

Adjusting for Regression Effect in Uncontrolled Studies

Joshua D. Naranjo and Joseph W. McKean

Department of Mathematics and Statistics, Western Michigan University

Kalamazoo, Michigan 49008, U.S.A.

SUMMARY

When clinical studies require enrolled patients to have abnormal assays, the natural tendency of repeat measurements to regress towards the mean can lead to a false assessment of effectiveness of therapy. We propose a method to more accurately estimate the true effect of therapy by adjusting for a component of improvement that can be attributed to regression effect. The model we use allows for a combination of additive and/or multiplicative effects of the therapy.

Key words: Additive effects; Dual effects; Multiplicative effects; Rank-based estimation; Regression effect; Sampling with selection; Uncontrolled studies.

1 INTRODUCTION

Regression effect is the tendency of extreme values to move closer to the mean when measured a second time. It explains why tall fathers tend to have sons who are shorter than them, and why placebo patients tend to show improvement in proportion to the severity of the symptoms. Consider a clinical trial where patients are selected have high values of a variable of interest X (e.g. level of blood cholesterol). Then their average value is expected to drop in a second measurement (Y). This drop in average due to regression effect can cause a misleading assessment of the potency of treatment (see, for example, Freedman et al. (1998) and Chuang-Stein (1993) for more detailed discussion).

Comparison with a control group does accurately measure effects of treatment. In this paper, though, we are concerned with situations where a placebo arm is not available (as in early phase studies), or unethical (as when standard treatments exist). For convenience, we assume that patients have *high* values of the variable of interest. The procedure works when patients are selected for low values as well.

There is considerable literature on the problem of adjusting for regression effect in uncontrolled studies. James (1973), Senn and Brown (1985, 1989), and Chen and Cox (1992) have proposed procedures for estimating treatment effect under a *multiplicative* model where treatment effect on a variable depends on the level of the variable at baseline. Curnow (1987), Mee and Chua (1991), and Lin and Hughes (1995) consider estimation of treatment effects under an *additive* model of constant treatment effect regardless of the level at baseline.

In this paper, we consider a dual effects model which allows for both patterns of treatment effect: additive and multiplicative. We estimate a measure of additive effect δ and a measure of multiplicative effect γ . Test procedures can be conducted against specific alternatives of nonzero additive effect or nonzero multiplicative effect, or against the general alternative of nonzero treatment effect regardless of pattern.

In Section 3, we discuss a clinical study for a cholesterol-reducing drug. Patients were selected for having LDL cholesterol levels greater than 165 mg/dl. Our proposed procedure detects significant treatment effects at high dose, and indicate the existence of both additive and multiplicative effects in cholesterol reduction. For this example, the procedure of Chen and Cox (1992) disagrees with our conclusions, while Mee and Chua (1991) agrees for the most part.

2 The Model and Estimates

Let (X, Y) denote Baseline and Posttreatment values for effectiveness of therapy. In the absence of treatment effects, we assume a bivariate normal distribution with common mean μ , common variance σ^2 , and correlation ρ . Express the bivariate normal model in the regression form

$$Y = \mu + \rho(X - \mu) + \varepsilon, \tag{1}$$

where $X \sim N(\mu, \sigma^2)$, $\varepsilon \sim N(0, \sigma^2(1 - \rho^2))$, and X and ε are independent. Since repeat measurements are generally positively correlated, we assume that $0 < \rho < 1$ without loss of generality. Equation (1) implies that $E(X - Y | X = x) = (1 - \rho)(x - \mu)$. Thus, a reduction of magnitude $(1 - \rho)(x - \mu)$ is expected due merely to the fact that the patient was $x - \mu$ units above average at baseline. For a sample of patients, the expected average reduction due to regression effect is $(1 - \rho)(\bar{x} - \mu)$, where \bar{x} is the average of baseline values.

