CONJOINT

A Mu Sigma University Publication, February 2016

Mu Sigma

In This Issue...

Abstract

In this article, we discuss the technique of Analysis of Covariance, often abbreviated as ANCOVA, which is an extension of the technique of Analysis of Variance and is used to control and eliminate the effect of a quantitative covariate in making comparisons.

Analysis of Covariance (ANCOVA)

- By T. Krishnan

Comparison of Means in the Presence of a Covariate

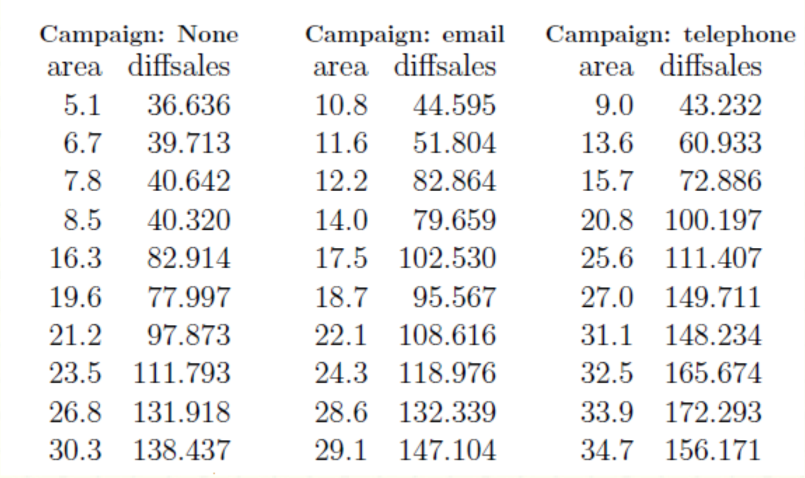
Suppose you are comparing the effects of the following three types of campaigns in a retail chain: none (control, no campaign), email, and telephone. Observational units are stores. Campaign effect is measured by a metric (difference in sales between campaign month and previous month in ’0000 $’s, denoted “diffsales”). Suppose there are 10 stores for each campaign. A standard analysis will involve a comparison of the means of the three groups by Analysis of Variance, testing the hypothesis of equality of the three means and presenting estimates of the differences in the means with their standard errors.

The study, the data from which is presented here, was designed by allotting the stores to the three campaigns at random. Sales, and hence the metric used depend on the size of the stores. In this study the sizes of stores for different campaigns were not completely controlled and hence may vitiate the comparisons of the campaign effects. The size of the stores here is represented by the store area. The technique of Analysis of Covariance (ANCOVA) is intended to offset such factors and make the comparisons more valid and effective. Such factors, as the area in this case, are called **covariates** or **concomitant variables**.

Example: Comparison of Different Campaign Effects

Consider the following data where the Store Area (area in ’000 sq. ft.) is also given.

Campaign: A: No campaign (Control); B: email campaign; C: Telephone campaign



Campaign: None		Campaign: email		Campaign: telephone	
area	diffsales	area	diffsales	area	diffsales
5.1	36.636	10.8	44.595	9.0	43.232
6.7	39.713	11.6	51.804	13.6	60.933
7.8	40.642	12.2	82.864	15.7	72.886
8.5	40.320	14.0	79.659	20.8	100.197
16.3	82.914	17.5	102.530	25.6	111.407
19.6	77.997	18.7	95.567	27.0	149.711
21.2	97.873	22.1	108.616	31.1	148.234
23.5	111.793	24.3	118.976	32.5	165.674
26.8	131.918	28.6	132.339	33.9	172.293
30.3	138.437	29.1	147.104	34.7	156.171

In This Issue...

The means and standard deviations of area and diffsales are given for each campaign below:

Campaign	area		diffsales	
	Mean	S.D.	Mean	S.D.
none	16.58	9.09	79.82	39.56
email	18.89	6.89	96.41	32.82
telephone	24.39	9.17	118.07	46.95

The means (campaign effects) are substantially different. The areas of stores for the campaigns are also substantially different and this may impact and vitiate the comparisons.

