Introduction to Statistical Methods
SOC-GA 2332

Lecture 6: Regression Extended I

Siwei Cheng

NEW YORK UNIVERSITY

Lecture Outline

- Multiple Regression
- Types of Multivariate Relationships
- Standardized Regression Coefficients
- Regression with Dummy Variables
- Analysis of Variance (ANOVA)

Multiple Regression (Quick Review)

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \beta_k X_k + \epsilon. \quad (1)$$

- β_0 is the intercept, that is, the expected value of Y when all $X's$ are set to zero.
- β_1 is the slope on X_1, which measures the change in Y with respect to X_1, *holding other factors fixed*.
- β_k is the slope on X_k, which measures the change in Y with respect to X_k, *holding other factors fixed*.

Testing whether *any* of the explanatory variables are statistically different from zero

- Here, we test whether the explanatory variables collectively have an effect on y.
- That is, we test the null hypothesis H_0: $\beta_1 = \beta_2 ... = \beta_k = 0$ (i.e. y independent of all the explanatory variables) against the alternative hypothesis H_a: At least one $\beta_j \neq 0$.
- The test judges whether using the linear regression function with $x_1, x_2, ...x_k$ together to predict y is better than using $\bar{y}$.

The F Statistic

- The test statistic (F-statistic) is written as:

$$F = \frac{R^2/k}{(1-R^2)/(n-k-1)}, \quad \frac{\text{explained v.}}{\text{unexplained v.}} \tag{2}$$

where R^2 is the coefficient of multiple determination (varying from zero to one), similar to what we have defined in the binary regression case:

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{\Sigma(y-\bar{y})^2 - \Sigma(y-\hat{y})^2}{\Sigma(y-\bar{y})^2} \tag{3}$$

- We can think of the F-statistic as the **explained** variance divided by the **unexplained** variance.

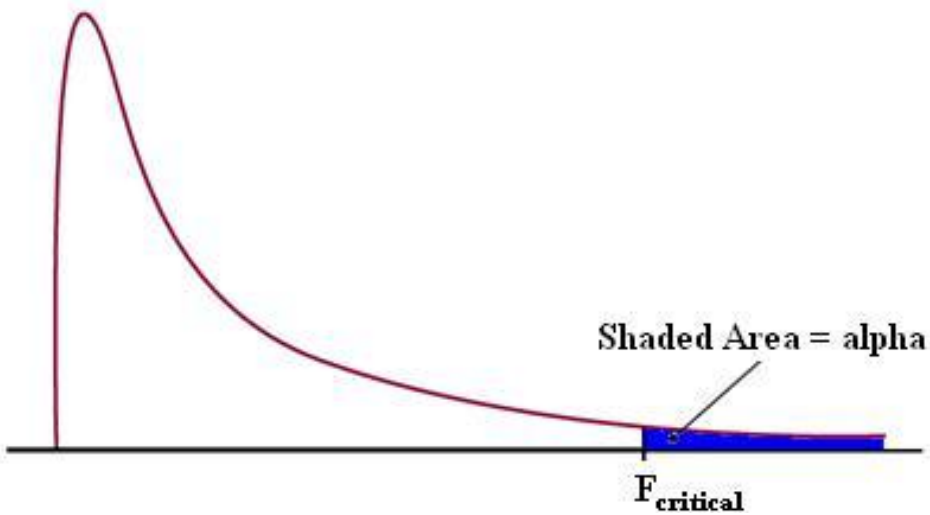
The F Statistic

n = sample size k = # of predictors

- When H_0 is true, the F-statistic follows F distribution with $df = (k, n - k - 1)$.

$df_1 = k$

$df_2 = n - k - 1$



- When the calculated F-statistic is greater than the critical value, we reject the null hypothesis, and the evidence is leaning towards our alternative hypothesis (at least one of the $\beta's$ is non-zero).

The F Statistic for Nested Models

- Two models are **nested** if both contain the same terms and one has at least one additional term.

$$\text{Restricted model: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon. \quad (4)$$

$$\text{Full model: } Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon. \quad (5)$$

- The first model is **nested** within second model.
- How do we decide whether the more complex (full) model contributes additional information about the association between y and the predictors?
- This is equivalent to testing the null hypothesis H_0: $\beta_3 = \beta_4 = 0$ against the alternative hypothesis H_a: At least one $\beta_j \neq 0, j = 3$ or 4..

The F Statistic for Nested Models

Seeing if full model adds anything

- The test statistic can be written as:

$$F = \frac{(SSE_R - SSE_F)/p}{SSE_F/(n-(k+p+1))} \quad (6)$$

- where k is the number of predictors in the restricted model and p is the number of **additional** predictors added in the full model.
- The F statistic follows an F distribution.
- Parsimonious *models are preferable to big models as long as both have similar predictive power.*
- **Important note**: If the models are not nested, we cannot use the F−test above to choose between one and another.

Coefficient of Multiple Determination

- Let's now go back a little bit to look at the coefficient of multiple determination R^2:

$$R^2 = \frac{TSS - SSE}{TSS} = \frac{\Sigma(y - \bar{y})^2 - \Sigma(y - \hat{y})^2)}{\Sigma(y - \bar{y})^2} \tag{7}$$

- R^2 equals the squared term of the correlation between outcome variable and our linear prediction: $R^2 = r_{y,\hat{y}}$.
- R^2 falls between 0 and 1. $R^2 = 1$ when all residuals are 0. $R^2 = 0$ when $\beta_1 = \beta_2 ... = \beta_k = 0$. The prediction line become a flat line with $\hat{y} = \bar{y}$.
- The larger the value of R^2, the better the set of explanatory variables collectively predict y.

Coefficient of Multiple Determination

- Note that if we add an additional explanatory variable to the model, R^2 will not decrease. That is, it is impossible to explain less variation in y by incorporating one additional variable.
- This raises an important problem: in multiple regression models, we can increase R^2 "artificially" by adding more and more explanatory variables but without meaningful interpretations.
- To account for the potentially spurious increase in R^2 by number of predictors, we often report *adjusted* R^2 as an alternative:

$$R^2_{adj} = R^2 - (1 - R^2)\frac{k}{n - k - 1}. \tag{8}$$

- The adjustment is done by placing a "**penalty**" on increase in k. To see this, note that the larger the k, the greater the term that we subtract from the original R^2.

