

Stataによる計量分析の実践

麦山 亮太 Ryota MUGIYAMA

学習院大学法学部政治学科

ryota.mugiyama@gakushuin.ac.jp

2021/8/31 2021年度CSRDA計量分析セミナー

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

目次

Stataの基礎とプロジェクト管理

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解釈を深める

学習のための参考文献

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

よくわからない分析

「回帰分析はなんだかたくさんの独立変数を入れたほうがよさそう」

「回帰分析をすると因果関係がわかる／回帰分析では因果関係はわからないが、xxx分析（任意の"モダンな"分析手法の名前を入れる）を使うと因果関係がわかる」

「決定係数は高いほどよい／高すぎるとよくない」

「従属変数が2値のときはロジスティック回帰分析をしないといけないらしい」



何が何だかわからない人のイラスト

出所) https://www.irasutoya.com/2019/01/blog-post_148.html

つらい作業

Stataでグループごとに平均値を計算→結果をExcelに貼り付け→Excelで棒グラフを作成→グラフをコピー→Wordに貼り付け→データを修正したので再度平均値を計算→Excelに貼り付け→Excelで棒グラフを作成→グラフをコピー.....

Stataで回帰分析を実施→結果をExcelに貼り付け→変数の名前を1つひとつ入力→有意かどうかを示す*印をp値をみてつけていく→Wordに結果を貼り付け→データを修正したので再度回帰分析を実施→Excelに貼り付け→*印を.....



忙しく仕事をしている会社員のイラスト

出所) https://www.irasutoya.com/2014/10/blog-post_35.html

よりよい計量分析の実践のために

よくわからない分析をよくわかる分析にすることで.....

- より信頼できる結果を提示できる
- よりインパクトがあり、見る側にとってわかりやすい結果を提示できる

つらい作業を減らすことで.....

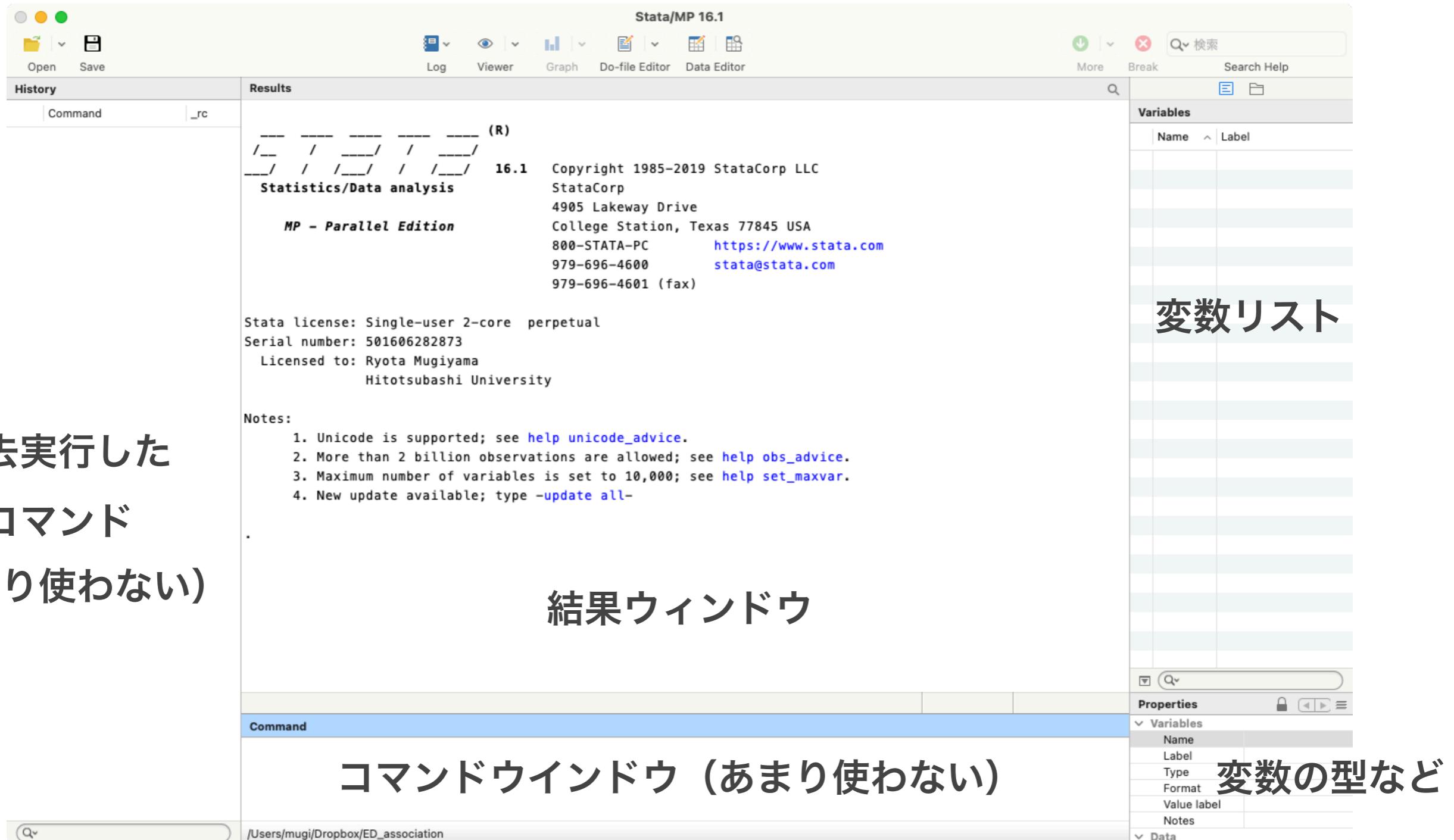
- 本質的なことを考える時間ができる
- ストレスが減る

分析手法を適切に理解し、Stataをうまく使えば、よくわかる分析に近づき、つらい作業を減らすことができる！

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



設定の変更

EditまたはStata/MP 16.1 → Preferences → General Preferences

→ Syntax highlight でコードハイライトの色を変更できる

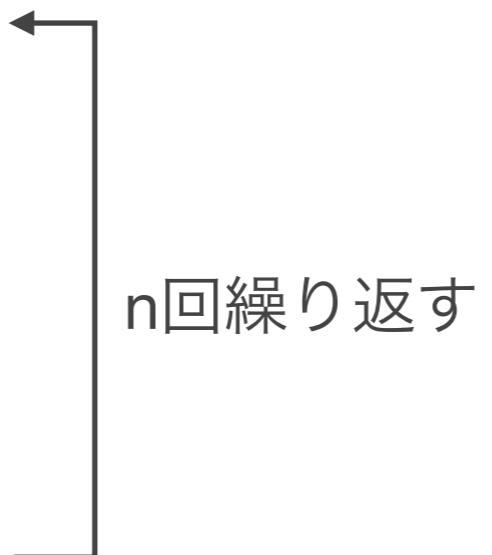
→ Windows で結果ウインドウやその他の箇所の色を変えられる

EditまたはStata/MP 16.1 → Preferences → User-interface language で言語
を変更できる

自分好みの設定にすると愛着が湧いてやる気が出るかも？

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
5. 加工したデータを保存
6. 加工したデータを分析
7. 分析結果の出力



プロジェクトフォルダの構成の例

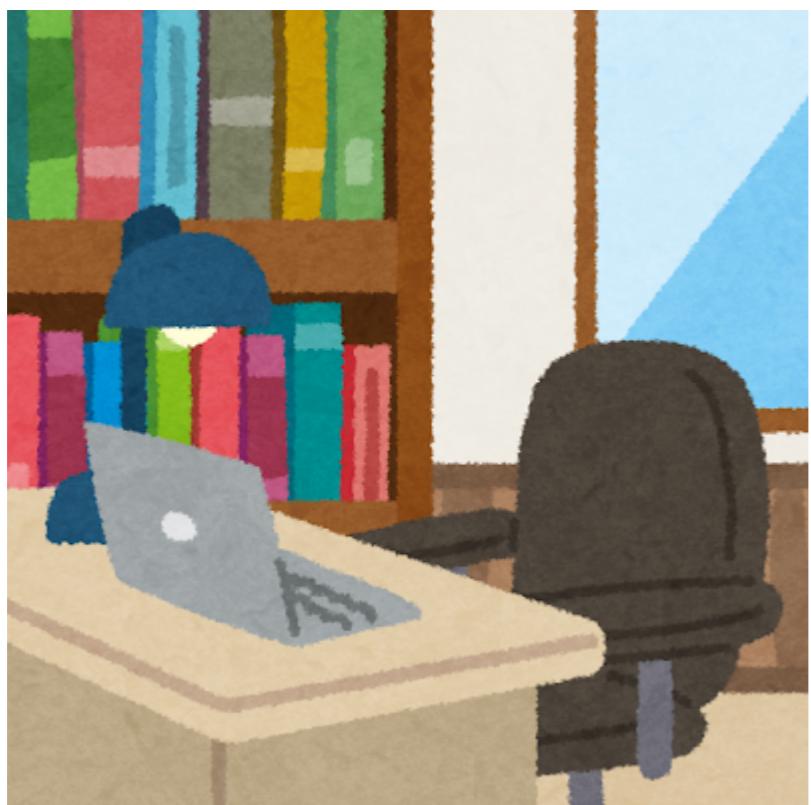
- project : あるプロジェクトに関するファイルをすべて入れる
- code : データの加工・分析に使用するコードを入れる
- codebook : データのコードブックを入れる
- data : 分析に使用するデータを入れる
- manuscript : 論文などの原稿を入れる
- presentation : 学会報告などで使用するスライドを入れる
- results : 分析の出力結果を入れる
- submission : 投稿したときのファイル、査読コメント・リプライ原稿などを入れる

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で、変更できる。

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。

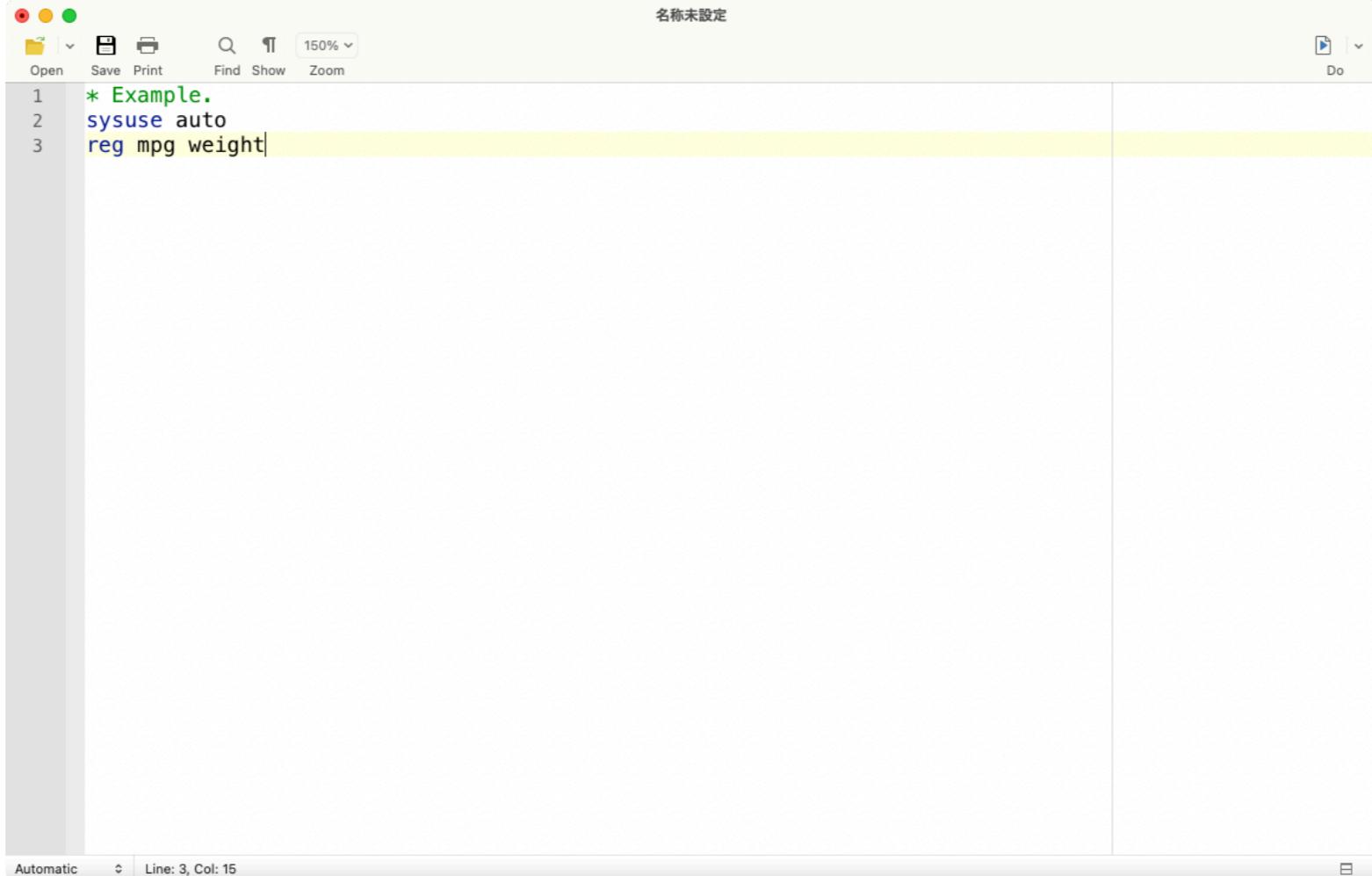


作業ディレクトリのイメージ？

出所) https://www.irasutoya.com/2013/11/blog-post_4497.html

doファイル

コマンドウィンドウに doedit と入力して実行 (Enter)



The screenshot shows a Stata doedit window titled "名称未設定". The window contains a script with the following content:

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

The window has a menu bar with "Open", "Save", "Print", "Find", "Show", and "Zoom" options. On the right side, there is a "Do" button. At the bottom, it shows "Automatic" and "Line: 3, Col: 15".

doファイルにコマンドを書いて実行するのが基本。

doファイルを保存する

doファイル上部の「Save」をクリック → (はじめてのときは) 名前をつけて保存。

保存先は、「code」フォルダとする。

doファイルの命名規則

2021-08-31statSeminar.do

handling2021-08-31.do

など、何に関する・いつのファイルなのか分かる名前がおすすめ。

もう一つの方法：プロジェクトの作成

File → New → Project... を選択し、先ほどの「project」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：
 _project_2021statSeminar.stpr

Windowsユーザ：doファイルを開く → File → New → Project... を選択

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。今回の授業で使うものは以下：

`estout`

`coefplot`

`descstable`

`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_install2021-08-31.doを開き、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

OECDが実施している国際成人力調査（Programme for the International Assessment of Adult Competencies, PIAAC）のデータを使ってみよう

The screenshot shows the homepage of the OECD Skills Surveys website. At the top, there is the OECD logo and a search bar. Below the logo, a blue banner reads "OECD Skills Surveys". The main navigation menu includes links for HOME, ABOUT, PIAAC DESIGN, EVENTS, DATA, PUBLICATIONS, and ONLINE ASSESSMENT. A sub-menu for "Survey of Adult Skills (PIAAC)" is visible under the "ABOUT" link. On the left, there is a section titled "About PIAAC" which provides an overview of the survey. On the right, there is a section titled "PIAAC Round 3 International Launch Webinar" featuring a chart titled "Effect of education, numeracy proficiency and numeracy use at work on wages". The chart compares data across various countries.

[Survey of Adult Skills \(PIAAC\)](https://www.oecd.org/skills/piaac/)

About PIAAC

The Programme for the International Assessment of Adult Competencies (PIAAC) is a programme of assessment and analysis of adult skills. The major survey conducted as part of PIAAC is the **Survey of Adult Skills**. The Survey measures adults' proficiency in key information-processing skills - literacy, numeracy and problem solving - and gathers information and data on how adults use their skills at home, at work and in the wider community.

This international survey is conducted in over 40 countries/economies and measures the key cognitive and workplace skills needed for individuals to participate in society and for economies to prosper.

[Learn more about how we measure and collect data.](#)

PIAAC Round 3 International Launch Webinar

The Survey of Adult Skills

Effect of education, numeracy proficiency and numeracy use at work on wages

Percentage change in wages associated with a one standard deviation increase in years of education, proficiency in numeracy and numeracy use at work

Years of education Proficiency (numeracy) Numeracy at work

Statistically significant differences are marked in a darker tone

OECD

見る YouTube

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

データを開き、中身を確認する

1_handling2021-08-31.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、プロジェクトフォルダ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。

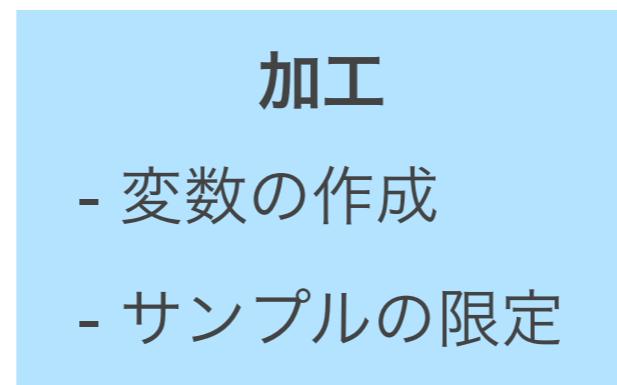
../というふうにすると、プロジェクトフォルダから1つ上の階層を指定することができる。例：use "../project_usa/data/piaacusa.dta", clear

データを加工して、加工後のデータを保存する

1_handling2021-08-31.doの残りの行を実行し、データを保存しよう

save "data/piaacjpn-analysis.dta", replaceで、名前をつけて
データを保存

	x1	x2	x3
1			
2	元々のデータ		
3	piaacjpn.dta		
...			



	x4	x5	x6
1			
2	加工後のデータ		
3	piaacjpn-analysis.dta		
...			

