

PBAF 528

Part I:

Correlation, Linear Relationships, & Causality

Part II:

Regression Analysis and the Research Process

Part I: Correlation, Linear Relationships, & Causality

- The Purpose of Research
- Correlation: Bivariate Linear Relationships
- Causation
- Motivating Example
 - Travel Time to Campus
 - Are travel time and distance from campus correlated?
 - Does one cause the other?
- Computer Directions
- Syllabus and Course Design

The Purposes of Research

- What is research?
- Why do we undertake it?

- Investigating a hunch
- Is what I think true really true?
- To build on previous knowledge.
- To find out why things happen
- To explore relationships—are two things related?
- To investigate the effectiveness of an intervention.
- To learn about something you don't know about
- To justify what you are doing? To persuade someone.
- To document conditions

Example

Time and Distance to Campus

- Assignment 1
 - Do I1 and I2

Correlation

- A starting point to explore whether two variables are linearly related:

Positively related (**positive correlation**)

- When x is high, y is usually high
- When x is low, y is usually low

Negatively related (**negative correlation**)

- When x is high, y is usually low
- When x is low, y is usually high

Coefficient of Correlation

Measures the degree of linear association.

ρ population correlation coefficient; it equals between -1 and $+1$

r sample correlation coefficient; it equals between -1 and $+1$

0 means that there is no relationship.

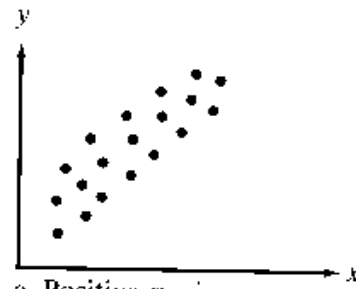
1 or -1 means a very strong linear relationship.

Correlation

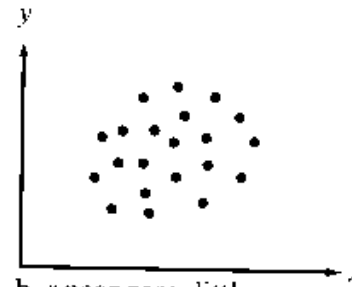
Correlations with Household Income

	HH income
Poverty Level	.770
Hours worked per week	.069
# of phone lines	.139
# of people living in house	.041
Age	-.020
Wages per week	.357

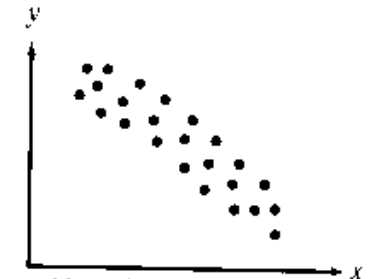
Correlation some examples



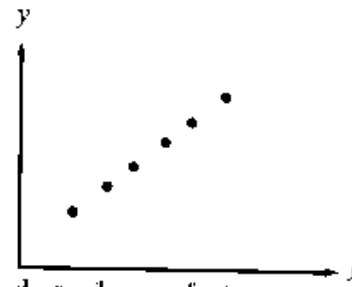
a. Positive r : y increases as x increases



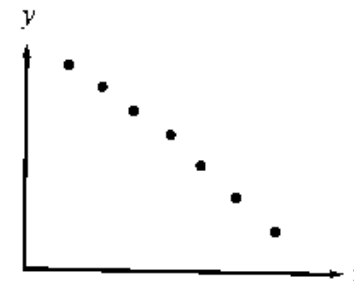
b. r near zero: little or no relationship between y and x



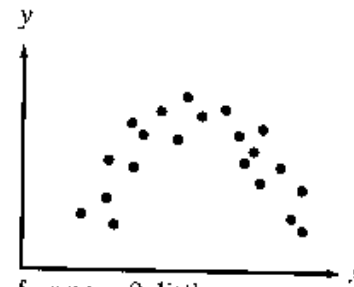
c. Negative r : y decreases as x increases



d. $r = 1$: a perfect positive relationship between y and x



e. $r = -1$: a perfect negative relationship between y and x



f. r near 0: little or no linear relationship between y and x

Part I.3-I.5

Are time and distance correlated?

$$H_0: \rho = 0$$

$$H_a: \rho \neq 0$$

Decision Rule: if $t > t_{\alpha}$ then **reject the null hypothesis.**
(we can be _____ confident that _____)

Critical Value: at $n-2$ df ?