Example: Multiple Regression

- Suppose a researcher collected a sample of 600 students with measures of students' reading test score and three psychological variables: *locus of control*, *self-concept* and *motivation*.
- The researcher is interested in knowing: (1) whether at least one of the three psychological factors has an effect on reading score; (2) whether each of the psychological variables has an affect on reading score; (3) How much variations in the student's reading score can be explained by the set of three psychological factors.

Example: Multiple Regression

- Let's summarize these variables:

```
> describe(score_minidata)
                vars   n mean    sd median trimmed   mad   min   max range  skew kurtosis   se
read               1 600 51.9 10.10  51.86   51.76 10.99 24.62 80.59 55.97  0.09    -0.25 0.41
motivation         2 600  0.0  0.82  -0.01   -0.01  0.79 -2.75  2.58  5.33  0.09    -0.07 0.03
self_concept       3 600  0.0  0.71   0.03    0.01  0.73 -2.53  2.09  4.63 -0.12    -0.09 0.03
locus_of_control   4 600  0.1  0.67   0.08    0.09  0.67 -2.00  2.21  4.20  0.08    -0.07 0.03
>
```

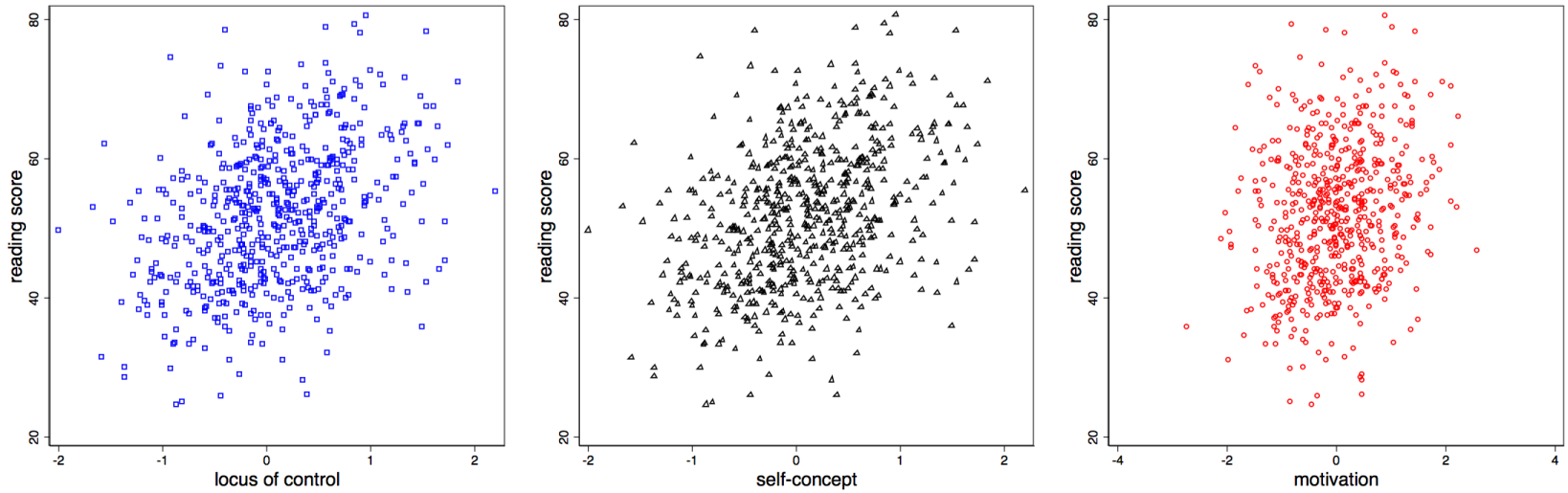
- Then we can look at the correlation between pairs of these variables.

```
> rcorr(as.matrix(score_minidata), type="pearson")
                  read motivation self_concept locus_of_control
read              1.00       0.21         0.06              0.37
motivation        0.21       1.00         0.29              0.25
self_concept      0.06       0.29         1.00              0.17
locus_of_control  0.37       0.25         0.17              1.00

n= 600
```

Example: Multiple Regression

- It's also helpful to show the scatter plots of these predictors and outcome variable:



Example: Multiple Regression

- Now, we estimate the multiple regression model in R:

```
> # regression
> linear_regression <- lm(read~motivation + self_concept + locus_of_control,data=score_minidata)
> summary(linear_regression)

Call:
lm(formula = read ~ motivation + self_concept + locus_of_control,
    data = score_minidata)

Residuals:
     Min       1Q   Median       3Q      Max
-28.3235  -6.5275  -0.3192   6.1475  29.2204

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)      51.3937     0.3840 133.832  < 2e-16 ***
motivation        1.6787     0.4936   3.401 0.000717 ***
self_concept     -0.5457     0.5662  -0.964 0.335498
locus_of_control  5.2241     0.5885   8.876  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 9.305 on 596 degrees of freedom
Multiple R-squared:  0.1559,    Adjusted R-squared:  0.1517
F-statistic:  36.7 on 3 and 596 DF,  p-value: < 2.2e-16
```

Example: Multiple Regression

- What proportion of the total variation in reading score is explained by the three psychological variables?
- From the results, can you tell me what would be the *correlation* between the **observed** reading score and the **predicted** reading score?
- We can check our calculation by doing the prediction and calculating the correlation in R:

```
> # prediction and correlation between read_hat and read
> score_minidata$read_hat <- fitted(linear_regression)
> cor(score_minidata$read_hat,score_minidata$read)
[1] 0.3948953
>
```

- Why is the adjusted R-squared slightly smaller than the original R-squared? → penalty for added predictors

Example: Multiple Regression

- We can calculate the adjusted R-squared by hand:

$$R^2_{adj} = R^2 - (1 - R^2)\frac{k}{n - k - 1} \tag{9}$$

$$= 0.1559 - (1 - 0.1559) \cdot \frac{3}{600 - 3 - 1} \tag{10}$$

$$= 0.1517 \tag{11}$$

Types of Multivariate Relationships

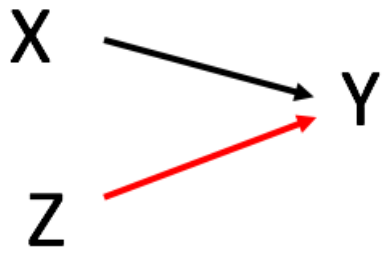
Types of Multivariate Relationships

- Up to now, we have introduced the setup and inferences on the multivariate regression model.
- Next, we turn to the understanding of different types of multivariate relationships.
- Why is it important to differentiate among these different types?