絶対にやるべきでないコードの書き方の例

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

```
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

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, 2011) 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）データの加工をしてはいけない

0_install2021-08-31.do

1_handling2021-08-31.do

データの加工

2_descriptive2021-08-31.do

データの分析

3_regression2021-08-31.do

4_logit2021-08-31.do

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効

logファイル

logファイルは普段見直すことはほとんどないが、分析の過程を見せてくださいと要求されたときに「研究ノート」としての役割を果たす

```
log using "...", replace
```

```
code ...
```

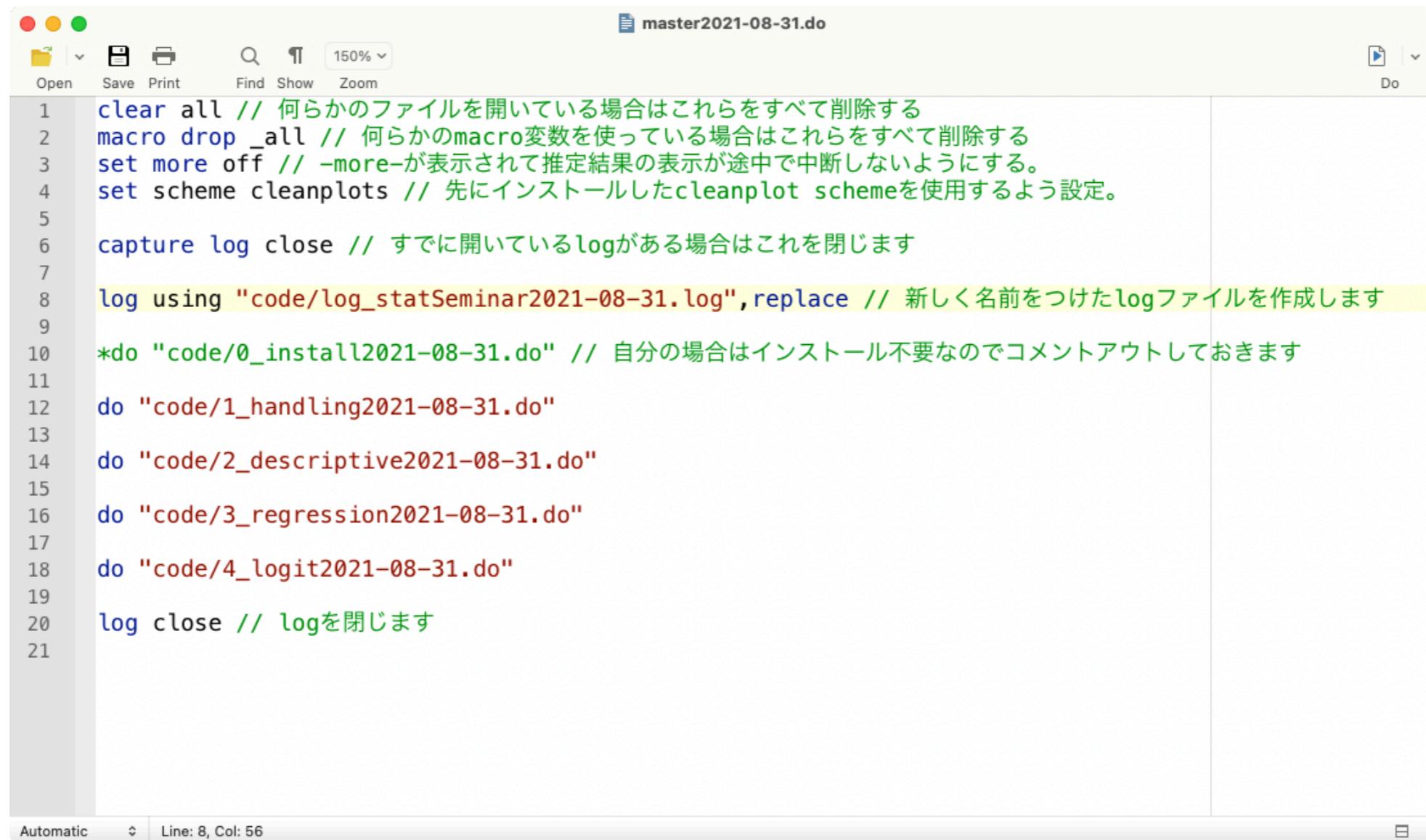
```
code ...
```

```
code ...
```

```
log close
```

Master do-fileから個別のdoファイルを実行する

master2021-08-31.doを開いて中身を確認してみよう



The screenshot shows a Stata do-file editor window titled "master2021-08-31.do". The window contains the following code:

