GETTING STARTED

Create a new project called `week-13-lecture` in ShareLaTeX

CHRISTOPHER PRENER, PH.D.

FALL, 2017

WEEK 13

LECTURE 14

QUANTITATIVE ANALYSIS

MULTIPLE REGRESSION (1)

AGENDA

1. Front Matter
2. Bibliographies in LaTeX
3. Multiple Regression Theory
4. Multiple Regression in R
5. Back Matter

1 FRONT MATTER

ANNOUNCEMENTS

Lab-13 is due next Monday

2 BIBLIOGRAPHIES IN LaTeX

BIBLATEX

```
\usepackage[backend=engine, style=styleName,
    datelabel=arg]{biblatex}
```

Parameters:

- back
  bibl
- styl

Available in biblatex
Included in sharelatex.com

- datelabel is a preference for how dates should appear

BIBLATEX

```
\usepackage[backend=engine, style=styleName,
    datelabel=arg]{biblatex}
```

Parameters:

- `backend` refers to one of the two “engines” used to generate bibliography entries
- `style` refers to the citation style being used
- `datelabel` is a preference for how dates should appear

BIBLATEX WITH APA

```
\usepackage[backend=engine, style=styleName,
    datelabel=arg]{biblatex}
```

Using apa:

```
\usepackage[backend=biber, style=apa,
    datelabel=terse]{biblatex}
```

Not all citation styles are available, some are available in other packages, and argument construction may vary! To use APA, include this in your preamble.

BIBLATEX WITH CHICAGO

```
\usepackage[style, backend=engine]{biblatex-chicago}
```

Parameters:

- styl
- back
  bibl

Available in `biblatex-chicago`
Included in `sharelatex.com`

BIBLATEX WITH CHICAGO

```
\usepackage[style, backend=engine]{biblatex-chicago}
```

Parameters:

- `style` refers to the variant of Chicago being used
- `backend` refers to one of the two "engines" used to generate bibliography entries

BIBLATEX WITH CHICAGO

```
\usepackage[style, backend=engine]{biblatex-chicago}
```

Using parenthetical citations:

```
\usepackage[authordate, backend=biber]{biblatex-chicago}
```

Include this in your preamble.

LINK TO BIBLIOGRAPHY FILE

`\addbibresource{`*bibliography*`.bib}`

Parameters:

- *bibl* e

All other commands available in `biblatex`
Included in `sharelatex.com`

LINK TO BIBLIOGRAPHY FILE

```
\addbibresource{fileName.bib}
```

Parameters:

- *fileName.bib* refers to the file name of your bibliography file

LINK TO BIBLIOGRAPHY FILE

```
\addbibresource{fileName.bib}
```

Using parenthetical citations:

```
\addbibresource{bibliography.bib}
```

Include this in your preamble and add the bibliography file to your project in sharelatex.

INSIDE YOUR BIBLIOGRAPHY FILE

```
@book{fisher1937design,
  title={The design of experiments},
  author={Fisher, Ronald Aylmer},
  year={1937},
  publisher={Oliver And Boyd; Edinburgh; London}
}
@article{mcgill1978variations,
  title={Variations of box plots},
  author={McGill, Robert and Tukey, John W and Larsen, Wayne A},
  journal={The American Statistician},
  volume={32},
  number={1},
  pages={12--16},
  year={1978},
  publisher={Taylor \& Francis Group}
}
```

INSIDE YOUR BIBLIOGRAPHY FILE

```
@book{fisher1937design,
  title={The design of experiments},
  author={Fisher, Ronald Aylmer},
  year={1937},
  publisher={Oliver And Boyd; Edinburgh; London}
}
@article{mcgill1978variations,
  title={Variations of box plots},
  author={McGill, Robert and Tukey, John W and Larsen, Wayne A},
  journal={The American Statistician},
  volume={32},
  number={1},
  pages={12--16},
  year={1978},
  publisher={Taylor \& Francis Group}
}
```

It is a good practice to keep it organized alphabetically

INSIDE YOUR BIBLIOGRAPHY FILE

```
@book{fisher1937design,
  title={The design of experiments},
  author={Fisher, Ronald Aylmer},
  year={1937},
  publisher={Oliver And Boyd; Edinburgh; London}
}
@article{mcgill1978variations,
  title={Variations of box plots},
  author={McGill, Robert and Tukey, John W and Larsen, Wayne A},
  journal={The American Statistician},
  volume={32},
  number={1},
  pages={12--16},
  year={1978},
  publisher={Taylor \& Francis Group}
}
```

Make sure that `cite_key` values do not conflict

INSIDE YOUR BIBLIOGRAPHY FILE

```
@book{fisher1937design,
  title={The design of experiments},
  author={Fisher, Ronald Aylmer},
  year={1937},
  publisher={Oliver And Boyd; Edinburgh; London}
}
@article{mcgill1978variations,
  title={Variations of box plots},
  author={McGill, Robert and Tukey, John W and Larsen, Wayne A},
  journal={The American Statistician},
  volume={32},
  number={1},
  pages={12--16},
  year={1978},
  publisher={Taylor \& Francis Group}
}
```

There are a number of standard templates including articles, books, and chapters (`@incollection`)

INSIDE YOUR BIBLIOGRAPHY FILE

```
@book{fisher1937design,
  title={The design of experiments},
  author={Fisher, Ronald Aylmer},
  year={1937},
  publisher={Oliver And Boyd; Edinburgh; London}
}
@article{mcgill1978variations,
  title={Variations of box plots},
  author={McGill, Robert and Tukey, John W and Larsen, Wayne A},
  journal={The American Statistician},
  volume={32},
  number={1},
  pages={12--16},
  year={1978},
  publisher={Taylor \& Francis Group}
}
```

The fields available will vary based on template

INSIDE YOUR BIBLIOGRAPHY FILE

```
@book{fisher1937design,
  title={The design of experiments},
  author={Fisher, Ronald Aylmer},
  year={1937},
  address={London, UK},
  publisher={Oliver and Boyd}
}
@article{mcgill1978variations,
  title={Variations of box plots},
  author={McGill, Robert and Tukey, John W and Larsen, Wayne A},
  journal={The American Statistician},
  volume={32},
  number={1},
  pages={12--16},
  year={1978},
  publisher={Taylor \& Francis Group}
}
```

The address field is not used by Google Scholar but is particularly useful for including publisher location in book cites!

INSIDE YOUR BIBLIOGRAPHY FILE

Braces around field values are used to preserve case

```
@book{fisher1937design,
  title={The design of experiments},
  author={Fisher, Ronald Aylmer},
  year={1937},
  address={London, UK},
  publisher={Oliver and Boyd}
}
@article{mcgill1978variations,
  title={Variations of box plots},
  author={McGill, Robert and Tukey, John W and Larsen, Wayne A},
  journal={The American Statistician},
  volume={32},
  number={1},
  pages={12--16},
  year={1978},
  publisher={Taylor \& Francis Group}
}
```

INSIDE YOUR BIBLIOGRAPHY FILE

```
@book{fisher1937design,
  title="The design of experiments",
  author="Fisher, Ronald Aylmer",
  year="1937",
  address="London, UK",
  publisher="Oliver and Boyd"
}
@article{mcgill1978variations,
  title="Variations of box plots",
  author="McGill, Robert and Tukey, John W and Larsen, Wayne A",
  journal="The American Statistician",
  volume="32",
  number="1",
  pages="12--16",
  year="1978",
  publisher="Taylor \& Francis Group"
}
```

You can replace braces with double quotes to allow style-based case

INSERT CITATIONS

`\cite[`*page*`]{`*cite_key*, *cite_key*`}`

Single citation:

```
\cite{fisher1937design}
```

Make sure your *cite_key* is included in your bibliography file!

