HUDM 5123 - Linear Models and Experimental Design
Notes 04 - Continuous Predictors and ANCOVA

1 Data

Data for today's notes come from a randomized experiment to study the efficacy of acupuncture for treating headaches. Results of the trial were published in the British Medical Journal in 2004. You may view the paper at the following link: `http://www.bmj.com/content/328/7442/744.full`. The data set includes 301 cases, 140 control (no acupuncture) and 161 treated (acupuncture). Participants were randomly assigned to groups. Variable names and descriptions are as follows:

- `age`; age in years
- `sex`; male = 0, female = 1
- `migraine`; diagnosis of migraines = 1, diagnosis of tension-type headaches = 0
- `chronicity`; number of years of headache disorder at baseline
- `acupuncturist`; ID for acupuncture provider
- `group`; acupuncture treatment group = 1, control group = 0
- `pk1`; headache severity rating at baseline
- `pk5`; headache severity rating 1 year later

2 One-Way ANOVA

Visual examination of parallel boxplots, shown in Figure 1, reveals that the distribution of posttest headache severity was lower, on average, for the group that received acupuncture than for the control group. We will run one-way ANOVA to test the effect of the acupuncture treatment on headache severity one year later. The full model is as follows:

$$\text{pk5}_i = \beta_0 + \beta_1 \text{group}_i + \epsilon_i$$

The reduced model is identical with the exception that β_1 will be constrained to be 0.

$$\text{pk5}_i = \beta_0 + \epsilon_i$$

The null hypothesis in this case is that the regression slope on a dummy-coded (or deviation coded) group indicator is 0.

$$H_0 : \beta_1 = 0$$

Because the treatment group variable is dummy-coded, the slope parameter, β_1, can be interpreted as the difference in group means, $\mu_1 - \mu_0$. Thus, another way of writing the null hypothesis is as follows.

$$H_0 : \mu_1 - \mu_0 = 0$$
$$H_0 : \mu_1 \neq \mu_0$$

If the categorical treatment variable here had three levels instead of two (say, treatment 1, treatment 2, and control), then we would have needed two dummies in the full model and the null hypothesis would have been $H_0 : \beta_1 = \beta_2 = 0$, or $H_0 : \mu_1 = \mu_2 = \mu_3$. If the categorical treatment variable had four levels, the null hypothesis for the one-way ANOVA would have used three betas (or four means). And so on.

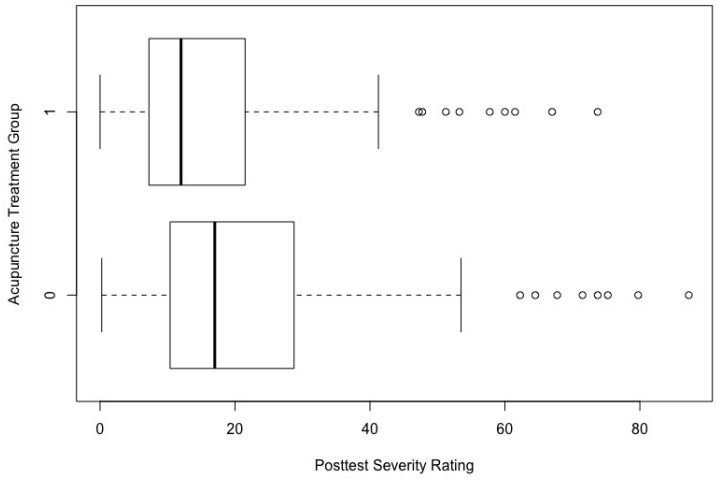

Figure 1: Side-by-side boxplots of headache severity measured after one year; 1 = acupuncture group, 0 = control group

The general ANOVA table: RSS_r - RSS_f

Source	Sum of Squares	df	Mean Square	F
Regression	RegSS	$df_R - df_F$	$\frac{RegSS}{df_R - df_F}$	$\frac{RegMS}{RMS}$
Residuals	RSS	df_F	$\frac{RSS}{df_F}$	

The ANOVA table for testing $H_0 : \beta_1 = 0$ against $H_1 : \beta_1 \neq 0$ with acupuncture data: Plugging in the estimated coefficients for the full model (with dummy-coded treatment variable) gives the prediction equation for the one-way ANOVA model:

$$\hat{\text{pk5}}_i = 22.3 - 6.1(\text{group}_i)$$
$$[\hat{\text{pk5}}_i | \text{group}_i = 0] = 22.3$$
$$[\hat{\text{pk5}}_i | \text{group}_i = 1] = 22.3 - 6.1 = 16.2$$

Table 1: ANOVA table for the test of acupuncture group.

