Analysis of Covariance

Preface

Analysis of covariance (ANCOVA) is a statistical methodology that combines the concept of analysis of variance (ANOVA) and the concept of linear regression analysis. In this paper, we will define and illustrate analysis of covariance by reviewing our previous example of one-way analysis of variance between subjects and by extending that example to include an instance of covariance. This paper presumes knowledge of simple linear correlation and regression. An appendix to this paper is a Microsoft Excel workbook that consists of the numerical calculations that are embedded in this paper.

Gerry Del Fiacco
Math Center
Metropolitan State University
St. Paul, Minnesota
August 4, 2014

Analysis of Variance

In our previous example of one-way analysis of variance between subjects, we examined whether or not the method of training factory workers had a significant effect on the proficiency of those workers. The independent variable for this study was a categorical variable with three values:

No Training

Basic Training

Enhanced Training

Three randomly chosen groups of workers were selected, that is, one group for each of the three respective training methods. The dependent variable for this study was the number of processing errors per worker when working with a new fabricating machine. The mean number of processing errors per group was determined to be as follows:

Group	**Training**	**Mean Number of Errors Per Worker for Each Group**	**Grand Mean Number of Errors for All Workers**
1	No Training	$\overline{X}_1 = 9.625$	$\overline{\overline{X}} = 6.583$
2	Basic Training	$\overline{X}_2 = 5.875$	
3	Enhanced Training	$\overline{X}_3 = 4.250$	

The parameters for the analysis of variance calculations were as follows:

Number of Groups	$k = 3$
Number of Workers Per Group	$n = 8$
Total Number of Workers in the Study	$N = 24$

The calculations for the analysis of variance had the following degrees of freedom:

Degrees of Freedom TOTAL	$df_{TOTAL} = N - 1 = 23$
Degrees of Freedom BETWEEN	$df_{BETWEEN} = k - 1 = 2$
Degrees of Freedom WITHIN	$df_{WITHIN} = N - k = 21$

The F-statistic and p-value for the ANOVA test of statistical significance were as follows:

$$F = \frac{MS_{BETWEEN}}{MS_{WITHIN}} = \frac{60.7917}{2.3929} = 25.405$$

$$\text{p} - \text{value} = 0.000002$$

These test results provided evidence that the expected number of errors per worker was not the same for all three methods of training. And, a post-hoc test identified the significant differences between the means:

Tukey HSD Test	Interpretation
$\left\|\overline{X}_1 - \overline{X}_2\right\| = \left\| 9.63 - 5.88 \right\| = 3.75 > 1.96$	Basic training is significantly better than no training.
$\left\|\overline{X}_1 - \overline{X}_3\right\| = \left\| 9.63 - 4.25 \right\| = 5.38 > 1.96$	Enhanced training is significantly better than no training.
$\left\|\overline{X}_2 - \overline{X}_3\right\| = \left\| 5.88 - 4.25 \right\| = 1.63 < 1.96$	There is no significant difference between basic training and enhanced training.

Concept of Covariance

Now, we will expand the scope and complexity of analysis of variance by introducing the concept of covariance into this study of factory workers. The example of covariance that we will use is the years of experience per factory worker. That is, if the average number of years of experience per factory worker was not the same for each of the three groups of workers in the study, we could reasonably conjecture that the mean number of errors per training group was influenced by the difference in the level of experience per group as well as by the difference in training method per group.

Covariate - Years of Experience							
V_1	V_1^2		V_2	V_2^2		V_3	V_3^2
13	169		11	121		6	36
6	36		9	81		10	100
12	144		12	144		8	64
4	16		8	64		10	100
15	225		10	100		8	64
9	81		18	324		11	121
17	289		9	81		13	169
10	100		14	196		11	121
86	1060		91	1111		77	775
$\overline{V}_1$	10.75		$\overline{V}_2$	11.375		$\overline{V}_3$	9.625
$\overline{\overline{V}}$	10.583						

