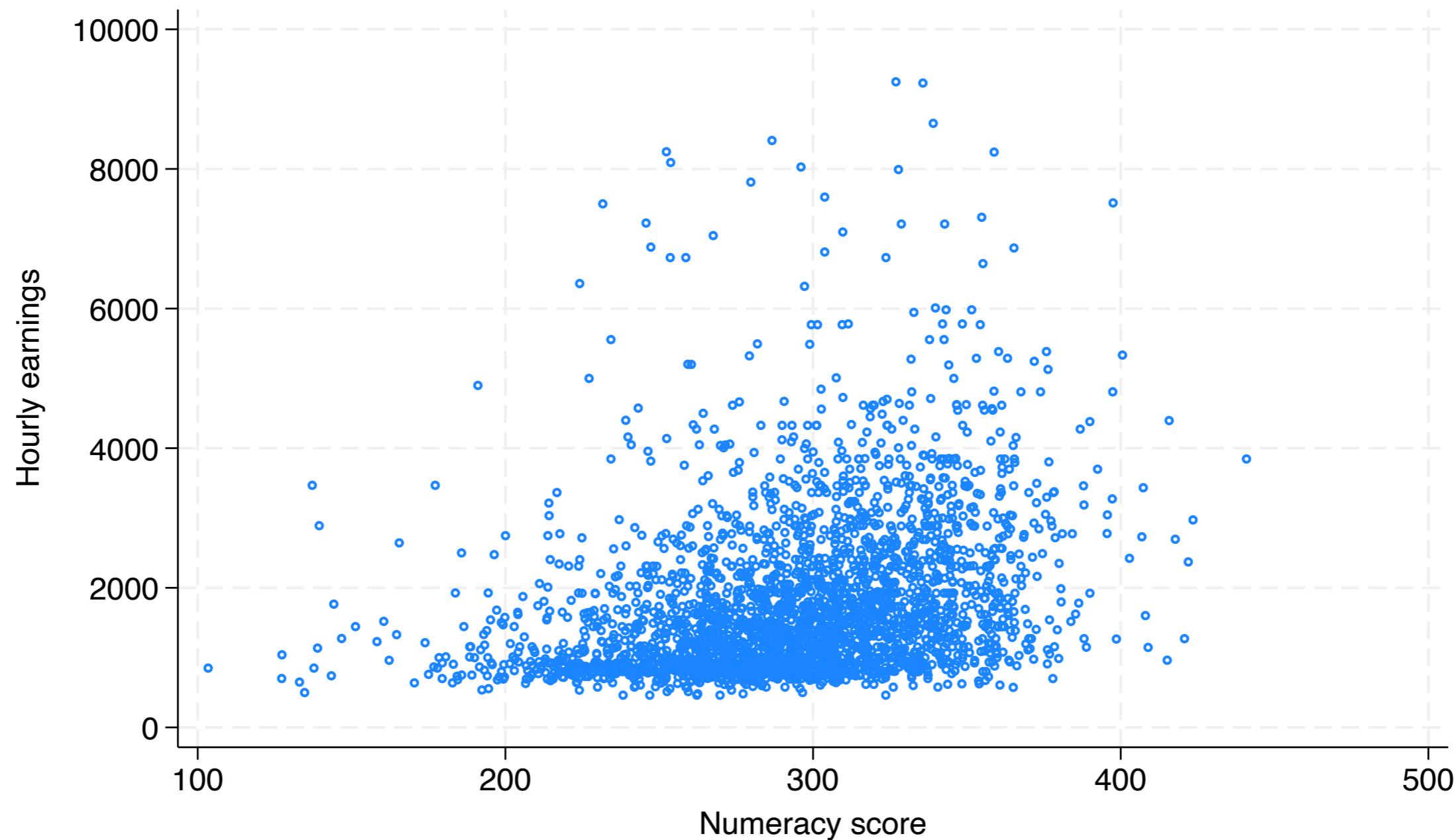
川口大司, 2017, 『労働経済学：理論と実証をつなぐ』 有斐閣.

認知的能力と賃金の関係について：

Hanushek, Eric A. and Ludger Woessmann. 2008. "The Role of Cognitive Skills in Economic Development." *Journal of Economic Literature* 46(3):607–68.

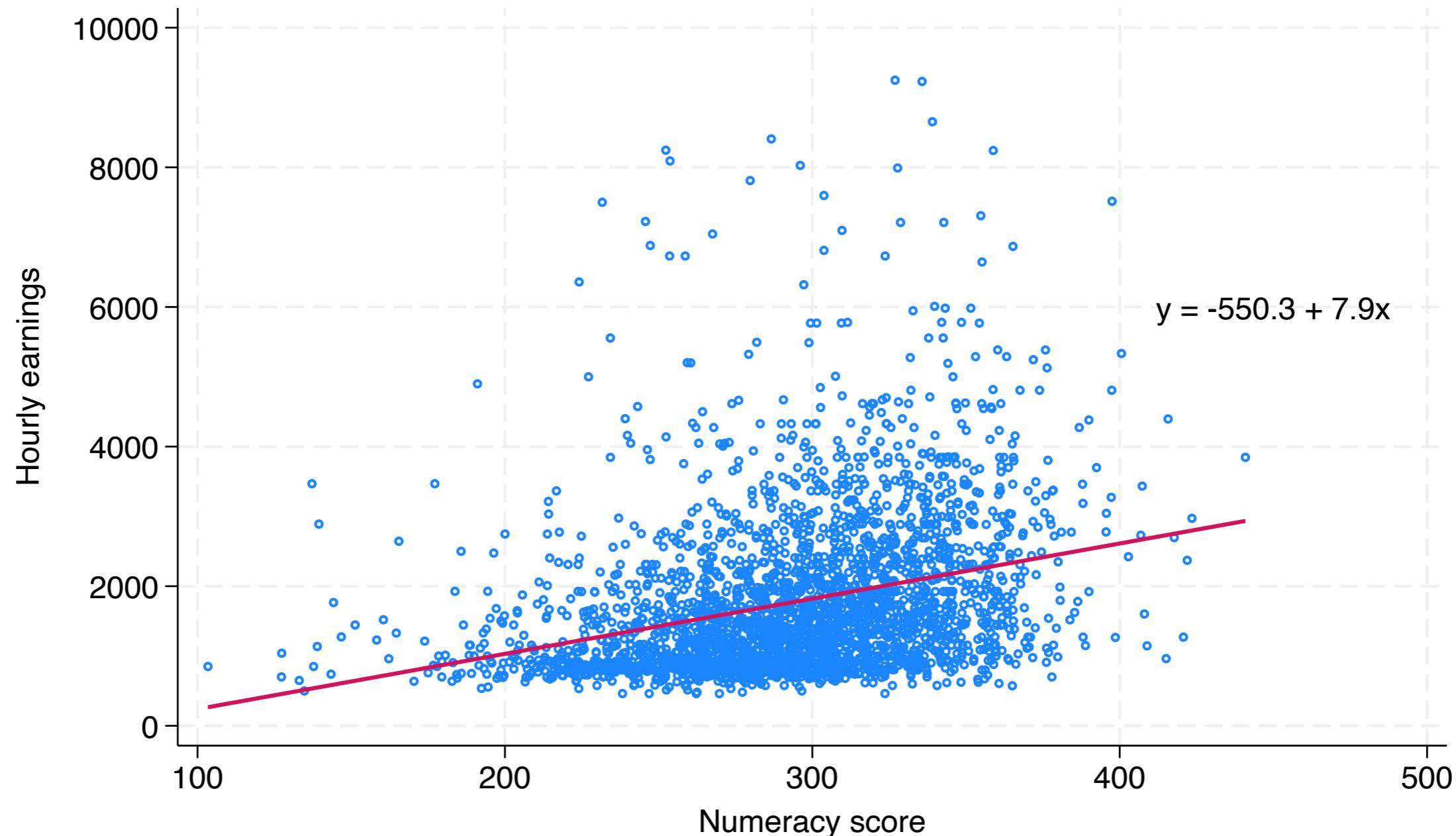
Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2015. "Returns to Skills around the World: Evidence from PIAAC." *European Economic Review* 73:103–30.

散布図を描いてみる



散布図の傾向を表す直線を引く

数的思考力 (x) が1ポイント高いと、賃金 (y) が7.9円高い



線形回帰分析 linear regression

従属変数Yと独立変数Xの間の関係を以下のような関数によって要約する方法のことをして、**線形回帰（分析／モデル）** という

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon$$

線形回帰分析の場合、各係数 $\beta_0, \beta_1, \dots, \beta_k$ は最小二乗法 Ordinary Least Squares; OLSによって推定される

回帰分析は条件付き期待値として解釈できる：

$$E(Y | X_1, \dots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

*ここで仮定： $E(\varepsilon | X_1, \dots, X_k) = E(\varepsilon), E(\varepsilon) = 0$

傾きの係数の解釈

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

X_1 が1単位増加したときの Y の增加分を ΔY とおく。

$$\begin{aligned} Y + \Delta Y &= \beta_0 + \beta_1(X_1 + 1) + \cdots + \beta_k X_k + \varepsilon \\ &= (\beta_0 + \beta_1 X + \cdots + \beta_k X_k + \varepsilon) + \beta_1 \\ &= Y + \beta_1 \end{aligned}$$

$$\Delta Y = \beta_1$$

傾きの係数は、 X が1単位増加したときの Y の增加分 ($\partial Y / \partial X_1$) を表す

X 1単位の変化に対する Y の変化量を**限界効果 partial effect/marginal effect**という

Stataでの回帰分析の出力結果

4_regression2024-09-04.doを開き、散布図を作成、および単回帰分析を推定してみよう（4.1.1）

. regress wage numeracy

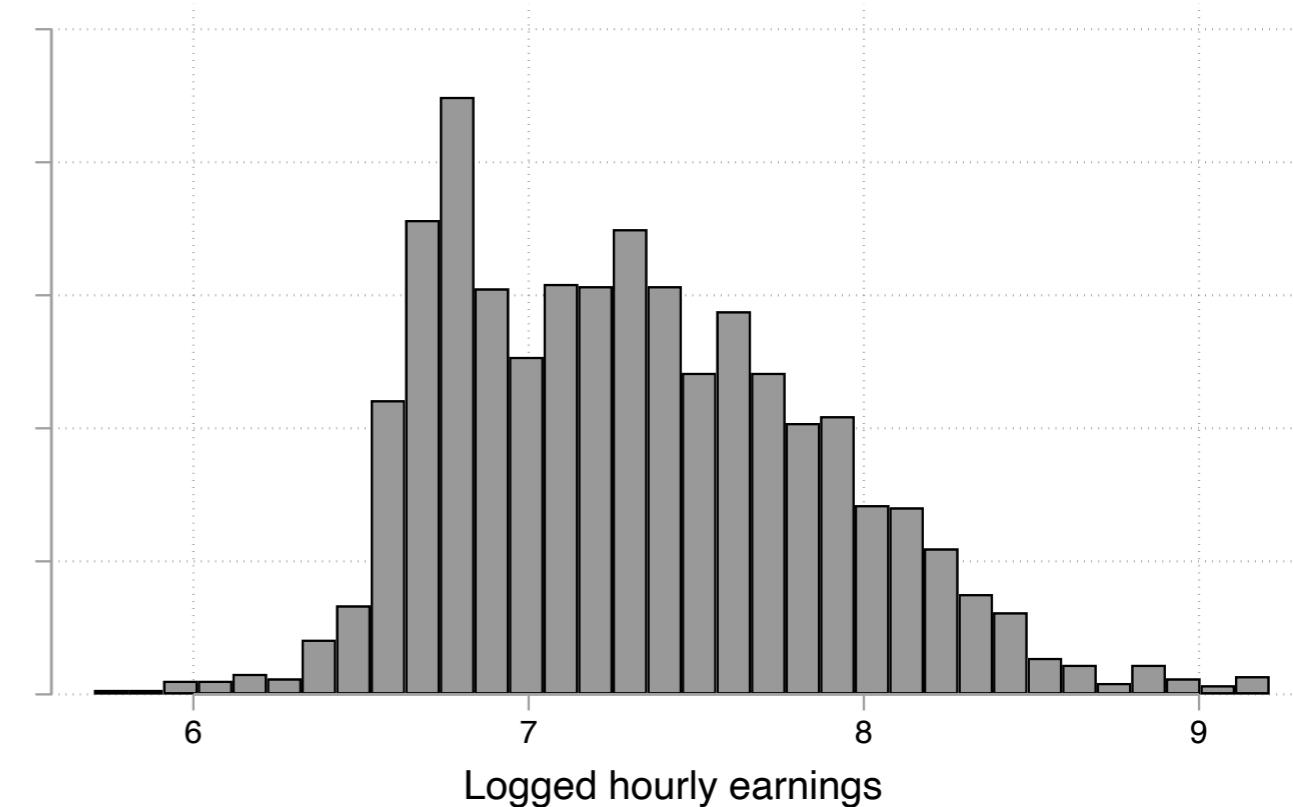
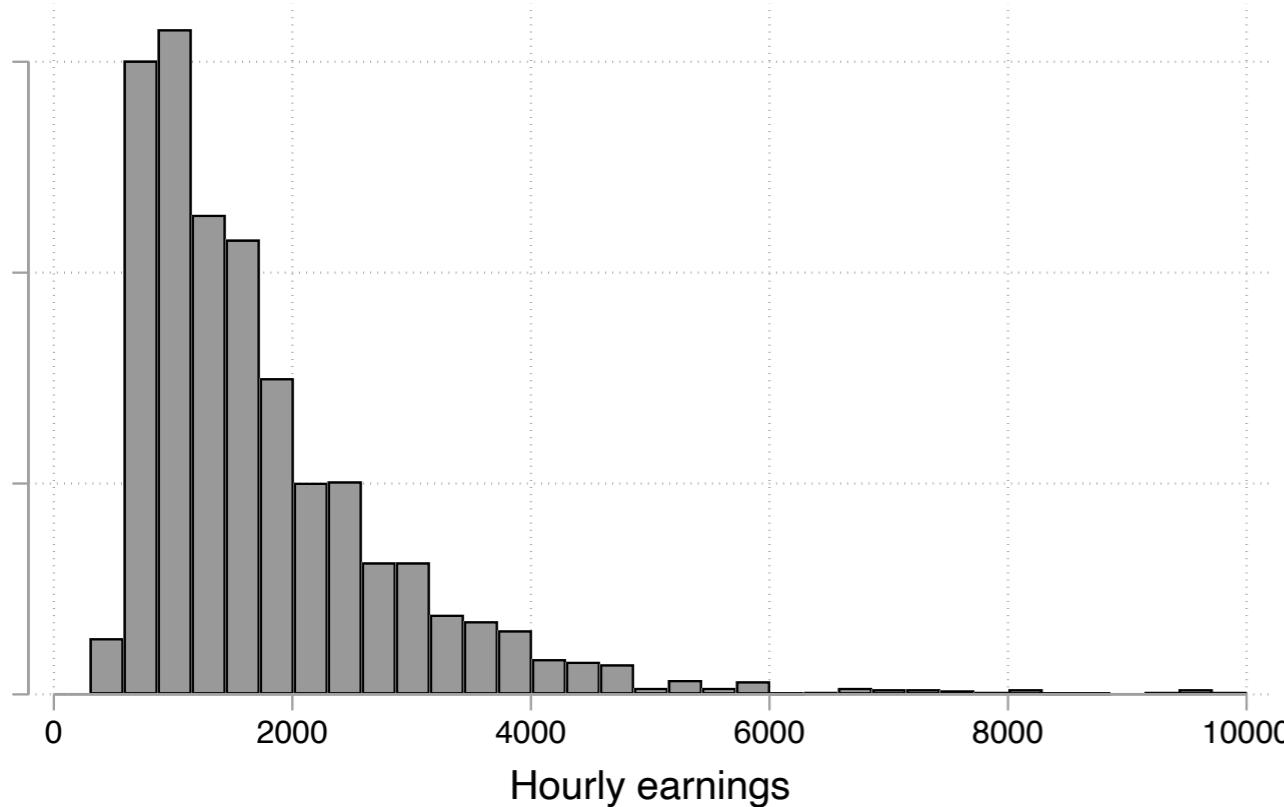
| Source | SS | df | MS | Number of obs | = | 2,805 |
|----------|-------------------|--------------|-------------------|---------------|---|---------------|
| Model | 329439217 | 1 | 329439217 | F(1, 2803) | = | 273.48 |
| Residual | 3.3765e+09 | 2,803 | 1204614.05 | Prob > F | = | 0.0000 |
| Total | 3.7060e+09 | 2,804 | 1321673.47 | R-squared | = | 0.0889 |
| | | | | Adj R-squared | = | 0.0886 |
| | | | | Root MSE | = | 1097.5 |

| wage | Coefficient | Std. err. | t | P> t | [95% conf. interval] |
|----------|------------------|-----------------|--------------|--------------|-----------------------------------|
| numeracy | 7.904687 | .4779924 | 16.54 | 0.000 | 6.967435 8.84194 |
| _cons | -550.2746 | 142.1371 | -3.87 | 0.000 | -828.9786 -271.5707 |