Source	Sum of Squares	df	Mean Square	F	p-value
Group	2783.4	1	$\frac{2783.4}{1} = 2783.4$	$\frac{2783.4}{235.2} = 11.8$	.0007
Residuals	70333	299	$\frac{70333}{299} = 235.2$		

Thus, there are only two distinct predicted values based on the full model: 22.3, the mean of the control group, and 16.2, the mean of the acupuncture group. R^2 for the full model is .038, which suggests that about 4% of the variability in the one-year headache severity rating can be explained by whether a study participant received the acupuncture treatment or not.

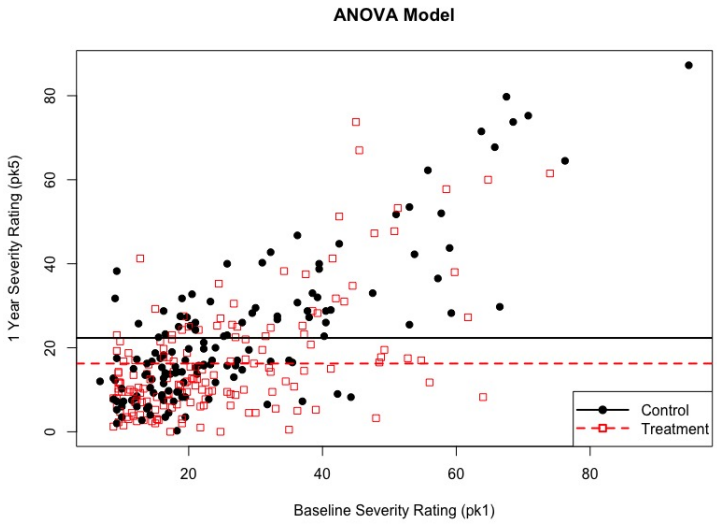

Figure 2: Scatterplot of headache severity at one year vs at baseline with ANOVA model predicted values

3 Checking Baseline Balance on Predictors

Because participants were randomly assigned to groups, there should be no systematic differences across treatment and control groups on *any* variables. This is very different from an observational study (i.e., a study lacking random assignment) where we might expect large and significant covariate differences across treatment groups. To verify that random assignment was carried out successfully, we can check for balance on the covariates across treatment groups at baseline.

Let $\bar{X}_T$ and $\bar{X}_C$ be the sample means, and s_T^2 and s_C^2 be the sample variances, for the treated and control groups, respectively. One way to check for balance across treatment conditions is to examine the mean difference across groups, $\bar{X}_T - \bar{X}_C$, where closer means indicate better balance. Means are not a perfect solution, however, because they provide no information about the scale or variability of the data. A solution is to use the *standardized*

mean difference, d, which is an effect size measure that quantifies the difference in group means scaled by the number of pooled standard deviations.

$$d = \frac{\bar{X}_T - \bar{X}_C}{s_{\text{pooled}}}, \text{ where } s_{\text{pooled}} = \sqrt{\frac{(N_T - 1)s_T^2 + (N_C - 1)s_C^2}{N_T + N_C - 2}},$$

where N_T and N_C are the sample sizes for the treated and control groups, respectively. The *variance ratio*, r, is another effect size measure that is useful for assessing balance, defined as follows.

$$r = \frac{s_T^2}{s_C^2}.$$

The standardized mean difference, d, allows us to assess imbalance across two groups in terms of the *center*, or *location*, of the data; while the variance ratio, r, measures imbalance by *spread*, or *variability*, across two groups. The larger the value of d, the more out of balance the two groups are with respect to their means. There are no strict limits for what values of the standardized mean difference are 'too large', though some authors (e.g., Steiner and Cook, 2011) have suggested that $|d| > .1$ or .2 be used to flag covariates that are imbalanced. For the variance ratio, r, values farther from 1 indicate more imbalance. Again, there are no absolute rules here either, though Rubin (2001) suggested that variance ratios outside of [4/5, 5/4] are indicative of a level of imbalance that is "of concern." The balance statistics for the acupuncture data follow in the table below.