1. Because it helps us decide whether to include some variables in our model.
2. Because it tells us *why* we include certain variables in our models, what kind of difference we expect that to make.
3. Because it helps us interpret and adjudicate among results presented by other researchers.

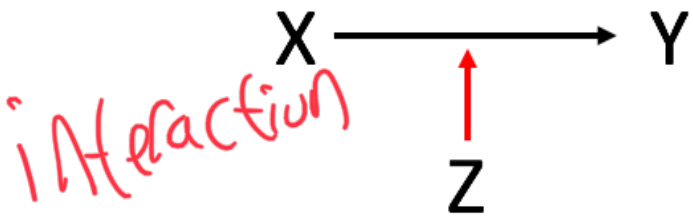
Types of Multivariate Relationships

- We focus on four types of multivariate relationships: (1) multiple causes; (2) mediation; (3) Interaction; (4) confounding.

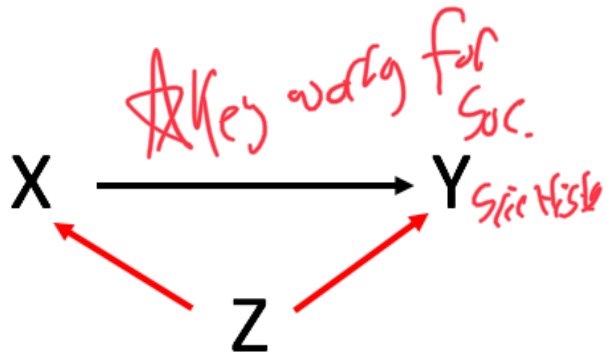


(1) Multiple causes

(2) Mediation



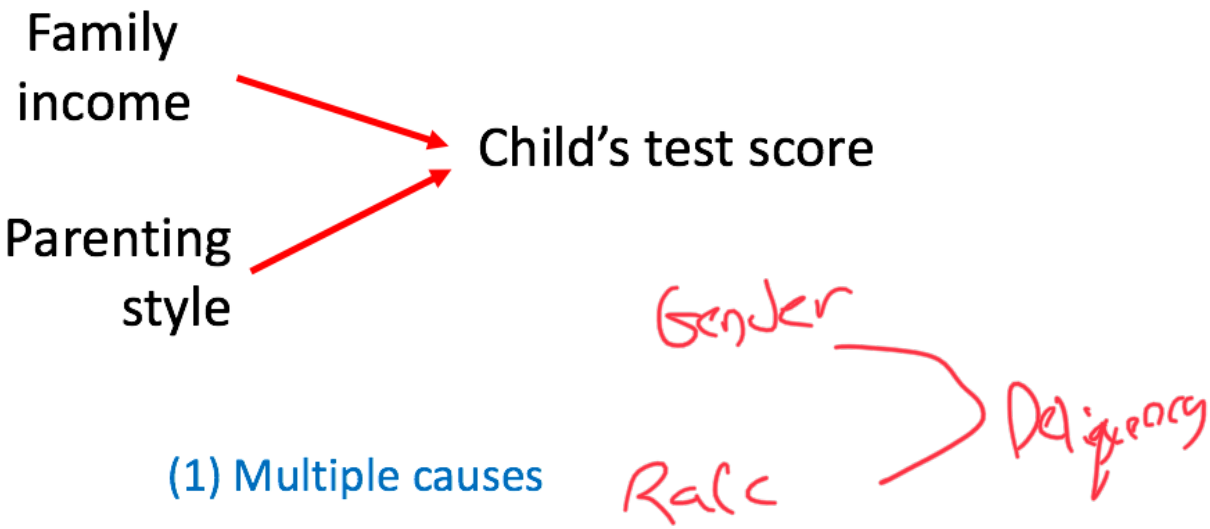
(3) Moderation



(4) Confounding

Multiple Causes

- **Multiple causes**: more than one factors may have contributed to the variation in the outcome variable.
- These explanatory variables may be independent or correlated.



Multiple Causes

- So, what would happen to the estimated β_1 and se_{β_1} if two causes are highly correlated?
- Recall that β_1 is expressed as:

$$\hat{\beta}_1 = \frac{Cov(x_1, y) \cdot Var(x_2) - Cov(x_2, y) \cdot Cov(x_1, x_2)}{Var(x_1) \cdot Var(x_2) - Cov(x_1, x_2)^2} \tag{12}$$

- If the two causes are uncorrelated, then $Cov(x_1, x_2) = 0$, and we have:

$$\hat{\beta}_1 = \frac{Cov(x_1, y) \cdot Var(x_2)}{Var(x_1) \cdot Var(x_2)} \tag{13}$$

$$= \frac{Cov(x_1, y)}{Var(x_1)} \tag{14}$$

The Problem of Multicollinearity

- If two or more explanatory variables in a multivariate regression model are highly linearly related, then we have a problem of multicollinearity.
- Multicollinearity will not affect much the overall predictive power of the model, but it will cause large margin of error to the effect of individual predictors.
- Recall from the previous lecture that the formula for the standard error of $\hat{\beta}_1$ is:

$$se_{\beta_1} = \sqrt{\frac{s_\epsilon^2}{(n-k-1)Var(x_1)(1-r_{x_1,x_2}^2)}} \tag{15}$$

- As the formula suggests, se is smaller when the correlation between the two independent variables are small.

Multiple Causes: An Example

- To see why including multiple predictors in a regression model would be useful, let's look at an example table from Weeden (2002), which examines the effect of occupational closure strategies on earnings of practitioners in the occupation.
- The next page will show the bivariate correlations of the key occupational variables in this article.