変数の（自然）対数変換

変数が正規分布から乖離しているときや、変数の単位に依存せず効果の大きさを測定したいときには、変数を対数変換することを検討するとよい

時間あたり賃金 Y と、その自然対数をとった値 $\log(Y)$ の分布を比較すると：



補足：ネイピア数・対数関数・自然対数

$e = \lim_{t \rightarrow 0} (1 + t)^{\frac{1}{t}} \simeq 2.7182818\dots$ で定義される数のことをネイピア数という。慣習上、

e を底とする指数 e^x を $\exp(x)$ と表記する。

$\log_a x$ のように表される関数を x の対数関数といい、次のように定義される：

$$a^y = x \leftrightarrow y = \log_a x$$

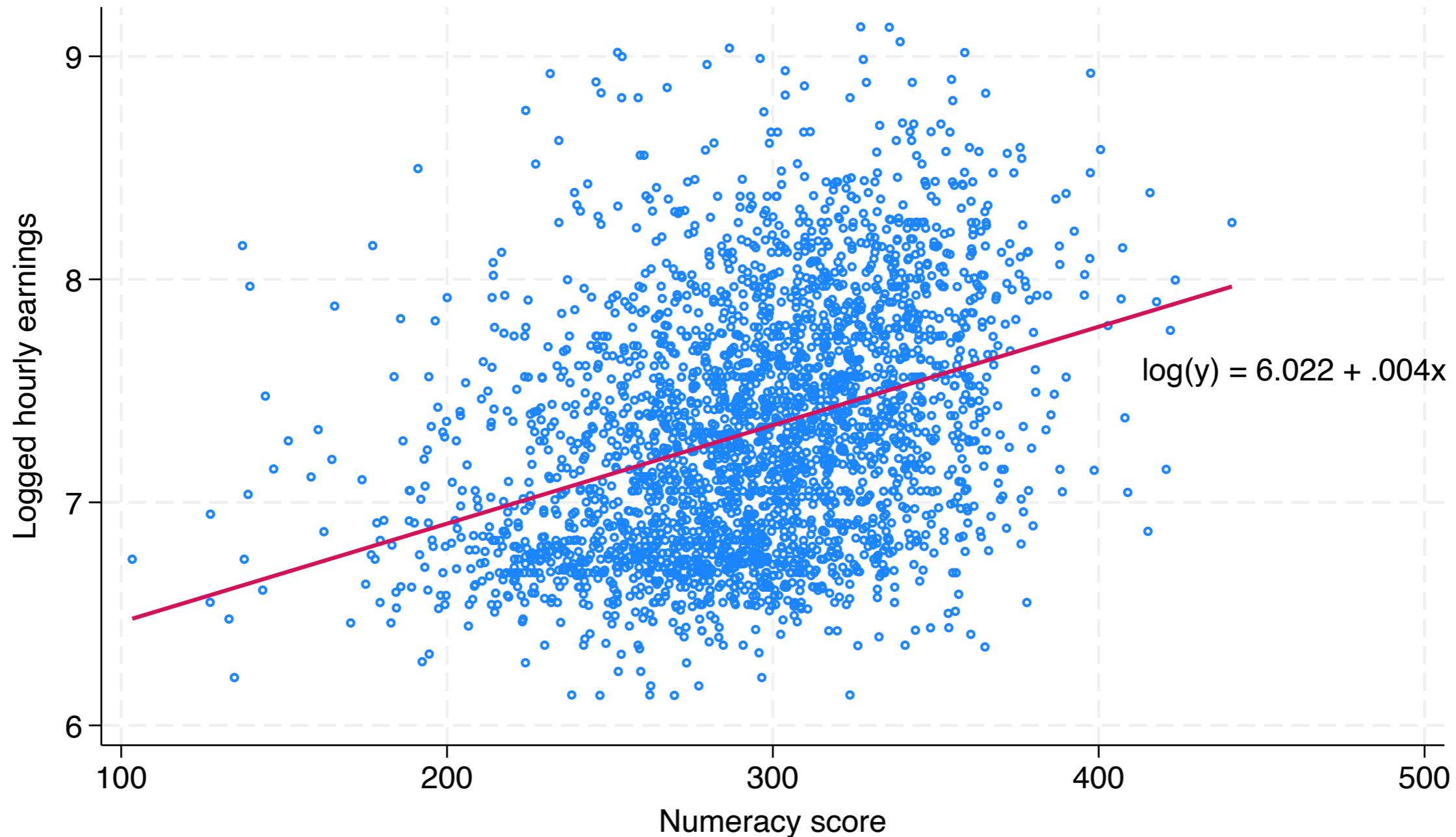
とくに底が e の対数関数を自然対数という。社会科学系の文脈では、この場合は底を省略して、 $e^y = x \leftrightarrow y = \ln(x)$ と書かれることが多い。 $\ln(x)$ の場合もある。

ネイピア数は以下のような便利な性質を持つ。

- 指数の微分： $[\exp(x)]' = \exp(x)$
- 自然対数の微分： $(\ln x)' = 1/x$

対数変換したときの散布図と回帰式

対数を取った変数を従属変数とするときの回帰式： $\log(Y) = \beta_0 + \beta_1 X + \varepsilon$



変数を対数変換したときの限界効果

$$\log(Y + \Delta Y) = \beta_0 + \beta_1(X + 1) + \varepsilon$$

$$\begin{aligned}Y + \Delta Y &= \exp(\beta_0 + \beta_1(X + 1) + \varepsilon) \\&= \exp(\beta_1)\exp(\beta_0 + \beta_1X + \varepsilon)\end{aligned}$$

$$\Delta Y = (\exp(\beta_1) - 1)Y$$

β_1 が小さい値のときは、おおむね「 X が1単位増加すると Y は $\beta_1 \times 100\%$ 増加する」と読める：

$$\beta_1 = 0.1 \leftrightarrow \exp(\beta_1) \simeq 1.11$$

$$\beta_1 = 0 \leftrightarrow \exp(\beta_1) = 1$$

$$\beta_1 = -0.1 \leftrightarrow \exp(\beta_1) \simeq 0.90$$

係数が大きくなるほど両者は一致しない。正確な値を知るには $[\exp(\beta_1) - 1] (\times 100\%)$ を計算する

変数を対数変換したときの係数の読み方

従属変数 独立変数 解釈

Y X Xが1単位高いと、Yが β_1 高い

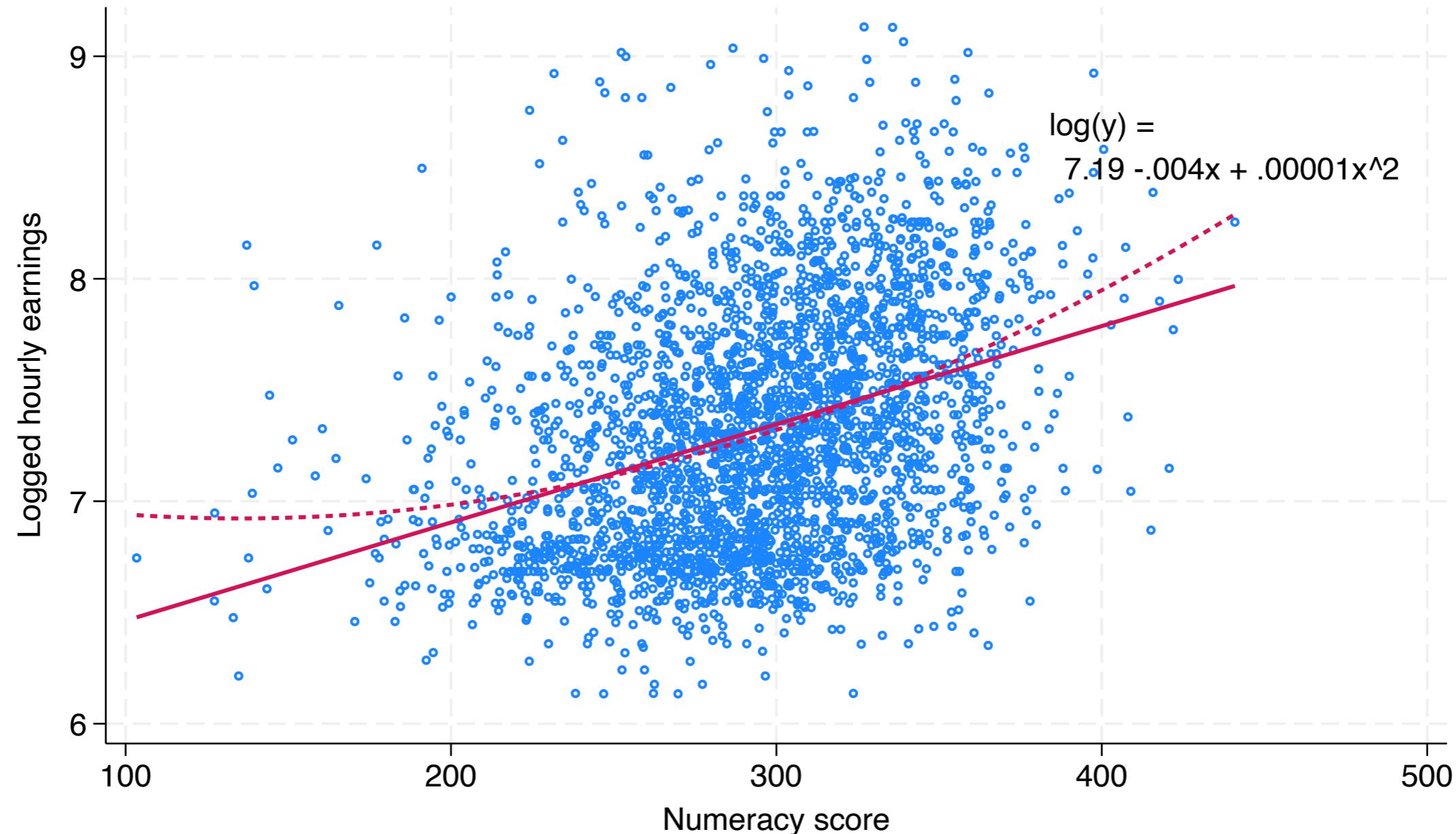
$\log(Y)$ X Xが1単位高いと、Yが $100 \times \beta_1\%^*$ 高い

Y $\log(X)$ Xが1%高いと、Yが $\beta_1/100$ 高い

$\log(Y)$ $\log(X)$ Xが1%高いと、Yが $\beta_1\%$ 高い

非線形の関係を考慮する：2乗項の投入

数的思考力がとくに高い人の間で正の関連が強い可能性がある。たとえばこのような回帰式を考えてみる： $\log(Y) = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$



2乗項を投入したときの限界効果

回帰式 $Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon$ において、 X が1単位増加したときの Y の増加量
(限界効果) は、もともとの X の値によって異なる

$$\begin{aligned} Y + \Delta Y &= \beta_0 + \beta_1(X + 1) + \beta_2(X + 1)^2 + \varepsilon \\ &= (\beta_0 + \beta_1 X + \beta_2 X^2) + \beta_1 + (2X + 1)\beta_2 \\ \Delta Y &= \beta_1 + (2X + 1)\beta_2 \end{aligned}$$

回帰式の形状：

$\beta_2 < 0$ ならば、 $-\beta_1/2\beta_2$ を底とする、上に凸な二次関数

$\beta_2 > 0$ ならば、 $-\beta_1/2\beta_2$ を底とする、下に凸な二次関数

対数や2次の項を含めた回帰分析を推定する

対数変換した変数を使ったり、2乗項を考慮した回帰分析を推定し、結果を出してみよう（4.1.2）

2乗項を含めた場合には、どのような形状になるかがぱっとはわからないので、その都度形状をmarginsコマンドを使って確認するとよい

複数の回帰分析の結果を比較する

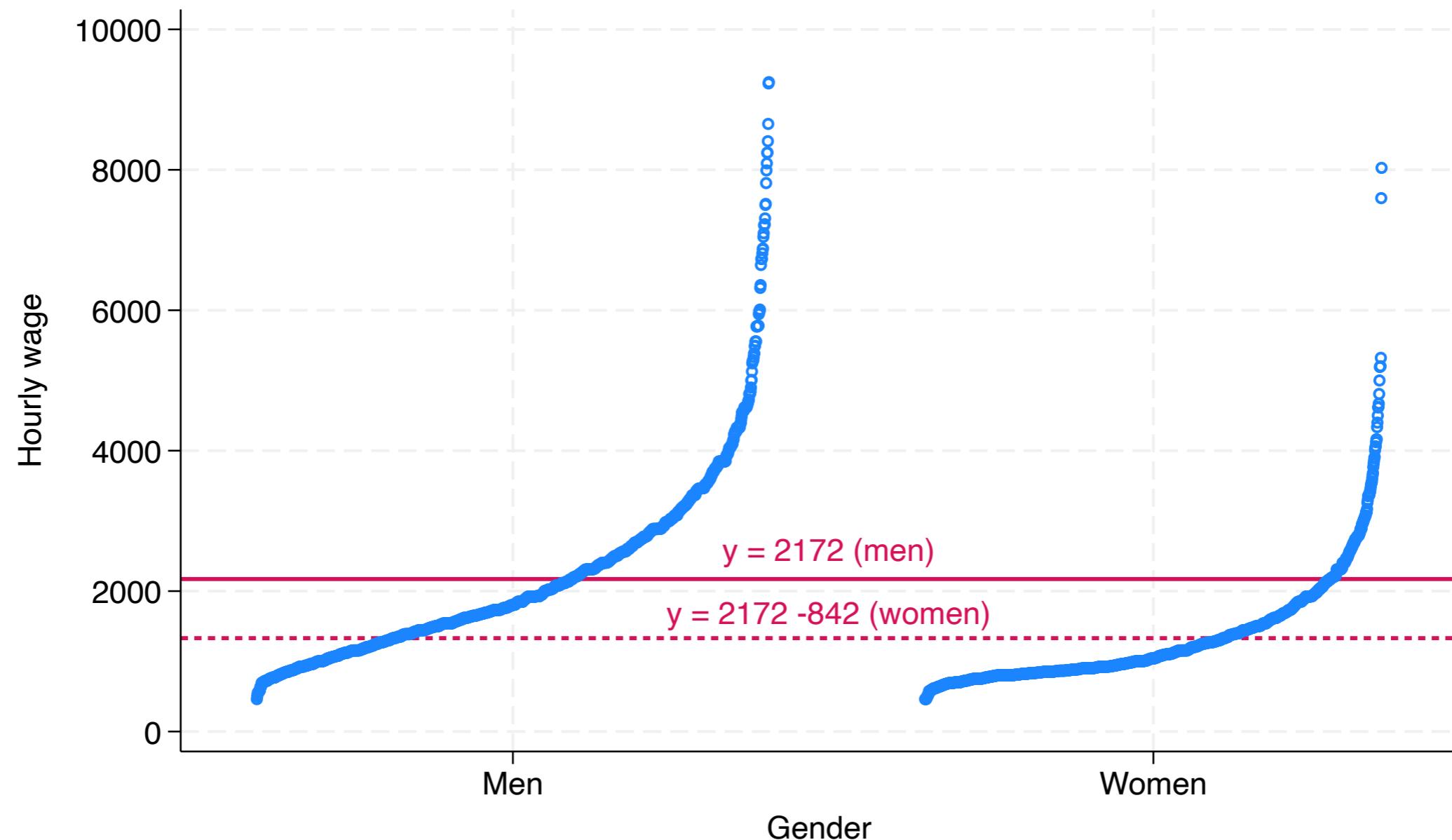
対数変換しない賃金を使ったときの結果、対数変換した賃金を使った結果、2乗項を使った結果の3つを並べて結果を比較してみよう（4.1.3）

`estout` <http://repec.sowi.unibe.ch/stata/estout/esttab.html> コマンドを使おう

1. 回帰分析を推定
2. `estimates store` で結果を保存
3. `esttab` で複数の結果を並べて表示

Xがカテゴリ変数の場合

独立変数がカテゴリ変数（性別など）の場合、独立変数ごとに賃金の散布図（ストリップ・プロット）を描くと次のようになる。切片の高さの差がグループ間の差を表す



ダミー変数と結果の解釈

男性であれば0、女性であれば1をとる変数 D （ダミー変数）を作り、 D を独立変数とする回帰式 $Y = \beta_0 + \beta_1 D + \varepsilon$ を推定する。

このときの傾き β_1 は、 $D = 0$ のグループ（参照カテゴリ）とくらべて $D = 1$ のグループの値がどの程度高いか（低いか）を表す。

$D = 0$ （男性）のとき： $Y = \beta_0 + \varepsilon$, $E(Y|D = 0) = \beta_0$

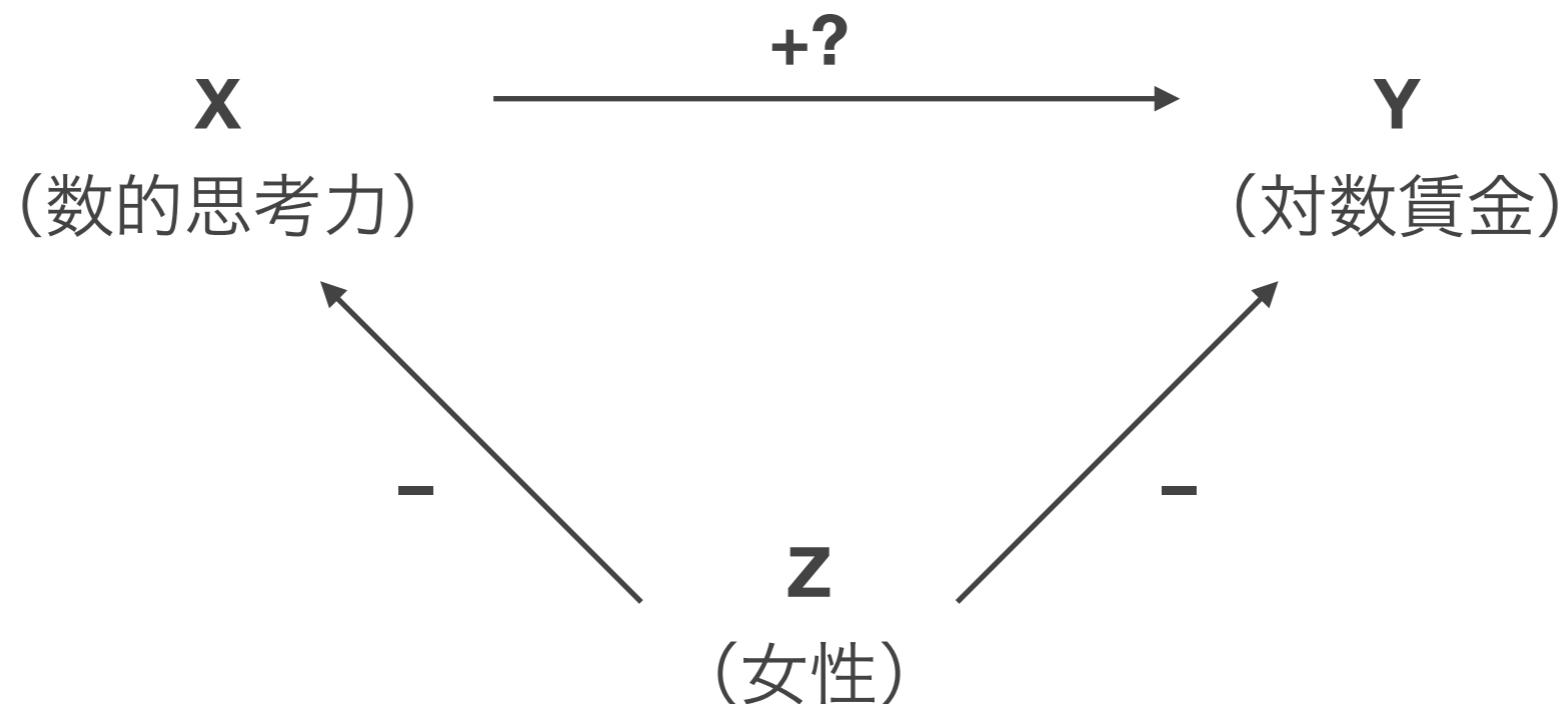
$D = 1$ （女性）のとき： $Y = \beta_0 + \beta_1 + \varepsilon$, $E(Y|D = 1) = \beta_0 + \beta_1$