INSERT CITATIONS

```
\cite[page]{cite_key, cite_key}
```

Single citation:

```
\cite{fisher1937design}
```

APA

(Fischer, 1937)

Chicago

(Fischer 1937)

INSERT CITATIONS

```
\cite[page]{cite_key, cite_key}
```

Single citation, year only:

```
\cite*{fisher1937design}
```

Make sure your *cite_key* is included in your bibliography file!

INSERT CITATIONS

```
\cite[page]{cite_key, cite_key}
```

Single citation, year only:

```
\cite*{fisher1937design}
```

APA

(1937)

Chicago

(1937)

INSERT CITATIONS

`\cite[page]{cite_key, cite_key}`

Single citation with page:

`\cite[25]{fisher1937design}`

Make sure your *cite_key* is included in your bibliography file!

INSERT CITATIONS

```
\cite[page]{cite_key, cite_key}
```

Single citation with page:

```
\cite[25]{fisher1937design}
```

APA

(Fischer, 1937, p. 25)

Chicago

(Fischer 1937, 25)

INSERT CITATIONS

`\cite[page]{cite_key, cite_key}`

Single citation with page range:

`\cite[25--30]{fisher1937design}`

Make sure your *cite_key* is included in your bibliography file! The double dash is an important detail for how the final product appears!

INSERT CITATIONS

```
\cite[page]{cite_key, cite_key}
```

Single citation with page range:

```
\cite[25--30]{fisher1937design}
```

APA	Chicago
(Fischer, 1937, pp. 25-30)	(Fischer 1937, 25-30)

INSERT CITATIONS

```
\cite[page]{cite_key, cite_key}
```

Single citation with page range:

```
\cite[25--30]{fisher1937design}
```

Make sure your *cite_key* is included in your bibliography file! The double dash is an important detail for how the final product appears!

INSERT CITATIONS

```
\cite[page]{cite_key, cite_key}
```

Two citations

```
\cite{fisher1937design, tukey1962future}
```

APA

(Fischer, 1937; Tukey, 1962)

Chicago

(Fischer 1937; Tukey 1962)

WORKS CITED

```
\printbibliography
```

Include cited resources on same page as other content:

```
\printbibliography
```

Handouts and “concept papers” are the only time that we typically do this in academia.

WORKS CITED

`\printbibliography`

Include cited resources on same page as other content:

`\printbibliography`

APA

References

Fisher, R. A. (1937). *The design of experiments*. London, UK: Oliver and Boyd.

WORKS CITED

```
\printbibliography
```

Begin on new page:

```
\newpage
\printbibliography
```

This is a more typical configuration for academic writing.

BIBLIOGRAPHY

```
\nocite{*}
```

Include all resources in your `.bib` file:

```
\nocite{*}
\printbibliography
```

Use this combination for the final project handout.

BIBLIOGRAPHY

Include all resources in your `.bib` file:

```
\nocite{*}
\printbibliography
```

APA

References

Fisher, R. A. (1937). *The design of experiments*. London, UK: Oliver and Boyd.

McGill, R., Tukey, J. W., & Larsen, W. A. (1978). Variations of box plots. *The American Statistician*, *32*(1), 12–16.

Tukey, J. W. (1962). The future of data analysis. *The annals of mathematical statistics*, *33*(1), 1–67.

CHANGE TITLE OF SECTION

```
\printbibliography[title = title name]
```

Change default title (“References”) to “Key Sources”:

```
\printbibliography[title = Key Sources]
```

