The University of Georgia

The Analysis of Covariance

ERSH 8310

Keppel and Wickens Chapter 15

Today's Class

- Initial Considerations
- Covariance and Linear Regression
- The Linear Regression Equation
- The Analysis of Covariance
- Assumptions Underlying the Analysis of Covariance
- Example ANCOVA Analysis

THE ANALYSIS OF COVARIANCE

The Analysis of Covariance

- In a completely randomized design subjects are randomly assigned to the experimental treatments
- The completely randomized design is relatively deficient in power
- If there is a variable available before the start of experiment that is reasonably correlated with the dependent variable (i.e., control variable, concomitant variable, or covariate; e.g., intelligence, grade point average, etc.), we may either employ blocking or statistical adjustment

The Analysis of Covariance

- The randomized-blocks design includes groups of homogeneous subjects drawn from respective blocks

- Advantages of the randomized-blocks design are:
  - Blocking helps to equate the treatment groups before the start of the experiment more effectively than is accomplished in the completely randomized design
  - The power is increased because smaller error term usually associated with the blocking design
  - Interactions can be assessed

The Analysis of Covariance

- Disadvantages may include:
  - There will be cost of introducing the blocking factor
  - It may be difficult to find blocking factors that are highly correlated with the dependent variable
  - Loss of power may occur if a poorly correlated blocking factor is used

The Analysis of Covariance

- The analysis of covariance reduces experimental error by statistical, rather than experimental, means

- Subjects are first measured on the concomitant variable called the covariate which consists of some relevant ability or characteristics

- Subjects are then randomly assigned to the treatment group without regard for their scores on the covariate

The Analysis of Covariance

- The analysis of covariance refines estimates of experimental error and uses the adjust treatment effects for any differences between the treatment groups that existed before the experimental treatments were administered

