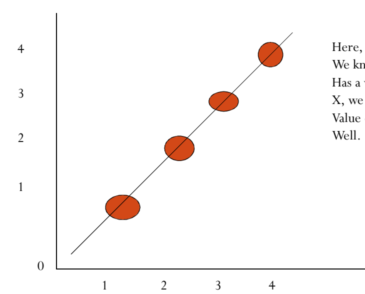
Outline for Today…

- **Linearity**
- **Dummy Variables**
- **Multiple Regression**
- **Interaction Terms**
- **Lab**

Where we are…

#	Date	Read (5th)	Due	Area	Lecture Topic	Lab #	Lab Assignment	T Lab	R Lab
1	Tue Aug 26			*Orientation*	Welcome & Orientation	-			
2	Thu Aug 28	1.1 to 1.4			Basic Terms	-		-	-
3	Tue Sep 2	2.1 & 2.5			Measurement Issues	1	Levels / Age 3x	1	-
4	Thu Sep 4	3.1			Data Reduction	2	Heaven & Hell		1, 2
5	Tue Sep 9	-	HW1	*Description*	Display & Analysis (Shapes & SPSS)	3	Total Miles	2, 3	
6	Thu Sep 11	3.2 & 3.5			Central Tendency	4	CT		3, 4
7	Tue Sep 16	3.3 & 3.4			Dispersion	5	Dispersion	4, 5	
8	Thu Sep 18	3.7			Indices & Data Cleaning	6	Music index		5, 6
9	Tue Sep 23	4.2		*Inference*	Probability & Z Scores	7	Standardizing Scores	6, 7	
10	Thu Sep 25	4.3	HW2		Zs & Ps	8	Table A		7, 8
11	Tue Sep 30	3.6 & 5.1			Parameters & Pt Estimation	9	Distances	8, 9	
12	Thu Oct 2	2.2 to 2.4 & 4.3			Sampling (Issues, Methods, Effects)	10	Sampling		9, 10
13	Tue Oct 7	4.4 to 4.6			The Central Limit Theorem	11-EC	CLT/World (EC)	10, 11*	
14	Thu Oct 9	5.3		*Estimation*	Confidence Intervals	12	CI for Intervals		11*, 12
15	Tue Oct 14		HW3		CIs for Proportions	13	CI for Proportions	12, 13	
16	Thu Oct 16	6.1 & 6.4			Hypotesting & Zs	14	Writing Hypotheses		13, 14
17	Tue Oct 21				Hypotesting for Large ns	15	Two Tests	14, 15	
18	Thu Oct 23	6.3 & 6.8	HW4		The "t" test, for small ns	16	CI & Test Ages		15, 16
19	Tue Oct 28	5.4			Sample Size Estimation	17	Estimating n Needed	16, 17	
20	Thu Oct 30	7.1, 7.3, & 10.1		*Covariation*	Differences in Means	18	Comparing Means		17, 18
21	Tue Nov 4	7.2	HW5		Differences in Proportions	19	Comparing Proportions	18, 19	
22	Thu Nov 6	12.1			Analysis of Variance	20, 21-EC	ANOVA (+ MODELS EC)		19, 20, 21*
	Tue Nov 11				(Veteran's Day)				
23	Thu Nov 13	9.4 & 9.5			Scatterplots & Correlation	22	Grade Correlations		22
[illegible]	[illegible]	[illegible]	HW6		Regression	[illegible]	Regression Lab	[illegible]	
5	**Thu Nov 20**	10.2 & 11.1			**Multiple Regression**	24-EC	Multiple Reg (EC)		23, 24*
26	**Tue Nov 25**	8.1	HW7	*Association*	**Crosstabulations**	25	TBA (any)	24*, 25	
	Thu Nov 27				(Thanksgiving)				
27	**Tue Dec 2**	8.2 & p.233			**Dependence**	26, 27-EC	TBA (SCU) (& 27-EC)	26, 27*	
28	**Thu Dec 4**	pp.238 to 243	HW8		**Association**	28-EC	Measures of Assoc (EC)		25,26,27*,28
x	Tue Dec 9				(no lecture - work session only)			28*	
x	Thu Dec 11				(no lecture - work session only)				
[illegible]	Thu Dec 18		HW9		[illegible]				



Here, for example, if
We know a new case
Has a value of 3.5 on
X, we can predict the
Value of 3.5 on Y as
Well.

