36-309/749
Experimental Design for Behavioral and Social Sciences

Sep. 29, 2015
Lecture 5: Multiple Regression

Review of ANOVA & Simple Regression

- Both
  - Quantitative outcome
  - Independent, Gaussian errors with equal variance
  - Group assignment assumed correct (fixed-x)
- One way (between-subjects) ANOVA
  - Categorical IV (k levels) with means μ_1 through μ_k
  - Best prediction: $\widehat{Y}_i = \overline{Y}_j$ for subject *i* in group *j*
- Simple (one IV) regression
  - Quantitative IV
  - Coefficient parameters are β_0 and β_1
  - True mean outcome at each x is $E(Y|x)=\beta_0+\beta_1 x$ (linearity)
  - Best prediction: $\widehat{Y}_i = b_0 + b_1 x$

Example

Team Problem Solving

Multiple Regression

- New Idea #1: extend the means model
  - IVs are x_1, x_2, …
  - Means model: $E(Y|x_1,x_2,\ldots) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots$
  - Prediction: $\hat{Y} = b_0 + b_1 x_1 + b_2 x_2 + \ldots$
  - Consequences:
    - β_0 is the **mean** of the DV when **all** IVs equal **0**
    - β_1 is the **change in the mean** of the DV when x_1 **goes up by one** and **all other** x's are held **constant**.
    - E.g., with 2 x's and x_2 fixed at c, Y vs. x_1 is a line:
      $E(Y|x1,x2=c) = \beta_0 + \beta_1 x_1 + \beta_2 c = (\beta_0 + \beta_2 c) + \beta_1 x_1$
      So Y vs. x_1 forms parallel lines at various fixed x_2 values

Multiple Regression: dummies

- New idea #2: Dummy variables
  - Multiple regression can accommodate categorical IVs but ***only if*** they are coded appropriately
  - **Indicator variable**: A categorical variable (factor) with 2 levels should be named for one level and coded with: 1=named level, 0=other level, e.g., a "Female" variable "F" is coded 0=Male, 1=Female.
  - E.g., x_1=Age, x_2=Female: $E(Y)=\beta_0+\beta_A A+\beta_F F$ is a means model of parallel lines:

    Males: $E(Y) = \beta_0 + \beta_A A$

    Females: $E(Y) = (\beta_0+\beta_F F) + \beta_A A = (\beta_0+\beta_F) + \beta_A A$

Multiple Regression: dummies, cont.

- Coding of a categorical IV with k>2 levels
  - Choose an arbitrary baseline (e.g., “control”)
  - Create ***indicator variables*** for all ***non-baseline*** levels
  - Throw away the original variable
  - Example:

Color (code)	Color (“value”)	Red	Blue
3	Red	1	0
1	Blue	0	1
1	Blue	0	1
2	Green	0	0

- “Green” is the arbitrary “baseline”. “Red” and “Blue” are the IVs used in the regression.

Multiple Regression: ANCOVA

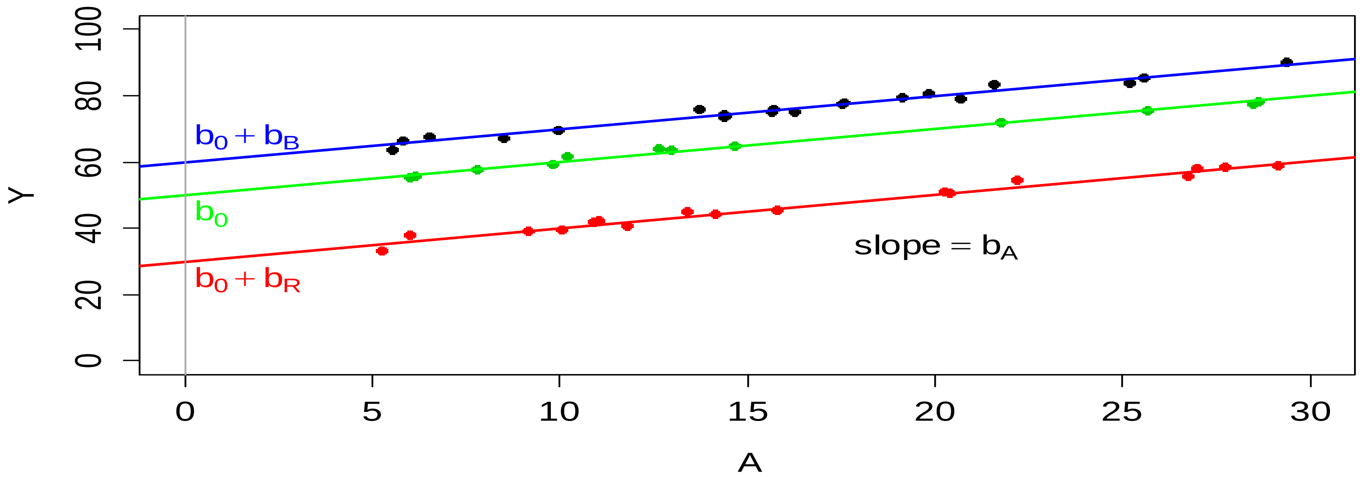
- Generally the term **ANCOVA** (analysis of covariance) refers to multiple regression with one quantitative IV ("covariate") and one categorical IV of primary interest coded as dummy variables.
- Example: covariate "Age" and factor "Color" (baseline=green)