The Dual Effects Model

For modeling treatment effects, consider the *dual effects model*

$$Y = \mu + \rho(X - \mu) - \delta - \eta(X - \mu) + \varepsilon, \quad (2)$$

where X and ε have the same distribution in Equation (1). The null hypothesis of no treatment effect is represented by the parameter values $(\delta = 0, \eta = 0)$. A value $\delta > 0$ measures the amount of reduction due to treatment that is *independent of severity of the disease*, while $\eta > 0$ measures the component reduction due to treatment effect that is *proportional to the severity of the disease*. Rewrite equation (2) as

$$Y = \mu - \delta + \gamma\rho(X - \mu) + \varepsilon, \quad (3)$$

where $\gamma = (\rho - \eta)/\rho$. If the additive component $\delta = 0$, then equation (3) reduces to the multiplicative model of Chen and Cox (1992), where treatment effect is represented by $\gamma \neq 1$. If $\gamma = 1$, then equation (3) reduces to the additive model of Mee and Chua (1991), where treatment effect is represented by $\delta > 0$. Henceforth, we refer to δ and γ as the additive and multiplicative components, respectively, of treatment effect.

Parameter Estimates

Following Chen and Cox (1992), we consider the case of selected sampling where the X measurement is recorded for a random sample of $n + m$ patients, but Y measurements are taken only on the subset of n X -values that exceeded a severity-of-disease criterion $X > a$. Thus we have data of the form $(x_1, y_1), \dots, (x_n, y_n), x_{n+1}, \dots, x_{n+m}$. In the discussion that follows, the parameters μ and σ^2 refer to estimates of the mean and variance of the healthy population based on the larger x -sample x_1, \dots, x_{n+m} . In the absence of a larger x -sample, we will assume that estimates for the mean and variance may be obtained from separate studies or independent sources of data. These assumptions are not unique to this paper. For instance, Mee and Chua (1991) assume that μ is known while Chen and Cox (1992) assume an X -sample large enough so that $n/m \rightarrow 0$.

As in Chen and Cox, consider the likelihood $L = \prod_{i=1}^n f(y_i|x_i)h(x_i) \prod_{j=n+1}^{n+m} h(x_j)$. The marginal distribution $h(x)$ does not depend on δ , γ , or ρ and may be excluded. We thus have

the log-likelihood (omitting terms without δ , γ , or ρ)

$$\ell(\delta, \gamma, \rho) = -\frac{n}{2} \ln(1 - \rho^2) - \frac{1}{2\sigma^2(1 - \rho^2)} \sum_{i=1}^n \{y_i - \mu + \delta - \gamma\rho(x_i - \mu)\}^2.$$

Taking partial derivatives and solving the resulting score equations give the MLE's

$$\hat{\delta} = \hat{\gamma}\hat{\rho}(\bar{x} - \mu) - (\bar{y} - \mu) \quad (4)$$

$$\hat{\gamma}\hat{\rho} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} \quad (5)$$

$$\hat{\rho} = \sqrt{1 - \frac{\sum_{i=1}^n \{y_i - \bar{y} - \hat{\gamma}\hat{\rho}(x_i - \bar{x})\}^2}{n\sigma^2}} \quad (6)$$

$$\hat{\gamma} = \hat{\gamma}\hat{\rho}/\hat{\rho} \quad (7)$$

Standard error estimates based on the information matrix are $SE(\hat{\delta}) = \sqrt{\hat{\sigma}^2(1 - \hat{\rho}^2)/n}$ and $SE(\hat{\gamma}) = \sqrt{(1 - \hat{\rho}^2)\{\hat{\gamma}^2(1 - \hat{\rho}^2) + 2\hat{\rho}^2\}/(2n\hat{\rho}^2)}$.

The estimate $\hat{\delta}$ in equation (4) is the same as the estimate proposed in Mee and Chua (1991). However, the standard error $SE(\hat{\delta})$ is different from Mee and Chua. The Mee and Chua approach based on LS regression gives simpler estimates of standard error but this approach will not work for the dual effects model which is not linear in the parameters. Hence, we took the likelihood approach to estimating standard errors. Unfortunately, the simulations in Section 4 show that the standard errors based on asymptotic likelihood theory underestimate the simulated standard errors by a large margin. Though we have included the standard error estimates for completeness, we recommend the alternative use of bootstrap-based confidence intervals in applications. These bootstrap confidence intervals (Efron and Tibshirani, 1993) performed well in the real data example of Section 3 and the simulations in Section 4.

