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

2023/09/05 2023年度CSRDA計量分析セミナー・夏@Zoom

自己紹介

現所属

2021/04– 学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

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

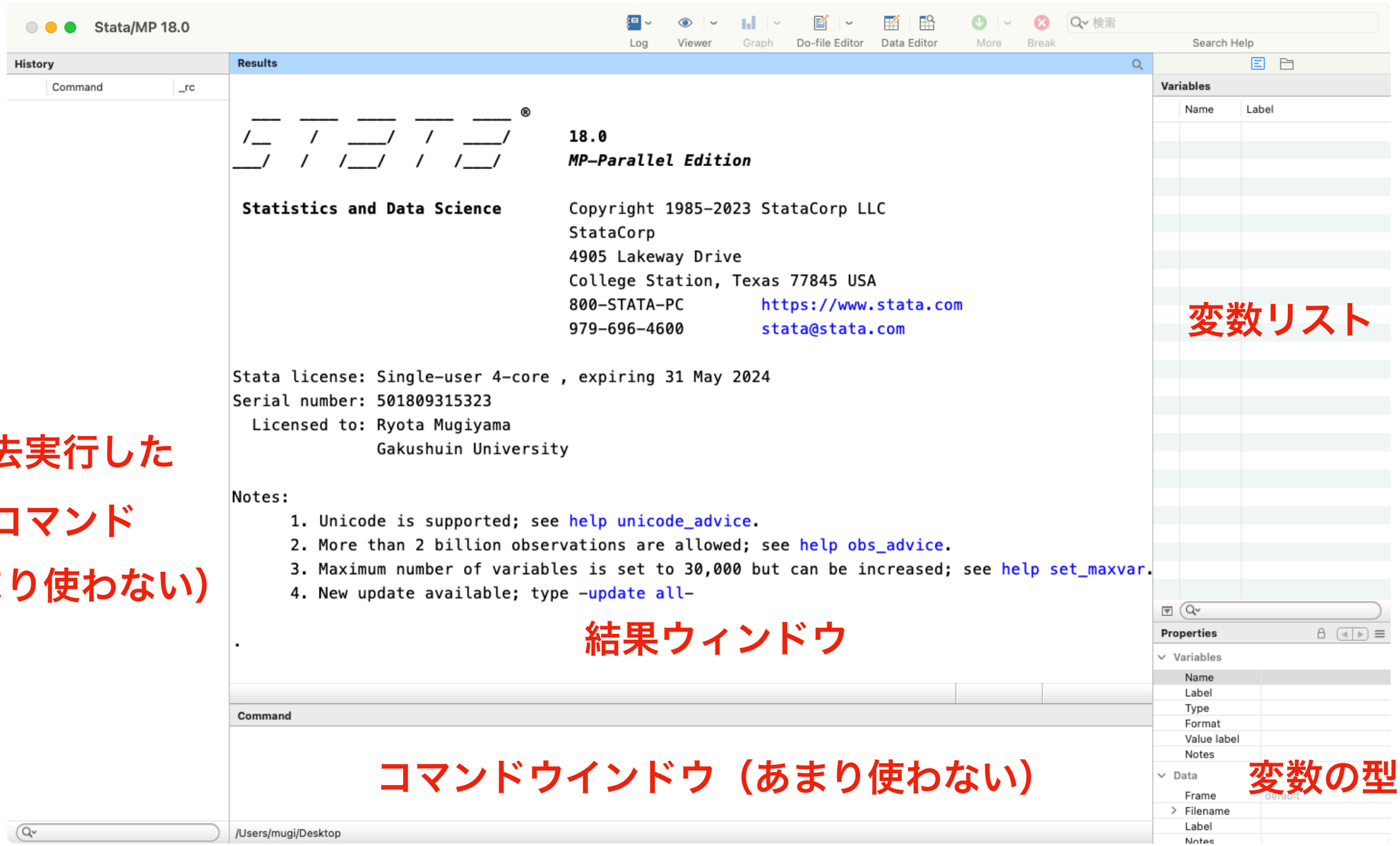
- 無意味な作業の時間を減らすことで、本質的なことを考える時間を取れる

「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



過去実行したコマンド（あまり使わない）

結果ウィンドウ

コマンドウインドウ（あまり使わない）

変数リスト

変数の型など

設定の変更

各種設定変更

EditまたはStata/MP 18.0* → Preferences →

General Preferences

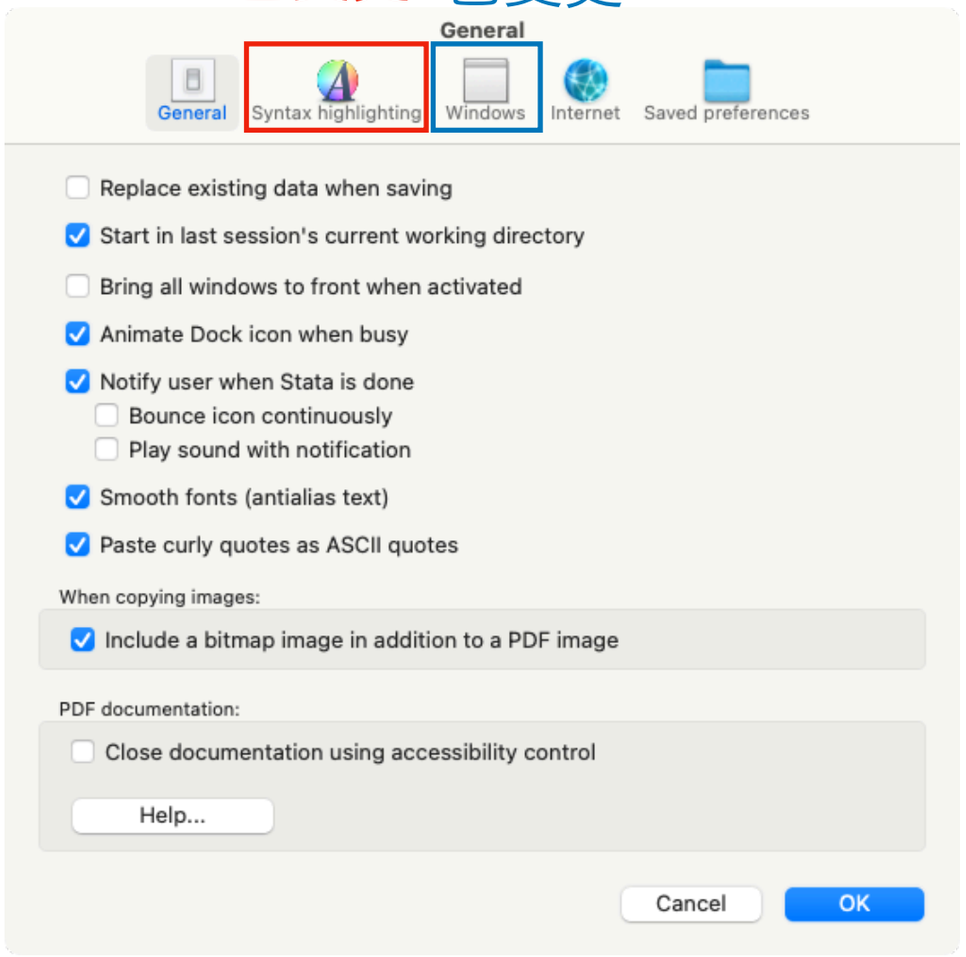
言語変更

EditまたはStata/MP 18.0* → Preferences →

User-interface language

ハイライトの色変更

結果ウインドウ等の色変更



*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

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

繰り返す

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：投稿したときのファイル、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

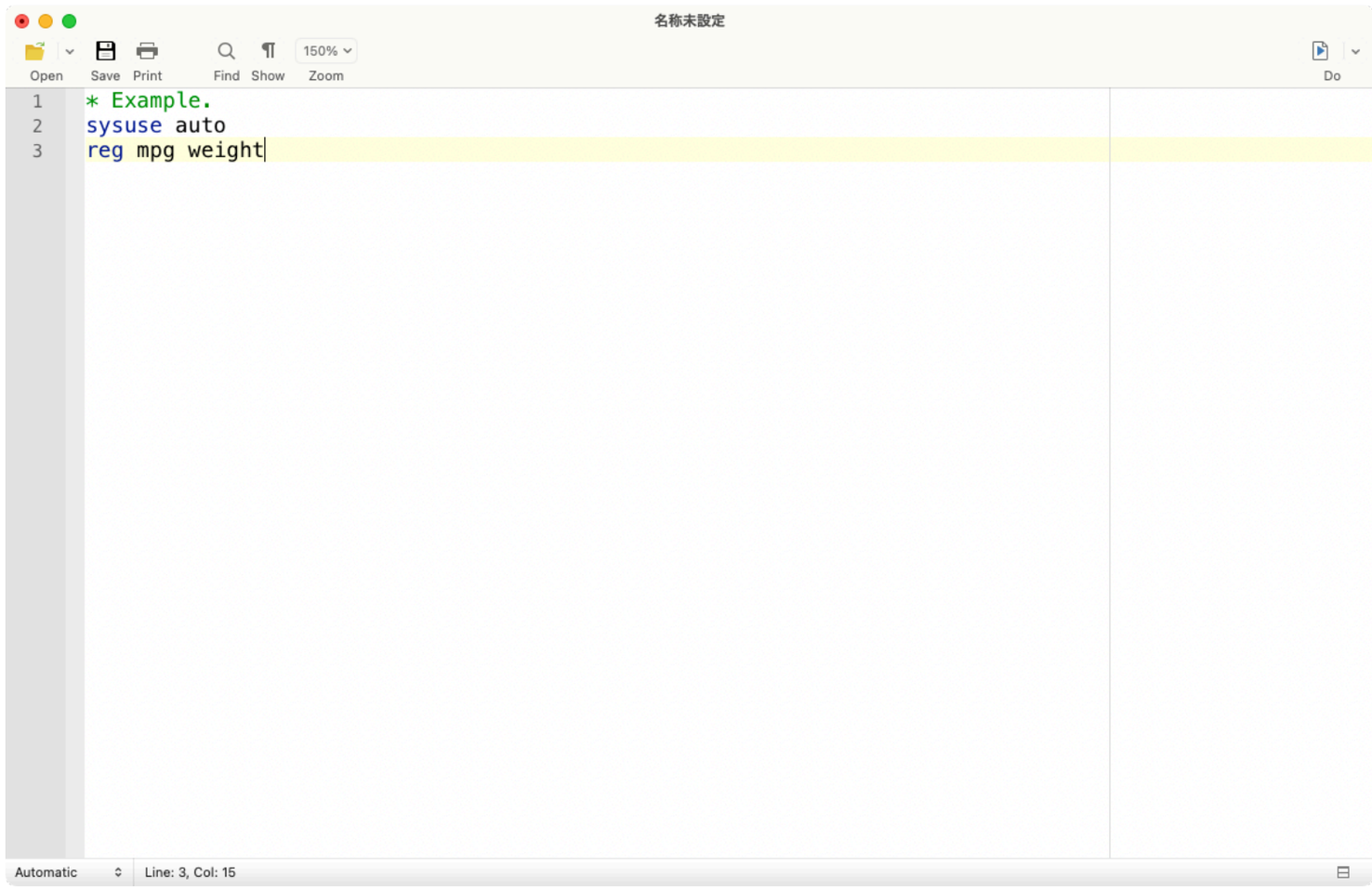
分析をする前に、分析のコードを走らせる場所（=作業ディレクトリ）をPCに教えてあげる。

- 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projectファイルまたはdoファイルを開く

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ファイルと保存の方法

コマンドウインドウに `doedit` と入力して実行（Enter）



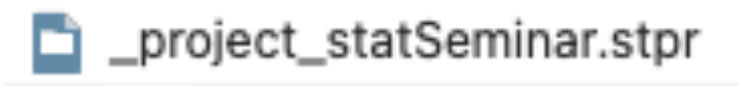
doファイル上部の「Save」をクリック → 名前をつけて保存

保存先は、「code」フォルダとする

もう一つの方法：プロジェクトの「しおり」

File → New → Project...を選択し、先ほどの「code」に当たるフォルダを選択し、任意の名前をつけて保存すると、フォルダに以下のようなファイルができる

（Windowsユーザ：doファイルを開いて、File → New → Project...を選択）



.stprファイルは、プロジェクトの場所（作業ディレクトリ）を記憶しておくための「しおり」だと思えばよい。

このファイルをクリックすることで、作業ディレクトリが変更される。**当該プロジェクトのデータ分析を実行するときには、プロジェクトファイルを起動する**

（Macの場合は上記ファイルをクリック、Windowsの場合はdoファイルを開く → Open → Project...で選択）

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`

`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

0_install2023-09-05.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

https://www.oecd.org/skills/piaac/

データを開き、中身を確認する

1_variables2023-09-05.doを開き、以下のコードを実行しよう：

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

describe

browse
```

作業ディレクトリを指定する、もしくはstprファイルを開いた場合には、それ以降、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。例：

```
use "../../project_usa/data/piaacusa.dta", clear
```

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
```

```
summarize i.ageg
```

データの分析

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ファイルを分ける

分析に関わるdoファイル内では（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_install2023-09-05.do

1_variables2023-09-05.do

2_sample2023-09-05.do

データの加工

3_descriptive2023-09-05.do

4_regression2023-09-05.do

5_logit2023-09-05.do

データの分析

事前にどのような分析をするかをきちんと計画することが有用

doファイル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**のが望ましい（100行目から120行目は飛ばして〜みたいなのはダメ）
- doファイルは何に関する、いつ作成した（編集した）ものなのかがわかるような名前をつけるのがおすすめ（たとえば「2023-09-05statSeminar.do」「handling2023-09-05.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ファイルを作るとよい
- 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ファイルを実行する