If you ignore the difference in areas, you might carry out a standard Analysis of Variance (ANOVA) with the following results, where the reference campaign for comparisons is “none”.

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    79.82      12.71   6.281 1.01e-06 ***
email          16.58      17.97   0.922   0.3644
telephone      38.25      17.97   2.128   0.0426 *
---
Signif. codes:  0*** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 40.19 on 27 degrees of freedom
Multiple R-squared:  0.1444,Adjusted R-squared:  0.08097
F-statistic: 2.278 on 2 and 27 DF,  p-value: 0.1219

Analysis of Variance Table

Response: diffsales
          Df Sum Sq    Mean Sq    F value   Pr(>F)
campaign   2  7358     3679.1    2.2776    0.1219
Residuals 27 43615     1615.4
```

The email effect is not significantly different from the no campaign effect whereas the telephone campaign is somewhat significantly different from the no campaign effect. It also appears that the overall campaigns effects are not much different.

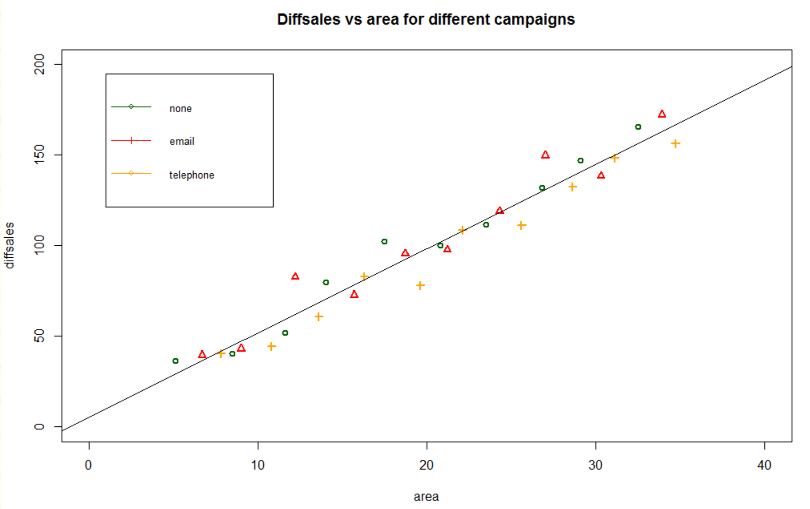
Covariate Effect

In the design of this study, although the stores for the campaigns have been allotted at random, no specific efforts have been made to control for store properties like the store area, which may influence the sales. This happens to be rather different for the different campaign stores, not completely taken care of by the random allocation. Of course, information on store area will be available when the study is being designed and it would be possible to take this into account while developing the design.

In This Issue. . .

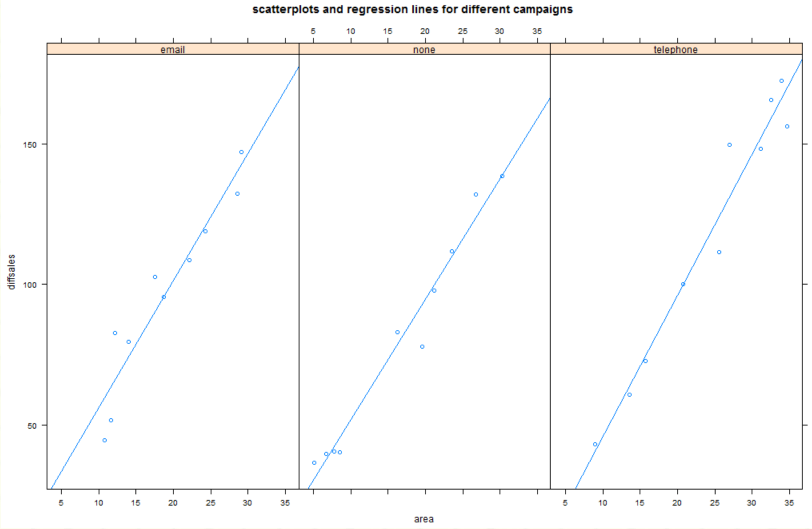
However, we consider the situation where this information is used at the time of analysis. This variable, the area, is called a covariate or a concomitant variable which is evidently correlated with the sales metric and influences the comparison of means. Thus there is a need to adjust for this campaign means for the differences in the covariate values of stores under different campaigns. This is precisely what the Analysis of Covariance (ANCOVA) is intended to do.

Let us study the relationship between diffsales and area. The following are scatterplots of diffsales on area for each of the campaigns and together.



We notice that diffsales is highly correlated with area (correlation = 0.975) overall and within each campaign. There is sufficient evidence from these graphs that the regression relationship of diffsales on area is linear. It also appears that the linear relationship is of the same kind for each of the three campaigns. Thus the covariate adjustment may be made by the same linear relationship in all the three campaigns.

A caveat here: This linearity and common regression may not always be the case; in that situation, appropriate relationships need to be developed for each campaign and made use of for making the comparisons. Here we use a common regression for all campaigns to make adjustment of comparisons.



In This Issue...

Analysis of Covariance

After a preliminary investigation as above the first step in the analysis of covariance is to include the covariate as an explanatory variable in the ANOVA. The results in this example are as follows:

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   3.0795     4.6459   0.663    0.513
campB         5.8887     4.2474   1.386    0.177
campC         2.0989     4.5398   0.462    0.648
area          4.6288     0.2148  21.546   <2e-16 ***
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 9.432 on 26 degrees of freedom
Multiple R-squared:  0.9546,Adjusted R-squared:  0.9494
F-statistic: 182.3 on 3 and 26 DF,  p-value: < 2.2e-16

Analysis of Variance Table

Response: diffsales
          Df Sum Sq Mean Sq F value    Pr(>F)
camp       2   7358    3679  41.352 8.385e-09 ***
area       1  41302   41302 464.216 < 2.2e-16 ***
Residuals 26   2313      89
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1
```

