ANALYSIS OF COVARIANCE

Analysis of covariance (ANCOVA)

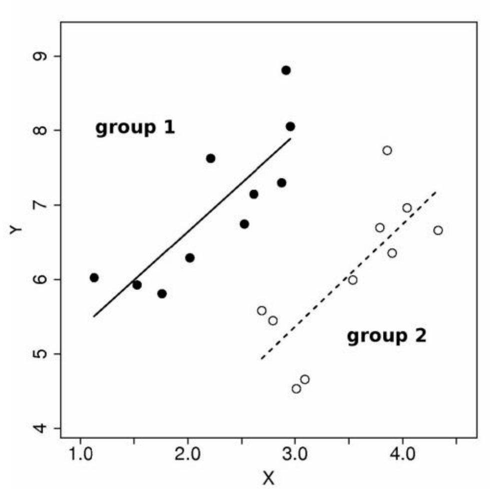
ANCOVA combines a qualitative or factor independent variable (“factor”) with a quantitative/continuous independent variable (“covariate”), to reduce the residual variance.

Note: Kuehl’s discussion of ANCOVA, on pp. 550-564, is somewhat different from here. Focus on the concepts, and on the specific issues we cover in lecture.

Analysis of covariance, or ANCOVA, combines a qualitative or factor independent variable with a quantitative or continuous independent variable to reduce the residual variance.

Kuehl’s discussion of ANCOVA is somewhat different from here. Focus on the concepts, and on the specific issues we cover in lecture.

The basic idea



The response Y is influenced by a qualitative/factor independent variable (the two groups) and a quantitative/continuous independent variable (covariate X)

This figure illustrates the basic idea behind an ANCOVA model.

In this figure, we see that the response variable Y is influenced by a factor independent variable, group, and a quantitative independent variable, X.

The basic idea (continued)

X is a covariate, or concomitant variable, that influences the response of interest. By "controlling for" X in our analysis of group effects, we

1. increase the precision of the comparison; and
2. avoid the effects of confounding (when X differs between groups), especially in observational studies.

If we didn't adjust for X here, we would conclude there is little or no difference between groups. (Imagine losing the X coordinate and collapsing all the points onto the the y axis.)

Analysis of covariance fit separate regression lines to different groups and allow us to compare the Y values from different groups at a common X value.

(Here it assumes the lines are parallel. This assumption must be tested!)

X is a covariate, or concomitant variable, that influences the response of interest. By "controlling for" X in our analysis of group effects, we increase the precision of the comparison; and avoid the effects of confounding when X differs between groups, especially in observational studies.

Recall that a variable is a confounding factor if it correlated with both the response and a predictor in the model.

If we didn't adjust for X here, we would conclude there is little or no difference between groups. Just imagine losing the X coordinate and collapsing all the points onto the the y axis.

Analysis of covariance fit separate regression lines to different groups and allow us to compare the Y values from different groups at a common X value.

(Here it assumes the lines are parallel. This assumption must be tested!)

The single-factor covariance model

One way to write the basic model is

$$y_{ij} = \mu + \tau_i + \beta(x_{ij} - \overline{x}_{..}) + e_{ij}, i = 1, \dots, t, j = 1, \dots, r.$$

where

- $\mu =$ the overall mean,
- $\tau_i =$ effect of ith treatment,
- $x_{ij} = ij$th covariate value,
- $\overline{x}_{..} =$ overall mean of covariate values, and
- β is a regression coefficient (indicating the dependence of y_{ij} on x_{ij}).
- e_{ij} are assumed i.i.d. $N(0, \sigma^2)$.

Comparing this model to the one-way ANOVA model, we see that

- the random part of the model has the same iid normal assumption, and
- the fixed part now has one more term, $\beta(x_{ij} - \overline{x}_{..})$, for the effect of the continuous covariate X.

Here we give the model equation for a single-factor covariance model.

Comparing this model to the one-way ANOVA model, we see that the random part of the model has the same i.i.d. normal assumption, and the fixed part now has one more term for the effect of the continuous covariate X.

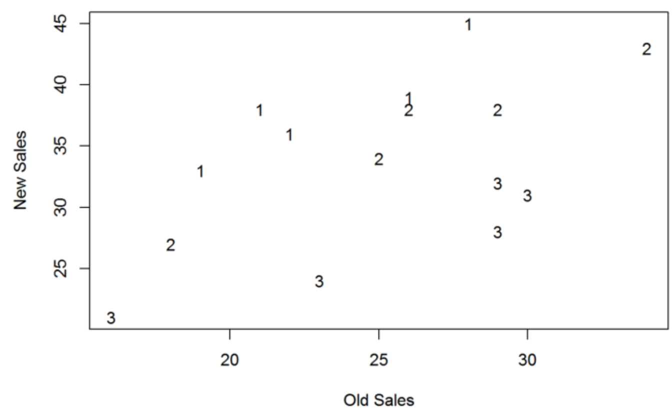
Example (Neter, Wasserman and Kutner, p. 854).

