Statistical formulas

This information is intended for readers with a statistical background who might be interested in some key formulae related to ANCOVA and its extension to situations involving a group difference in the slope of an outcome Y on a covariate X. For more detailed information, see Zar's Biostatistical Analysis [1], a highly recommended and practical treatment of statistics for investigators in the biomedical sciences. References [2-5] are some recommended papers on EE analysis. An excellent web resource for statistical information relevant to biomedical investigators is The Handbook of Biological Statistics (http://udel.edu/~mcdonald/statintro.html).

Key:

N=total sample size

n_1 = sample size for group 1

n_2 = sample size for group 2

$\bar{Y}_1$= mean of Y variable (the response variable) for group 1

$\bar{Y}_2$= mean of Y variable (the response variable) for group 2

$\bar{X}_1$= mean of X variable (the covariate) for group 1

$\bar{X}_2$= mean of X variable (the covariate) for group 2

b_c= pooled slope estimate = $\frac{\sum x_1^2 b_1 + \sum x_2^2 b_2}{\sum x_1^2 + \sum x_2^2}$

b_1= slope estimate for group 1

b_2= slope estimate for group 2

$\sum x_1^2$= sum of squared deviations from $\bar{X}_1$

$\sum x_2^2$= sum of squared deviations from $\bar{X}_2$

S^2= residual variance estimate (estimated as mean squared error of the regression fit) = $\frac{residual\ SS1 + residual\ SS\ 2}{residual\ df1 + residual\ df2}$

residual SS1 = residual sum of squares from regression of Y on X in group 1

residual SS2 = residual sum of squares from regression of Y on X in group 2

residual df1 = residual degrees of freedom for regression of Y on X in group 1

residual df2 = residual degrees of freedom for regression of Y on X in group 2

Student's t-test for the significance of the adjusted group comparison The test to compare the height of two parallel elevations in ANCOVA is the test for the adjusted group comparison. This test can be done using a Student's t-test on N-3 degrees of freedom as follows:

$$t = \frac{\bar{Y}_1 - \bar{Y}_2 - b_c(\bar{X}_1 - \bar{X}_2)}{\sqrt{S^2(\frac{1}{n_1} + \frac{1}{n_2} + \frac{(\bar{X}_1 - \bar{X}_2)^2}{\sum x_1^2 + \sum x_2^2})}}$$

Notes:

1. The numerator of the t statistic is an expression for the difference in the Y-intercepts and is derived from the fact that the regression line of each group passes through the coordinates $(\bar{X}_i, \bar{Y}_i)$
2. If the groups have the same X means, the t-test reduces a t-test on the Y means, but may have improved precision over a standard unadjusted t-test on the Y means.

Standard error of the mean adjusted Y The SE of the mean adjusted Y for a group (here group 1) is:

$$\sqrt{S^2(\frac{1}{n_1} + \frac{\left(\bar{X}_1 - \bar{X}_{grand}\right)^2}{\sum x_1^2 + \sum x_2^2})}$$

Student's t-test to compare two slope estimates The Student's t test (on n1 – n2 – 4 degrees of freedom) to compare two slope estimates is:

$$t = \frac{b_1 - b_2}{\sqrt{S^2(\frac{1}{\sum x_1^2} + \frac{1}{\sum x_2^2})}}$$

A simple way to test whether $\hat{Y}$ differs between two groups at a particular value of a covariate $X = X_0$ when the slopes of the regression lines differ but the unexplained variances about the regression lines do not differ (a simple alternative to the Johnson-Neyman technique)

1. Compute a transformed covariate X* that is "centered" on Xo by subtracting Xo from every value of X, i.e., X*= (X-Xo).
2. Run a regression that includes as predictor variables *group, X*, and the group by X* interaction term*.

3. The p-value for group correctly indicates whether the groups differ significantly at Xo. The t statistic and p-value will be exactly the same as the values calculated using the formula shown on p.361 of Zar (1996) [1] for comparing points on two regression lines (discussed below).
4. CAUTIONARY NOTE: In a regression model that includes as predictor variables *group, X and the group by X interaction term* where X is the untransformed covariate, the p-value for group refers to the significance of group at X=0, which is not useful. Many people have incorrectly concluded that the groups do or do not differ significantly in accordance with an incorrect understanding of meaning of the p-value for group in models with the interaction term [6].

Testing whether $\hat{Y}$ differs between two groups at a particular value of a covariate $X = X_0$ when the slopes of the regression lines differ and the unexplained variance about the regression lines differ.

$\hat{Y}_k = a_k + b_k X$ is the equation for the regression of Y on X for the k-th group (k=1,2),
a_k is the Y-intercept for group k, and b_k is the slope coefficient for group k.

$\hat{Y}_1(X_0) = \hat{Y}_1\ evaluated\ at\ X = X_0$

$\hat{Y}_2(X_0) = \hat{Y}_2\ evaluated\ at\ X = X_0$

$\hat{\sigma}_1^2 = unexplained\ variance\ estimate\ (mean\ square\ residual)\ for\ regression\ fit\ 1$

$\hat{\sigma}_2^2 = mean\ square\ residual\ for\ regression\ fit\ 2$

$X_0 = value\ of\ X\ at\ which\ \hat{Y}_1\ and\ \hat{Y}_2\ are\ co\square pared$

The t-statistic for the comparison is calculated as:

$t = \frac{\hat{Y}_1(X_0) - \hat{Y}_2(X_0)}{SE_{\hat{Y}_1(X_0) - \hat{Y}_2(X_0)}}$ (Zar, 1996, p.361) [1]

If $\hat{\sigma}_1^2 \approx \hat{\sigma}_2^2$ such that one can pool $\hat{\sigma}^2$, then degrees of freedom are calculated as $n_1 + n_2 - 4$; however, if pooling is not justified, then degrees of freedom may need to be adjusted using an appropriate method, e.g., the Welch-Satterthwaite equation (shown below).

