CHRISTOPHER PRENER, PH.D.
FALL, 2016

WEEK 14
LECTURE 15

QUANTITATIVE ANALYSIS

BIVARIATE REGRESSION

AGENDA

1. Follow-up
2. Tables and Plots
3. The Failings of Simple Models
4. Bivariate Regression Theory
5. Bivariate Regression in Stata

1 FOLLOW-UP

2 TABLES AND PLOTS

TABLES

- Table format will depend on medium - print vs. projection

	mpg	weight	foreign
mpg	1.00	-	-
weight	-.807***	1.00	-
foreign	.393***	-.593***	1.00

* - $p < .05$; ** - $p < .01$; *** - $p < .001$

TABLES

- Tables for projection should have less detail and emphasize readability

	mpg	weight	foreign
mpg	1.00	-	-
weight	-.807***	1.00	-
foreign	.393***	-.593***	1.00

* - p < .05; ** - p < .01; *** - p < .001

TABLES

- Print tables use minimum horizontal lines
- Tables are included at end of paper with references left in text where they should be placed: <<<<< TABLE 2 ABOUT HERE >>>>>

TABLE 2 - Pearson's *r* Correlation Results

	miles per gallon	weight (lbs.)	foreign
miles per gallon	1.00		
weight (lbs.)	-.807***	1.00	
foreign	.393***	-.593***	1.00

Notes: * - p < .05; ** - p < .01; *** - p < .001

TABLES

- Use your word processor of choice to create a table and edit it to match requested format for journal submissions

TABLE 2 - Pearson's *r* Correlation Results

	miles per gallon	weight (lbs.)	foreign
miles per gallon	1.00		
weight (lbs.)	-.807***	1.00	
foreign	.393***	-.593***	1.00

Notes: * - p < .05; ** - p < .01; *** - p < .001

PLOTS

- Plot format will depend on medium - print vs. projection
- Plots for projection can (and often should!) use color with fonts sized appropriately for slides
- Many journals have a specific standard for how plots are designed, and typically required black and white or grayscale plots
- Fonts for print plots need to be sized appropriately for printing

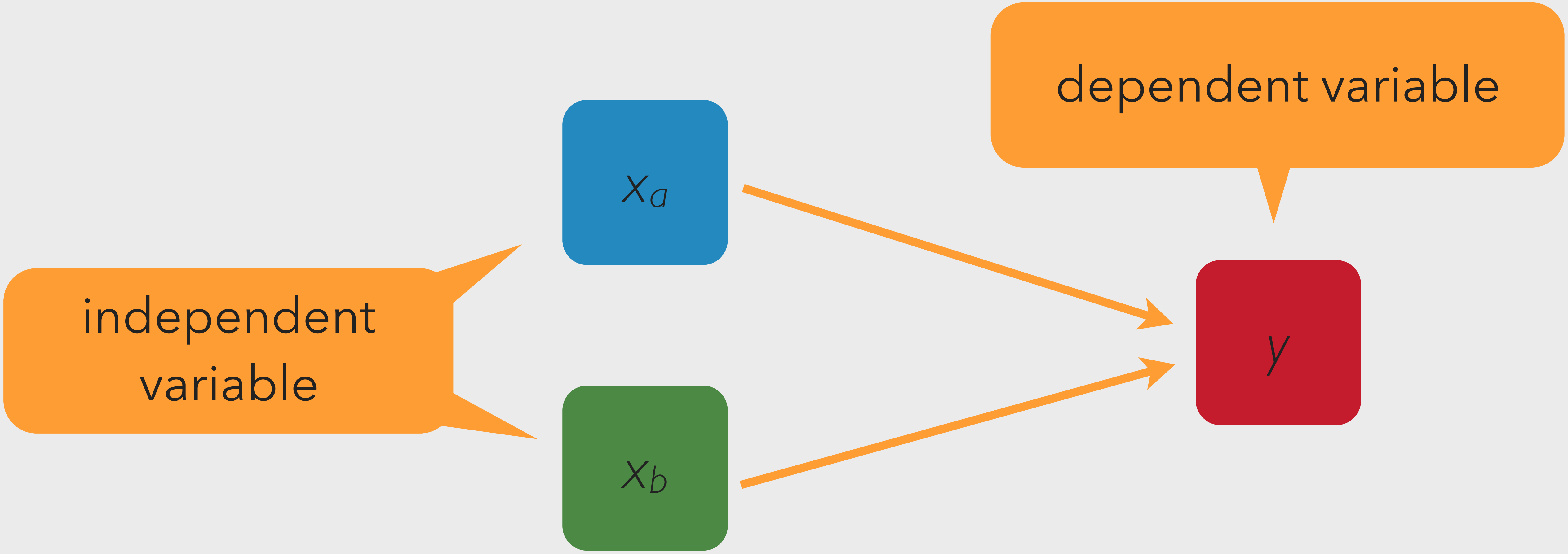
PRESENTATIONS

- See:

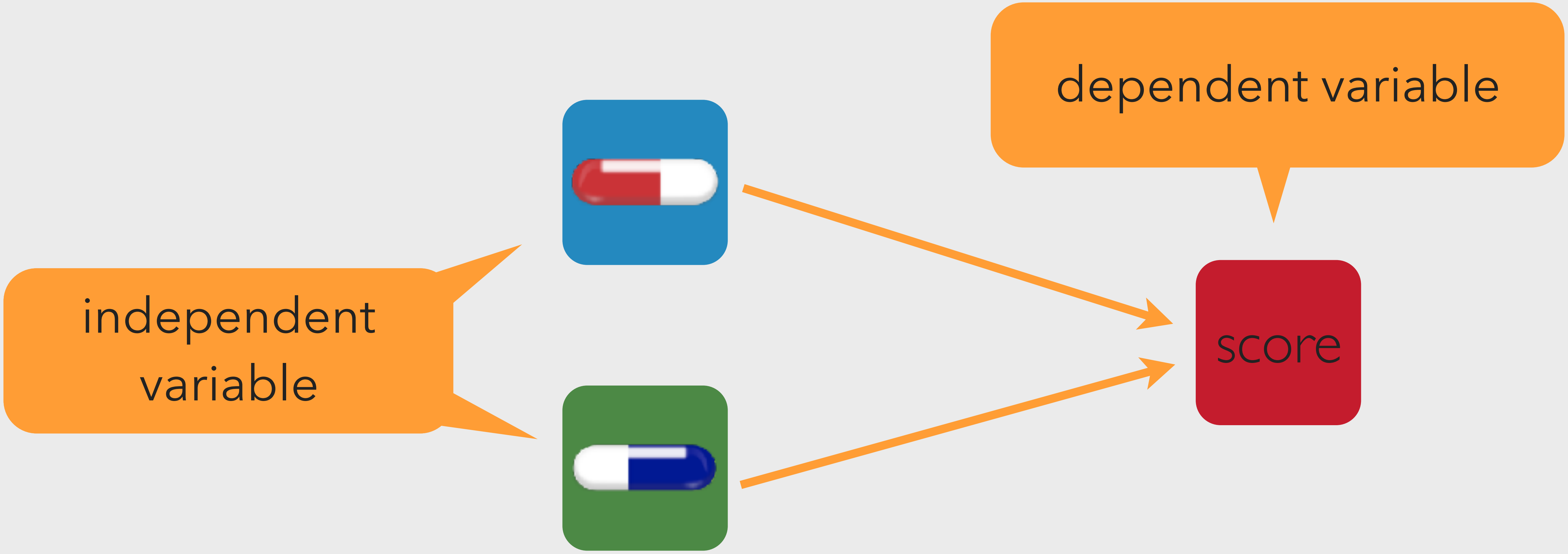
https://github.com/chris-prener/Presentations