$E(Y|Age, Color) = E(Y|A,B,R) = \beta_0 + \beta_A A + \beta_B B + \beta_R R$

$E(Y|Age, Color{=}Green) = E(Y|A, B{=}0, R{=}0) = \beta_0 + \beta_A A$

$E(Y|Age, Red) = \beta_0 + \beta_A A + \beta_B 0 + \beta_R 1 = (\beta_0+\beta_R) + \beta_A A$

$E(Y|Age, Blue) = \beta_0 + \beta_A A + \beta_B 1 + \beta_R 0 = (\beta_0+\beta_B) + \beta_A A$



Multiple Regression: ANCOVA, cont.

Coefficients						
Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	49.992	.318		157.009	.000
	A	.996	.017	.572	58.144	.000
	B	9.875	.301	.345	32.806	.000
	R	-19.793	.293	-.721	-67.568	.000

In ANCOVA as regression, dummy variables' "***slopes***" reflect different ***intercept offsets*** from the intercept of the baseline category. ***As opposed to individual regressions, inference for comparing lines is provided.***

Fear and Anger Example

- This is loosely based on *Constraints for emotion specificity in fear and anger: The context counts* by Stemmler, et al., **Psychophysiology**, 38, 275–291 (2001). One hundred and sixty-nine adult female subjects were randomized to a control condition or to induction of fear or anger. The outcome of interest is the subjects' combined ratings on three 0-10 point scales of "negativity". The "covariate" is a quantitative measure called heart-period-variability (HPV), which is measured before the emotion induction and is taken as a measure of individual physiological sensitivity to one's surroundings.
  - Experiment or observational study? Experimental units? Interpretability? Generalizability? Power? Construct validity? EDA?
  - Model? Null hypotheses? Alternative hypotheses?

Example, cont.

- Regression output

Model Summary[b]				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.891[a]	.794	.790	3.181
a. Predictors: (Constant), Anger induction, Heart period variability, Fear induction				
b. Dependent Variable: Feelings of negativity				

ANOVA						
Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	6438.971	3	2146.324	212.119	.000
	Residual	1669.550	165	10.118		
	Total	8108.521	168			

Example, cont.

	Unstandardized Coefficients		t	Sig.	95% Confidence Interval for B	
	B	Std. Error			Lower Bound	Upper Bound
(Constant)	2.707	.691	3.920	.000	1.343	4.071
Heart period variability	1.442	.128	11.263	.000	1.189	1.694
Fear induction	13.233	.618	21.397	.000	12.012	14.454
Anger induction	12.118	.602	20.141	.000	10.930	13.306