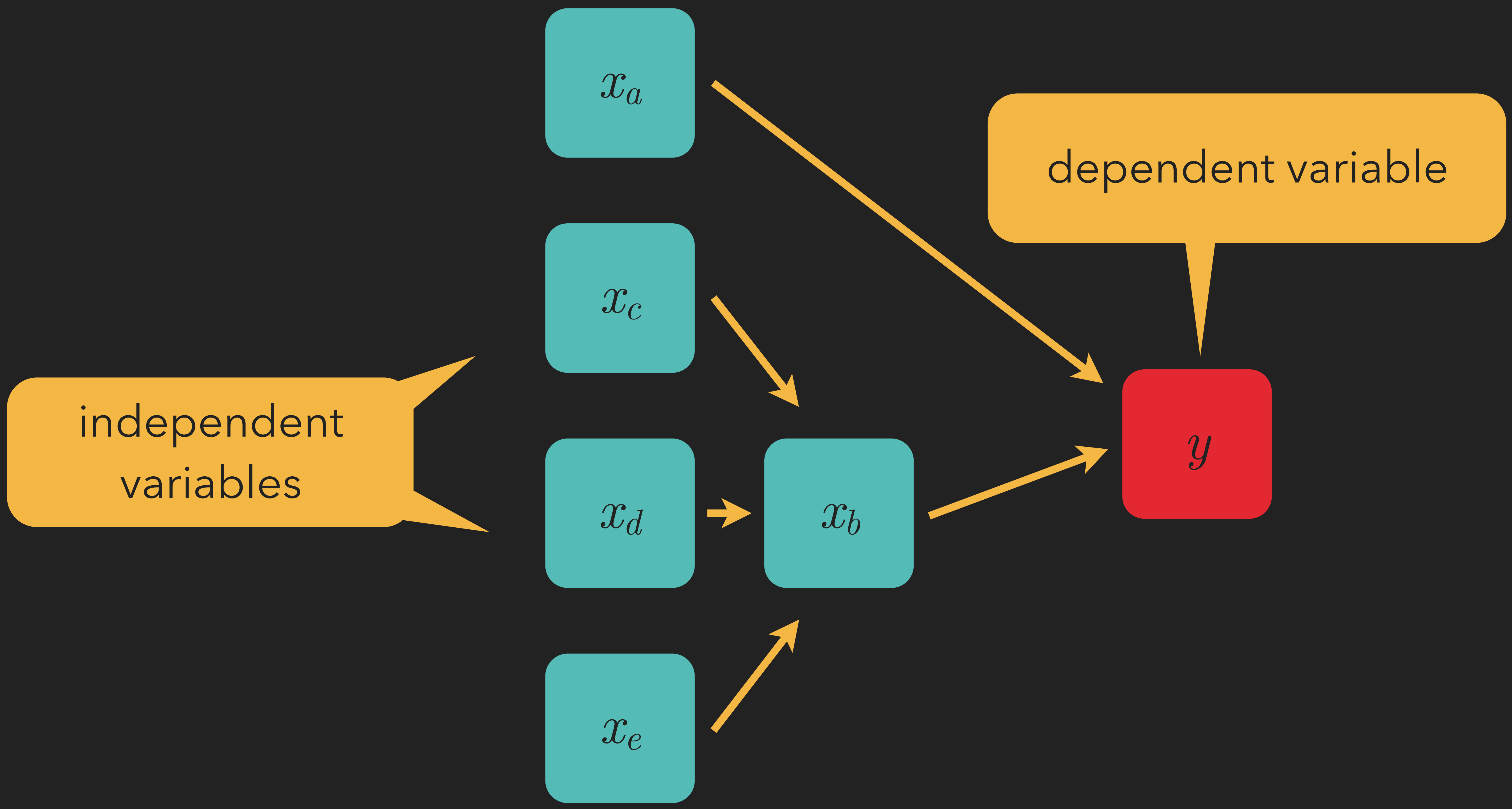
APA

Key Sources

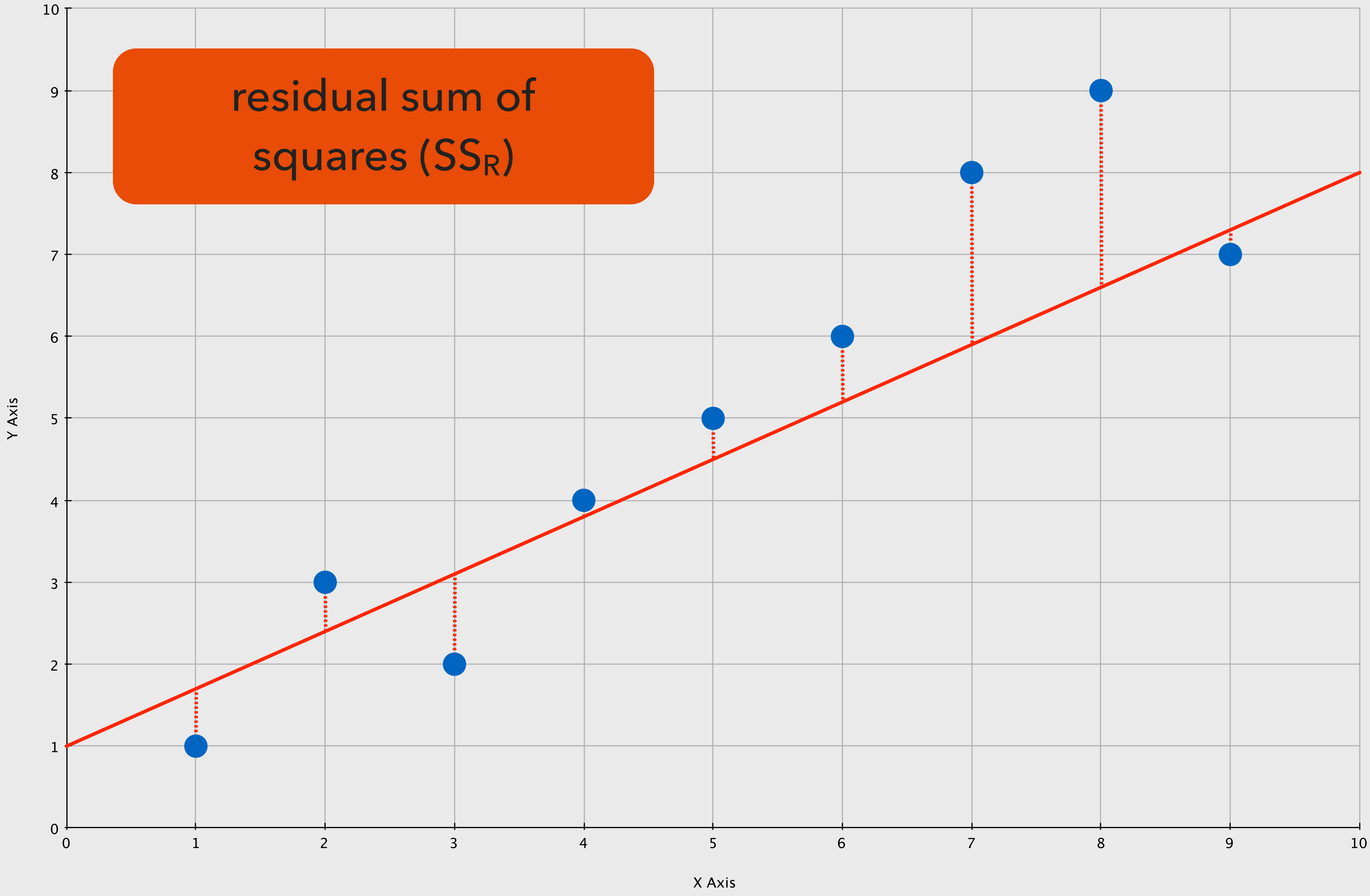
Fisher, R. A. (1937). *The design of experiments*. London, UK: Oliver and Boyd.