Study the effect of three promotion methods on sales of crackers.

Three methods of promotion were randomly assigned to 15 stores (5 stores per method). The responses are sales of crackers in 15 stores after the promotion. The covariate is the sales of crackers in some period before the promotion.

Now let’s look at an example where a researcher studies the effect of three promotion methods on sales of crackers. Three methods of promotion were randomly assigned to 15 stores (5 stores per method). The responses are sales of crackers in 15 stores after the promotion. The covariate is the sales of crackers in some period before the promotion.

Example (continued), look at the data



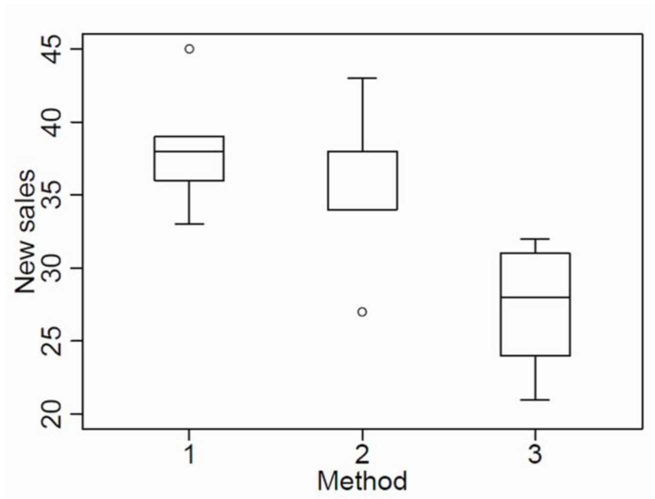
Scatter plot, new sales versus old sales

In real data analysis, it’s always a good idea to look at the data first.

In this Figure, we plot the new sales volume against the old sales volume, each point in the figure corresponds to a store. The labels 1, 2, and 3 mark the three promotional methods.

We see that within each group receiving the same promotion method, the new sales is increasing with old sales.

Example (continued), look at the data



Box plot, the new sales by promotion method

If we ignore the old sales volume, we can use side-by-side boxplots to summarize the relation between new sales and the promotion method.

From the boxplots, we see that new sales volume after using promotion method 3 seems significantly lower than using the other two methods.

Example (continued)

From the graphs,

1. It appears that new sales (1) > (2) > (3), but there's lots of overlap.
2. New sales are strongly related to old sales.
3. Old sales differ among groups, so this is a potential confounding factor.

A quick recap of what we see from the graphs:

It appears that new sales (1) > (2) > (3), but there's lots of overlap.

New sales are strongly related to old sales.

Old sales differ among groups, so this is a potential confounding factor.

Fitting regression models

Let y = new sales, x_3 = old sales, and let

$$x_1 = \begin{cases} 1 & \text{if method 1} \\ 0 & \text{otherwise} \end{cases}, \quad x_2 = \begin{cases} 1 & \text{if method 2} \\ 0 & \text{otherwise} \end{cases}.$$

x_1 and x_2 are two indicator variables (needed for coding a factor variable with three levels) and x_3 is a quantitative/continuous variable encoding the covariate, old sales.

We will fit a series of regression models to the data set, from simple to complex, to see which one is "the best": if a model is too simple (few parameters) it may not be adequate to capture enough variation in the data; if the model is too complex, it may "overfit".

The trick is to find the right balance: usually, we want to find the simplest model that adequately capture all variation in the data.

(See cracker.html for more details on the R code and output.)

Let y be the response variable, new sales; (x 1) and (x 2) be indicator variables for coding the promotional methods; and (x 3) be the continuous covariate, old sales.

Recall that to code a factor variable with three levels, we need two indicator variables. In this example, promotional method 3 is used as the reference group, it corresponds to when the indicator variables (x 1) and (x 2) are both 0.

We will fit a series of regression models to the data set, from simple to complex, to see which one is "the best": if a model is too simple (few parameters) it may not be adequate to capture enough variation in the data; if the model is too complex, it may "overfit".

The trick is to find the right balance: usually, we want to find the simplest model that adequately capture all variation in the data.

On the next few pages, we will go over some of the models we fit. See cracker.html for more details on the R code and output.

Again, I want to remind you that learning to use R to analyze data is an important part of this class.

MODEL 00, no explanatory variables

Model equation:

$$y = \beta_0 + e.$$

where e iid $N(0, \sigma^2)$. (For simplicity we have suppressed the subscript indexing the experimental units.)