- **Prediction equations:**

  $\hat{Y}_i$= 2.71 +1.44 HPV_i + 13.23 $Fear_i$ + 12.12 $Anger_i$

  Controls (Fear=0, Anger=0): $\hat{Y}_i$ = 2.71 + 1.44 HPV_i

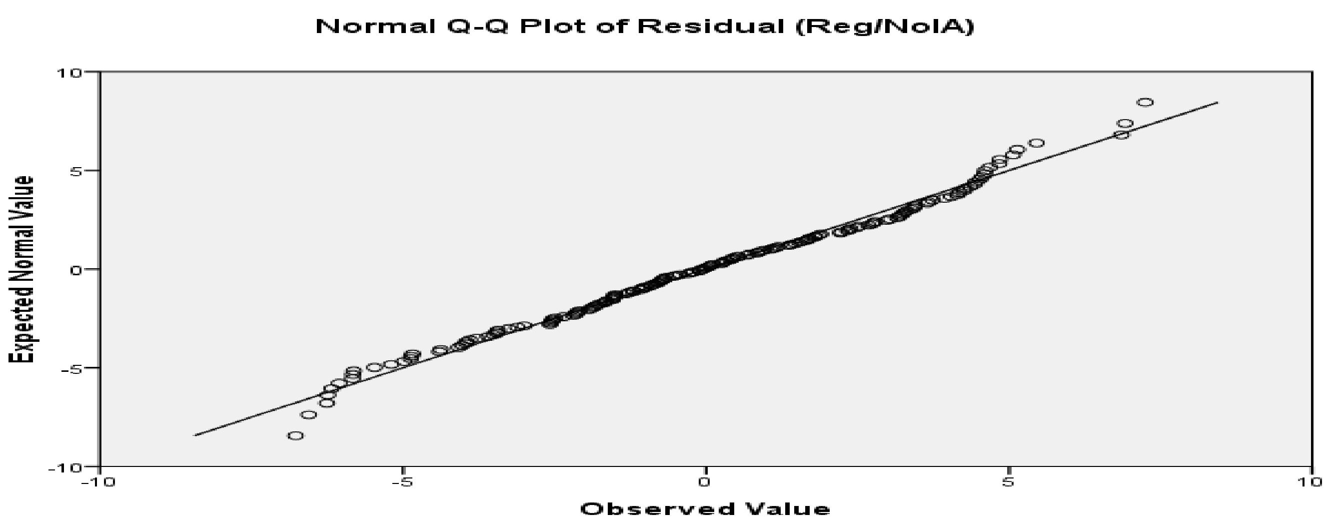
  Fear subjects: $\hat{Y}_i$ = 2.71 + 1.44 HPV_i + 13.23 (1) = 15.94 + 1.44 HPV_i

  Anger subejcts: $\hat{Y}_i$ = 2.71 + 1.44 HPV_i + 12.12(1) = 14.83 + 1.44 HPV_i

- Hidden assumption of the (non-interaction) ANCOVA means model:

Example, cont.

- **Standardized coefficients**: coefficients from running regression on standardized x's and Y: $x_{ij}{}^{*} = \left(x_{ij} - \overline{x_j}\right)/s_{x_j}$
  $Y_i{}^{*} = \left(Y_i - \bar{Y}\right)/s_Y$
- **Residual plots** for assumption checking
  - Residual = obs – exp = $Y_i - \widehat{Y_i}$ (estimated error)
  - Residual quantile normal plot: random scatter around reference line → Normality OK



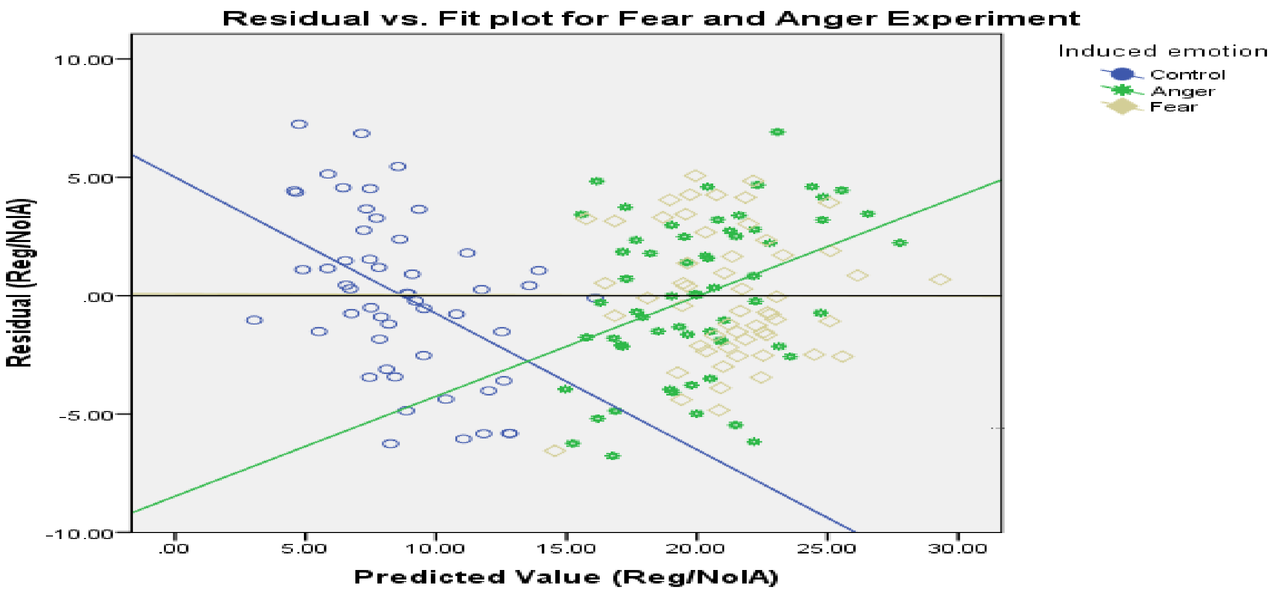
Example: residual analysis, cont.