Calculate r

$$r = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

Test Statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

Conclusion?

Calculate r

Student	X (distance to campus in miles)	Y (time to campus in minutes)	XY	X ²	Y ²
Sum					

TABLE B-1: Critical Values of the t-Distribution

Degrees of Freedom	Level of Significance				
	One Sided: 10% Two Sided: 20%	5% 10%	2.5% 5%	1% 2%	0.5% 1%
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947
16	1.337	1.746	2.120	2.583	2.921
17	1.333	1.740	2.110	2.567	2.898
18	1.330	1.734	2.101	2.552	2.878
19	1.328	1.729	2.093	2.539	2.861
20	1.325	1.725	2.086	2.528	2.845
21	1.323	1.721	2.080	2.518	2.831
22	1.321	1.717	2.074	2.508	2.819
23	1.319	1.714	2.069	2.500	2.807
24	1.318	1.711	2.064	2.492	2.797
25	1.316	1.708	2.060	2.485	2.787
26	1.315	1.706	2.056	2.479	2.779
27	1.314	1.703	2.052	2.473	2.771
28	1.313	1.701	2.048	2.467	2.763
29	1.311	1.699	2.045	2.462	2.756
30	1.310	1.697	2.042	2.457	2.750
40	1.303	1.684	2.021	2.423	2.704
60	1.296	1.671	2.000	2.390	2.660
120	1.289	1.658	1.980	2.358	2.617
(Normal) ∞	1.282	1.645	1.960	2.326	2.576

Source: Reprinted from Table IV in Sir Ronald A. Fisher, *Statistical Methods for Research Workers*, 14th ed. (copyright © 1970, University of Adelaide) with permission of Hafner, a division of the Macmillan Publishing Company, Inc.

Causation

- Conditions for establishing causation
 - ⇒ The cause precedes the effect in time
 - ⇒ Empirically correlated with each other
 - ⇒ The observed correlation cannot be explained in terms of some third variable that causes both of them
- Necessary cause
- Sufficient cause
- What other factors?

Correlation in SPSS

SPSS

ANALYZE>CORRELATE>BIVARIATE

Correlations

Correlations

		BRTHCT NL	FERTRA TE
BRTHCTNL	Pearson Correlation	1.000	-.850**
	Sig. (2-tailed)	.	.000
	N	25	25
FERTRATE	Pearson Correlation	-.850**	1.000
	Sig. (2-tailed)	.000	.
	N	25	25

** . Correlation is significant at the 0.01 level (2-tailed).

Correlation in Excel

= CORREL(array1, array2)

or

= PEARSON(array1,array2)

The screenshot shows a Microsoft Excel spreadsheet titled "Microsoft Excel - Fertrate.dat". The formula bar displays the formula `=PEARSON(A1:A27,B1:B27)`. The spreadsheet data is as follows:

	A	B	C	D	E	F
1	76	2.2		-0.86502		
2	69	2.3				
3	66	2.9				
4	71	3.5				
5	63	2.7				
6	62	3.4				
7	60	3.5				
8	55	4				
9	55	2.9				
10	50	3.1				
11	51	4.3				
12	48	4.5				
13	42	4				
14	46	5.4				
15	40	4.5				
16	40	5.5				
17	35	4.8				
18	35	5.5				
19	28	6.5				
20	24	5.5				
21	16	5.8				
22	14	6				
23	13	5				
24	13	6.5				
25	10	4.8				
26	9	7				
27	7	5.7				
28						
29						
30						

Recap Today, Part I

- The Purpose of Research
- Correlation: Bivariate Linear Relationships
- Causation
- Computer Directions
- Syllabus and Course Design

TAKE A BREAK

Part II: Regression Analysis and the Research Process

- SO FAR
 - Correlation
 - Linear Relationships
 - Conditions needed to establish causation
- Part II
 - Simple Linear Regression
 - Final Project and the Research Process

Are time and distance correlated?