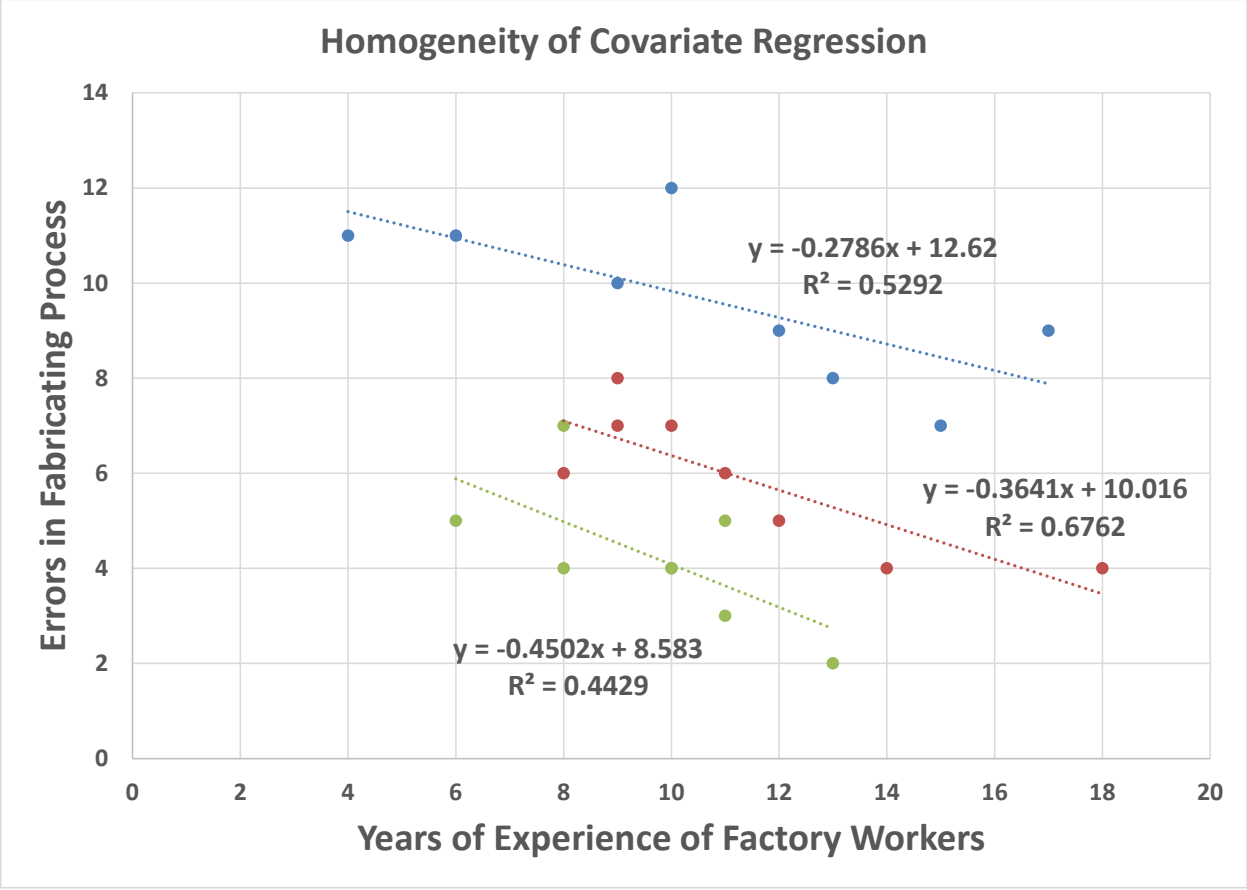
This table of data depicts the number of years of experience for each worker. In this framework, the years of experience per worker is known as a covariant. A covariant is a quantitative variable that is correlated with the quantitative values of the dependent variable within each of the independent groups in the analysis of variance. The scale of measure for the covariant and the scale of measure for the dependent variable do not have to be of the same dimension.

Homogeneity of Regression

Analysis of covariance (ANCOVA) requires several assumptions and limitations upon the research data. Most of these assumptions and limitations are similar to those imposed upon research data for analysis of variance (ANOVA). Principally, these assumptions and limitations deal with such issues as outliers, normality, independence and homogeneity of variance between groups.