3 THE FAILINGS OF SIMPLE MODELS

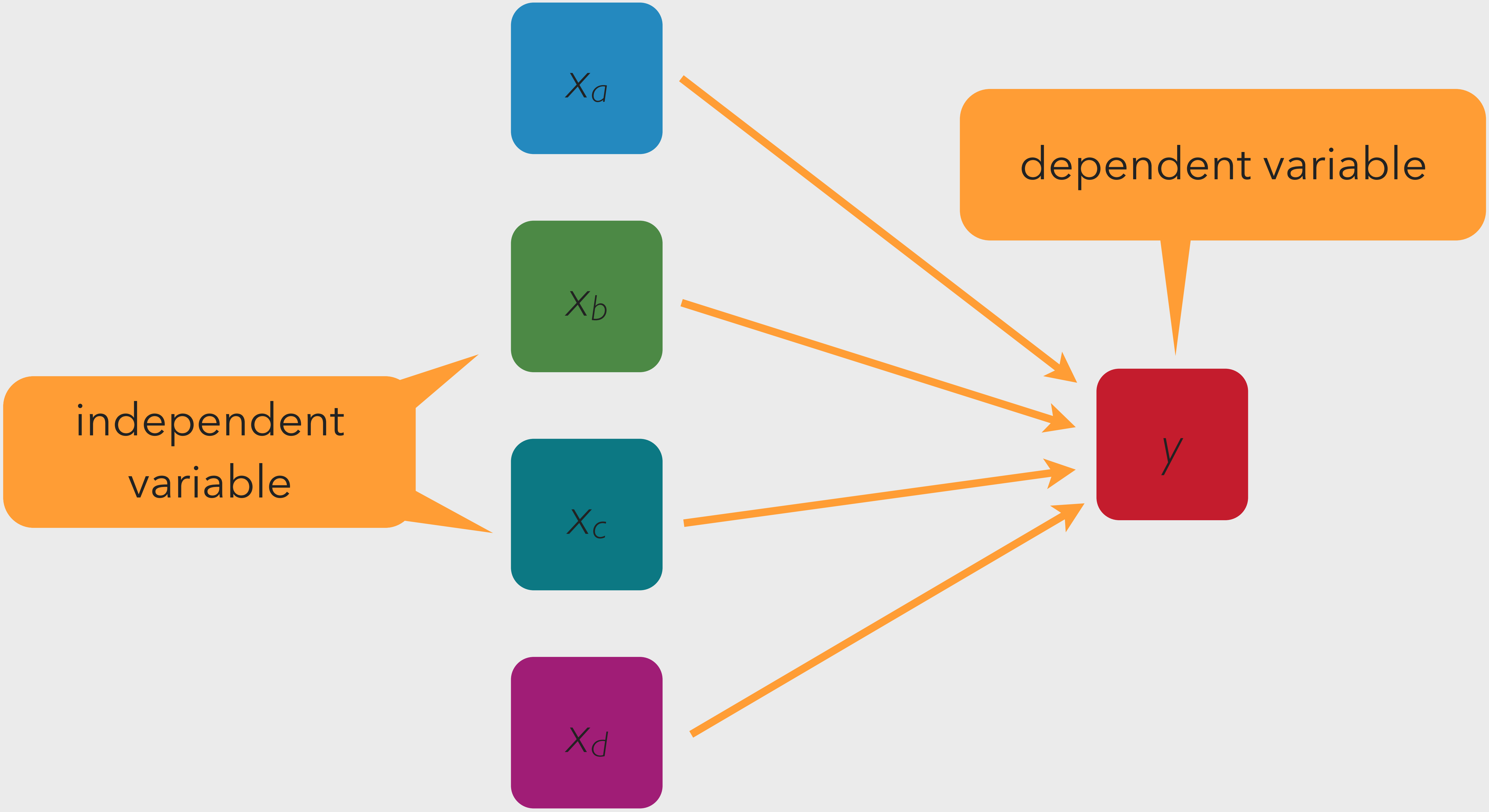
DIFFERENCE OF MEANS



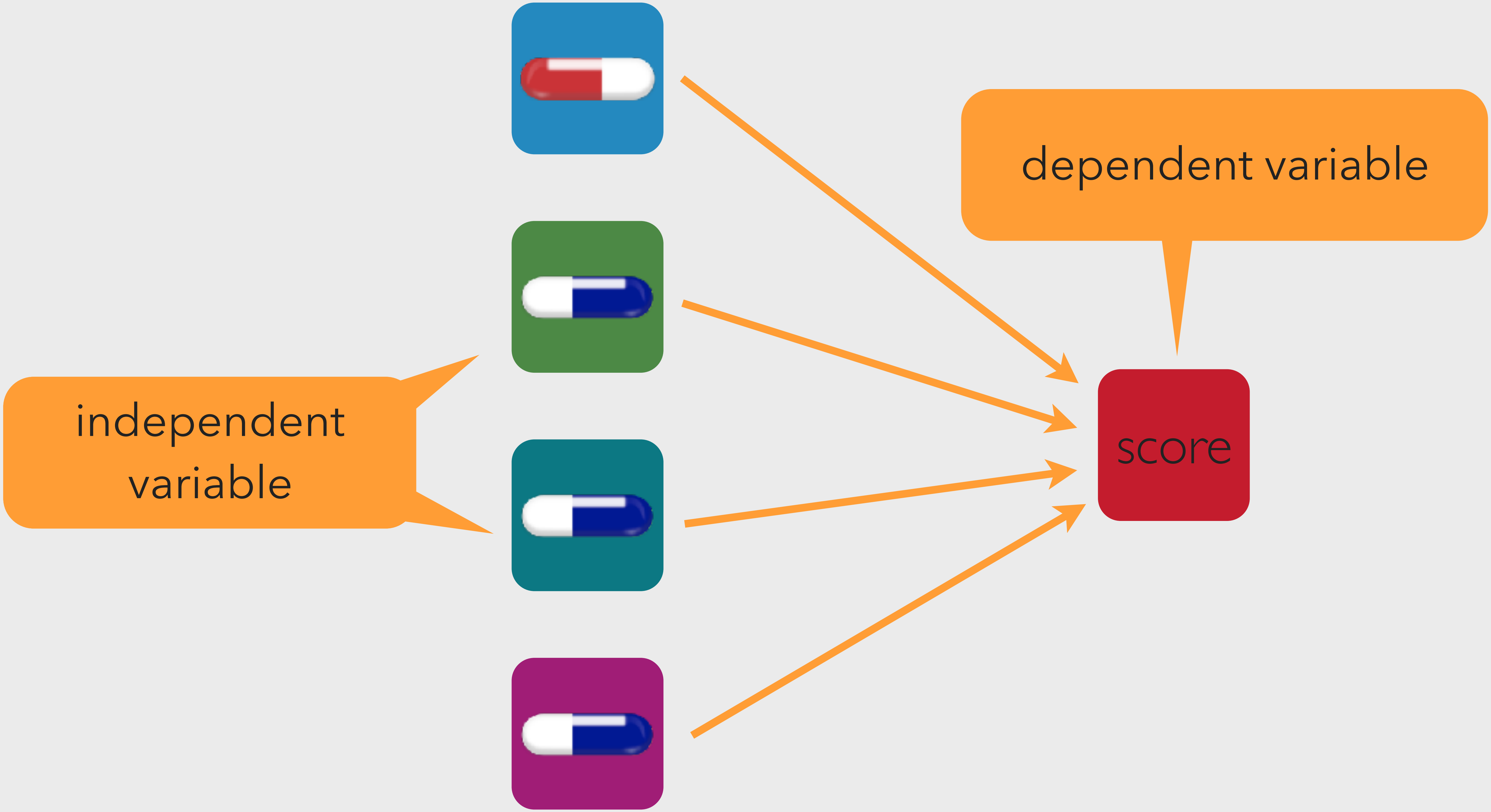
DIFFERENCE OF MEANS



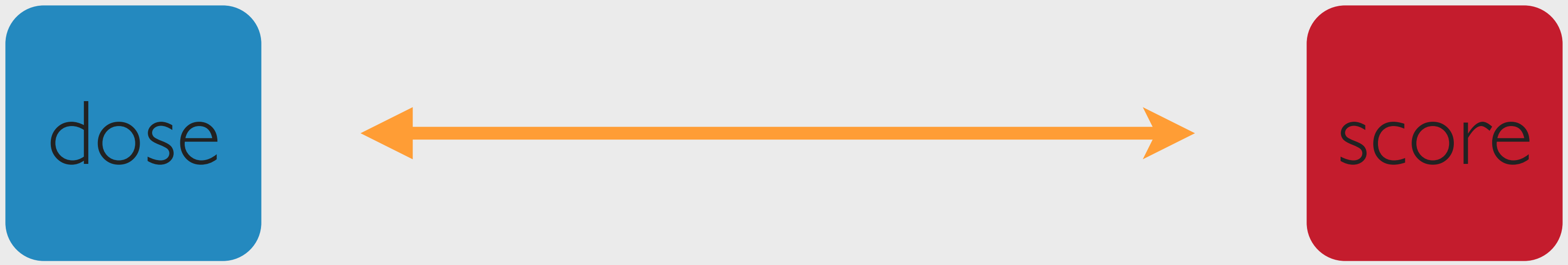
ANOVA



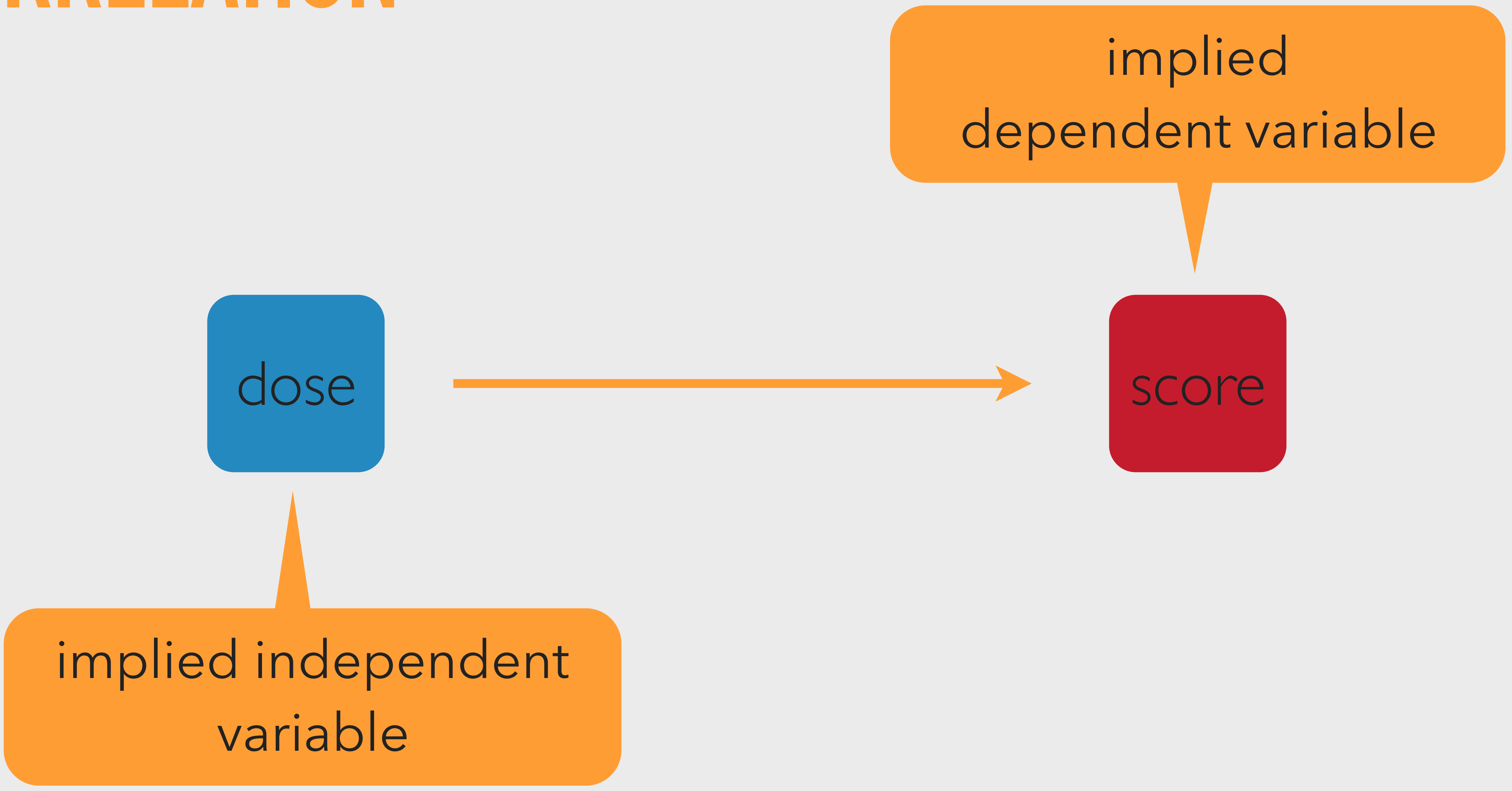
ANOVA



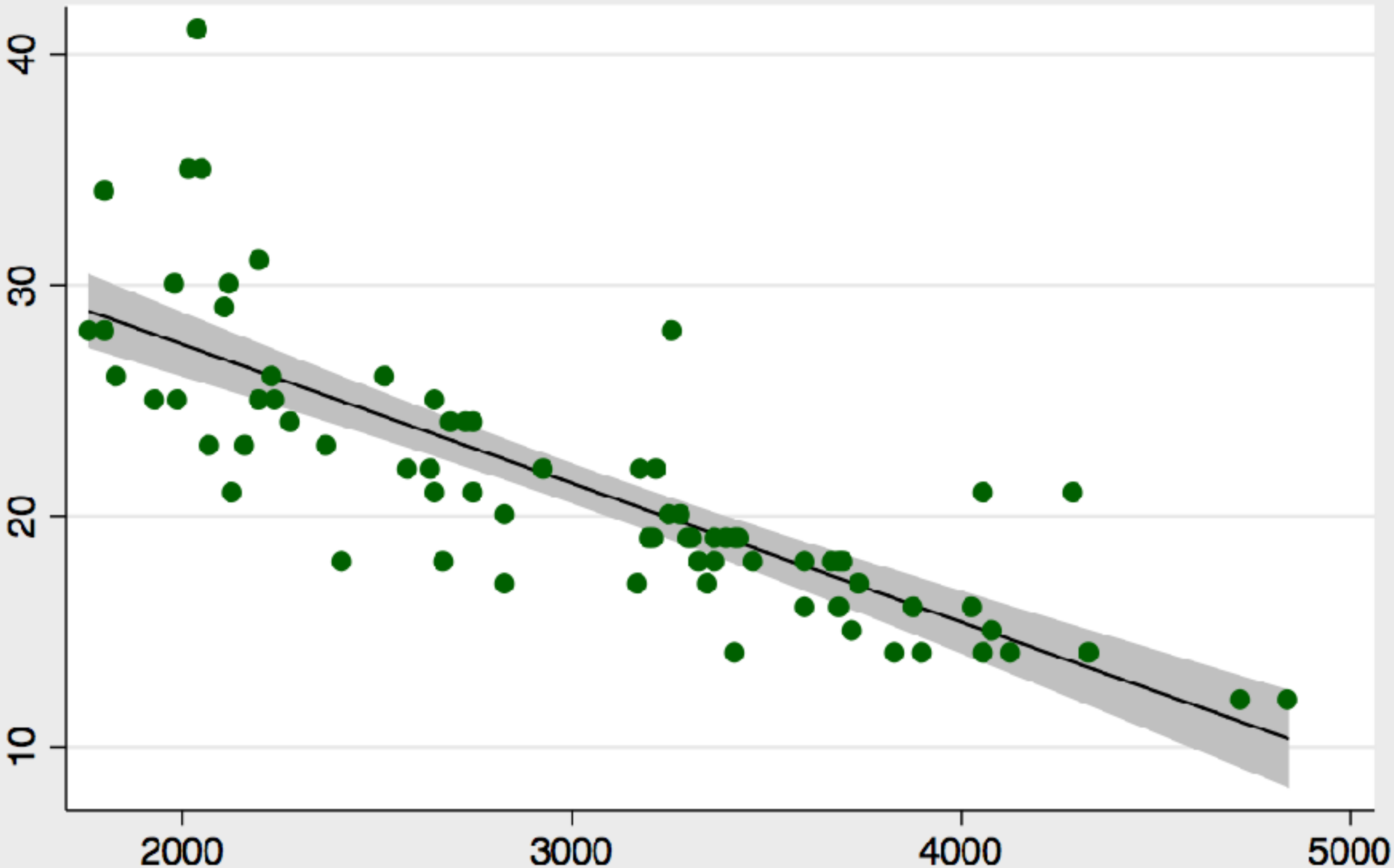
CORRELATION



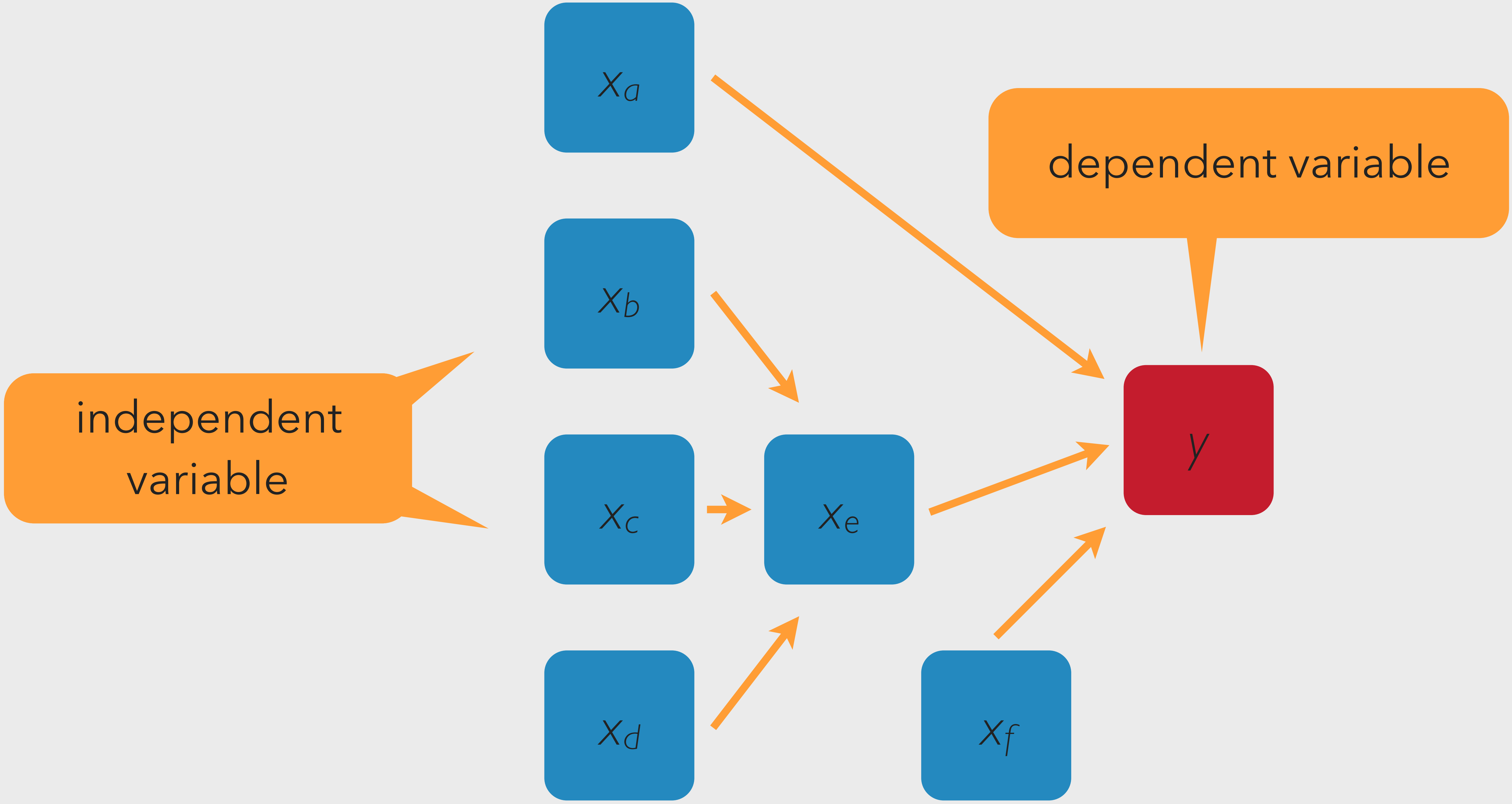
CORRELATION



CORRELATION

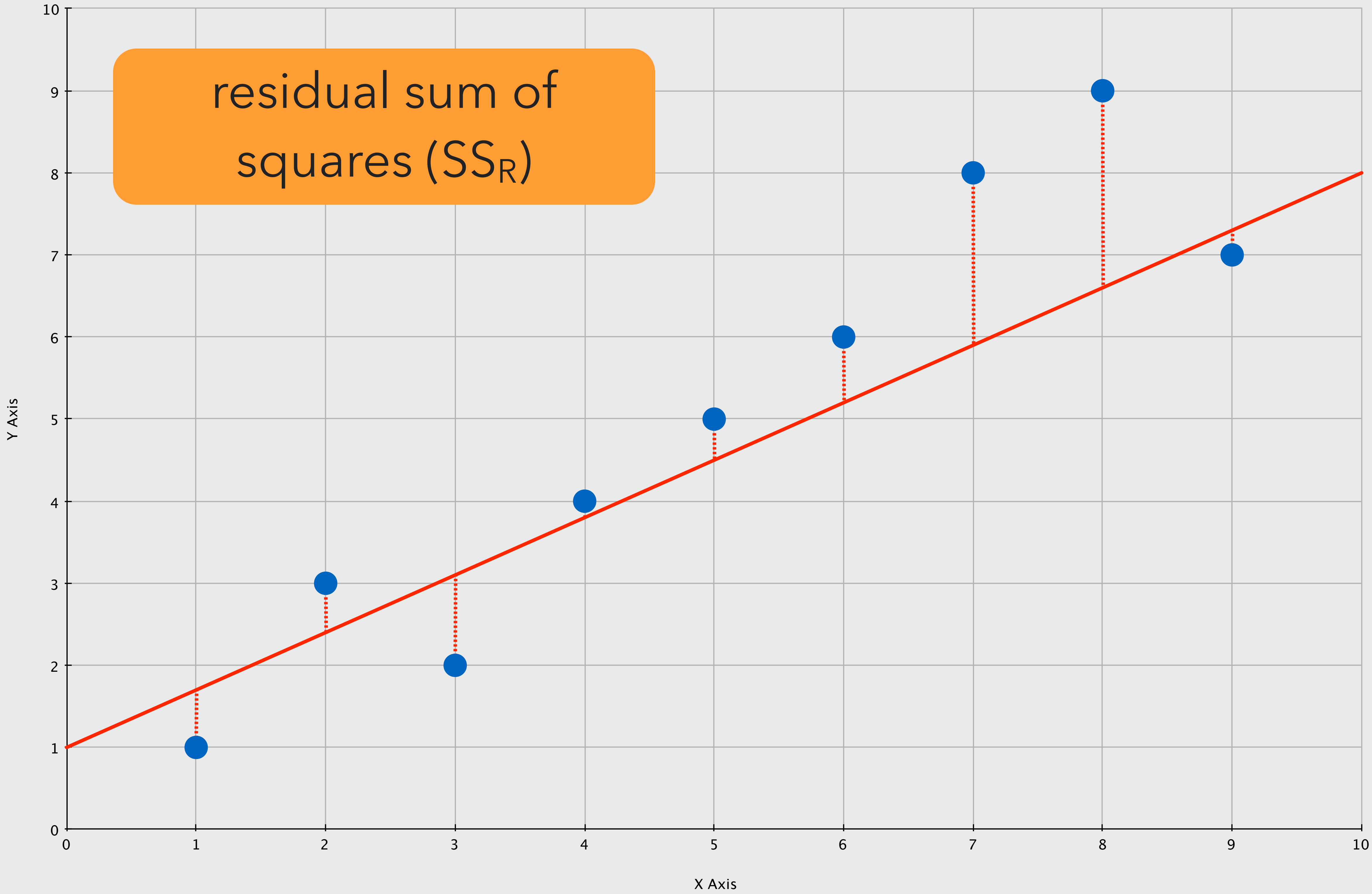


THE “REAL” WORLD

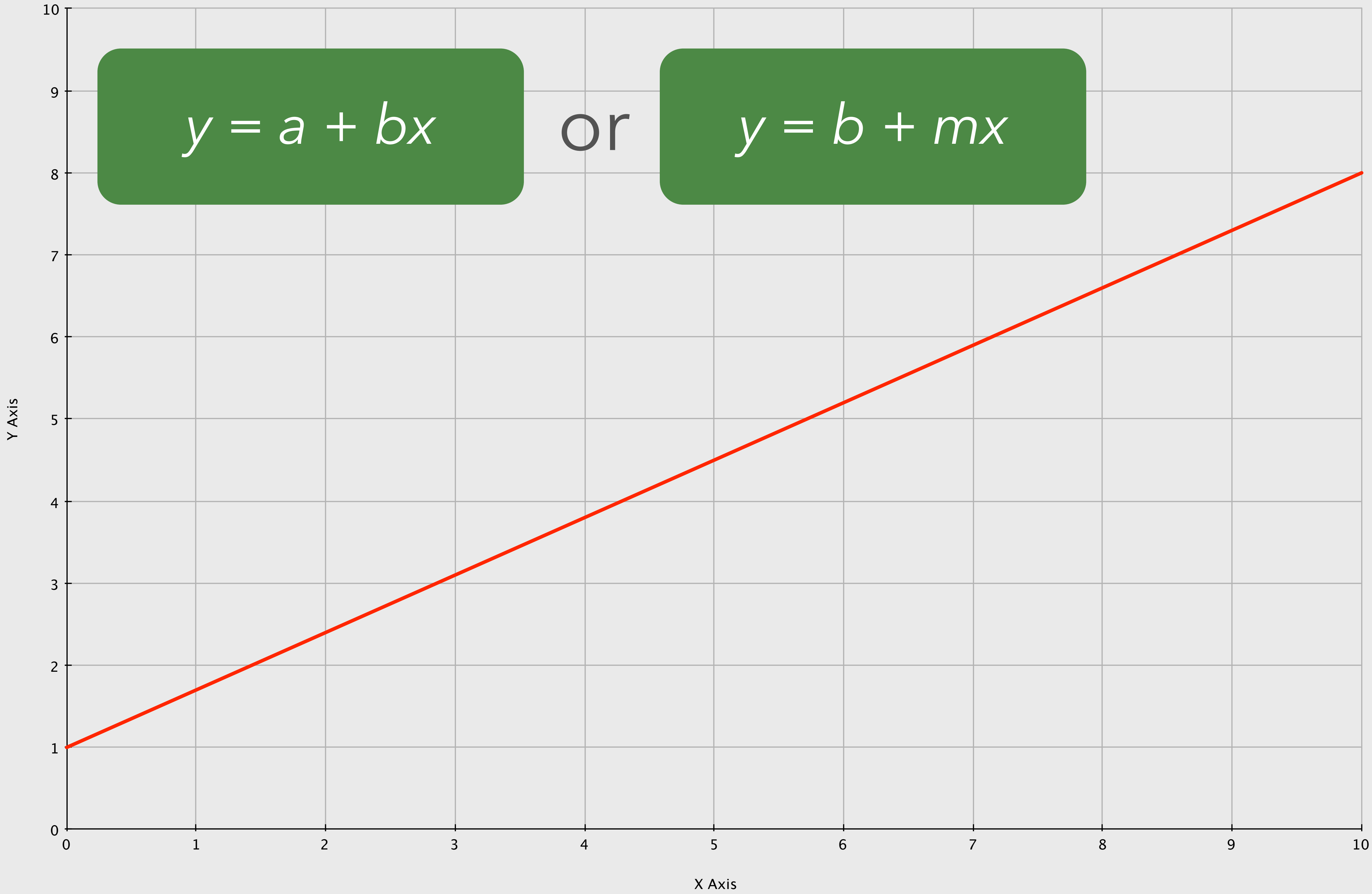


4 BIVARIATE REGRESSION THEORY

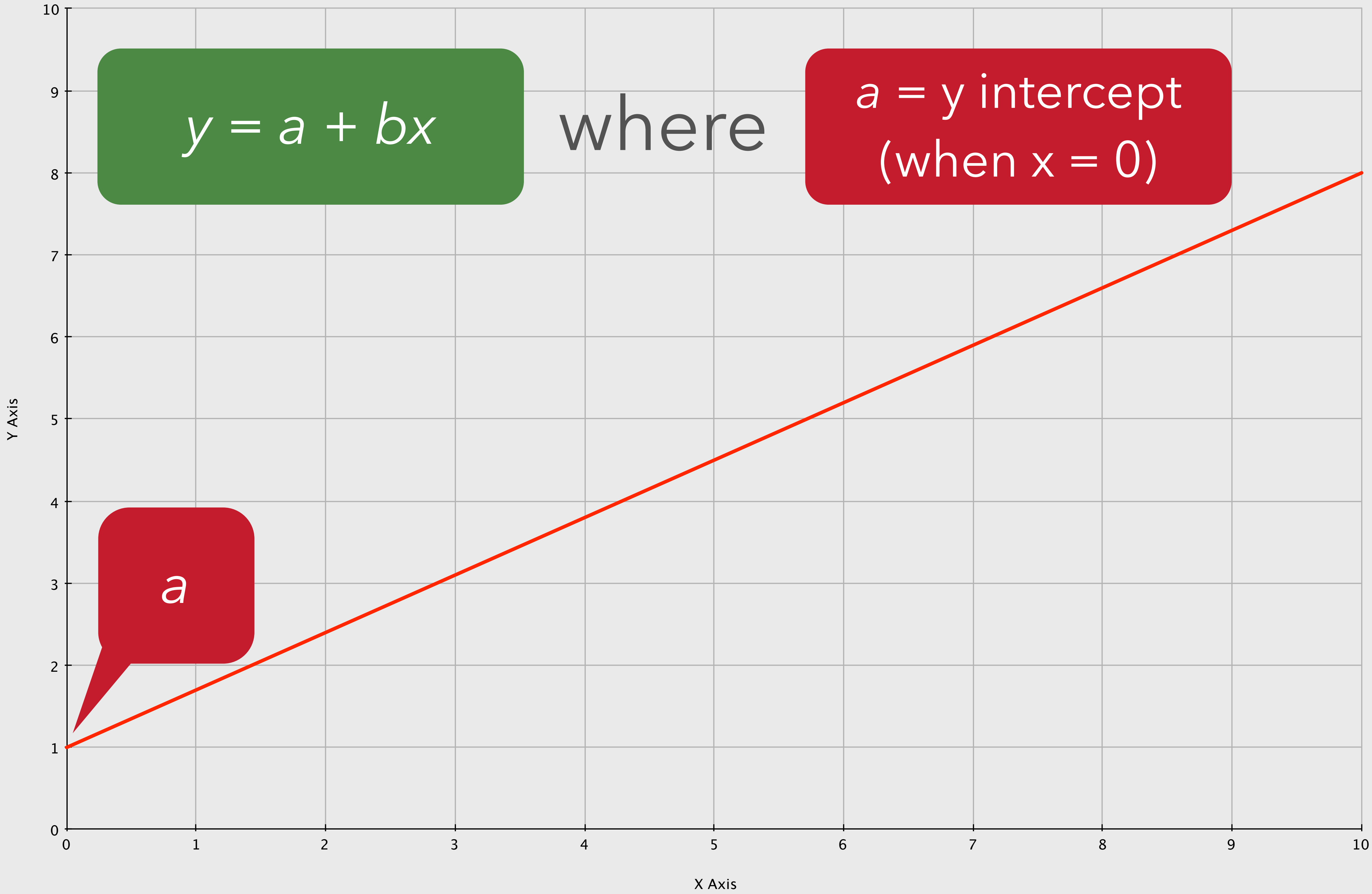
THE GOAL OF OLS REGRESSION



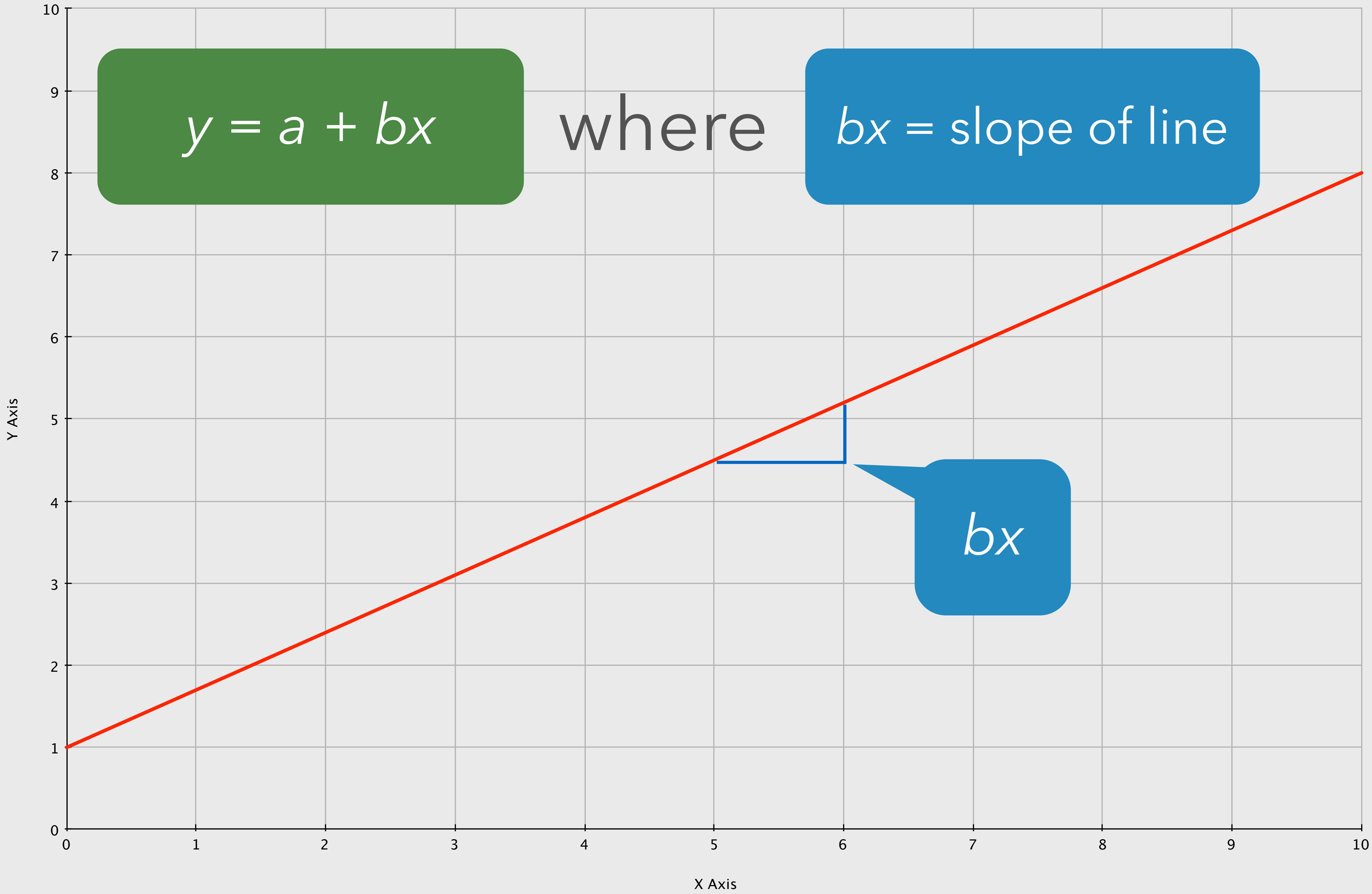
OLS REGRESSION BASICS