$H_0: \rho=0$ Decision Rule: if $t > t_\alpha$ then reject the null hypothesis.
Critical Value of t at $n-2$ df ?
 $H_a: \rho \neq 0$

Calculate r

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Test Statistic:

$$t = \frac{r}{\sqrt{\frac{1-r^2}{n-2}}}$$

For the entire class, do we have enough evidence to reject the null hypothesis?

$$r = \frac{\sum xy - \frac{(\sum x)(\sum y)}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Class Data and Correlation

$\alpha = .05$

$n = ?$, $n - 2 = ?$

$t_{\alpha} = ?$ at ? df

$r = ?$

$t = ?$

$p = ?$

Simple Linear Regression

Algebra

$$Y = MX + B$$

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

usually

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

(Regress Y on X)

For a sample:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$$

Simple Linear Regression

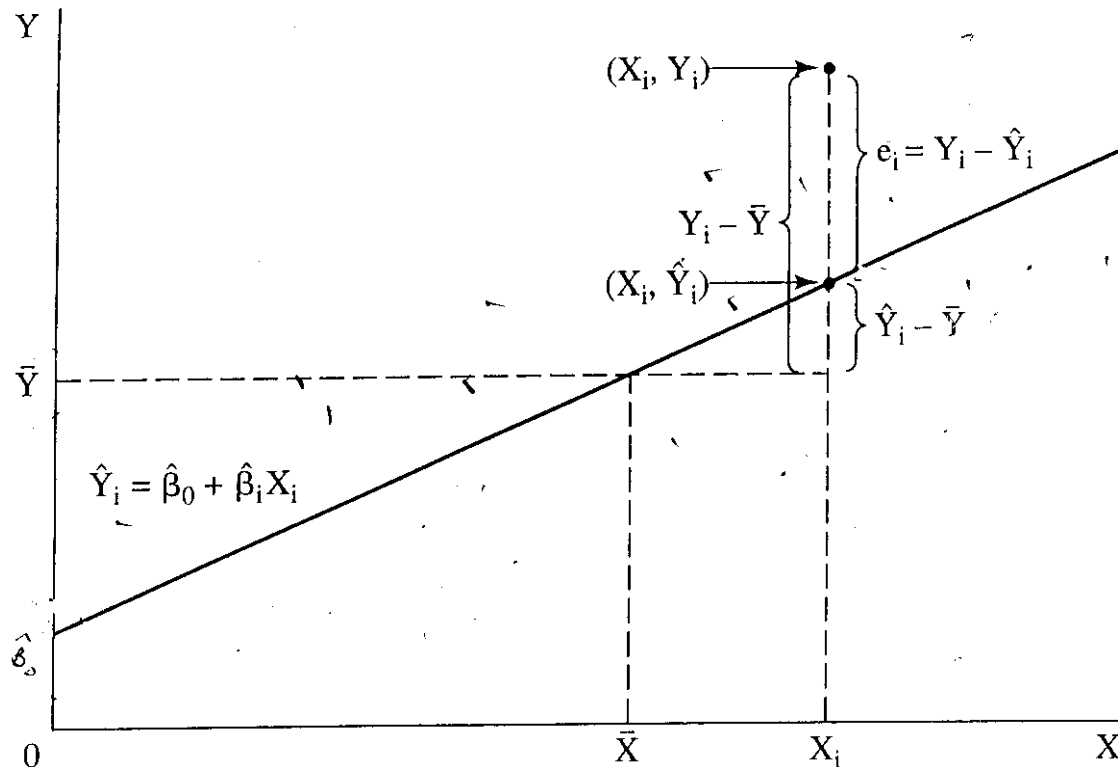


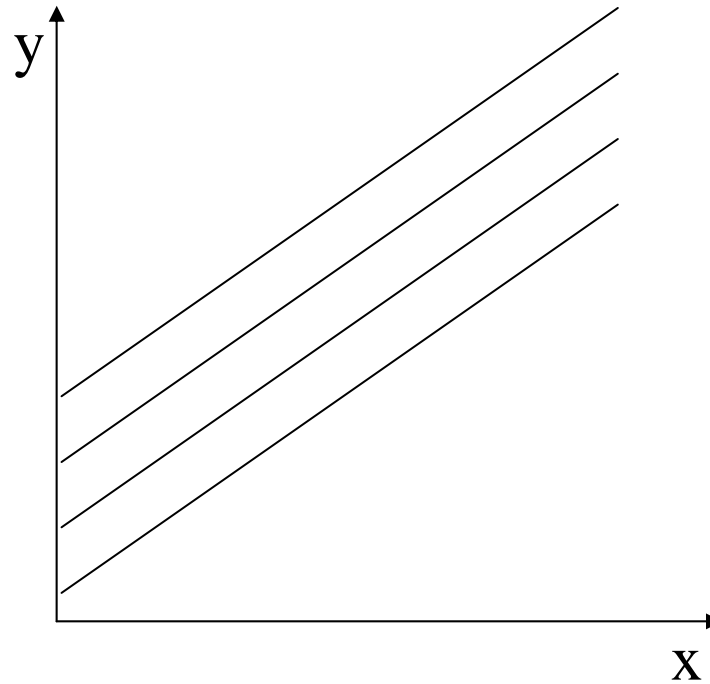
Figure 2.1 DECOMPOSITION OF THE VARIANCE IN Y

The variation of Y around its mean ($Y_i - \bar{Y}$) can be decomposed into two parts: 1) $(\hat{Y}_i - \bar{Y})$, the difference between the estimated value of Y (\hat{Y}) and the mean value of Y (\bar{Y}); and 2) $(Y_i - \hat{Y}_i)$, the difference between the actual value of Y and the estimated value of Y .

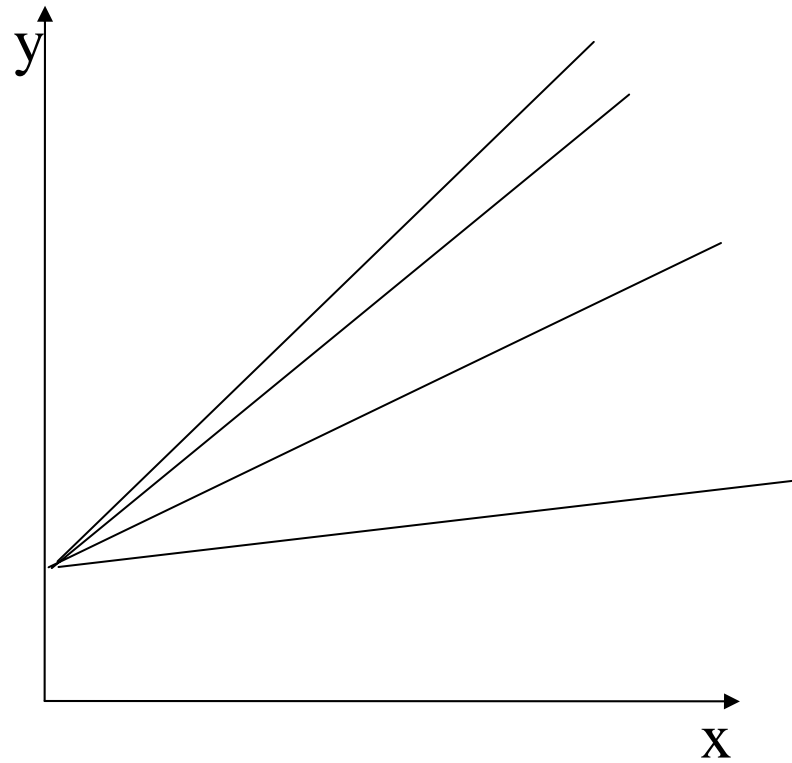
Simple Linear Regression

- For a sample $Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i$
- Assumptions
 - ⇒ X and Y are linearly related
 - ⇒ Randomness in Y comes from the error term
 - ⇒ Errors are normally distributed
 - ⇒ Errors are not related to each other