Multiple Causes: An Example

TABLE 3
MEANS, SDs, AND BIVARIATE CORRELATIONS OF OCCUPATION-LEVEL SKILL, GENDER COMPOSITION, AND CLOSURE VARIABLES

VARIABLE	MEAN	SD	BIVARIATE CORRELATIONS										
			1	2	3	4	5	6	7	8	9	10	11
Skills and working conditions:[a]													
1. Cognitive skills	.014	.848	. . .										
2. Authority	.160	.366	.36										
3. Nurturing	.312	.880	.35	.14									
4. Physical demands	−.163	.803	−.03	−.32	−.30								
5. Hazardous conditions	−.175	.737	−.43	−.14	−.46	.25							
Gender composition:													
6. %female[b]	.466	.302	−.03	−.16	.46	−.08	−.49						
Closure strategies (proportion of occupation members covered):[c]													
7. Licensing	.330	.404	.39	.24	.43	−.19	−.17	.06					
8. Educational credentialing	.574	.257	.83	.18	.57	−.20	−.56	.23	.48				
9. Voluntary certification	.078	.220	.23	−.09	.12	.10	−.10	.07	.26	.29			
10. Association representation	.153	.344	.48	*.01*	.19	.10	−.18	−.01	.26	.49	.35		
11. Unionization	.161	.154	−.22	−.25	.10	.08	.33	−.18	.01	−.11	−.12	−.06	. . .

NOTE.—N = 97,084 individuals in 488 occupations. Values are weighted by occupation size. All correlations except the value in italics are significant at $P < .01$.

[a] Scale items are as follows: (*a*) cognitive skills—general educational development, data complexity, numerical aptitude, intelligence, training time; (*b*) nurturing—face-to-face service (England 1992), talking/hearing, dealing with people; (*c*) physical demands—motor coordination, finger dexterity, form perception, spatial perception, visual requirements; (*d*) hazardous conditions—exposure to hazards, exposure to fumes, need to stoop, need to climb, environmental disamentities. Authority is a binary variable coded "1" if the occupation requires supervisory or managerial skills, "0" otherwise (see England 1992). Except as noted, items are from the *DOT* (see England and Kilbourne 1988).

[b] Calculated from the 1990 EEO census file (U.S. Department of Commerce 1993).

[c] See the appendix for variable definitions and data sources.

Multiple Causes: An Example

- As we can see in the above correlation matrix, some of the variables for closure stratigies and the occupational skills and working conditions variables are correlated to each other.
- Hence, if we only run bi-variate regressions of earnings on each of these occupational characteristics, we will not be able to tell whether the effects are effects of the variables, or the effects through their association with other "omitted" variables.
- If we run a multivariate regression model with all these variables included, then this will help us determine the "partial effect" of each variable, "holding constant" the values on other variables.
- The regression results are presented on the next page.

Multiple Causes: An Example

TABLE 4

ESTIMATED OCCUPATION-LEVEL EFFECTS FROM THE REGRESSION OF LOGGED WEEKLY EARNINGS ON INDIVIDUAL ATTRIBUTES AND OCCUPATIONAL SKILLS AND CLOSURE CHARACTERISTICS

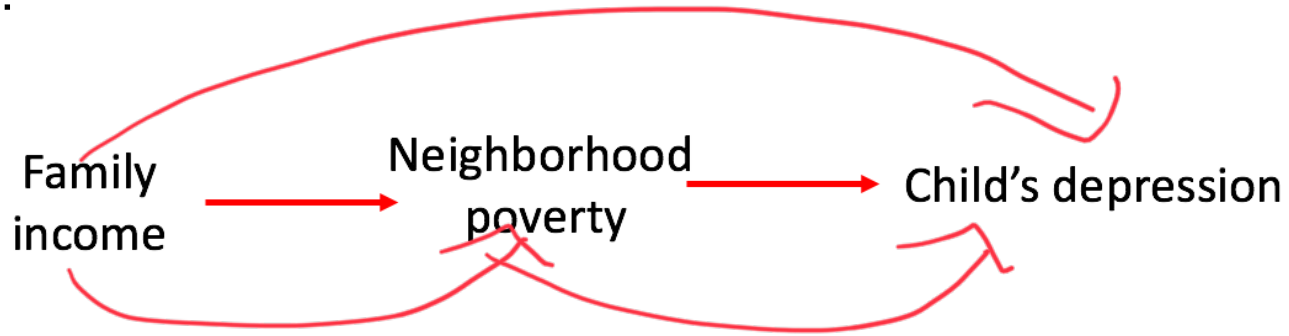
Independent Variable	Model 1	Model 2	Model 3
Fixed effects:[a]			
Constant	5.984***	5.711***	5.805***
	(.014)	(.023)	(.030)
Skills and working conditions:			
Cognitive skills	.142***	. . .	.072***
	(.014)		(.015)
Authority	.118***	. . .	.127***
	(.026)		(.024)
Nurturing	−.005	. . .	−.047***
	(.011)		(.011)
Physical demands	.058***	. . .	.047***
	(.010)		(.010)
Hazardous conditions	.010	. . .	.008
	(.012)		(.012)
Proportion female	−.074*	. . .	−.132***
	(.032)		(.030)
Closure strategies:			
Licensing	. . .	.077**	.085***
		(.028)	(.023)
Educational credentialing	. . .	.363***	.317***
		(.037)	(.054)
Voluntary certification	. . .	.102**	.116***
		(.035)	(.028)
Association representation	. . .	.012	−.033*
		(.016)	(.014)
Unionization	. . .	.131**	.047
		(.050)	(.044)
Random effects:			
Occupation-level variance	.019	.026	.014
	(.002)	(.002)	(.001)
Individual-level variance	.220	.220	.220
	(.001)	(.001)	(.001)
−2 log-likelihood	129,605	129,725	129,517
χ^2 test statistic[b]	301***	179***	88***
df	6	5	5

NOTE.—N = 97,084 individuals in 488 occupations. See table 3 for data sources.

[a] Both models fit all individual-level covariates (see table 2) and their interactions with sex as grand-mean centered fixed effects.