_master2023-09-05.doを開いて中身を確認してみよう

master2022-03-02.do

Open Save Print Find Show 158% Zoom Do

```
/*-------------------------
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さまざまに加工して先に立てた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

Altmetric 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.

データ加工のフロー

元々のデータ：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_variable2023-09-05.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_sample2023-09-05.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
- 最近では欠損除外前のサンプルサイズ（何の欠損でどれくらいサンプルサイズが減ったか）と除外後のサンプルサイズを併記するのがnormとなりつつある
- リストワイズ削除をする場合、欠損が完全にランダムに生じている（NCAR）と仮定している。記述統計量にはバイアスが生じるが、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（Stataの仕様は最大32文字）：たとえば学歴ならedではなくeducation、最低でもeducという名前をつけるのがよい。全角文字は一応使えるがおすすめしない
- 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ファイルは（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
…				
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_descriptive2023-09-05.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

summarizeで出力した結果は、そのまま論文に載せるには少し手作業が必要なうえ、せっかくつけたラベルの情報が失われてしまう

このようなときには `desctable` https://www.trentonmize.com/software/desctable が便利

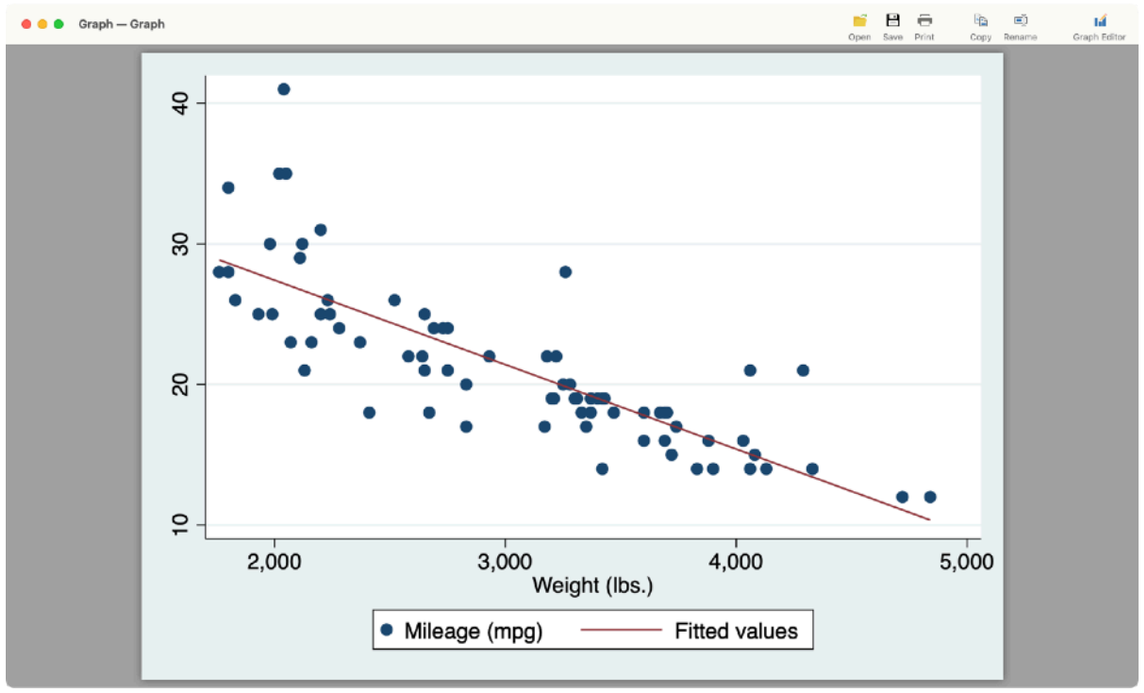
Table #: Descriptive Statistics (N = 2805)

	n	Mean/Prop.	SD	Min.	Max.
Hourly wage	2805	1775.17	1149.64	461.54	9248.55
Age	2805	43.93	10.88	25.00	64.00
Gender	2805	.47			
Level of education	2805				
Junior high		.09			
Senior high		.35			
Junior college		.25			
University		.31			
Numeracy score	2805	2.94	.43	1.03	4.41

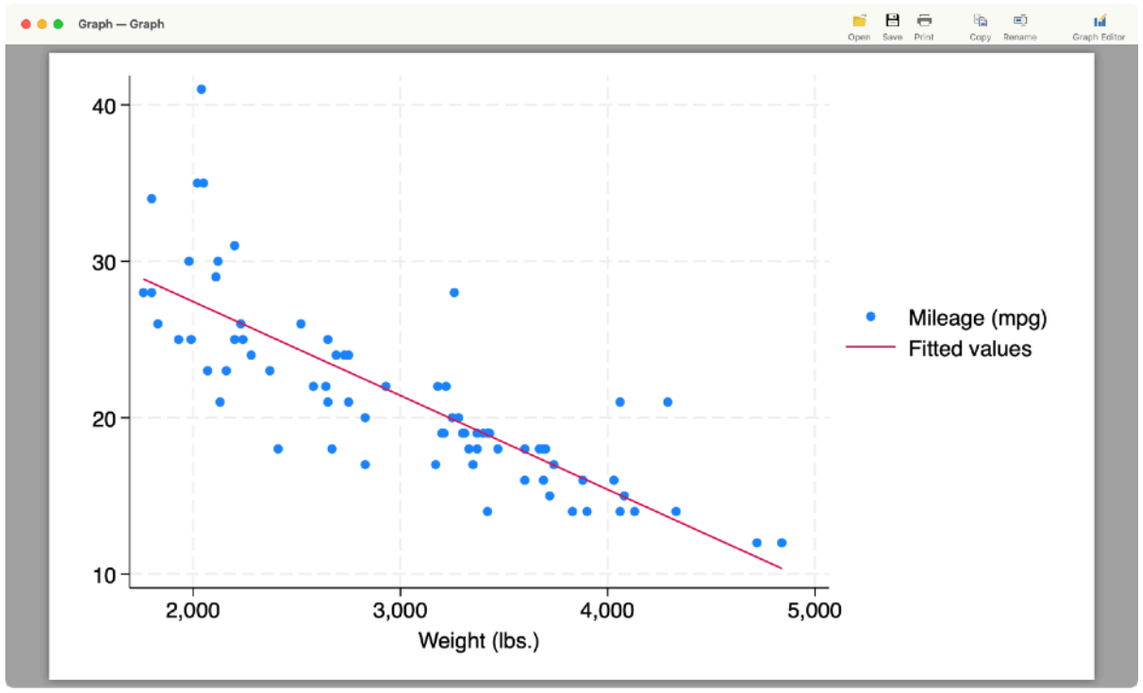
desctableコマンドを使って要約統計量をExcelに書き出してみよう（3.1.2）

Stata 18のグラフデザイン変更

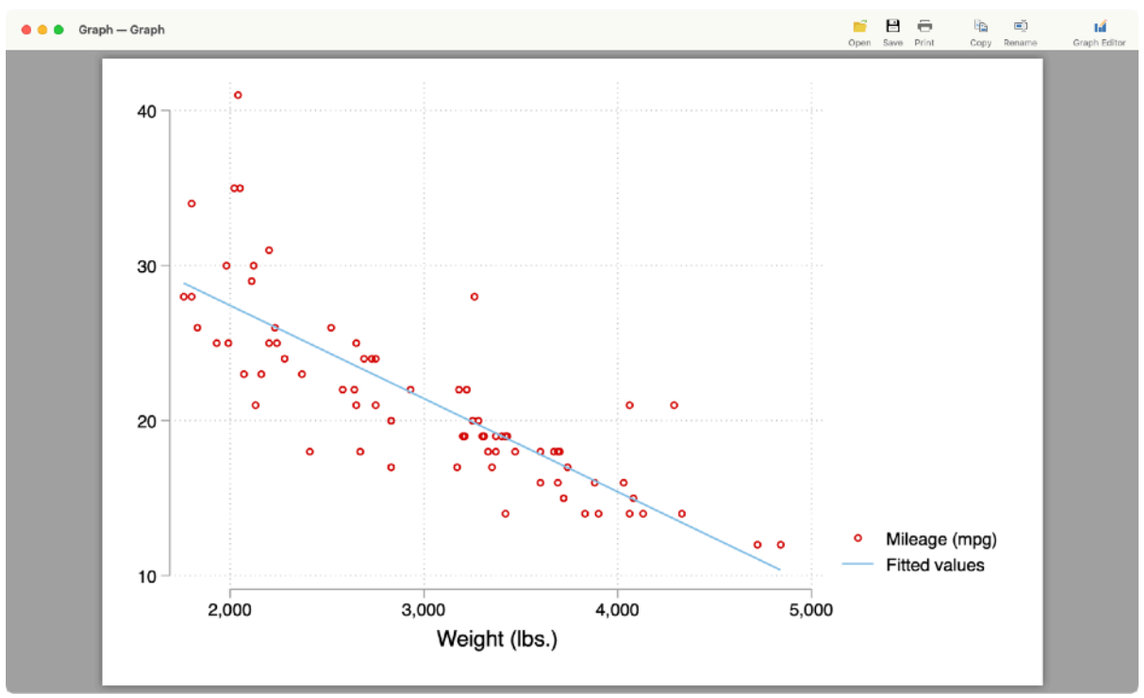
Stata 17まで



Stata 18以降

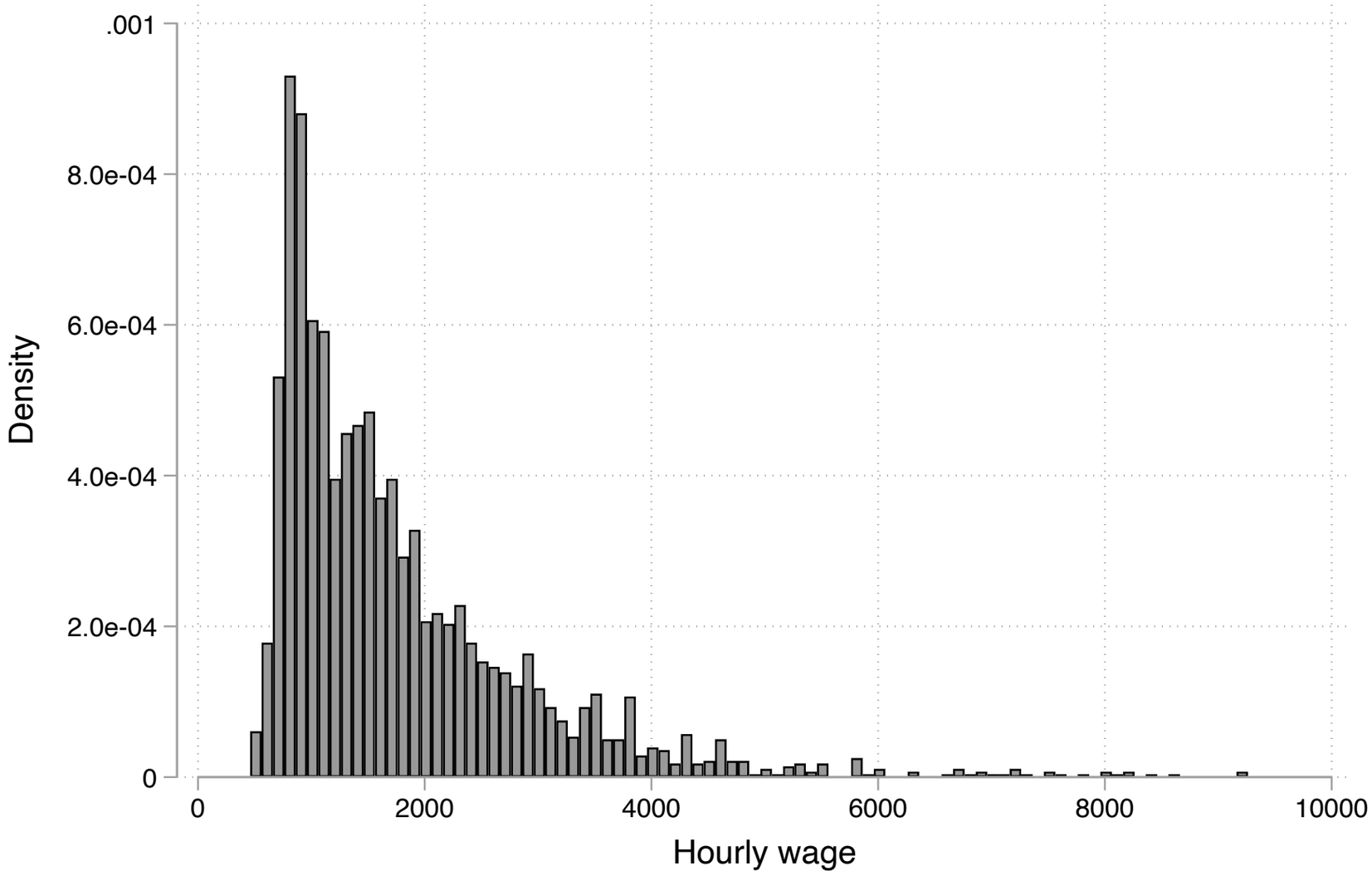


scheme(cleanplots)



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）

2変数関係

変数の分布や、その集計値を異なるグループごとに比べることで、**比較の問い**に答えることができる。

1変数分布	2変数関係
Y?	X ⟶ Y
Yはどのような分布？	Yの分布はXによってどれくらい違う？
平均や中央値はどれくらい？	Yの平均値はXごとにどれくらい違う？

連続変数をグループ間で比較する

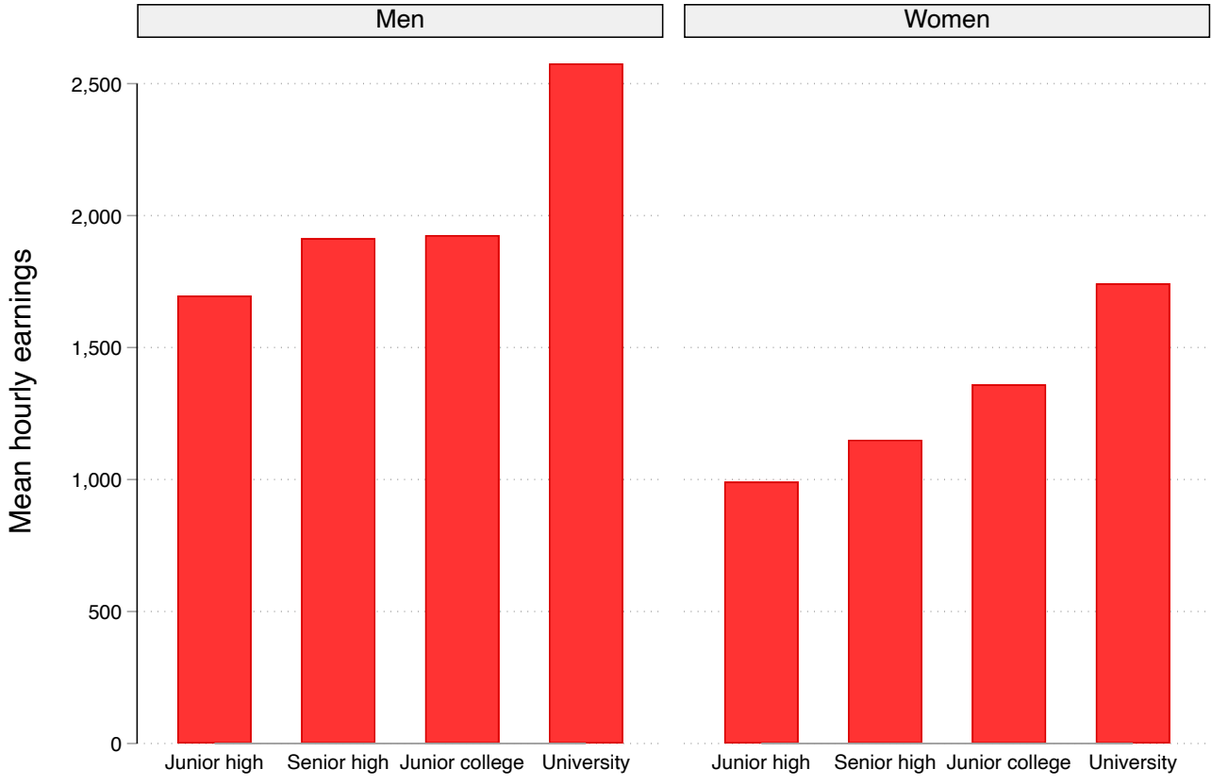
男性と女性で、変数の平均値や標準偏差にはどの程度違いがあるだろうか？

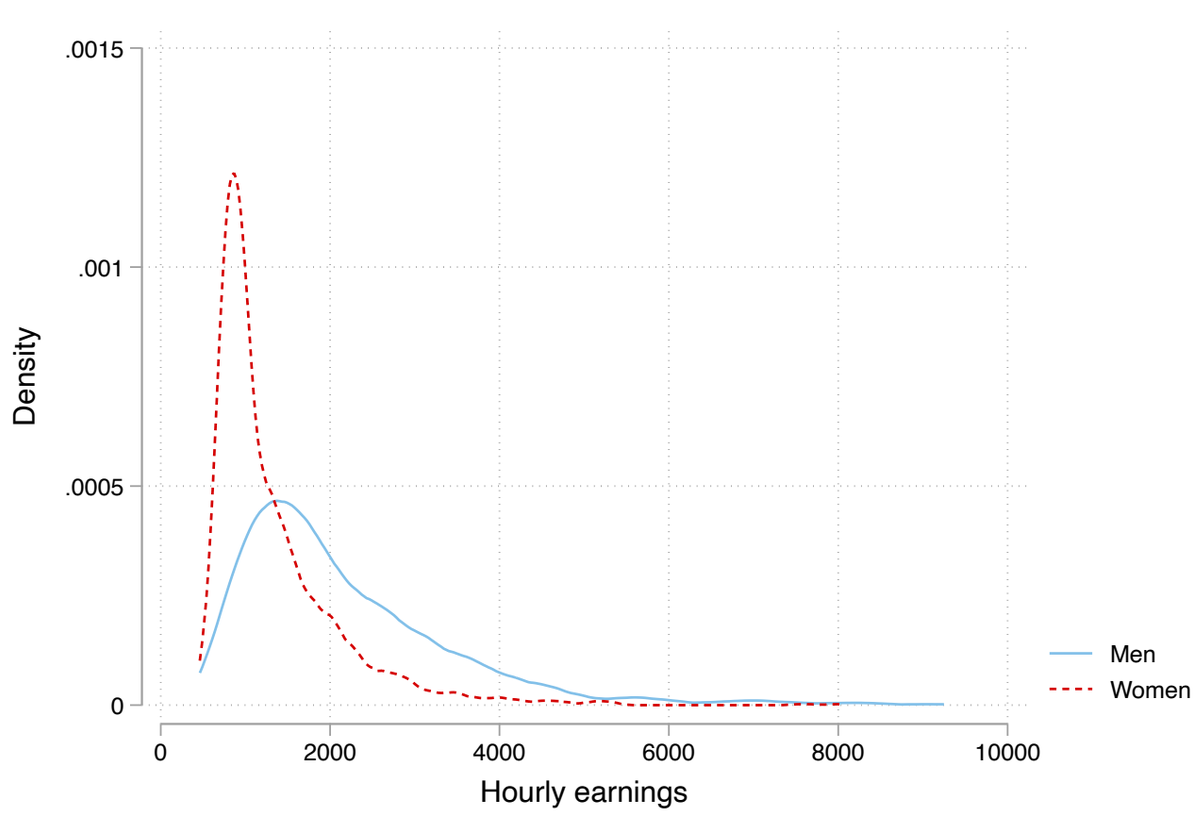
tabstatコマンドを使ってグループ別の集計をしてみよう（3.2.1）

desctableコマンドを使ってグループ別の集計をしてみよう（3.2.2）

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（3.2.3）





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？

カテゴリ変数（Y）の度数およびその分布をグループ（X）別に集計した表のことを指して、クロス集計表という。

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）

クロス集計などの結果をcsvファイルに出力

tabulateで出力した結果は、そのまま論文に載せるには少し手作業が必要

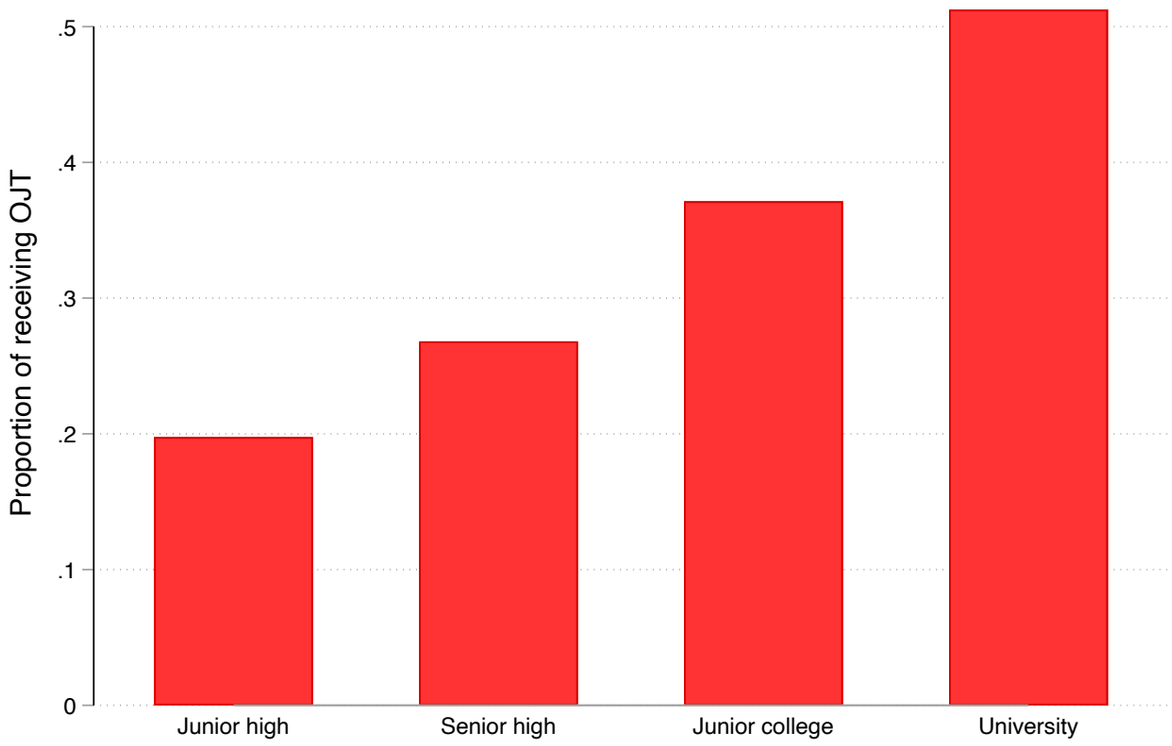
summarize, tabulate, regressなどの出力結果をcsvファイルにエクスポートできる便利なコマンドがesttab http://repec.sowi.unibe.ch/stata/estout/

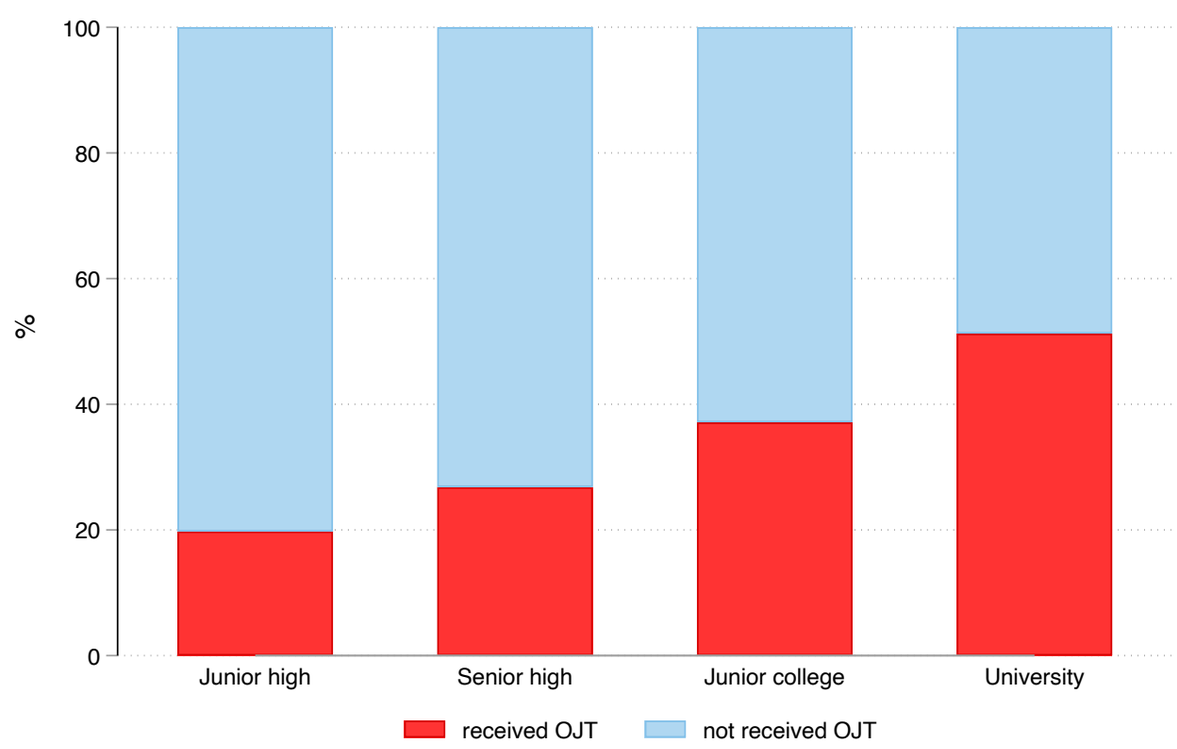
	No	Yes	Total
	b/rowpct	b/rowpct	b/rowpct
Junior high	192	48	240
	80.0	20.0	100.0
Senior high	726	268	994
	73.0	27.0	100.0
Junior college	435	257	692
	62.9	37.1	100.0
University	428	451	879
	48.7	51.3	100.0
Total	1781	1024	2805
	63.5	36.5	100.0

esttabは回帰分析の表を作るときに本領を発揮するが、クロス集計表でも使えないことはない

クロス集計表を図で表す

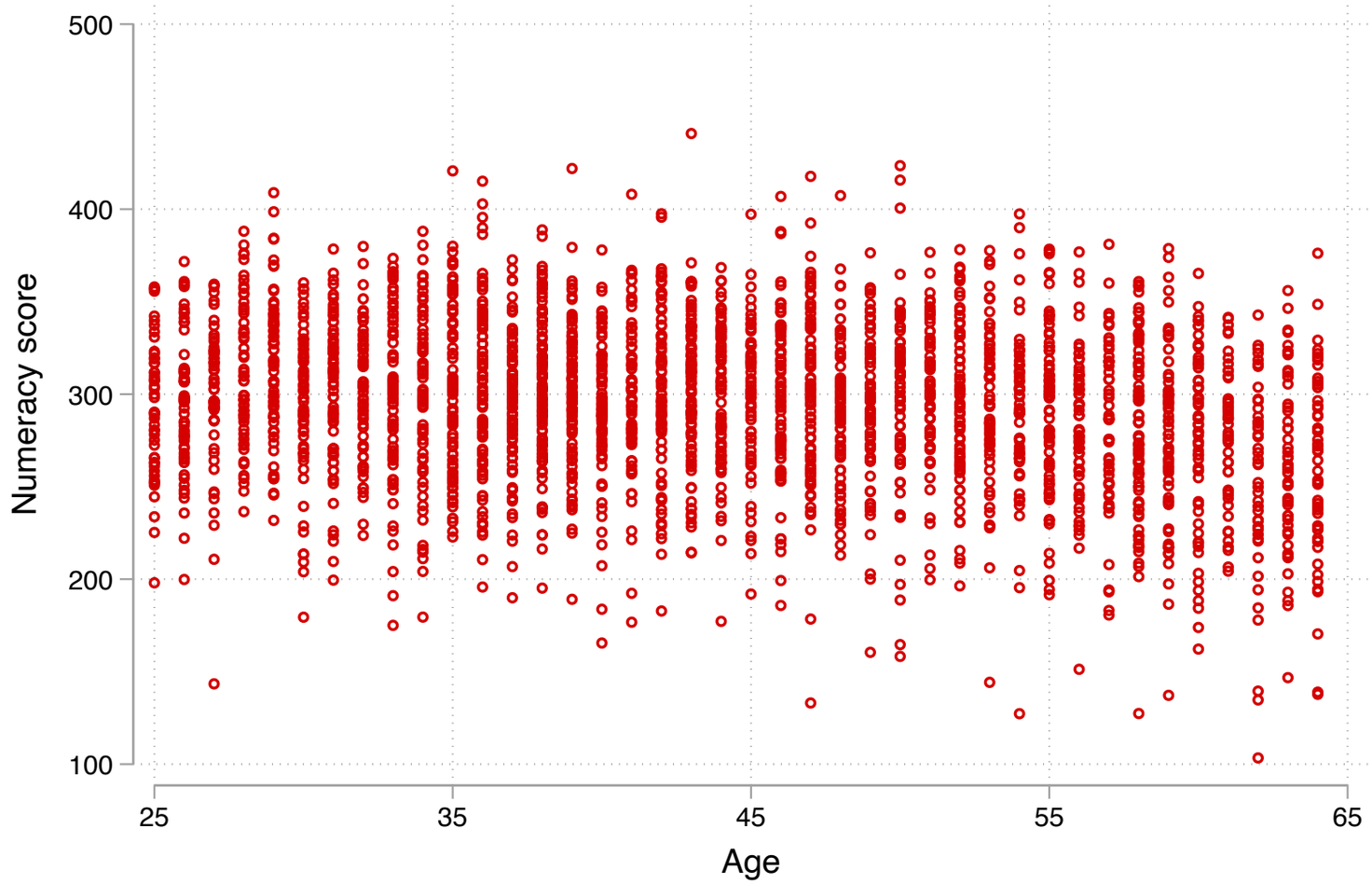
列変数が2値のときのクロス集計表を棒グラフで表してみよう（3.3.2）





散布図と相関係数

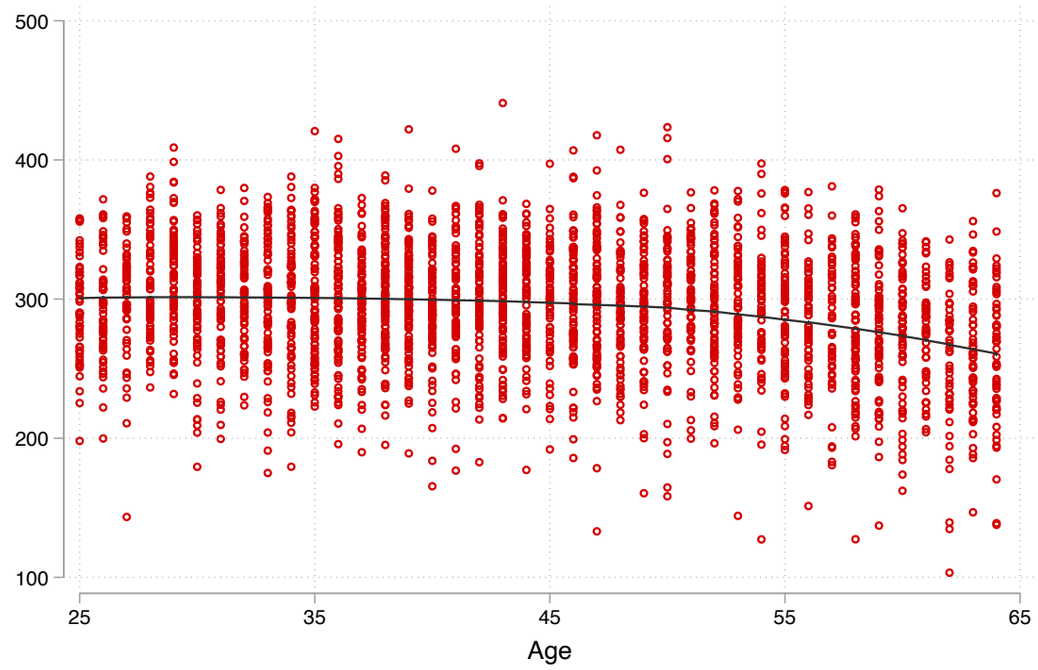
連続変数の値ごとに連続変数の値を比較する場合には、相関係数を計算したり、散布図を作成するのがよい。

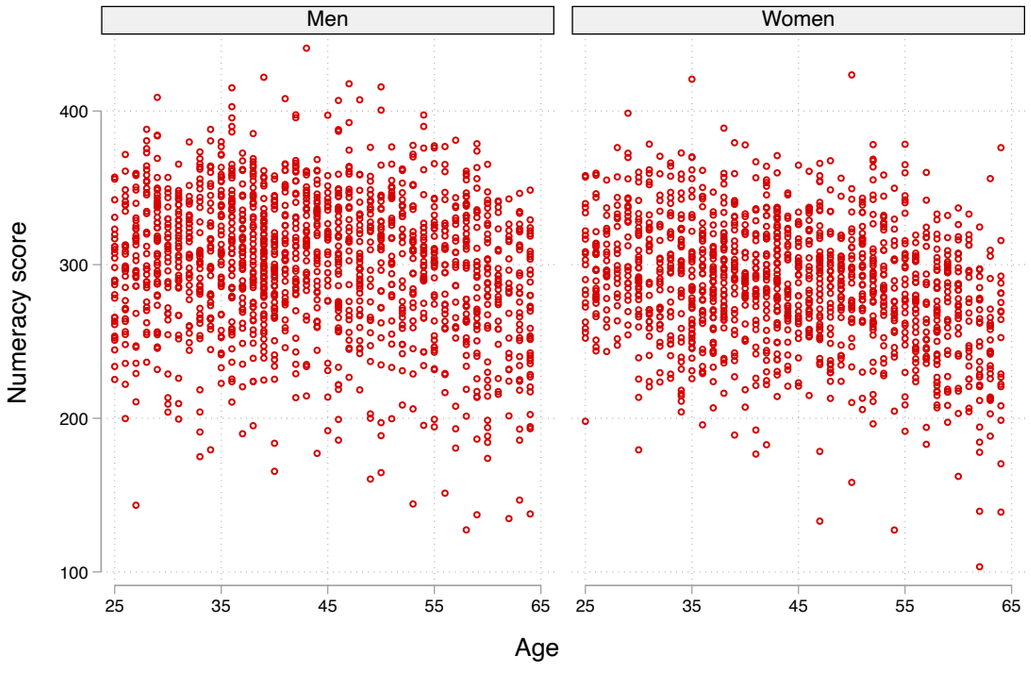


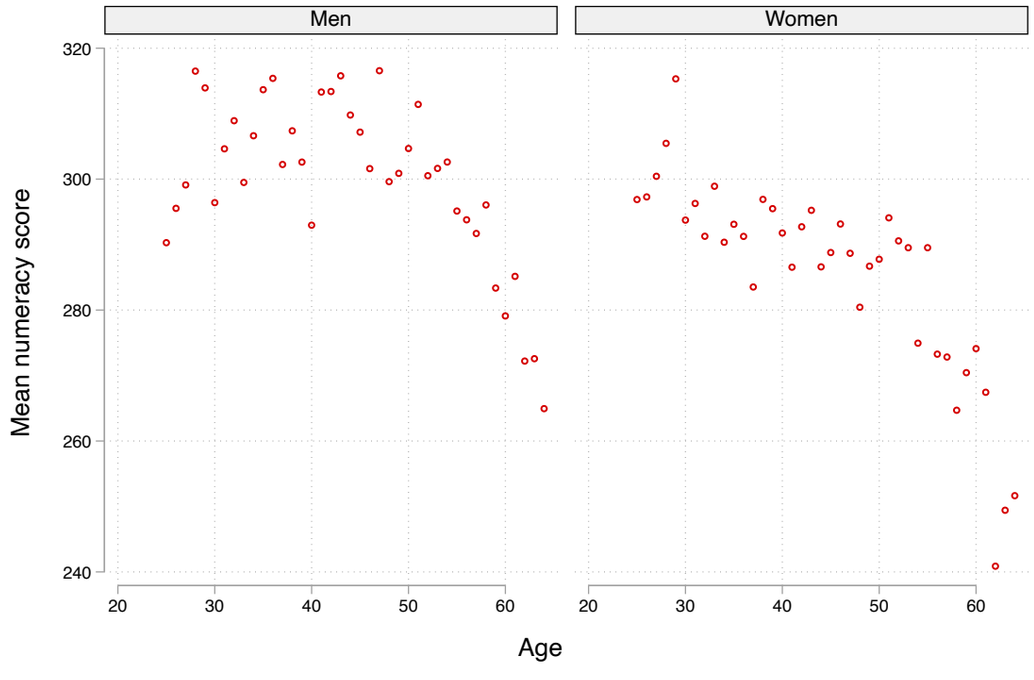
相関係数を計算し、散布図を作成してみよう（3.4.1）

散布図から変数間の関係性を探索する

いろいろなパターンの散布図を作成してみよう（3.4.2）

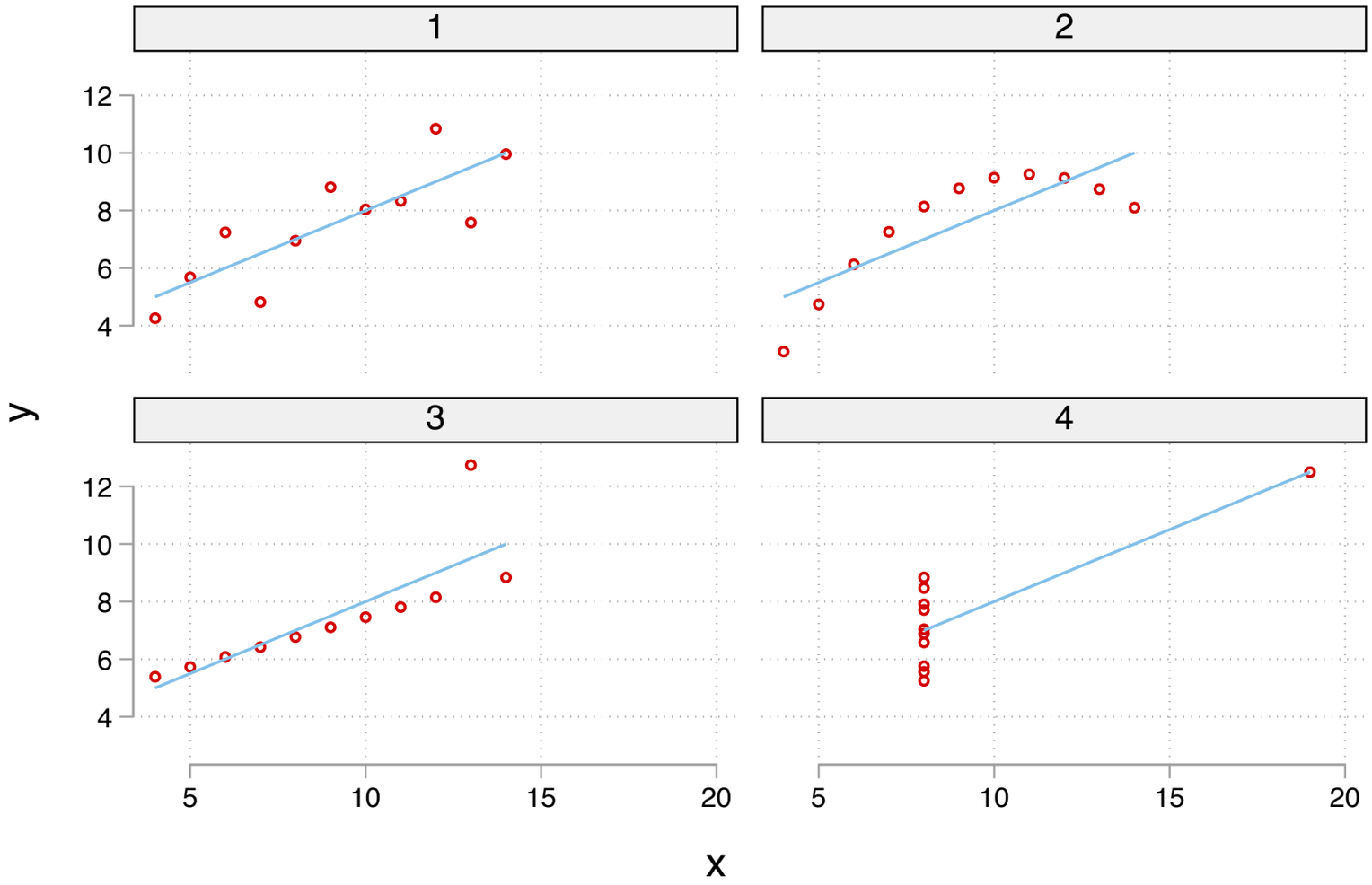






相関係数の罠（Anscombe's Quartet）

たとえ同じ相関係数であったとしても、それが同じような線形の関係を表しているとは限らない。常にデータを可視化して確かめることが重要



Anscombe's Quartetの散布図を作成してみよう（3.4.3）

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。

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：

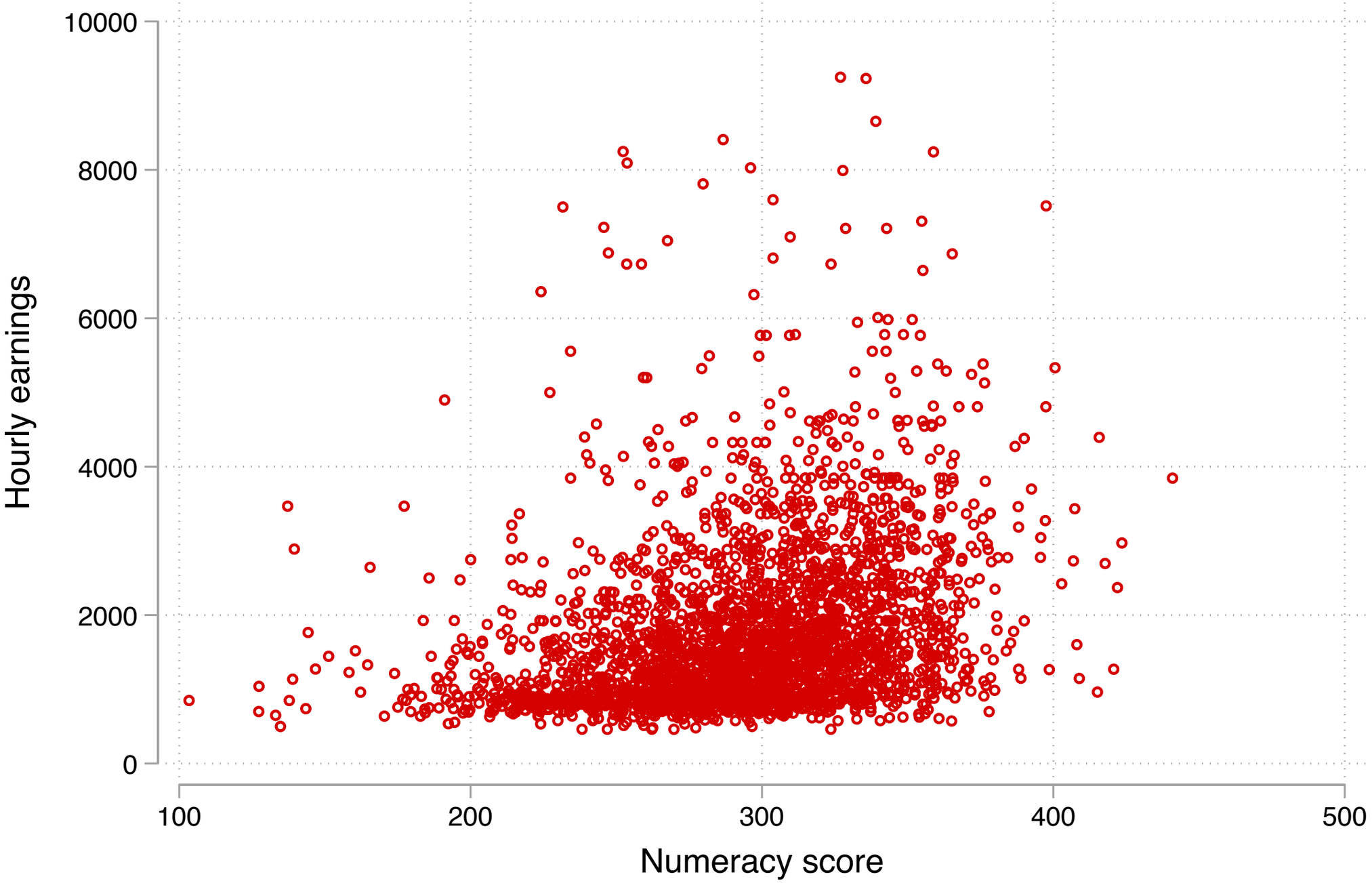
川口大司，2017，『労働経済学：理論と実証をつなぐ』有斐閣.

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.

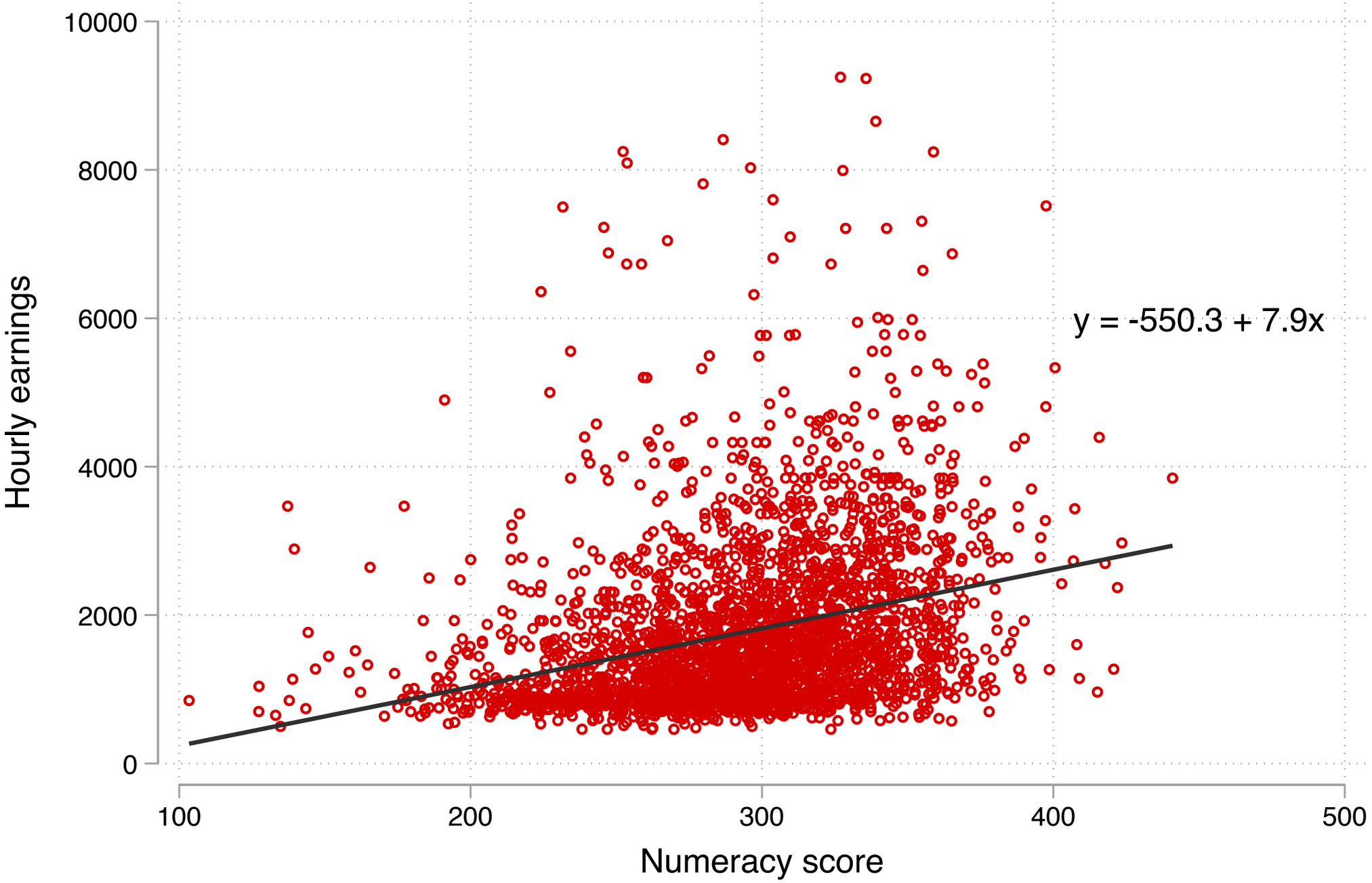
Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2015. “Returns to Skills around the World: Evidence from PIAAC.” European Economic Review 73:103–30.

散布図を描いてみる


10000
8000
6000
4000
2000
0
Hourly earnings
100
200
300
400
500
Numeracy score

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指して、**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 $\beta_0, \beta_1, \cdots, \beta_k$ 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することができる：

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

*ここでの仮定： $E(\varepsilon \mid X_1, \cdots, X_k) = E(\varepsilon),\ E(\varepsilon) = 0$

傾きの係数の解釈

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

Xが1単位増加したときのYの増加分をΔYとおく*。

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

$$\Delta Y = \beta_1$$

傾きの係数は、Xが1単位増加したときのYの増加分を表す。

X1単位の変化に対するYの変化量を**限界効果 marginal effect**という。

*高校数学II/IIIを勉強したことがある人はxで微分すると言うことと同じです（以下同じ）。

Stataでの回帰分析の出力結果

4_regression2023-09-05.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
```