3 MULTIPLE REGRESSION THEORY

THE "REAL" WORLD



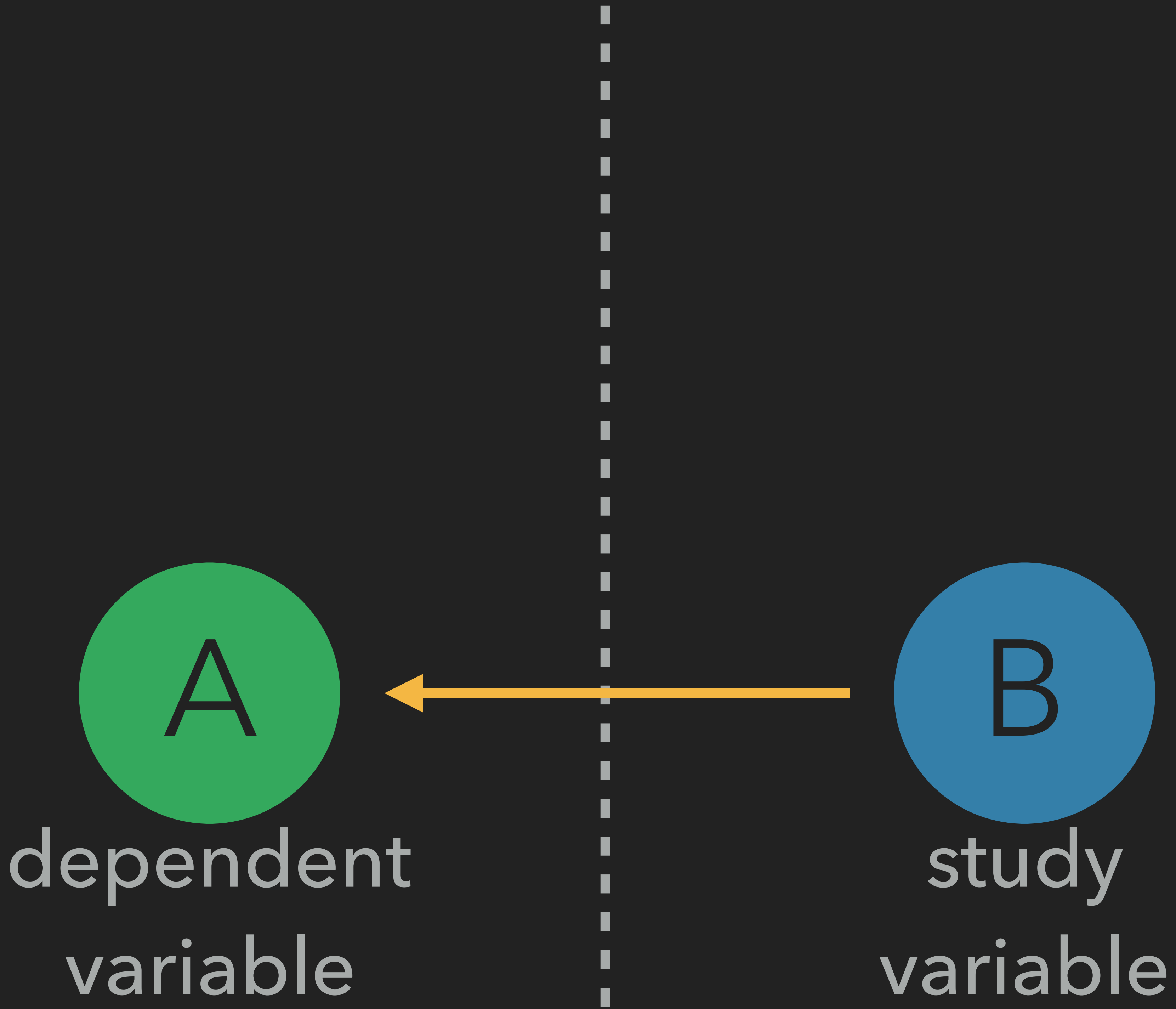
THE GOAL OF OLS REGRESSION



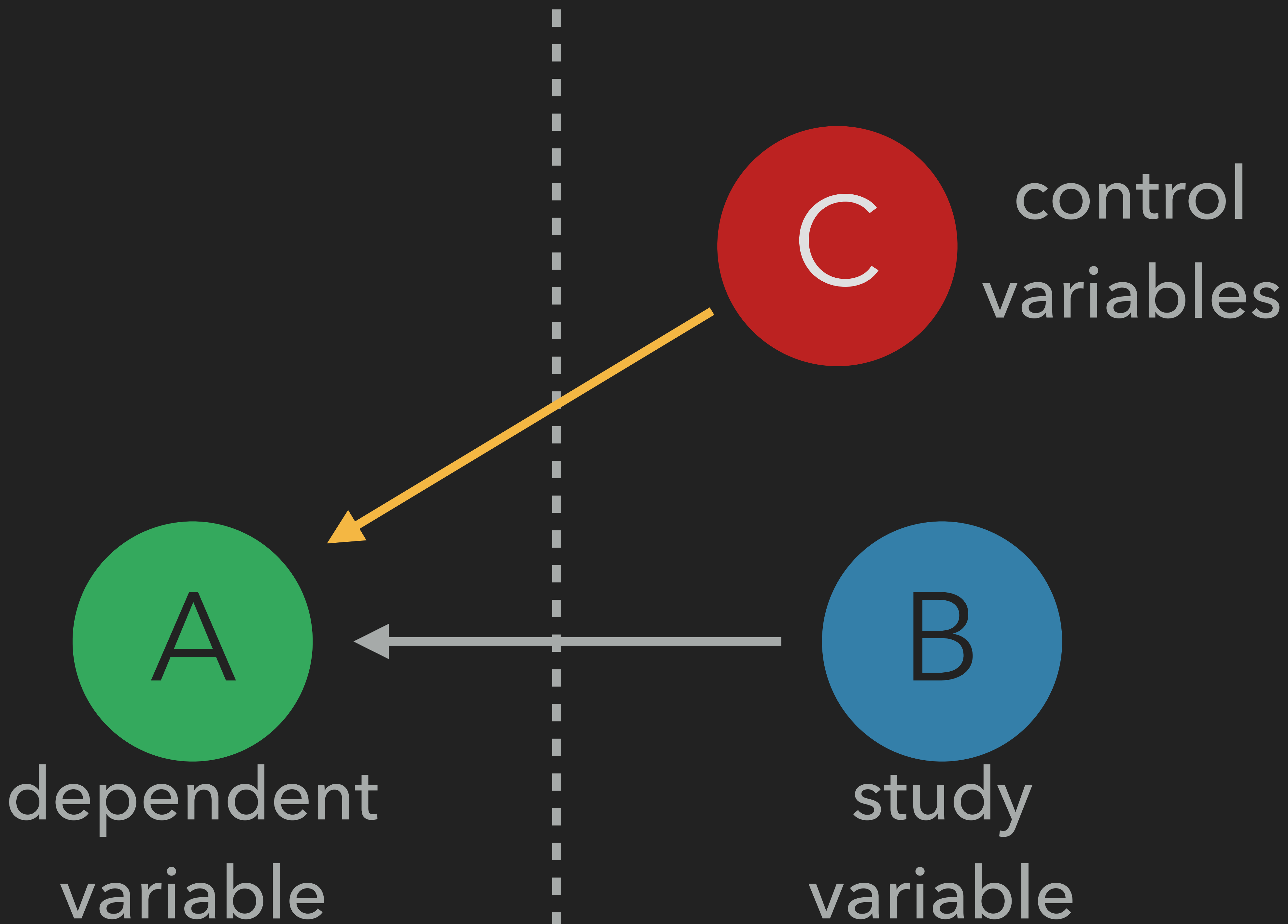
IN OTHER WORDS...

We want to explain as much of the variation in y as we can while also minimizing the residual error in the regression line.