Under this model, all responses are iid normally distributed with mean β_0 and a constant variance σ^2. So this model assumes there is not difference in sales volume among the three groups and β_0 is the grand mean of the new sales.

Regression equation (fitted model):

$$\hat{y} = \hat{\beta}_0 = 33.8,$$

for all three promotion methods.

$$SSE = \text{Total SS} = 646.4 \quad (14\ d.f.)$$

We start from the simplest model where the fixed part of the model has only one parameter, (beta 0).

The random term is assumed to be iid normal with mean 0 and constant variance.

For simplicity we have suppressed the subscript indexing the experimental units in the model equation.

Under this model, all responses are iid normally distributed with mean (beta 0) and a constant variance. So this model assumes there is not difference in sales volume among the three groups and (beta 0) is the grand mean of the new sales.

The estimated regression equation is (y hat equals beta 0 hat equals 33.8) for all stores.

The SSE from this model is 646.4 with 14 degrees of freedom.

The SSE from this model is the same as the total sum of squares for this data set since we simply fit a grant mean to the data set.

MODEL 0, method effect only

Model equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + e.$$

Regression equation (fitted model):

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 = 27.2 + 11.0 x_1 + 8.8 x_2.$$

$$\text{Method } 1, x_1 = 1, x_2 = 0: \hat{y} = 27.2 + 11 = 38.2$$

$$\text{Method } 2, x_1 = 0, x_2 = 1: \hat{y} = 27.2 + 8.8 = 36.0$$

$$\text{Method } 3, x_1 = 0, x_2 = 0: \hat{y} = 27.2$$

$$SSE = 307.6 \quad (12\ d.f.)$$

Note that if we force β_1 and β_2 to be 0, then this model will reduce to MODEL 00. So in this sense we say that MODEL 00 is nested in MODEL 0.

Next up, we fit a model using only the two indicator variables for promotional methods as the predicting variables.

The random part of the model has the same iid normal assumption as before.

In the fixed part of the model, the three regression coefficients (beta 0), (beta 1) and (beta 2) together with the two indicator variables, will give us the mean response value for the three treatment groups.

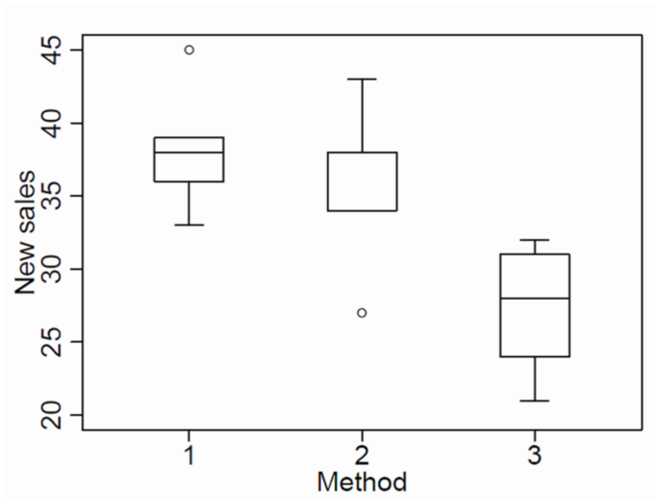
This model is essentially the one-way ANOVA model. If you have studied the R code in script1, you will remember that when fitting a one-way ANOVA model, we create indicator variables to code the factor levels—that is what the R function “factor” does.

The fitted model is give by (y hat equal 27.2 + 11.0 x_1 + 8.8 x_2) Under this model, the group means corresponding to the three promotional methods are 38.2, 36.0 and 27.2. Make sure you understand how we can get the three estimated group means from the this fitted regression model.

The SSE from this model is 307.6 with 12 degrees of freedom.

Note that if we force (beta 1) and (beta 2) to be 0, then this model will reduce to

MODEL 00. So in this sense we say that MODEL 00 is nested in MODEL 0.



MODEL 0

In MODEL 0, we fit three separate means to the three groups. We can use the side-by-side boxplots to help us think about the fitted model.

But remember that the boxplots shows the median and quantiles, not means and standard deviations, of each group.

MODEL 1, old sales effect only

Model equation:

$$y = \beta_0 + \beta_3 x_3 + e.$$

Regression equation (fitted model):

$$\hat{y} = 15.61 + 0.73 x_3.$$
$$SSE = 455.7 \quad (13\ d.f.)$$

Note that if we force β_3 to be 0, then this model is reduced to MODEL 00. So MODEL 00 is nested in this model too.