Mediation

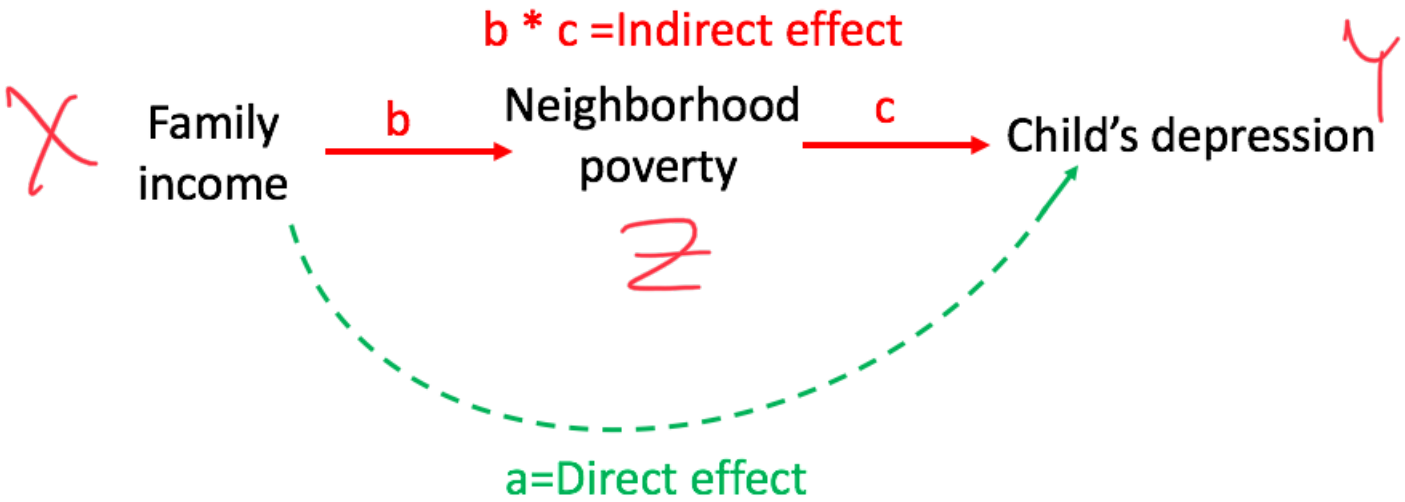
- **Mediation**: a third variable helps explain the relationship between X and Y.
- For example, once we know that X affects Y, we may want to know why X affects Y. So we want to introduce mediating variables to "explain" the effect of X on Y.
- Example:



(2) Mediation

Mediation

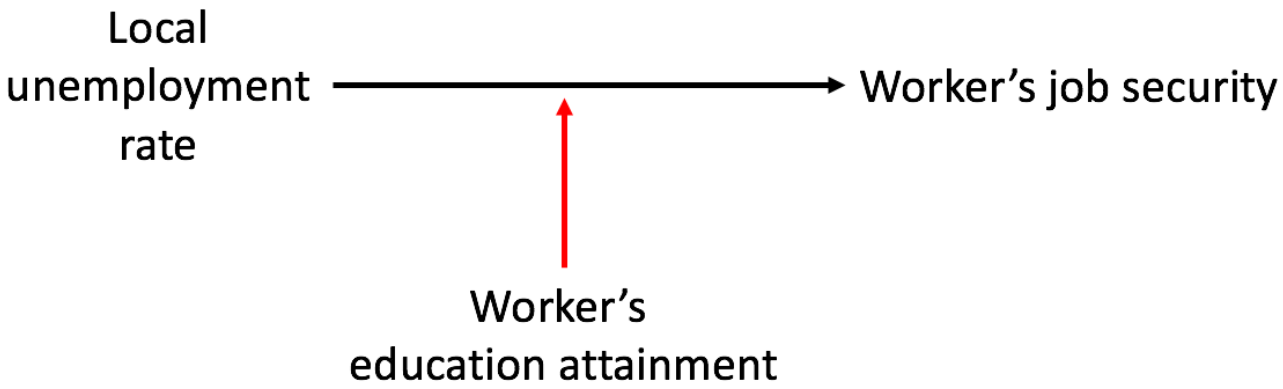
- In many cases, not all of the effect of X on Y can be mediated by Z. So we want to separate out the "direct effect" of X and the "indirect effect" of X on Y:




CHEETAH CONSERVATION FUND

Moderation

- **Moderation**: the effect of X on Y is different for different values of Z or distinct groups of the population.
- Z is called a "moderator" or "effect moderator."
- Example:



(3) Moderation

Moderation

- As illustrated earlier, a moderator can be captured by an **interaction** term between X and Z.
- When you suspect that there is an interaction effect in your analysis, you can run a regression among different groups of the population (e.g. by gender) separately (sub-group analysis), but to test whether the effect of a predictor varies significantly between the two groups, it is always useful to run a model with the *pooled sample*, and test whether the coefficient on the *interaction* term is significantly different from zero.

Moderation

- For example, if you are interested in whether the effect of English proficiency on hourly wage differs between male and female immigrants, you could break down your sample into two sub-samples (male and female), and run the following regression model for each sample:

$$Wage = \beta_0^{male} + \beta_1^{male} \cdot English + \epsilon \tag{16}$$

$$Wage = \beta_0^{female} + \beta_1^{female} \cdot English + \epsilon \tag{17}$$

Moderation

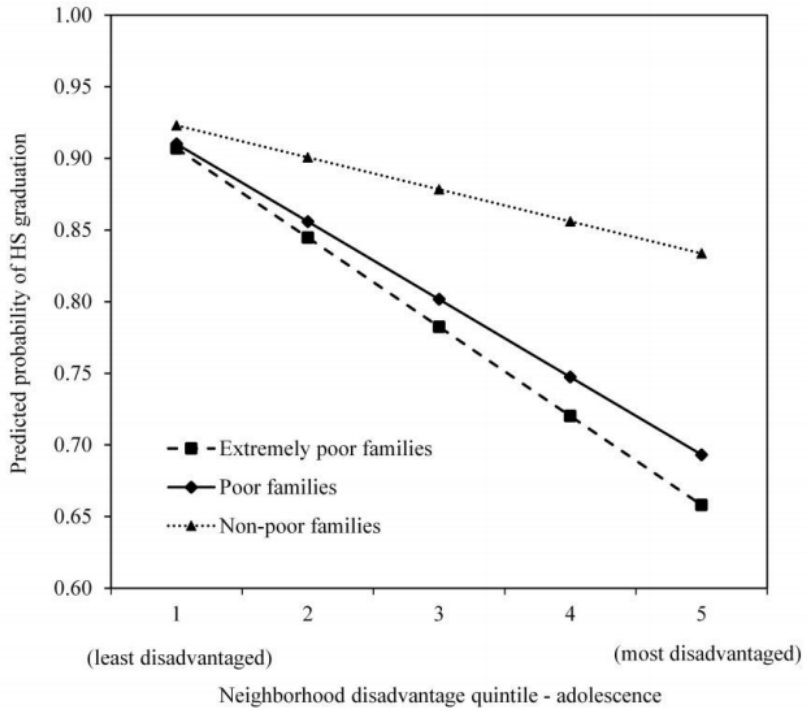
- You will get β_1^{male} and β_1^{female}, and they may be different.
- But in order to test whether their difference is statistically significant, you can pool the male and female sample togehter, and run the following regression model with an interaction term:

$$Wage = \beta_0 + \beta_1 \cdot English + \beta_2 \cdot Gender + \beta_3 \cdot Gender \times English + \epsilon \tag{18}$$

- If the null hypothesis $\beta_3 = 0$ is rejected, then the gender difference in the effect of English proficiency is statistically significant.