- **Residual plots** for assumption checking
  - Residual vs. fit (predicted) plot

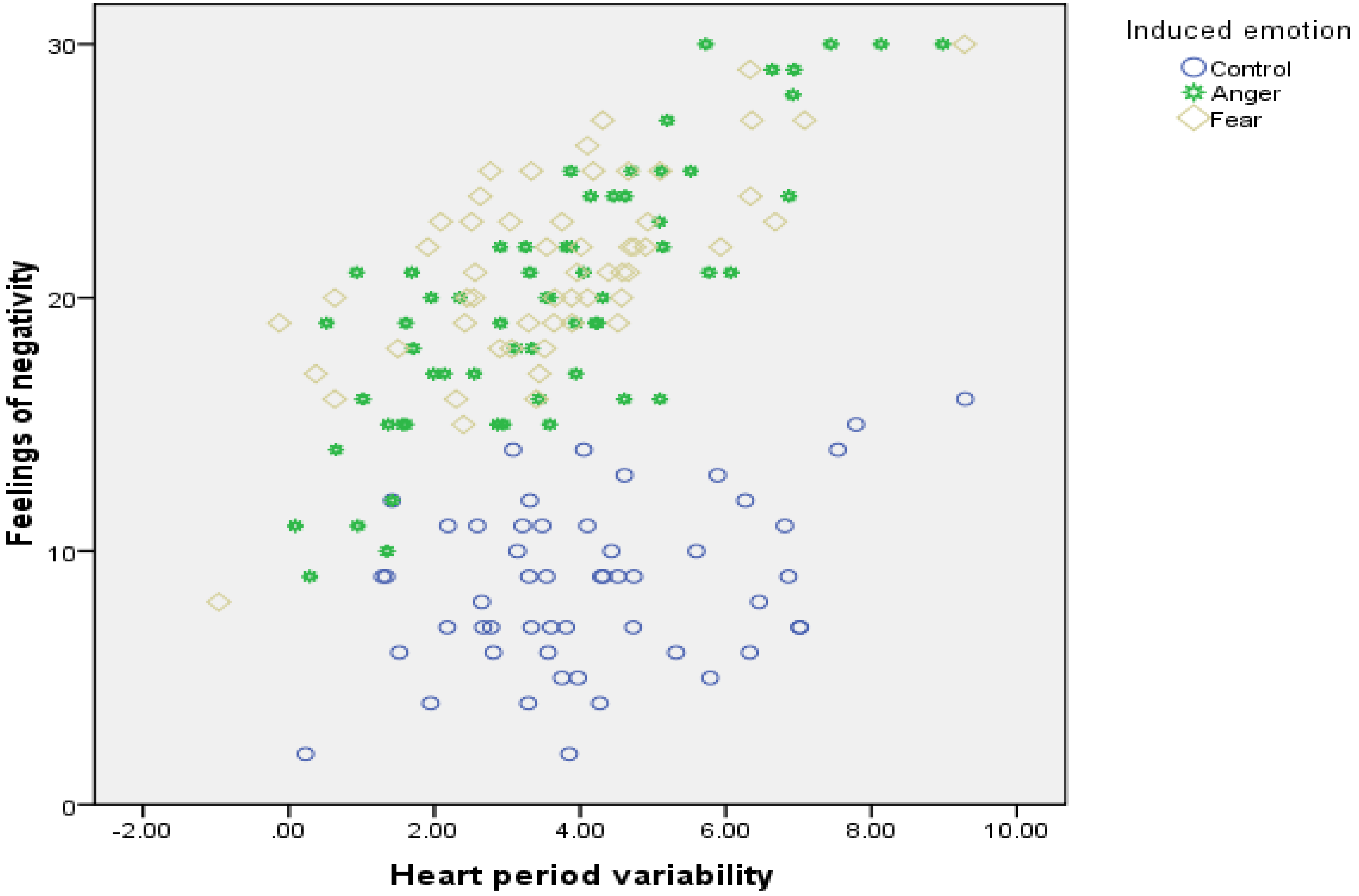
    y-axis: residuals, x-axis: fitted values