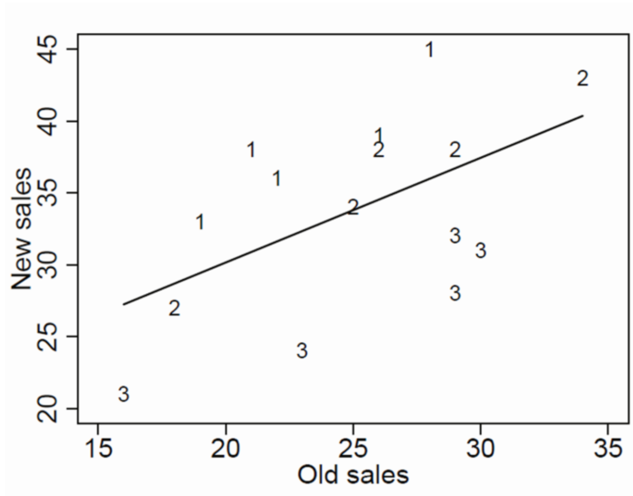
In MODEL 1, we consider only the effect of the continuous covariate (x 3), the old sales volume.

This is a simple regression model. The estimated regression equation is

(y hat) equals 15.61 + 0.73 (x 3)

The SSE from this model is 455.7 with 13 degrees of freedom.

Note that if we force (beta 3) to be 0, then this model is reduced to MODEL 00. So MODEL 00 is nested in this model too.



MODEL 1

MODEL 1 corresponds to fitting a single regression line into the data set. It ignores the effects of different promotional methods.

MODEL 2, method + old sales

Model equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + e.$$

Regression equation (fitted model):

$$\hat{y} = 4.38 + 12.98x_1 + 7.90x_2 + 0.90x_3.$$

$$\text{Method 1: } \hat{y} = 17.36 + 0.90x_3$$

$$\text{Method 2: } \hat{y} = 12.28 + 0.90x_3$$

$$\text{Method 3: } \hat{y} = 4.38 + 0.90x_3$$

$$SSE = 38.57 \quad (11\ d.f.)$$

- This model fits a separate regression line to each treatment group (promotion method). The slope for x_3 is the same in all three groups: so the three regression lines are parallel with different intercepts.
- The intercepts are not too meaningful here since they correspond to $x_3 = 0$, which means the old sales volume is 0. That is not a useful reference point. If we want to compare the three treatment groups, we can compare them at a more meaningful common x_3 value, e.g., $x_3 = 25$.

In model 2, we include both the treatment-group effects coded by (x 1) and (x 2) and the continuous covariate (x 3) in the model equation.

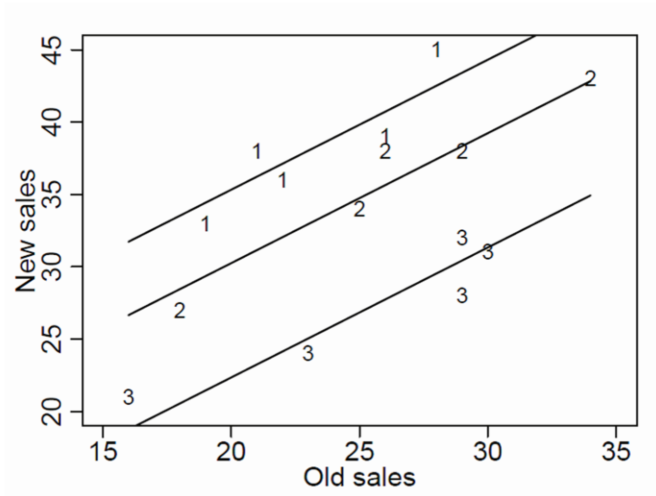
The fitted model is given by

(y hat) equals 4.38 + 12.98 x_1 + 7.90 x_2 + 0.90 x_3

This model fits a separate regression line to each treatment group.

Note that the slope for (x 3) is the same for the three regression lines: so the three lines are parallel with different intercepts.

Note that the intercepts are not too meaningful here since they correspond to (x 3 equals 0), which means the old sales volume is 0. That is not a useful reference point. If we want to compare the three treatment groups, we can compare them at a more meaningful common (x 3) value, e.g., (x 3 equals 25).



MODEL 2

MODEL 2 corresponds to fitting three parallel regression lines to the three treatment groups.