```
1 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
2 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
3 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。
4 set scheme cleanplots // 先にインストールしたcleanplot schemeを使用するよう設定。
5
6 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
7
8 log using "code/log_statSeminar2021-08-31.log", replace // 新しく名前をつけたlogファイルを作成します
9
10 *do "code/0_install2021-08-31.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
11
12 do "code/1_handling2021-08-31.do"
13
14 do "code/2_descriptive2021-08-31.do"
15
16 do "code/3_regression2021-08-31.do"
17
18 do "code/4_logit2021-08-31.do"
19
20 log close // logを閉じます
21
```

The status bar at the bottom indicates "Automatic" mode, line 8, column 56.

わかりやすいコードを書くための一般的な注意点

変数の名前は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。全角文字はおすすめしない

変数には必ず変数ラベルをつける

- lab var educ "学歴"

カテゴリ変数の値には必ず値ラベルをつける

- lab def educlab 1 "中学" 2 "高校" 3 "短大高専" 4 "大学"

こまめにやっている作業のメモを残す

- *や//、/**/などでコメントアウト

doファイルを分割する

- 長すぎるdoファイルは見づらいしどこに何があるかも分からなくなる

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？

2_descriptive2021-08-31.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 (2.1.1)

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

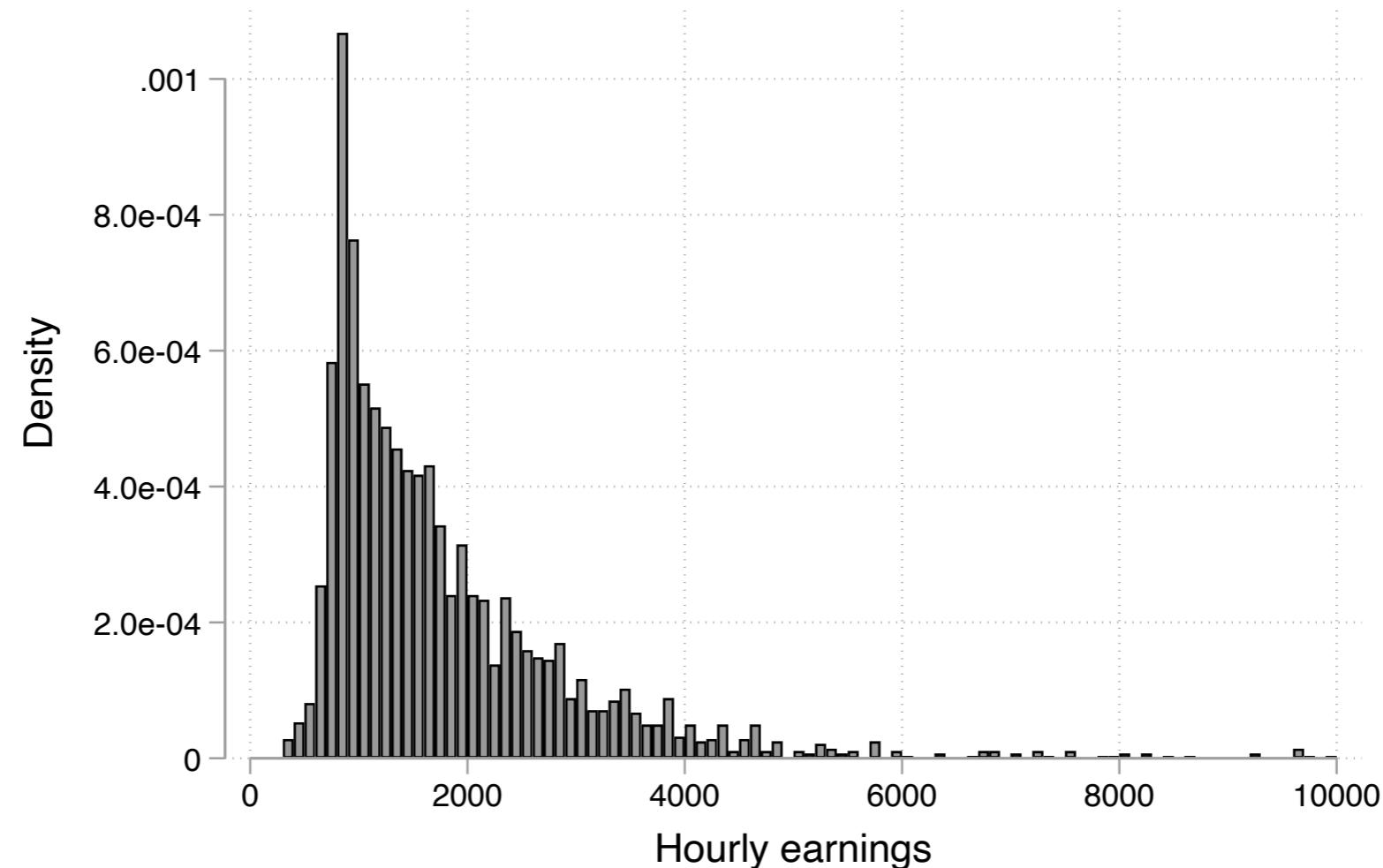
それを解決してくれる最近のブレークスルーがdesctable <https://www.trentonmize.com/software/desctable>

Table #: Descriptive Statistics (N = 3329)					
	n	Mean/Prop.	SD	Min.	Max.
Hourly earnings	2828	1784.35	1206.38	300.00	10000.00
Age	3329	44.67	11.00	25.00	64.00
Gender	3329	.45			
<i>Level of education</i>	3329				
Junior high		.09			
Senior high		.36			
Junior college		.24			
University		.31			
Numeracy score	3329	2.94	.43	1.03	4.41
Literacy score	3329	2.99	.39	1.26	4.13

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 (2.1.2)

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 (2.1.3)

2変数関係

1変数分布

Y?

Yはどのような分布？

平均や中央値はどれくらい？

2変数関係

X → Y

Yの分布はXによってどれくらい違う？

Yの平均値はXごとにどれくらい違う？

変数の分布や、その集計（平均値etc）を異なるグループごとに比べることで、**比較の問い合わせ**に答えることができる。

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？

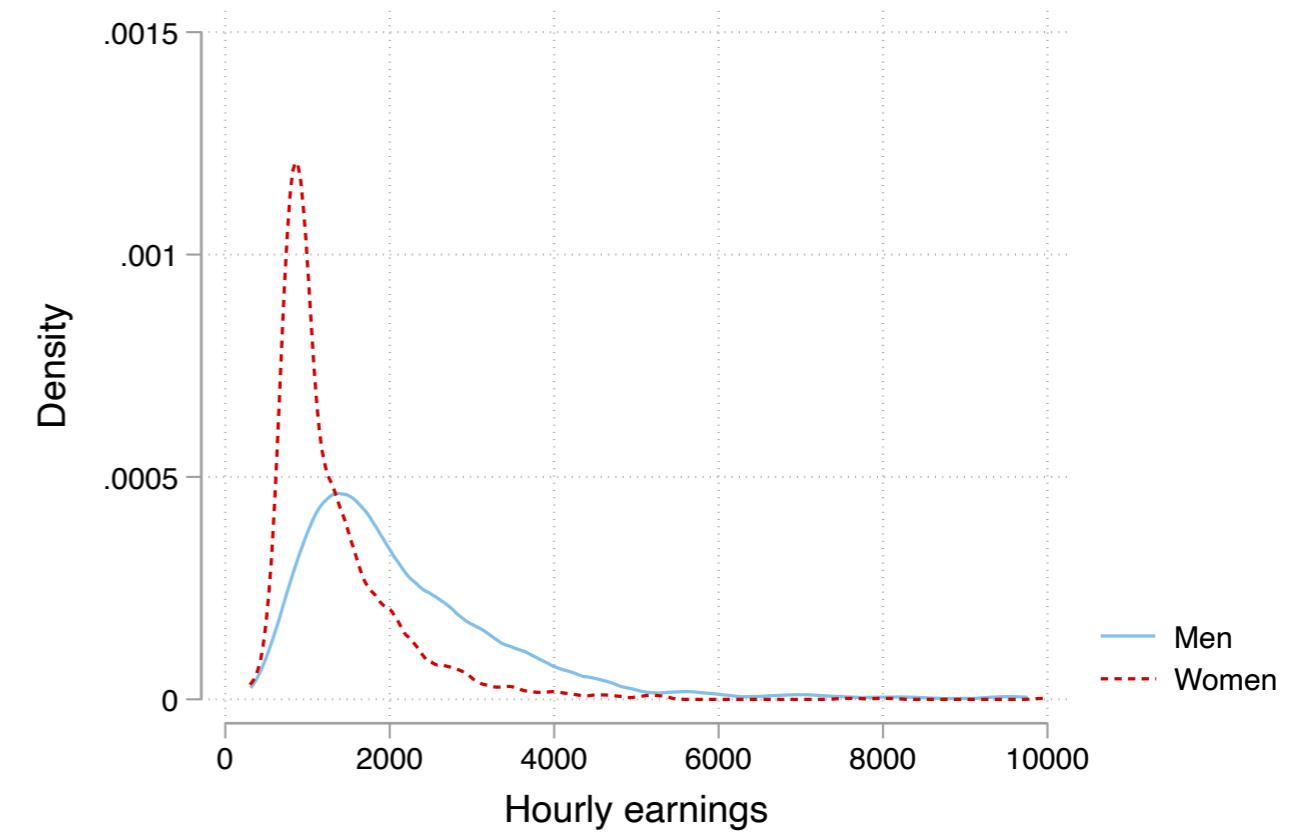
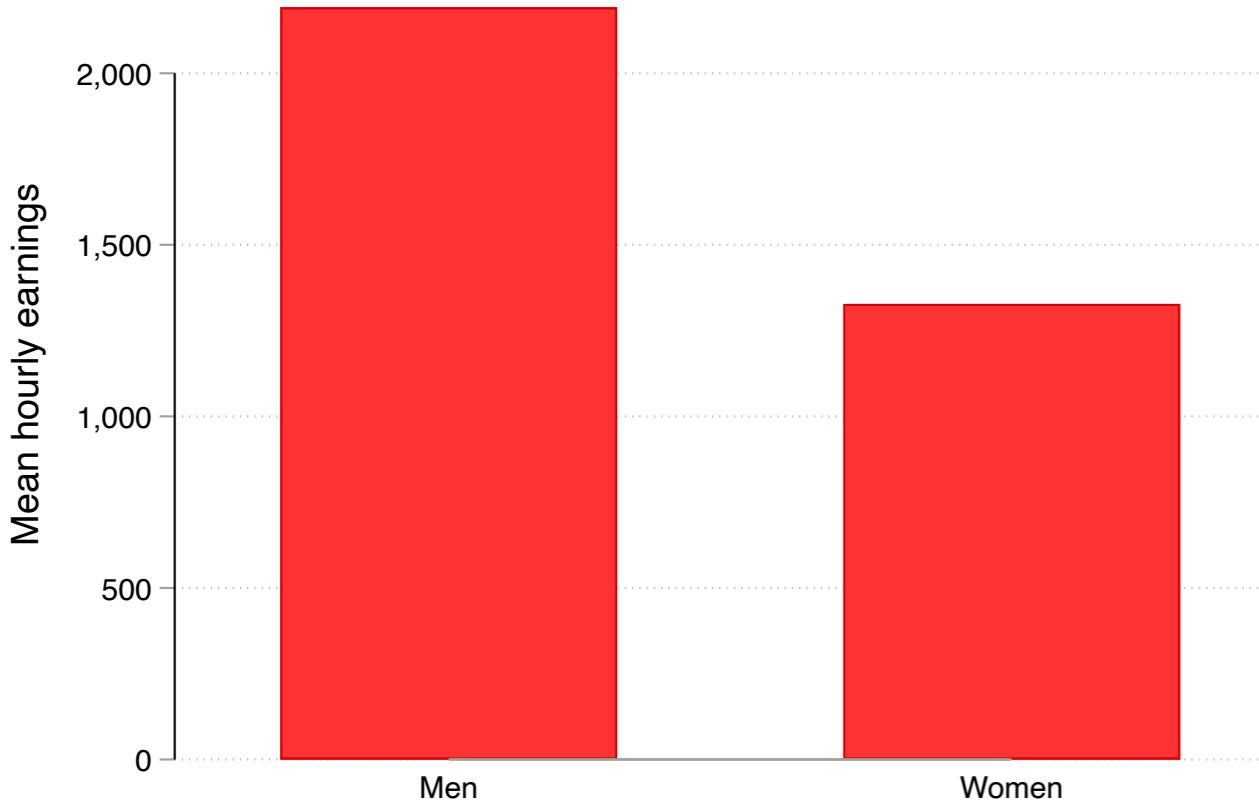
tabstatコマンドを使ってグループ別の集計をしてみよう (2.2.1)

desctableコマンドを使ってグループ別の集計をしてみよう (2.2.2)

より効果的なプレゼンテーション

Stataを使って楽しておしゃれなグラフを作ろう

グループ別平均値の棒グラフ、箱ひげ図、カーネル密度グラフを作成してみよう (2.2.3)



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）を受けるかどうかはどれくらい違うだろうか？

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（2.3.1）

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。