Table 2: Balance Statistics for Baseline Covariates and Outcome (pk5)

Variable	$\bar{X}_C$	s_C	$\bar{X}_T$	s_T	d	r	sig
Age	46.23	10.83	46.43	10.03	0.02	0.86	.86
Female	0.86	0.35	0.83	0.38	-0.08	1.17	.56
Migraine	0.94	0.23	0.94	0.23	0.01	0.98	1.00
Chronicity	21.91	13.30	21.33	14.53	-0.04	1.20	.72
Baseline Severity (pk1)	26.71	16.78	24.58	14.12	-0.14	0.71	.24
1 yr Severity (pk5)	22.34	17.01	16.25	13.72	-0.41	0.65	$< .001$

Note: $\bar{X}$ = sample mean, s = sample SD, d = sample standardized mean difference, r = sample variance ratio, sig = p-value from Welch's t test for continuous covariates and chi-square test for categorical covariates.

The value of $d = -0.14$ for pk1 tells us that at baseline (i.e., before any acupuncture treatments) the treatment group average severity score was about 14% of a pooled SD lower than the control group. Furthermore, a value of $r = 0.71$ says that the estimated variance of severity in the treated group was only 70% as large as that in the control group at baseline. We can use appropriate statistical tests to test for baseline differences, though the usual caveats regarding power apply. That is, if you have a very small sample, you won't find significant differences even when they exist, and if you have a very large sample, you will find differences to be significant even though, practically speaking, they are not meaningful. Running two-sample t-tests for the continuous covariates and chi-square tests of independence for the categorical covariates we find no evidence of significant differences

across treatment conditions at baseline. Nevertheless, the values of d and r for baseline severity indicate imbalance at a level the causes some concern.

The imbalance on the pretest measure of severity at baseline (pk1) is such that those in the control group *already had* higher severity ratings even before acupuncture. The ANOVA analysis we just ran did not account for those baseline differences in any way. In fact, all information other than group membership (acupuncture vs not) and the outcome (pk5) was ignored. This is problematic because we are trying to show that the acupuncture treatment was effective in reducing headache severity. If the control group already had higher severity at baseline, how will we be able to attribute any difference at posttest to the treatment?

4 Incorporating a Continuous Covariate

This study was designed so that participants were *randomly assigned* to either receive the acupuncture treatment or to be in the control condition. Because of random assignment, it *should* be the case that the participant profiles in both groups are similar. This helps to rule out alternative explanations for any observed treatment effect based on mean differences across groups. In contrast, in a non-randomized experiment, where participants self-select into treatment arms, the unadjusted mean difference cannot be trusted as an estimate of the causal effect of the treatment because there could have been imbalance on important variables across the two groups. Here, however, as we noted above, there is some imbalance on baseline headache severity despite the randomization.

The difference in group means (6.1 in this case) is referred to as the *unadjusted* difference in group means. Through ANOVA, we found that the acupuncture treatment had a significant effect on headache severity rating ($p = .0007$). When the effect of a categorical predictor on an outcome is of primary interest (as it is here with the acupuncture data) it is also possible to include one or more additional predictors in the full and reduced models. When a model is formulated to include such additional predictors, we say we *control for them*, and they are sometimes referred to as *covariates*; furthermore, the omnibus test for a categorical predictor while controlling for one or more continuous predictors is referred to as the analysis of covariance (ANCOVA).