ダミー変数を使った回帰分析を推定し、結果を比較してみよう（4.1.4）

重回帰分析を活用する

重回帰分析による交絡要因confounderの除去

単回帰分析で数的思考力が高い人ほど賃金が高い傾向があることがわかった。しかし、この相関を即因果関係と呼ぶことはできない



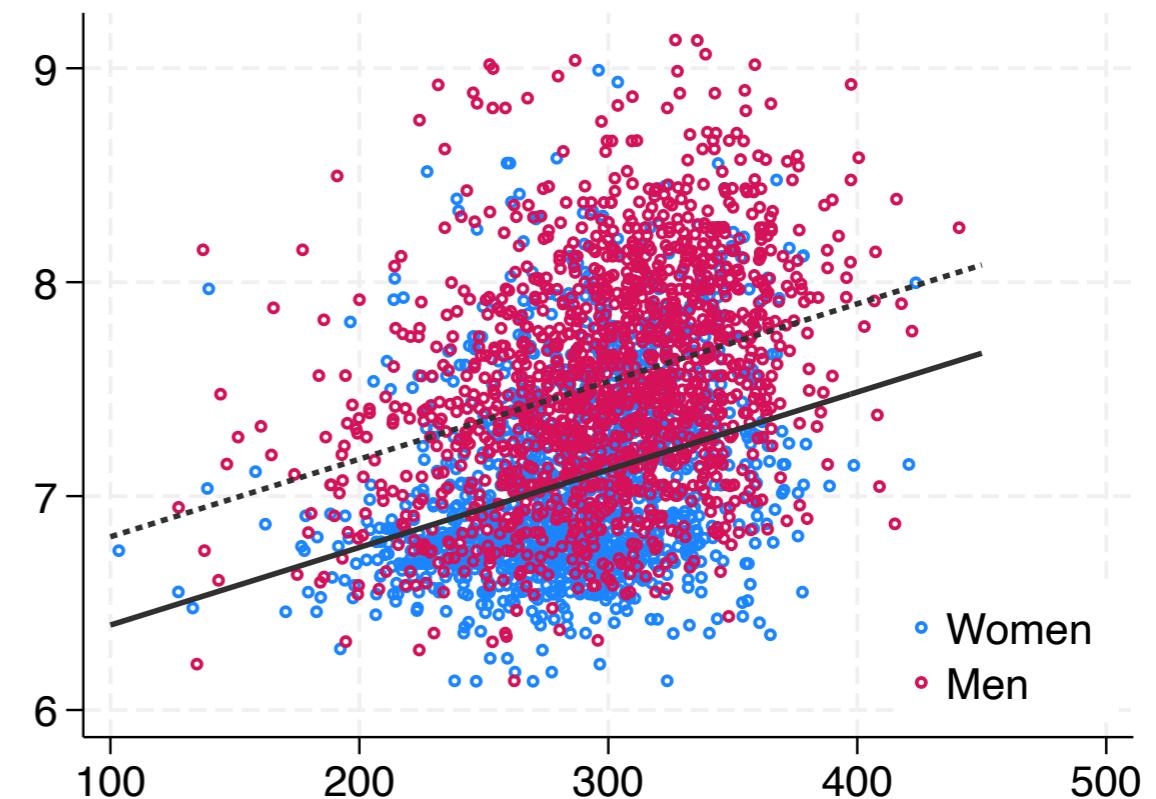
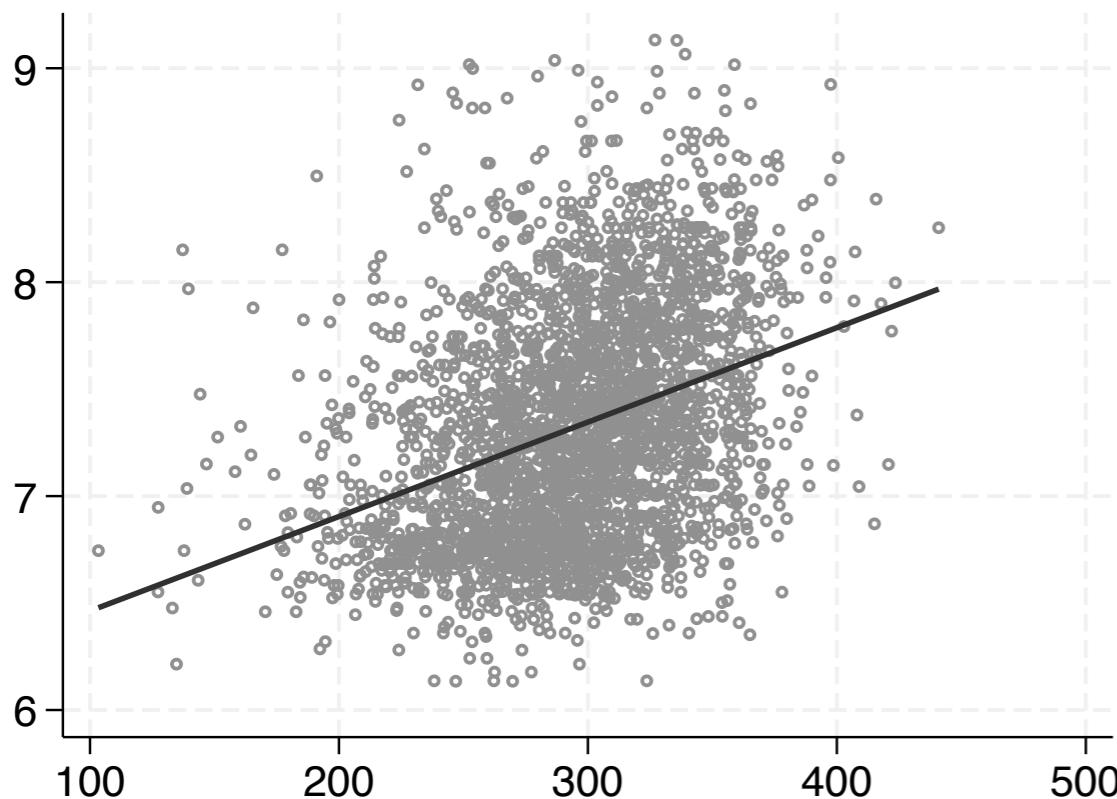
XとYの両者に影響する交絡要因Zを統制することで、Yに対するXの因果効果に近づくことができる（条件付きの関連を推定できる）

単回帰分析と重回帰分析

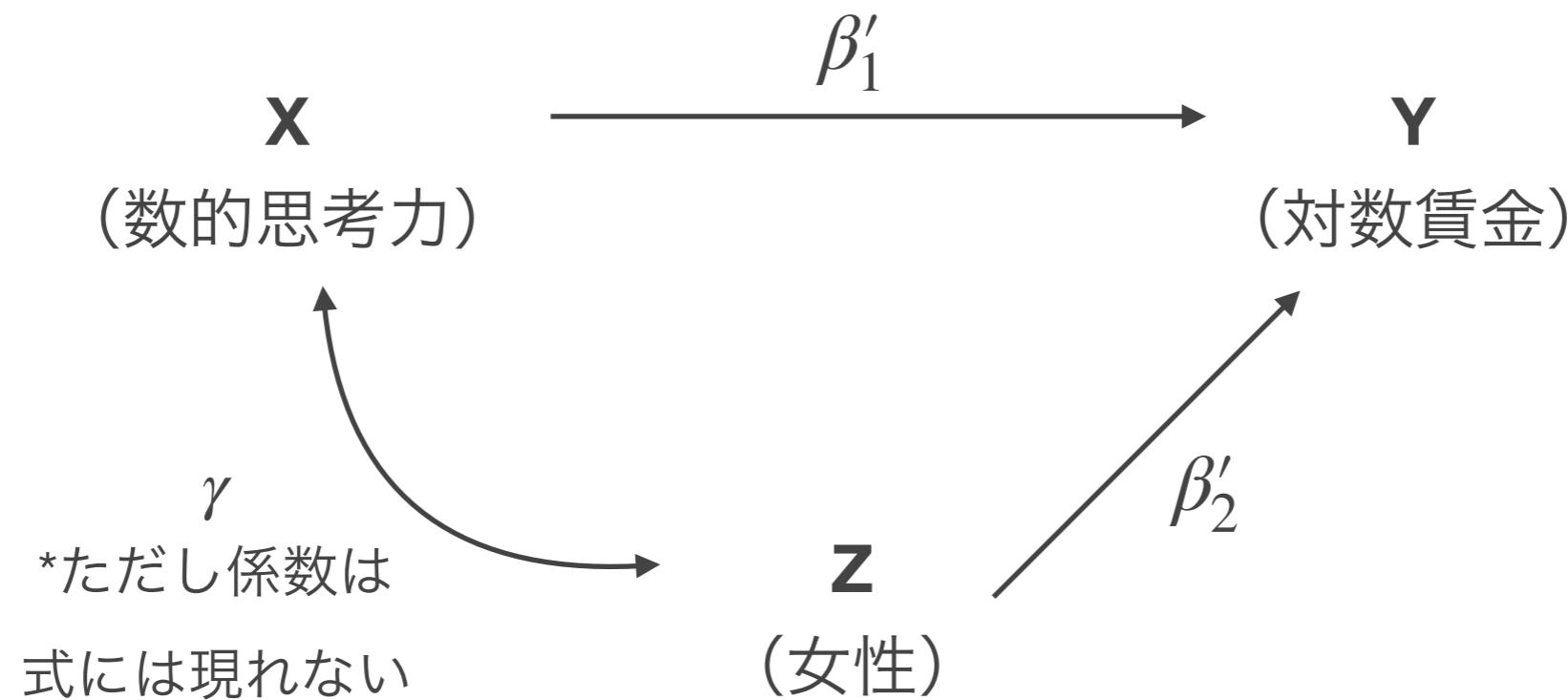
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

$$Y = \beta'_0 + \beta'_1 X + \beta'_2 Z + \varepsilon'$$

ZがXとYの両方と何らかの相関を示す場合、両回帰式でXの係数は一致しない



重回帰分析の推定結果と統制前係数のバイアス



XとZの相関 ZとYの相関 Z統制前の係数と統制後のXの係数の大小

$\gamma > 0$ $\beta'_2 > 0$ $\beta_1 > \beta'_1$ —— 統制しないと過大推計

$\gamma < 0$ $\beta'_2 < 0$ $\beta_1 > \beta'_1$ —— 統制しないと過大推計

$\gamma < 0$ $\beta'_2 > 0$ $\beta_1 < \beta'_1$ —— 統制しないと過小推計

$\gamma > 0$ $\beta'_2 < 0$ $\beta_1 < \beta'_1$ —— 統制しないと過小推計

単回帰分析と重回帰分析で主張できる内容が異なる

単回帰分析からいえること：

数的思考力が高いほど賃金が高い傾向がある

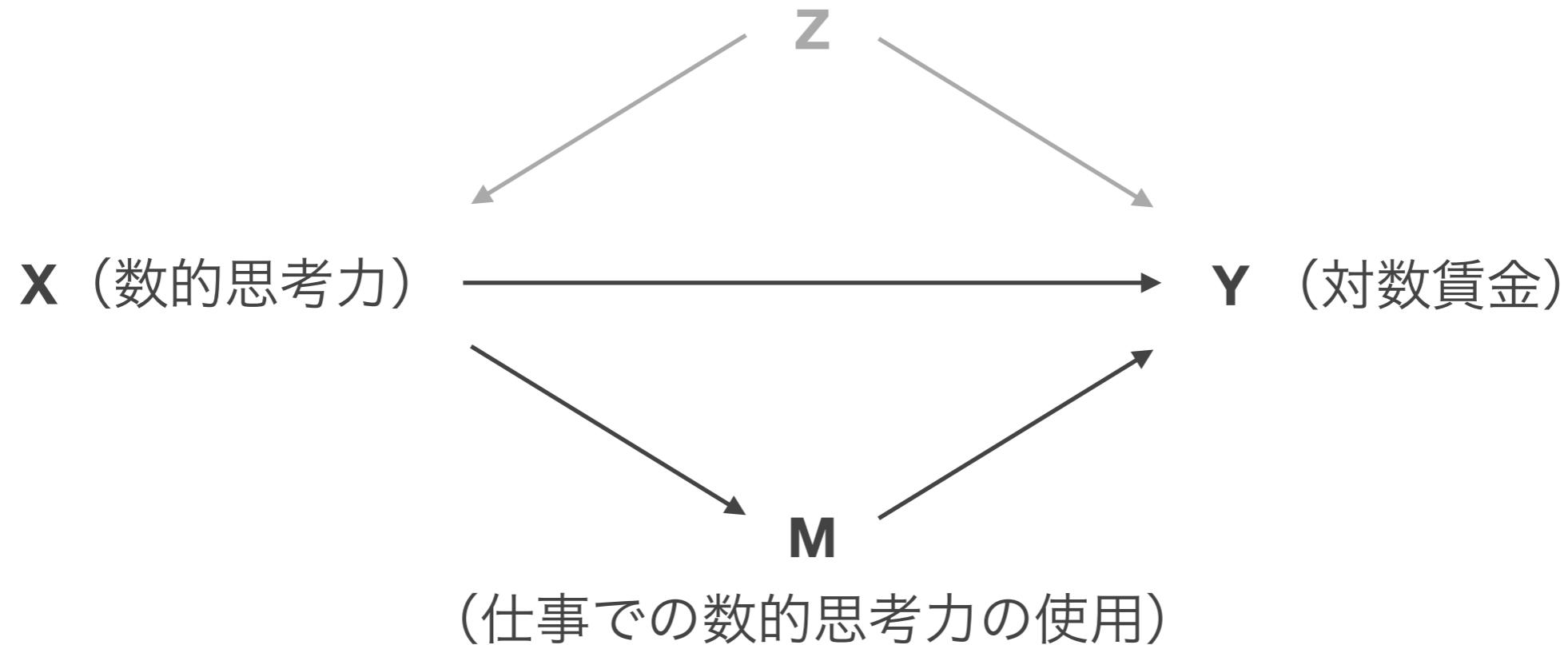
→独立変数が無作為に割り当てられていない限り、「数的思考力が高いと賃金が高くなる」とはいえない