If the unexplained variance about each regression line does differ, then the derivation of the standard error for the t statistic is as follows:

$VAR\left(\hat{Y}_1(X_0)\right) = \hat{\sigma}_1^2\left(\frac{1}{n_1} + \frac{(X_0 - \bar{X}_1)^2}{\sum_{i=1}^{n1}(X_i - \bar{X}_1)^2}\right)$ (Zar, 1996, p.333) [1].

$VAR\left(\hat{Y}_2(X_0)\right) = \hat{\sigma}_2^2\left(\frac{1}{n_2} + \frac{(X_0 - \bar{X}_2)^2}{\sum_{j=1}^{n2}(X_j - \bar{X}_2)^2}\right)$

Because VAR(A ± B) = VAR(A) + VAR(B),

$$VAR\left(\hat{Y}_1(X_0) - \hat{Y}_2(X_0)\right) = \hat{\sigma}_1^2\left(\frac{1}{n_1} + \frac{(X_0 - \bar{X}_1)^2}{\sum_{i=1}^{n1}(X_i - \bar{X}_1)^2}\right) + \hat{\sigma}_2^2\left(\frac{1}{n_2} + \frac{(X_0 - \bar{X}_2)^2}{\sum_{j=1}^{n2}(X_j - \bar{X}_2)^2}\right)$$

Therefore,

$$SE\left(\hat{Y}_1(X_0) - \hat{Y}_2(X_0)\right) = \sqrt{\hat{\sigma}_1^2\left(\frac{1}{n_1} + \frac{(X_0 - \bar{X}_1)^2}{\sum_{i=1}^{n1}(X_i - \bar{X}_1)^2}\right) + \hat{\sigma}_2^2\left(\frac{1}{n_2} + \frac{(X_0 - \bar{X}_2)^2}{\sum_{j=1}^{n2}(X_j - \bar{X}_2)^2}\right)}$$

(We note that if $\hat{\sigma}_1^2 = \hat{\sigma}_1^2$, this SE simplifies to the formula shown on p.361 of Zar (1996) [1] for comparing points on two regression lines.)

Welch-Satterthwaite equation for degrees of freedom.

For a standard two-group t-test involving disparate group variances, the degrees of freedom (d.f.) are calculated as follows:

$$d.f. = \frac{\left(\frac{\hat{\sigma}_1^2}{n_1} + \frac{\hat{\sigma}_2^2}{n_2}\right)^2}{\frac{\left(\frac{\hat{\sigma}_1^2}{n_1}\right)^2}{(n_1 - 1)} + \frac{\left(\frac{\hat{\sigma}_2^2}{n_2}\right)^2}{(n_2 - 1)}};$$

For comparing points on two regression lines that have different unexplained variance estimates, $(n_k - 1)$ should be $(n_k - 2)$.

Moreover, in reviewing Zar's comments on the considerable robustness of t-tests to the equal variance assumption when two-tailed tests are employed (Zar, 1996, p.128) [1], it is not clear that one needs to adjust degrees of freedom.

Acknowledgements: This is a companion to the energy expenditure modules written by Karl J. Kaiyala[1], Charles F. Spiekerman[2], Robert H. Podolsky[3] and Owen P. McGuinness[4]. The MMPC ANCOVA web tool was created by Richard McIndoe[5].

[1]School of Dentistry, Department of Oral Health Sciences, University of Washington, Seattle, WA; [2]School of Public Health and Department of Biostatistics, University of Washington, Seattle, WA; [3]Department of Family Medicine and Public Health Sciences, Wayne State University School of Medicine, Detroit MI; [4]Department of Molecular Physiology and Biophysics, Vanderbilt University School of Medicine, Nashville, TN; [5]Coordinating and Bioinformatics Unit, Georgia Regents University, Augusta, GA.

The authors were supported by funding from the National Institute of Health (KJK: U24DK076169, P30DK035816; CFS: U24DK076169; RHP: R03-DA033453; OPM: DK59637, DK20593, DK078188, DK043748; RM: U24DK076169) and the NIDDK Mouse Metabolic Phenotyping Center Program.

References

1. Zar J Biostatistical analysis. Third Ed. Upper Saddle River, New Jersey. Prentice Hall, 1996.
2. Tschop MH, Speakman JR, Arch JR, Auwerx J, Bruning JC, et al. (2012) A guide to analysis of mouse energy metabolism. Nat Methods 9: 57-63.
3. Speakman JR, Fletcher Q, Vaanholt L (2013) The '39 steps': an algorithm for performing statistical analysis of data on energy intake and expenditure. Dis Model Mech 6: 293-301.
4. Arch JR, Hislop D, Wang SJ, Speakman JR (2006) Some mathematical and technical issues in the measurement and interpretation of open-circuit indirect calorimetry in small animals. Int J Obes (Lond) 30: 1322-1331.
5. White CR (2003) Allometric analysis beyond heterogeneous regression slopes: use of the Johnson-Neyman technique in comparative biology. Physiol Biochem Zool 76: 135-140.
6. Engqvist L (2005) The mistreatment of covariate interaction terms in linear model analyses of behavioural and evolutionary ecology studies. Animal Behaviour 70: 967-971.