If you compare the Analysis of Variance tables with and without “area” you will see that a great deal of variation in the diffsales is because of the variation due to the area of the stores. The residual variation has come down considerably. Now the difference between the telephone and the no campaign which was significant at 5% level when area was ignored is no longer significant. So, it seems the differences were caused by the differences in the areas.

Adjusted Means

The formula for covariate-adjusted mean is as follows:

$$\bar{Y}_j^* = \bar{Y}_j - \beta(\bar{X}_j - \bar{X}),$$

where j indicates the campaign index
$\bar{Y}_j^*$ is the adjusted mean of diffsales for campaign j
$\bar{Y}_j$ is the unadjusted mean of diffsales for campaign j
β is the regression coefficient (this could be different for different groups and the regressions may even be nonlinear)
$\bar{X}_j$ is the mean of area of campaign j
$\bar{X}$ is the grand mean of the covariate

In This Issue...

The common regression coefficient is 4.642.

Campaign	Unadjusted Mean	Adjusted Mean	Area
None	79.82	95.46	16.58
Email	96.41	101.33	18.89
Telephone	118.07	97.46	24.39

Note that the adjusted means make Email better than Telephone in view of the fact that the area on which Telephone was based is larger than that of Email. Standard errors should be computed for adjusted means and hypothesis tests on the adjusted means could be computed. The hypothesis test would now be as follows:

H_0: the campaign means are equal after controlling for the covariate
H_1: the campaign means are not equal after controlling for the covariate

References

- Anne Boomsma(2012): *Analysis of Covariance with R*. Department of Statistics & Measurement Theory. University of Groningen. http://www.ppsw.rug.nl/~boomsma/apstatdata/ANCOVA_R.pdf