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（後述）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

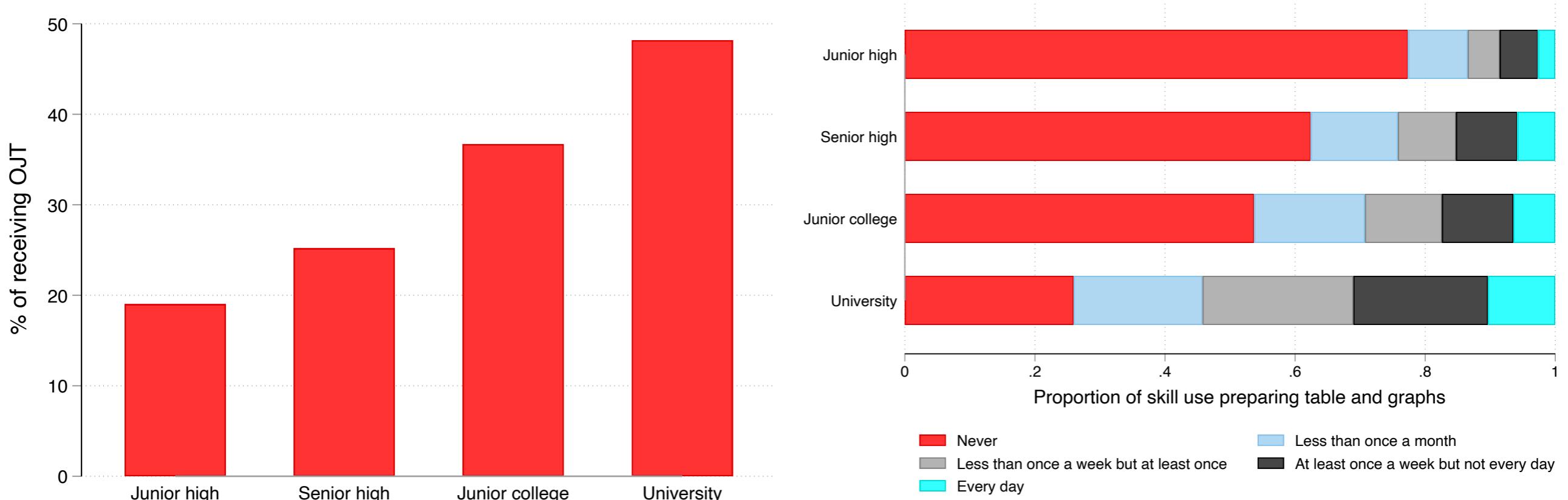
	0	1	Total
	b/rowpct	b/rowpct	b/rowpct
Junior high	247	58	305
	81.0	19.0	100.0
Senior high	888	299	1187
	74.8	25.2	100.0
Junior college	502	291	793
	63.3	36.7	100.0
University	541	503	1044
	51.8	48.2	100.0
Total	2178	1151	3329
	65.4	34.6	100.0

esttabは回帰分析の表を作ると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

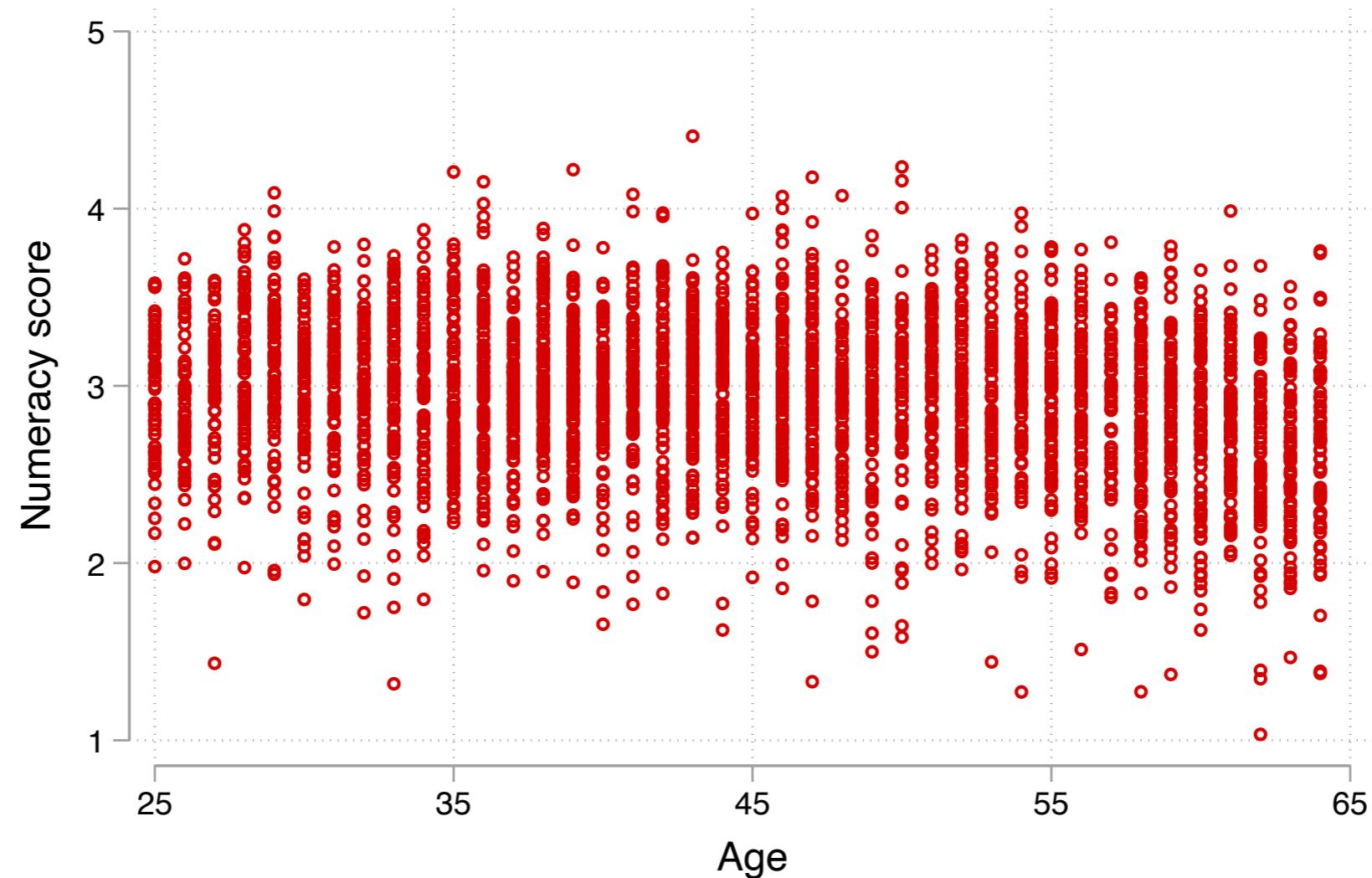
列変数（分布を示したい変数Y）が2値の場合と、列変数が多値の場合で少し効果的な示し方が異なる

列変数が2値のときと、列変数が多値のときのクロス集計表を棒グラフで表してみよう（2.3.2）



散布図と相関係数

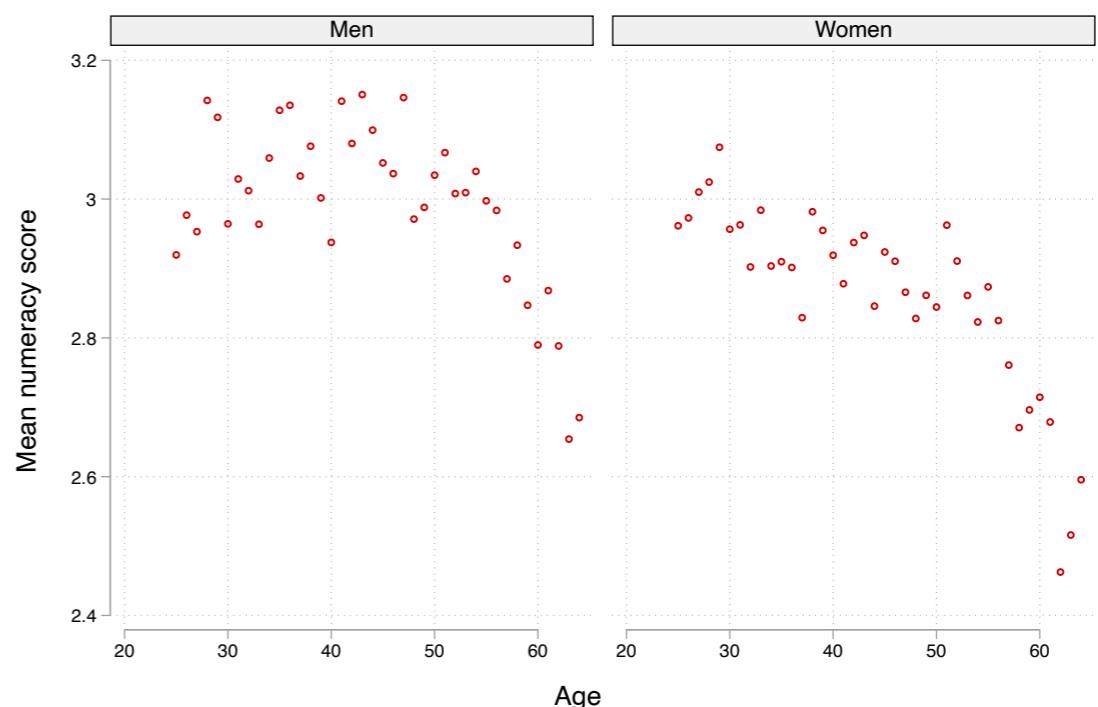
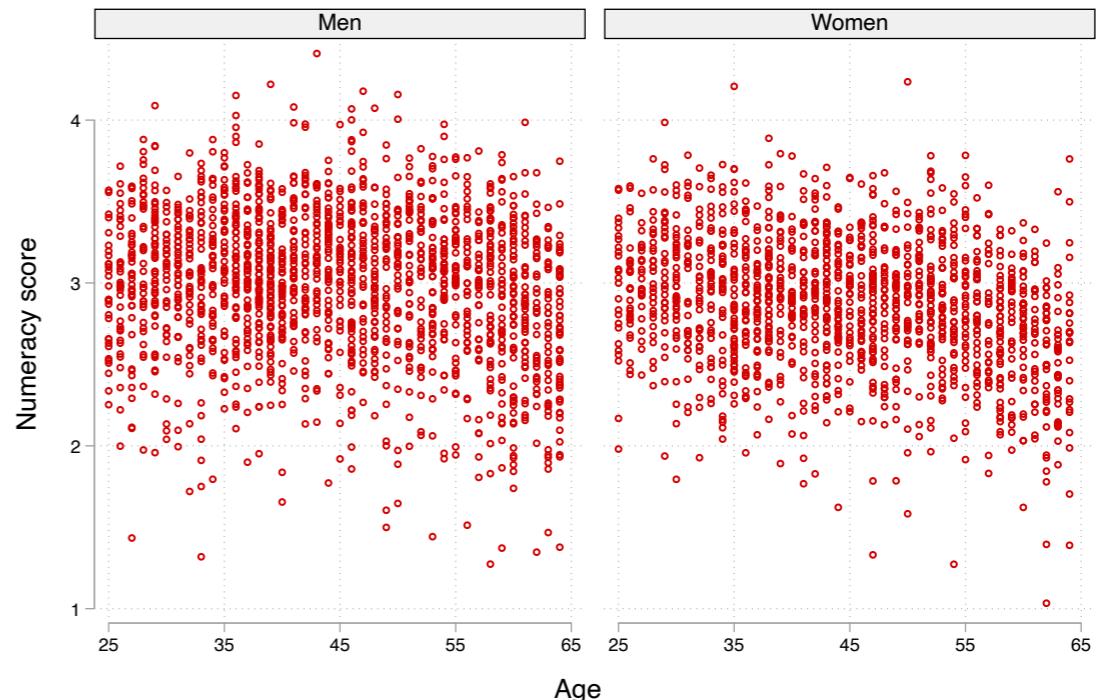
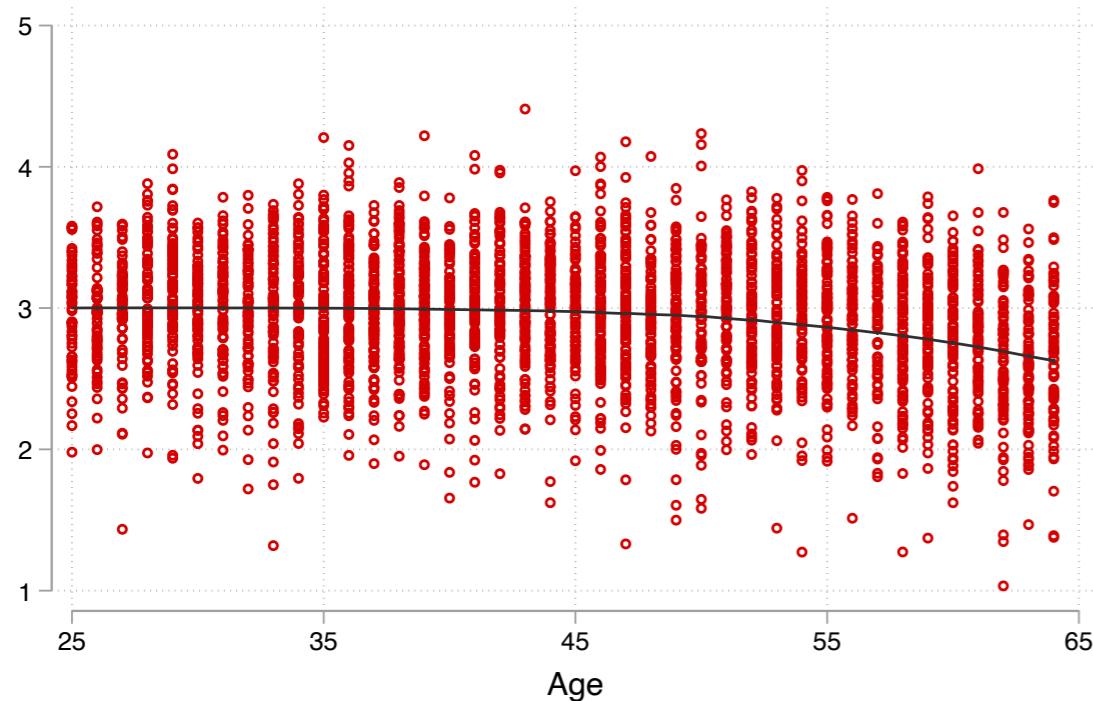
連續変数の値ごとに連續変数の値を比較する場合には、相関係数を計算したり、散布図を作成するのがよい。



相関係数を計算し、散布図を作成してみよう (2.4.1)

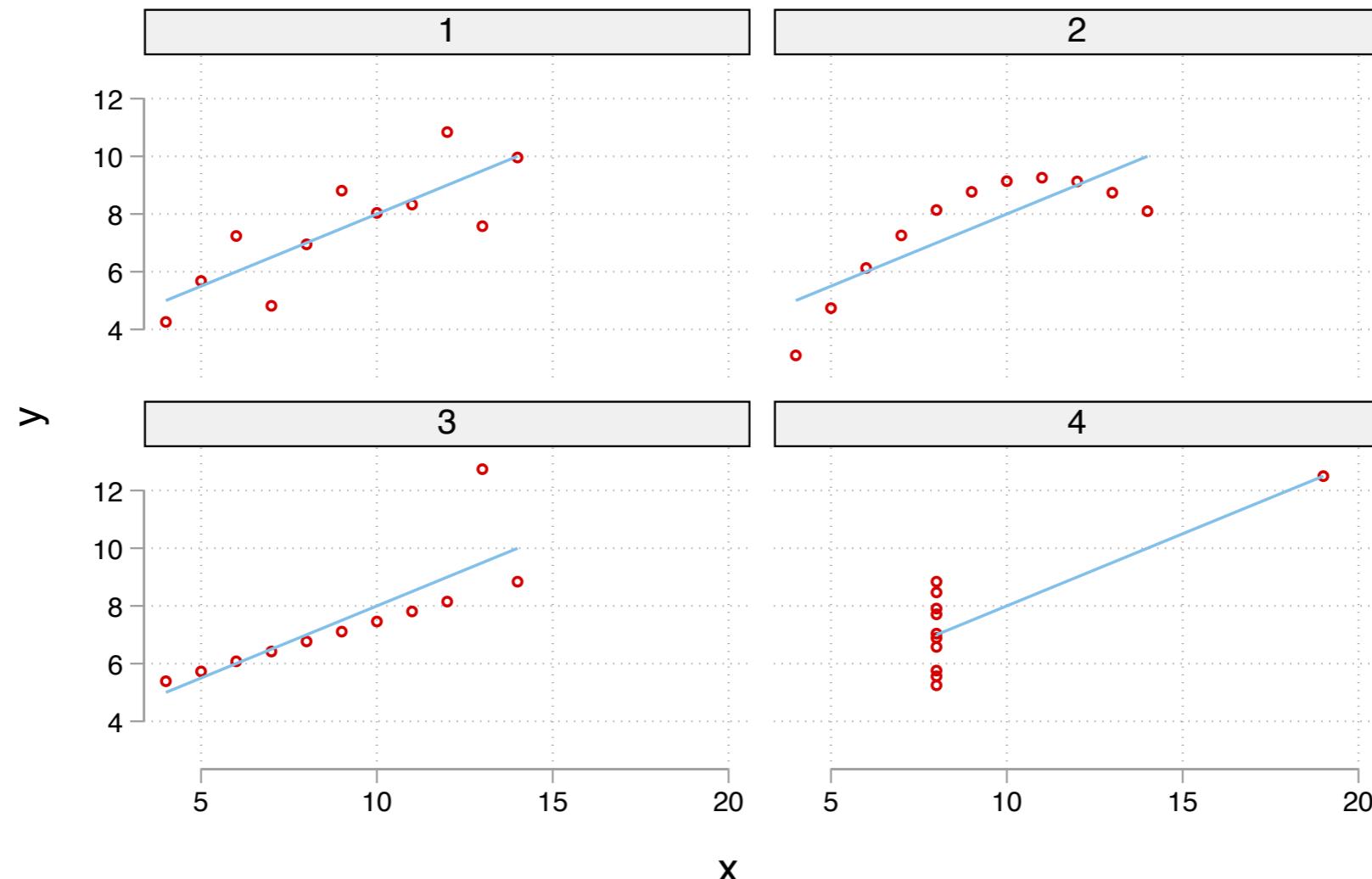
散布図から変数間の関係性を探索する

いろいろなパターンの散布図を作成してみよう (2.4.2)



相関係数だけみるのはあぶない：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 (2.4.3)

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：

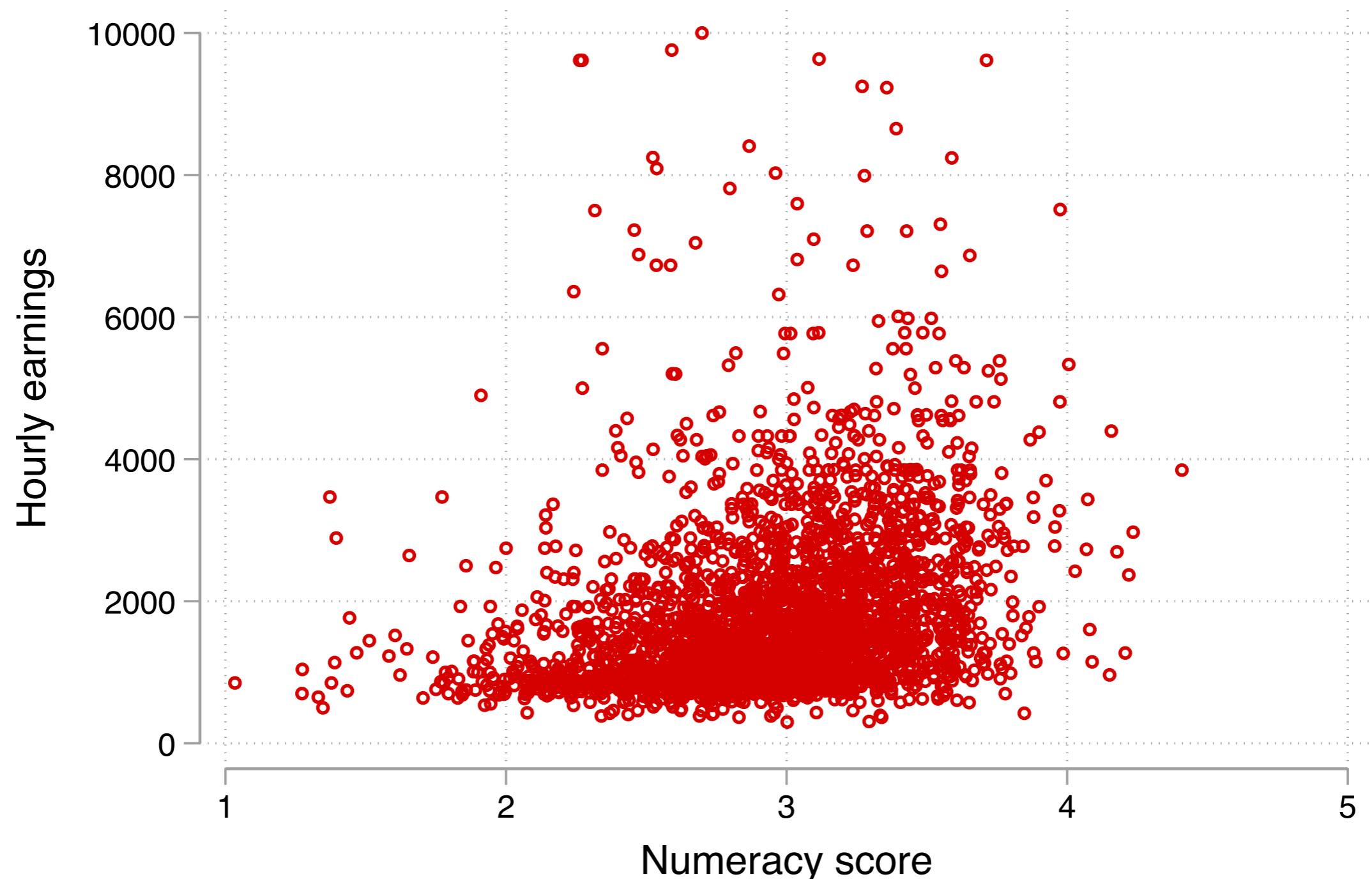
川口大司, 2017, 『労働経済学：理論と実証をつなぐ』有斐閣.

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.

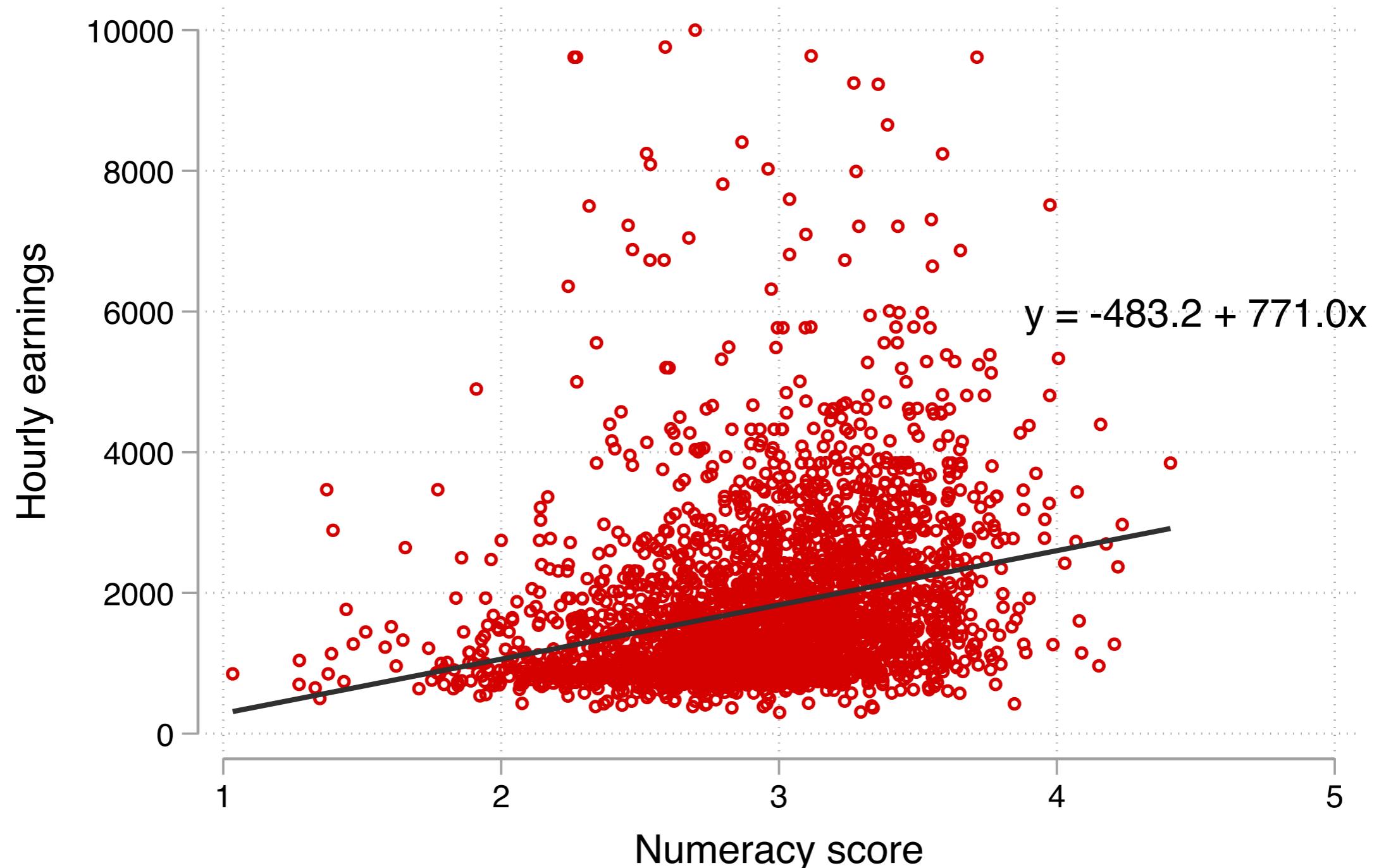
Hanushek, Eric A., Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann. 2015. "Returns to Skills around the World: Evidence from PIAAC." *European Economic Review* 73:103–30.

散布図を描いてみる



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7.7円高い



線形回帰分析

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 linear regression analysis** 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することができる：

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

*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

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

X が1単位増加したときの Y の增加分を ΔY とおく。

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

$$\Delta Y = \beta_1$$

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。

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

Stataでの回帰分析の出力結果

3_regression2021-08-31.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（3.1.1）

. regress earnhrbonus numeracy

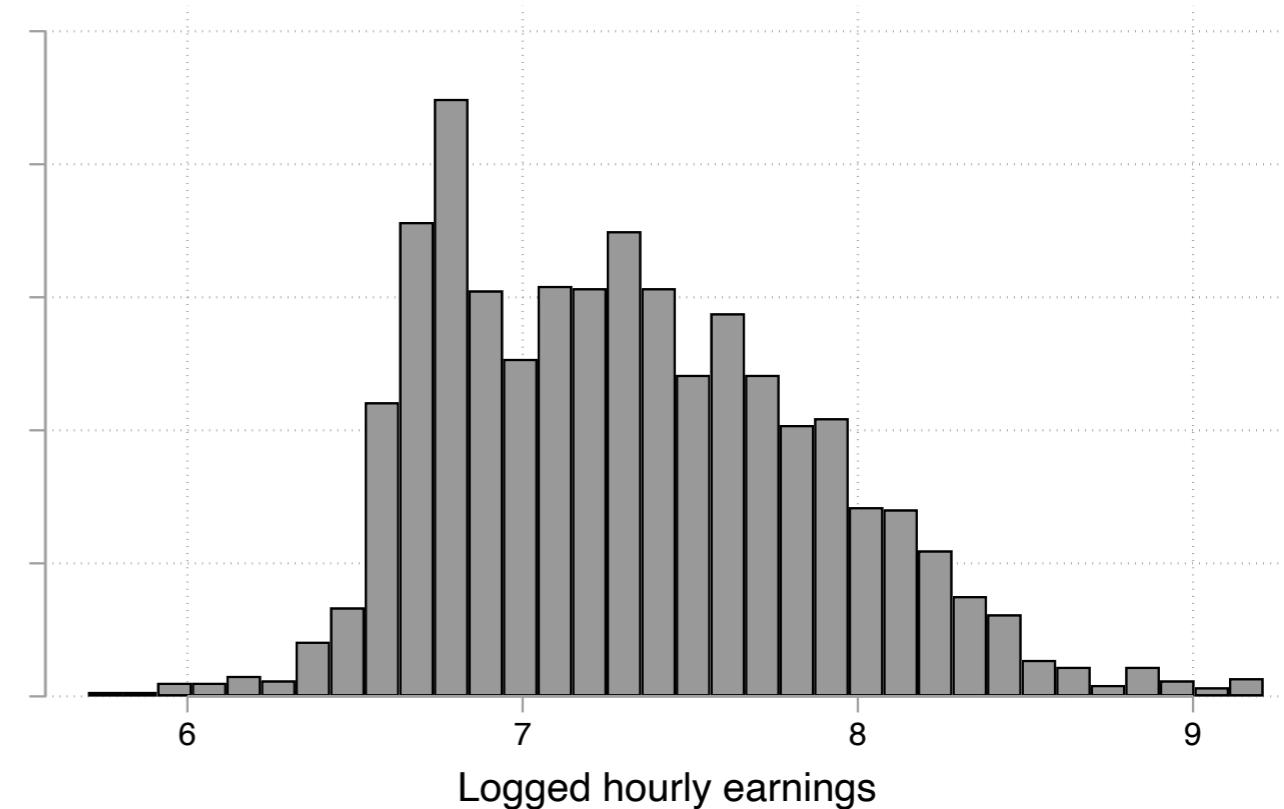
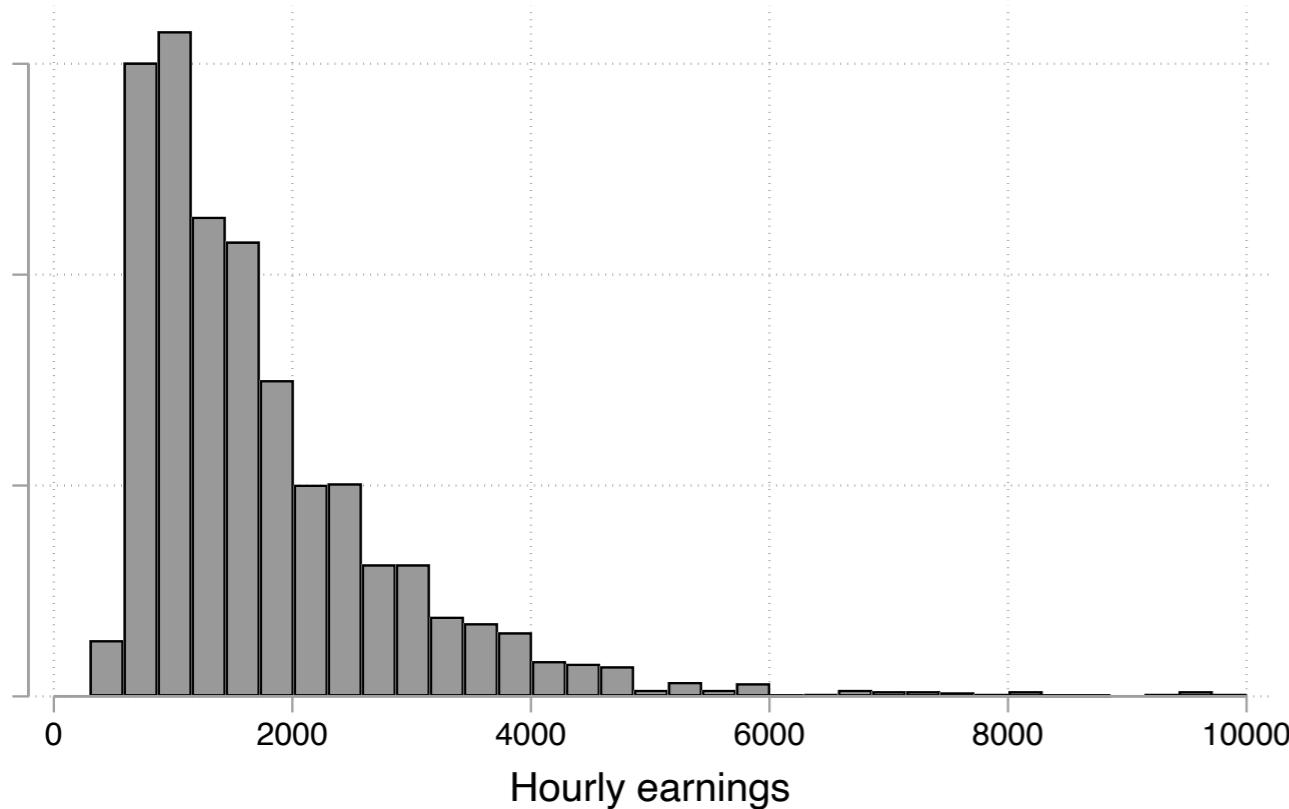
Source	SS	df	MS	Number of obs	=	2,828
Model	316276501	1	316276501	F(1, 2826)	=	235.33
Residual	3.7980e+09	2,826	1343943.61	Prob > F	=	0.0000
Total	4.1143e+09	2,827	1455345.29	R-squared	=	0.0769
				Adj R-squared	=	0.0765
				Root MSE	=	1159.3

	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