重回帰分析からいえること：

性別が同じでも、数的思考力が高いほど賃金が高い傾向がある

→すべての交絡要因を統制していない限り、「数的思考力が高いと賃金が高くなる」とはいえない（たぶん近づいてはいる）

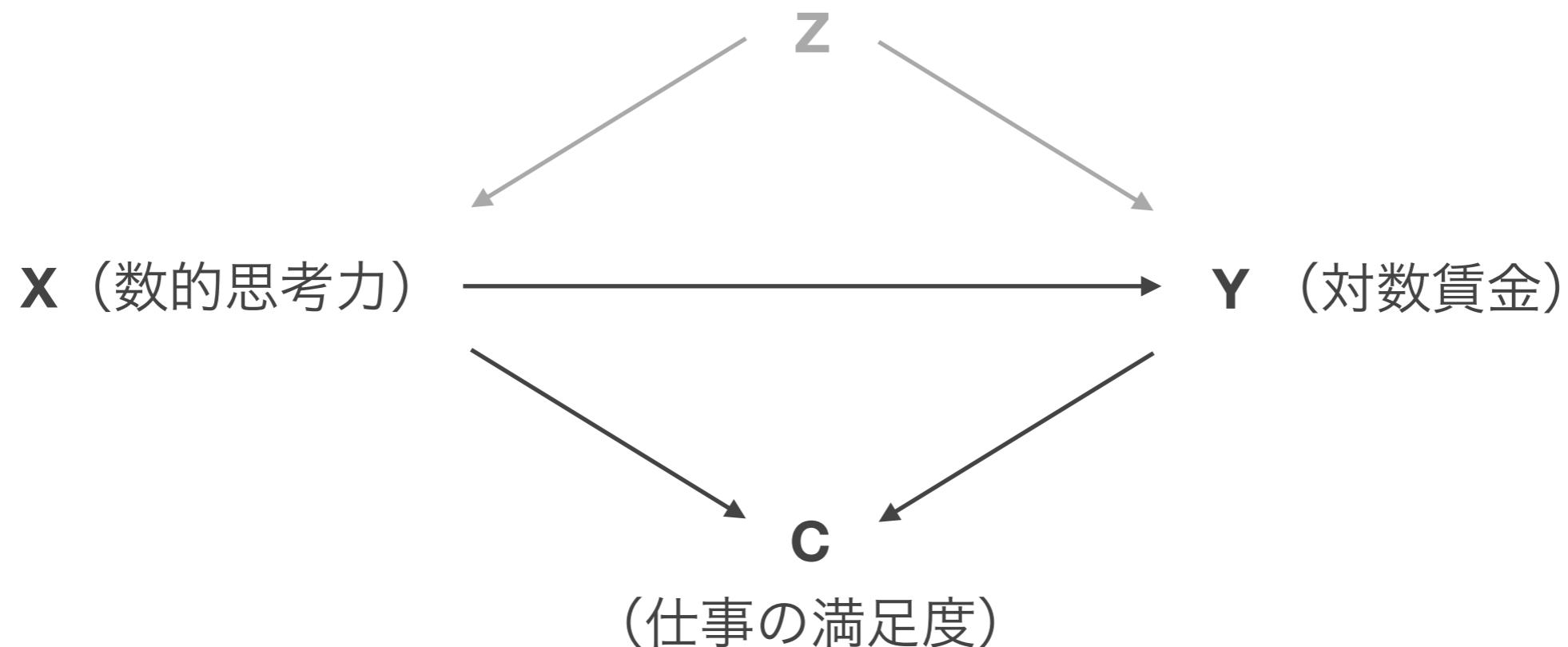
統制すべき変数を吟味する：媒介要因 mediator



Mのような変数を投入するかどうかは知りたい効果の中身に依存する

- もし知りたい効果が「同じくらい数的思考力を使う仕事をしていたとしてもなお数的思考力が賃金を高める効果」ならば、**M**は統制すべき
- 「数的思考力が賃金を高める効果」ならば、**M**は統制すべきではない

統制すべき変数を吟味する：合流点バイアス



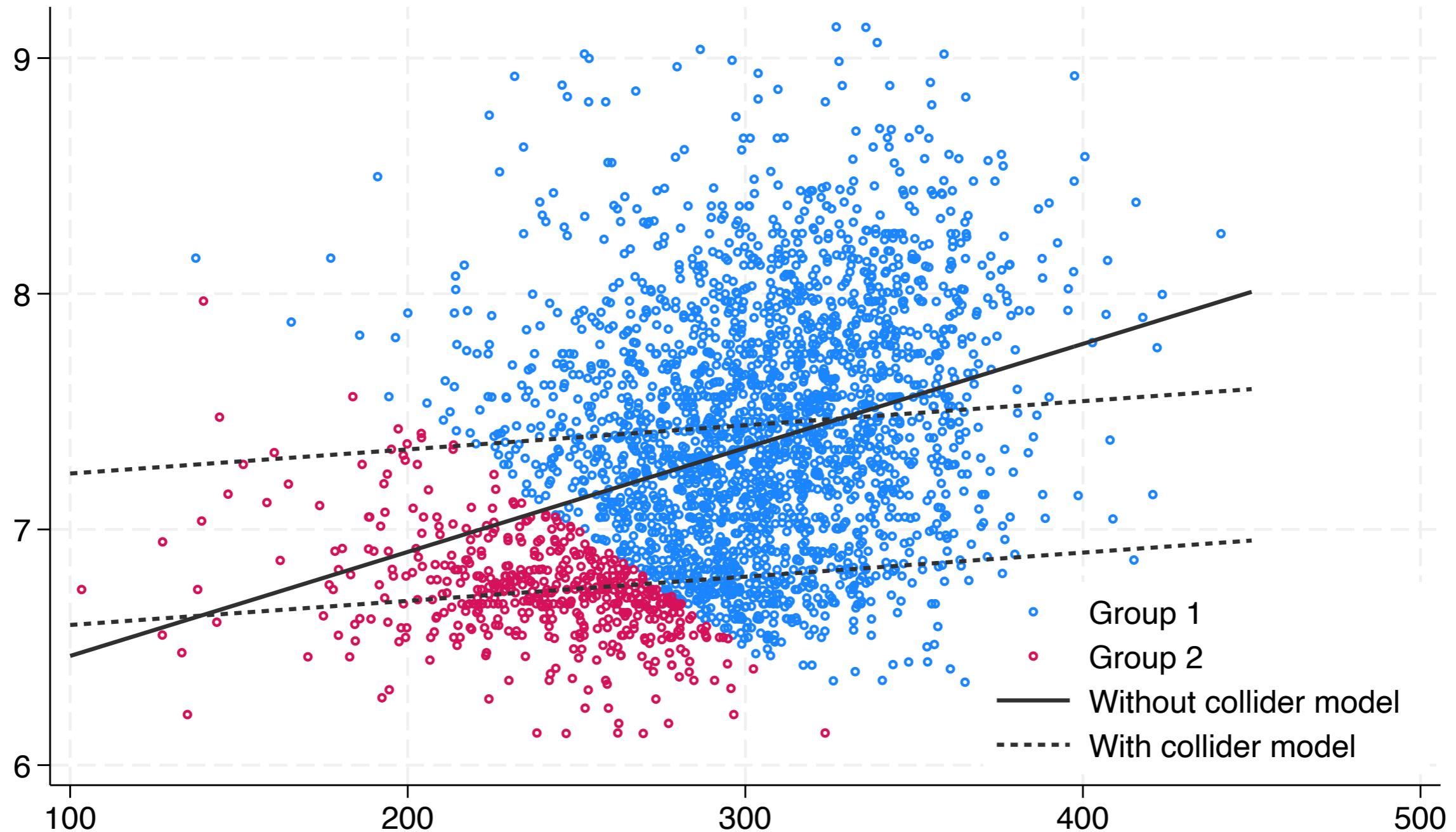
重回帰分析を変数の効果を知るために使う場合、**Cのような変数は投入してはいけない**

**Cのような変数を統制することによってXの係数にバイアス（合流点バイアス
Collider bias）が生じる**

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology*, 40, 31–53.

合流点バイアスの仮想例

合流点となる変数を統制すると、数的思考力の係数にバイアスが生じる



小括：回帰分析の使い方

回帰分析は、適切に交絡要因を統制することで、自分の研究で知りたいことに近づける。しかし、適切でない要因を統制すると、かえって遠ざかってしまう

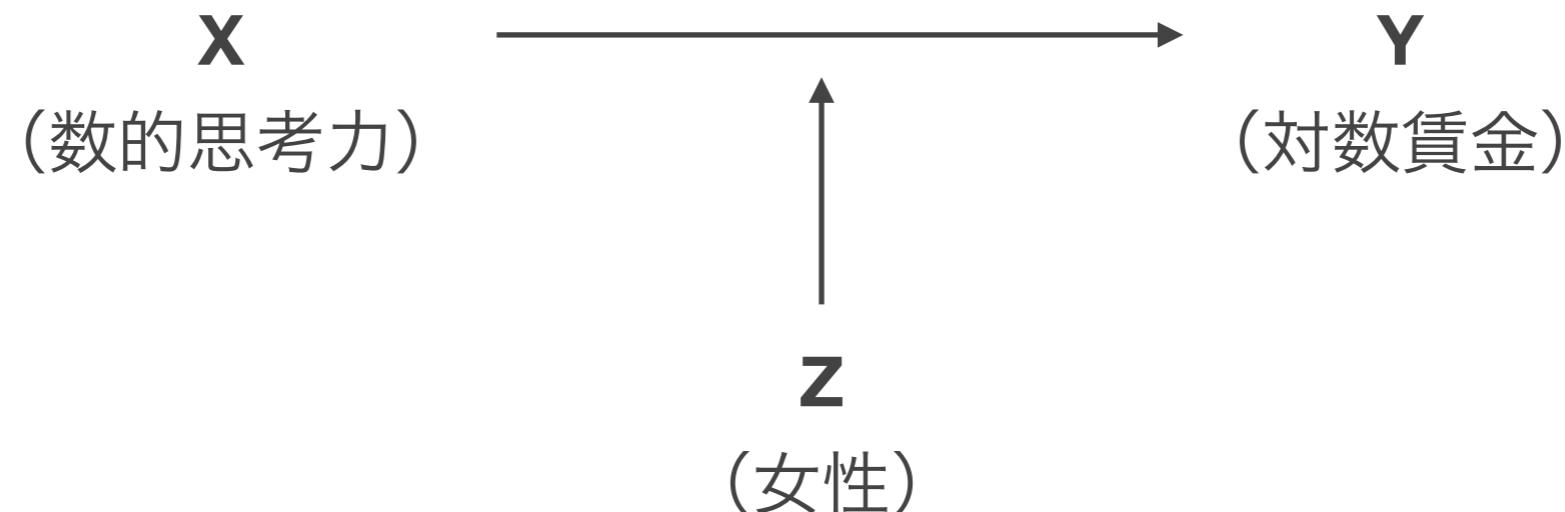
どのような効果が知りたいのか

そのために、どのような交絡要因や媒介要因を統制すればよいか

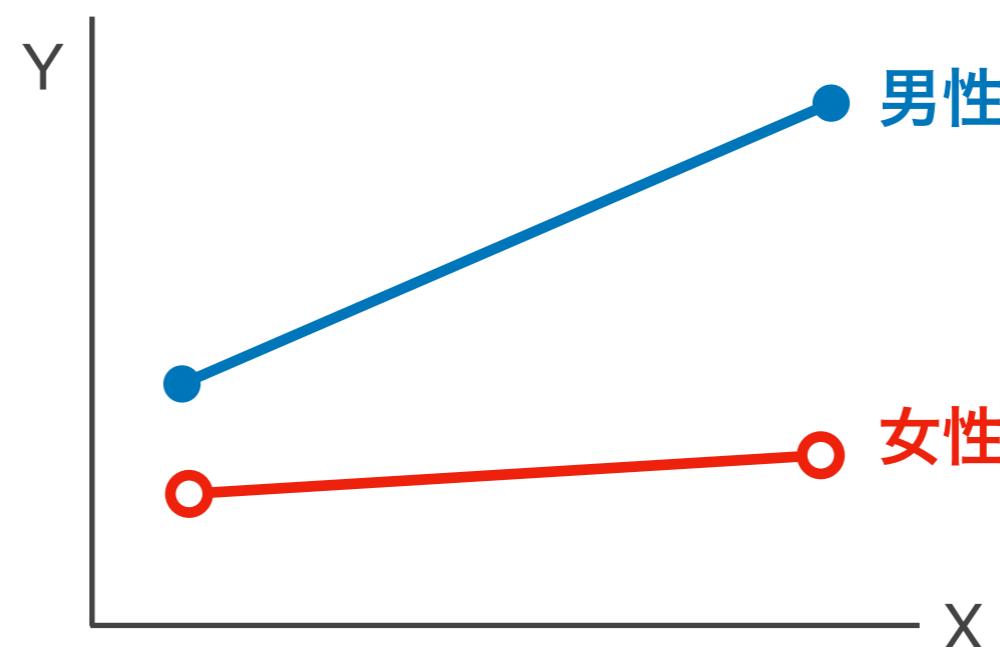
(手持ちのデータでは考慮できないとしても) どのようなバイアスがありうるか
を考えることが大事

交絡要因や媒介要因、合流点（と考えられる）を統制した重回帰分析をそれぞれ
推定し、結果を比較してみよう（4.2.1）

調整効果 moderation / 交互作用 interaction



変数の効果が別の変数の水準によって異なることが考えられる。このような関連を指して、**調整効果**あるいは**交互作用（効果）**という



調整効果を推定するためのモデル

見たい変数Xと、調整変数Zをかけ算した変数を独立変数として投入する。

Zがダミー変数のときを考える：

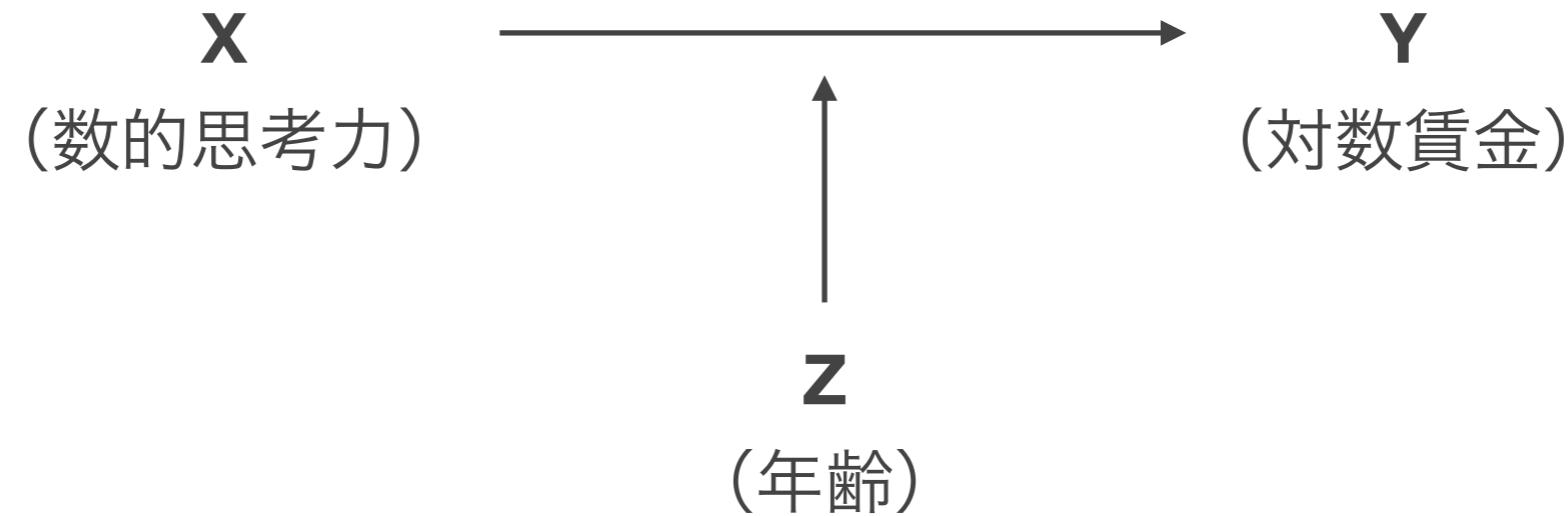
$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

Z = 0 (男性) のとき : $Y = \beta_0 + \beta_1 X + \varepsilon$. $\partial Y / \partial X = \beta_1$

Z = 1 (女性) のとき : $Y = \beta_0 + \beta_2 + (\beta_1 + \beta_3)X + \varepsilon$. $\partial Y / \partial X = \beta_1 + \beta_3$

β_3 は、男性におけるXの傾きとくらべて、女性におけるXの傾きがどの程度大きいか（小さいか）を表す。

調整変数が連続変数のとき



Zが連続変数のときを考える：

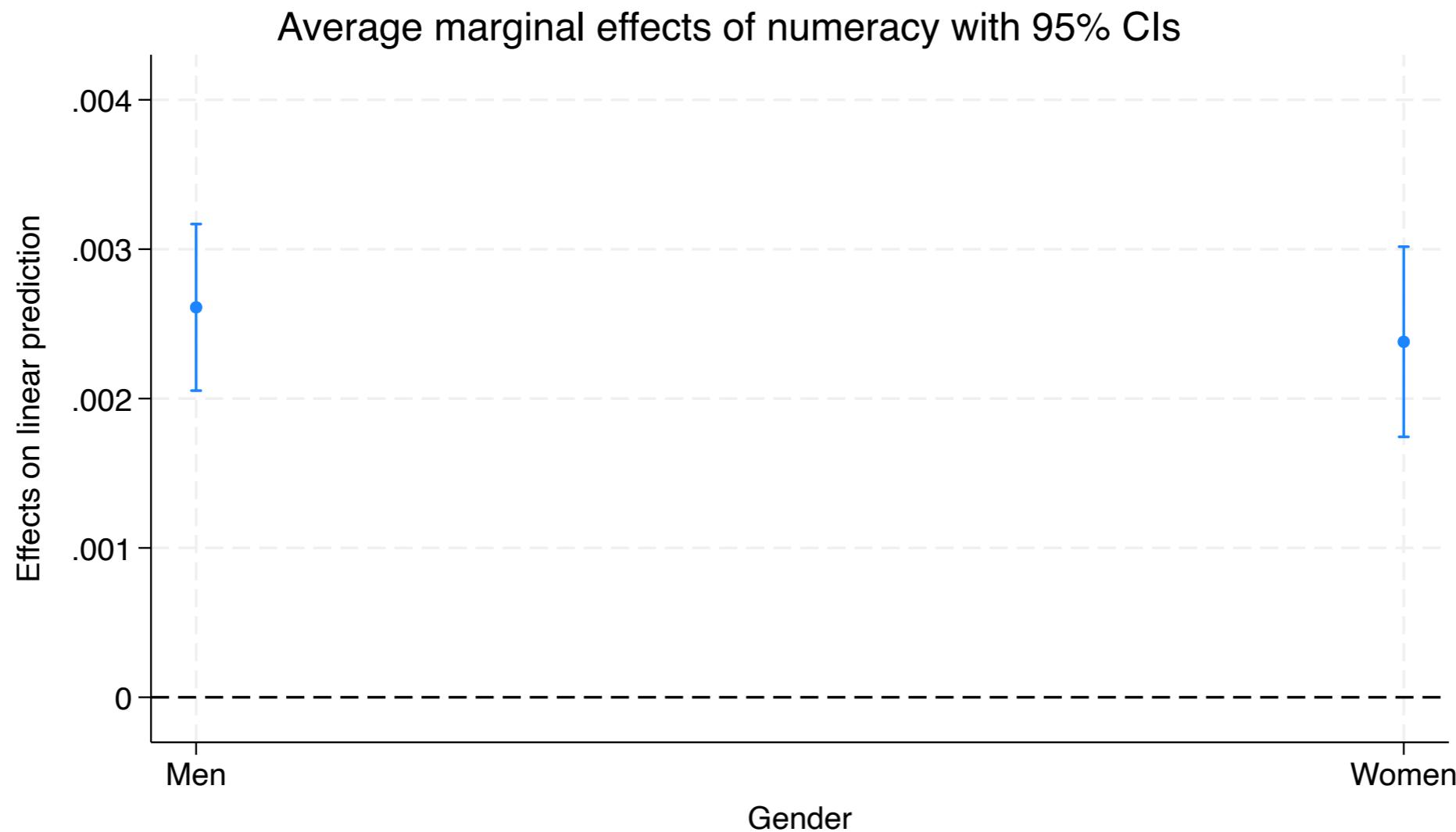
$$\begin{aligned}Y &= \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon \\&= \beta_0 + (\beta_1 + \beta_3 Z)X + \beta_2 Z + \varepsilon\end{aligned}$$