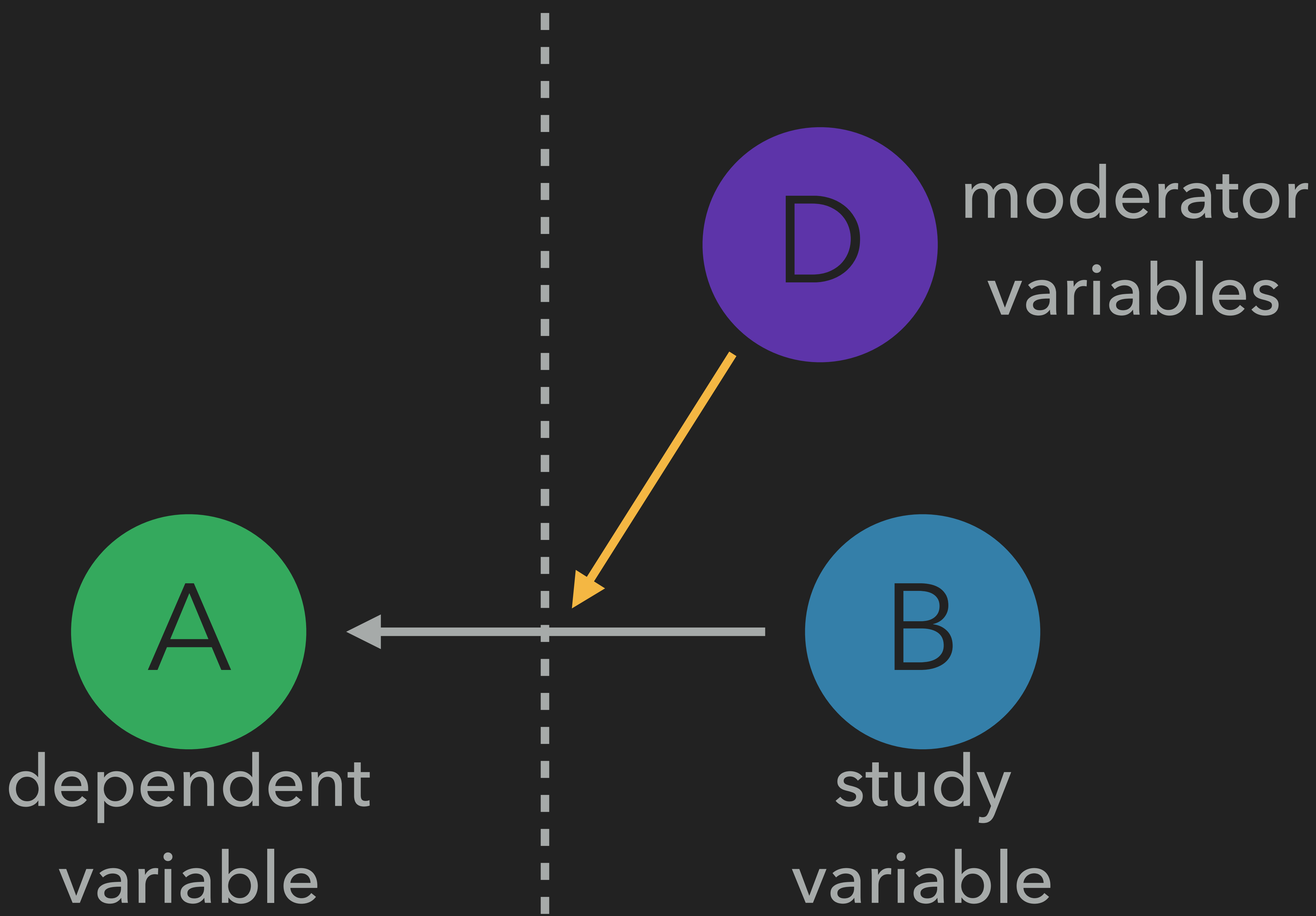
ACCOUNTING FOR VARIATION



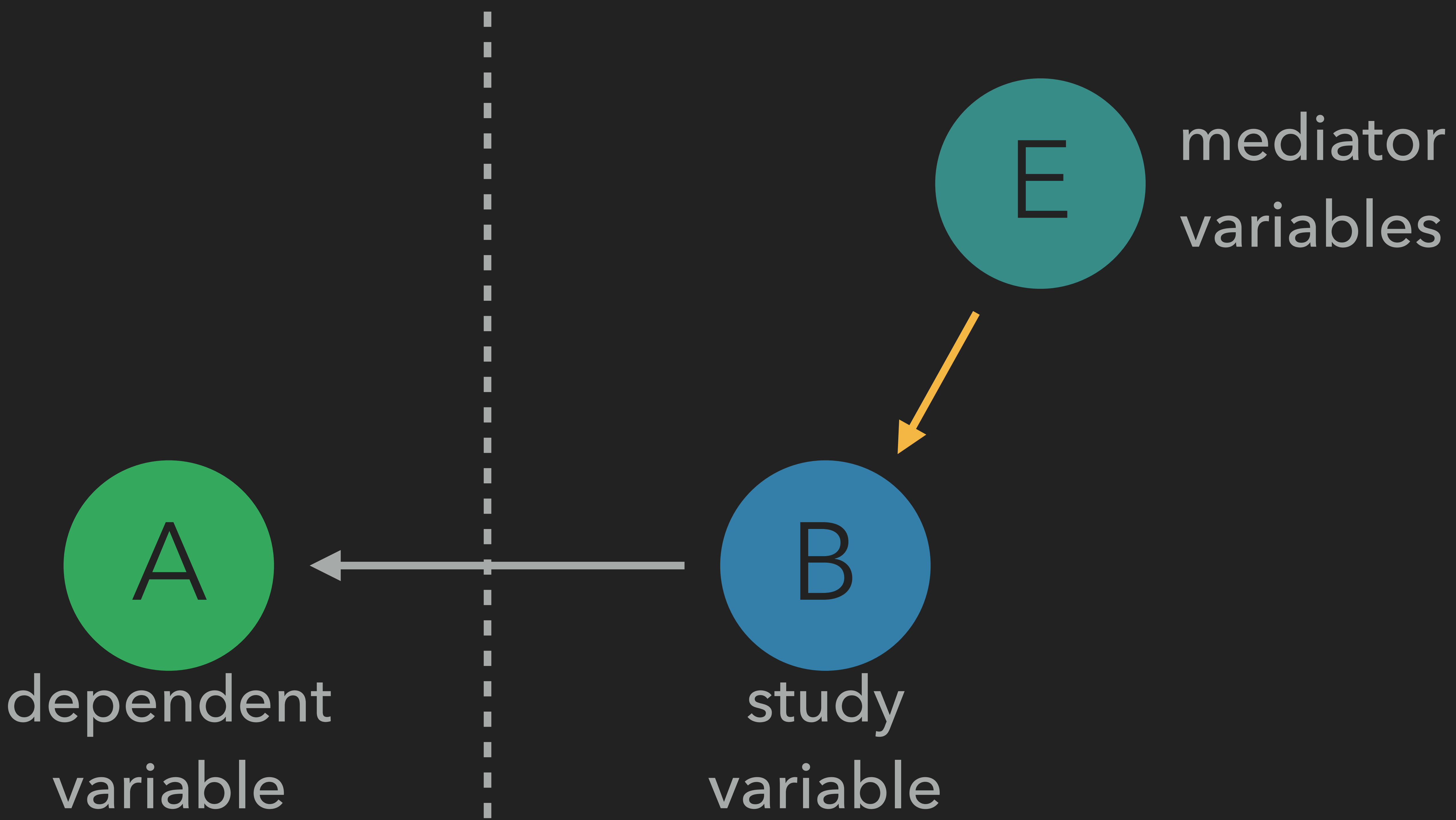
ACCOUNTING FOR VARIATION



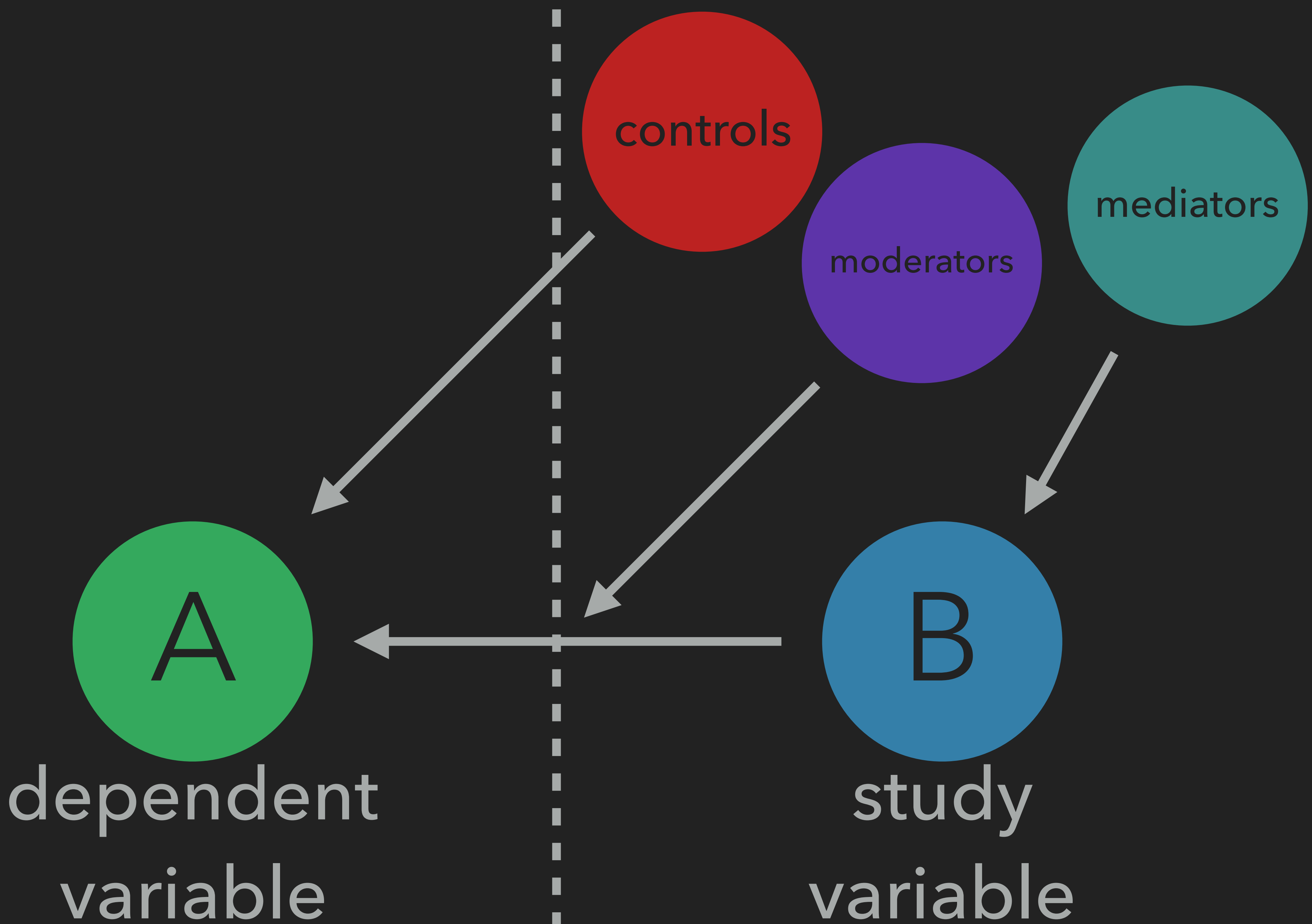
ACCOUNTING FOR VARIATION



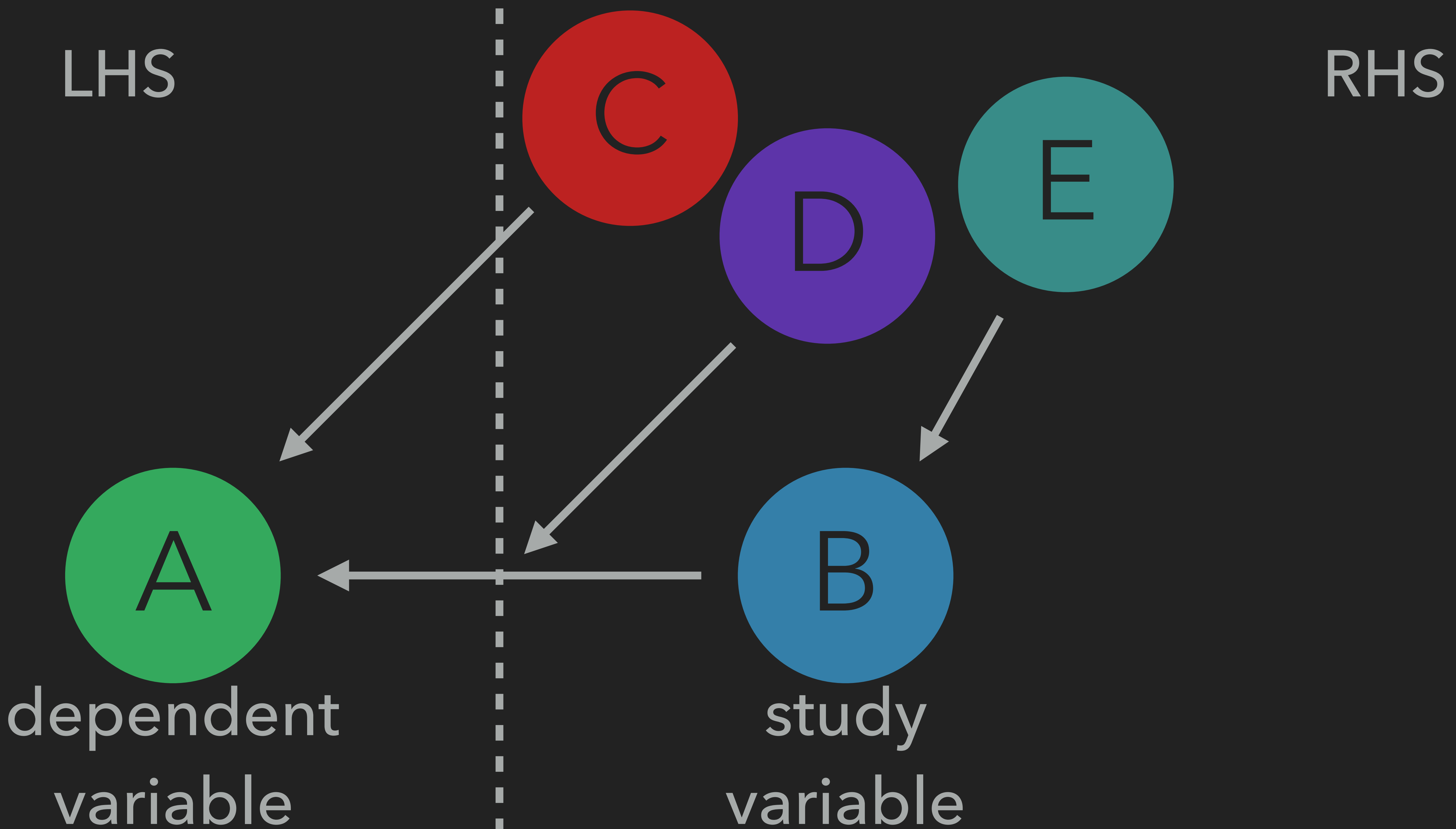
ACCOUNTING FOR VARIATION



ACCOUNTING FOR VARIATION



ACCOUNTING FOR VARIATION



ACCOUNTING FOR VARIATION

$$y_{height} = \alpha + \beta_1 x_{female} + \epsilon$$

LHS

RHS

VARIABLES FOR THE RHS

What other measures are included with your data (or are accessible) that might help account for variation in y?

LIMITS ON THE RHS

For every RHS variable, we need 10 to 15 observations. If we exceed that rule of thumb, we consider the model “overfitted”.

LIMITS ON THE RHS

```
> library(ggplot2)

> autoData <- mpg

> nrow(mpg)
[1] 234
```

How many RHS predictors could we include in our model?

AN EXAMPLE

What is the effect of grade point average on test scores, controlling for the effects of time spent studying and socioeconomic status?