The bootstrap confidence intervals are calculated as follows. A sample of size n is drawn with replacement from $(x_1, y_1) \dots, (x_n, y_n)$. This gives new estimates $(\hat{\delta}_1^*, \hat{\gamma}_1^*)$. The bootstrap joint distribution of $(\hat{\delta}, \hat{\gamma})$ is obtained by repeating the procedure B times to get $(\hat{\delta}_1^*, \hat{\gamma}_1^*), \dots, (\hat{\delta}_B^*, \hat{\gamma}_B^*)$. A 95% confidence interval for, say, δ is given by the 2.5th and 97.5th percentiles of the bootstrap distribution of $\hat{\delta}$. These bootstrap-percentile confidence intervals are denoted 95% BCI in Table 2.

Tests of Significance

The treatment effect components δ and γ may be tested separately using t -ratios based on the likelihood estimates and standard errors. However, as in estimation, we recommend the use of bootstrap-based inference over the likelihood-based standard errors above. A marginal test with 5% level of significance against the alternative hypothesis $\delta \neq 0$ can be conducted by constructing a 95% bootstrap-percentile confidence interval I_δ for δ to see if it contains the value 0. Similarly, the alternative hypothesis $\gamma \neq 1$ can be detected if its 95% confidence interval I_γ does not contain 1. An overall test against the general alternative ($\delta \neq 0$ or $\gamma \neq 1$) is conducted as follows. Let I_δ^* and I_γ^* be bootstrap-percentile 97.5% confidence intervals for δ and γ , respectively. A Bonferroni test with at most a 5% level of significance for testing $H_o : \delta = 0, \gamma = 1$ versus $H_a : \delta \neq 0$ or $\gamma \neq 1$ is given by

$$IB_{\delta\gamma}: \text{Reject } H_o \text{ if } 0 \notin I_\delta^* \text{ or } 1 \notin I_\gamma^* . \quad (8)$$

In the simulation studies of Section 4, this test achieved the desired level of significance under the null hypothesis, and had good power properties to detect the presence of treatment effects.

3 Example

We look at a clinical study to assess effectivity of a drug in reducing cholesterol levels. At the screening phase, eligible patients with elevated LDL cholesterol greater than 165 mg/dl were selected for the study. Selected patients were randomized into a high dose treatment group and a low dose group. A follow-up lipid profile was taken two weeks after taking the drug. Baseline and Week 2 data are summarized in Table 1.

We calculated estimates and confidence intervals for the additive model of Mee and Chua (1991), the multiplicative model of Chen and Cox (1992), and our dual effects model. The results are summarized in Table 2. For these calculations, the required population parameters for total cholesterol and LDL were estimated from lab normal ranges in the *Clinical Guide to Laboratory Tests*, by treating population reference ranges as a 4 standard deviation range. Total cholesterol was estimated as having mean $\mu=225$ mg/dL with a standard deviation of $\sigma=45$ mg/dL. For LDL, the estimates are $\mu=150$ mg/dL and $\sigma=40$ mg/dL.

We do not know whether the underlying treatment effect, if any, is additive or multiplicative so each test is considered as a test for the significance of drug effect, regardless of pattern. Note that the Mee and Chua additive model declares significant treatment effects in the high dose group for both total cholesterol and LDL. The low dose groups show no statistical significance for either total cholesterol or LDL. Similar results are seen when we look at the additive component δ of the the dual effects model. This is not surprising, since the estimates $\hat{\delta}$ are equal (though the confidence intervals are not). The multiplicative component γ of the dual effects model are also significant for the not detected by the purely additive model.

The Chen and Cox multiplicative model gives contrary results to the additive and dual effects models. It sees no treatment effect on total cholesterol in the high dose group, contrary to the competitors. On the other hand, it declares significant treatment effect on total cholesterol for low dose, again contrary to the competitors.