Moderation

- Let's look at an example from Wodtke, Elwert, and Harding (2016), in which the authors explored how the effects of disadvantaged neighborhoods on child educational outcomes depend on a family's economic resources and the timing of neighborhood exposures during the course of child development.
- Here is one of the figure they used to illustrate their results:



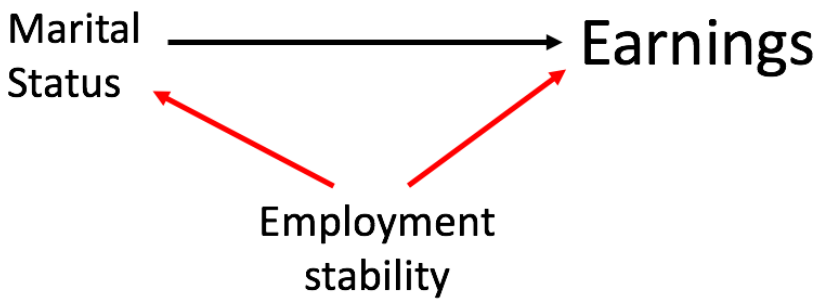
Confounding

- **Confounding**: X and Y have a mutual cause (or many mutual causes).
- The confounder may be observed (measured) or unobserved.
- This is, by far, the biggest problem with regression analysis and its variants. It will cause spuriousness in the "claimed" effect of X on Y.
- A classic example would be the "effect" of ice-cream sale and number of swimmers in the sea - they are caused by a mutual factor: temperature.

- Controlling for Z removes relationship b/tw X & Y

Confounding

- Example:



(4) Confounding

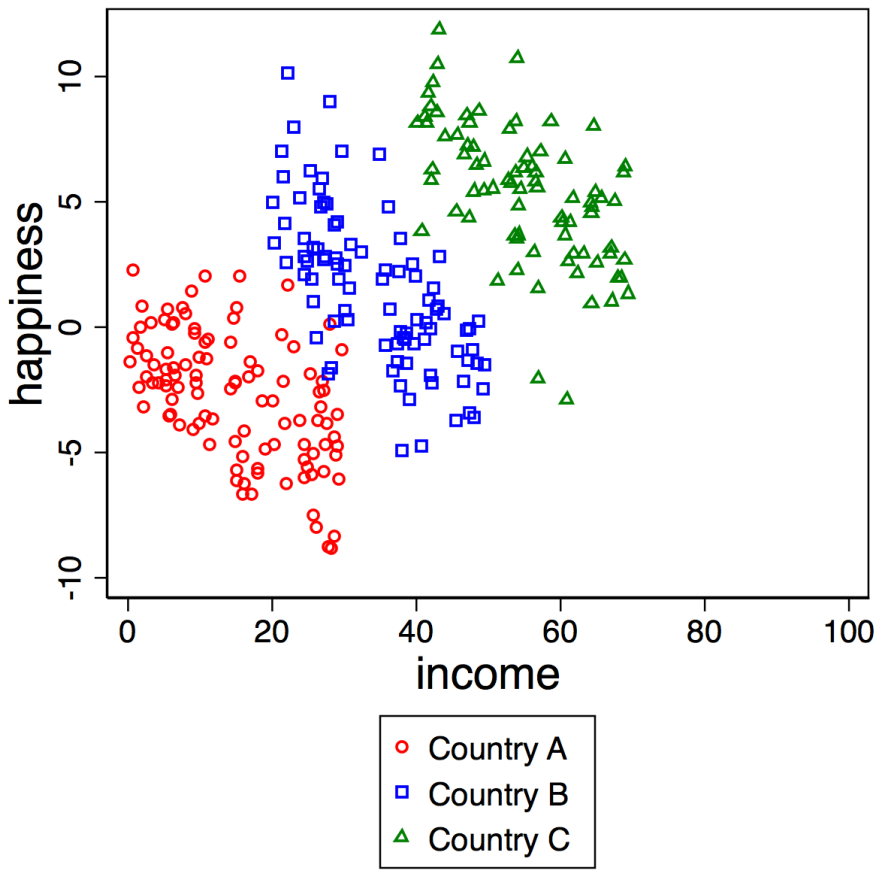
- We know from the data that generally (in the US), marriage is associated with higher earnings for men.
- However, we also know that men with greater employment stability, greater earnings prospect, and are simply smarter are more likely to get married.

Confounding

- If we can measure all the potential confounders, such as "employment stability," "future earnings," "reliability," "IQ," the effect of marriage on men's earnings may drop dramatically or even disappear. Still an ongoing debate...
- Fixed-effect model may be a solution, but not all confounders are fixed over time...We will come back to this when we discuss panel data and longitudinal models.