earnhrbonus	7.709565	.5025588	15.34	0.000	6.724145	8.694984
_cons	-483.1706	149.4107	-3.23	0.001	-776.1357	-190.2056

変数の対数変換

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。

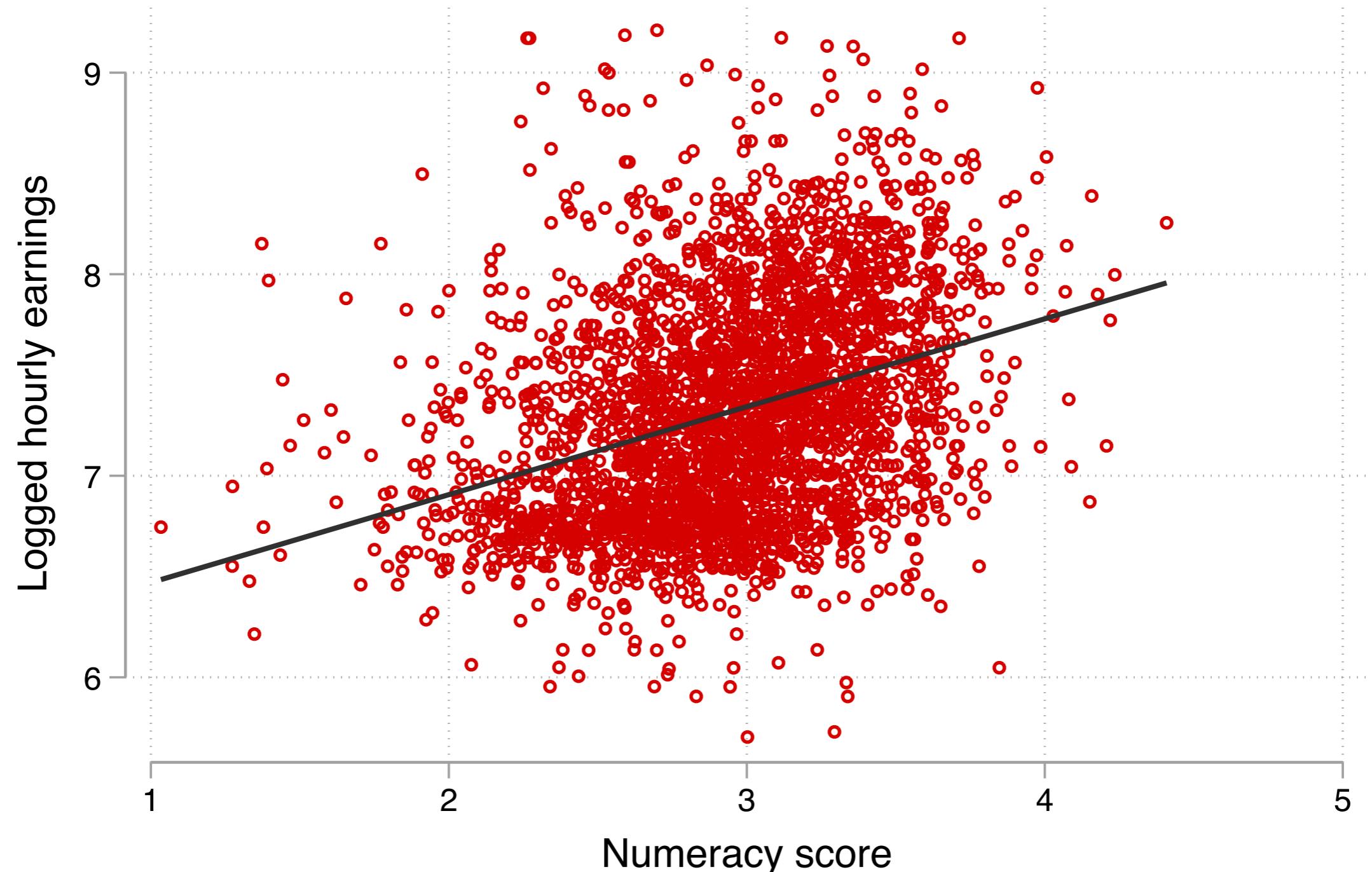
ネイピア数は以下のような便利な性質を持つ。

- 指数の微分： $[\exp(x)]' = \exp(x)$

- 自然対数の微分： $(\ln x)' = 1/x$

自然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

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

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

β_1 が0に近い値のときは、おおむね「Xが1単位増加するとYは $\beta_1 \times 100\%$ 増加する」ことを表す：

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

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

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

*より正確な値を知りたいときは $\exp(\beta_1) - 1$ を計算する。

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

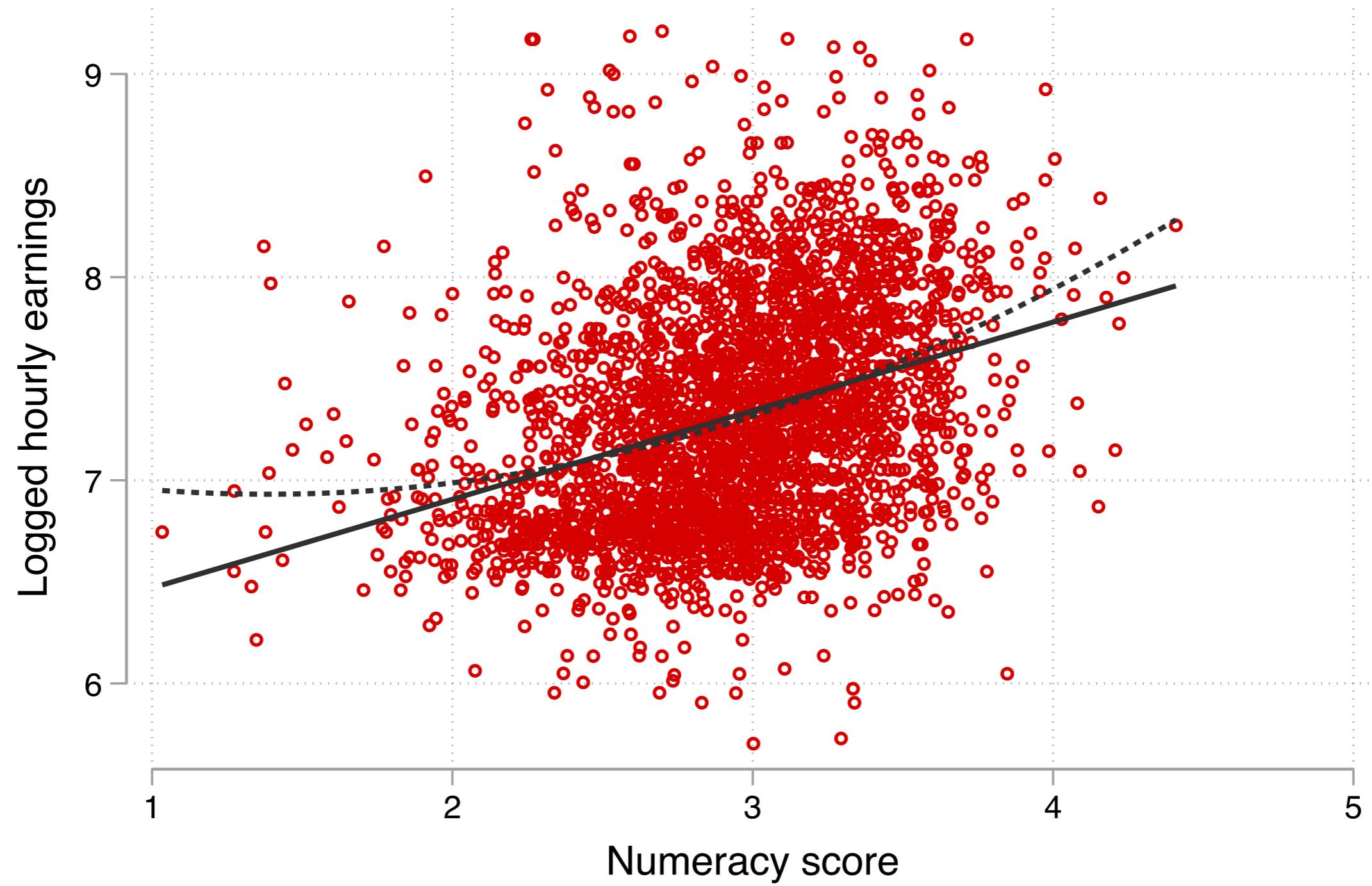
$\log(Y)$ X X が1単位高いと、Yが $100 \times \beta_1\%$ 高い

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

$$\begin{aligned}Y + \Delta Y &= \beta_0 + \beta_1(X + 1) + \beta_2(X + 1)^2 + \varepsilon \\&= (\beta_0 + \beta_1X + \beta_2X^2) + \beta_1 + (2X + 1)\beta_2 \\ \Delta Y &= \beta_1 + (2X + 1)\beta_2\end{aligned}$$

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（3.1.2）

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

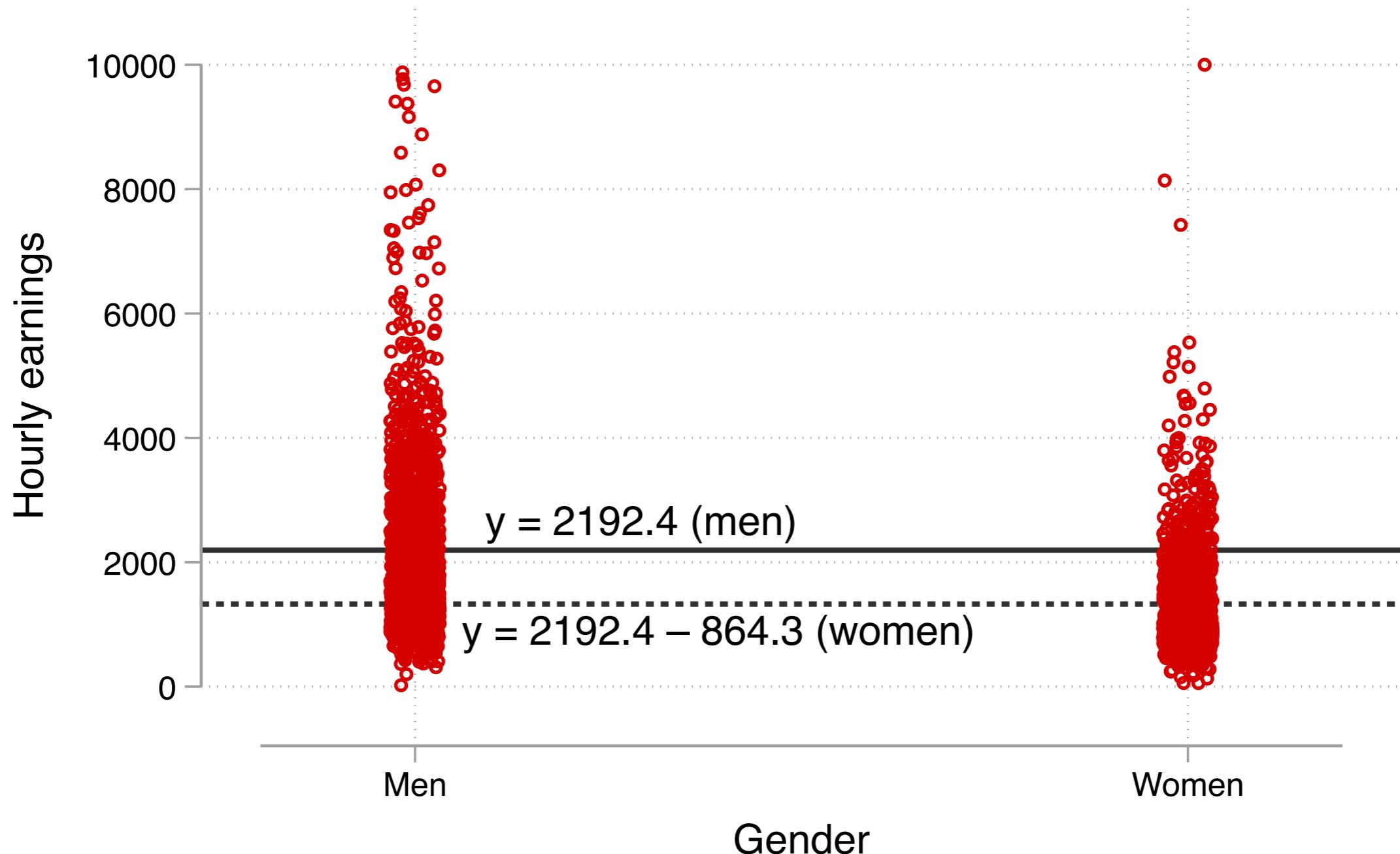
対数変換しない賃金を使ったときの結果、対数変換した賃金を使った結果、
2乗項を使った結果の3つを並べて結果を比較してみよう（3.1.3）

estoutパッケージで、こうした比較に便利な関数が提供されている。<http://repec.sowi.unibe.ch/stata/estout/esttab.html>

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。



Note: 図を見やすくするため各点にゆらぎ（jitter）を加えている

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 （ダミー変数とよぶ）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（参照カテゴリとよぶ）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： $Y = \beta_0 + \beta_1 + \varepsilon, E(Y|D=1) = \beta_0 + \beta_1$

複数の回帰分析の結果を比較する

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（3.1.4）

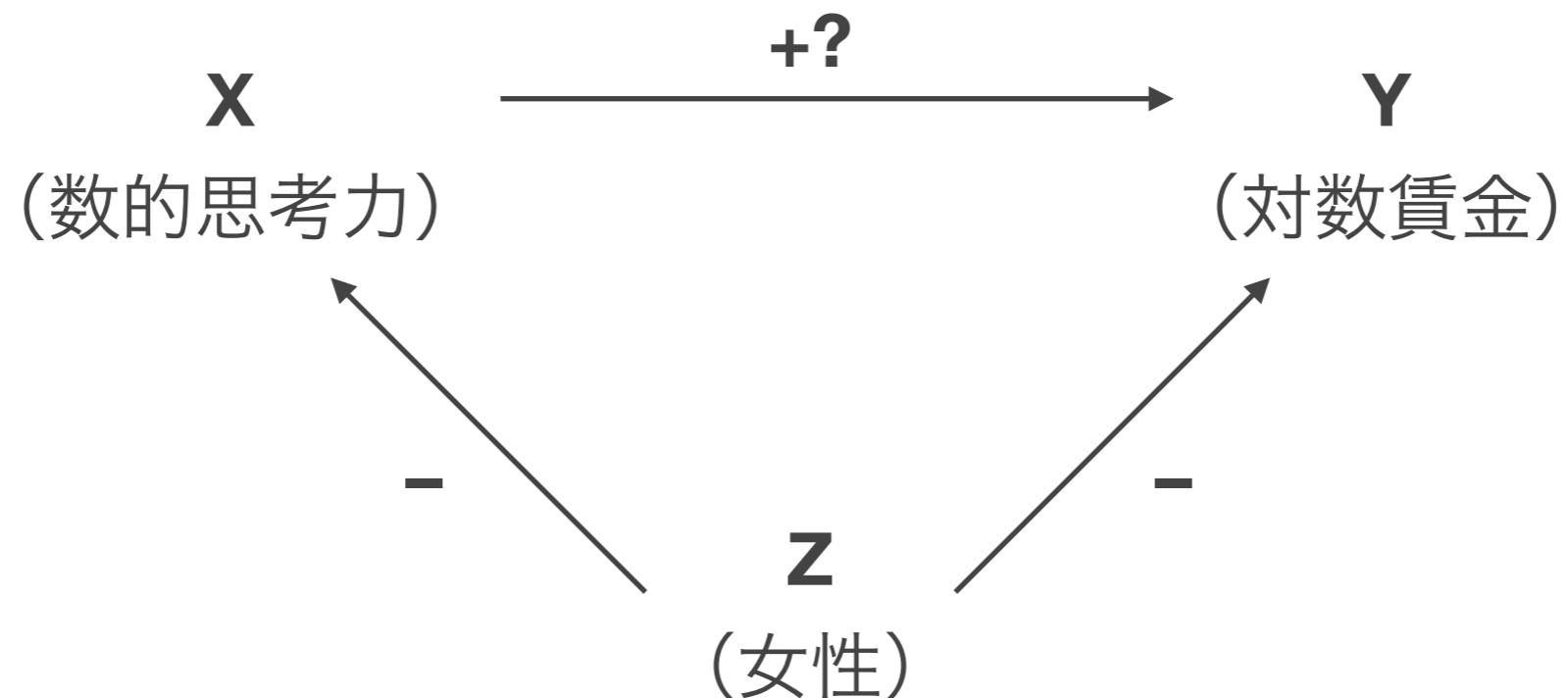
estoutパッケージで、こうした比較に便利な関数が提供されている。<http://repec.sowi.unibe.ch/stata/estout/esttab.html>

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

重回帰分析を活用する

重回帰分析による交絡要因の除去

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。



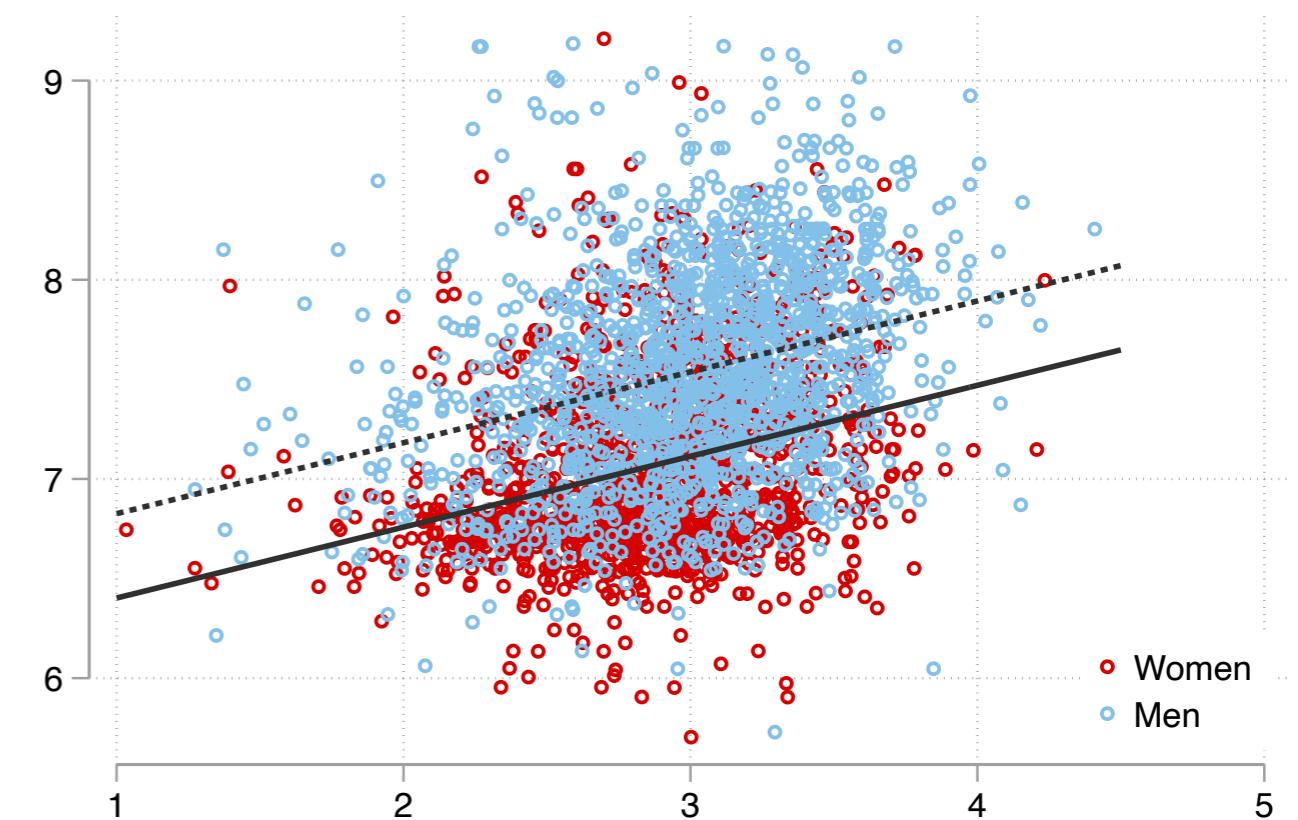
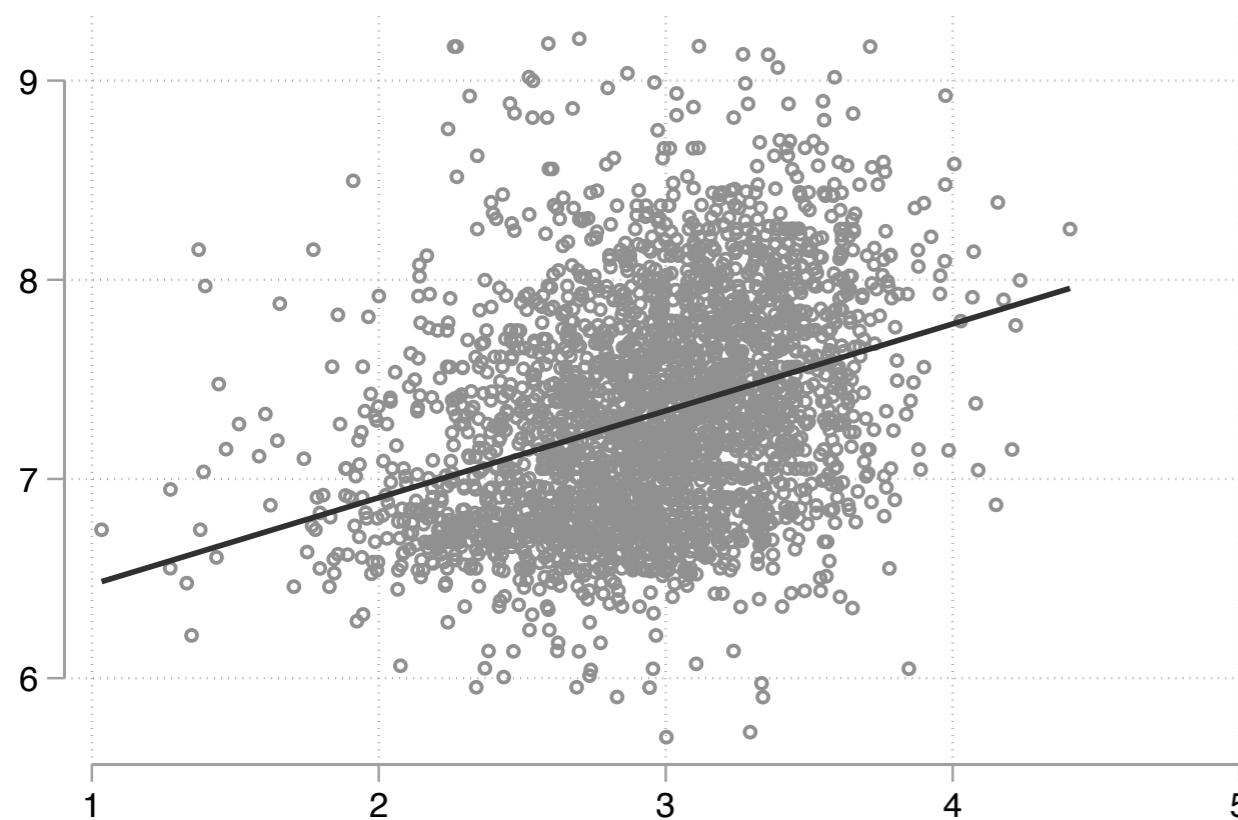
XとYの両者に影響する交絡要因Zを統制することで、Yに対するXの因果効果に近づくことができる。

単回帰分析と重回帰分析を比較してみる

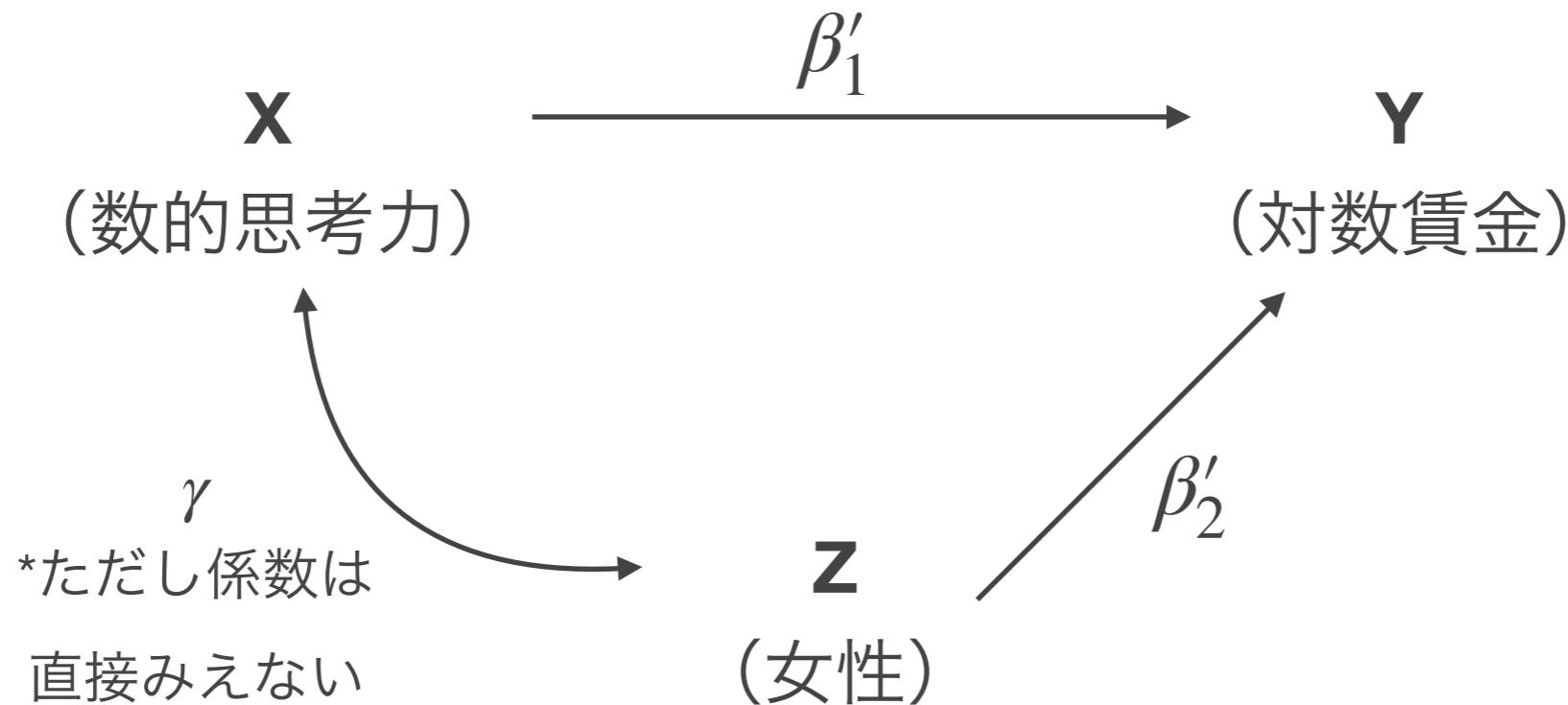
$$Y = \beta_0 + \beta_1 X + \varepsilon$$

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

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の係数は異なる。