MODEL 3, method + old sales + interaction

Model equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 (x_1 x_3) + \beta_5 (x_2 x_3) + e.$$

Regression equation (fitted model):

$$\hat{y} = 8.56 + 4.32x_1 + 1.27x_2 + 0.73x_3 + 0.36(x_1 x_3) + 0.26(x_2 x_3)$$

$$\text{Method 1: } \hat{y} = 8.56 + 4.32 + (0.73 + 0.36)x_3 = 12.88 + 1.09x_3$$

$$\text{Method 2: } \hat{y} = 8.56 + 0.36 + (0.73 + 0.26)x_3 = 9.83 + 0.99x_3$$

$$\text{Method 3: } \hat{y} = 8.56 + 0.73x_3$$

$$SSE = 31.52 \quad (9\ d.f.)$$

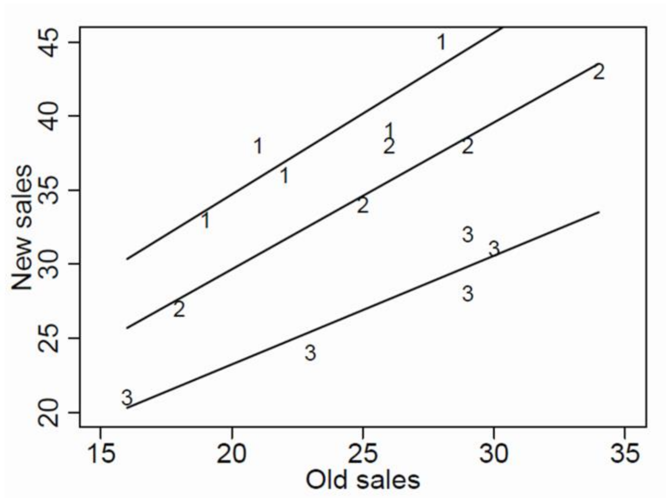
MODEL 3 fits separate regression lines to the three groups, with each regression line having a different slope.

Our next model, MODEL 3, is the most complex of the batch. It includes the method effect, the old sales effect, and their interactions.

In the model equation, the interactions are represented by the terms (x 1 times x 3) and (x 2 times x 3). These interaction terms simply mean that in different treatment groups, the slopes for (x 3) can be different.

So essentially, MODEL 3 fits separate regression lines to the three groups, with each regression line having a different slope.

MODEL 3 include all previous models as special cases. The SSE from this model is 31.52 with 9 degrees of freedom.



MODEL 3

This figure shows the three separate regression lines fitted to the three groups under MODEL 3, with each regression line having a different slope.

Analysis (one possible approach)

Start with the most complicated model (MODEL 3), and test

$$H_0: \beta_4 = \beta_5 = 0, \text{equality of slopes (or no interaction).}$$

Under H_0, MODEL 3 reduces to MODEL 2.

The general formula for a F-test statistic comparing two **nested** regression model is

$$F_0 = \frac{SSE_r - SSE_f}{df_r - df_f} \div \frac{SSE_f}{df_f} = \frac{38.57 - 31.52}{11 - 9} \div \frac{31.52}{9} = 1.01$$

Comparing this F-test statistic value to a F distribution with d.f. $(2,9)$ gives a p-value of 0.4.

```
pf(1.01, 2, 9, lower.tail = FALSE)
## [1] 0.4020445
```

So we do not reject H_0, and we adopt MODEL 2 as more parsimonious.

After fitting this series of models to the data set, the next question is, naturally, which model is the best.

One possible approach to answering this question is to start with the most complicated model (MODEL 3), and ask whether a simpler model, e.g., MODEL 2, can fit the data almost as well with fewer parameters.

In other words, we can compare two nested models using a statistical test. Note that the more complex model will always explain more data variation and thus has a smaller SSE. The question is whether the reduction in SSE is worth the additional parameters added to the model.

To answer this question, we can use a F test to test the null hypothesis that the additional parameters in the more complex are all 0. Under the null hypothesis, the more complex model is reduced to the simpler model.

The general form of the F test is given in this formula: In the numerator, we have the reduction in SSE divided by the reduction in the number of d.f. for error estimation. The reduction in the number of d.f. for the SSE is the same as the increase in the number of parameters in the fixed part of the regression model.

In the denominator, we have the SSE from the full model divided by its corresponding number of d.f.

If the null hypothesis is true, this F statistic will have a F distribution.

In this example, the F-statistic value is 1.01. Under the null, it has a F-distribution with 2 and 9 numerator and denominator degrees of freedom. The p-value of the F-test is thus 0.4.

So there is no evidence that MODEL 3 is significantly better than MODEL 2, and we can adopt MODEL 2 as more parsimonious.

Analysis (continued)

"Analysis of covariance"—is there an effect of old sales, after we have accounted the effects for method (i.e., after we have included the indicator variables for method)?

To answer this question, we compare MODEL 2 to MODEL 0, i.e., test

$$H_0: \beta_3 = 0.$$

We use the same general formula for a F-test comparing two nested regression models, except that now the full model is MODEL 2 and the reduced model is MODEL 0:

$$F_0 = \frac{SSE_r - SSE_f}{df_r - df_f} \div \frac{SSE_f}{df_f} = \frac{307.6 - 38.57}{12 - 11} \div \frac{38.57}{11} = 76.72$$

Comparing F_0 to a F distribution with d.f. $(1,11)$ gives a p-value of 2.7×10^{-6}.