4.1 What Does it Mean to 'Control For' a Variable?

To control for baseline headache severity (pk1) in examining the effect of acupuncture on posttest headache severity (pk5), the full and reduced models would be modified as follows. The full model:

$$\text{pk5}_i = \beta_0 + \beta_1 \text{pk1}_i + \beta_2 \text{group}_i + \epsilon_i$$

The reduced model is identical with the exception that β_2 will be constrained to be 0.

$$\text{pk5}_i = \beta_0 + \beta_1 \text{pk1}_i + \epsilon_i$$

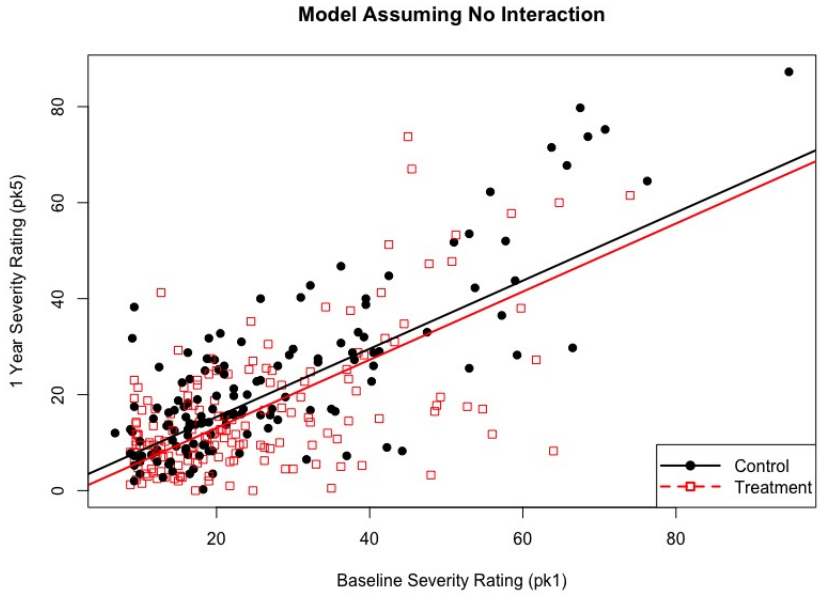

Figure 3: Scatterplot of the effect of acupuncture on headache severity, controlling for baseline headache severity

The ANOVA table for testing $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ with acupuncture data (this is the ANCOVA model):

Table 3: ANOVA table for the test of acupuncture group controlling for baseline headache severity

Source	Sum of Squares	df	Mean Square	F	p-value
Group	1568	1	$\frac{1568}{1} = 1568$	$\frac{1568}{116.8} = 13.4$	.0003
Residuals	34799	298	$\frac{34799}{298} = 116.8$		

Notice several changes going from ANOVA to ANCOVA. For each of these changes, consider why it happened and what the implications are for the p-value for group.

- The residual SS dropped from 70333 to 34799. Why? The model that accounts for pretest makes much better predictions.
- The residual df dropped by one from 299 to 298. Why? It took one additional parameter to include the linear relationship between pk1 and pk5.
- The SS for the treatment group also decreased from 2783 to 1568. Why? Recall that in this case, there was initial imbalance with respect to the baseline headache score across groups such that the control group had worse headaches to begin with. After linearly controlling for baseline, some of the group differences were 'controlled away'.
- The F statistic for the treatment group went up from 11.83 to 13.43. Why? See the formula for F_0.

- The p-value for group went down from .00066 to .00029. Why? Larger F_0 translates to a smaller p-value.

Thus, with ANCOVA, there are competing forces at work. On the one hand, an additional degree of freedom is used up for the covariate which makes the denominator of the incremental F test slightly larger. In addition, in this case, the inclusion of the covariate reduced the SS for the treatment group. Look at baseline differences to see why. Nevertheless, the huge reduction in residual SS was enough to swamp the other factors and bring an overall increase in F and a decrease in the p-value. In randomized experiments, ANCOVA is often used to increase precision of estimation by carving away a large chunk of the residual SS which, ultimately, makes the F statistic larger. Note that only covariates that are linearly related to the outcome will be successful in reducing the residual SS, this fact makes a pretest a particularly good candidate for a covariate.