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

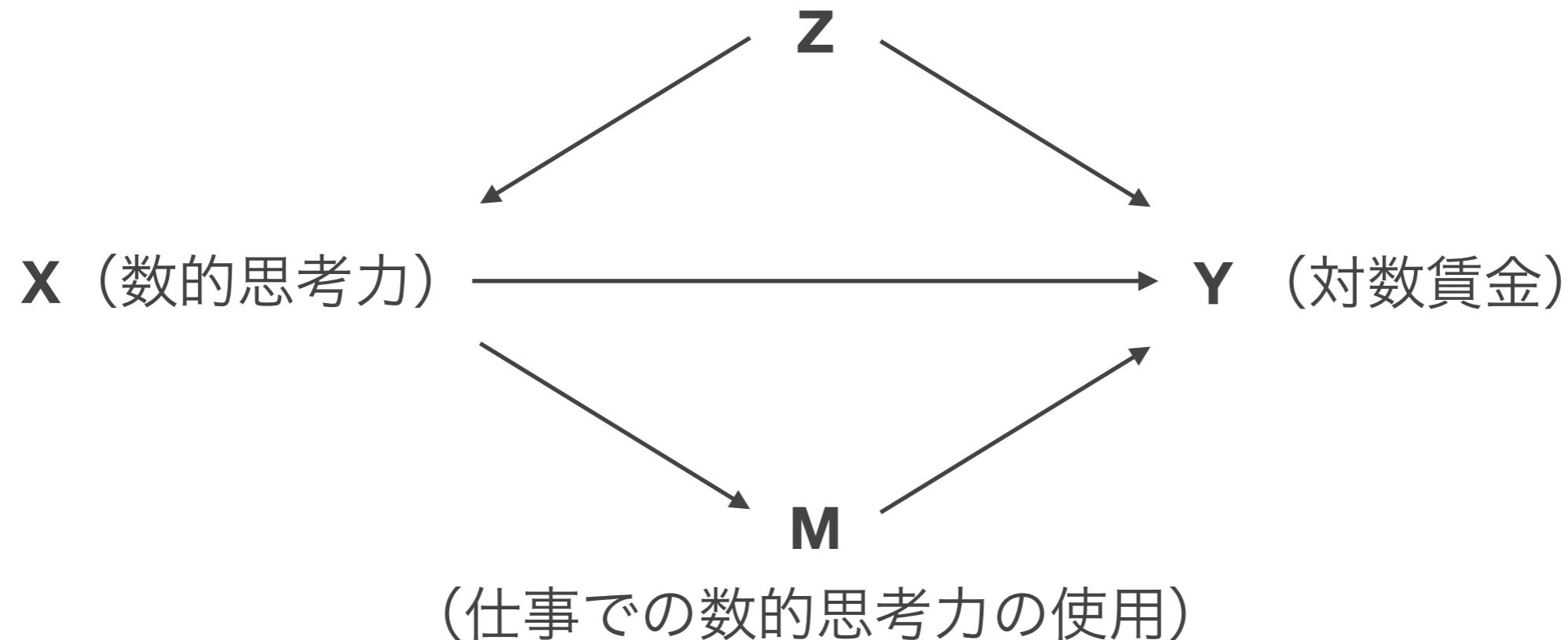
→独立変数が無作為に割り当てられていない限り、「数的思考力が高いと賃金が高くなる」といってしまうと誤り

重回帰分析から言えること：

性別が同じであったとしても、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」といってしまうと誤り（近づいてはいる）

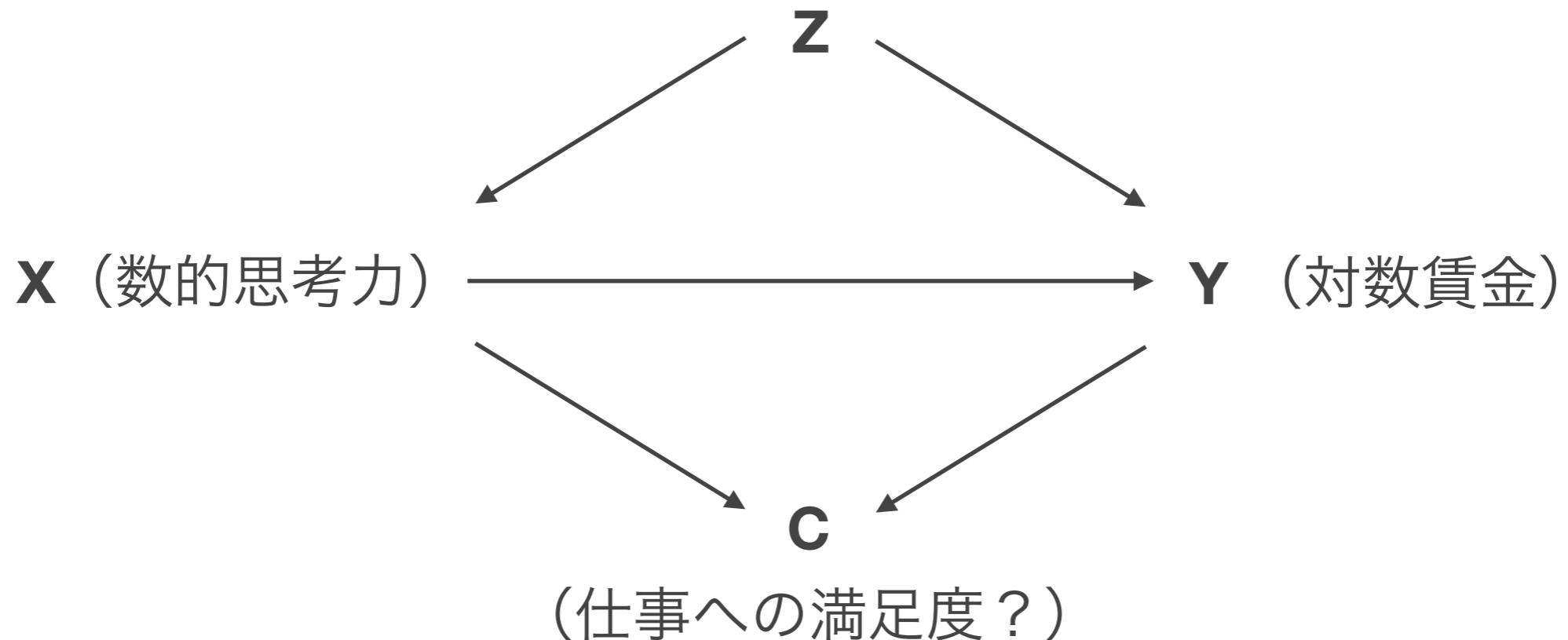
統制すべき変数を吟味する：媒介要因



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

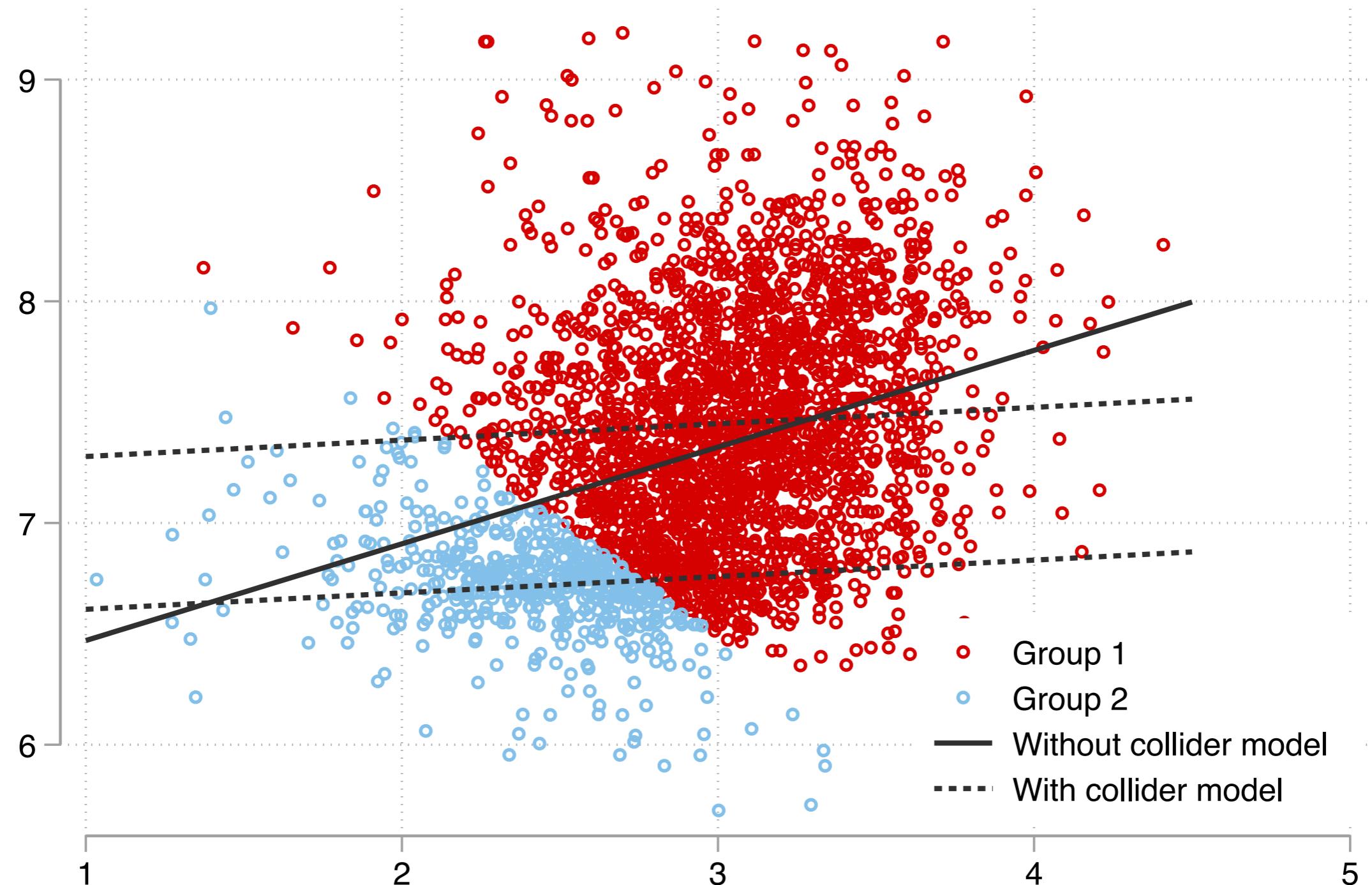


重回帰分析を因果効果を知るために使う場合、**Cのような変数は投入してはいけない**

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**や内生性バイアス **Endogeneity bias**などという (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

適切に交絡要因を統制すれば、回帰分析を使って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？

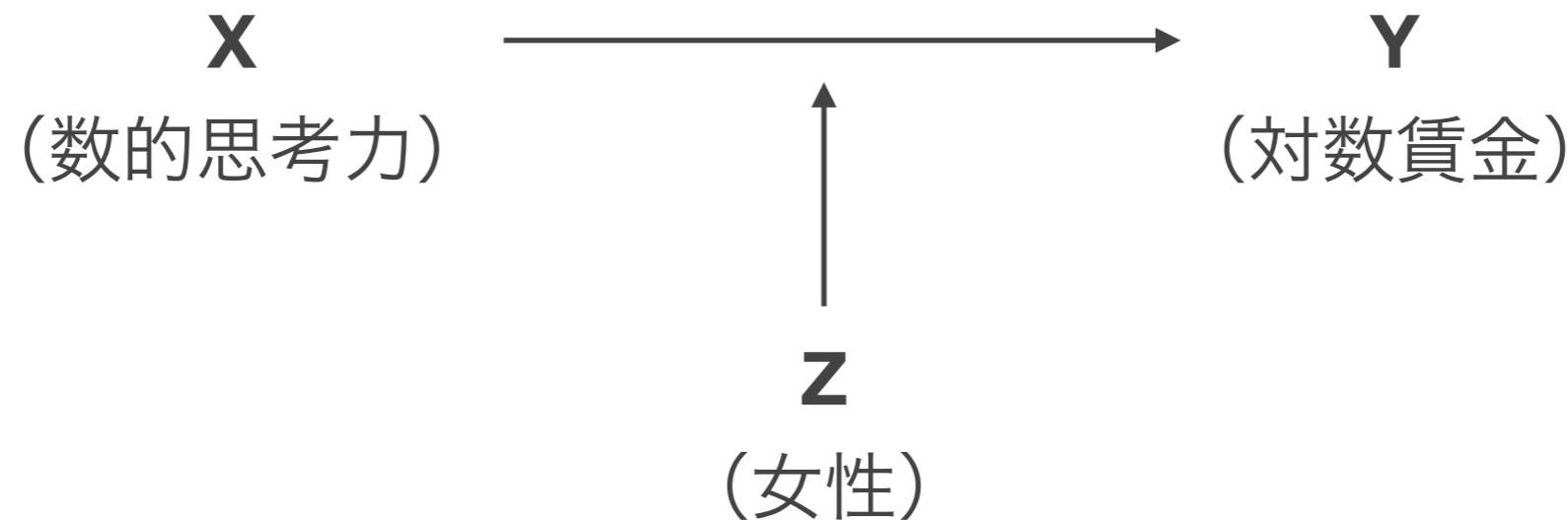
そのために、どのような交絡要因や媒介要因を統制すればよいか？

（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？

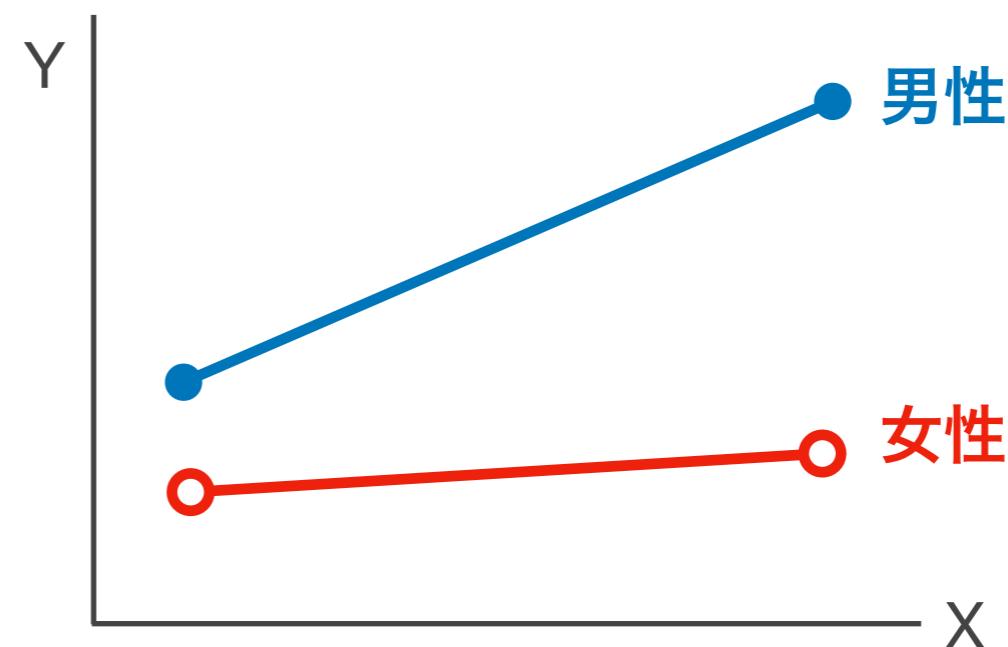
を考えることが大事

交絡要因や媒介要因、合流点（と考えられる）要因と統制した重回帰分析を推定し、結果を比較してみよう（3.2）

調整効果



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 Moderation**あるいは**交互作用効果 Interaction**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：

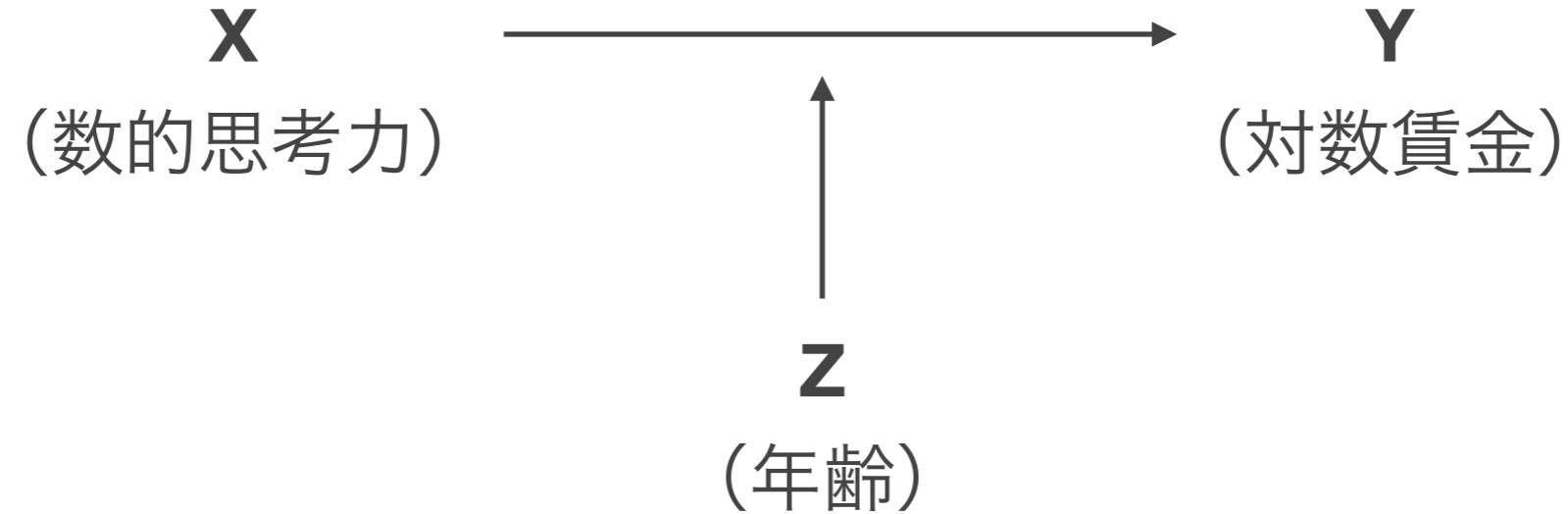
$$Y = \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon$$

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

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

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：

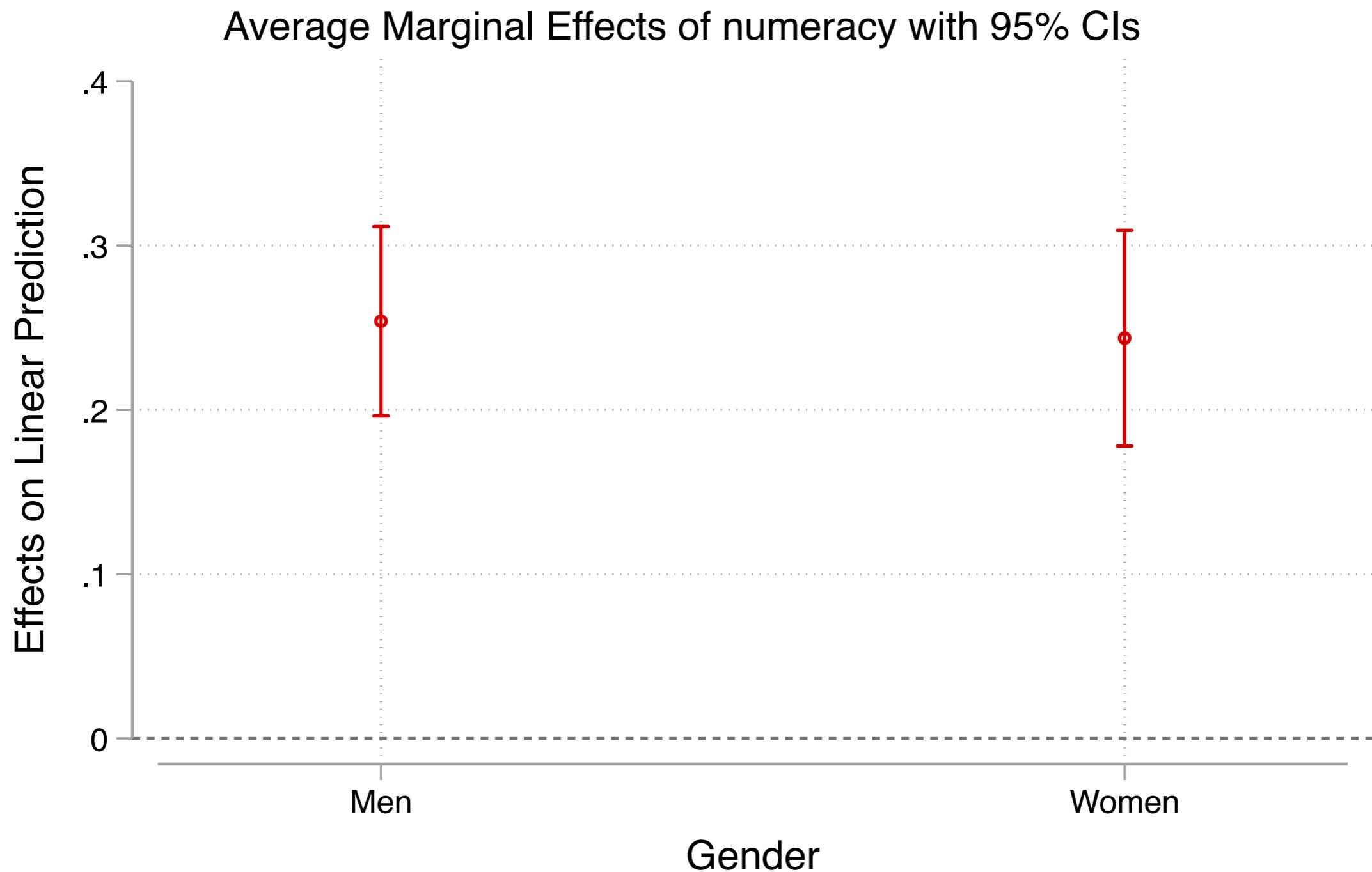
$$\begin{aligned} Y &= \beta_0 + \beta_1 X + \beta_2 Z + \beta_3 XZ + \varepsilon \\ &= \beta_0 + (\beta_1 + \beta_3 Z)X + \beta_2 Z + \varepsilon \end{aligned}$$

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

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。

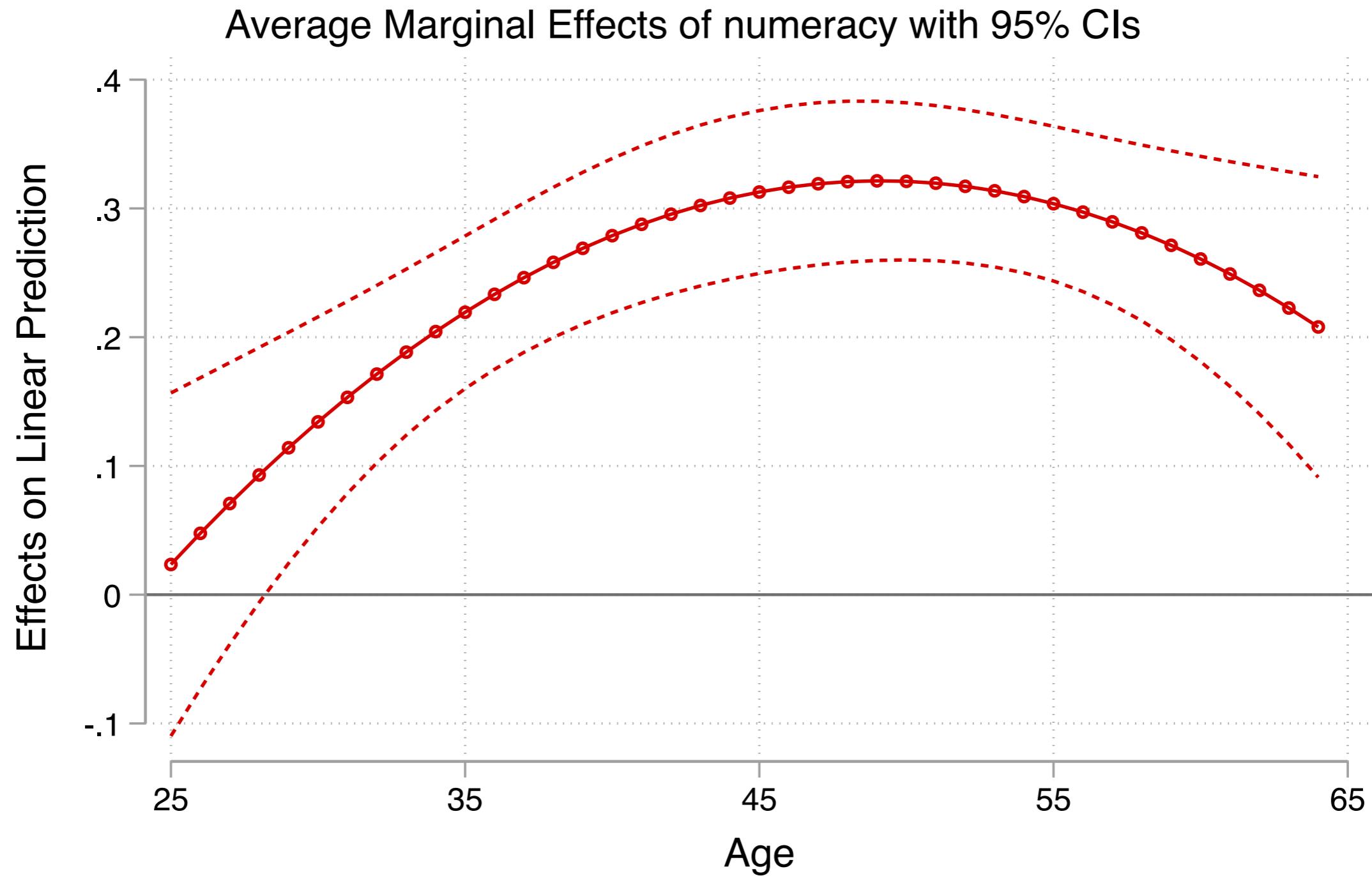
性別ごとにみた限界効果

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示すとよい



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（3.3.1）

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（3.3.2）

よく話題になる点：多重共線性

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 $\text{Cor}(X_2, X_3 | X_1)$ が非常に高いと、 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` でチェックできる。

多重共線性を気にする必要はあるか？→問い合わせに依存する。

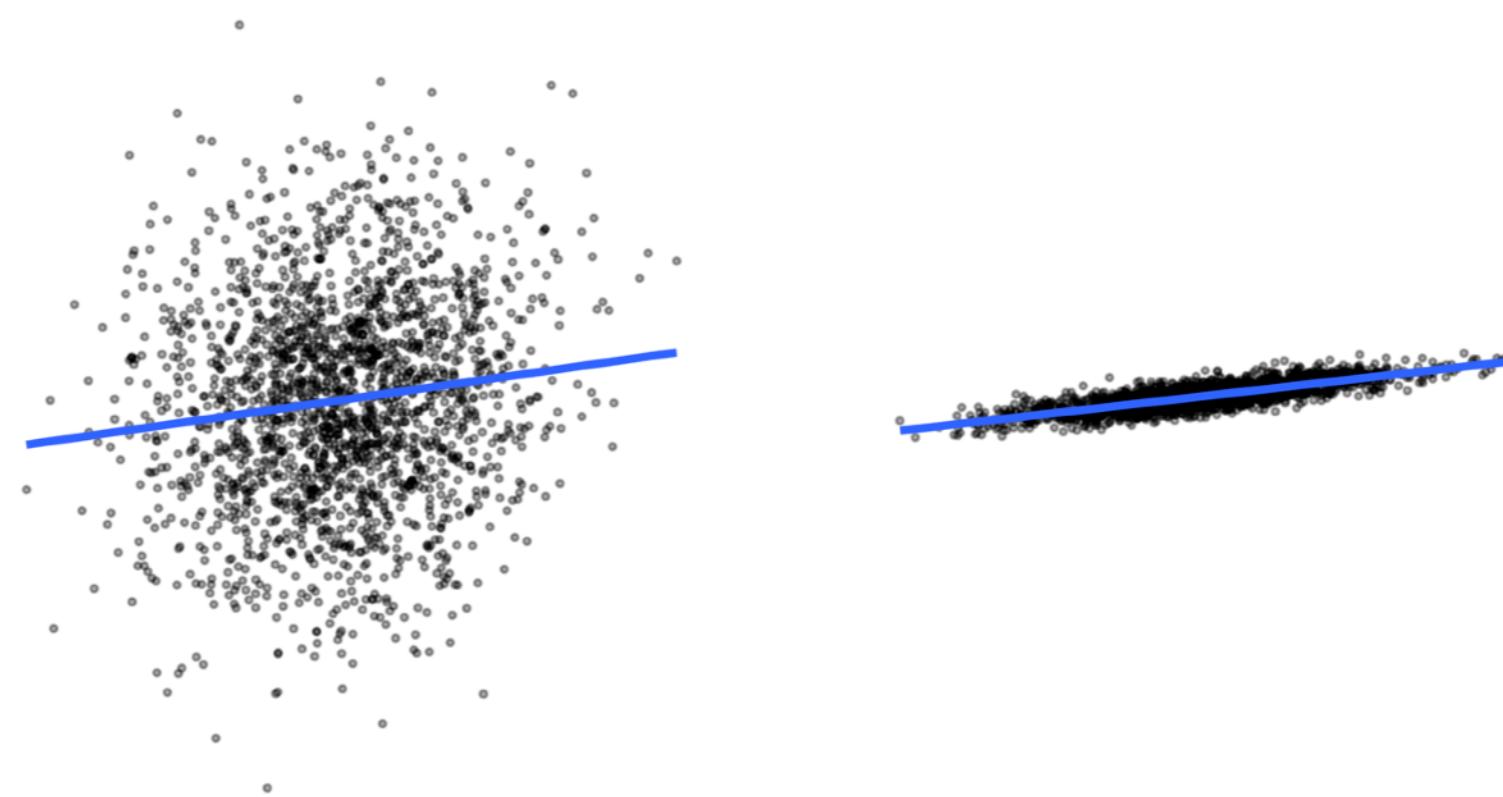
- 知りたい係数が β_1 であるなら、気にする必要はない。
- 知りたい係数が β_2 または β_3 のどちらかなら、VIF以外にも標準誤差が過剰に大きくなったりしていないかなどを確認する。こうしたことがなさそうなら、気にする必要はない。

よく話題になる点：決定係数

決定係数：回帰式により得られる予測分散がYの分散をどの程度説明しているか。

$$R^2 = \frac{\text{Var}(\hat{Y})}{\text{Var}(Y)} = 1 - \frac{\text{Var}(\varepsilon)}{\text{Var}(Y)}$$
 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、標準誤差が小さくなる。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】