Today's Example Data

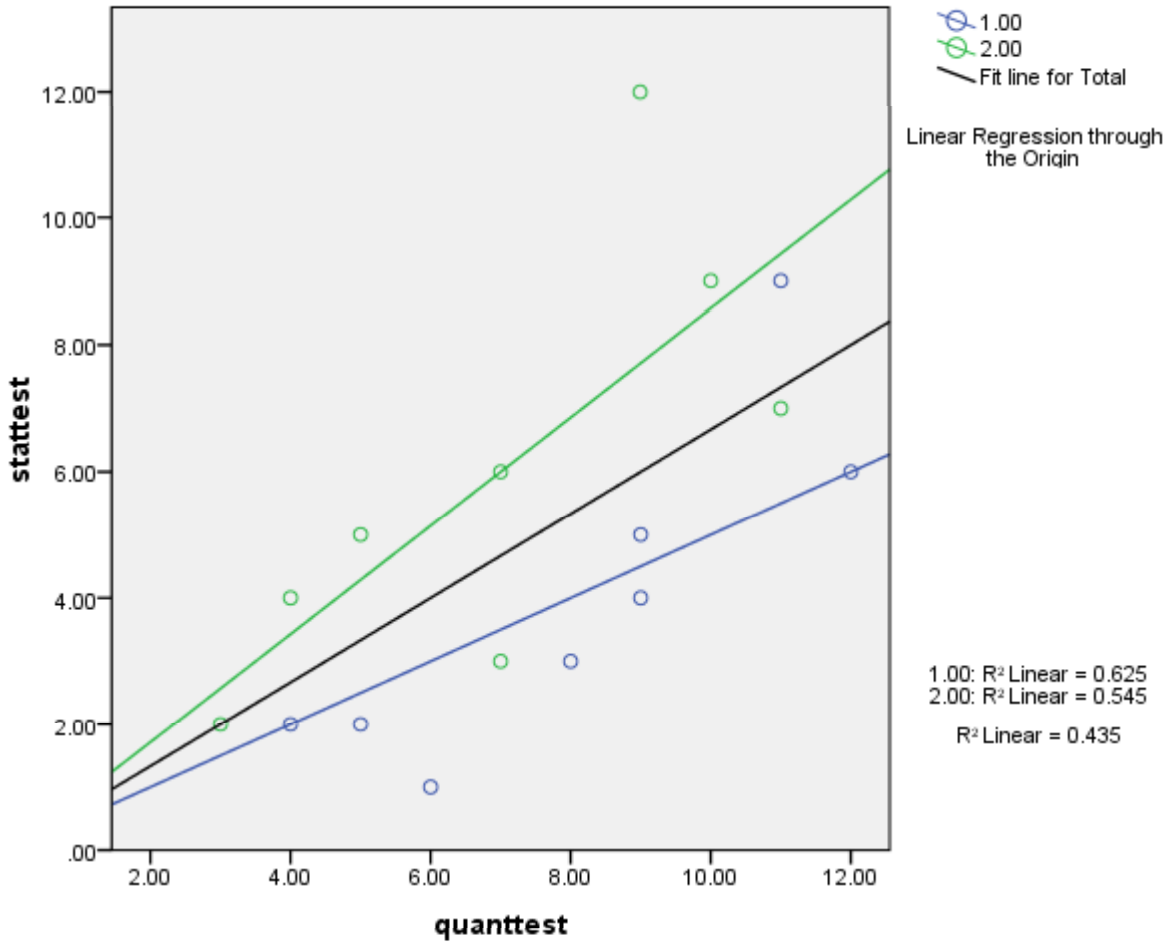
- We will rehash data from Section 11.5

- A researcher was studying the effects of instructional material on how well college students learn basic concepts in statistics
  - Two instructional material groups (a = 2)
  - Pretest given – General quantitative ability (now called the covariate X)
  - DV – Scores on a test on basic statistics test (Y)

The Data

	Group a_1		Group a_2	
	X	Y	X	Y
	11	9	3	2
	6	1	7	3
	5	2	11	7
	8	3	9	12
	9	5	5	5
	4	2	10	9
	9	4	7	6
	12	6	4	4
Mean:	8	4	7	6

Our Data – As a Plot…



COVARIANCE AND LINEAR REGRESSION

Covariance and Linear Regression

- The correlation coefficient between two variables X and Y is:

$$r_{XY} = \frac{s_{xy}}{s_x s_y}$$

- The standard deviation of X is:

$$s_x = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N-1}}$$

- Where the covariance between X and Y is:

$$s_{XY} = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{N-1}$$

- The standard deviation of Y is:

$$s_Y = \sqrt{\frac{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}{N-1}}$$

With Our Data Set

- The correlation coefficient between two variables X and Y is:

$$r_{XY} = \frac{s_{xy}}{s_x s_y} = \frac{5.667}{2.805 * 3.055} = 0.661$$

- The standard deviation of X is:

$$s_x = \sqrt{\frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2}{N-1}} = 2.805$$

- Where the covariance between X and Y is:

$$s_{XY} = \frac{\sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{N-1} = 5.667$$

- The standard deviation of Y is:

$$s_Y = \sqrt{\frac{\sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2}{N-1}} = 3.055$$

Other Useful Terms (Will Help in ANCOVA)

- We may define the sum of products:

$$SP_{XY} = \sum_{i=1}^{N} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right) = 85$$

- And the sums of squares:

$$SS_X = \sum_{i=1}^{N} \left(X_i - \overline{X}\right)^2 = 118 \qquad SS_Y = \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2 = 140$$

- Consequently:

$$r_{XY} = \frac{SP_{XY}}{\sqrt{SS_X SS_Y}} = \frac{85}{\sqrt{118 * 140}} = 0.661$$

THE LINEAR REGRESSION EQUATION

The Linear Regression Equation

- The linear regression line relating the dependent variable Y to the covariate X is:

$$Y_i = \beta_0 + \beta_1 X_i + E_i$$

- And the prediction equation for i is:

$$\hat{Y}_i = b_0 + b_1 X_i$$

- Where:

$$b_0 = \overline{Y} - b_1 \overline{X} \qquad b_1 = r_{XY} \frac{S_Y}{S_X} = \frac{SP_{XY}}{SS_X}$$

With Our Data…

- The linear regression line relating the dependent variable Y to the covariate X is:

$$Y_i = \beta_0 + \beta_1 X_i + E_i$$

- And the prediction equation for i is:

$$\hat{Y}_i = b_0 + b_1 X_i = -.403 + .720 X_i$$

- Where:

$$b_0 = \overline{Y} - b_1 \overline{X} = -.403 \qquad b_1 = r_{XY} \frac{S_Y}{S_X} = \frac{SP_{XY}}{SS_X} = .720$$

Regression Hypothesis Tests

- In regression, the key hypothesis test we are worried about is that for the slope, b_1
  - H_0: $\beta_1 = 0$
  - H_1: $\beta_1 \neq 0$

- Using these hypotheses, and our method for calculating sums of squares from Chapter 14 (the GLM chapter), we can then create an ANOVA-like table for the regression.
- We need: $SS_{unexp}^{H_0}$ $SS_{unexp}^{H_1}$