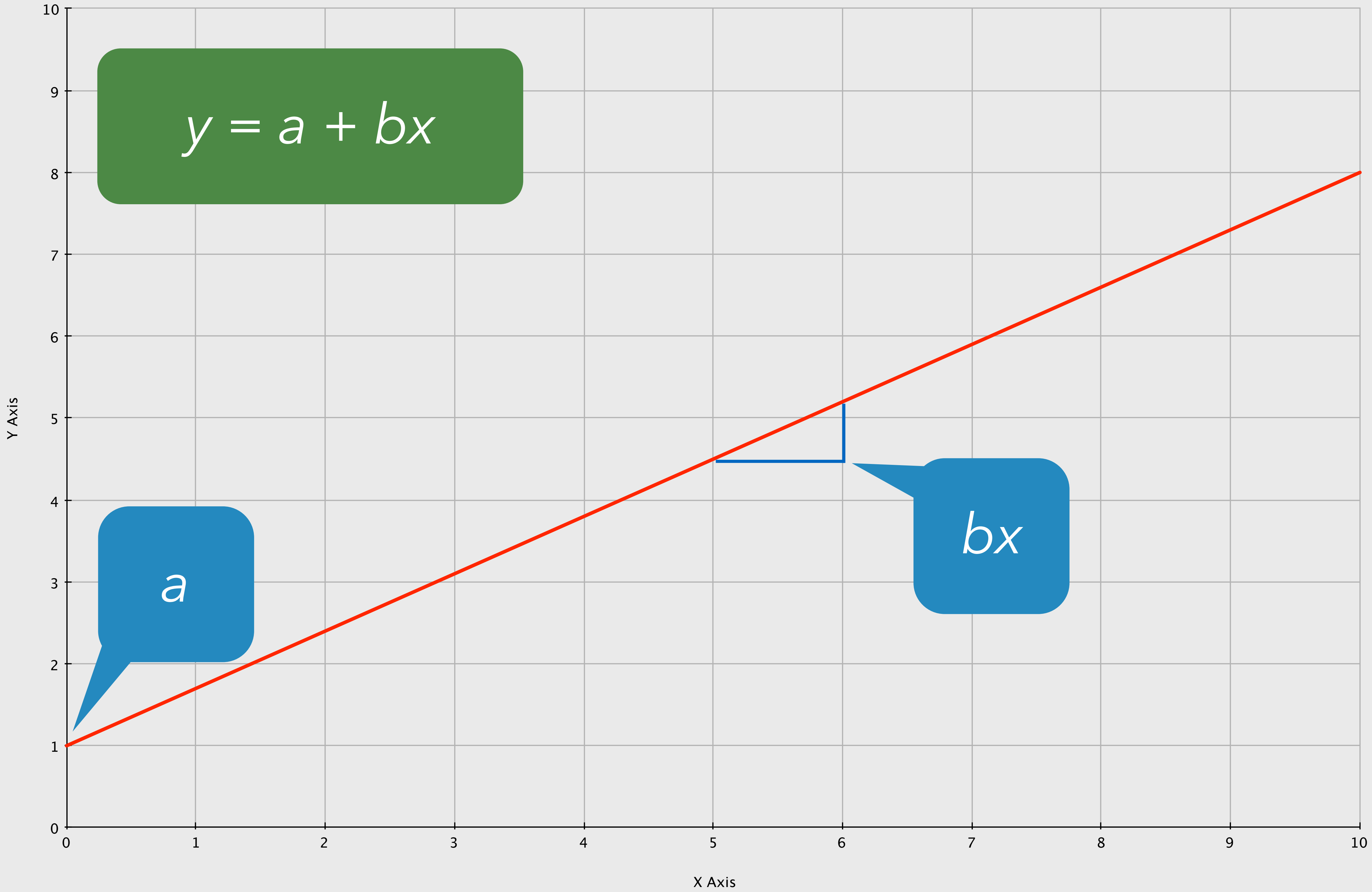
OLS REGRESSION BASICS



OLS REGRESSION BASICS



OLS REGRESSION BASICS



OLS REGRESSION BASICS

$$y = a + bx$$

y is the dependent variable (DV) in regression analysis

$$y = b_0 + b_i x_i + \varepsilon$$

b_0 is called the ‘constant’ rather than the ‘intercept’

OLS REGRESSION BASICS

subscript used because the slope of y is dependent on multiple factors

$$y = a + bx$$

$$y = b_0 + b_i x_i + \varepsilon$$

ε is included because we are estimating the line, there may be unexplained variation in y

MODEL BUILDING

y = dependent variable
b_0 = constant
x_i = independent variable i
b_i = beta value of IV i

DV = height
IV = gender
(where 0 = male