Multiple Regression

SOC424 – Statistics w/ Dr. Ellis Godard

Assumption of Linearity

- **Describing the <u>relationship between two variables</u>**
- **Working with a <u>general formula</u>**
  - Y = f(X)
  - Not same as X "causes" Y
  - Not that the value of X *determines* the value of Y.
  - But that there's a general trend or pattern
    - The <u>values on one *associated* with the values of the other</u>
    - The values of the two variables have a statistical *association*
- But does not assert (or even assume) consistent linearity
  - Most points are NOT "on the line," but
  - The vertical averages are the best estimate (*if* the model is significant)

Benefits of that "Assumption"

- **Describes the relationship between 2 variables**
- **Inferential values: Allows us to predict…**
  - …the *average* value of the DV, for each value of the IV
  - … the avg *increase* in DV, *for each increase of 1* in the IV

Coefficients[a]

Model		Unstandardized Coefficients B	Unstandardized Coefficients Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	223.190	1.393		160.193	.000
	Index of online media use	-5.102	.366	-.201	-13.932	.000

a. Dependent Variable: weight

  - …how much of the variance in the DV is explained by variance in the IV

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.201[a]	.041	.040	79.830

a. Predictors: (Constant), Index of online media use

- **Can write equation:**
  - **Weight = 223.190 – (5.102*OLUSE)**
  - **Note p values**

Note that equations include variable names!

Exceptions to Strict Linearity

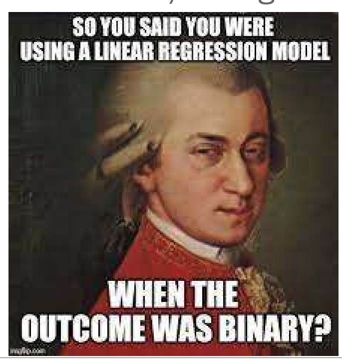
- **Dummy Variables**
  - Binary (dichotomous) IV, instead of interval
  - 1 is the "on" state; 0 is "off" (i.e. *not* that characteristic)
- **Multiple Regression**
  - More than one interval variable, but still just ONE equation, for all of them @ once
- **Interaction Terms**
  - Some IVs have a joint effect, beyond individual effects
  - Easy and powerful (great combination!)

Limits of that Assumption

- **Not all relationships are linear**
  - Some are curvilinear, U-linear, S curves*, etc.
  - Check scatterplot; does it *look* linear?
- **Don't know beyond range of data**
  - Avoid assuming trend continues!
- **Most points aren't on the line**
  - May be that *none* are
  - Model still "works", if there's a (sig.) trend
  - There's an implicit error term in formula
    - Every prediction is *known* to be wrong by some estimate average amount

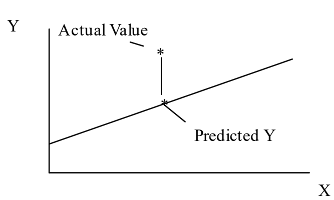
*** google it ;)**

Don't use SLR w/ Categorical DVs



Residuals…

The vertical distance between the line (which represents predicted values) and each actual observed data point is the "residual" (a regression "error", though different from the implied "error term")



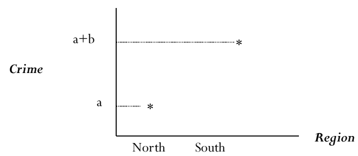
Regression w/ a Dummy Variable

- Consider an equation to predict crime by region.
- Region isn't interval, so it doesn't make sense to have a slope if there are, say, five categories.

  Nominal values could be reordered without changing the variable's meaning, but the slope assumes settled order.
- We could instead have a variable for each region.

Regional Differences in Crime

- For example, SOUTH has two values, 0 and 1.
  1 means the state is in the South; 0 means it isn't
- Equation w/ only that IV would be: Crime=a + b(South)
- Crime rate in the South is: Crime = a+b(1) = a+b
- Crime rate in the North is: Crime=a+b(0)=a.