We have strong evidence that β_3 is not 0 and the full model is a better fit.

Next, we can ask is there an effect of old sales, after we have accounted the effects for method—i.e., after we have included the indicator variables for method)?

To answer this question, we compare MODEL 2 to MODEL 0, i.e., test the null hypothesis (beta 3) equals 0 in MODEL 2.

We use the same general formula for a F-test comparing two nested regression models, except that now the full model is MODEL 2 and the reduced model is MODEL 0.

Comparing the F value to a F distribution with degrees of freedom (1, 11) gives a p-value of 2.7 times 10^{-6}.

We have strong evidence that (beta_3) is not 0 and the full model is a better fit.

Analysis (continued)

Is there an effect of method, after adjusting for old sales?

To answer this question, we compare MODEL 2 to MODEL 1, i.e., test

$$H_0: \beta_1 = \beta_2 = 0$$

Again, we use the same general formula for a F-test comparing two **nested** regression models. The full model is MODEL 2 and the reduced model is MODEL 1:

$$F_0 = \frac{SSE_r - SSE_f}{df_r - df_f} \div \frac{SSE_f}{df_f} = \frac{455.7 - 38.57}{13 - 11} \div \frac{38.57}{11} = 59.48$$

Comparing F_0 to a F distribution with d.f. $(2,11)$ gives a p-value of 1.26×10^{-6}.

So there is a difference among different promotion methods even after we have accounted for old sales effect, and MODEL 2 is better.

Overall, MODEL 2 is our “best” model of the data.

We can also ask, is there an effect of method, after adjusting for old sales (using common slope)?

To answer this question, we compare Model 2 to Model 1, and test the null hypothesis that (beta 1) and (beta 2) are both 0 in Model 2.

Again, we use the same general formula for a F-test comparing two nested regression models, except that now the full model is Model 2 and the reduced model is Model 1:

Comparing the F-statistic value, 59.48, to an F distribution with d.f. (2, 11) gives a p-value of 1.26 times 10^{-6}.

So there is a difference among different promotion methods even after we have accounted for old sales effect, and MODEL 2 is better.

Overall, Model 2 is our “best” model of the data.

Analysis (continued)

Model	$\hat{y}$ (meth. 1) - $\hat{y}$ (meth. 2)	$\hat{y}$ (meth. 2) - $\hat{y}$ (meth. 3)	Extra-SS tests for the method effect		
			Comparison	F-statistic	P-value
0	2.2	8.8	0 vs. 00	6.61	0.012
2	5.1	7.9	2 vs. 1	59.5	< 0.0001

Adjustment for old sales has given us different, and more precise, estimates of the method effect.

This table illustrate that the contrasts between different promotion methods become more significant under MODEL 2, where we included the old-sales effect.

One way to see this is that the F-test p-value for comparing MODEL 2 to MODEL 1 is much more significant than the F-test p-value for comparing MODEL 0 to MODEL 00.

If we assume that the promotion methods have no effect, then MODEL 2 will reduce to MODEL 1 and MODEL 0 will reduce to MODEL 00. So the two F-tests in the table are both testing the effect of promotion methods. The difference is the starting full model used: MODEL 2 includes the old-sales effect, MODEL 0 does not.

Adjusted treatment means (Kuehl, p. 556-667)

What is a good estimate of the mean response in the ith treatment?

$\overline{y}_{i\cdot}$ will vary, depending on what x values occur in the ith group (where x is the covariate).

It seems reasonable to compare all methods at a common x value. One option is to adjust the mean to the value expected at $\overline{x}_{\cdot\cdot}$ (the mean of the x values):

$$\overline{y}_{i\cdot}(\text{adj}) = \overline{y}_{i\cdot} - \hat{\beta}(\overline{x}_{i\cdot} - \overline{x}_{\cdot\cdot}),$$

where $\hat{\beta}$ is the common slope estimate.

(*Note:* The raw group means of the responses correspond to the response at the group means of the x values.)