Confounding

- One (extreme) example of confounding is the Simpson's Paradox:



Confounding

- The existence of potentially unobservable confounders is one reason that some researchers turn to **experimental methods**. For example, Pager (2003) used an audit study design to address the "omitted variable bias" in observational studies, as shown in the following figure:

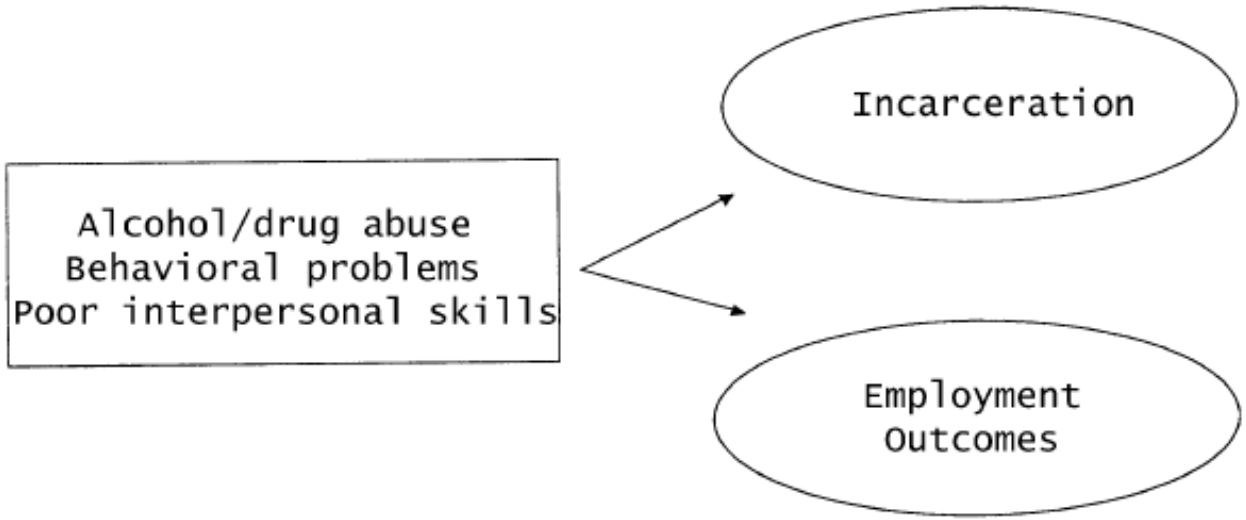


FIG. 2.—Model of spurious effects

Standardized Regression Coefficients

Standardized Regression Coefficients

- As we discussed in the previous lecture, the size of a regression coefficient depends on the scale at which the independent and dependent variables are measured. For example, if we use $X^* = 0.1 \cdot X$ as the new independent variable, the coefficient on X will be ten times the original coefficient.
- So, is there a way to generate coefficients that are NOT sensitive to the scale at which the variables are measured?

Standardized Regression Coefficients

- In fact, there are two ways to create "**standardized regression coefficients**," that is, coefficients that tell us the *change in y, in y standard deviation units, for a one-standard-deviation increase in x.*
- The first method is to adjust the original regression coefficient by the sample standard deviations of x and y:

Standardized

$$\beta_1^* = \beta_1 \cdot \frac{s_x}{s_y} \tag{19}$$

Standardized Regression Coefficients

- The second method is to create the "**standardized variables**" for y and x by calculating their z-scores:

$$z_y = \frac{y - \bar{y}}{s_y} \tag{20}$$

$$z_x = \frac{x - \bar{x}}{s_x} \tag{21}$$

And then we run the regression of z_y on z_x without a constant term:

$$\hat{z_y} = \hat{\beta}_1^* \cdot z_x \tag{22}$$

- Why we don't need a constant term here? Because we know that when z_x=0, $x = \bar{x}$, and we know that when $x = \bar{x}$, $y = \bar{y}$, and thus $z_y = \frac{y-\bar{y}}{s_y} = 0$.

Standardized Regression Coefficients

- But we should be careful using standard coefficients, because they are sensitive to the **standard deviations** of the variables in the sample.
- If our sample standard deviation is not representative of the population-level dispersion (e.g. we collected a very homogeneous sample), then the sample standardized coefficients cannot be generalized to the population standard coefficients.

Regression with Dummy Variables

- So far, we have been talking about continuous variables in our models. What do we do if we have categorical variables?
- If our outcome variable is categorical, then we will use a more complicated type of regression models. We will get back to this in later lectures.
- Today, let's focus on the case of categorical independent variables.

Regression with Dummy Variables

- When our independent variable is categorical, we create *a set of artificial variables* (called "dummy variables"). Each of these artificial variables will equal 1 if an observation belongs to a particular group and 0 otherwise.
- For example, the GSS 2014 data contains a categorical of marital status:

```
. tab marital
```

marital status	Freq.	Percent	Cum.
married	1,158	45.70	45.70
widowed	209	8.25	53.95
divorced	411	16.22	70.17
separated	81	3.20	73.36
never married	675	26.64	100.00
Total	2,534	100.00	

Regression with Dummy Variables

- We can generate five dummy variables, indicating whether this individual come from a particular group:

Marital status	D1	D2	D3	D4	D5
Married	1	0	0	0	0
Widowed	0	1	0	0	0
Divorced	0	0	1	0	0
Separated	0	0	0	1	0
Never married	0	0	0	0	1

- D1, D2, D3, D4, and D5 are called dummy variables that indicate the person's marital status.

Regression with Dummy Variables

- It is important to note that when we run the regression of an outcome variable, say, general happiness (y=happiness) on marital status, we should include only 4 of the dummy variables. For example, if we exclude D1, then the regression model is as follows:

$$y = \beta_0 + \beta_2 D2 + \beta_3 D3 + \beta_4 D4 + \beta_5 D5 + \epsilon. \qquad (23)$$

- We exclude one dummy variable here, because these categories are mutually exclusive and thus when we know a person's status on any of these four dummy variables, we know directly what the person's status on the fifth dummy variable is.
- In general, for a categorical with m groups, we include $m - 1$ dummy variables (i.e. exclude 1 group as a "**reference group**") in the regression model.

Regression with Dummy Variables

- Interpretation: If D1 is treated as the reference group and hence omitted from the model, then β_2 captures the difference between the group corresponding to D2 and the group corresponding to D1, holding other factors constant.

$$y = \beta_0 + \beta_2 D2 + \beta_3 D3 + \beta_4 D4 + \beta_5 D5 + \epsilon. \qquad (24)$$

- From Equation (24), we can obtain the predicted group means:

Marital status	Predicted Group Mean
Married	$\hat{\mu}_1 = \hat{\beta}_0$
Widowed	$\hat{\mu}_2 = \hat{\beta}_0 + \hat{\beta}_2$
Divorced	$\hat{\mu}_3 = \hat{\beta}_0 + \hat{\beta}_3$
Separated	$\hat{\mu}_4 = \hat{\beta}_0 + \hat{\beta}_4$
Never married	$\hat{\mu}_5 = \hat{\beta}_0 + \hat{\beta}_5$

Interaction between Dummy Variables

- Now, let's consider the *interaction* between two dummy variables.
- Suppose we are interested in two dummy variables, **race** (white, black) and **education** (no college, college).
- Separately, we can create 2 dummy variables for race, and 2 dummy variables for education.
- Remember that we should omit the reference categories in our regression models. To do so, include the "main effect" of the two variables separately, omitting "white" and "no college" as the reference groups. In other words, we will include D_{black} and D_{college}.
- Then, because we are interested in the interaction effect of the two variables, we should include an additional variable, called $D_{\text{black} \times \text{college}}$.