However, a unique constraint that is required before we can conduct analysis of covariance is that the linear regression equations that represent the linear relationship between the covariant and the dependent variable should have substantially the same slope for each of the groups. This constraint is known as homogeneity of regression and can be ascertained by a formal test of statistical significance.

In our example of analysis of covariance, the meaning and presence of homogeneity of regression can be depicted in the following diagram:



In this diagram, the x-axis represents years of experience per worker and the y-axis represents errors per worker in the fabricating process. The linear relationship between these two variables is captured in the following linear regression equations for the respective training groups:

Group	Training	Linear Regression Equation
1	No Training	$y = -0.2786 \cdot x + 12.62$
2	Basic Training	$y = -0.3641 \cdot x + 10.016$
3	Enhanced Training	$y = -0.4502 \cdot x + 8.583$

The downward slope of the three lines in the diagram is essentially the same. The presence of homogeneity of regression essentially tells us that an increase in the years of experience is associated in a consistent manner with a reduced number of errors in the fabricating process within each of the three given training groups. This leads to the question of whether or not the three training groups have the same average years of experience among the workers in each group:

Group	Training	Average Years of Experience Among Workers in Each Group	Grand Average Years of Experience for All Workers
1	No Training	$\overline{V}_1 = 10.75$	$\overline{\overline{V}} = 10.583$
2	Basic Training	$\overline{V}_2 = 11.375$	
3	Enhanced Training	$\overline{V}_3 = 9.625$	

This table tells us that the average years of experience are significantly different across the three training groups. This suggests that the mean number of errors per training group was influenced by the difference in the years of experience as well as by the difference in training method per group. It is the purpose of analysis of covariance to investigate this conjecture and to contend with its consequences.

Partitioning of the Variation in Data for ANOVA

In analysis of variance, the variation in the research data is partitioned in a simple and easily understandable manner as follows:

ANOVA		
	Variation Between Groups	**Variation Within Groups**
Variation in the Dependent Variable	$SS_{BETWEEN}$	SS_{WITHIN}

In this partitioning, the variation between groups is due to the effects of the independent variable (method of training) and the variation within groups is attributed to statistical error. The calculations of these sources of variation were as follows in our previous one-way ANOVA between subjects study:

$$SS_{TOTAL} = \sum X^2 - \frac{\left(\sum X_1 + \sum X_2 + \sum X_3\right)^2}{N}$$

$$= 1212 - \frac{(77 + 47 + 34)^2}{24} = 171.8333$$

$$SS_{WITHIN} = \sum X^2 - \frac{\left(\sum X_1\right)^2 + \left(\sum X_2\right)^2 + \left(\sum X_3\right)^2}{n}$$

$$= 1212 - \frac{(77)^2 + (47)^2 + (34)^2}{8} = 50.250$$

$$SS_{BETWEEN} = SS_{TOTAL} - SS_{WITHIN} = 121.5833$$

As we proceed with our analysis of covariance, we will derive adjusted values of the variation between groups and the variation within groups.

Partitioning of the Variation in Data for ANCOVA

In analysis of covariance, the variation in the research data is partitioned in a more elaborate manner:

ANCOVA		
	Variation Between Groups	**Variation Within Groups**
Variation in the Covariate Data	$SSV_{BETWEEN}$	SSV_{WITHIN}
Covariance Between the Dependent Variable and the Covariate Data	$SP_{BETWEEN}$	SP_{WITHIN}
Variation in the Adjusted Dependent Variable	$SS^{*}_{BETWEEN}$	SS^{*}_{WITHIN}

This partitioning of the variation for ANCOVA is based on the premise that we intend to adjust the ANOVA results so as to neutralize the effects of the covariate. That is, we will conduct the ANCOVA study by estimating what the ANOVA results would have been if the average value of the covariate had been equal in each of the groups. In our example, adjusted ANOVA results will be derived from an assumption that each of the three training groups consisted of factory workers with the same average years of experience.