Regression Sums of Squares

- Recall that $SS_{unexp}^{H_0}$ came from calculating the sums of squares where the predicted value came from the linear model when H_0 was true
  - It was treatment + error
  - In regression the corresponding predicted values are:

$$\hat{Y}_i = b_0 = \overline{Y}$$

  - This comes from:

$$b_0 = \overline{Y} - b_1\overline{X} = \overline{Y} - 0 * \overline{X} = \overline{Y}$$

- The $df_{unexp}^{H_0}$ comes from the number of observations (here 16) minus the number of parameters (1)…so…

$$df_{unexp}^{H_0} = 15$$

Regression Sums of Squares

- Recall that $SS_{unexp}^{H_1}$ came from calculating the sums of squares where the predicted value came from the linear model when H_1 was true
  - It was error
  - In regression the corresponding predicted values are:

$$\hat{Y}_i = b_0 + b_1 X_i$$

- The $df_{unexp}^{H_1}$ comes from the number of observations (16) minus the number of parameters (2), so…

$$df_{unexp}^{H_1} = 14$$

Residual Variation and the Linear Model

- Another way of looking at Sums of Squares:
- The sum of the squared deviation from the mean is the Sums of Squares Total:

$$SS_T = SS_Y = \sum_{i=1}^{N} \left(Y_i - \bar{Y}\right)^2 = SS_{un\exp}^{H_0}$$

- The sum of the squared residuals from the regression line is:

$$SS_{\mathrm{Re}\,gression} = SS_{Y|X} = \sum_{i=1}^{N} \left(\hat{Y}_i - \bar{Y}\right)^2 = SS_{un\exp}^{H_0} - SS_{un\exp}^{H_1}$$

- Finally, the sum of squares for error is:

$$SS_{Error} = \sum_{i=1}^{N} \left(Y_i - \hat{Y}\right)^2 = SS_{un\exp}^{H_1}$$

Putting it Together in an ANOVA Table:

- See Excel worksheet on ELC for computations

Observed	Predicted (H1)	Predicted (H0)		Deviations			Squared Deviations	
Y	Y-hat	Y-bar		Y-Y-hat	Y-Y-bar		Y-Y-hat	Y-Y-bar
9	7.521	5		1.479	4		2.187441	16
1	3.919	5		-2.919	-4		8.520561	16
2	3.199	5		-1.199	-3		1.437601	9
3	5.36	5		-2.36	-2		5.5696	4
5	6.081	5		-1.081	0		1.168561	0
2	2.479	5		-0.479	-3		0.229441	9
4	6.081	5		-2.081	-1		4.330561	1
6	8.242	5		-2.242	1		5.026564	1
2	1.758	5		0.242	-3		0.058564	9
3	4.64	5		-1.64	-2		2.6896	4
7	7.521	5		-0.521	2		0.271441	4
12	6.081	5		5.919	7		35.03456	49
5	3.199	5		1.801	0		3.243601	0
9	6.801	5		2.199	4		4.835601	16
6	4.64	5		1.36	1		1.8496	1
4	2.479	5		1.521	-1		2.313441	1
						Sums	78.76674	140
						df	14	15
		SS	df	MS	F	p-value		
	Regression	61.233261	1	61.23326	10.8836	0.005272		
	Error	78.766739	14	5.626196				
	Total	140	15					

Residual Variation and the Linear Model

- The amount of the Y variability that can be attributed to the regression equation is:

$$SS_{regression} = SS_Y - SS_{Y|X}$$

- The squared correlation coefficient (or what we call the effect size eta-squared) is:

$$r_{XY}^2 = \frac{SS_{regression}}{SS_Y} = .437$$

Now with SPSS

Model Summary[b]

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.661[a]	.437	.397	2.37203

a. Predictors: (Constant), quanttest

b. Dependent Variable: stattest

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	61.229	1	61.229	10.882	.005[a]
	Residual	78.771	14	5.627		
	Total	140.000	15			

a. Predictors: (Constant), quanttest

b. Dependent Variable: stattest

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	-.403	1.742		-.231	.821
	quanttest	.720	.218	.661	3.299	.005

a. Dependent Variable: stattest

THE ANALYSIS OF COVARIANCE

The Analysis of Covariance

- Now that we know a little about regression, we can use it to improve power to detect group differences in an ANOVA-like experimental design

- ANCOVA tests for differences between groups by comparing a description of the data based on a single regression line to one based on lines with the same slope and different intercepts for each group

In the Beginning…There was the ANOVA Model

- Recall our GLM – representation of ANOVA:

- The H_1 model:

$$Y_{ij} = \mu_T + \alpha_j + E_{ij}$$

- The H_0 model:

$$Y_{ij} = \mu_T + E_{ij}$$

- ANCOVA adds in the covariate X to the model…

The Analysis of Covariance and the General Linear Model

- For the analysis of covariance, the alternative-hypothesis model is:

$$Y_{ij} = \beta_0 + \alpha_j + \beta_1 X_{ij} + E_{ij}$$

- And the null-hypothesis model (testing the effect of the treatment means) is:

$$Y_{ij} = \beta_0 + \beta_1 X_{ij} + E_{ij}$$

Decomposing ANCOVA

- If the H_1 model holds in ANCOVA, it means that the intercepts of the regression line differ across groups:

- For our 2-group example that means, for group 1:

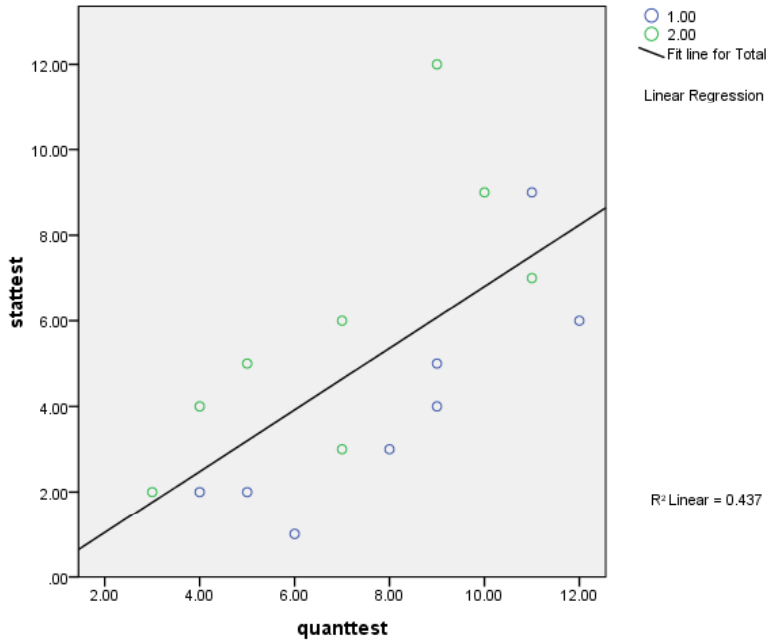
$$Y_{i1} = (\beta_0 + \alpha_1) + \beta_1 X_{i1} + E_{i1}$$

- For group 2:

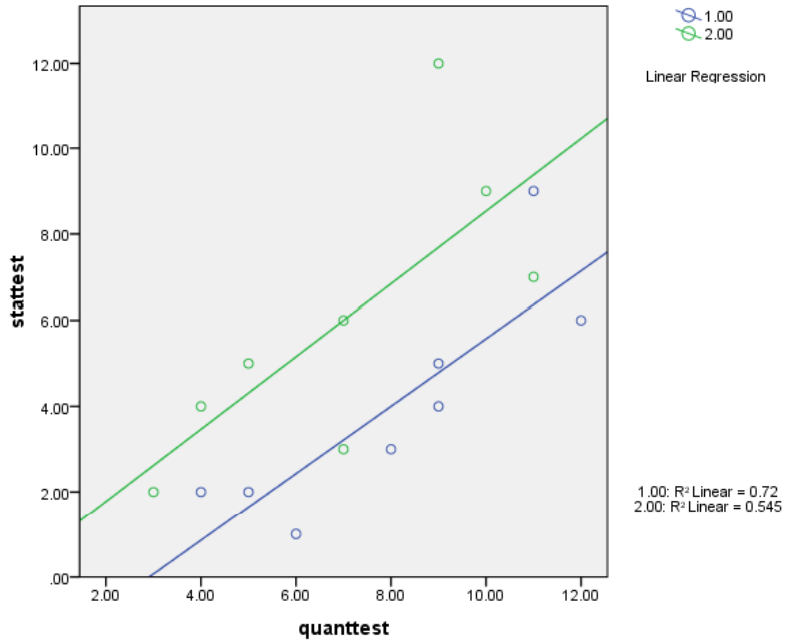
$$Y_{i2} = (\beta_0 + \alpha_2) + \beta_1 X_{i2} + E_{i2}$$

Does it? Hmm…

H_0 Model – same intercepts



H_1 Model – Diff. Intercepts



...Its a Question of Statistical Significance

- Using our knowledge of GLM, we could form SS for each of the possible categories…or we could use SPSS:

Tests of Between-Subjects Effects

Dependent Variable:stattest

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	91.868[a]	2	45.934	12.407	.001
Intercept	2.250	1	2.250	.608	.450
quanttest	75.868	1	75.868	20.492	.001
group	30.640	1	30.640	8.276	.013
Error	48.132	13	3.702		
Total	540.000	16			
Corrected Total	140.000	15			

a. R Squared = .656 (Adjusted R Squared = .603)