This same clinical study contained a placebo arm, which we use for confirmation. In the presence of a placebo group, the two-sample t -test based on $\bar{C}_{trt} - \bar{C}_{pl}$ where \bar{C} denotes average change from baseline can detect either additive or multiplicative treatment effects. The p -values for the two-sample tests against placebo are $p < .0001$ for high dose total cholesterol, $p = .34$ for low dose cholesterol, $p < .0001$ for high dose LDL, and $p = .11$ for low dose LDL. The first three p -values agree with the additive and dual effects tests of significance. The fourth value $p = .11$ for low dose LDL indicates some possibly weak treatment effect. We believe that this picture is more appropriately painted by the joint confidence intervals for dual effects, rather than the additive effects model which flatly proclaims that there are no treatment effects present.

In summary, the tests based on the additive model and dual effects model agree with two-sample t -tests against placebo in two out of four cases. In these two cases (high dose total cholesterol and high dose LDL), the test based on the multiplicative model reaches the opposite conclusion. In one case (LDL low dose), the dual effects and multiplicative model see some degree of treatment effect while the additive model does not.

4 Monte Carlo Study

To investigate the behavior of these different procedures, we performed a simulation study. For each simulation, we obtained 200 independent observations from a bivariate normal distribution, $(x_i, y_i), i = 1, \dots, 200$. The selected sample was based on the upper 25% of the x_i 's. The observations were generated under the dual treatment effects model (3). We considered four situations: (S1) $\delta = 0, \gamma = 1$; (S2), $\delta = 0, \gamma = .6$; (S3) $\delta = .7, \gamma = 1$; and (S4) $\delta = .7, \gamma = .6$. In all situations we set $\sigma^2 = 1$ and $\rho = .8$. For each situation 1000 simulations were run. For each simulation we used the larger x -sample with $n = 200$ to estimate μ and σ^2 .

We considered four procedures: (i) NML, our dual model estimates $\hat{\delta}$ and $\hat{\gamma}$ with likelihood based standard errors; (ii) NMB, our dual model estimates $\hat{\delta}$ and $\hat{\gamma}$ with inference based on the bootstrap; (iii) CC, the Chen and Cox multiplicative model with likelihood based estimates; and (iv) MC, the Mee and Chua additive model with LS standardization.

The CC procedure performed poorly overall. The estimate of multiplicative effect γ was very biased whenever an additive effect was present (situations S3 and S4). Its corresponding test statistic had adequate Type I behavior in the null situation S1, but was extremely liberal in S3 in which $\gamma = 1$ but there is an additive effect. Thus it cannot be used to discern between multiplicative and additive effects. On the other hand, the MC procedure had good Type I behavior in both the null situation S1 and the marginal situation S2. Thus it did not exhibit power in the presence of purely multiplicative effects.

The estimates of the NML procedure had little bias in the four situations. The standardization by the usual likelihood theory, however, was extremely poor. The likelihood-based asymptotic variances were too small, typically 3 to 8 times smaller than the simulated empirical variances; hence, the resulting tests are unreliable and cannot be recommended. This may be due to the fact that the likelihood equation does not reflect the selected-sampling nature of the data.

On the other hand the bootstrap procedures NMB performed very well over all four situations. The dual effects confidence intervals I_δ and I_γ had good simulated coverage probabilities of their respective parameters in all four situations. The dual effects Bonferroni bootstrap test $IB_{\delta\gamma}$ also did well against the overall alternative H_a for existence of a treatment effect. Its empirical α levels were quite close to nominal levels in situation S1 and it exhibited high power

in situations S2-S4. Thus, it can be used as a general test for the presence of treatment effect, whether additive or multiplicative. Furthermore, it appears to have the added advantage of being able to tell whether the underlying treatment effect is additive or multiplicative by looking at I_δ and I_γ separately. This is because in situation S2, I_γ has high rejection rate while I_δ does not, and in situation S3, I_δ has high rejection rate while I_γ does not. Thus, the three tests $IB_{\delta\gamma}$, I_δ , and I_γ based on the bootstrap together provide a test against all three types of alternative hypotheses as given in situations S2-S4.

5 Summary

For sampling with selection, the literature contains two lines of approach to estimating treatment effects after controlling for regression effect. The approaches differ according to the pattern of treatment effect: additive or multiplicative. The simulations show that under bivariate normality, the multiplicative model by Chen and Cox (1992) can detect the presence of treatment effects regardless of pattern but it cannot distinguish between them. Curiously, in the real data example, it gives contrary results to the additive and dual effects competitors *and* the two-sample t -tests against placebo. This could be a sign that the estimate is sensitive to deviations from normality.