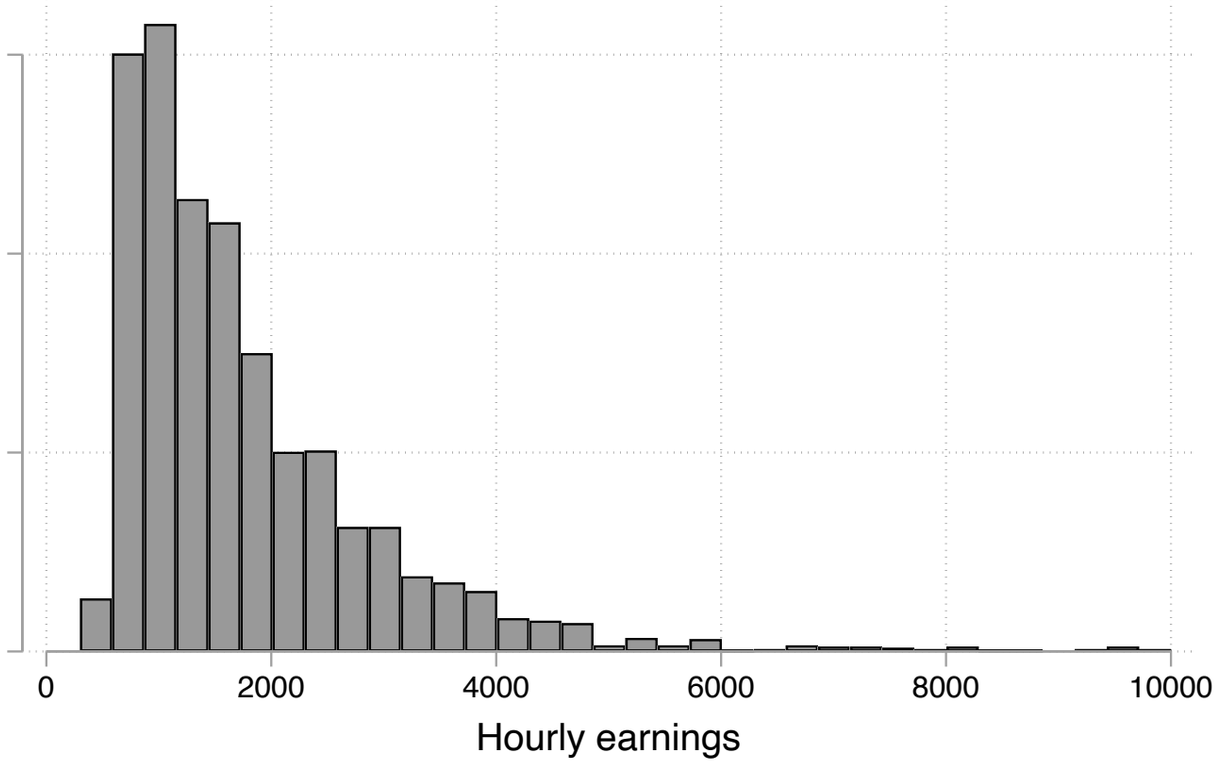
Source	SS	df	MS	Number of obs	=	2,805
				F(1, 2803)	=	273.48
Model	329439217	1	329439217	Prob > F	=	0.0000
Residual	3.3765e+09	2,803	1204614.05	R-squared	=	0.0889
				Adj R-squared	=	0.0886
Total	3.7060e+09	2,804	1321673.47	Root MSE	=	1097.5