Multiple Regression

- More than one independent variable (X)
- A different slope (b) for each one, e.g.
  - INCOME = 2.4 + 535*EDUC - 158*KIDS
    - Each IV has its own coefficient (e.g. 535, 158)
  - Each slope is the *average predicted* change in Y for each increase of 1 in *that* X
- ONE equation for ALL of the variables
  - NOT a separate equation for each variable
  - No limit to the number or type of variables
  - Can plug in values of IV, to predict value of DV

Multiple Regression w/ 1 Dummy: Region, Urbanization and Crime

- We hypothesize that crime is a function of urbanity and southerness – that is, that it will be higher in cities and in the south.
- Assuming the relationship is linear, the regression equation would be:
  Crime=a + b_1(Urban) + b_2(South).
  For the non-south cases, one part of the equation "falls out", leaving 2 equations:

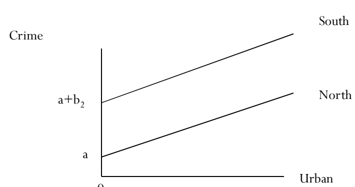
$$Crime_{South} = a + b1(Urban) + b2(1) \qquad = a + b1(Urban) + b2$$

$$Crime_{North} = a + b1(Urban) + b2(0) \qquad = a + b1(Urban)$$

Multiple X's shouldn't be highly correlated

Multiple Regression Plots

- Since we have 2 equations, we have 2 lines:



- Same slope (b_1), 2 intercepts (a & a+b_2)

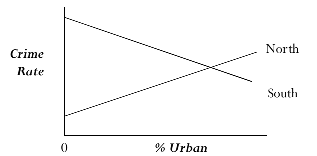
Regression w/ Interaction Term: Region, Urbanization and Crime

- Sometimes variables *interact*
  - Maybe urbanity is different in the north & south
  - Maybe it has a different *effect* in the south
- Crime is a *joint* function, of independent variables *and* interactions among them
  - Crime = *f*(Urban, Region, Urban*Region)
- Create "interaction" terms w/ compute, e.g.
  - COMPUTE=SOUTH*URBAN
- Then, include them in the (regression) model
  - Crime=a + b_1(Urban) + b_2(South) +b_3(Urban*South)

Note that equations include variable names!

Plot with Interaction Term Included: Region, Urbanization and Crime

Now, 2 lines with 2 *slopes*



Assumes different slopes across regions

The next slide shows a meaty example with *three* lines….

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.760[a]	.578	.556	19.82766

a. Predictors: (Constant), absent

We can explain 57.8% of the variance in midterm grades with the variance in days absent

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10233.114	1	10233.114	26.029	.000[b]
	Residual	7469.589	19	393.136		
	Total	17702.703	20			

a. Dependent Variable: mtgrade

b. Predictors: (Constant), absent

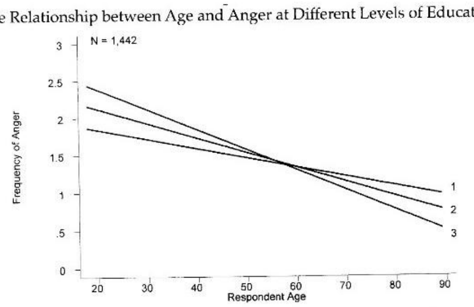
mtgrade=80.308 – 14.630*absent

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	80.308	5.045		15.918	.000
	absent	-14.630	2.868	-.760	-5.102	.000

a. Dependent Variable: mtgrade

The Relationship between Age and Anger at Different Levels of Education



Note: Lines correspond to equation 2 of Table 1 and reflect predicted anger frequency scores for white men. Intercepts will vary slightly with other combinations of sex and race. Lines 1, 2, and 3 reflect predictions for individuals with 17, 13, and 10 years of education, respectively. These education values are the 90th, 50th, and 10th percentiles.

Scott Schieman, "Socioeconomic Status and the Frequency of Anger Across the Life Course". In *Sociological Perspectives* (journal of the PSA), Vol 46, No. 2 (Summer 2003), pp. 207-222.