小池和男, 2005, 『仕事の経済学（第3版）』

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.

クロス集計表を見る

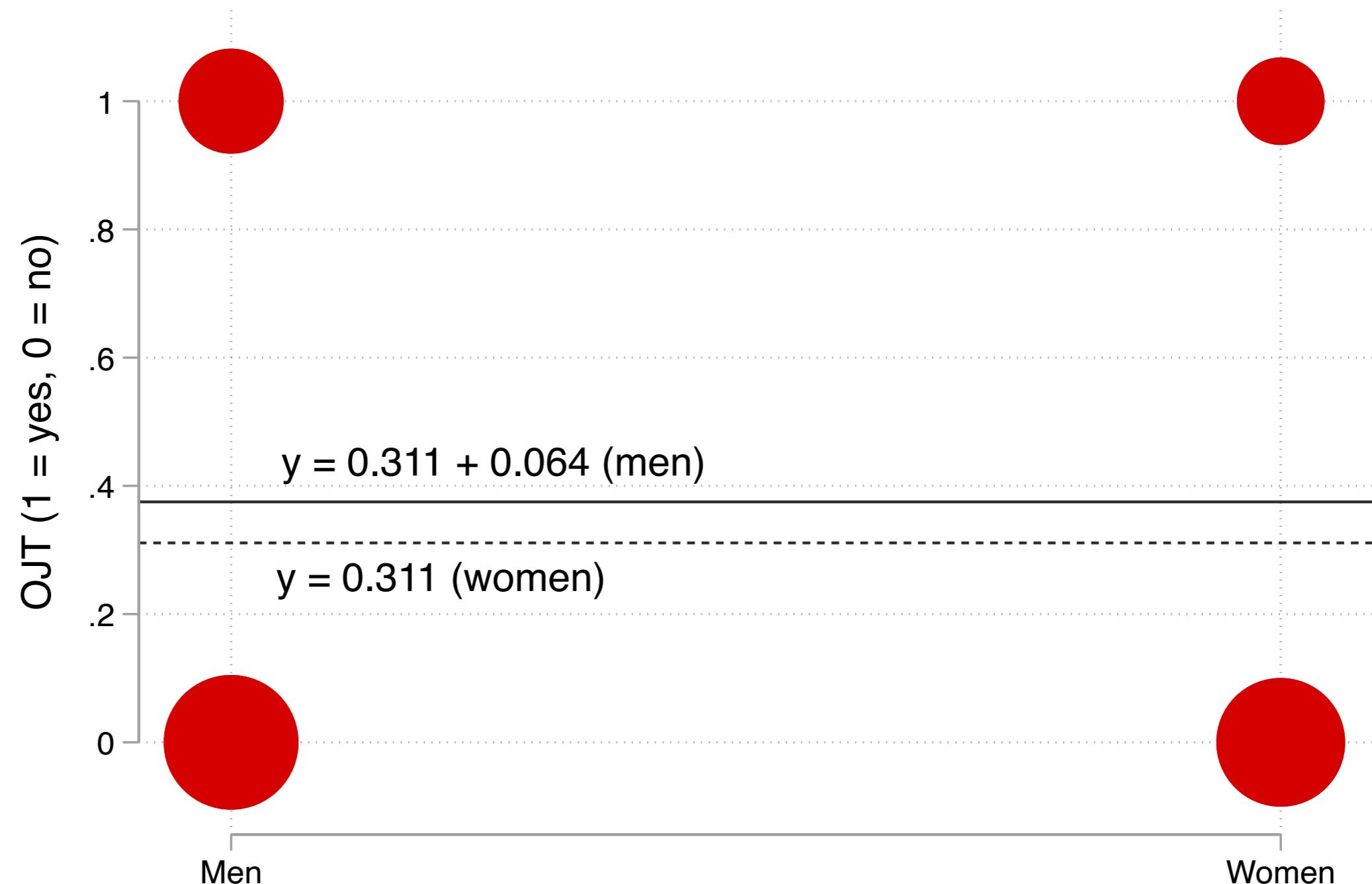
Gender	OJT		Total
	0	1	
Men	1,139 62.51	683 37.49	1,822 100.00
Women	1,039 68.94	468 31.06	1,507 100.00
Total	2,178 65.43	1,151 34.57	3,329 100.00

女性よりも男性のほうがこの1年にOJTを受けている割合が6.4%ポイント高い。

線形回帰分析を使って表現すると？

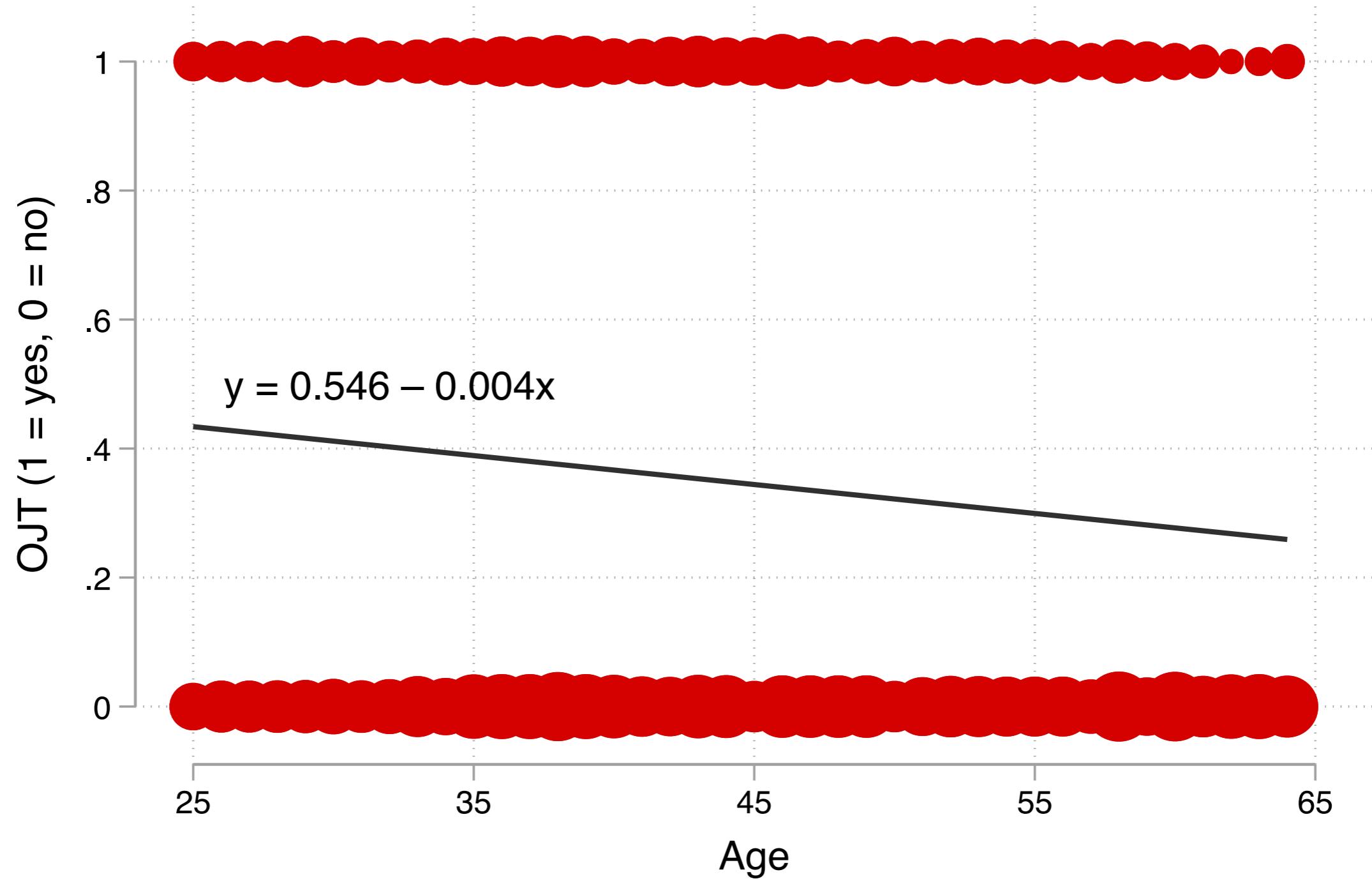
線形回帰分析を当てはめてみる

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



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

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

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

線形確率モデルを推定する

4_logit2021-08-31.doを開き、線形確率モデルを推定してみよう（4.1.1）

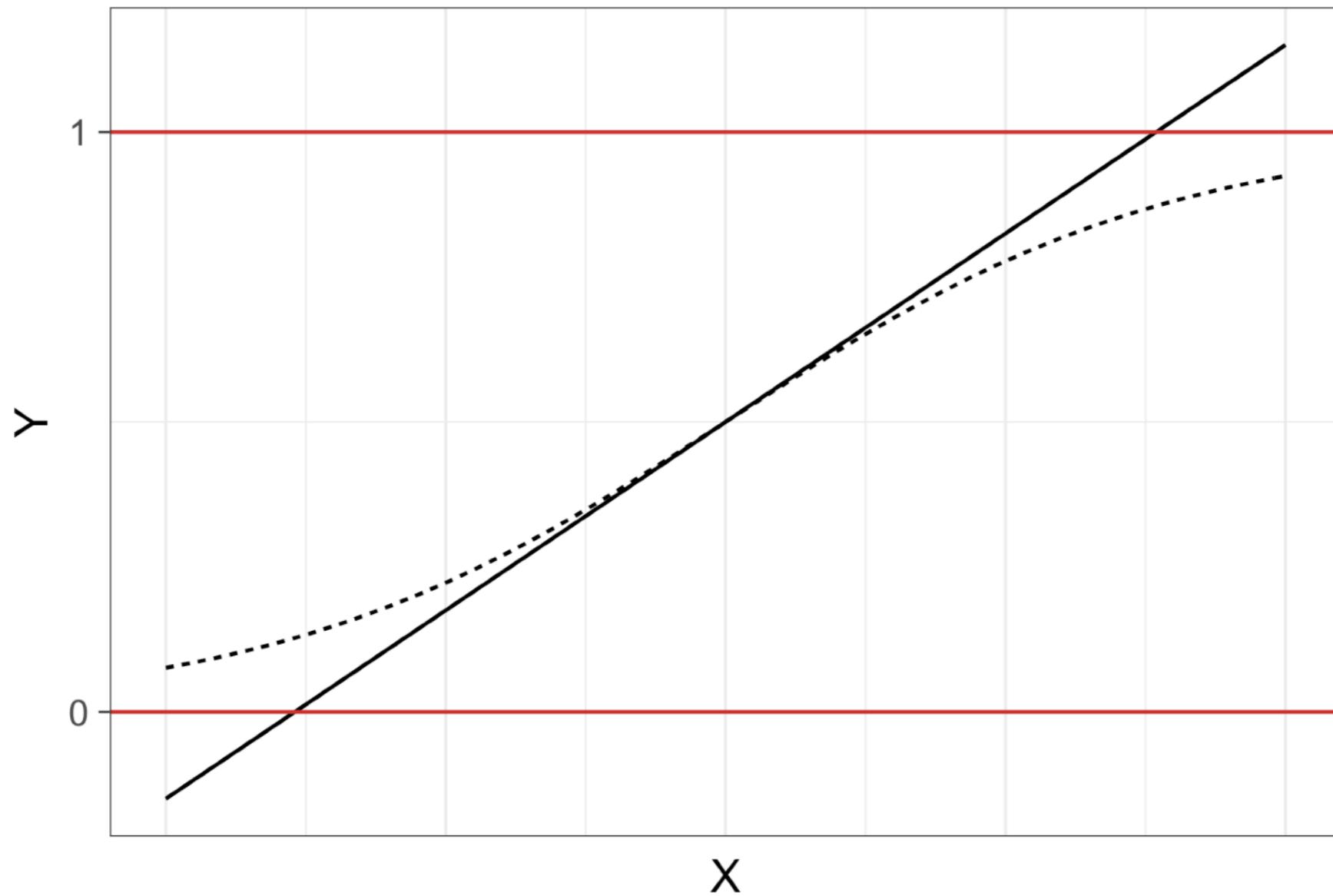
. reg ojt ib2.gender age, vce(robust) // ロバスト標準誤差						
Linear regression		Number of obs = 3,329 F(2, 3326) = 27.44 Prob > F = 0.0000 R-squared = 0.0153 Root MSE = .47216				
ojt		Robust				
		Coef.	Std. Err.	t	P> t	[95% Conf. Interval]
gender						
Men		.064827	.0163637	3.96	0.000	.0327431 .096911
age		-.0044988	.000723	-6.22	0.000	-.0059165 -.0030811
_cons		.5112172	.0351361	14.55	0.000	.4423266 .5801078

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、1より大きい）になることがある。ただしこれは普通の回帰分析でも起こりうる
2. 残差が正規分布しない（残差の不均一分散）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

線形の式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？



— Linear: $y = a + bx$ ----- Logistic: $y = \exp(a + bx)/(1 + \exp(a + bx))$

ロジスティック回帰分析 / ロジットモデル

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 **Logistic regression analysis** あるいはロジットモデル **Logit model** とよぶ。

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

$$\log \frac{\Pr(Y = 1)}{1 - \Pr(Y = 1)} = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

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 = 1におけるオッズ : $p_1/(1 - p_1)$

X	Y		1
	Success	Failure	
1	p_1	$1 - p_1$	
2	p_2	$1 - p_2$	

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 :

$$\begin{aligned} & \log \frac{p_1/(1 - p_1)}{p_2/(1 - p_2)} \\ &= \log(p_1/(1 - p_1)) - \log(p_2/(1 - p_2)) \end{aligned}$$

具体例

Gender	OJT		Total
	0	1	
Men	1,139 62.51	683 37.49	1,822 100.00
Women	1,039 68.94	468 31.06	1,507 100.00
Total	2,178 65.43	1,151 34.57	3,329 100.00

男性のオッズ（OJTなしに対するOJTありの比）： $37.49 / 62.51 = 0.60$

女性のオッズ（OJTなしに対するOJTありの比）： $31.06 / 68.94 = 0.45$

男性の対数オッズ： $\log(0.60) = -0.51$

女性の対数オッズ： $\log(0.45) = -0.80$

対数オッズ比（男性の対数オッズ – 女性の対数オッズ）： $\log(0.60) - \log(0.45) = 0.29$

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 (4.2.1)