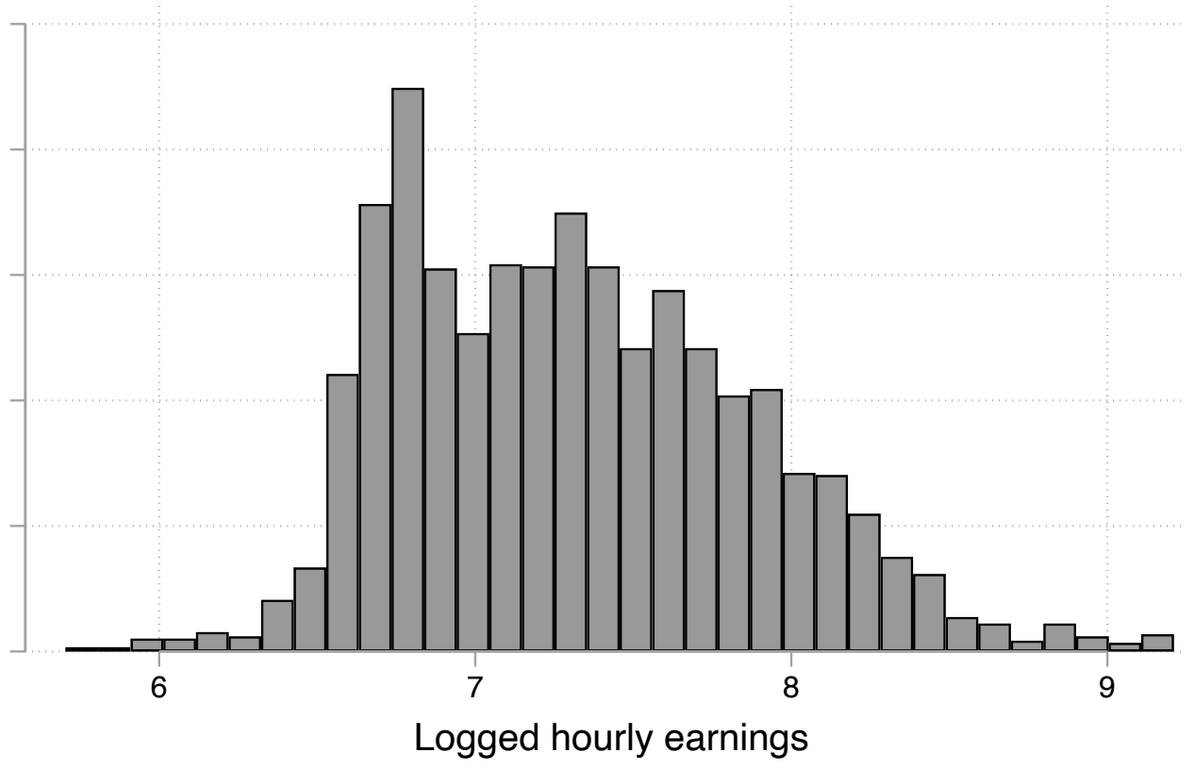
wage	Coefficient	Std. err.	t	P>|t|	[95% conf.	interval]
numeracy	7.904687	.4779924	16.54	0.000	6.967435	8.84194
_cons	-550.2746	142.1371	-3.87	0.000	-828.9786	-271.5707

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

$e = \lim_{t \to 0}(1+t)^{\frac{1}{t}} \simeq 2.7182818\cdots$ 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：

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

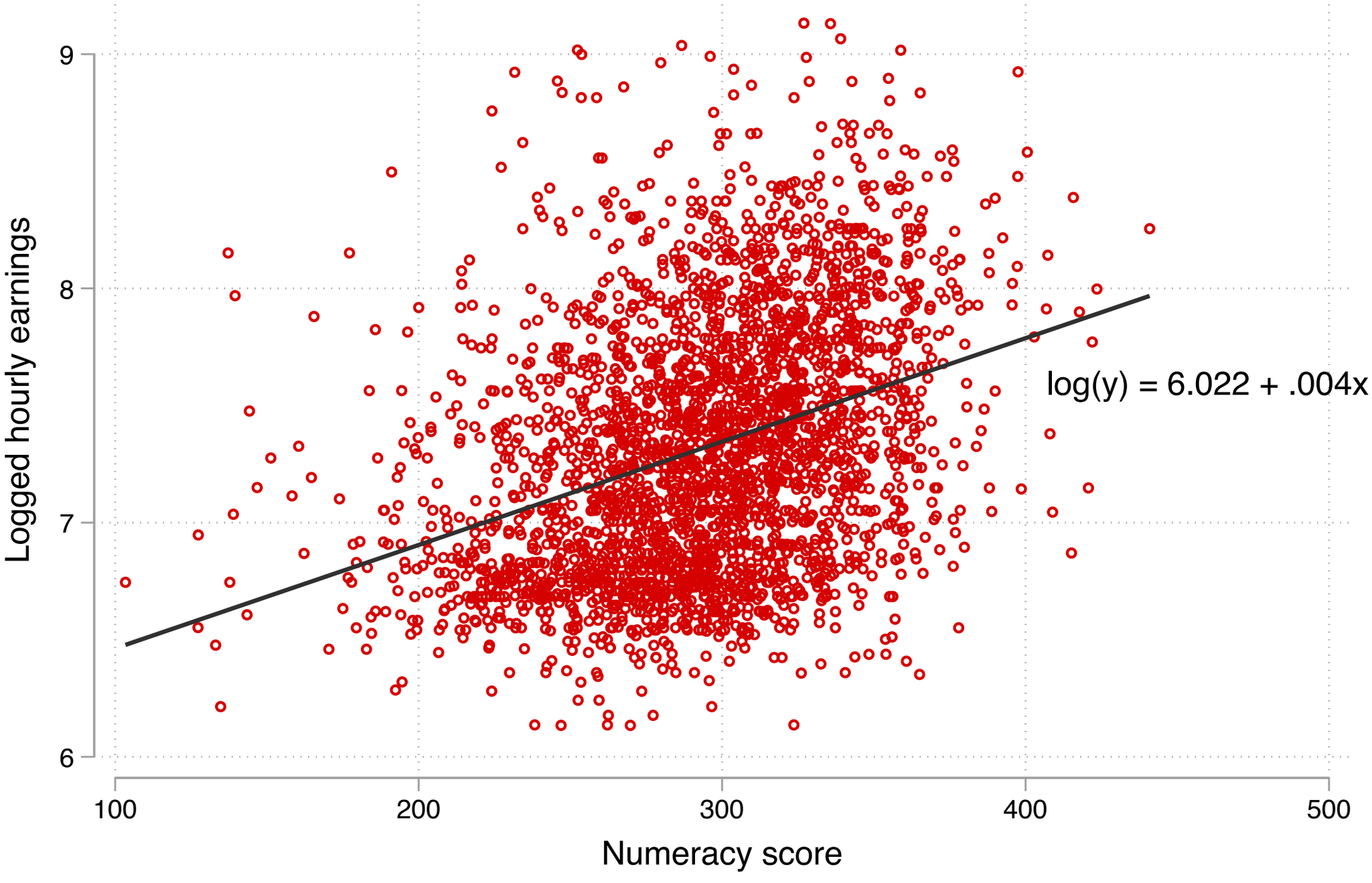
とくに底がeの対数関数を自然対数という。社会科学系の文脈では、この場合は底を省略して、$e^y = x \leftrightarrow y = \log(x)$と書かれることが多い。$\ln(x)$の場合もある。

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：　$(\log x)' = 1/x$

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

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

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：

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

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

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

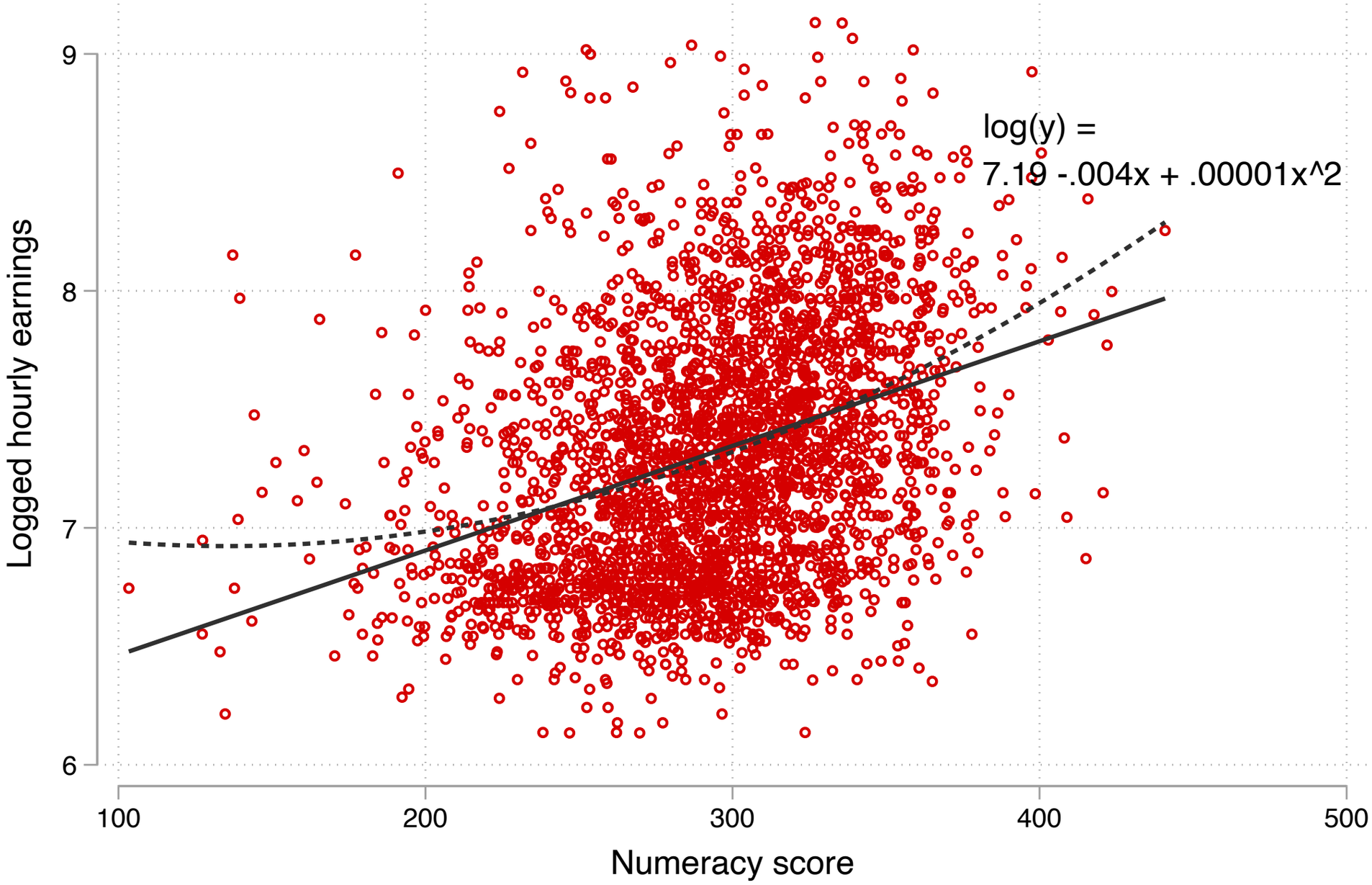
係数が大きくなるほど両者は一致しない。正確な値を知るには$[\exp(\beta_1) - 1]$ $(\times 100\%)$ を計算する

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β_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

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

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。

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、その都度確認するとよい。

`margins`コマンドを使うと、指定した独立変数の値ごとに予測値を計算してくれて便利。

複数の回帰分析の結果を比較する

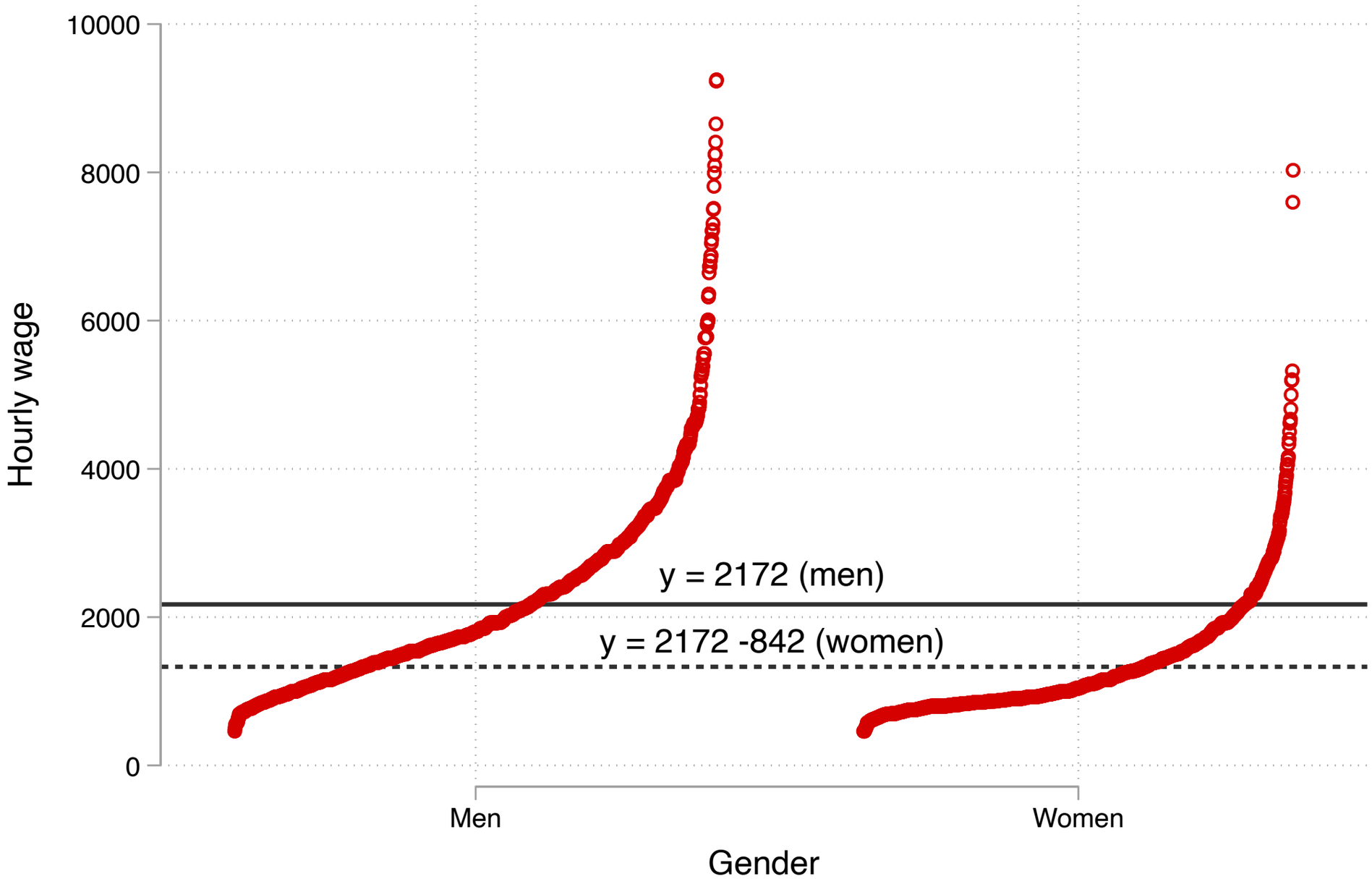
対数変換しない賃金を使ったときの結果、対数変換した賃金を使った結果、2乗項を使った結果の3つをならべて結果を比較してみよう（4.1.3）