$$\frac{\partial Y}{\partial X} = \beta_1 + \beta_3 Z$$

β_3 は、Zの値に応じてXの傾きがどの程度加算されるかを表す。

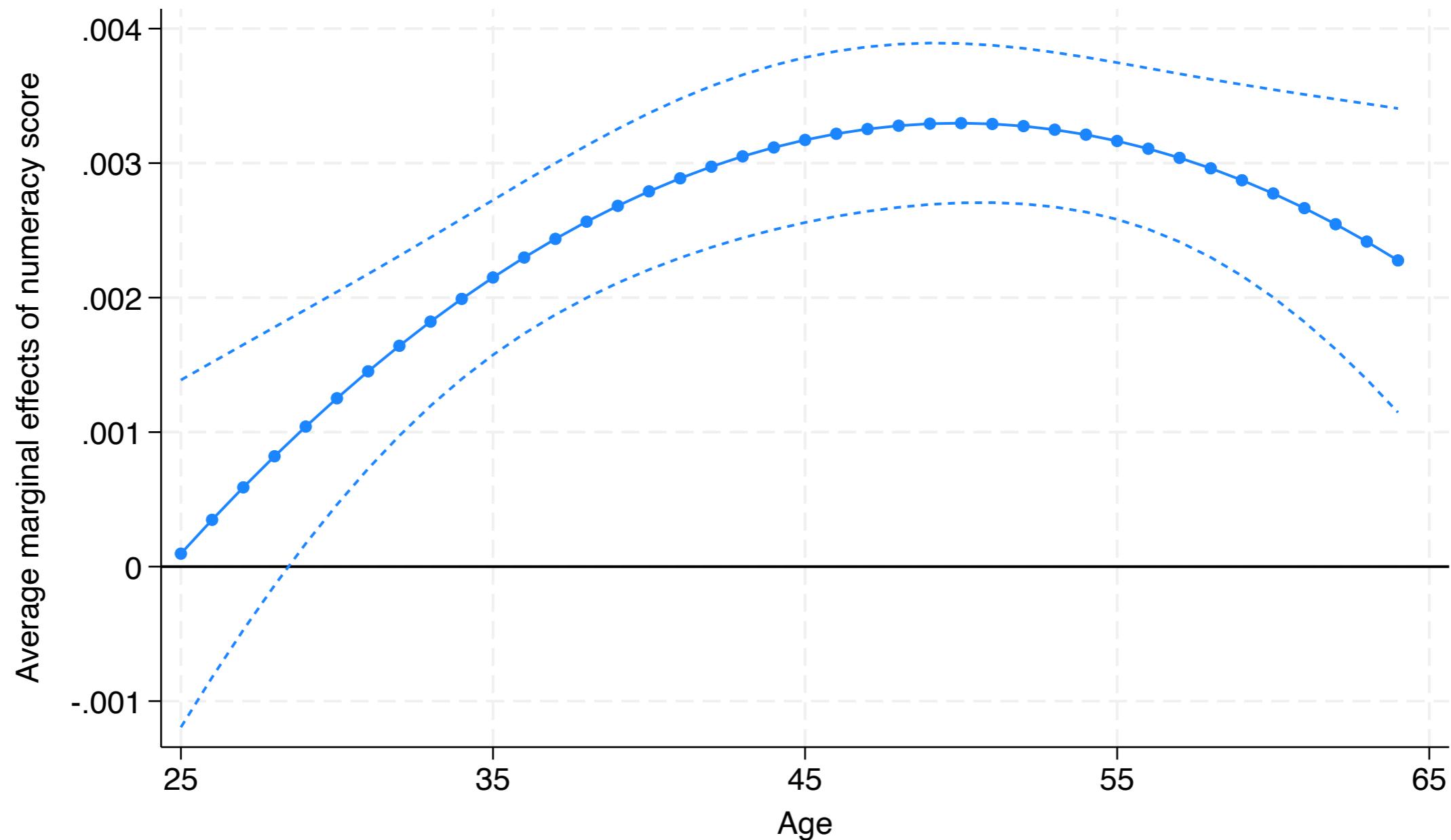
性別ごとにみた限界効果

交互作用項を含めた場合には各係数の解釈が少し煩雑になるため、以下のようにZの値別の限界効果を確認するとよい



年齢ごとにみた限界効果

限界効果を具体的に図示することで、どのくらいの年齢ではどの程度の効果があるのかを効果的に示すことができる



調整効果の推定

性別と数的思考力、年齢と数的思考力、年齢 2 乗と数的思考力の交互作用項を含むモデルを推定し、結果を比較してみよう（4.3.1）

調整効果（交互作用）をより解釈しやすくするため、限界効果に関するグラフを作成しよう（4.3.2）

多重共線性 Multicollinearity

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$ において、 $\text{Cor}(X_2, X_3 | X_1)$ が非常に高い場合、 β_2, β_3 の係数が不安定となりその標準誤差も大きくなる。これを多重共線性という

Stataでは、`regress y x, vif` で多重共線性の程度をチェックできる

多重共線性を気にする必要はあるかどうかは、問い合わせに依存する

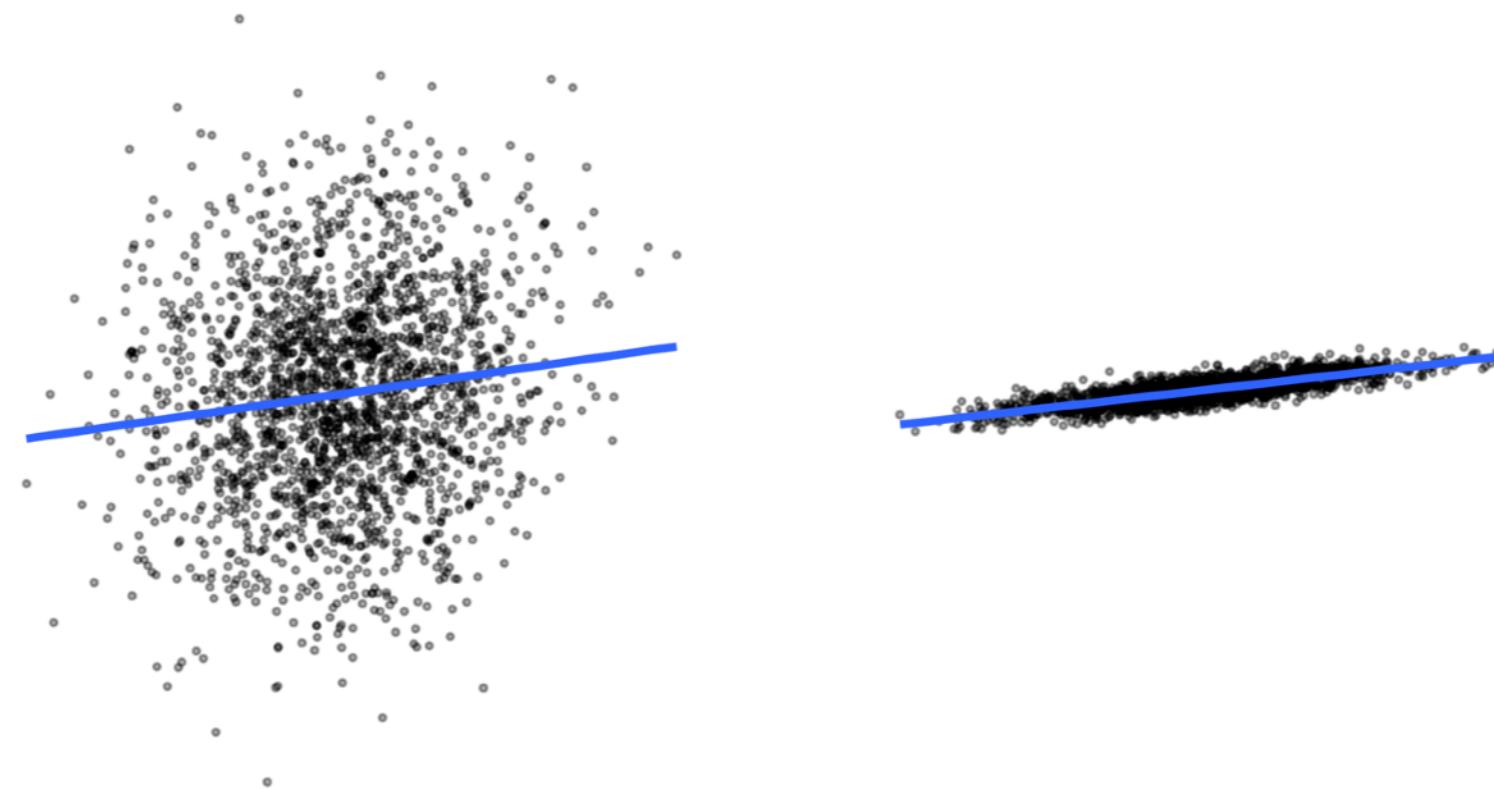
- 知りたい係数が β_1 であるなら、気にする必要はない
- 知りたい係数が β_2 または β_3 のどちらかまたは両方なら、
 - VIF以外にも標準誤差が過剰に大きくなったりしているかなどをチェックし、問題なさそうなら、気にする必要はない (VIFがxxx以下なら大丈夫 (じゃない) ... という基準は全く意味がない)
 - 問題ありそうなら、理論的な妥当性などを再検討し、どちらかを除外する

決定係数 R²

決定係数：回帰式により得られる予測分散がYの分散に占める割合。

$$R^2 = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)} = 1 - \frac{\text{Var}(\varepsilon)}{\text{Var}(Y)}$$
 で定義される。

- 決定係数が高い = 残差が小さいということなので、決定係数を高くできれば標準誤差を小さくできる。しかしそのためだけに独立変数を増やすのは本末転倒
- 異なるサンプル間で決定係数の大きさは直接比較できない



ロジスティック回帰分析

男性は女性よりも職場で多くの訓練を受けているか？

日本の労働市場では、企業内訓練（OJT）によって技能を培うことが重要視されている。男女間の技能の差、ひいては賃金格差を生む要因として、男性が女性よりもOJTを受けやすいことがあるかもしれない

「この1年間に、実践研修（OJT）や上司または同僚による研修に参加したことがありますか」という質問項目をOJT受講の有無とみなし、性別とOJT受講の関係を分析してみよう

(参考文献)

Estevez-Abe, Margarita, Torben Iversen, and David Soskice. 2001. “Social Protection and the Formation of Skills: A Reinterpretation of the Welfare State.” Pp. 145–83 in *Varieties of Capitalism: The Institutional Foundations of Comparative Advantage*. Oxford University Press.

クロス集計表

| Gender | OJT | | Total |
|--------|-------|-------|--------|
| | No | Yes | |
| Men | 885 | 598 | 1,483 |
| | 59.68 | 40.32 | 100.00 |
| Women | 896 | 426 | 1,322 |
| | 67.78 | 32.22 | 100.00 |
| Total | 1,781 | 1,024 | 2,805 |
| | 63.49 | 36.51 | 100.00 |

男性は女性と比べてこの1年にOJTを受けている割合が高い（男性は女性と比べてOJTを受けやすい）。

効果を定義する：差を見るか、比を見るか

| Gender | OJT | | Total |
|--------|-------|-------|--------|
| | No | Yes | |
| Men | 885 | 598 | 1,483 |
| | 59.68 | 40.32 | 100.00 |
| Women | 896 | 426 | 1,322 |
| | 67.78 | 32.22 | 100.00 |
| Total | 1,781 | 1,024 | 2,805 |
| | 63.49 | 36.51 | 100.00 |

二値変数を従属変数とする場合、「効果」の測り方に2つの見方がある

差を見る：男性は女性と比べて8.1%ポイント ($=40.32 - 32.22$) OJTを受ける割合が高い

比を見る：男性は女性と比べてxx倍（後述）、OJTを受けやすい

差に着目して効果を推定する：線形確率モデル

2値の従属変数に対して線形回帰分析を当てはめるモデルを指して**線形確率モデル**

(**Linear Probability Model, LPM**) という。

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$$

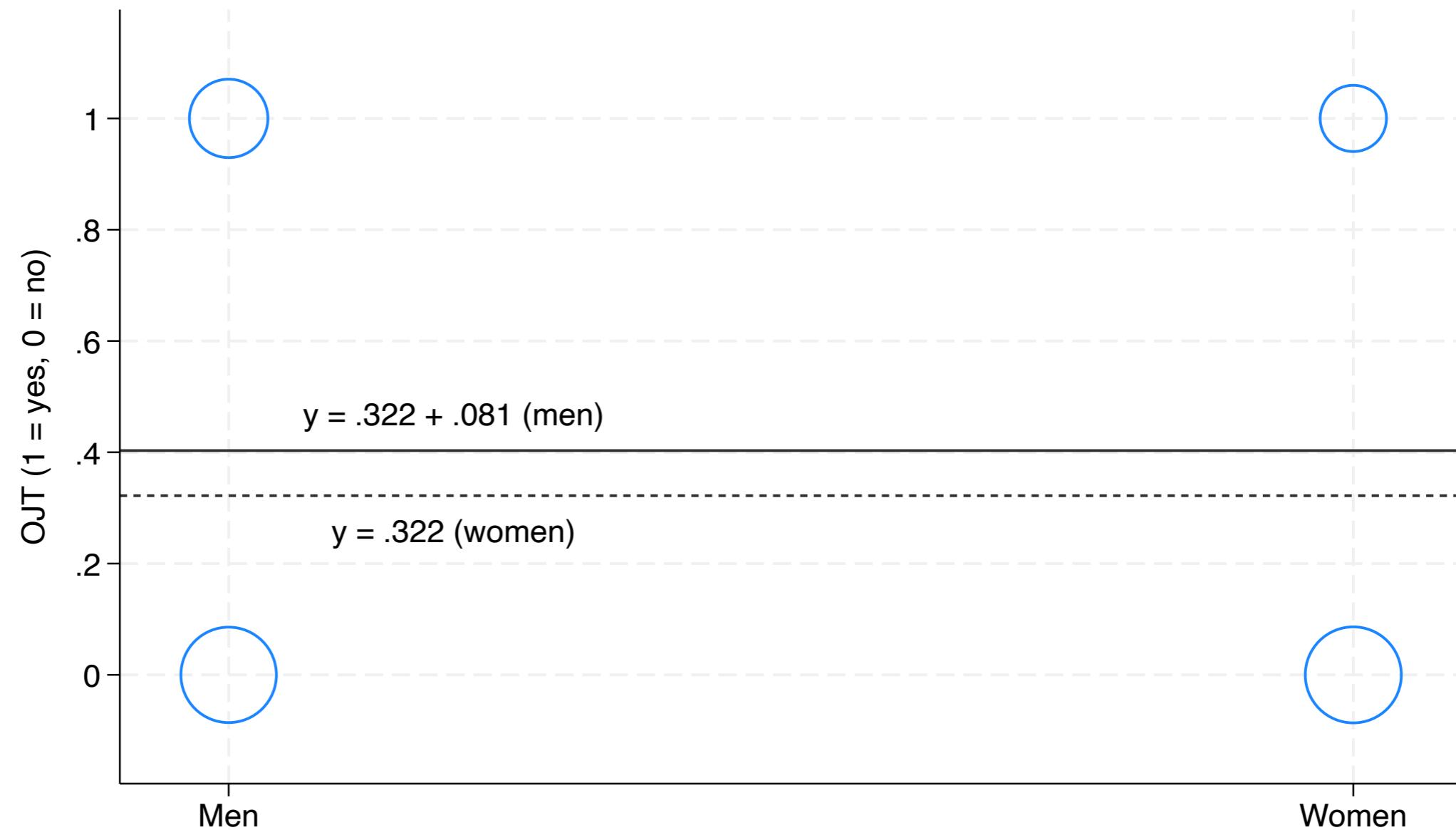
期待値を取ると

$$E(Y|X_1, \dots, X_k) = \Pr(Y|X_1, \dots, X_k) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