AN EXAMPLE

H_1 = higher grade point averages are associated with higher test scores, holding constant both effort and SES

AN EXAMPLE

$$y = \alpha + \beta_i x_i + \epsilon$$

$y =$ dependent variable

$\alpha =$ constant

$x_i =$ independent variable i

$\beta_i =$ beta value of IV i

DV = test score

ME = gpa

IV = hours studying

IV = free lunch eligible

AN EXAMPLE

$$y = \alpha + \beta_i x_i + \epsilon$$

$y =$ dependent variable

$\alpha =$ constant

$x_i =$ independent variable i

$\beta_i =$ beta value of IV i

DV = test score

ME = gpa

IV = hours studying

IV = free lunch eligible

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

AN EXAMPLE

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

AN EXAMPLE

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

AN EXAMPLE

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

? What do you think the constant (or intercept) represents?

AN EXAMPLE

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

The average test score for a student with a GPA of “0” who studied for “0” hours and does not get a free lunch.

AN EXAMPLE

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

The average test score for a student with a GPA of “0” who studied for “0” hours and does not get a free lunch.

INTERPRETING BETAS

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      68.7166     5.7262  12.000 1.81e-15 ***
gpa               6.1242     0.0811   1.531   0.0009 ***
studyHrs        -1.17074    0.28129  -4.162 0.000148 ***
freeLunch        -7.8843     3.7484  -2.103   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.728 on 43 degrees of freedom
Multiple R-squared:  0.433,     Adjusted R-squared:  0.3935
F-statistic: 10.95 on 3 and 43 DF,  p-value: 1.811e-05
```

How would you interpret the effect of GPA on test scores?

INTERPRETING BETAS

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      68.7166     5.7262  12.000 1.81e-15 ***
gpa               6.1242     0.0811   1.531   0.0009 ***
studyHrs        -1.17074    0.28129  -4.162 0.000148 ***
freeLunch        -7.8843     3.7484  -2.103   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.728 on 43 degrees of freedom
Multiple R-squared:  0.433,	Adjusted R-squared:  0.3935
F-statistic: 10.95 on 3 and 43 DF,  p-value: 1.811e-05
```

A unit change in GPA is associated with 6.124 ($p = .0009$) increase in test scores. Higher GPAs are associated with better test scores, controlling for hours spent studying and free lunch eligibility.

INTERPRETING BETAS

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      68.7166     5.7262  12.000 1.81e-15 ***
gpa               6.1242     0.0811   1.531   0.0009 ***
studyHrs        -1.17074    0.28129  -4.162 0.000148 ***
freeLunch        -7.8843     3.7484  -2.103   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.728 on 43 degrees of freedom
Multiple R-squared:  0.433,     Adjusted R-squared:  0.3935
F-statistic: 10.95 on 3 and 43 DF,  p-value: 1.811e-05
```

? How would you interpret the effects of hours spent studying and free lunch eligibility on test scores?

STANDARD ERROR OF BETA

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      68.7166     5.7262  12.000 1.81e-15 ***
gpa               6.1242     0.0811   1.531   0.0009 ***
studyHrs        -1.17074    0.28129  -4.162 0.000148 ***
freeLunch        -7.8843     3.7484  -2.103   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.728 on 43 degrees of freedom
Multiple R-squared:  0.433,     Adjusted R-squared:  0.3935
F-statistic: 10.95 on 3 and 43 DF,  p-value: 1.811e-05
```

The standard error is an indicator of amount of uncertainty in the estimate, representing the amount of variation present across observations. It is also used to find *t*.

MEASURES OF MODEL FIT

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      68.7166     5.7262  12.000 1.81e-15 ***
gpa               6.1242     0.0811   1.531   0.0009 ***
studyHrs        -1.17074    0.28129  -4.162 0.000148 ***
freeLunch        -7.8843     3.7484  -2.103   0.0412 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.728 on 43 degrees of freedom
Multiple R-squared:  0.433,    Adjusted R-squared:  0.3935
F-statistic: 10.95 on 3 and 43 DF,  p-value: 1.811e-05
```

These are based on calculations of the total sum of squares and the residual sum of squared error.

INTERPRETING R-SQUARED

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     68.7166     5.7262  12.000 1.81e-15 ***
gpa              6.1242     0.0811   1.531   0.0009 ***
studyHrs       -1.17074    0.28129  -4.162 0.000148 ***
freeLunch       -7.8843     3.7484  -2.103   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.728 on 43 degrees of freedom
Multiple R-squared:  0.433,	Adjusted R-squared:  0.3935
F-statistic: 10.95 on 3 and 43 DF,  p-value: 1.811e-05
```

We use adjusted R^2 with multiple regression to account for artificial increases in R^2 due to added RHS parameters.

INTERPRETING R-SQUARED

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     68.7166     5.7262  12.000 1.81e-15 ***
gpa              6.1242     0.0811   1.531   0.0009 ***
studyHrs       -1.17074    0.28129  -4.162 0.000148 ***
freeLunch       -7.8843     3.7484  -2.103   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.728 on 43 degrees of freedom
Multiple R-squared:  0.433,     Adjusted R-squared:  0.3935
F-statistic: 10.95 on 3 and 43 DF,  p-value: 1.811e-05
```

? How would you interpret the adjusted R^2 value?

INTERPRETING R-SQUARED

```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)     68.7166     5.7262  12.000 1.81e-15 ***
gpa              6.1242     0.0811   1.531   0.0009 ***
studyHrs       -1.17074    0.28129  -4.162 0.000148 ***
freeLunch       -7.8843     3.7484  -2.103   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.728 on 43 degrees of freedom
Multiple R-squared:  0.433,     Adjusted R-squared:  0.3935
F-statistic: 10.95 on 3 and 43 DF,  p-value: 1.811e-05
```

The adjusted R^2 value indicates that these factors together account for 39.35% of the variation in test scores.

RESIDUAL STANDARD ERROR

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      68.7166     5.7262  12.000 1.81e-15 ***
gpa               6.1242     0.0811   1.531   0.0009 ***
studyHrs        -1.17074    0.28129  -4.162 0.000148 ***
freeLunch        -7.8843     3.7484  -2.103   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.728 on 43 degrees of freedom
Multiple R-squared:  0.433,    Adjusted R-squared:  0.3935
F-statistic: 10.95 on 3 and 43 DF,  p-value: 1.811e-05
```

Also known as the root mean squared error. The average error per observation. We want to minimize this value.

THE F-STATISTIC

```
Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      68.7166     5.7262  12.000 1.81e-15 ***
gpa               6.1242     0.0811   1.531   0.0009 ***
studyHrs        -1.17074    0.28129  -4.162 0.000148 ***
freeLunch        -7.8843     3.7484  -2.103   0.0412 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9.728 on 43 degrees of freedom
Multiple R-squared:  0.433,	Adjusted R-squared:  0.3935
F-statistic: 10.95 on 3 and 43 DF,  p-value: 1.811e-05
```

Evaluation of the null hypothesis that all the betas are equal to zero. It is a measure of the reliability of the model.

CONFIDENCE INTERVALS

```
                   2.5 %     97.5 %
(Intercept)  57.17611384 80.2570179
gpa           8.03926404  4.2876193
studyHours    3.06346803  0.5734902
freeLunch   -15.43865689 -1.3300016
```

These are measures of the accuracy of each beta estimate. If they include zero, the estimate will not be statistically significant.