The lines can have different intercepts, :



The lines can have different slopes, :



Simple Linear Regression

- To estimate coefficients

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

$$\hat{\beta}_1 = \frac{SS_{xy}}{SS_{xx}} = \frac{\sum_{i=1}^n [(x_i - \bar{x}) \cdot (y_i - \bar{y})]}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sum x_i^2 - \frac{(\sum x_i)^2}{n}}$$

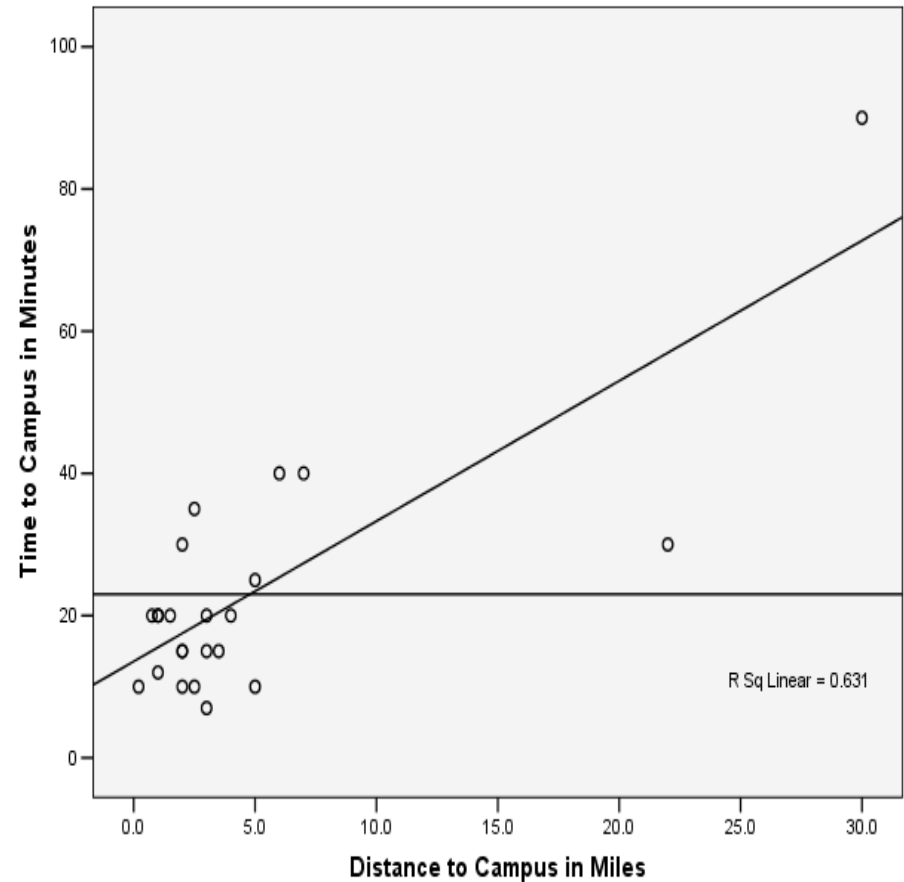
Time & Distance From Campus

$$\hat{Y}_i = 13.571 + 1.972 X_i$$

$$\bar{Y} = 23$$

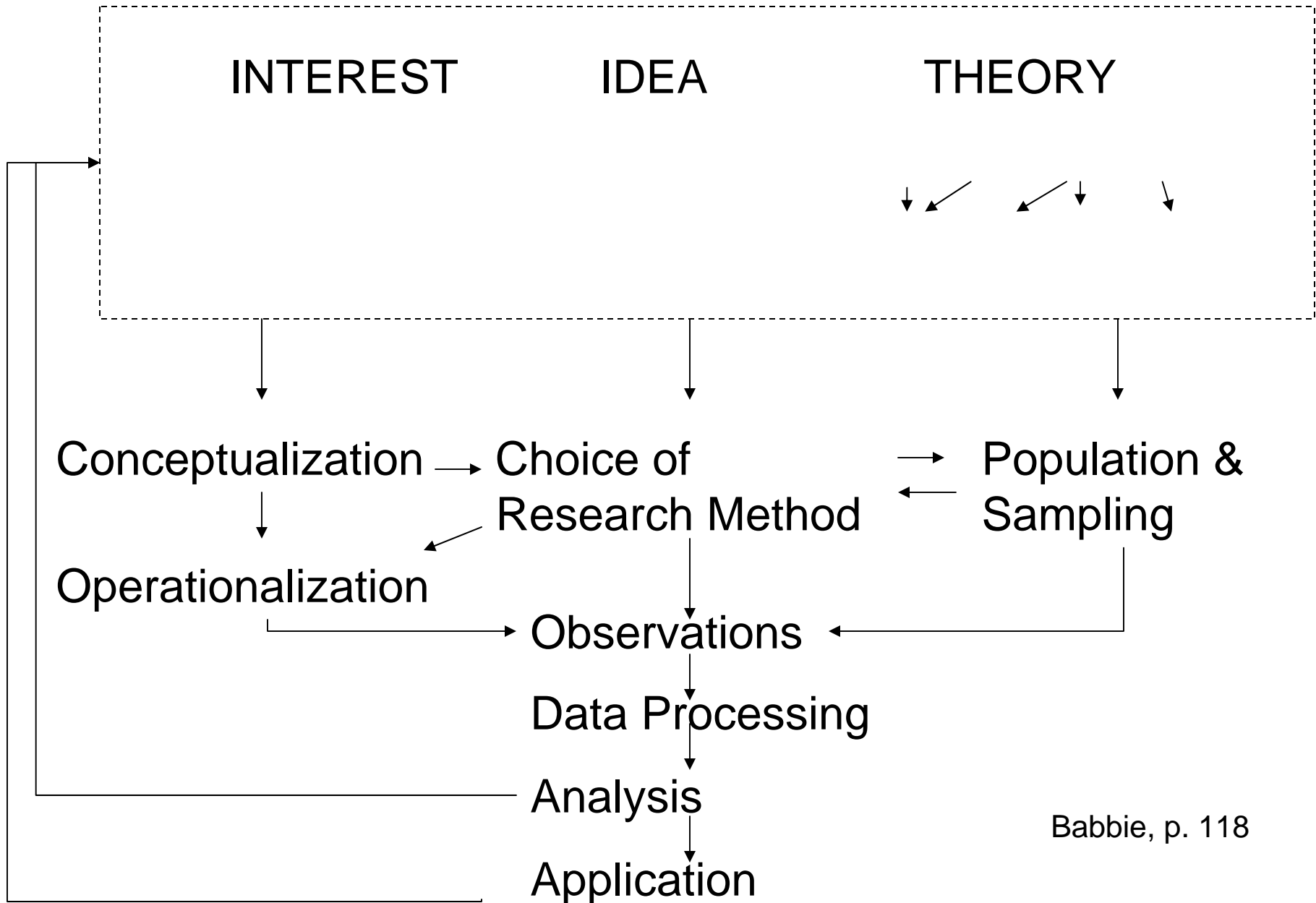
$$(Y_i - \bar{Y}) = (\hat{Y}_i - \bar{Y}) + (Y_i - \hat{Y}_i)$$

$$= (\hat{Y}_i - \bar{Y}) + e_i$$



Research Process & Proposal

- Formulate a question/problem
- Review the literature, develop a theoretical model
- Specify the model: independent and dependent variables
- Hypothesize about the expected signs of the coefficients
- Collect the data/operationalize
- Analysis
- Document



Babbie, p. 118

Today Recap

- Correlation
- Causation
- Simple Linear Regression
- Final Project and the Research Process
- Next time:
 - What about other causal factors?
 - Interpreting results