The additive model by Mee and Chua (1991) gives more consistent and predictable results in the presence of additive effects. However, the test has no power to detect treatment effect if the underlying effect pattern turns out to be multiplicative.

We have proposed a dual effects model that can detect either additive or multiplicative effects, and it can distinguish between the two patterns. It can be used as a tool for formal inference, or as a diagnostic aid to determine which treatment effect pattern holds. Inference based on bootstrap percentiles were shown to be more reliable than standard errors based on likelihood theory.

REFERENCES

- Chen, S. and Cox, C. (1992). Use of baseline data for estimation of treatment effects in the presence of regression to the mean. *Biometrics* **48**, 593-598.
- Chuang-Stein, C. (1993). The regression fallacy. *Drug Information Journal* **27**, 1213-1220.
- Curnow, R.N. (1987). Correcting for regression in assessing the response to treatment in a selected population. *Statistics in Medicine* **6**, 113-117.
- Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. New York: Chapman & Hall.
- Freedman, D., Pisani, R. and Purves, R. (1998). *Statistics*, 3rd Ed. New York: Norton & Company, Inc.
- James, K.E. (1973). Regression toward the mean in uncontrolled clinical trials. *Biometrics* **29**, 121-130.
- Lin, H.M. and Hughes, M.D. (1995). Use of historical marker data for assessing treatment effects in Phase I/II trials when subject selection is determined by baseline marker level. *Biometrics* **51**, 1053-1063.
- Mee, R.W. and Chua, T.C. (1991). Regression toward the mean and the paired sample t test. *The American Statistician* **45**, 39-41.
- Senn, S.J. and Brown, R.A. (1985). Estimating treatment effects in clinical trials subject to regression to the mean. *Biometrics* **41**, 555-559.
- Senn, S.J. and Brown, R.A. (1989). Maximum likelihood estimation of treatment effects for samples subject to regression to the mean. *Communications in Statistics* **18**, 3389-3406.
- Tietz, N.W., and Finley, P.R., Eds. (1983). *Clinical Guide to Laboratory Tests*. W.B. Saunders Co., Philadelphia.

Table 1: Summary statistics for Total Cholesterol and LDL

			MEAN	SD	MIN	MAX
Total Cholesterol	High Dose (n=19)	Baseline	287.0	54.40	235	433
		Week 2	235.4	37.68	184	339
	Low Dose (n=19)	Baseline	271.5	28.57	233	366
		Week 2	247.2	31.13	188	294
LDL	High Dose (n=19)	Baseline	217.5	55.70	166.4	364.6
		Week 2	160.8	39.10	99.4	262.0
	Low Dose (n=19)	Baseline	195.2	22.28	171.8	258.8
		Week 2	167.4	26.14	131.8	220.2

Table 2: Estimates and Confidence Intervals for the Cholesterol Reduction Study. 95% BCI's are bootstrap confidence intervals.

		Total Cholesterol		LDL	
		High dose	Low dose	High dose	Low dose
Additive:	$\hat{\delta}$:	27.38	-2.53	31.52	-9.33
	95% CI:	(13.65, 41.09)*	(-30.22, 25.16)	(17.60, 45.44)*	(-43.37, 24.71)
Multiplicative:	$\hat{\gamma}$:	.42	.59	.45	.60
	95% CI:	(-.31, 1.15)	(.26, .92)*	(.23, .67)*	(.32, .88)*
Dual Effects:	$\hat{\delta}$:	27.38	-2.53	31.52	-9.33
	95% BCI:	(11.10, 45.50)*	(-44.79, 28.29)	(14.54, 50.09)*	(-63.52, 20.51)
	$\hat{\gamma}$:	.66	.54	.69	.41
	95% BCI:	(.41, .84)*	(.01, 1.49)	(.36, .88)*	(.01, .88)*

*Statistically significant (CI for δ does not contain 0 or CI for γ does not contain 1).