傾きの係数は、 X が1単位増加したときの $\Pr(Y)$ の增加分を表す。

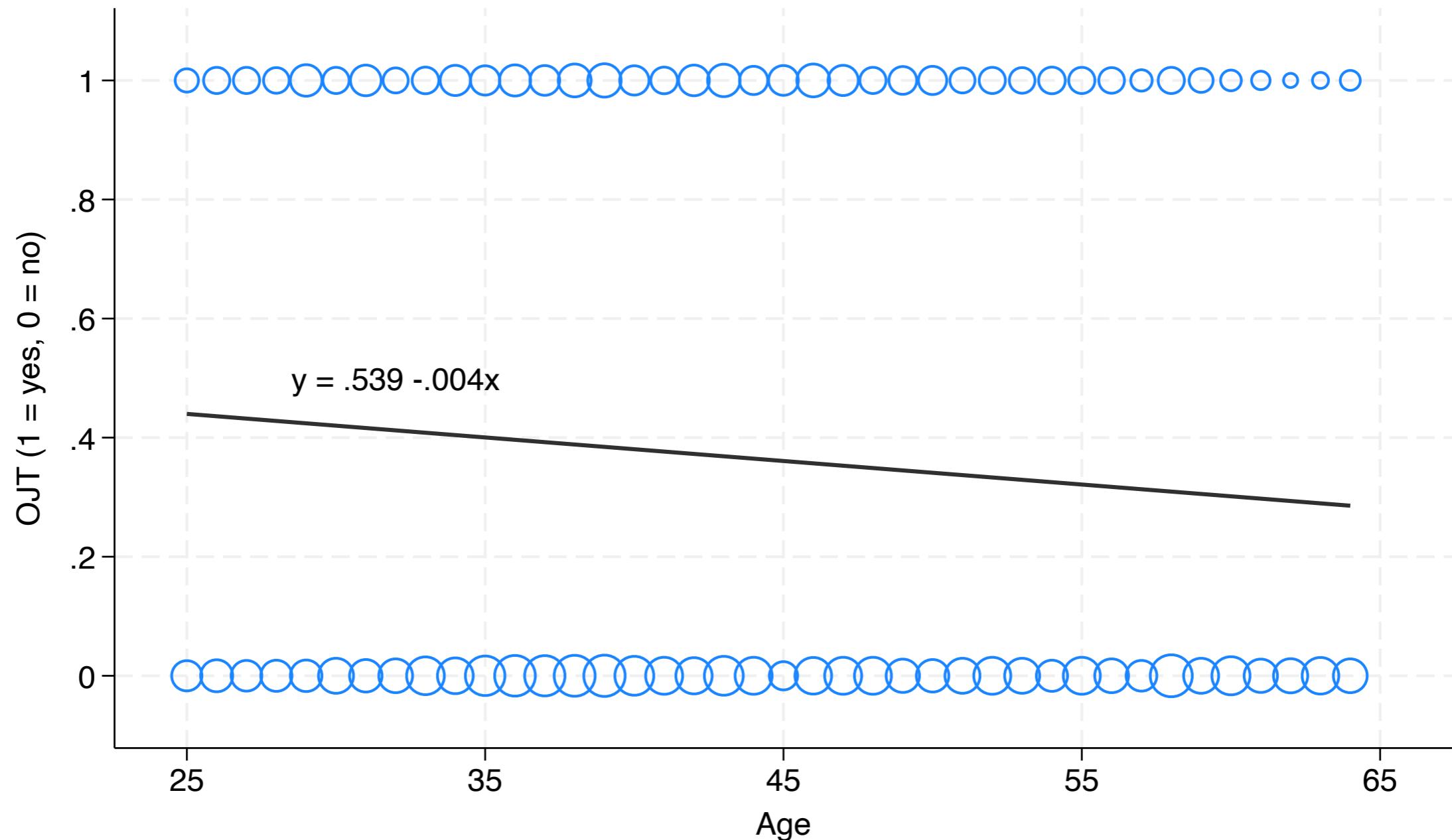
Yが二値変数、Xが二値変数の場合

散布図および、最小二乗法によって引かれた回帰直線は次のようになる



Yが二値変数、Xが連続変数の場合

同じように散布図に回帰直線を引くことで関係性を表現できる



線形確率モデルを推定する

5_logit2024-09-04.doを開き、線形確率モデルを推定してみよう（5.1.1）

```
. reg ojt ib2.gender age, vce(robust) // ロバスト標準誤差
```

Linear regression

| | Robust | | | | | |
|---------------|--------|-----------|---|------|----------------------|----------|
| | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
| Number of obs | | | | | | = 2,805 |
| F(2, 2802) | | | | | | = 21.90 |
| Prob > F | | | | | | = 0.0000 |
| R-squared | | | | | | = 0.0147 |
| Root MSE | | | | | | = .47815 |

| ojt | Coef. | Std. Err. | t | P> t | [95% Conf. Interval] | |
|--------|-----------|-----------|-------|-------|----------------------|-----------|
| gender | | | | | | |
| Men | .0791581 | .0180424 | 4.39 | 0.000 | .0437804 | .1145357 |
| age | -.0038766 | .0008104 | -4.78 | 0.000 | -.0054656 | -.0022876 |
| _cons | .4935005 | .0387239 | 12.74 | 0.000 | .4175703 | .5694307 |

線形確率モデル（差による効果測定）の問題点

1. 残差が正規分布しない（不均一分散）ため標準誤差にバイアスが生じる
→ロバスト標準誤差 (heteroskedasticity-robust standard error) を使うことで対処可能
2. 予測値が確率の定義上あり得ない数値（0未満、あるいは1より大きい）になることがある
→従属変数が1を取る割合が0または1に近いほどそのリスクが高くなる
3. 関数型の誤り：もし真の関係が非線形——従属変数が1をとる確率が異なる個人間で、ある独立変数が1単位増えることによる確率の増加量が異なる（天井効果 ceiling effect ないし床効果 flooring effect）——のであれば、変数の「効果」の推定として不適切

Mood, C. (2010). Logistic regression: Why we cannot do what we think we can do, and what we can do about it. *European Sociological Review*, 26(1), 67–82.

比に着目して効果を推定する：(対数) オッズ比

| | | Y | |
|---|---|-------------|-------------|
| X | | Failure (0) | Success (1) |
| 1 | 1 | $1 - p_1$ | p_1 |
| | 2 | $1 - p_2$ | p_2 |

X = 1におけるオッズ : $p_1/(1 - p_1)$

X = 2におけるオッズ : $p_2/(1 - p_2)$

X = 2に対するX = 1のオッズ (= オッズ比) : $\frac{p_1/(1 - p_1)}{p_2/(1 - p_2)}$

対数オッズ比 : $\log \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} = \log(p_1/(1 - p_1)) - \log(p_2/(1 - p_2))$

オッズ比を計算する

| Gender | OJT | | Total |
|--------|----------------|----------------|-----------------|
| | No | Yes | |
| Men | 885 59.68 | 598 40.32 | 1,483 100.00 |
| Women | 896 67.78 | 426 32.22 | 1,322 100.00 |
| Total | 1,781 63.49 | 1,024 36.51 | 2,805 100.00 |

男性のオッズ（OJTなしに対するOJTありの比）： $40.32 / 59.68 = 0.676$

女性のオッズ（OJTなしに対するOJTありの比）： $32.22 / 67.78 = 0.475$

オッズ比： $0.676 / 0.475 = 1.42$

→女性と比べて男性はオッズでみて**1.42倍**OJTを受けやすい

対数オッズ比： $\log(0.676 / 0.475) = \log(0.676) - \log(0.475) = 0.352$

ロジスティック回帰分析 Logistic regression

以下のような式を当てはめる分析を指してロジスティック回帰分析あるいはロジットモデル **Logit model** とよぶ

$$\Pr(Y = 1) = \frac{\exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)}{1 + \exp(\beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k)} \quad \text{または}$$

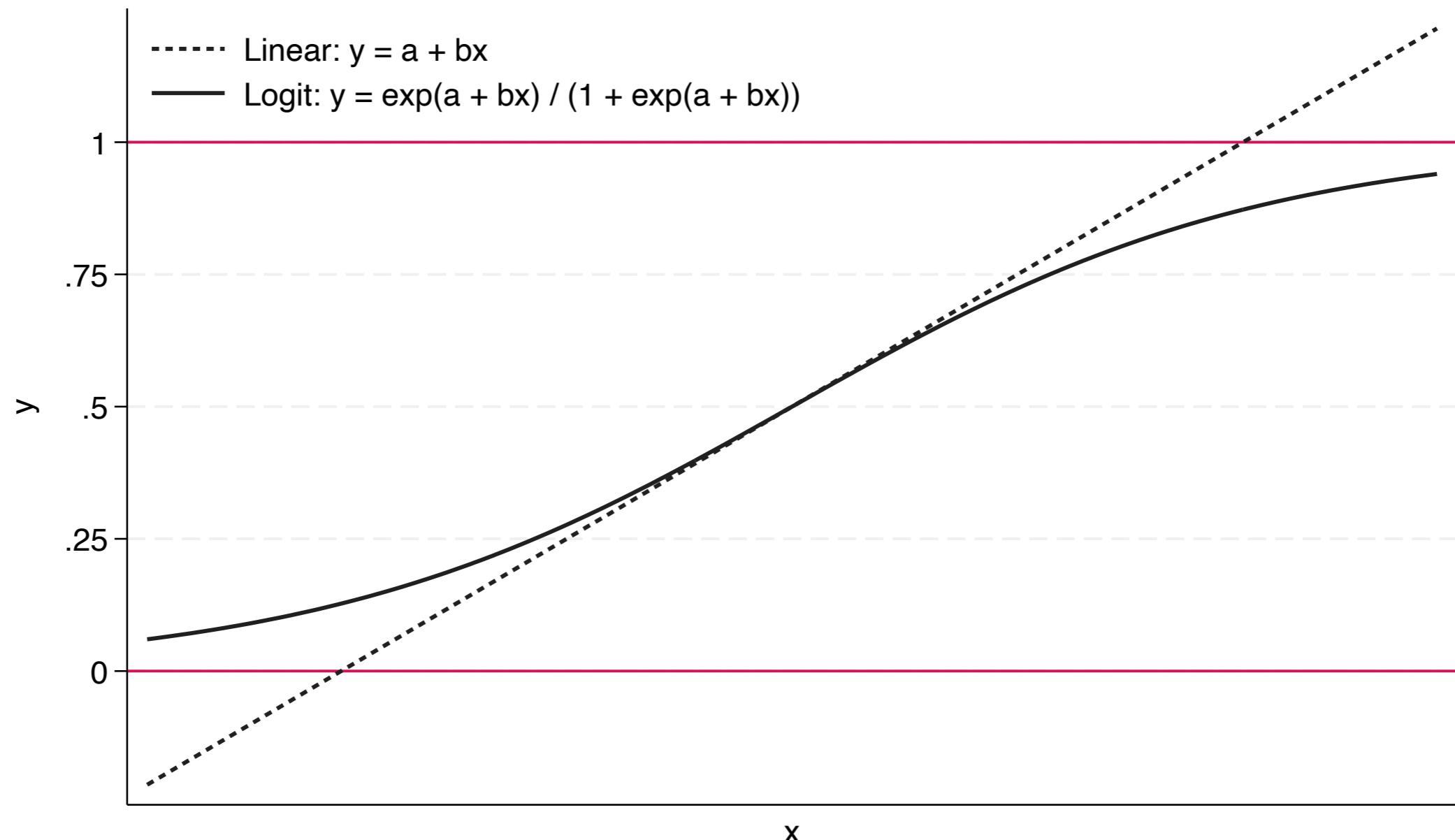
$$\log \frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k \quad \text{と表記する。}$$

各係数は最尤法 Maximum likelihood estimation によって推定される。

係数 β_k は、 X_k が 1 単位増加したときの従属変数の対数オッズの増加量を示す。

直線とロジスティック曲線の違い

ロジスティック曲線は、yの値が0.5から離れて0または1に近づくほど、x1単位の変化に対して緩やかに確率が上昇するS字形になる



ロジットモデルを推定する

ロジットモデルを推定してみよう (5.2.1)

```
. logit ojt ib2.gender
```

```
Iteration 0:  log likelihood = -1840.8525
Iteration 1:  log likelihood = -1830.9335
Iteration 2:  log likelihood = -1830.9276
Iteration 3:  log likelihood = -1830.9276
```

```
Logistic regression                                         Number of obs     =      2,805
                                                               LR chi2(1)       =      19.85
                                                               Prob > chi2      =     0.0000
Log likelihood = -1830.9276                                Pseudo R2       =     0.0054
```

| ojt | Coef. | Std. Err. | z | P> z | [95% Conf. Interval] |
|--------|-----------|-----------|--------|-------|----------------------|
| gender | | | | | |
| Men | .3515042 | .079156 | 4.44 | 0.000 | .1963613 .5066471 |
| _cons | -.7435011 | .0588514 | -12.63 | 0.000 | -.8588477 -.6281544 |

女性の対数オッズに対する男性の対数オッズ
(対数オッズ比)

注意点は線形回帰分析と共通

線形確率モデルもロジットモデルも、モデルを作り解釈するうえで基本的に注意すべきことは同じ

- 係数が正（負）であると、従属変数が1をとる確率が高い（低い）
- 回帰分析のときと同じく、2乗項や対数変換した変数を必要に応じて使用する
- 複数の独立変数を投入する場合には、何を使うかを吟味する
- 変数どうしをかけ算した変数を投入することで調整効果（交互作用効果）を検討できる

線形確率モデルとロジットモデルの結果の比較

線形確率モデルとロジットモデルを推定し、結果を比較してみよう（5.2.2）

| | LPM | | Logit | |
|----------------|----------|---------|-----------|---------|
| main | | | | |
| Men | 0.054** | (0.018) | 0.255** | (0.085) |
| Women | 0.000 | (.) | 0.000 | (.) |
| Junior high | 0.000 | (.) | 0.000 | (.) |
| Senior high | 0.052 | (0.030) | 0.309 | (0.179) |
| Junior college | 0.154*** | (0.033) | 0.788*** | (0.185) |
| University | 0.278*** | (0.032) | 1.286*** | (0.178) |
| Age | 0.018* | (0.007) | 0.087** | (0.033) |
| Age # Age | -0.000** | (0.000) | -0.001** | (0.000) |
| Constant | -0.127 | (0.151) | -3.005*** | (0.709) |
| Observations | 2805 | | 2805 | |
| r2 | 0.062 | | | |
| r2_p | | | 0.048 | |

Standard errors in parentheses

* p<0.05, ** p<0.01, *** p<0.001

確率の差による解釈とオッズ比による解釈

線形確率モデル：確率の差

他の要因を一定として、男性がOJTを受ける確率は女性より5.4%ポイント高い

ロジットモデル：（対数）オッズ比

他の要因を一定として、男性がOJTを受けるオッズは女性の $1.29 = \exp(0.255)$ 倍である

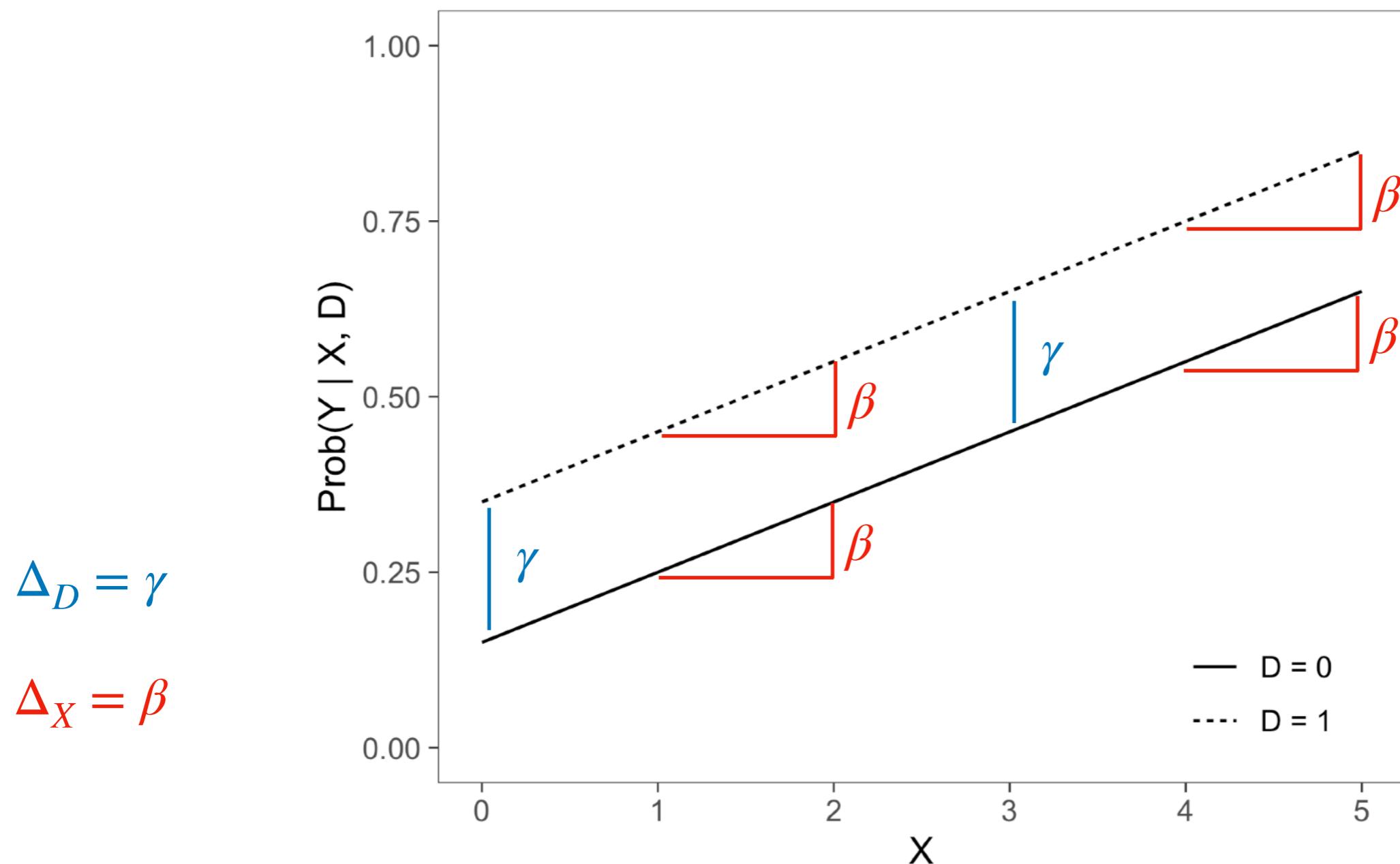
| | LPM (差) | Logit (比) |
|---------------------|---------|-----------|
| 確率への効果の非線形性 | 考慮しない | 考慮する |
| 異なるサンプル間の係数比較 | できる | できない |
| 異なる独立変数を含むモデル間の係数比較 | できる | できない |