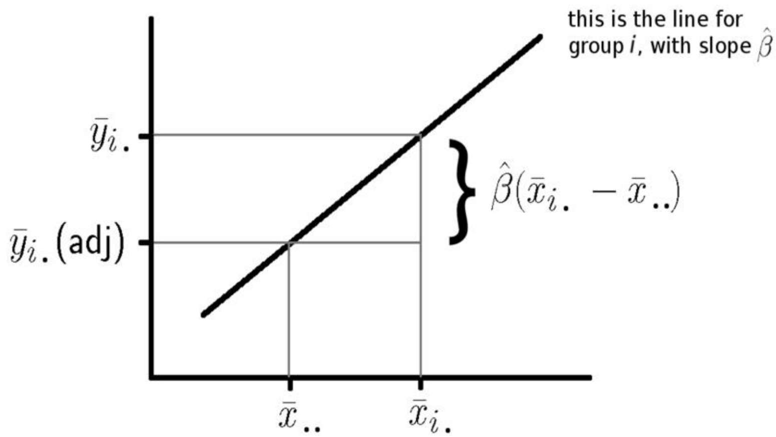
Under the ANCOVA model, even within each group, the mean response will still vary with the covariate.

One may ask, then, what is a good estimate of the mean response in the i-th treatment?

It seems reasonable to compare all methods at a common x value. One option is to adjust the group mean to the value expected at the mean of all x values.

Illustration of adjusted treatment mean

$$\bar{y}_{i\cdot}(\text{adj}) = \bar{y}_{i\cdot} - \hat{\beta}(\bar{x}_{i\cdot} - \bar{x}_{\cdot\cdot})$$



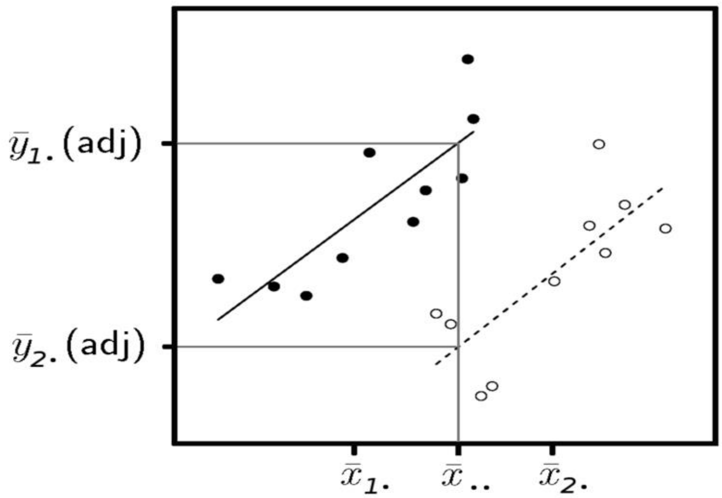
This plot illustrates the difference between the raw group mean (y bar i dot) and the adjusted treatment mean (y bar i dot adjust)

Under the fitted ANCOVA model, the mean response in group i corresponds to the predicted response at the group mean of the x values.

In general, the group mean of x, (x bar i), will be different for different treatment groups.

The adjusted treatment mean for group i is the predicted group i response at the overall mean of the x values.

Illustration of adjusted treatment mean



Thus figure shows that by using the adjusted treatment mean, we are comparing the predicted responses from each group at a common x value.

Adjusted treatment means for the cracker data

For the cracker data (using $\bar{x}_{\bullet\bullet} = 25$, and $\hat{\beta}_3 = 0.90$, the common slope estimate from Model 2):

Method	$\bar{y}_{i\bullet}$	$\bar{x}_{i\bullet}$	$\bar{y}_{i\bullet}$(adj)
1	38.2	23.2	39.8
2	36.0	26.4	34.7
3	27.2	25.4	26.8

Kuehl, p. 557, gives expressions for the standard errors of adjusted treatment means.

This table shows the raw group means and adjusted treatment means for the cracker data under the fitted MODEL 2.

We see that the there is more separation between Method 1 and Method 2 among the adjusted treatment means.