The covariance that exists between the dependent variable and the covariate is the key to making these adjustments. That is, the covariance is a measure of the linear relationship between the dependent variable and the covariate. This linear relationship, which was derived from the original research data, allows us to conjure appropriate adjustments to the dependent variable (errors in the fabricating process) based on a presumption of the same average value of the covariate (years of experience) in each of the groups.

Variation in the Covariate Data

The calculations of variation in the covariate data are summarized as follows:

$$SSV_{TOTAL} = \sum V^2 - \frac{\left(\sum V_1 + \sum V_2 + \sum V_3\right)^2}{N}$$
$$= 2946 - \frac{(86+91+77)^2}{24} = 257.8333$$

$$SSV_{WITHIN} = \sum V^2 - \frac{\left(\sum V_1\right)^2 + \left(\sum V_2\right)^2 + \left(\sum V_3\right)^2}{n}$$
$$= 2946 - \frac{(86)^2 + (91)^2 + (77)^2}{8} = 245.25$$

$$SSV_{BETWEEN} = SSV_{TOTAL} - SSV_{WITHIN} = 12.5833$$

Covariance Between the Dependent Variable and the Covariate Data

The calculations of the covariance are summarized as follows:

$$SP_{TOTAL} = \sum X \cdot V - \frac{\left(\sum X \cdot \sum V\right)}{N}$$
$$= 1609 - \frac{(158 \cdot 254)}{24} = -63.1667$$

$$SP_{WITHIN} = \sum X \cdot V - \frac{\left(\sum X_1 \cdot \sum V_1\right) + \left(\sum X_2 \cdot \sum V_2\right) + \left(\sum X_3 \cdot \sum V_3\right)}{n}$$
$$= 1609 - \frac{(77 \cdot 86) + (47 \cdot 91) + (34 \cdot 77)}{8} = -80.625$$

$$SP_{BETWEEN} = SP_{TOTAL} - SP_{WITHIN} = 17.4583$$

Variation in the Adjusted Dependent Variable

Next, we address the crux of analysis of covariance. We have to adjust the dependent variable so as to neutralize the effect of the covariate. Fortunately, we do not have to adjust each individual value of the dependent variable. We only have to account for the adjustment in the variation of the mean values of the dependent variable that occur between and within each of the three training groups. The necessary calculations are summarized as follows. The adjusted sum of squares between groups is:

$$SS^{*}_{BETWEEN} = SS_{BETWEEN} - \left(\frac{\left(SP_{TOTAL}\right)^2}{SSV_{TOTAL}} - \frac{\left(SP_{WITHIN}\right)^2}{SSV_{WITHIN}} \right)$$

$$= 121.5833 - \left(\frac{\left(-63.1667\right)^2}{257.8333} - \frac{\left(-80.625\right)^2}{245.25} \right) = 132.6133$$

The adjusted sum of squares within groups is:

$$SS^{*}_{WITHIN} = SS_{WITHIN} - \frac{\left(SP_{WITHIN}\right)^2}{SSV_{WITHIN}}$$

$$= 50.250 - \frac{\left(-80.625\right)^2}{245.25} = 23.74484$$

Degrees of Freedom for ANCOVA

The degrees of freedom for this analysis of covariance are:

Degrees of Freedom TOTAL	$df^{*}_{TOTAL} = N - 2 = 22$
Degrees of Freedom BETWEEN	$df^{*}_{BETWEEN} = k - 1 = 2$
Degrees of Freedom WITHIN	$df^{*}_{WITHIN} = N - k - 1 = 20$

ANCOVA Test Results

The mean square calculations of the adjusted sums of squares are:

$$MS^{*}_{BETWEEN} = \frac{SS^{*}_{BETWEEN}}{df^{*}_{BETWEEN}} = \frac{132.6133}{2} = 66.30664$$

$$MS^{*}_{WITHIN} = \frac{SS^{*}_{WITHIN}}{df^{*}_{WITHIN}} = \frac{23.74484}{20} = 1.187242$$

The resulting F-statistic and p-value are:

$$F = \frac{66.30664}{1.187242} = 55.849$$

$$\text{p} - \text{value} = 0.000000$$

These ANCOVA test results provide evidence that the expected number of errors per worker was not the same for all three methods of training. This also was the outcome of the ANOVA test. However, the analysis of covariance adjusts the expected mean values of the dependent variable in each of the training groups.