USING MULTIPLE MODELS

$$y = \alpha + \beta_i x_i + \epsilon$$

$y =$ dependent variable

$\alpha =$ constant

$x_i =$ independent variable i

$\beta_i =$ beta value of IV i

DV = test score

ME = gpa

IV = hours studying

IV = free lunch eligible

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

MULTIPLE MODELS

Model 1, Main Effects:

$$y_{score} = \alpha + \beta_1 x_{gpa} + \epsilon$$

Model 2, Full Model:

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

MODEL BUILDING

$$y = \alpha + \beta_i x_i + \epsilon$$

$y =$ dependent variable

$\alpha =$ constant

$x_i =$ independent variable i

$\beta_i =$ beta value of IV i

DV = test score

ME = gpa

IV = hours studying

IV = free lunch eligible

IV = gender

IV = race (white, black, other)

MODEL BUILDING

$y = \alpha + \beta_i x_i + \epsilon$

$y =$ dependent variable

$\alpha =$ constant

$x_i =$ independent variable i

$\beta_i =$ beta value of IV i

DV = test score

ME = gpa

IV = hours studying

IV = free lunch eligible

IV = gender

IV = race (white, black, other)

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \beta_4 x_{female} + \beta_5 x_{white} + \beta_6 x_{black} + \epsilon$$

MULTIPLE MODELS

Model 1, Main Effects:

$$y_{score} = \alpha + \beta_1 x_{gpa} + \epsilon$$

Model 2, Main + other educational measures :

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

Model 3, Full Model:

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \beta_4 x_{female}$$
$$+ \beta_5 x_{white} + \beta_6 x_{black} + \epsilon$$

COMPARING MODEL FIT

Adjusted R2 should increase, indicated increasing explanatory power.

Model 1, Main Effects:

$$y_{score} = \alpha + \beta_1 x_{gpa} + \epsilon$$

Model 2, Main effect + other educational measures :

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

Model 3, Full Model:

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \beta_4 x_{female} + \beta_5 x_{white} + \beta_6 x_{black} + \epsilon$$

COMPARING MODEL FIT

We can also use AIC and BIC “information criterion” values, which should decrease.

Model 1, Main Effects:

$$y_{score} = \alpha + \beta_1 x_{gpa} + \epsilon$$

Model 2, Main effect + other educational measures :

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \epsilon$$

Model 3, Full Model:

$$y_{score} = \alpha + \beta_1 x_{gpa} + \beta_2 x_{studyHrs} + \beta_3 x_{freeLunch} + \beta_4 x_{female} + \beta_5 x_{white} + \beta_6 x_{black} + \epsilon$$

4 MULTIPLE REGRESSION IN R

OLS MODEL

f(x)

```
lm(y ~ x1+x2+x3, data = dataFrame)
```

Parameters:

- tilde
  - *y*
  - *x1*, *x2*, *x3* are the independent variables
- *dataFrame* is the data source (can be a tibble)

All of the functions discussed are available in `stats`

OLS MODEL

f(x)

```
lm(y ~ x1+x2+x3, data = dataFrame)
```

Parameters:

- tilde (~) used in the construction of the formula where:
  - *y* is the dependent variable
  - *x1*, *x2*, *x3* are the independent variables
- *dataFrame* is the data source (can be a tibble)

OLS MODEL

f(x)

```
lm(y ~ x1+x2+x3, data = dataFrame)
```

Using the `hwy`, `cyl` and `displ` variables from `ggplot2`'s `mpg` data:

```
model <- lm(hwy ~ displ+cyl, data = autoData)
```

CONFIDENCE INTERVALS

f(x)

`confint(`*`modelObject`*`)`

Using the object created on the previous slide:

```
confint(model)
```

Model must be estimated and saved to a model object.

MULTIPLE OLS MODELS

f(x)

```
lm(y ~ x1+x2+x3, data = dataFrame)
```

Using the `hwy`, `cyl` and `displ` variables from `ggplot2`'s `mpg` data:

```
> model1 <- lm(hwy ~ displ, data = autoData)
> model2 <- lm(hwy ~ displ+cyl, data = autoData)
```

AKAIKE'S INFORMATION CRITERION

`AIC(`*`modelObject`*`)`

Using the object created on the previous slide:

```
> AIC(model1)
[1] 2274.479
```

Model must be estimated and saved to a model object.

BAYESIAN INFORMATION CRITERION

f(x) `BIC(modelObject)`

Using the object created on the previous slide:

```
> BIC(model1)
[1] 2424.296
```

Model must be estimated and saved to a model object.

REGRESSION TABLES IN LATEX

`stargazer(models, title = “table title”)`

Using the object created on the previous slide:

```
stargazer(model1, model2, title = "Effects of Engine
    Size on Fuel Efficiency”)
```

<<<<< OUTPUT OMITTED >>>>>

Models must be estimated and saved to a model object. AIC and BIC values have to be added manually.

ROUNDING

f(x)

`round(x, digits = val)`

Storing rounded AIC values:

```
> aic1 <- round(AIC(model1), digits = 3)
> aic2 <- round(AIC(model2), digits = 3)
```

ADDING NEW SUMMARY STATS

```
stargazer(models, title = “table title”, add.lines =
    list(c("text", value, value)))
```

Using the objects created on the previous slide:

```
stargazer(model1, model2, title = "Effects of Engine
    Size on Fuel Efficiency”, add.lines =
    list(c("AIC", aic1, aic2)))
```

<<<<< OUTPUT OMITTED >>>>>

Make sure you round the stored values, otherwise they will not be truncated and will have to be edited manually.

ROUNDING

`round(`x`, digits = `val`)`

Storing rounded AIC values:

```
> bic1 <- round(BIC(model1), digits = 3)
> bic2 <- round(BIC(model2), digits = 3)
```

ADDING NEW SUMMARY STATS

```
stargazer(models, title = “table title”, add.lines =
    list(c("text", value, value)))
```

Using the objects created on the previous slide:

```
stargazer(model1, model2, title = "Effects of Engine
    Size on Fuel Efficiency”, add.lines =
    list(c("AIC", aic1, aic2), c(“BIC”, bic1, bic2)))
```

<<<<< OUTPUT OMITTED >>>>>

OMITTING SUMMARY STATS

```
stargazer(models, title = "table title", omit.stat =
    stat, df = FALSE)
```

Omitting un-adjusted R^2 and degrees of freedom:

```
stargazer(model1, model2, title = "Effects of Engine
    Size on Fuel Efficiency", omit.stat = "rsq",
    df = FALSE)
```

<<<<< OUTPUT OMITTED >>>>>

ROTATING REGRESSION TABLES

```
\begin{landscape}
```

Available in `pdflscape`
Included in sharelatex.com

<<<<<< TABLE >>>>>>>

```
\end{landscape}
```

ROTATING REGRESSION TABLES

```
\begin{landscape}
```

Rotating a table into landscape:

```
\newpage
\begin{landscape}
                    <<<<<< TABLE >>>>>>>
\end{landscape}
```

5 BACK MATTER

WHAT WE COVERED TODAY

2. Bibliographies in $\LaTeX$
3. Multiple Regression Theory
4. Multiple Regression in R

REMINDERS

Lab-13 is due next Monday

My slack availability will be significantly reduced until Sunday night.