    Smile or frown suggests non-linearity (bad means model)

    Funneling suggests unequal variance



Example: skipped EDA



Multiple Regression: interaction

- An **interaction** between ***two IVs*** in their ***effect on the DV*** implies ***non-additivity***. The effect of a one unit increase in x_1 depends on the level of x_2 (and vice versa).
- Interaction is ***coded*** by adding a new IV which is the product of the two original IVs. If x_1 and x_2 are both quantitative there is one new IV, x_1*x_2. If one is a k-level factor there are k-1 new IVs.
- ANCOVA with interaction
  - Structural model: $E(Y|x_1,x_2) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_{12} x_1 x_2$
  - Prediction: $\widehat{Y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_{12} x_1 x_2$

Example with interaction

- E(Negativity|HPV, emotion) = E(N|H,A,F) $= \beta_0 + \beta_H H + \beta_A A + \beta_F F + \beta_{H*A} HA + \beta_{H*F} HF$

  Key step: simplification

  Key concept: β's are fixed; H,A,F are data values

  Controls: $E(N|H,A=0,F=0) = \beta_0 + \beta_H H$

  Anger: $E(N|H,A=1,F=0) = \beta_0 + \beta_H H + \beta_A + \beta_{H*A} H$

  $= (\beta_0 + \beta_A) + (\beta_H + \beta_{H*A})H$

  Fear: $E(N|H,A=0,F=1) = (\beta_0 + \beta_F) + (\beta_H + \beta_{H*F})H$

Example with Interaction: Results

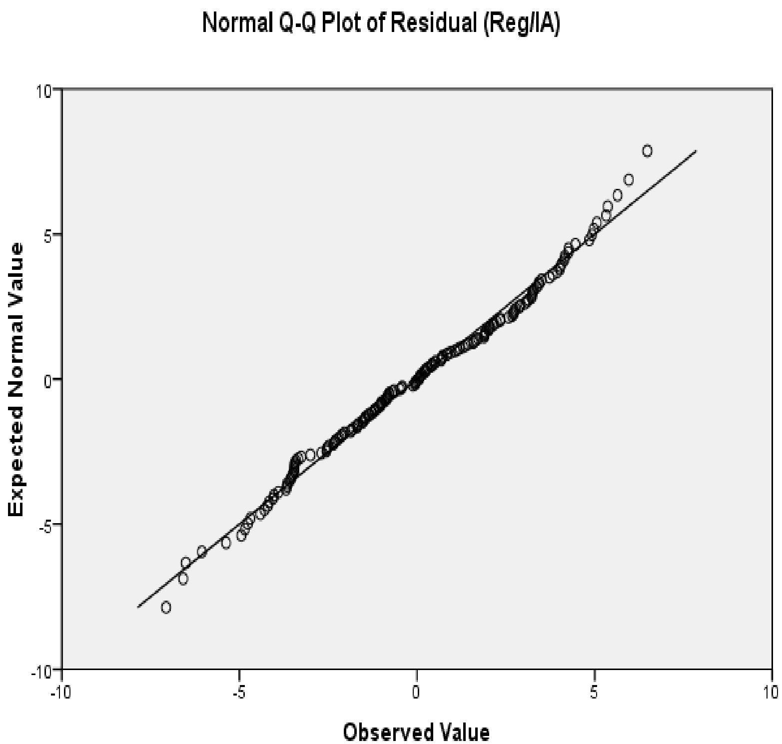
Model Summary				
Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	0.891	0.794	0.792	3.181
2	0.906	0.821	0.816	2.982