Adjusted Mean Values of the Dependent Variable

The average years of experience among all factory workers in the study was:

$$\overline{\overline{V}} = 10.583$$

In our analysis of covariance, we revised the original analysis of variance results under the premise of having each of the training groups be comprised of workers with this average years of experience. This resulted in adjusted mean values of the number of errors per worker in each training group. The calculations of the adjusted mean values are based on a pooled value of the slope of the three regression equations. This is essentially a weighted average of the three slopes of those regression equations:

$$B_{POOLED} = \frac{SP_{WITHIN}}{SSV_{WITHIN}} = \frac{-80.625}{245.25} = -0.328746$$

The calculations of the adjusted mean values are:

$$\overline{X}_1^* = \overline{X}_1 - B_{POOLED} \cdot \left(\overline{V}_1 - \overline{\overline{V}}\right) = 9.625 - B_{POOLED} \cdot (10.75 - 10.583) = 9.680$$

$$\overline{X}_2^* = \overline{X}_2 - B_{POOLED} \cdot \left(\overline{V}_2 - \overline{\overline{V}}\right) = 5.875 - B_{POOLED} \cdot (11.375 - 10.583) = 6.135$$

$$\overline{X}_3^* = \overline{X}_3 - B_{POOLED} \cdot \left(\overline{V}_3 - \overline{\overline{V}}\right) = 4.250 - B_{POOLED} \cdot (9.625 - 10.583) = 3.935$$

The comparison between original and adjusted means is as follows:

Group	Training	Original Mean Number of Errors Per Worker for Each Group	Adjusted Mean Number of Errors Per Worker for Each Group
1	No Training	$\overline{X}_1 = 9.625$	$\overline{X}_1^* = 9.680$
2	Basic Training	$\overline{X}_2 = 5.875$	$\overline{X}_2^* = 6.135$
3	Enhanced Training	$\overline{X}_3 = 4.250$	$\overline{X}_3^* = 3.935$

ANCOVA Post-Hoc Test Results

An adjustment in the mean number of errors per group that is of particular interest is that the workers in the basic training group had an average years of experience that was much higher than the other two groups. This influenced the original mean number of errors for the basic training group. After having adjusted for this disparity, the mean number of errors for the basic training group was increased accordingly. This also meant that there is a significant difference between basic training and enhanced training. This difference was not uncovered in the original ANOVA results. Here is a summary of the ANCOVA post-hoc test results:

Tukey HSD Test	Interpretation
$\left\|\overline{X}_1^* - \overline{X}_2^*\right\| = \left\| 9.680 - 6.135 \right\| = 3.545 > 1.38$	Basic training is significantly better than no training.
$\left\|\overline{X}_1^* - \overline{X}_3^*\right\| = \left\| 9.680 - 3.935 \right\| = 5.745 > 1.38$	Enhanced training is significantly better than no training.
$\left\|\overline{X}_2^* - \overline{X}_3^*\right\| = \left\|6.135 - 3.935 \right\| = 2.200 > 1.38$	Enhanced training is significantly better than basic training.

General Importance of Analysis of Covariance

The example of analysis of covariance in this paper is an elementary example of the concept of covariance. Nevertheless, it illustrates the importance of accounting for the presence of covariance. In experimental studies, researchers have to be wary of covariate conditions that are not of interest as independent variables but that may affect the dependent variable. For example, if medical researchers are studying the effects of smoking and alcohol consumption on esophageal cancer, they have to ensure that they have neutralized the variation of other conditions among the subjects of the study, such as, age, ethnicity, etc., that may be correlated with the experimental results.

References

The following textbook includes a clear explanation of the purpose, methods and calculations for conducting analysis of covariance:

"Experimental Designs Using ANOVA," by Barbara G. Tabachnick and Linda S. Fidell, Copyright 2007 by Thomson Brooks/Cole, ISBN 0534405142, pp. 379-394, 424-426.