Mood, Carina. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It." *European Sociological Review* 26(1):67–82. Table 6より一部抜粋

ロジスティック回帰分析の実質的意味

線形モデルの場合

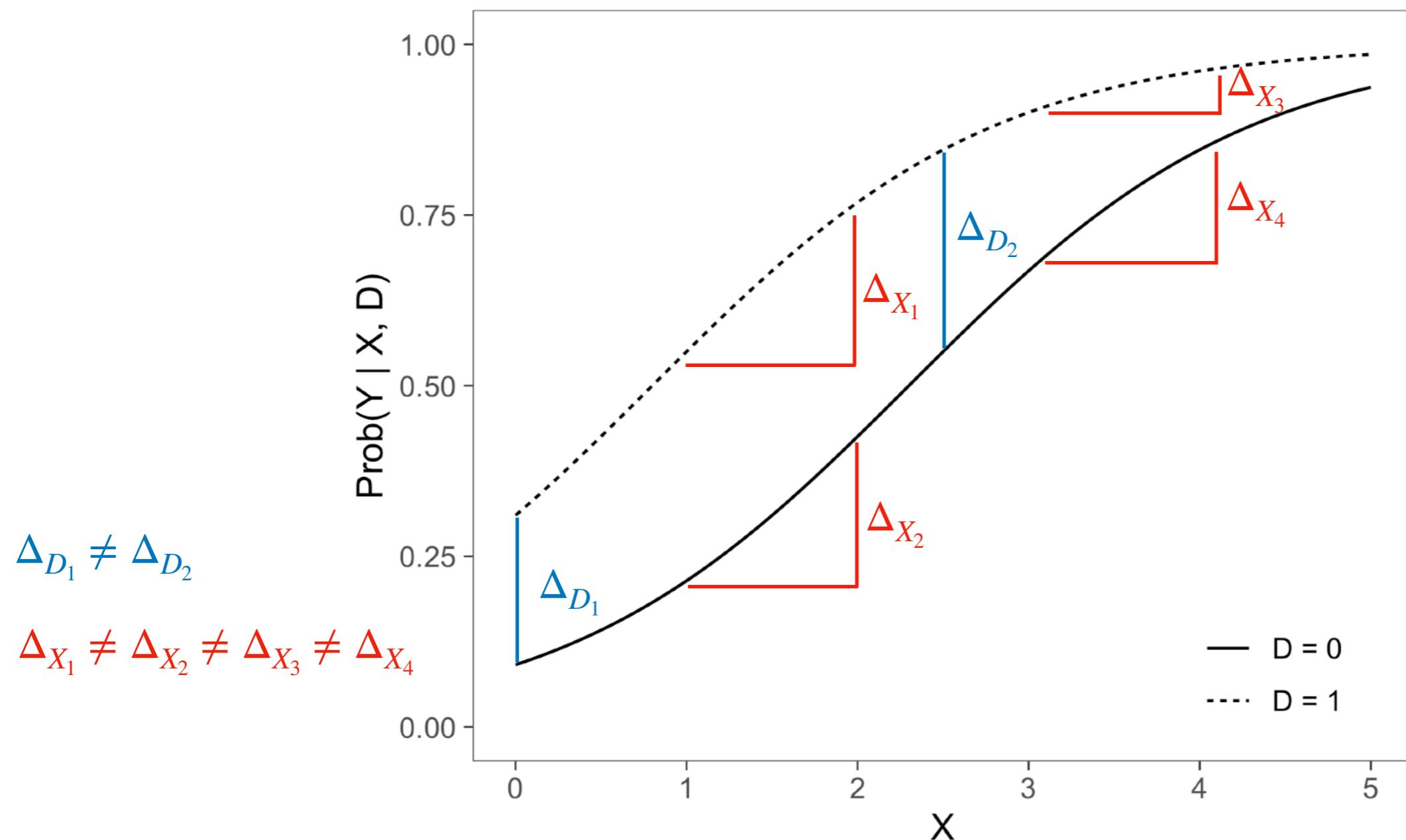
$\Pr(Y) = \alpha + \beta X + \gamma D$ (X は連続変数、 D は2値変数)において、係数は独立変数1単位の変化に対する確率の変化量を表す



非線形モデル（ロジット）の場合

$\Pr(Y) = \frac{\exp(\alpha + \beta X + \gamma D)}{1 + \exp(\alpha + \beta X + \gamma D)}$ において、独立変数1単位の変化に対する確率

の変化量は変化前の値により異なる（係数は確率の差ではなくオッズ比のため）



線形確率モデルとロジットモデルの違い

両者で係数の正負が変わることはほぼ*ないため、まずは線形確率モデル (+口バ
スト標準誤差) を使って分析してもよい

- 線形確率モデル：係数は確率の差を表し、より直感的に解釈しやすい
- ロジットモデル：係数は対数オッズ比を表し、直感的な解釈は難しい

とはいえ、ロジットモデルの結果から「確率の差による解釈」も提示できれば、
結果を解釈したり伝えたりするのに役立つ**

*ただし交互作用項やグループ間比較の場合は係数が反転することがあり得る。この問題についてはBloome, Deirdre, and Shannon Ang. 2022. “Is the Effect Larger in Group A or B? It Depends: Understanding Results from Nonlinear Probability Models.” *Demography* 59(4):1459–88.などに詳しい。

**その他様々な解釈のバリエーションについて簡潔にかつStataコード付きで解説している資料として例えばUberti, L. J. (2022). Interpreting logit models. *The Stata Journal*, 22(1), 60–76.

非線形モデルにおける限界効果

非線形モデルで限界効果を計算する場合には、何らかの基準点を決める必要がある。次の3つの方法がある：

平均値における限界効果 Marginal effect at the mean, MEM :

独立変数をすべて平均値に固定したうえで、そこから1単位の変化を見る

代表値における限界効果 Marginal effect at representative values, MER

独立変数を何らかの関心にもとづく特定の値に固定して、1単位の変化を見る

平均限界効果 Average Marginal Effect, AME

一人ひとりの実際の値ごとに限界効果を計算し、それらの平均をとる

平均値における限界効果 MEM / 代表値における限界効果 MER

平均値における限界効果 MEM

$$MEM = \frac{\Delta \Pr(Y = 1 | X_1 = \bar{X}_1, \dots, X_k = \bar{X}_k)}{\Delta X_k}$$

独立変数をすべて平均値に固定したとき（すべてが平均的な個人において）、 X_k が1単位増加したときに確率がどの程度変化するか

代表値における限界効果 MER

$$MER = \frac{\Delta \Pr(Y = 1 | X_1 = x_1, \dots, X_k = x_k)}{\Delta X_k}$$

ある属性をもつ集団において X_k が1単位増加したときに確率がどの程度変化するか

平均限界効果 AME

平均限界効果 AME

$$AME = \frac{1}{N} \sum_{i=1}^N \frac{\Delta \Pr(Y_i = 1 | X_1 = x_{1i}, \dots, X_k = x_{ki})}{\Delta X_k}$$

X_k が1単位増加したときに確率がどの程度変化するかを、すべての個人について平均した値

MEMと異なり、実在の個人の値を計算しているという利点がある（例：離散変数が独立変数に含まれているとき、その平均をとった個人—たとえば0.6だけ男性な人—というのは論理的に存在しない）

平均限界効果の計算の概略

$\log \frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} = -0.5 + 0.3X + 0.8D$ という推定結果が得られたとする。

この推定結果をもとに、各個人について $D = 1$ のときの予測確率 (1) と $D = 0$ のときの予測確率 (2) を計算し、両者の差 (1) – (2) をとる。

| id | X | D | (1) | (2) | (1) – (2) |
|----|-----|---|-------------------------|-------------------------|---------------------------|
| | | | $\Pr(Y = 1 X, D = 1)$ | $\Pr(Y = 1 X, D = 0)$ | $\Delta\Pr(Y = 1 X, D)$ |
| 1 | 2.4 | 1 | 0.735 | 0.555 | 0.180 |
| 2 | 3.1 | 1 | 0.774 | 0.606 | 0.168 |
| 3 | 1.5 | 1 | 0.679 | 0.488 | 0.192 |
| 4 | 0.5 | 0 | 0.611 | 0.413 | 0.197 |
| 5 | 4.3 | 0 | 0.831 | 0.688 | 0.143 |
| 6 | 2.2 | 0 | 0.723 | 0.540 | 0.183 |

AMEは、(1) – (2)の平均値 **0.177**。

3種類の限界効果の比較

MEM, MER, AMEの3種類の限界効果をそれぞれ計算し、結果を比較してみよう

(5.3.1)

限界効果はどれを使うのがよい？

- 平均的な限界効果を知りたいときは**AME**を使う
- 特定の集団における限界効果を知りたいときは**MER**を使う

平均限界効果と予測確率

- . margins, dydx(gender) 平均限界効果を表示する

```
Average marginal effects                               Number of obs     =      2,805  
Model VCE    : OIM  
  
Expression   : Pr(ojt), predict()  
dy/dx w.r.t. : 1.gender
```

| | Delta-method | | | | | |
|--------|--------------|-----------|------|-------|----------------------|----------|
| | dy/dx | Std. Err. | z | P> z | [95% Conf. Interval] | |
| gender | | | | | | |
| Men | .0555924 | .0185711 | 2.99 | 0.003 | .0191936 | .0919911 |

$$\Pr(Y = 1 | X, D = 1) - \Pr(Y = 1 | X, D = 0)$$

- . margins gender 予測確率 (2つ前のページの(1)と(2)にあたる) を表示する

```
Predictive margins                               Number of obs     =      2,805  
Model VCE    : OIM  
  
Expression   : Pr(ojt), predict()
```

| | Delta-method | | | | | |
|--------|--------------|-----------|-------|-------|----------------------|----------|
| | Margin | Std. Err. | z | P> z | [95% Conf. Interval] | |
| gender | | | | | | |
| Men | .391041 | .0125754 | 31.10 | 0.000 | .3663936 | .4156884 |
| Women | .3354486 | .0130598 | 25.69 | 0.000 | .3098519 | .3610452 |

$$\leftarrow \Pr(Y = 1 | X, D = 1)$$
$$\leftarrow \Pr(Y = 1 | X, D = 0)$$

線形確率モデルとロジットモデルの平均限界効果の比較

線形確率モデルとロジットモデルの平均限界効果を比較しよう (5.3.3)

| | LPM | Logit - AME | | |
|-------------|-----------------|-------------|-----------------|---------|
| 1.gender | 0.054** | (0.018) | 0.056** | (0.019) |
| 2.gender | 0.000 | (.) | 0.000 | (.) |
| 1.educ | 0.000 | (.) | 0.000 | (.) |
| 2.educ | 0.052 | (0.030) | 0.057 | (0.031) |
| 3.educ | 0.154*** | (0.033) | 0.160*** | (0.034) |
| 4.educ | 0.278*** | (0.032) | 0.280*** | (0.033) |
| age | 0.018* | (0.007) | -0.002* | (0.001) |
| c.age#c.age | -0.000** | (0.000) | | |
| _cons | -0.127 | (0.151) | | |
| N | 2805 | | 2805 | |

Standard errors in parentheses

* p<0.05, ** p<0.01, *** p<0.001

2乗項の平均限界効果が表示されない理由

$$\log \frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Age}^2$$

β_2 をそのまま解釈する場合：Ageを一定としたうえでAge²が1単位増加したときの対数オッズの増分

個人のレベルでは「Ageを一定としたうえでAge²が1単位増加する」ことは定義上起こり得ない。そのため、Age²の限界効果を単独で求めることはできない

marginsコマンドを使ってさまざまな年齢における予測値を求めてその実質的な意味を掴むのがよい

調整効果

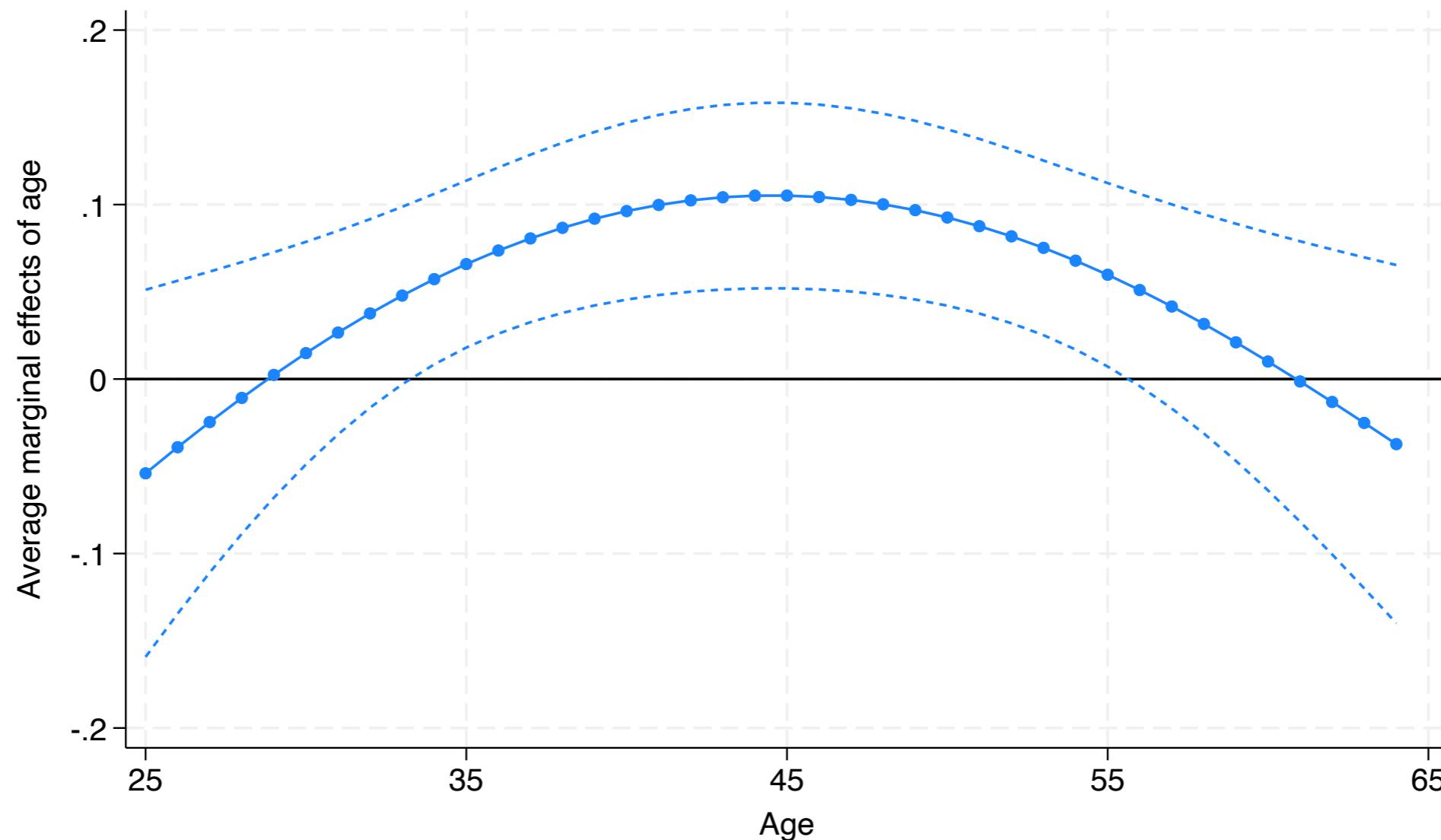
ロジットモデルでも線形回帰分析のときと同様に調整効果（交互作用効果）を考えることができる

$$\log \frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ$$

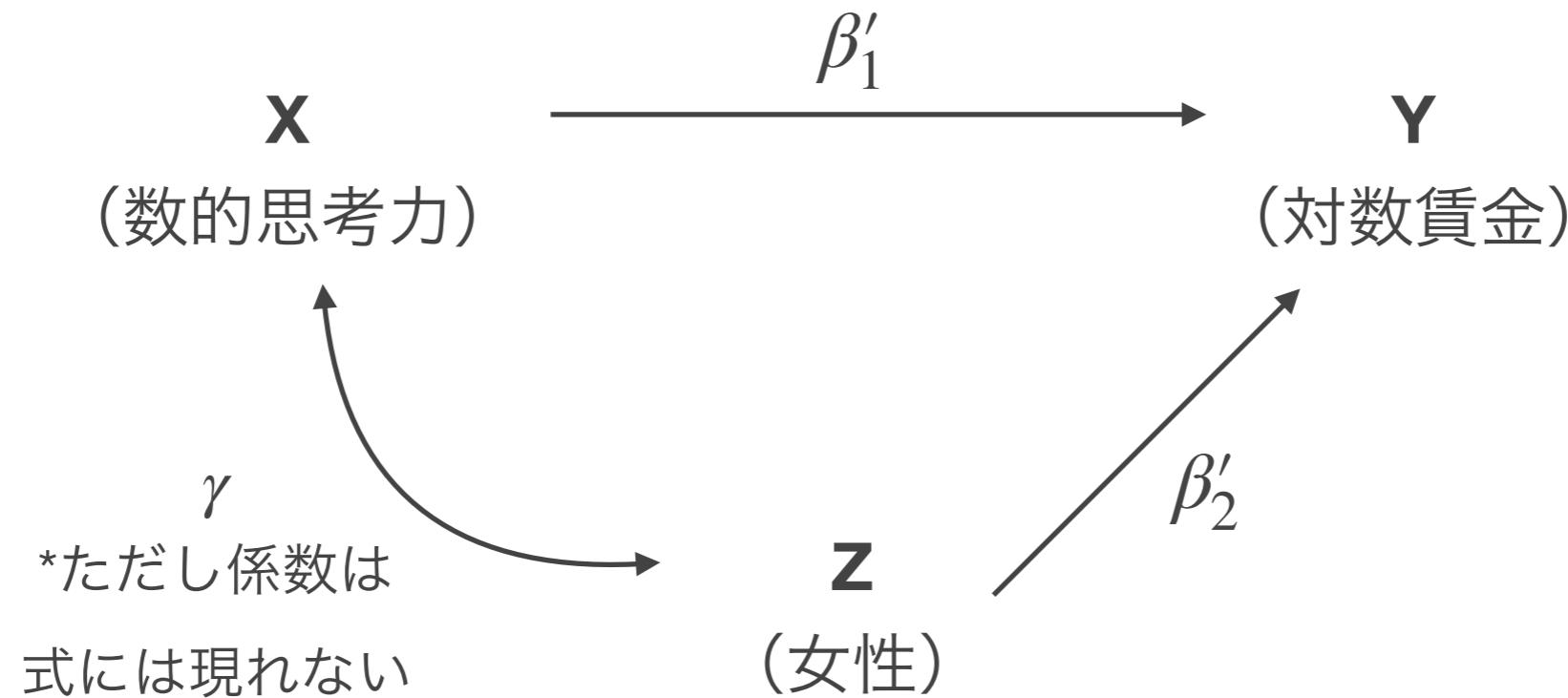
この場合も、対数オッズ比だけでなく、確率がどの程度異なるかを平均限界効果を用いてチェックするとよい *ただし注意すべき点あり...後述の「ロジットモデルにおけるサンプル間比較の問題」スライドを参照