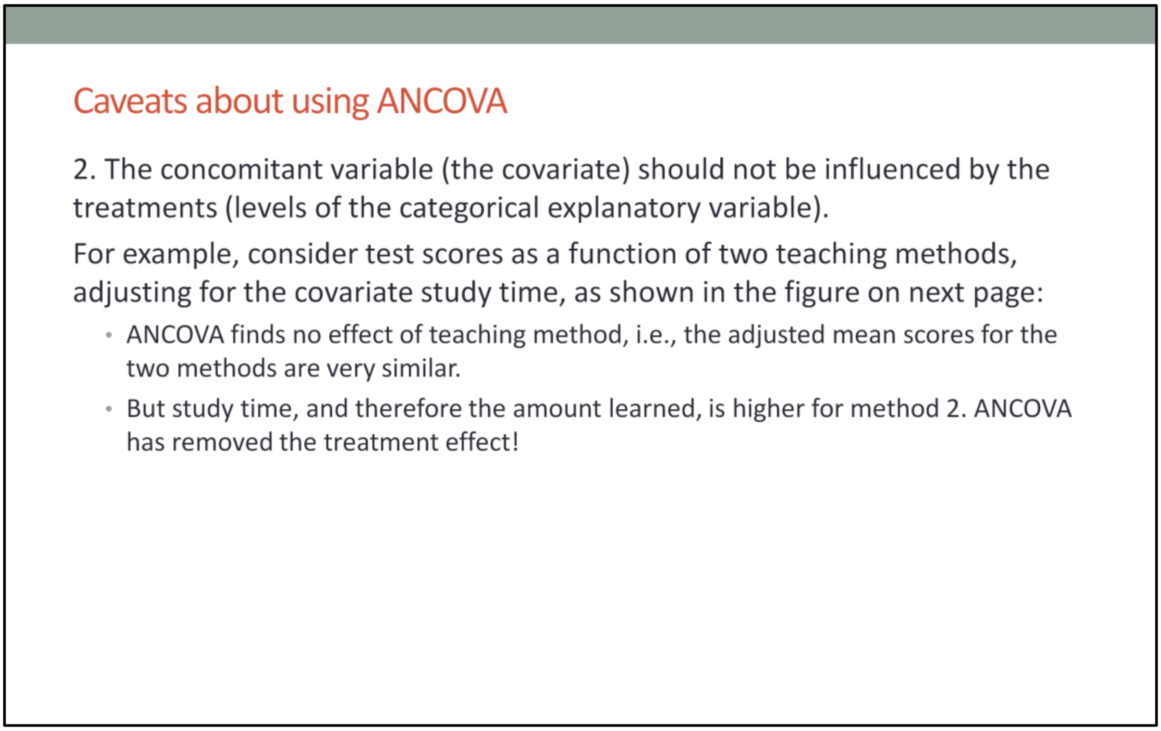
Caveats about using ANCOVA

1. Watch out for potential extrapolation.

It’s risky to compare regression lines with little or no overlap in their ranges of x-values. Computation and comparison of adjusted means would likely involve **extrapolation** of a line(s) beyond the range of x’s upon which it is based.

There are some potential caveats when using the ANCOVA model.

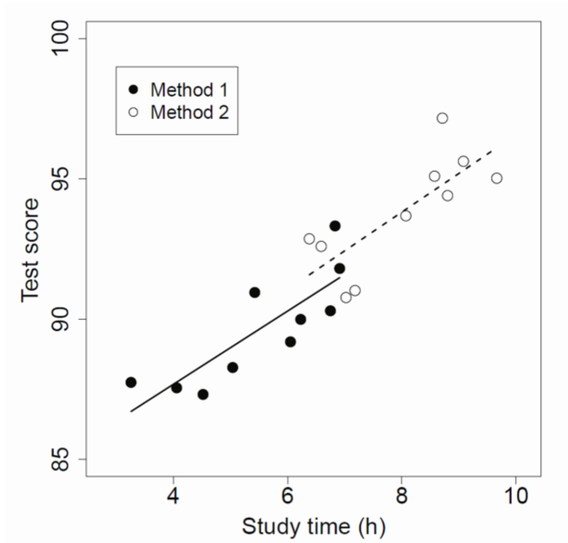
First, it’s risky to compare regression lines with little or no overlap in their ranges of x-values. Computation and comparison of adjusted means would likely involve extrapolation of a line(s) beyond the range of x’s upon which it is based.



Second, the concomitant variable, i.e., the covariate, should not be influenced by the treatments (levels of the categorical explanatory variable).

For example, consider test scores as a function of two teaching methods, adjusting for the covariate study time, as shown in the figure on next page:

> ANCOVA finds no effect of teaching method, i.e., the adjusted mean scores for the two methods are very similar.
>
> But study time, and therefore the amount learned, is higher for method 2. ANCOVA has removed the treatment effect!



In this example, the Study Time is a potential effect of the teaching method. By including it as a covariate in the ANCOVA model, we will remove the effect of the teaching method.

Summary

1. The ANCOVA model allows us to include an additional continuous covariate into the one-way ANOVA model—which can potentially explain part of the variation among the residuals and improve the precision of the treatment comparisons.

2. We can use a F-test to compare two nested regression models. There is a general formula for such a F test that involves the SSE from the full and the reduced models and their numbers of d.f..

3. Be careful not to extrapolate or include a potential treatment effect as a covariate when using the ANCOVA model.

Summary

The ANCOVA model allows us to include an additional continuous covariate into the one-way ANOVA model—which can potentially explain part of the variation among the residuals and improve the precision of the treatment comparisons.

We can use a F-test to compare two nested regression models. There is a general formula for such a F test that involves the SSE from the full and the reduced models and their numbers of degrees of freedom.

Be careful not to extrapolate or include a potential treatment effect as a covariate when using the ANCOVA model.