4.2 Treatment by Covariate Interaction

Our analyses so far have assumed that the treatment effect is constant across all values of the covariate. This may or may not be the case, so it is necessary to check the assumption. To test for interaction we may use the following full and reduced models. The full model:

$$\text{pk5}_i = \beta_0 + \beta_1 \text{pk1}_i + \beta_2 \text{group}_i + \beta_3 pk1 \times \text{group}_i + \epsilon_i$$

and the reduced model:

$$\text{pk5}_i = \beta_0 + \beta_1 pk1 + \beta_2 \text{group}_i + \epsilon_i$$

may be used to test $H_0 : \beta_3 = 0$.
The ANOVA table for testing $H_0 : \beta_2 = 0$ against $H_1 : \beta_2 \neq 0$ with acupuncture data (this is the ANCOVA model):

Table 4: ANOVA table for the test of baseline headache by treatment group interaction

Source	Sum of Squares	df	Mean Square	F	p-value
Group	1141	1	$\frac{1141}{1} = 1141$	$\frac{1141}{113.3} = 10.1$	.002
Residuals	33658	297	$\frac{33658}{297} = 113.3$		

The p-value for the test of treatment group by baseline severity interaction is significant ($p = .002$). Thus, there is a statistically significant interaction between group and baseline severity. We will discuss interactions in more detail next class, but for now, suffice it to say that the presence of an interaction implies that the effect of acupuncture treatment varies with different levels of baseline severity. The plot is helpful in understanding how the effect varies.
The estimated coefficients yield the following prediction equation:

$$\text{pk5}_i = 0.4 + 0.8\text{pk1}_i + 1.9\text{group}_i - 0.25\text{pk1}_i \times \text{group}_i$$

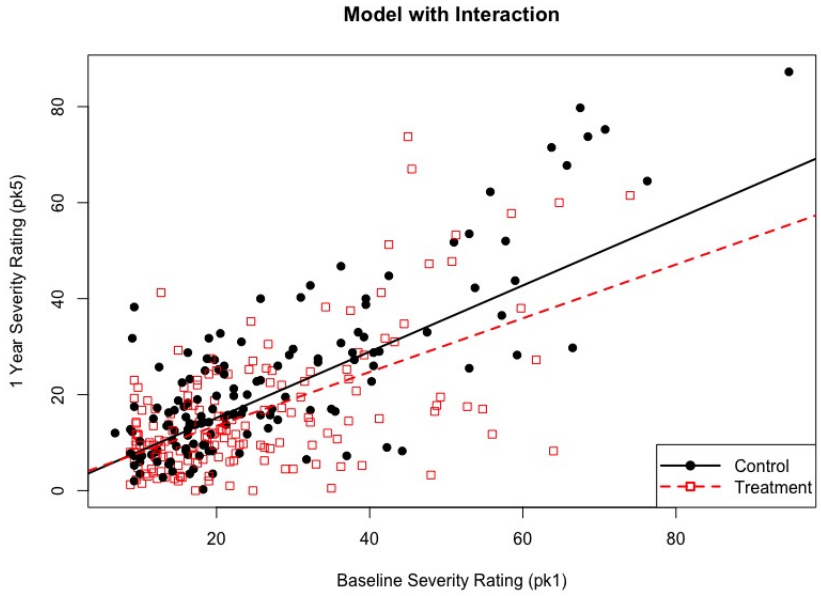

Figure 4: Scatterplot of the effect of acupuncture on headache severity, controlling for baseline headache severity, and the interaction between group and baseline severity

By substituting values of 0 and 1 for the group, we can see how the model-implied predictions differ by group.

$$
\begin{aligned}
[\text{pk5}_i|\text{group}_i = 0] =& 0.4 + 0.8\text{pk1}_i \\
[\text{pk5}_i|\text{group}_i = 1] =& (0.4 + 1.9) + (0.8 - 0.25)\text{pk1}_i \\
[\text{pk5}_i|\text{group}_i = 1] =& 2.3 + 0.55\text{pk1}_i
\end{aligned}
$$

According to this model, the acupuncture treatment had very little effect for those with low baseline levels of headache severity. For those with high baseline levels, the treatment was increasingly more effective.