限界効果のプロット：性別の効果は年齢によって異なるか？

性別、年齢、年齢^{2乗}、学歴、性別×年齢、性別×年齢^{2乗}を独立変数とするロジットモデルを推定し、限界効果を図示しよう（5.3.4）



再掲：重回帰分析の推定結果と統制前係数のバイアス



XとZの相関 ZとYの相関 Z統制前の係数と統制後のXの係数の大小

$\gamma > 0$ $\beta'_2 > 0$ $\beta_1 > \beta'_1$ —— 統制しないと過大推計

$\gamma < 0$ $\beta'_2 < 0$ $\beta_1 > \beta'_1$ —— 統制しないと過大推計

$\gamma < 0$ $\beta'_2 > 0$ $\beta_1 < \beta'_1$ —— 統制しないと過小推計

$\gamma > 0$ $\beta'_2 < 0$ $\beta_1 < \beta'_1$ —— 統制しないと過小推計

ロジットモデルの注意点と対策

$$\log[\Pr(Y = 1)/(1 - \Pr(Y = 1))] = \beta_0 + \beta_1 X$$

$$\log[\Pr(Y = 1)/(1 - \Pr(Y = 1))] = \beta'_0 + \beta'_1 X + \beta'_2 Z$$

のように、異なる独立変数を含む2つのモデルの係数を比較し、統制前の変数の変化をもって過大推計／過小推計や媒介要因の寄与を判断することはできない。このような場合には、以下の方法がある：

- それぞれのモデルについてAMEを計算して比較する (Mize et al. 2019)
- 対数オッズ比の変化について関心がある場合は、Karlson-Holm-Breenの要因分解法(など)を使う (Kohler et al., 2011; Hagenaars et al., 2024)

Mize, Trenton D., Long Doan, and J. Scott Long. 2019. "A General Framework for Comparing Predictions and Marginal Effects across Models." *Sociological Methodology* 49(1):152–89.

Kohler, Ulrich, Kristian Bernt Karlson, and Anders Holm. 2011. "Comparing Coefficients of Nested Nonlinear Probability Models." *The Stata Journal* 11(3):420–38.

Hagenaars, Jacques A. P., Steffen Kühnel, and Hans-Jürgen Andreß. 2024. *Interpreting and Comparing Effects in Logistic, Probit, and Logit Regression*. London, England: SAGE Publications.

独立変数の個数

ロジスティック回帰分析においてモデルに含めることのできる独立変数の個数の目安は従属変数0と1のうち少ないほうのケース数を10で割った値とされており、それを超えると推定結果が不安定になるとされている。

今回のOJTは0: 1781ケース, 1: 1024ケースなので、102個

Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstein. 1996. "A Simulation Study of the Number of Events per Variable in Logistic Regression Analysis." *Journal of Clinical Epidemiology* 49(12):1373–79.

完全予測の問題

| | | Y | |
|---|---|--------------|----------|
| | | X | 1 0 |
| X | 1 | 150 300 | 450 |
| | 2 | 0 400 | 400 |

完全予測をしてしまう独立変数がある場合には、当該カテゴリに該当するケースは分析から自動的に除外される

完全予測に近い変数（度数が1のセルなど）が複数含まれている場合には計算が収束しなかったり、係数が極端に大きくなる。ロバスト標準誤差を使っている場合には非常に強く統計的に有意になったりする

カテゴリ変数を独立変数として用いる場合にはクロス表などでこのような変数がないかを確認し、あればカテゴリの統合などを考える

ロジットモデルにおける決定係数

Stataの出力におけるPseudo R²は疑似決定係数と呼ばれる指標であり、以下のように定

$$\text{義される: Pseudo } R^2 = 1 - \frac{\log L(M_{full})}{\log L(M_{intercept})}$$

疑似決定係数にはいろいろな種類があり、まれにCox & Snell's R²やNagelkerke's R²といった指標が使われることもある

```
ssc install fitstat // install package
```

```
logit y x
```

```
fitstat
```

もちろん、決定係数の大小を気にすることにあまり意味はない

ロジットモデルにおけるサンプル間比較の問題

2つの異なるサンプルAとBで同じ独立変数からなるロジットモデルを推定すると、Xの係数はAのほうがBより大きいにもかかわらず、Xの平均限界効果を計算したときにはむしろ、AよりもBのほうがXの係数が大きいという逆転現象が起こることがある。交互作用項を含む場合にも、同じ問題が起こり得る

サンプル間比較を行う場合には基本的にはまず平均限界効果を使うことが推奨されている (Mize et al., 2019; Bloome & Ang, 2022)

一方、オッズ比も依然として重要であるとする派閥 (?) もある (Kuha & Mills, 2020; Hagenaars et al., 2024)

Mize, Trenton D., Long Doan, and J. Scott Long. 2019. "A General Framework for Comparing Predictions and Marginal Effects across Models." *Sociological Methodology* 49(1):152–89.

Bloome, Deirdre, and Shannon Ang. 2022. "Is the Effect Larger in Group A or B? It Depends: Understanding Results from Nonlinear Probability Models." *Demography* 59(4):1459–88.

Kuha, Jouni, and Colin Mills. 2020. "On Group Comparisons with Logistic Regression Models." *Sociological Methods & Research* 49(2):498–525.

Hagenaars, Jacques A. P., Steffen Kühnel, and Hans-Jürgen Andreß. 2024. *Interpreting and Comparing Effects in Logistic, Probit, and Logit Regression*. London, England: SAGE Publications.

補足資料

プロジェクト管理・作図

プロジェクト管理

Long, Scott J. 2009. *The Workflow of Data Analysis Using Stata*. Stata Press.

作図ほか

Mitchell, Michael N. 2021. *A Visual Guide to Stata Graphics, Fourth Edition*. Stata Press.

Mitchell, Michael N. 2021. *Interpreting and Visualizing Regression Models Using Stata, Second Edition*. Stata Press.

Visual overview for creating graphs. <https://www.stata.com/support/faqs/graphics/gph/stata-graphs/>

Stata Visual Library <https://worldbank.github.io/stata-visual-library/index.html>

Stata Cheat Sheets <https://www.stata.com/bookstore/stata-cheat-sheets/>

Stataでの回帰分析・ロジスティック回帰分析

線形回帰分析の基礎

Gordon, Rachel A. 2015. *Regression Analysis for the Social Sciences, Second Edition*. Routledge.

松浦寿幸, 2021, 『Stataによるデータ分析入門：経済分析の基礎から因果推論まで』 東京図書.

ロジスティック回帰分析

Long, Scott J. and Jeremy Freese. 2014. *Regression Models for Categorical Dependent Variables Using Stata, Third Edition*. Stata Press.

全般 (2冊合計で1700頁くらいありますが、回帰分析関連の方法はほぼすべて載っています)

Cameron, A. Colin and Pravin K. Trivedi. 2022. *Microeconometrics Using Stata, Second Edition*. Stata Press.

回帰分析の使い方

吉田寿夫・村井潤一郎, 2021, 「心理学的研究における重回帰分析の適用に関する諸問題」『心理学研究』92(3): 178–87.

Elwert, Felix, and Christopher Winship. 2014. “Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable.” *Annual Review of Sociology* 40:31–53.

Keele, Luke, Randolph T. Stevenson, and Felix Elwert. 2020. “The Causal Interpretation of Estimated Associations in Regression Models.” *Political Science Research and Methods* 8:1–13.

因果推論

松林哲也, 2021, 『政治学と因果推論：比較から見える政治と社会』 岩波書店.

安井翔太・株式会社ホクソエム, 2019, 『効果検証入門：正しい比較のための因果推論／計量経済学の基礎』 技術評論社.

Huntington-Klein, Nick. 2021. *The Effect: An Introduction to Research Design and Causality*. <https://theeffectbook.net/>

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press.
<https://mixtape.scunning.com/>

Morgan, Stephan and Christopher Winship. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research, 2nd Edition*. Cambridge University Press. (落海浩訳, 2024, 『反事実と因果推論』 朝倉書店.)

do-file editorに代わるテキストエディタ

Sublime text 3 <https://www.sublimetext.com/3>

日本語の解説：「Stata用のIDE（統合開発環境）もどきを導入してみた」<https://ryukius-hitties.hatenablog.com/entry/2019/04/21/180008>

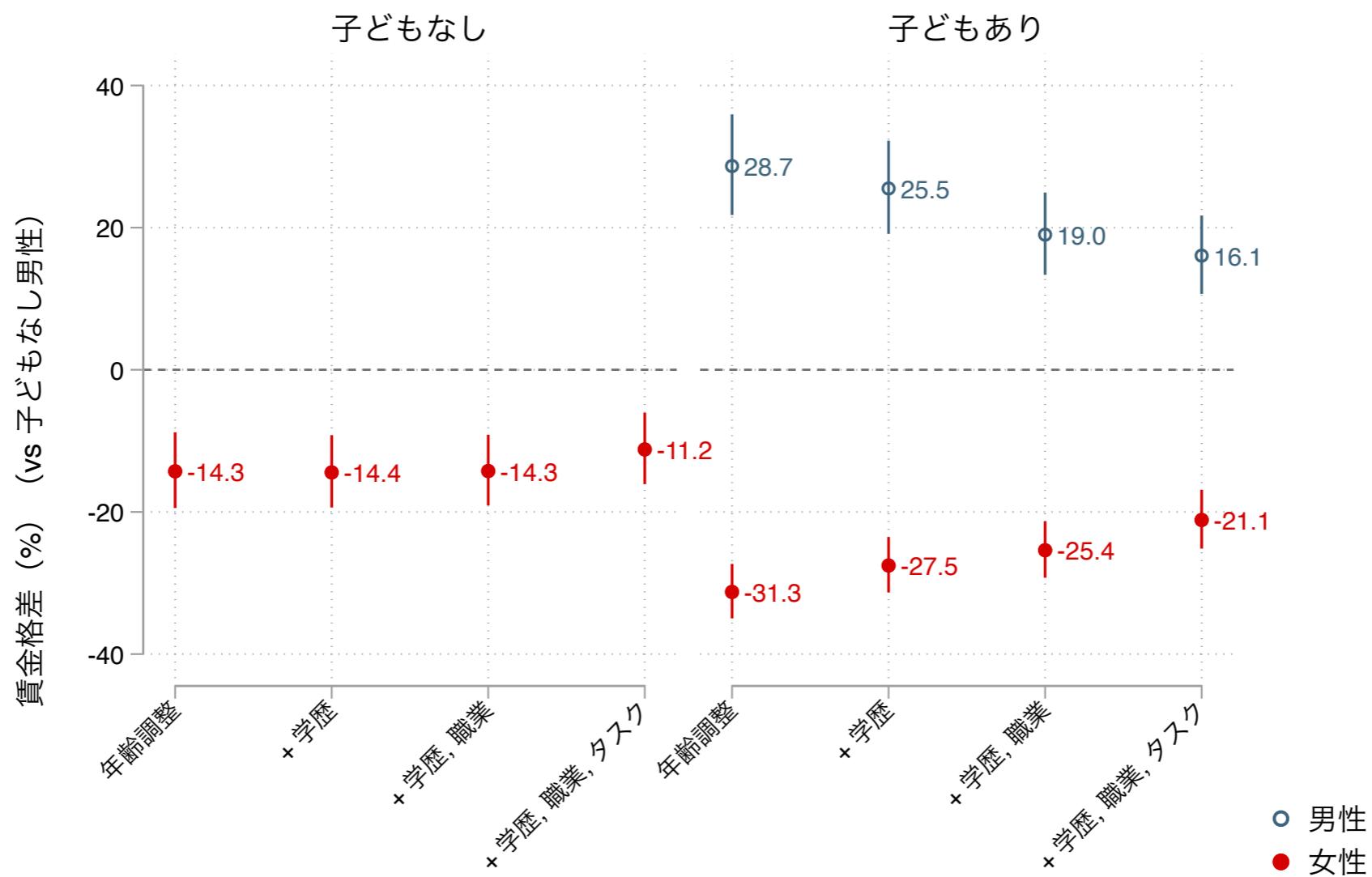
Atom <https://atom.io/>

解説：language-stata <https://atom.io/packages/language-stata>

Hydrogenを使う方法とstata-execを使う方法が紹介されているが、パソコンに強くない人はstata-execのほうがたぶん簡単。Windowsだとできないかも？

分析の実例

6_advanced2024-09-04.doを実行して、麦山（2022）のp.37–39の図表を再現してみよう（6.1–6.5）



麦山亮太, 2022, 「職業とタスクからみる仕事と賃金のジェンダー格差」財務総合政策研究所『「仕事・働き方・賃金に関する研究会：一人ひとりが能力を発揮できる社会の実現に向けて」報告書』20–41. https://www.mof.go.jp/pri/research/conference/fy2021/shigoto_report.html

Excelファイルを開いて結合する（パターン1）

同じ構造の複数のsheetからなるExcelファイルを順次読み込み、1つのデータに結合したい場合の手順

1. sheetを指定して読み込む
2. そのsheetを表す変数を作成して、保存する
3. 次のsheetについても同じように読み込み、変数を作成、保存
4. 2つのデータをappendで結合

7_loop_macro2024-09-04.doを開き、Excelファイルを読み込んでみよう（5.1）

Excelファイルを開いて結合する（パターン2）

同じ構造の複数のファイルからなるExcelファイルを順次読み込み、1つのデータに結合したい場合の手順

1. sheetを指定して読み込む
2. そのsheetを表す変数を作成して、保存する
3. 次のsheetについても同じように読み込み、変数を作成、保存
4. 以上の手順をforvaluesを用いて繰り返し
5. 作成したデータをappendで結合

複数のExcelファイルを繰り返し読み込んでみよう（7.2）

繰り返し処理：forvalues/foreach

forvaluesは変化する数値に対して繰り返し処理を実行する

```
forvalues i = ... {  
}  
}
```

foreachは変化する文字列に対して繰り返し処理を実行する

```
foreach w in ...{  
}  
}
```

繰り返し中身が変わる部分（上記の例では*i*や*w*）については`'で囲む

繰り返し呼び出し：local/global macro

localマクロは一時的にのみ呼び出せる、glocalマクロは一度実行するとStataを開いている限りは永続的に何度も呼び出すことができるという点で異なる。

localマクロは`'で囲み、globalマクロは\$をつける（または\${}で囲む）

使い分けた

- 一時的に呼び出す（それ以降使わない）場合にはlocalマクロとして定義する
- 一連のコード内で何度も繰り返し呼び出す場合はglobalマクロとして定義する。またglobalマクロはmasterファイル内で定義するほうが安全（どこかに書いたglobalが他のdo-file内の命令に影響する可能性があるため）

Rにもチャレンジしてみる



The screenshot shows a website for a guide to analyzing survey data using R. The header includes the title "Rによる社会調査データ分析の手引き" and navigation icons (menu, search, A, i, social media). The main content area displays the title again, author information ("麦山 亮太 (学習院大学法学部政治学科) / Ryota Mugiyama (Department of Political Studies, Gakushuin University)"), and a last update date ("Last update: 2022-02-28"). On the left sidebar, there is a table of contents for "まえがき" (Foreword) and "1 研究計画を立てる" (1. Planning the Research), which includes sections 1.1 through 1.5.

Rによる社会調査データ分析の手引き

麦山 亮太 (学習院大学法学部政治学科) / Ryota Mugiyama (Department of Political Studies, Gakushuin University)

Last update: 2022-02-28

| まえがき |
|----------------|
| 1 研究計画を立てる |
| 1.1 研究計画とは |
| 1.2 研究背景 |
| 1.3 研究目的・問い合わせ |
| 1.4 方法 |
| 1.5 参考文献 |

学部生／院生向けゼミで使っている資料をウェブページで公開しています

このセミナーで扱っている内容と同程度の内容を扱っています