```
. logit ojt ib2.gender

Iteration 0:  log likelihood = -2146.458
Iteration 1:  log likelihood = -2138.8921
Iteration 2:  log likelihood = -2138.8882
Iteration 3:  log likelihood = -2138.8882

Logistic regression                                         Number of obs      = 3,329
                                                               LR chi2(1)        = 15.14
                                                               Prob > chi2       = 0.0001
Log likelihood = -2138.8882                                Pseudo R2        = 0.0035


```

ojt	Coef.	Std. Err.	z	P> z	[95% Conf. Interval]
gender					
Men	.2861346	.0737652	3.88	0.000	.1415574 .4307118
_cons	-.7975457	.0556706	-14.33	0.000	-.9066581 -.6884333

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

回帰分析と同じことに注意する

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ように、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（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（4.2.2）

	LPM		Logit	
main				
Men	0.048**	(0.017)	0.233**	(0.079)
Women	0.000	(.)	0.000	(.)
Junior high	0.000	(.)	0.000	(.)
Senior high	0.040	(0.026)	0.260	(0.163)
Junior college	0.152***	(0.030)	0.798***	(0.169)
University	0.253***	(0.028)	1.209***	(0.162)
Age	0.019**	(0.006)	0.096**	(0.030)
Age # Age	-0.000***	(0.000)	-0.001***	(0.000)
Constant	-0.132	(0.138)	-3.135***	(0.659)
Observations	3329		3329	
r2	0.060		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解釈を深める

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4.8%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 $1.26 = \exp(0.233)$ 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

線形確率モデルとロジットモデル、どちらを使うか？

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を使って分析してみるとよい（計算も早い）

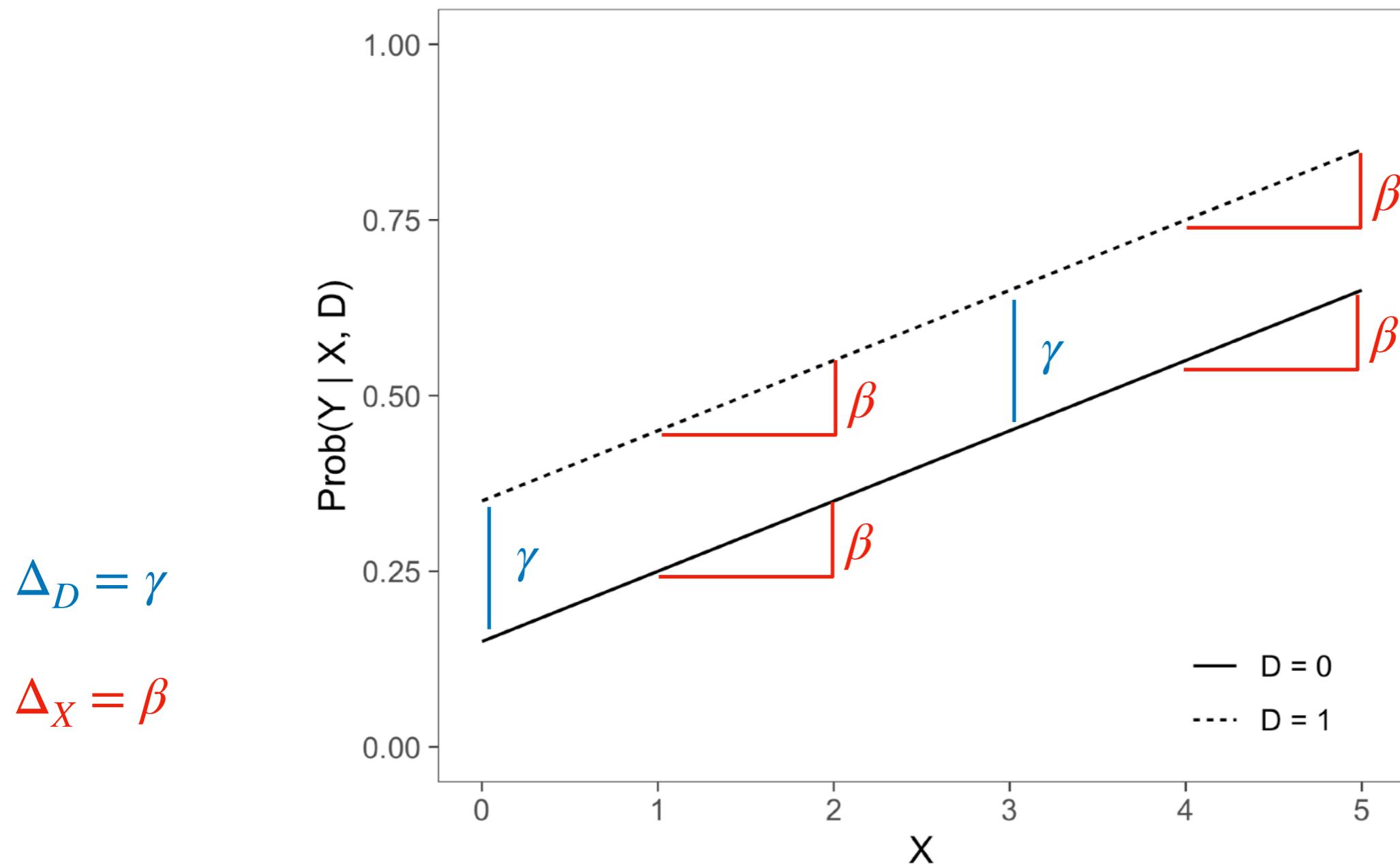
線形確率モデル：係数が解釈しやすい

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

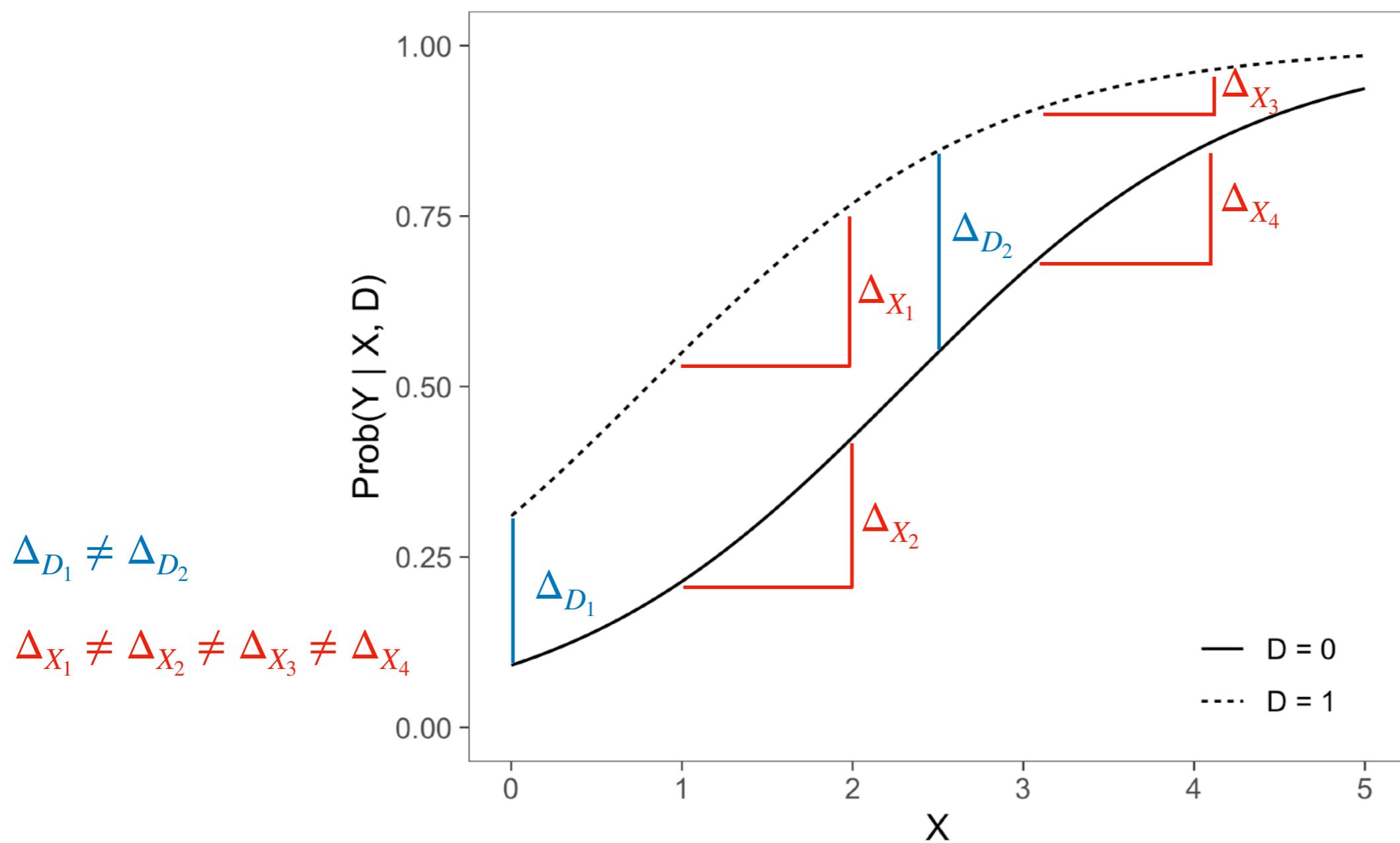
線形モデルの場合

$\Pr(Y) = \alpha + \beta X + \gamma D$ (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



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变化量は $\Pr(Y)$ 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）



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

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

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。

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

(1) – (2) をサンプル内で平均した値 **0.177** が、平均限界効果AMEとなる。

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 (4.3.1)

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 = 3,329
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	.0495588	.0167844	2.95	0.003	.016662 .0824556

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

. margins gender

予測確率 (つまり、2つ前のページの(1)と(2)を表示する)

Predictive margins
Number of obs = 3,329
Model VCE : OIM

Expression : Pr(ojt), predict()

Delta-method					
	Margin	Std. Err.	z	P> z	[95% Conf. Interval]
gender					
Men	.3680757	.0111904	32.89	0.000	.346143 .3900085
Women	.3185169	.012017	26.51	0.000	.294964 .3420699

$$\begin{aligned} &\leftarrow \Pr(Y = 1 | X, D = 1) \\ &\leftarrow \Pr(Y = 1 | X, D = 0) \end{aligned}$$

調整効果

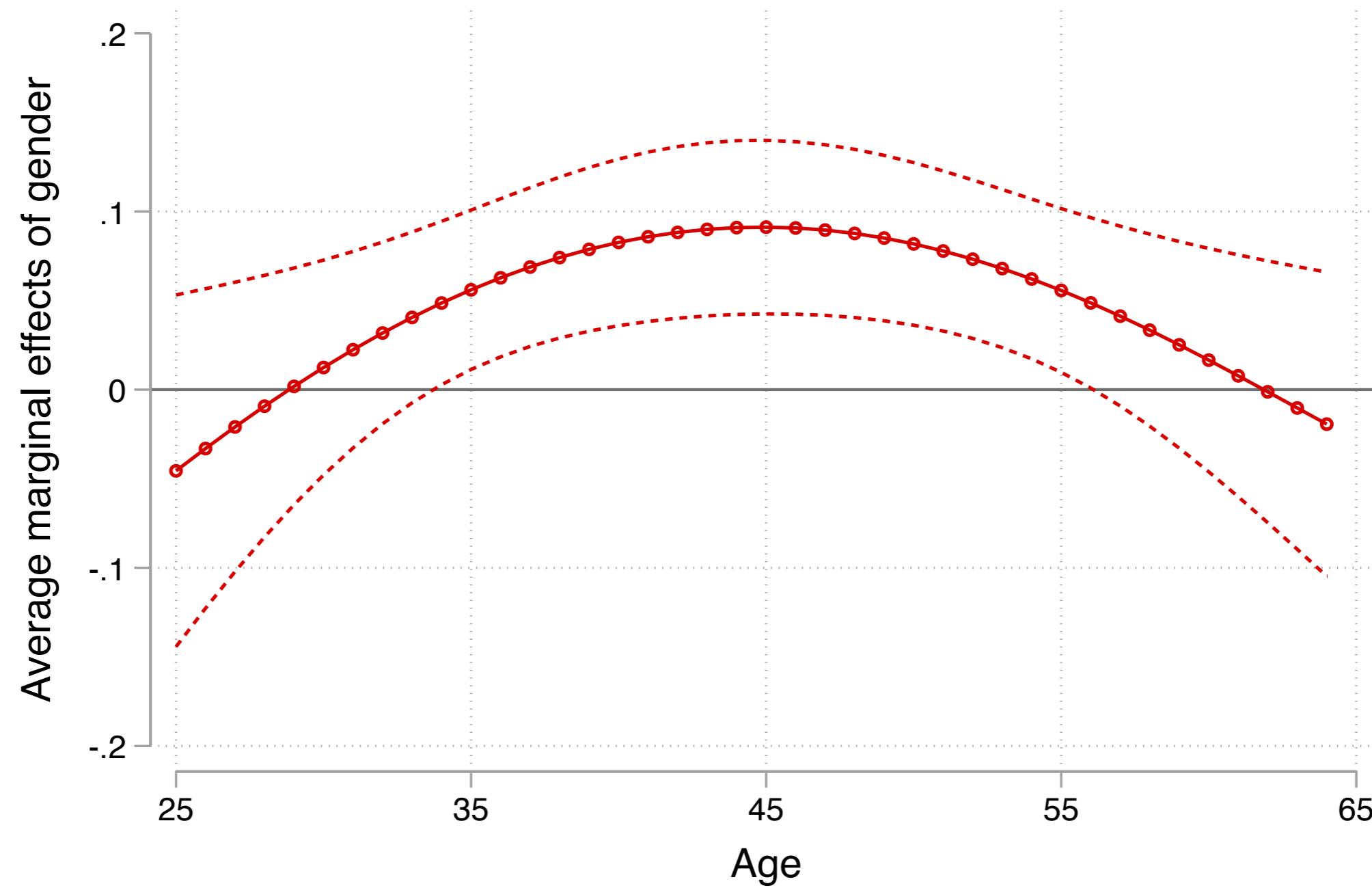
ロジットモデルでも線形回帰分析のときと同様に調整効果（交互作用効果）を考えることができる。

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

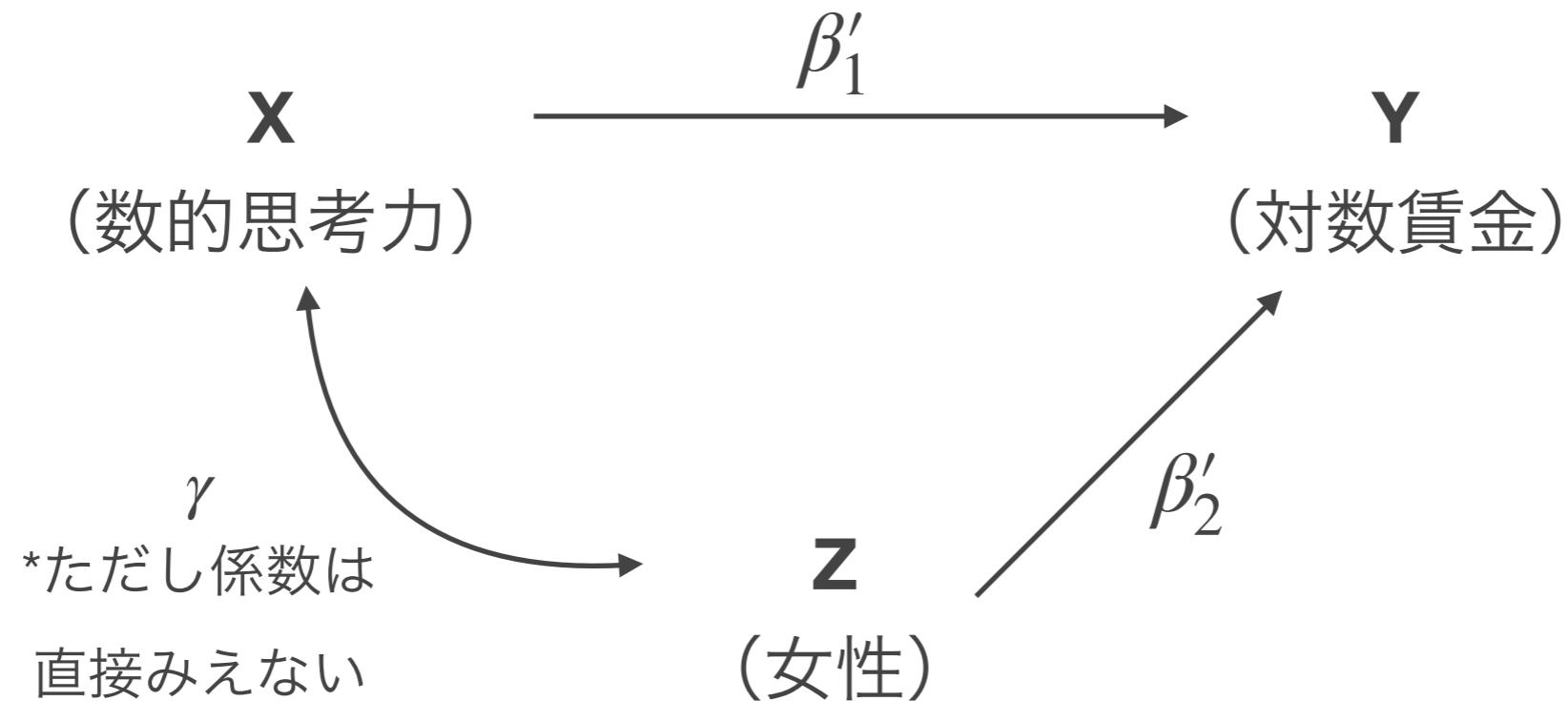
この場合も、対数オッズ比だけでなく、確率がどの程度異なるかを平均限界効果を用いてチェックするとよい。

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（4.3.3）



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
- 媒介要因の寄与分を計算したいときは、Imai-Keeleの方法やKarlson-Holm-Breenの方法を使う：

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.

よく話題になる点：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
いる。

今回のOJTは0: 2178ケース, 1: 1151ケースなので、115個

参考) 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.

よく話題になる点：完全予測

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

このように、 $X = 2$ のときには必ずYの値が決まるということを指して完全予測という。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が複数含まれている場合には計算が収束しないことがある。カテゴリ変数を独立変数として用いる場合、このような変数が含まれていないかを確認し、あればカテゴリの統合などを考える。

よく話題になる点：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

学習のための参考文献

Stataでのプロジェクト管理・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

Mitchell, Michael N. 2012. *A Visual Guide to Stata Graphics, Third 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.

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.

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.

Mize, Trenton D. 2019. “Best Practices for Estimating, Interpreting, and Presenting Nonlinear Interaction Effects.” *Sociological Science* 6:81–117.

回帰分析の使い方と因果関係

Morgan, Stephan and Christopher Winchip. 2015. *Counterfactuals and Causal Inference: Methods and Principles for Social Research, 2nd Edition*. Cambridge University Press.

Cunningham, Scott. 2021. *Causal Inference: The Mixtape*. Yale University Press.

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.

Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. “What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory.” *American Sociological Review* 86(3):532–65.

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』早期公開.