& 1 = female)

$y = b_0 + b_i x_i + \varepsilon$

$$y_{height} = b_0 + b_1 x_{female} + \varepsilon$$

MODEL BUILDING

DV

constant

IV

error

$$y_{height} = b_0 \quad \varepsilon$$

constant is "male"

MODEL BUILDING

$y = b_0 + b_i x_i + \varepsilon$

y = dependent variable
b_0 = constant
x_i = independent variable i
b_i = beta value of IV i

DV = height
IV = school year
(where 0 = pre-school,
1 = elementary,
& 2 = middle school)

$$y_{height} = b_0 + b_1 x_{elementary} + b_2 x_{middle} + \varepsilon$$

MODEL BUILDING

$y = b_0 + b_i x_i + \varepsilon$

y = dependent variable
b_0 = constant
x_i = independent variable i
b_i = beta value of IV i

DV = occupational prestige
IV = race/ethnicity
(where 0 = white,
1 = African American,
& 2 = Hispanic)

$$y_{prestige} = b_0 + b_1 x_{black} + b_2 x_{hispanic} + \varepsilon$$

SOLVING REGRESSION EQUATIONS

$$\hat{y}_{prestige} = 40 + (-5)x_{black} + (-7)x_{hispanic} + \varepsilon$$

what is the mean occupational prestige for African American respondents?

$$y_{prestige} = b_0 + b_1 x_{black} + b_2 x_{hispanic} + \varepsilon$$

SOLVING REGRESSION EQUATIONS

$$y_{prestige} = b_0 + b_1 x_{black} + b_2 x_{hispanic} + \varepsilon$$

$$\downarrow$$

$$\hat{y}_{prestige} = 40 + (-5)(1) + (-7)(0) + \varepsilon$$

what is the mean occupational prestige for African American respondents?

SOLVING REGRESSION EQUATIONS

$$y_{prestige} = b_0 + b_1 x_{black} + b_2 x_{hispanic} + \varepsilon$$

$$\hat{y}_{prestige} = 40 + (-5) + 0 + \varepsilon$$

what is the mean occupational prestige for African American respondents?

SOLVING REGRESSION EQUATIONS

$$y_{prestige} = b_0 + b_1 x_{black} + b_2 x_{hispanic} + \varepsilon$$

$$\hat{y}_{prestige} = 35$$

what is the mean occupational prestige for African American respondents?

SOLVING REGRESSION EQUATIONS

$$y_{prestige} = b_0 + b_1 x_{black} + b_2 x_{hispanic} + \varepsilon$$

$$\hat{y}_{prestige} = 40 + (-5)(1) + (-7)(0) + \varepsilon$$

what is the mean occupational prestige for Hispanic respondents?