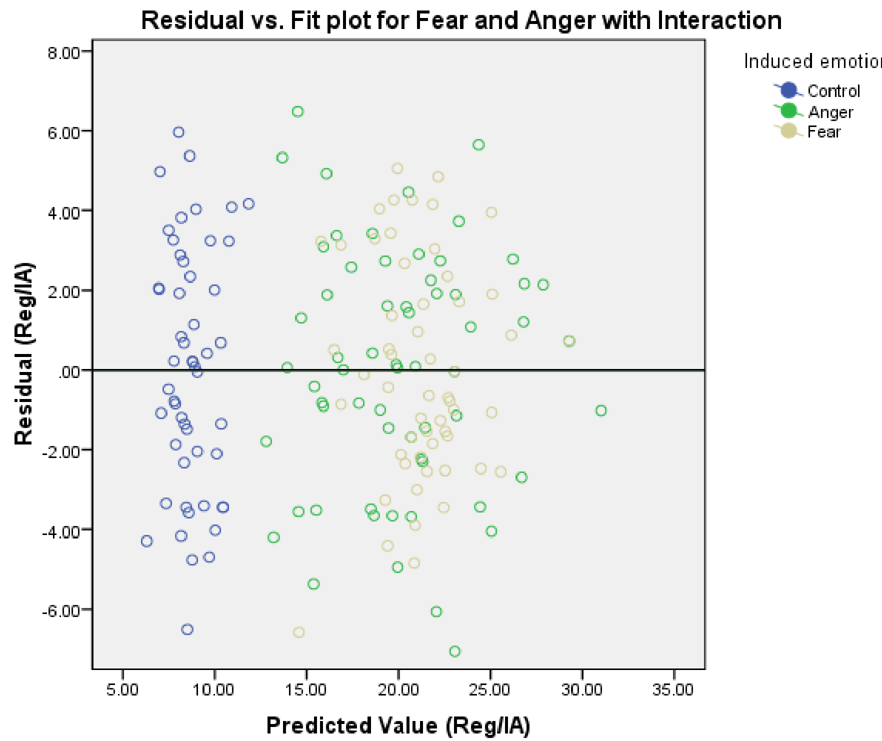
Model	Change Statistics				
	R Square Change	F Change	df1	df2	Sig. F Change
1	.794	212.119	3	165	.000
2	.027	12.359	2	163	.000

Coefficients[a]					
Model		Unstandardized Coefficients		t	Sig.
		B	Std. Error		
1	(Constant)	2.707	.691	3.920	.000
	Heart period variability	1.442	.128	11.263	.000
	anger	12.118	.602	20.141	.000
	fear	13.233	.618	21.397	.000
2	(Constant)	6.153	1.003	6.135	.000
	Heart period variability	.612	.220	2.780	.006
	anger	6.454	1.272	5.073	.000
	fear	9.807	1.350	7.264	.000
	HPV*Anger	1.439	.289	4.972	.000
	HPV*Fear	.825	.312	2.643	.009

a. Dependent Variable: Feelings of negativity

Example with interaction: Diagnostics





QN plot: OK for Normality

Res. vs. Fit: OK for linearity and equal spread

Example: Subject Matter Conclusions

- There is a statistically significant interaction (F_{change}=12.4, df=2, 163, p<0.0005) between HPV and emotion in their effects on negativity (N).
- Heart period variability is positively associated with negativity (t=2.78, df=163, p=0.006) in controls, and the estimated mean change in N is a rise of 0.612 points for each 1 unit rise in HPV (95% CI=[0.18,1.05].
- The estimated mean N when HPV=0 is 6.15 for controls, and is 6.45 higher for induced anger (t=5.07, df=163, p<0.0005) and 9.81 higher for induced fear (t=7.26, df=163, p<0.0005).
- The change in N associated with a 1 point rise in HPV is estimated to be 1.44 points greater for anger compared to control (t=4.97, df=163, p<0.0005) and 0.82 points greater for fear compared to control (t=2.63, df=163, p=0.009).
- Overall compared to control inducing fear and anger increases N, and the increase is greater when HPV is greater.

Class Summary

- In multiple regression, the means model ***adds terms of the form*** $\boldsymbol{\beta_v V}$ when variable V is added.
- Any ***k-level categorical variable*** must be ***replaced with k-1 indicator variables*** [or similar].
- ***Without interaction, a "parallel" means model is produced***: at each level of one IV the slope of the DV vs. the other IV is the same.
- ***With interaction (adding product variables) different slopes are accommodated***.
- You can deduce the meanings of parameters by ***simplifying the means model for each category*** to a Y=a+bX form where a is the intercept and b is the slope.
- Continue the deduction by finding equations that differ only in a single parameter. The p-value for that parameter is the null hypothesis that that β=0 which is equivalent to the two lines being the same.