`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：$Y = \beta_0 + \varepsilon,\ E(Y|D = 0) = \beta_0$

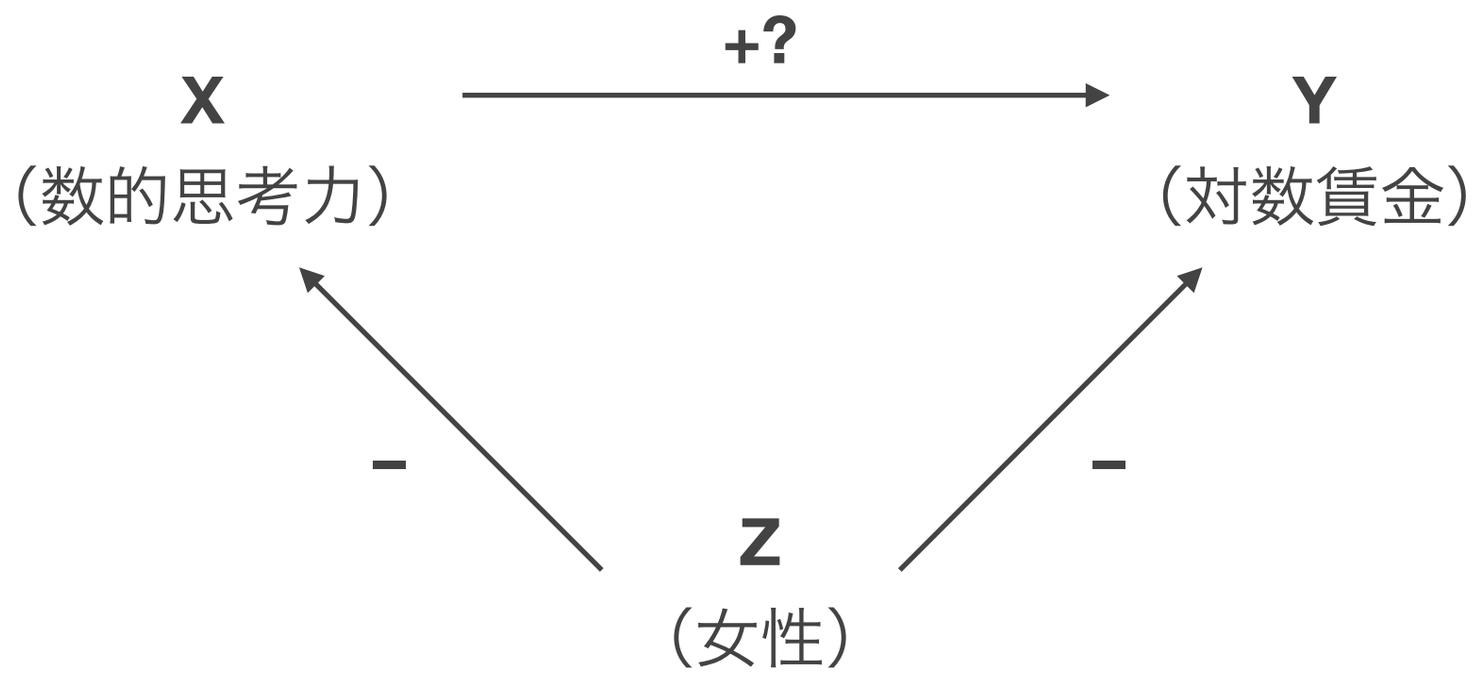
$D = 1$（女性）のとき：$Y = \beta_0 + \beta_1 + \varepsilon,\ E(Y|D = 1) = \beta_0 + \beta_1$

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。



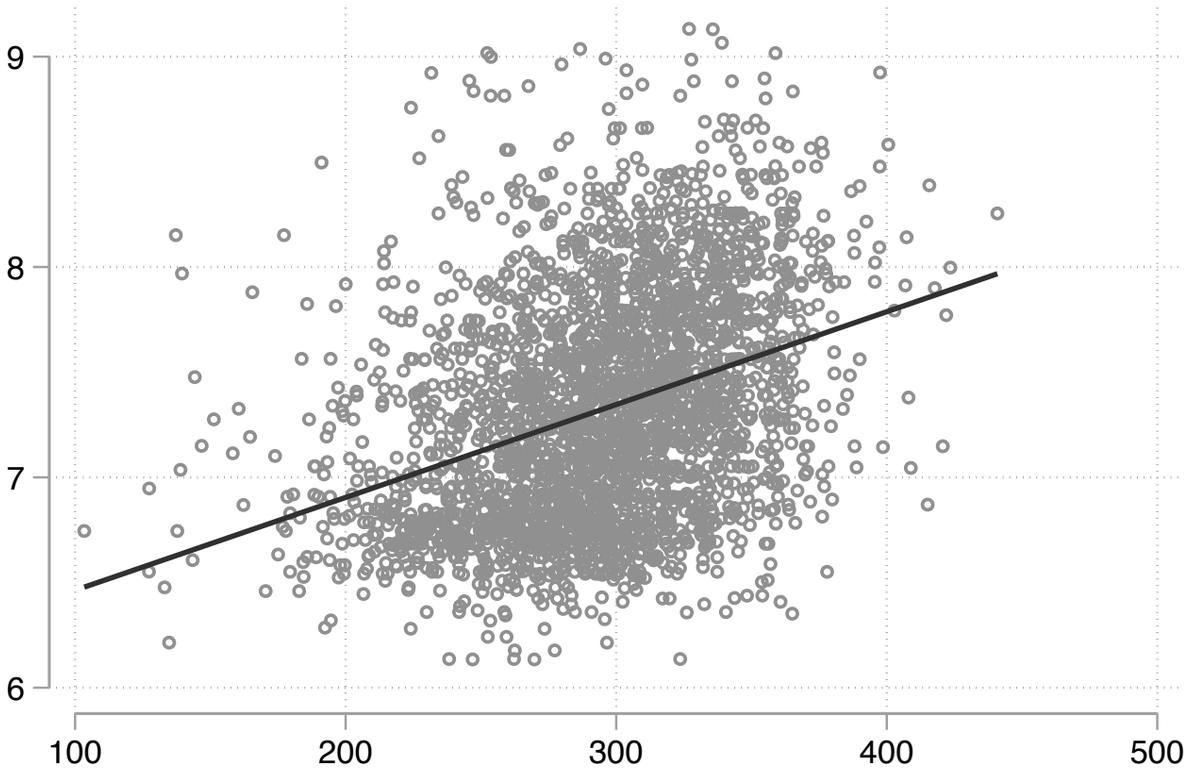
XとYの両者に影響する**交絡要因Z**を統制することで、Yに対するXの**因果効果に近づく**ことができる。

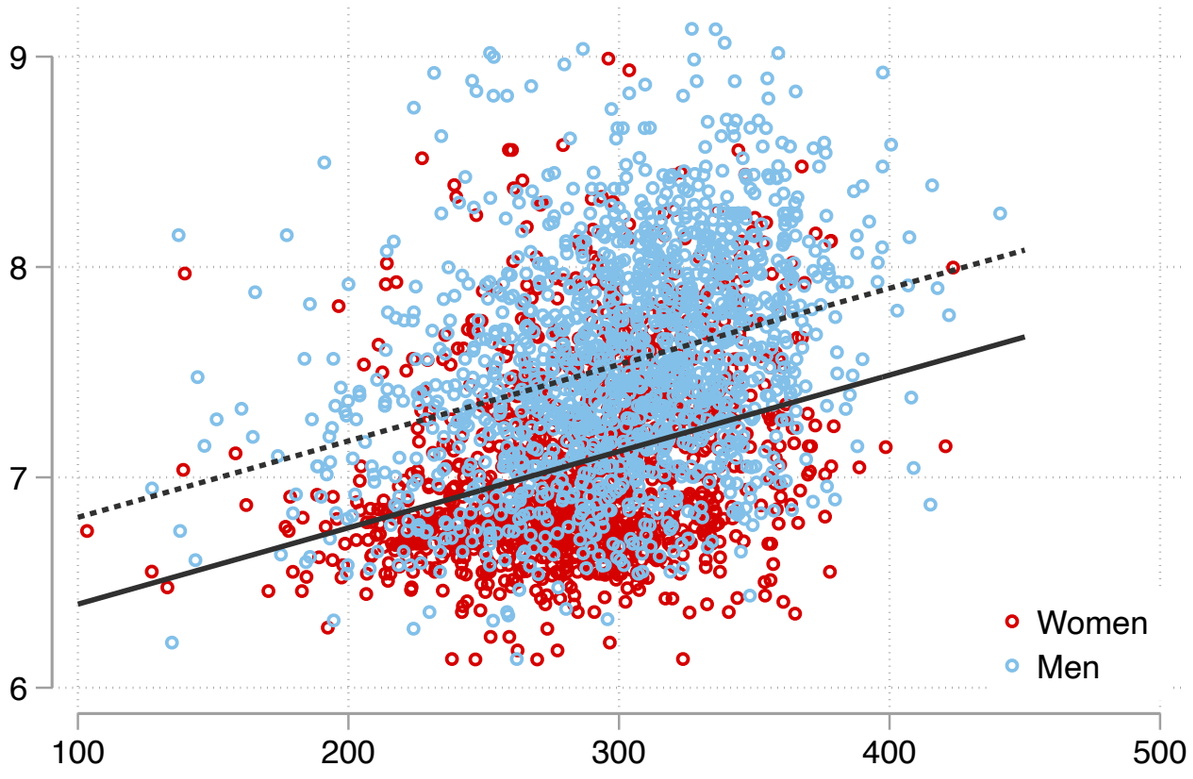
単回帰分析と重回帰分析を比較してみる

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

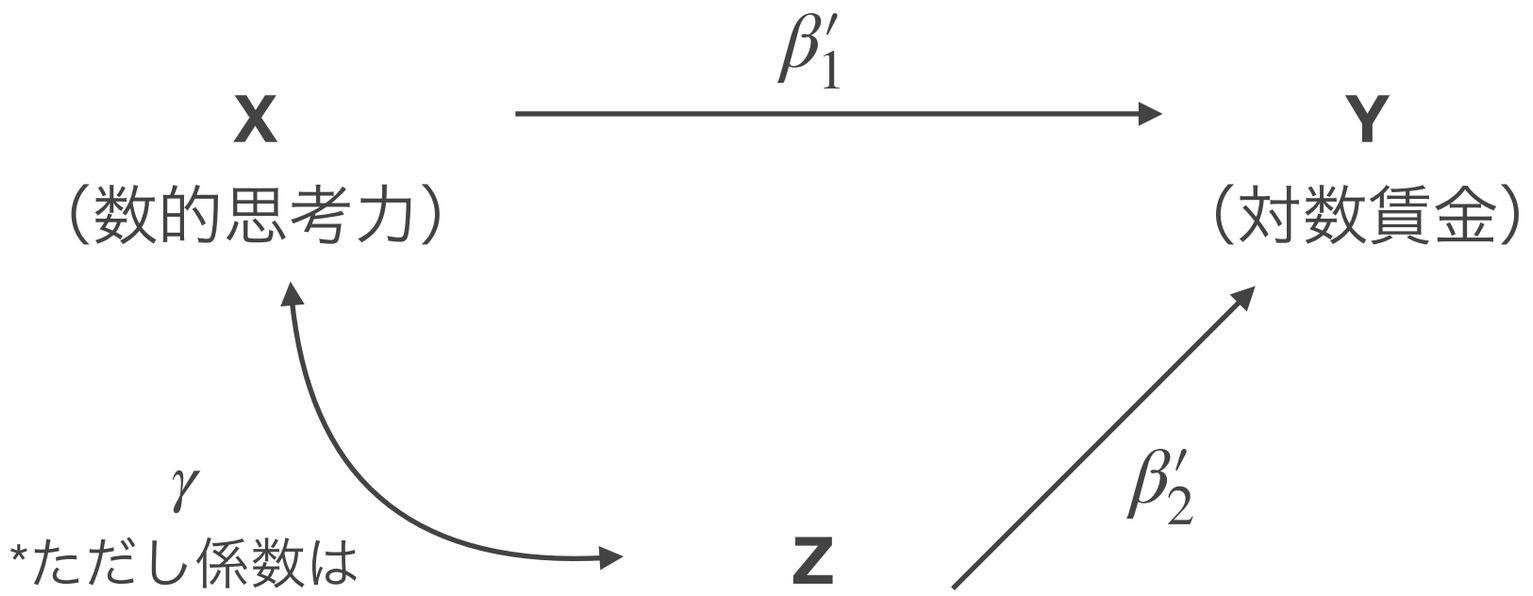
$$Y = \beta_0' + \beta_1' X + \beta_2 Z + \varepsilon'$$

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

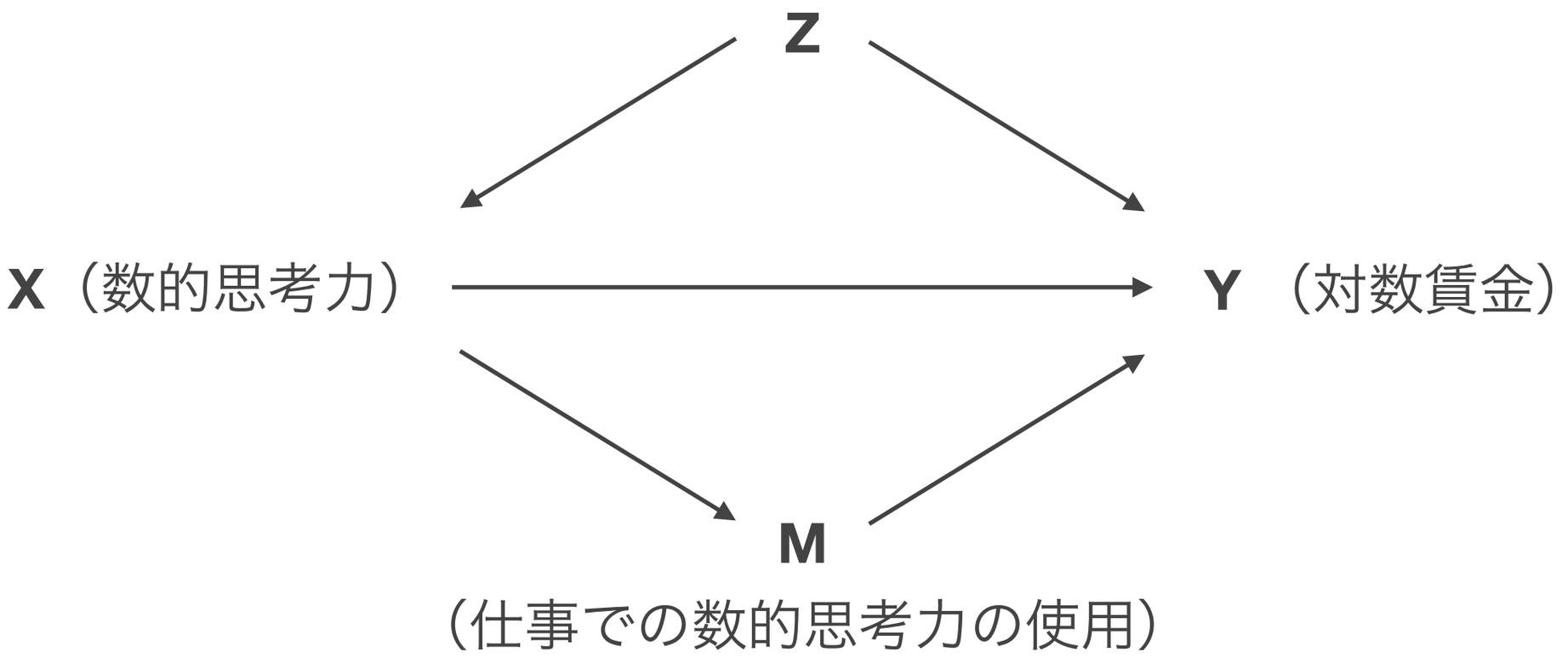
→独立変数が無作為に割り当てられていない限り、「数的思考力が高いと賃金が高く**なる**」とはいえない

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）

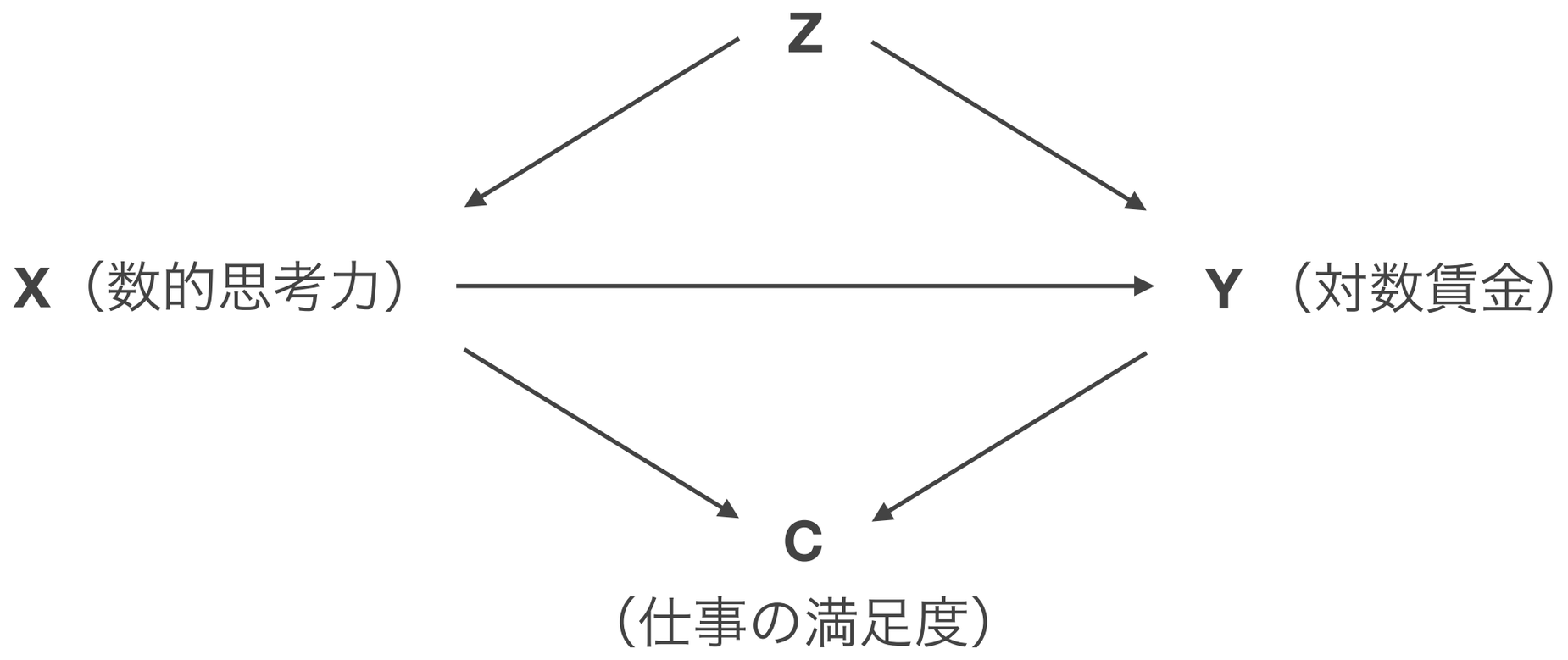
統制すべき変数を吟味する：媒介要因mediator



重回帰分析を因果効果を知るために使う場合、**Mのような変数を投入するかどうかは知りたい因果効果の内容に依存する**

- もし知りたい因果効果が「同じくらい仕事で数的思考力を使っていたとしてもなお数的思考力が賃金を高める効果」であるなら、Mは統制すべき
- 「数的思考力が賃金を高める効果」であるなら、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