SOLVING REGRESSION EQUATIONS

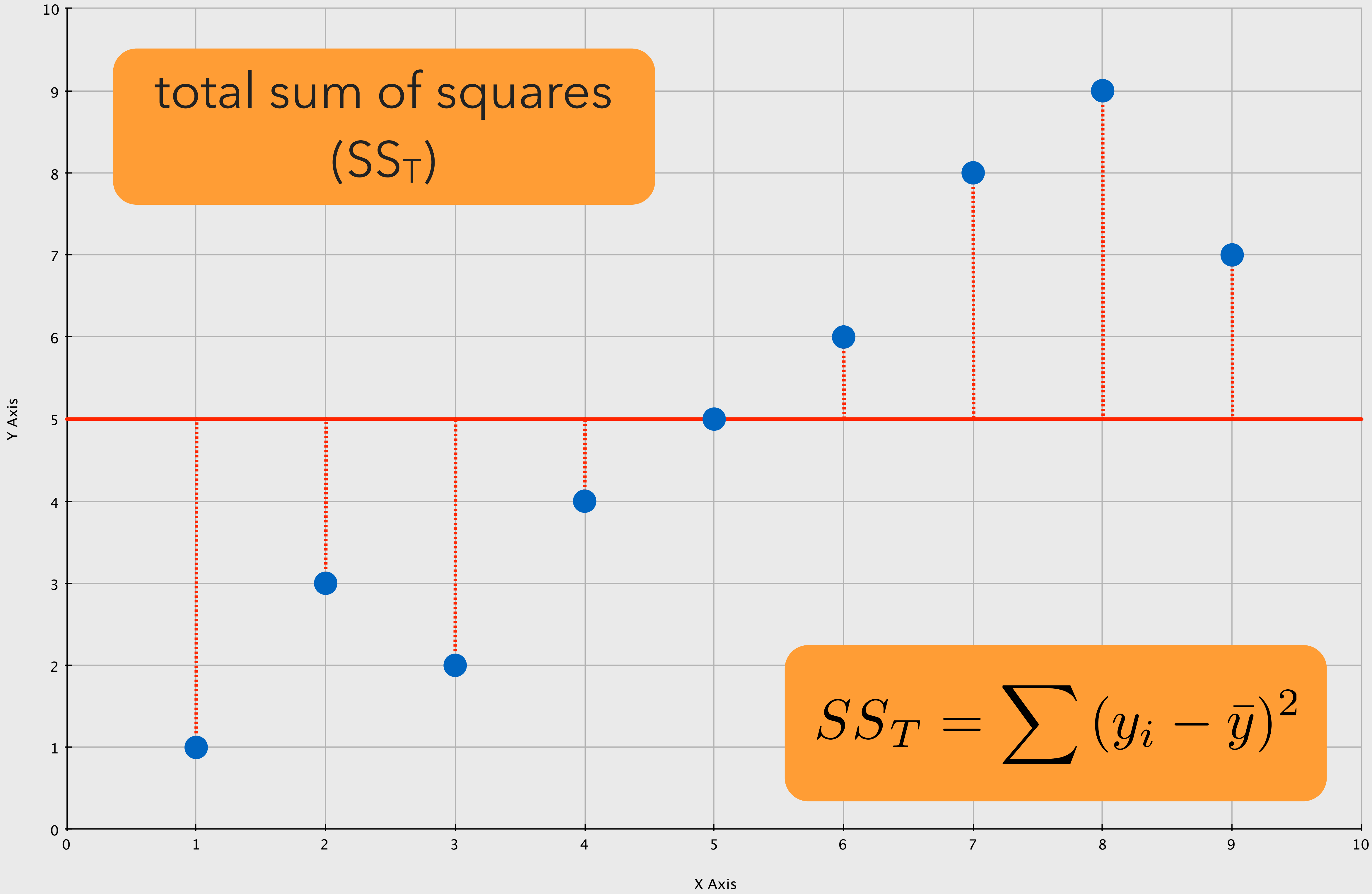
$$y_{prestige} = b_0 + b_1 x_{black} + b_2 x_{hispanic} + \varepsilon$$

$$\downarrow$$

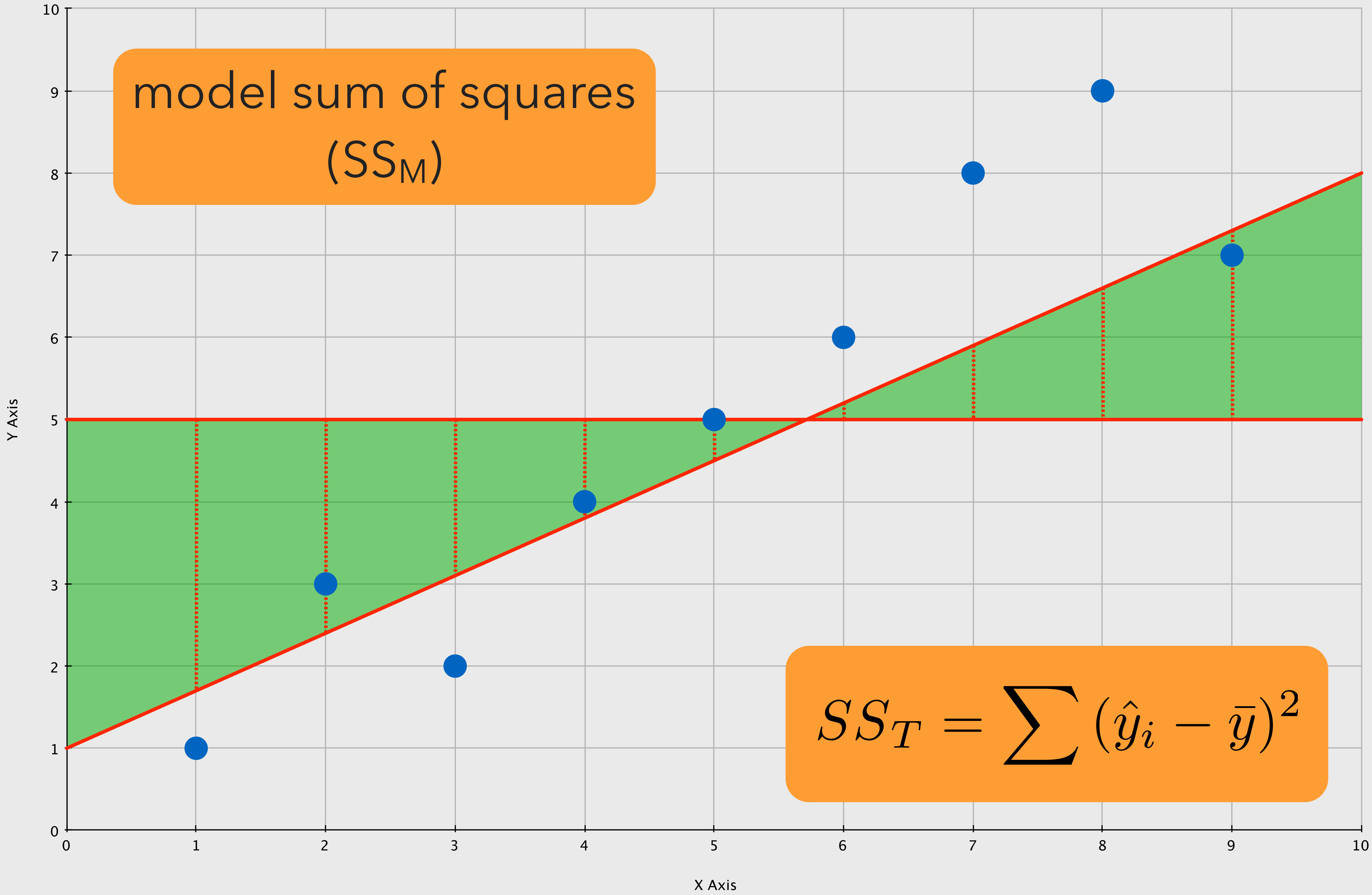
$$\hat{y}_{prestige} = 33$$

what is the mean occupational prestige for Hispanic respondents?