Covariate (X – the score on the general quantitative test)

Treatment group (A)

Comparing ANOVA and ANCOVA

ANOVA

Tests of Between-Subjects Effects

Dependent Variable:stattest

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	16.000[a]	1	16.000	1.806	.200
Intercept	400.000	1	400.000	45.161	.000
group	16.000	1	16.000	1.806	.200
Error	124.000	14	8.857		
Total	540.000	16			
Corrected Total	140.000	15			

a. R Squared = .114 (Adjusted R Squared = .051)

Estimated Marginal Means

group

Dependent Variable:stattest

group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1.00	4.000	1.052	1.743	6.257
2.00	6.000	1.052	3.743	8.257

ANCOVA

Tests of Between-Subjects Effects

Dependent Variable:stattest

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	91.868[a]	2	45.934	12.407	.001
Intercept	2.250	1	2.250	.608	.450
quanttest	75.868	1	75.868	20.492	.001
group	30.640	1	30.640	8.276	.013
Error	48.132	13	3.702		
Total	540.000	16			
Corrected Total	140.000	15			

a. R Squared = .656 (Adjusted R Squared = .603)

Estimated Marginal Means

group

Dependent Variable:stattest

group	Mean	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
1.00	3.592[a]	.686	2.110	5.075
2.00	6.408[a]	.686	4.925	7.890

a. Covariates appearing in the model are evaluated at the following values: quanttest = 7.5000.

The Difference? Adjusted Means

- For ANCOVA, the adjusted mean (or Least Squares Mean) is:

$$\overline{Y}' = \overline{Y}_j - b_1\left(\overline{X}_i - \overline{X}_T\right)$$

What about the Rest of ANOVA

- All of the rest of ANOVA happens on the adjusted means…like contrasts…or post hoc tests
  - Both of which are made easy in SPSS
- The design can be extended to include more factors and more covariates
- Of course, things get more tedious to understand as you add more variables
- The key is to realize that if your covariate significantly predicts Y (above and beyond the group variable), you will have more power to detect group differences

ASSUMPTIONS UNDERLYING THE ANALYSIS OF COVARIANCE

Assumptions Underlying the Analysis of Covariance

- Three assumptions in addition to the usual analysis of variance assumptions are:
  1. The assumption of linear regression: The deviations from regression are normally and independently distributed in the population, with means of zero and homogeneous variances
     - Hard to check definitively
  2. The assumption of homogeneous group regression coefficients: The within groups regression coefficient is actually an average of the regression coefficients for the respective treatment groups
     - Can be checked by adding an interaction term between the covariate and group variable
  3. The exact measurement of the covariate: The covariate is measured without error
     - Ubiquitous term

Checking the Homogeneity of Slopes Assumption

- You can easily check the homogeneity of slopes assumption by testing the interaction between your group IV and the covariate

- The H_1 model becomes:

$$Y_{ij} = \beta_0 + \alpha_j + \beta_1 X_{ij} + (\alpha\beta)_j X_{ij} + E_{ij}$$

- If true, it implies, for group 1:

$$Y_{i1} = (\beta_0 + \alpha_1) + [\beta_1 + (\alpha\beta)_1] X_{i1} + E_{i1}$$

- And for group 2:

$$Y_{i2} = (\beta_0 + \alpha_2) + [\beta_1 + (\alpha\beta)_2] X_{i2} + E_{i2}$$

In SPSS

- Does it meet our assumption of homogeneity of slopes?

Tests of Between-Subjects Effects

Dependent Variable:stattest

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	91.968[a]	3	30.656	7.659	.004
Intercept	2.162	1	2.162	.540	.476
group	2.514	1	2.514	.628	.443
quanttest	75.749	1	75.749	18.925	.001
group * quanttest	.100	1	.100	.025	.877
Error	48.032	12	4.003		
Total	540.000	16			
Corrected Total	140.000	15			

a. R Squared = .657 (Adjusted R Squared = .571)

Wrapping Up…

- Today's class covered a method for controlling for important variables in an experiment: ANCOVA
- ANCOVA is a general technique that adds additional (continuous/quantitative) variables to a model and adjusts for the values of such variables
- Any ANOVA design can include such variables

Up Next…

- In lab tonight:
  - How to do ANCOVA in SPSS
- Homework:
  - Assigned in the morning, due next week before class
- Reading for next week:
  - Chapter 16: Within Subject Designs
- Only two lectures left!