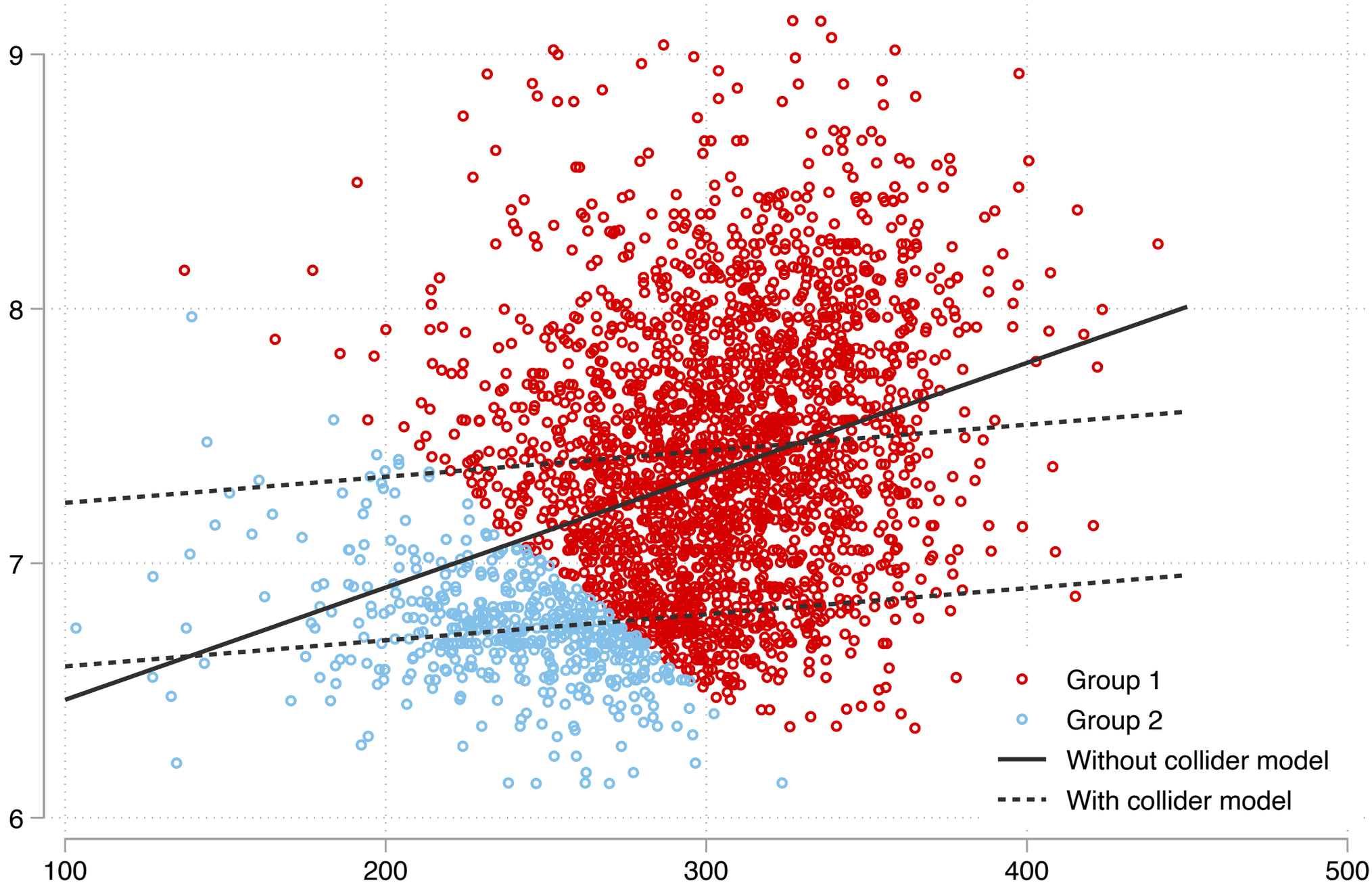
重回帰分析を因果効果を知るために使う場合、**Cのような変数は投入してはいけない**

Cのような変数を統制することによってXの係数にバイアスが生じる。これを指して**合流点バイアス Collider bias**、分野によっては**選択バイアス Selection bias**などともいう

（Elwert and Winship, 2014）

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相関関係から因果関係に近づくことができる。しかし、適切でない要因を統制してしまうと、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？

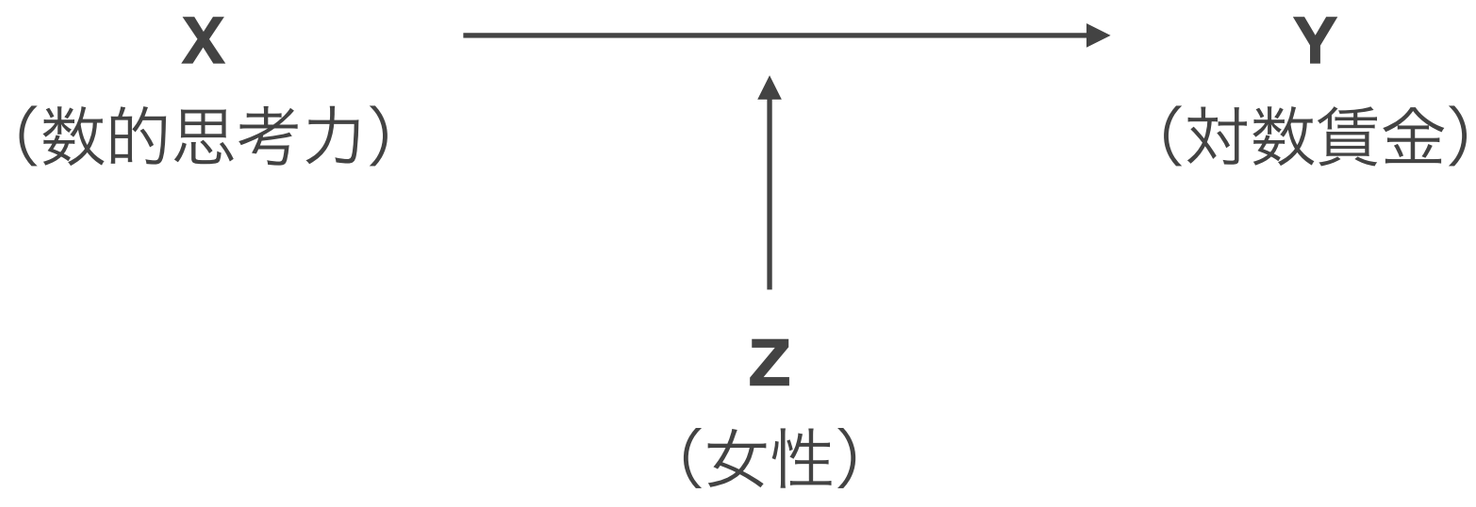
そのために、どのような交絡要因や媒介要因を統制すればよいか？

（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？

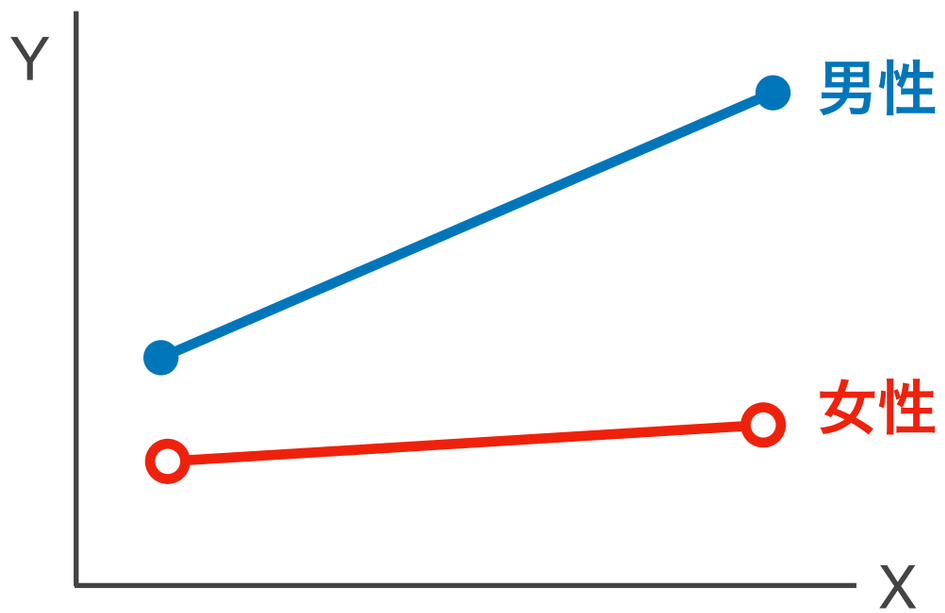
を考えることが大事

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ということが考えられる。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：

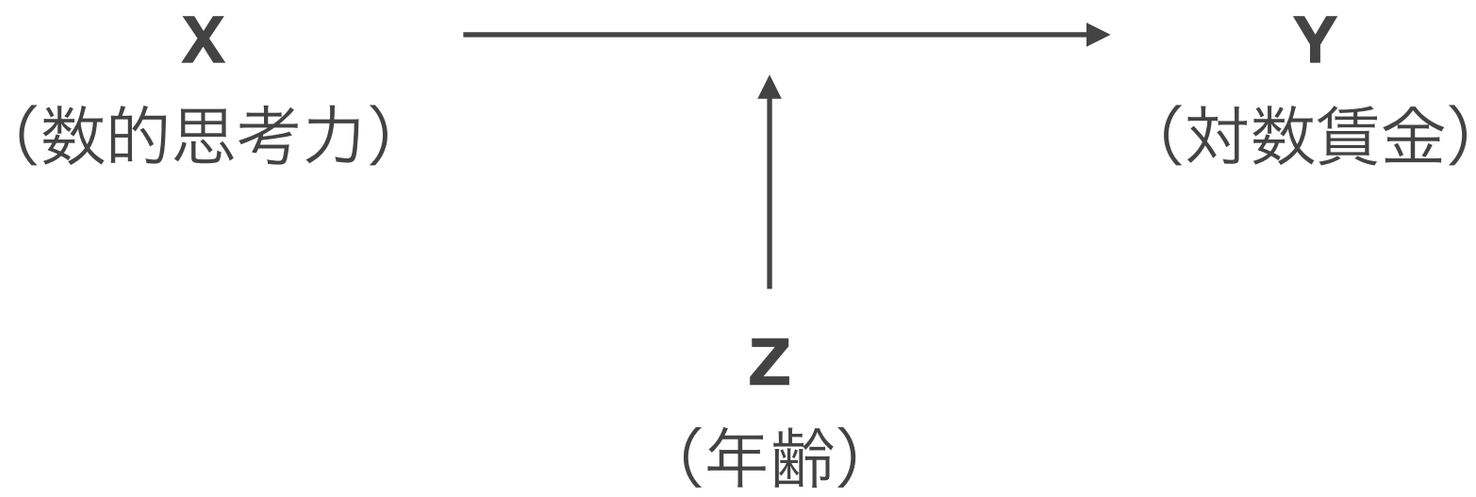
$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

$Z = 0$（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



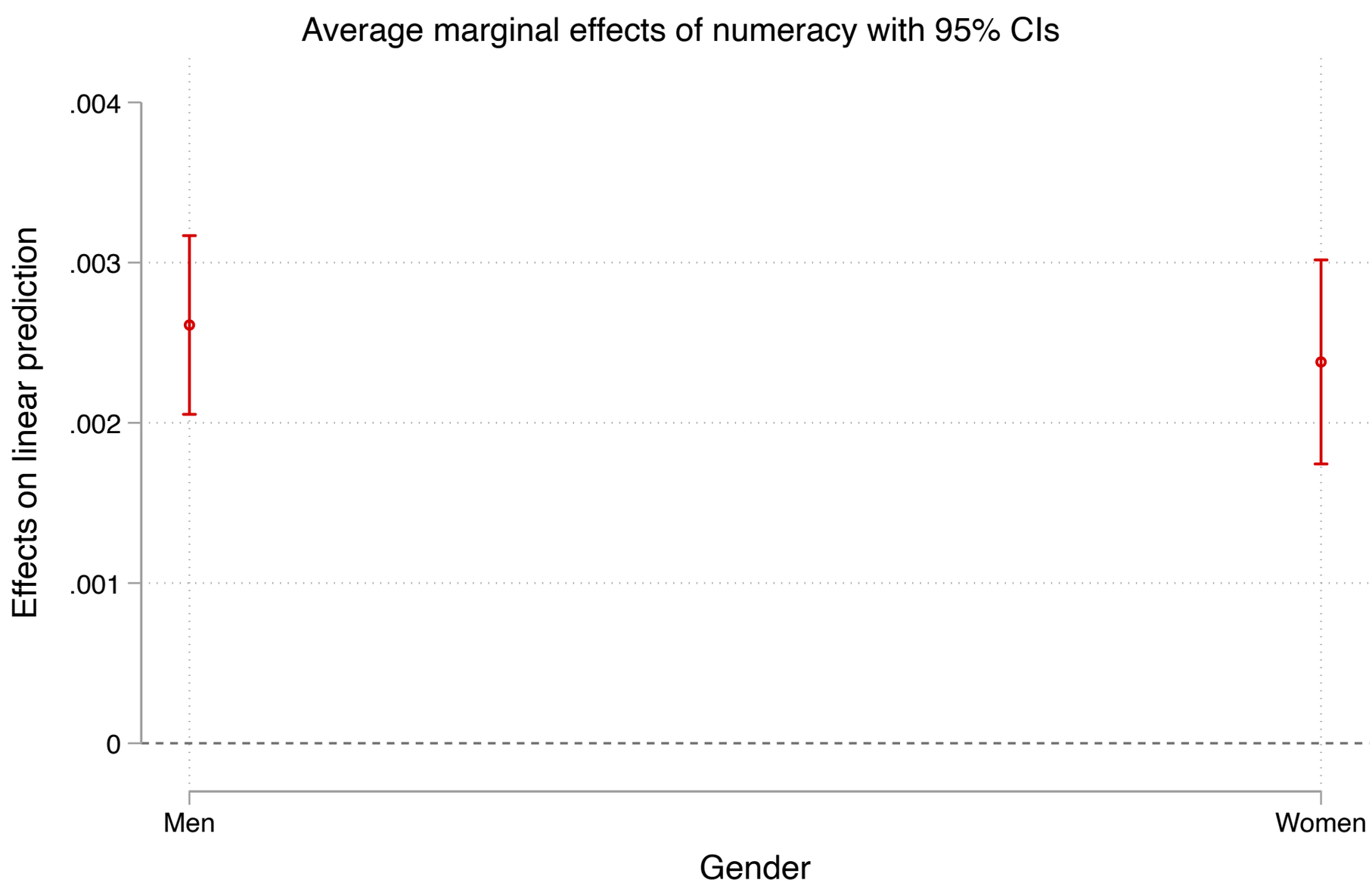
Zが連続変数のときを考える：

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

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。

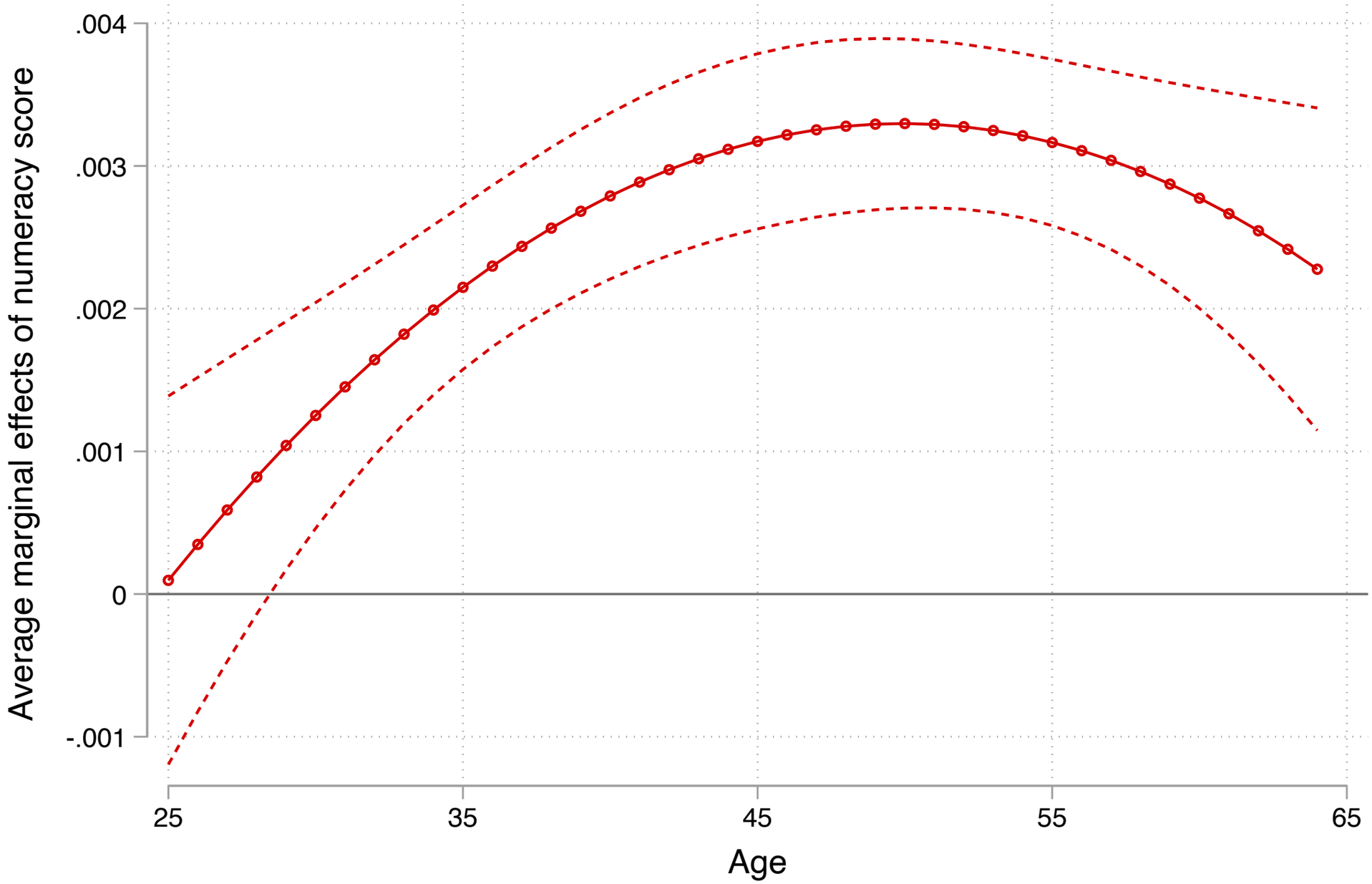
性別ごとにみた限界効果

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

たとえば$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon$において、$\mathrm{Cor}(X_2, X_3 \mid X_1)$ 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。