Interaction between Dummy Variables

- The regression equation becomes:

$$y = \beta_0 + \beta_1 D_{black} + \beta_2 D_{college} + \beta_3 D_{black \times college} + \epsilon \qquad (25)$$

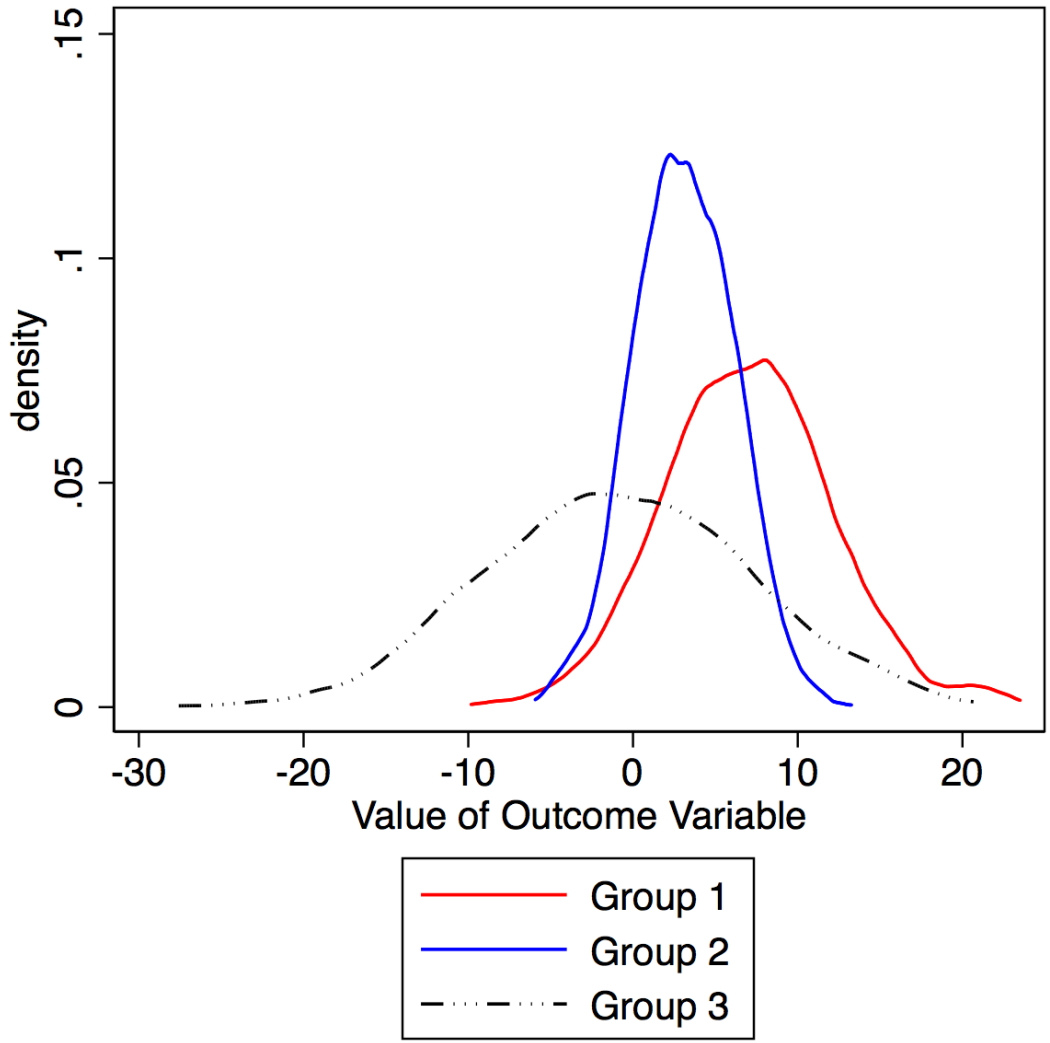
- β_1 and β_2 are main effects, and β_3 is the interaction effect.
- Questions to think about:
  1. What is the predicted group mean for black people without a college degree?
  2. What is the predicted group mean for white people with a college degree?
  3. What is the predicted group mean for black college graduates?
  4. If $\beta_3 > 0$, what does this mean?
  5. How to interpret β_0?

Analysis of Variance

Analysis of Variance

- Earlier, we have learned how to compare the means of two groups (two-sample t-test). Now, let's turn to the problem of comparing the means of three or more groups.
- The Analysis of Variance (ANOVA) provides a F test to determine **whether the means of an outcome variable are different across groups**.
- Formally, let G denote the number of groups to compare, and the $\mu's$ be group means.
- The null hypothesis is: $H_0 : \mu_1 = \mu_2 = ... = \mu_g$.
- The alternative hypothesis is: H_a: at least two of the population means are unequal.
- We can also think of ANOVA as a test of independence between the continuous outcome variable and the categorical explanatory variable (# of groups=G).

Analysis of Variance



Analysis of Variance

- The F-test statistic is a ratio of two variance estimates: between-group variance and within-group variance.

$$F = \frac{\text{between-group variance}}{\text{within-group variance}} \tag{26}$$

- Between-group variance is calculated as:

$$\text{Between-group variance} = \frac{\sum_{g=1}^{G} n_g (\bar{Y}_g - \bar{Y})^2}{G - 1} \tag{27}$$

$$\text{Within-group variance} = \frac{\sum_{i=1}^{N} (Y_i - \bar{Y}_g)^2}{N - G} \tag{28}$$

- Under H_0, F-statistic will follow F-distribution with df=(G-1, N-G).

Analysis of Variance

- In fact, ANOVA is closely related to regression analysis — We can think of ANOVA as regressing the outcome variable on dummies of the categorical variable.
- "Between-group variance" can be seen as the "model sum of squares," and "within-group variance" can be seen as the "residual sum of squares."
- The F-test in a regression model will look the same as the F-test in ANOVA, if we do not include other variables in the regression model.