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.730[a]	.533	.509	20.85342

a. Predictors: (Constant), labsdid

We can explain 53.3% of the variance in midterm grades with the variance in labs done

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	9440.266	1	9440.266	21.708	.000[b]
	Residual	8262.437	19	434.865		
	Total	17702.703	20			

a. Dependent Variable: mtgrade

b. Predictors: (Constant), labsdid

mtgrade=9.304 + 6.158*labsdid

Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	9.304	13.207		.704	.490
	labsdid	6.158	1.322	.730	4.659	.000

a. Dependent Variable: mtgrade

Multiple Regression Demo

- Using regmulti-demo.sav
  - Labs done, Days absent, and Midterm grade for Summer 2020
- Two *separate* simple linear regressions:
  - Labsdid predicting MTGrade & Absent predicting MTGrade
  - Compare r-squared values
  - Compare coefficient directions (one is negative!)
  - Can't compare coefficient sizes – different scale, intercept, etc.
  - Compare p values for coefficients
- One multiple regression
  - Look at change in r-squared
  - *Can* compare coefficients, because each takes other in consideration
  - *Do* compare p values for coefficients

Not in lab – might demo, time permitting

- Compute interaction term
  - LABATT = labsdid * absent
- New multiple regression
  - Test interaction on its own
  - New regression w/ everything (but not if p>.05)
  - Write new equation

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.776[a]	.602	.558	19.78261

a. Predictors: (Constant), labsdid, absent

We can explain 60.2% of the variance in midterm grades with the variance in BOTH days absent AND labs done

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	10658.375	2	5329.188	13.617	.000[b]
	Residual	7044.327	18	391.352		
	Total	17702.703	20			

a. Dependent Variable: mtgrade

b. Predictors: (Constant), labsdid, absent

mtgrade=52.254 - 9.721*absent + 2.517*labsdid

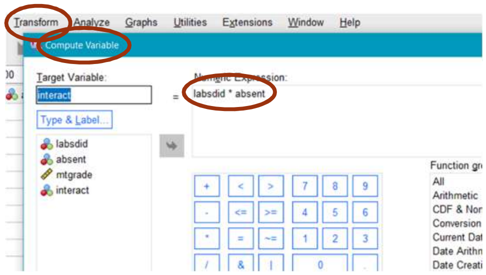
Coefficients[a]

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	52.254	27.379		1.909	.072
	absent	-9.721	5.510	-.505	-1.764	.095
	labsdid	2.517	2.415	.298	1.042	.311

a. Dependent Variable: mtgrade

But now none of the slopes are significant

Interaction term (demo only; NOT lab)



Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.822[a]	.676	.619	18.37289

a. Predictors: (Constant), interact, labsdid, absent

We can explain 67.6% of the variance in midterm grades when we also add that interaction term

ANOVA[a]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	11964.132	3	3988.044	11.814	.000[b]
	Residual	5738.571	17	337.563		
	Total	17702.703	20			

a. Dependent Variable: mtgrade

b. Predictors: (Constant), interact, labsdid, absent

mtgrade=65.445 - 15.827*absent + 1.007*labsdid + 1.606*interact

Model		Unstandardized Coefficients B	Std. Error	Standardized Coefficients Beta	t	Sig.
1	(Constant)	65.445	26.298		2.489	.023
	absent	-15.827	5.985	-.822	-2.644	.017
	labsdid	1.007	2.370	.119	.425	.676
	interact	1.606	.816	.331	1.967	.066

a. Dependent Variable: mtgrade

But two of these slopes are not significant

Multiple Regression Lab Exercise

- Using regmulti-lab.sav, find the parameters to predict R's level of education with five possible independent variables:
  - Their father's level of education
  - Their mother's level of education
  - The number of siblings the respondent has
  - The age of the respondent
  - The respondent's age when they first married
- Write the equation, using all of the intercepts and slopes.
- With which variables is education inversely proportional (that is, when they go up, predicted education goes down)?
- For which variable(s) is the probability too high that the slope differs from zero only due to sampling error?
- Re-run the regression using only those variables whose coefficients are significant.
- Rewrite the equation using only those (*new!*) coefficients.
- Using this new equation, what level of education do we predict for a respondent whose mother got 16 years, whose father got 20, who married at 23, who has 2 siblings, and who is 29 years old?