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

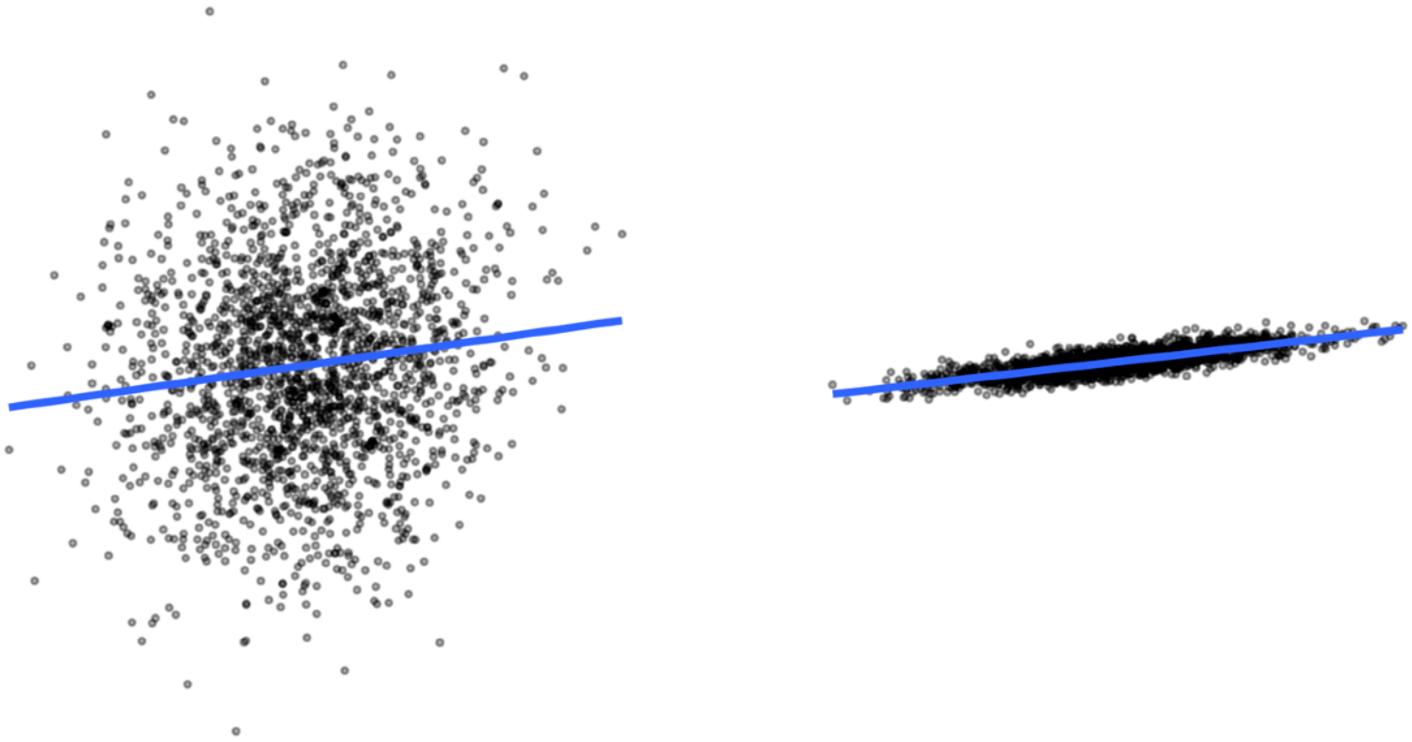
- 知りたい係数がβ_1であるなら、気にする必要はない。
- 知りたい係数がβ_2またはβ_3のどちらかまたは両方なら、
  - VIF以外にも標準誤差が過剰に大きくなったりしているかなどをチェックし、問題なさそうなら、気にする必要はない。
  - 問題ありそうなら、理論的な妥当性などを鑑みつつ、どちらかを除外する

決定係数 R^2

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。

$R^2 = \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)} = 1 - \frac{\mathrm{Var}(\varepsilon)}{\mathrm{Var}(Y)}$ 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を受けやすいという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.

クロス集計表をみる

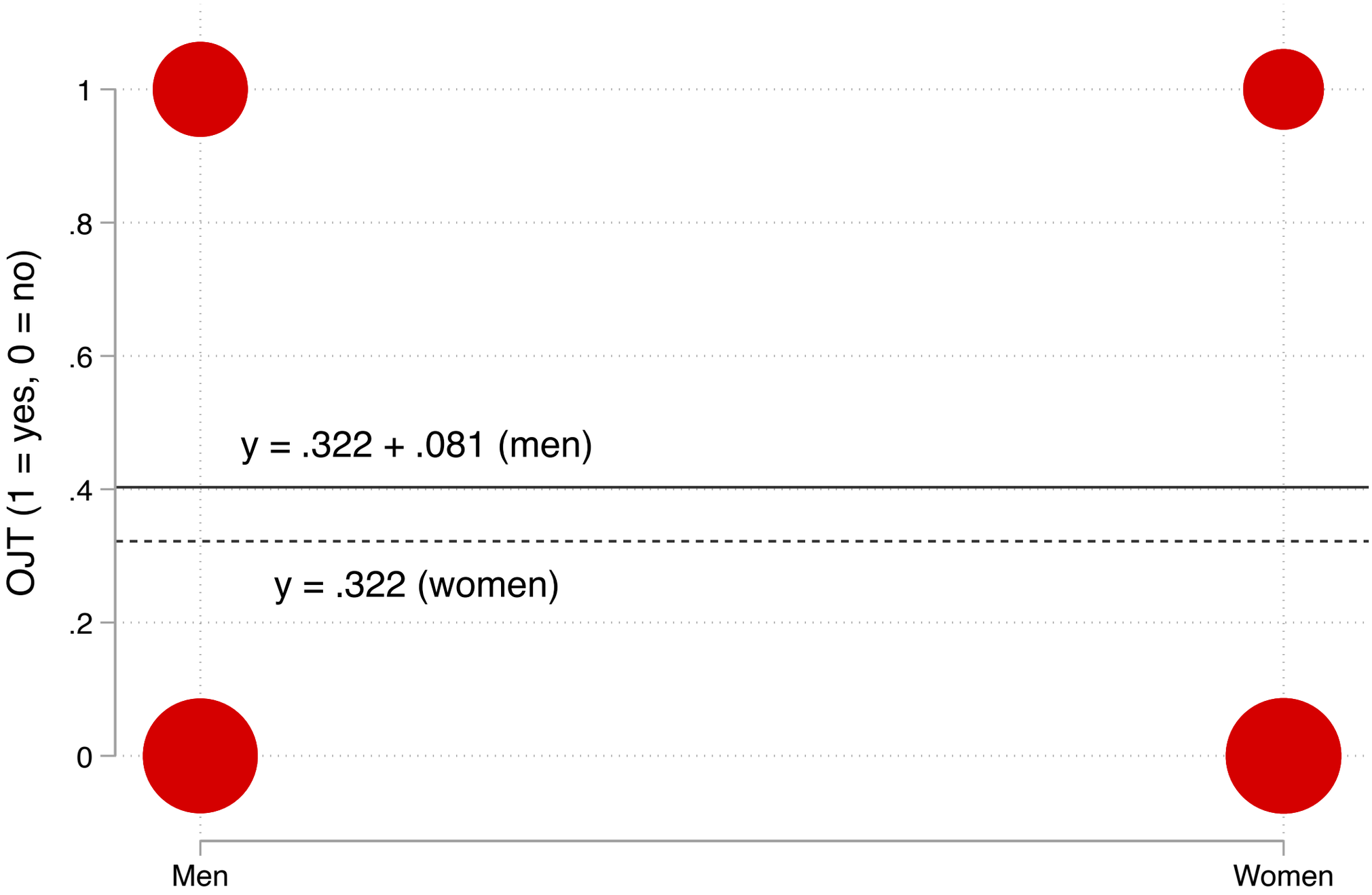
Gender	OJT No	OJT Yes	Total
Men	885 59.68	598 40.32	1,483 100.00
Women	896 67.78	426 32.22	1,322 100.00
Total	1,781 63.49	1,024 36.51	2,805 100.00

男性は女性と比べてこの1年にOJTを受けている割合が8.1%ポイント高い。

線形回帰モデルを使って表現すると？

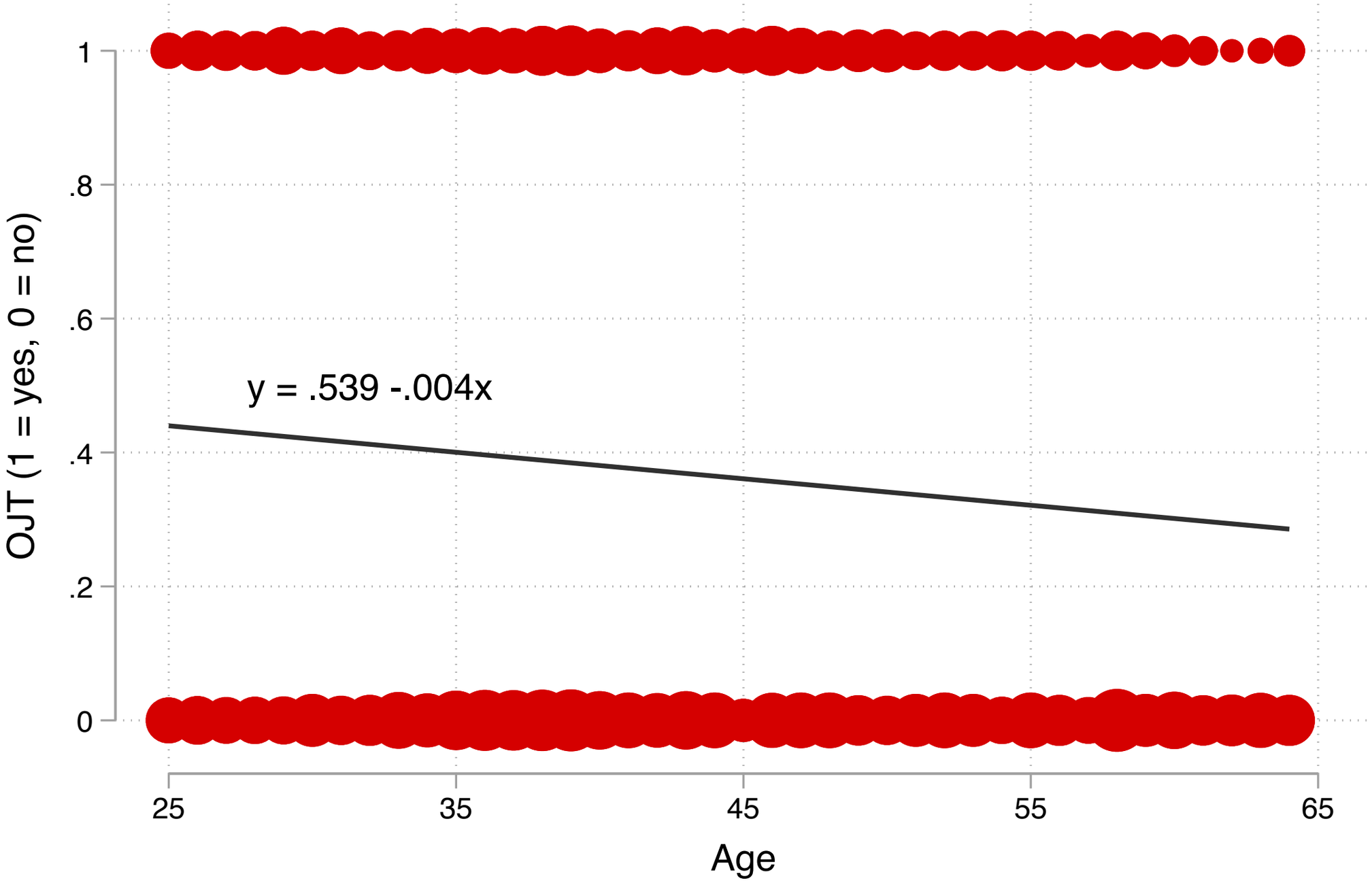
Yが二値変数、Xが二値変数の場合

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 Linear Probability Model

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

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

傾きの係数は、Xが1単位増加したときの$\Pr(Y)$の増加分を表す。

線形確率モデルを推定する

5_logit2023-09-05.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                               Number of obs     =      2,805
                                                F(2, 2802)        =      21.90
                                                Prob > F          =     0.0000
                                                R-squared         =     0.0147
                                                Root MSE          =     .47815

------------------------------------------------------------------------------
             |               Robust
         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
------------------------------------------------------------------------------
```

よく言われている線形確率モデルの注意点（Mood, 2010）

1. 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

   →普通の回帰分析でもこういうことはある

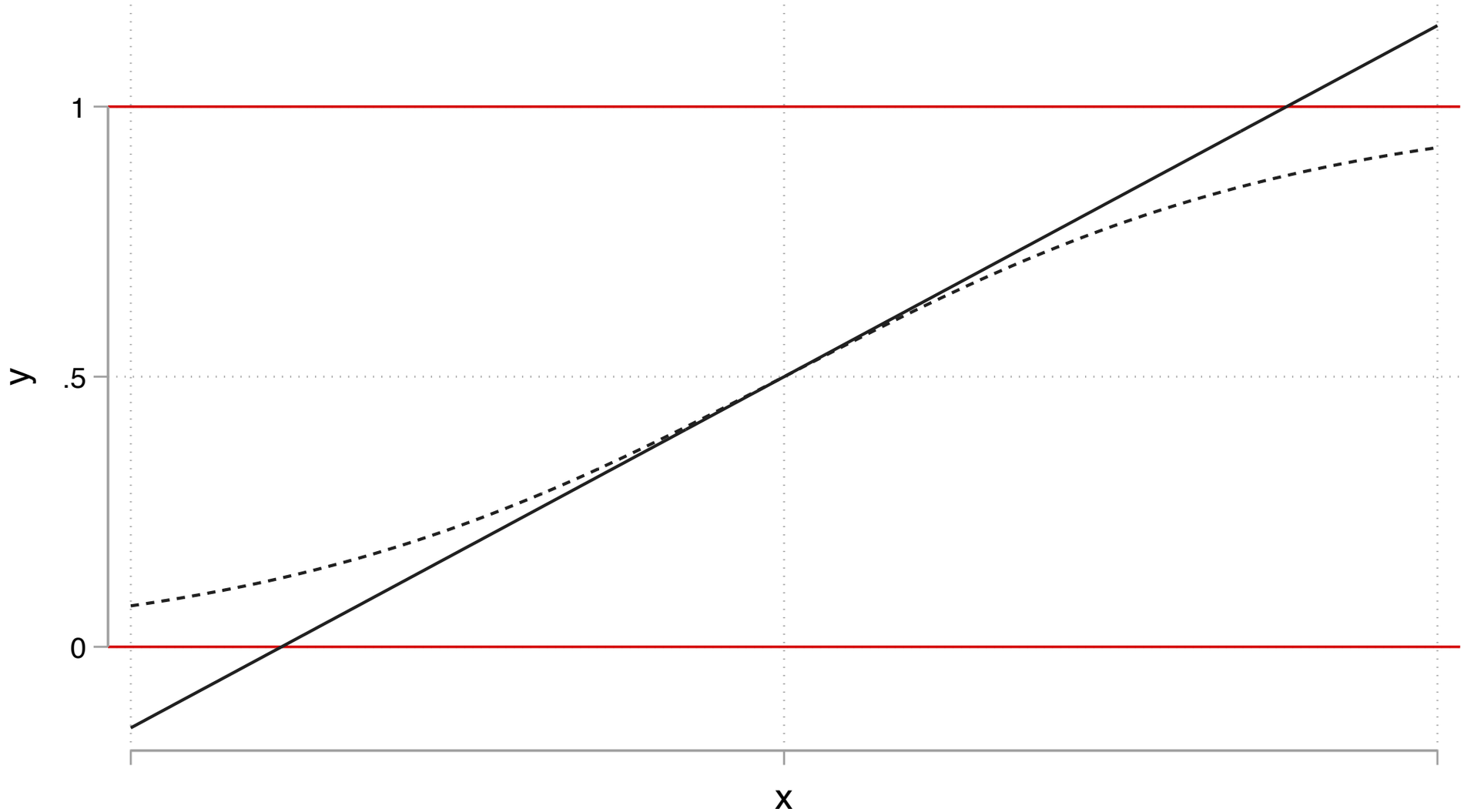
2. 残差が正規分布しない（不均一分散）ため標準誤差にバイアスが生じる

   →ロバスト標準誤差（頑健標準誤差）を使うことで対処可能

3. **関数型の誤り**：もし真の関係が非線形――従属変数が1をとる確率が低い個人と中程度の個人で、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――のであれば、変数の効果を正しく推定できない

ロジスティック曲線の当てはめ

線形ではなく以下のような曲線を当てはめられれば、先の問題（1や3）に対処することができるのではないか？



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。

（対数）オッズとは何か

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

X = 1におけるオッズ：$p_1/(1-p_1)$

X = 1における対数オッズ：$\log(p_1/(1-p_1))$

X = 2におけるオッズ：$p_2/(1-p_2)$

X = 2における対数オッズ：$\log(p_2/(1-p_2))$

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：

$$\log \frac{p_1/(1-p_1)}{p_2/(1-p_2)}$$
$$= \log(p_1/(1-p_1)) - \log(p_2/(1-p_2))$$

具体例

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

男性の対数オッズ：log(0.676) = **-0.391**

女性の対数オッズ：log(0.475) = **-0.744**

対数オッズ比（男性の対数オッズ – 女性の対数オッズ）：log(0.676) - log(0.475) = **0.352**

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
------------------------------------------------------------------------------
```

切片：女性の対数オッズ

係数：女性の対数オッズと比べて男性の対数オッズがどの程度高いか

注意点は線形回帰分析と共通

LPMもロジットも以下の点は共通しており、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して調整効果を検討できる