THE MECHANICS OF OLS REGRESSION

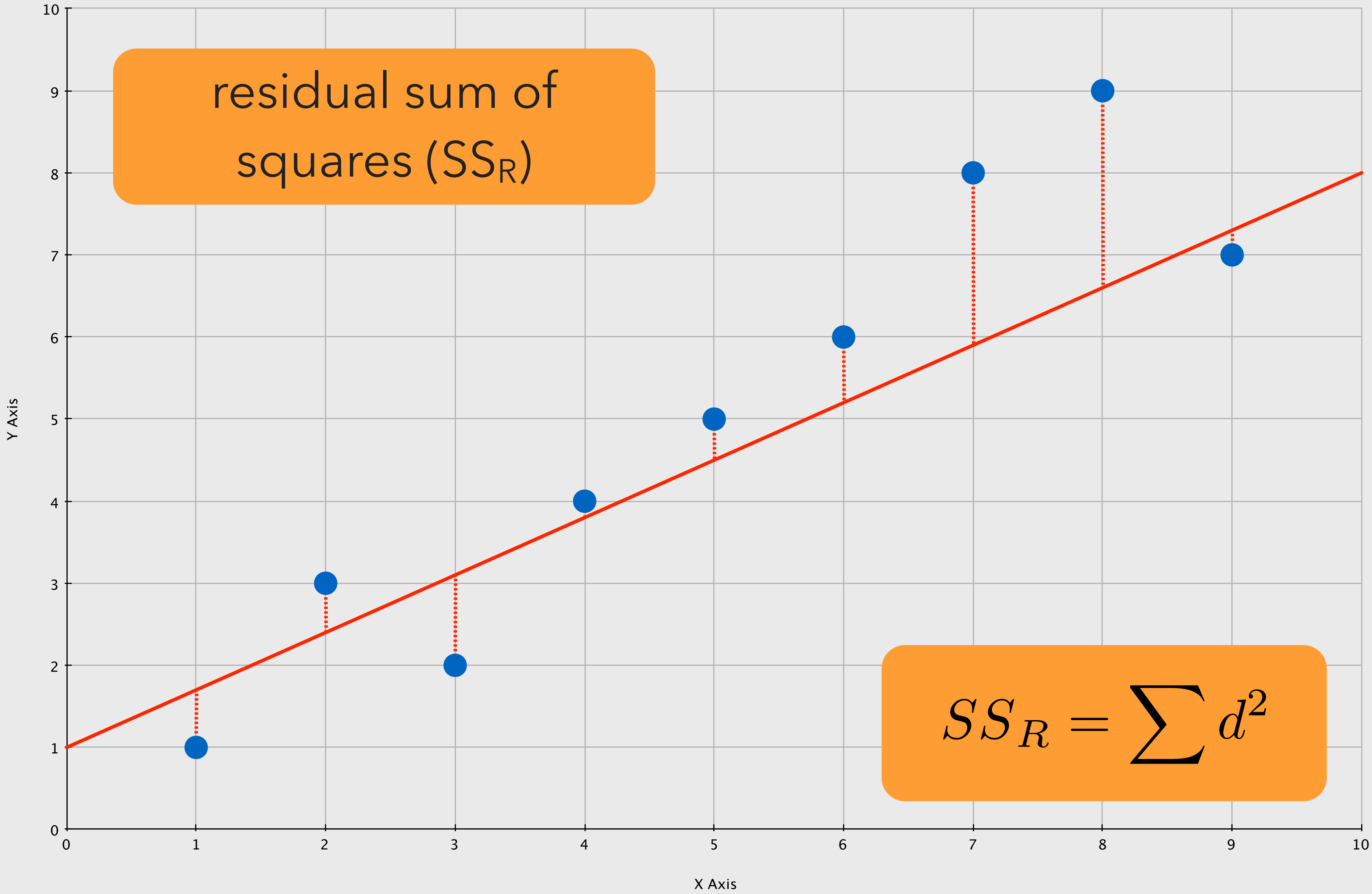


THE MECHANICS OF OLS REGRESSION



$$SS_T = \sum (\hat{y}_i - \bar{y})^2$$

THE MECHANICS OF OLS REGRESSION



EXPLAINED VARIATION

the amount of variation in *y* that our model explains, known as r^2

$$r^2 = \frac{SS_M}{SS_T}$$

ASSUMPTIONS

- DV must be continuous
- IV may be:
  - Dummy
  - Ordinal
  - Continuous
- IV's must have a variance > 0
- Relationship is linear
- DV should be normal
- No significant outliers

There are also assumptions about *multicollinearity* and *autocorrelation*, but we'll address these next week.

5 BIVARIATE REGRESSION IN STATA

ESTIMATING REGRESSION MODELS

```
regress yvar xvar [, robust]
```

```
. regress mpg weight, robust

Linear regression                               Number of obs     =         74
                                                F(1, 72)          =     105.83
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6515
                                                Root MSE          =     3.4389

------------------------------------------------------------------------------
             |               Robust
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005841   -10.29   0.000     -.007173   -.0048443
       _cons |   39.44028    1.98832    19.84   0.000     35.47664    43.40393
------------------------------------------------------------------------------
```

RELATIONSHIP WITH BIVARIATE TESTS

```
. pwcorr mpg weight

             |      mpg   weight
-------------+------------------
         mpg |   1.0000
      weight |  -0.8072   1.0000
```

```
. display -0.8072^2
-.65157184
```

```
. regress mpg weight, robust

Linear regression                               Number of obs     =         74
                                                F(1, 72)          =     105.83
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6515
                                                Root MSE          =     3.4389

------------------------------------------------------------------------------
             |               Robust
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005841   -10.29   0.000     -.007173   -.0048443
       _cons |   39.44028    1.98832    19.84   0.000     35.47664    43.40393
------------------------------------------------------------------------------
```

```
. ttest mpg, by(foreign)

Two-sample t test with equal variances
------------------------------------------------------------------------------
   Group |     Obs        Mean    Std. Err.   Std. Dev.   [95% Conf. Interval]
---------+--------------------------------------------------------------------
Domestic |      52    19.82692     .657777    4.743297    18.50638    21.14747
 Foreign |      22    24.77273     1.40951    6.611187    21.84149    27.70396
---------+--------------------------------------------------------------------
combined |      74     21.2973    .6725511    5.785503     19.9569    22.63769
---------+--------------------------------------------------------------------
    diff |           -4.945804    1.362162               -7.661225   -2.230384
------------------------------------------------------------------------------
    diff = mean(Domestic) - mean(Foreign)                         t =  -3.6308
Ho: diff = 0                                     degrees of freedom =       72

    Ha: diff < 0                 Ha: diff != 0                 Ha: diff > 0
 Pr(T < t) = 0.0003         Pr(|T| > |t|) = 0.0005          Pr(T > t) = 0.9997
```

```
. regress mpg foreign, robust

Linear regression                               Number of obs     =         74
                                                F(1, 72)          =      10.26
                                                Prob > F          =     0.0020
                                                R-squared         =     0.1548
                                                Root MSE          =     5.3558

------------------------------------------------------------------------------
             |               Robust
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   4.945804   1.544418     3.20   0.002     1.867062    8.024546
       _cons |   19.82692    .660407    30.02   0.000     18.51043    21.14342
------------------------------------------------------------------------------
```

INTERPRETATIONS: MODEL FIT

```
. regress mpg weight

      Source |       SS           df       MS      Number of obs   =        74
-------------+----------------------------------   F(1, 72)        =    134.62
       Model |   1591.9902         1   1591.9902   Prob > F        =    0.0000
    Residual |  851.469256        72  11.8259619   R-squared       =    0.6515
-------------+----------------------------------   Adj R-squared   =    0.6467
       Total |  2443.45946        73  33.4720474   Root MSE        =    3.4389

------------------------------------------------------------------------------
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005179   -11.60   0.000    -.0070411   -.0049763
       _cons |   39.44028   1.614003    24.44   0.000     36.22283    42.65774
------------------------------------------------------------------------------
```

The results of the ANOVA test (f = 134.62, p < .001) indicate that the model is a good fit for the data. Overall, the model explains 64.67% of the variation in miles per gallon.

INTERPRETATIONS: MODEL FIT

```
. regress mpg weight, robust

Linear regression                               Number of obs     =         74
                                                F(1, 72)          =     105.83
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6515
                                                Root MSE          =     3.4389

------------------------------------------------------------------------------
             |               Robust
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005841   -10.29   0.000     -.007173   -.0048443
       _cons |   39.44028    1.98832    19.84   0.000     35.47664    43.40393
------------------------------------------------------------------------------
```

The results of the ANOVA test (f = 105.83, p < .001) indicate that the model is a good fit for the data. Overall, the model explains 64.67% of the variation in miles per gallon.

INTERPRETATIONS: CONTINUOUS PREDICTORS

```
. regress mpg weight, robust

Linear regression                               Number of obs     =         74
                                                F(1, 72)          =     105.83
                                                Prob > F          =     0.0000
                                                R-squared         =     0.6515
                                                Root MSE          =     3.4389

------------------------------------------------------------------------------
             |               Robust
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
      weight |  -.0060087   .0005841   -10.29   0.000     -.007173   -.0048443
       _cons |   39.44028    1.98832    19.84   0.000     35.47664    43.40393
------------------------------------------------------------------------------
```

The model finds that a unit increase in weight is associated with a -0.006 (p < .001) miles per gallon decrease. Heavier cars are, on average, less fuel efficient than lighter cars.

INTERPRETATIONS: BINARY PREDICTORS

```
. regress mpg foreign, robust

Linear regression                               Number of obs     =         74
                                                F(1, 72)          =      10.26
                                                Prob > F          =     0.0020
                                                R-squared         =     0.1548
                                                Root MSE          =     5.3558

------------------------------------------------------------------------------
             |               Robust
         mpg |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     foreign |   4.945804   1.544418     3.20   0.002     1.867062    8.024546
       _cons |   19.82692    .660407    30.02   0.000     18.51043    21.14342
------------------------------------------------------------------------------
```

The model finds that foreign cars are associated with a 4.946 ($p < .001$) miles per gallon increase. Foreign cars are, on average, more fuel efficient than domestic cars..

DOCUMENT DETAILS

Document produced by Christopher Prener, Ph.D for the Saint Louis University course SOC 5050: QUANTITATIVE ANALYSIS - APPLIED INFERENTIAL STATISTICS. See the course wiki and the repository README.md file for additional details.