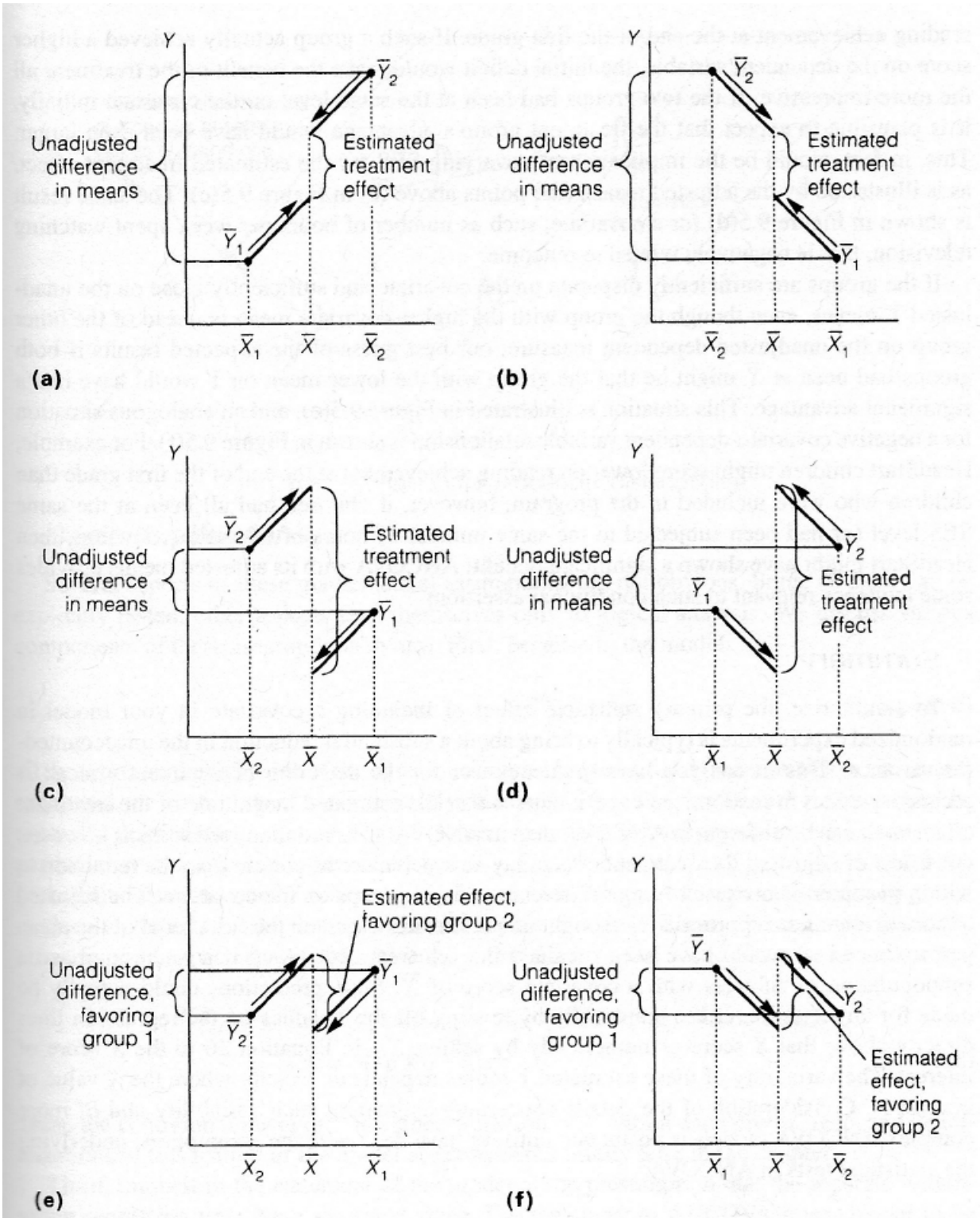

FIG. 9.5. Some possible relationships between unadjusted and adjusted estimates of treatment effect: (a and b) An apparent treatment benefit due primarily to preexisting differences; (c and d) estimate of treatment effect increased by adjusting for preexisting difference; (e and f) an apparent harmful effect of treatment seen as benefit by adjusting for preexisting differences.

Figure 5: Figure 9.5 from Maxwell and Delaney