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

ロジスティック回帰分析の実質的意味

確率による解釈とオッズによる解釈の対比

線形確率モデル：確率による解釈

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）オッズによる解釈

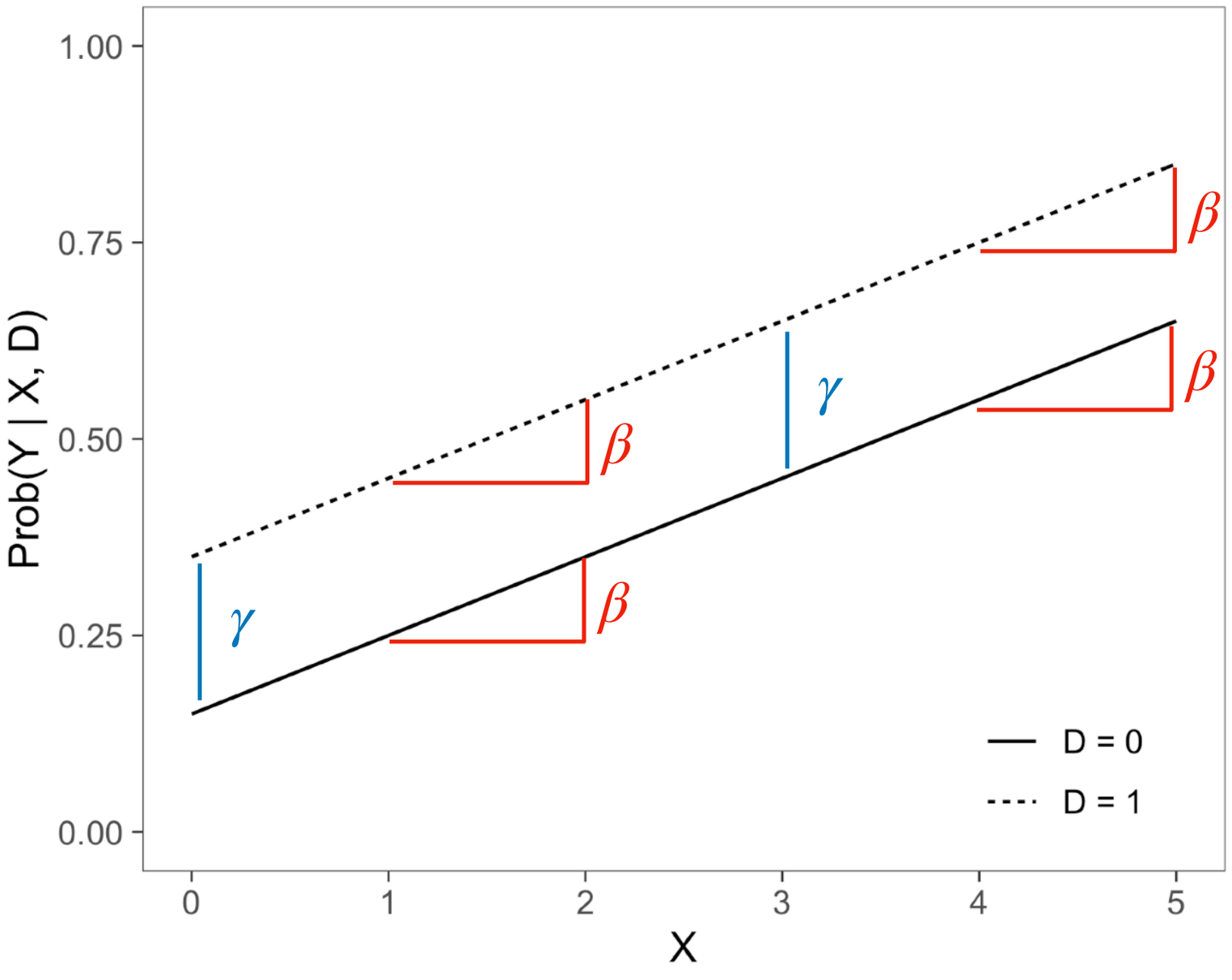
他の要因を一定として、男性がOJTを受けるオッズは女性の1.29 = exp(0.255) 倍である

	LPM	Logit
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

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）を当てはめた場合、$\Pr(Y)$の値によらず1単位の変化に対する確率の変化量は係数に一致



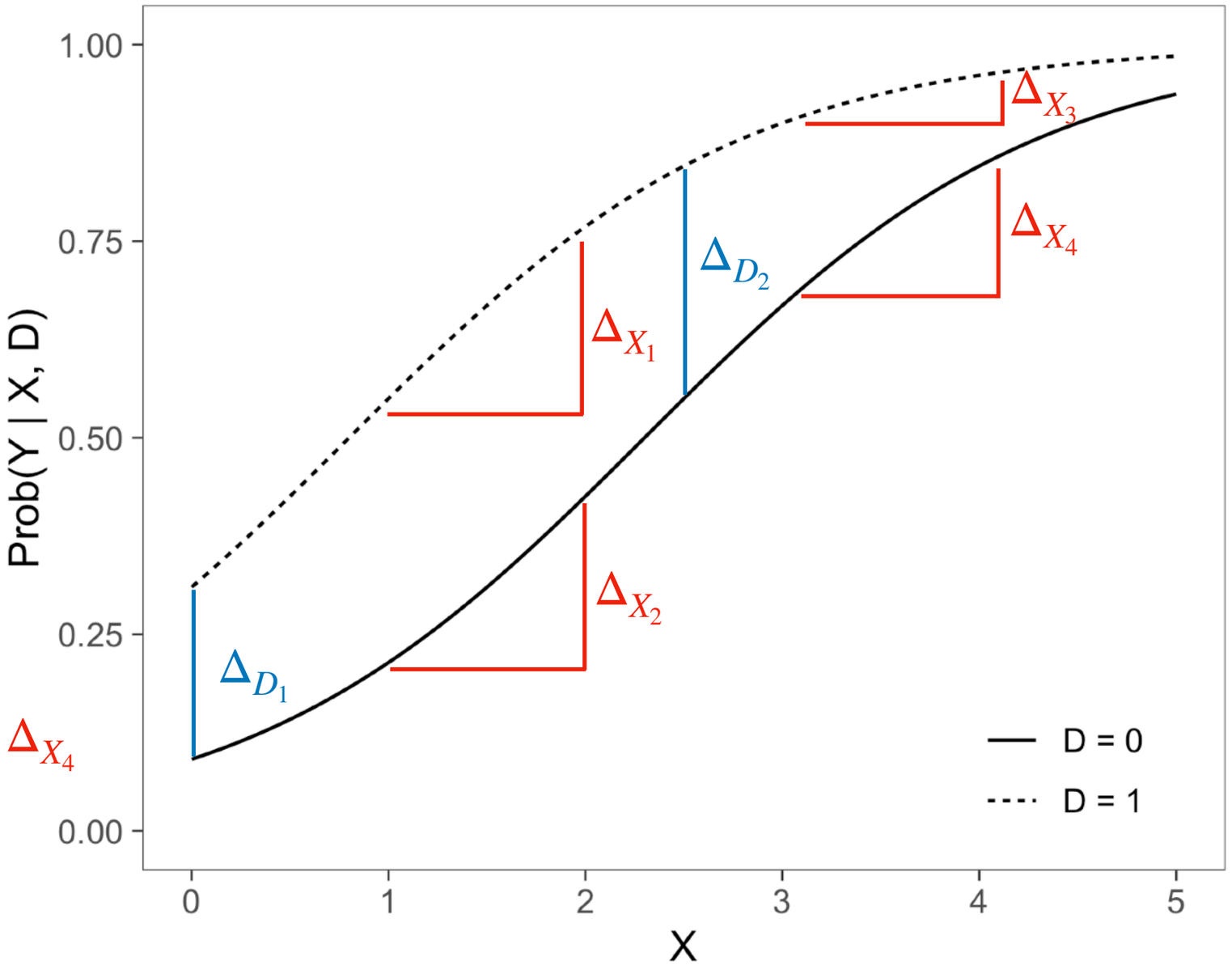
$\Delta_D = \gamma$

$\Delta_X = \beta$

非線形モデル（ロジット）の場合

$$\Pr(Y) = \frac{\exp(\alpha + \beta X + \gamma D)}{1 + \exp(\alpha + \beta X + \gamma D)}$$を当てはめた場合、1単位の変化に対する確率の

変化量は$\Pr(Y)$の値によって異なる（係数そのものは確率の変化を意味しない）



$\Delta_{D_1} \neq \Delta_{D_2}$

$\Delta_{X_1} \neq \Delta_{X_2} \neq \Delta_{X_3} \neq \Delta_{X_4}$

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**が変わることはないため、まずは線形確率モデル（+ロバスト標準誤差）を使って分析してもよい

線形確率モデル：係数が解釈しやすい

ロジットモデル：確率特有の非線形性を適切に表しているかもしれない

ロジットモデルの結果から「確率による解釈」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

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単位の変化をみる

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単位の変化をみる

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

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

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

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

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

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

平均的に、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。

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。

			(1)	(2)	(1) – (2)
id	X	D	Pr(Y = 1 | X, D = 1)	Pr(Y = 1 | X, D = 0)	ΔPr(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

← Pr(Y = 1 | X, D = 1)
← 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を出さない理由

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

β_2をそのまま解釈しようとするなら：Ageを一定としたうえでAge^{2}が1単位増加したときの対数オッズの増分

しかし、「Ageを一定としたうえでAge^{2}が1単位増加する」ことは定義上あり得ない。そのため、Age^{2}の限界効果だけを単独で考えることには意味がない

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

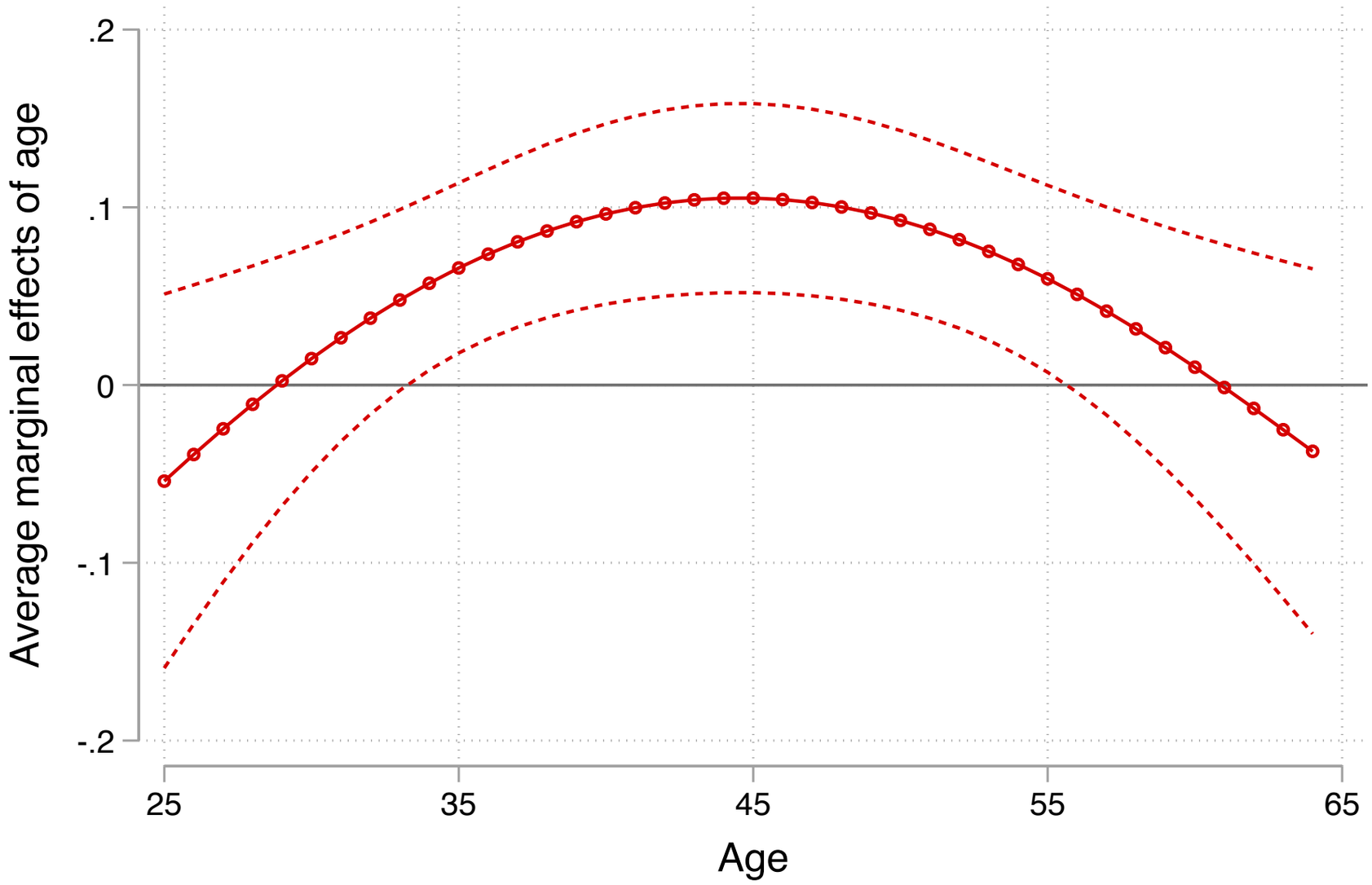
ロジットモデルでも線形回帰分析のときと同様に調整効果（交互作用効果）を考えることができる。

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

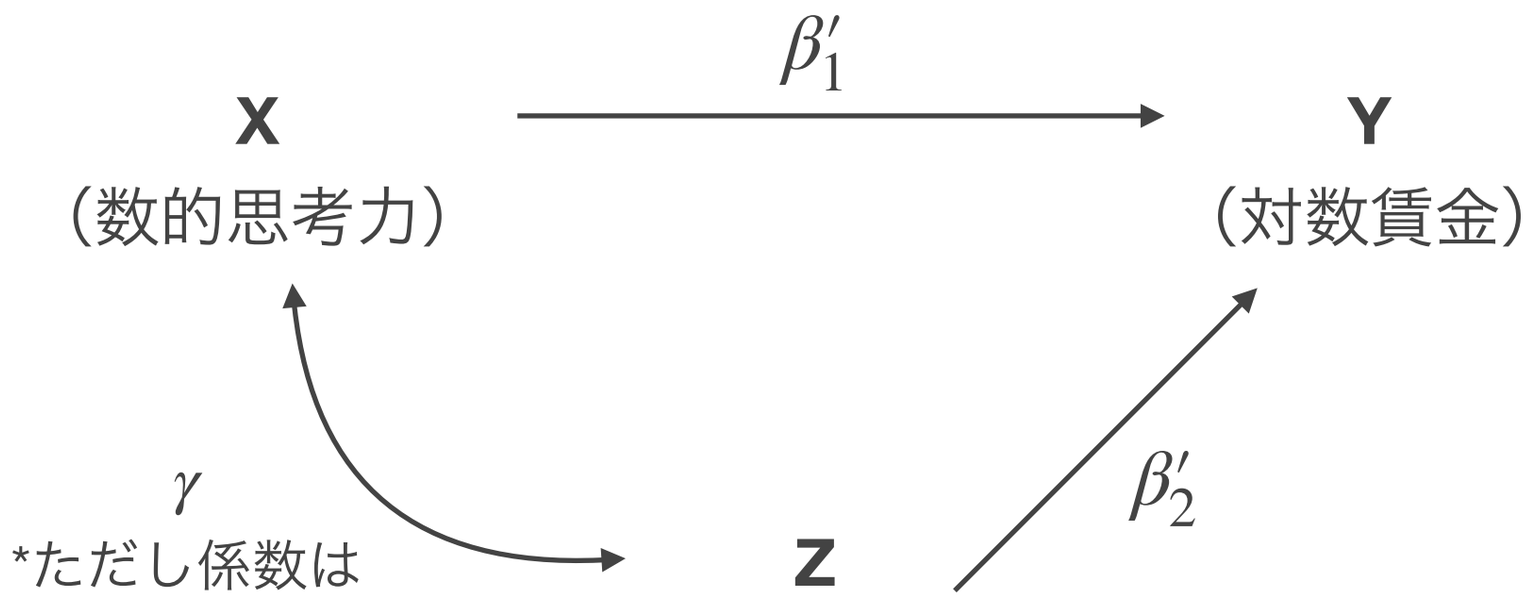
この場合も、対数オッズ比だけでなく、確率がどの程度異なるかを平均限界効果を用いてチェックするとよい。

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**。

どうすればいい？

- 線形確率モデルを使う
- それぞれのモデルについてAMEを計算する
- Imai-KeeleやKarlson-Holm-Breenの要因分解法を使う

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.

Hicks, Raymond, and Dustin Tingley. 2011. “Causal Mediation Analysis.” *The Stata Journal* 11(4):605–19.

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

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

参考）Peduzzi, Peter, John Concato, Elizabeth Kemper, Theodore R. Holford, and Alvan R. Feinstem. 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	
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、係数が異常に大きくなる。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、以下のように定義される：$\text{Pseudo } R^2 = 1 - \frac{\log L(M_{full})}{\log L(M_{intercept})}$

疑似決定係数にはいろいろな種類があり、まれにCox & Snell's R^2やNagelkerke's R^2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と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より大きいのに、Xの平均限界効果を計算すると、AよりもBのほうが大きいという逆転現象 flipped-sign phenomenon が起こることがある。交互作用項を含めたモデルでも同様。

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（Bloome & Ang, 2022）

Bloome, D., & Ang, S. (2022). Is the Effect Larger in Group A or B? It Depends: Understanding Results From Nonlinear Probability Models. *Demography*, *59*(4), 1459–1488.

さらなる学習のために

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/

線形回帰分析の基礎

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

田中隆一，2015, 『計量経済学の第一歩：実証分析のススメ』有斐閣.

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.

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.

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わる諸問題」『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/

Morgan, Stephan and Christopher Winchip. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research, 2nd Edition.* Cambridge University Press.（落海浩訳，2022，『反事実と因果推論』朝倉書店.）

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

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

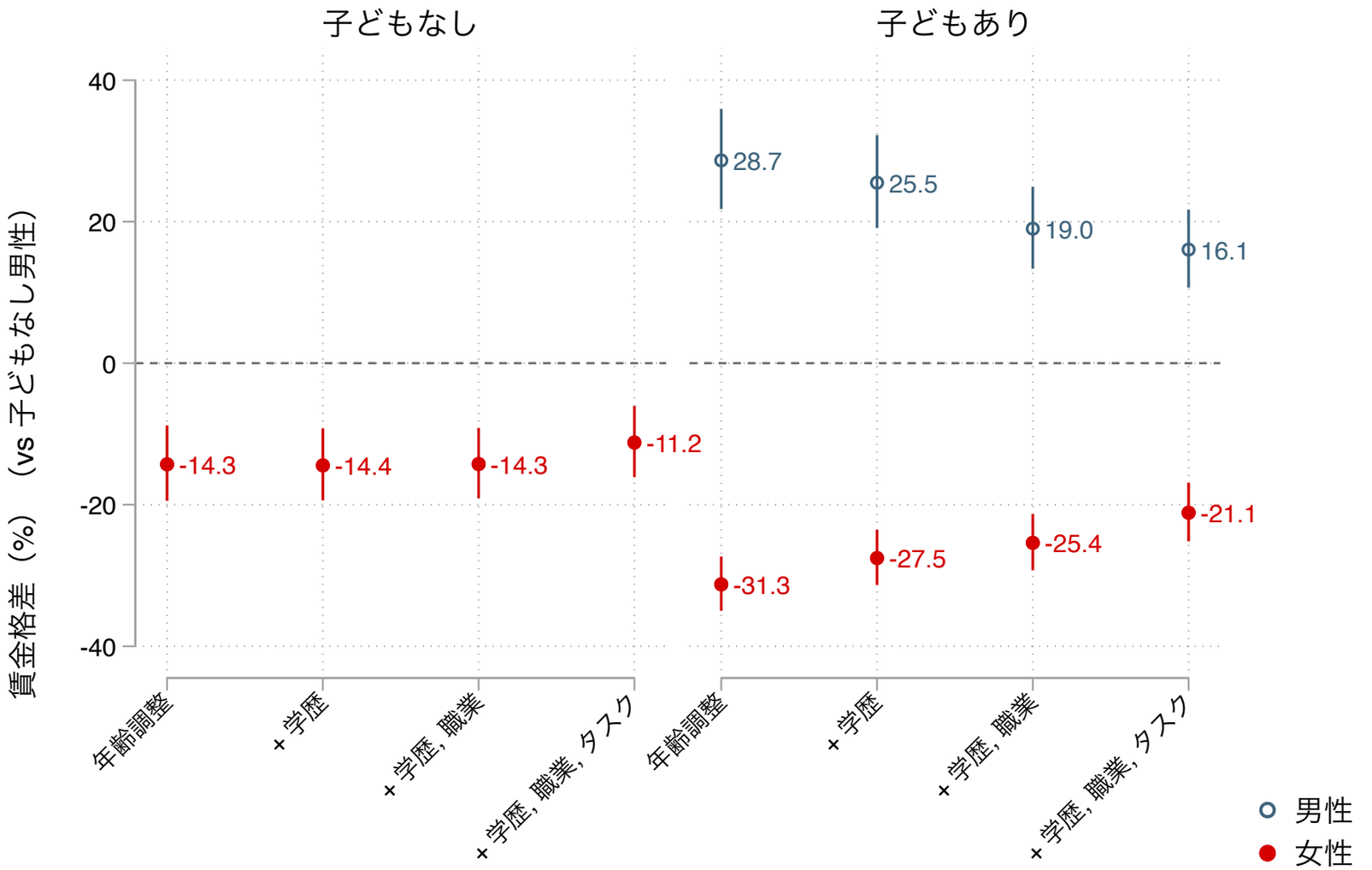
Atom　https://atom.io/

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？

セミナー後の学習用コード

6_advanced2023-09-05.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、変数を作成、保存
4. 2つのデータをappendで結合

7_loop_macro2023-09-05.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で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）

使い分けかた

- 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

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のうち、回帰分析までの内容を扱っています