1. Using tlc data, reproduce results in S95-108.

For reproducing results I need to use the “lead-data”, where there’s no covariate besides time. So here it is:

Variance covariance matrix (it was not clear from the code that variances were allowed to vary for each time…In R, besides setting correlation as unstructured, you have to specify heteroscedasticity):

Marginal variance covariance matrix

[,1] [,2] [,3] [,4]

[1,] 25.210 15.466 15.138 22.986

[2,] 15.466 58.867 44.029 35.965

[3,] 15.138 44.029 61.657 33.022

[4,] 22.986 35.965 33.022 85.494

Model results:

Generalized least squares fit by REML

Model: lead ~ week

Data: d

AIC BIC logLik

1308.337 1354.231 -640.1687

AIC, BIC, and -2loglikelihood slightly different than in SAS. I presume is an algorithm issue

Now for the null model likelihood ratio test, I need to fit two models, one with and one without the time covariate, and using ML instead of REML

Model df AIC BIC logLik Test L.Ratio p-value

m3 1 11 1381.866 1418.148 -679.9331

m2 2 14 1314.459 1360.635 -643.2294 1 vs 2 73.40745 <.0001

Again, this is different from SAS output. Degrees of freedom here are 3, while in the SAS output were 9. 3 df makes more sense to me, as the only difference between null and full model would be the regression coefficients for the three measurement occasions not included in the intercept.

Now, for the parameter estimates and standard errors, they look the same…finally some reassurance.

Coefficients:

Value Std.Error t-value p-value

(Intercept) 20.762 1.307627 15.877619 0e+00

week0 5.778 1.137826 5.078105 0e+00

week1 -7.240 1.203585 -6.015360 0e+00

week4 -5.248 1.273641 -4.120471 1e-04

Residual standard error: 7.560013

Degrees of freedom: 200 total; 196 residual I’ve a question here…are there really 196 residual df here? We have estimated 14 parameters (10 for the covariance matrix and 4 regression coefficients, so I would guess in reality there are 186 df left). Maybe I’m just divagating here.

Now for the test of fixed effects, for some reason df of the denominator are different in SAS (why 49, if we have 200, although repeated, observations):

Denom. DF: 196

numDF F-value p-value

(Intercept) 1 1250.4896 <.0001

week 3 54.5735 <.0001

Regarding the contrast, I think that in this case, where the reference level is set at week 6, making a contrast of week 6 minus week 0 is a bit redundant, as from the regression coefficients we know it would be a significant difference, with an estimate of -5.78. Nevertheless, for the sake of doing a contrast, here it is:

Linear Hypotheses:

Estimate Std. Error z value Pr(>|z|)

1 == 0 -5.778 1.138 -5.078 3.81e-07 \*\*\*

Now, if we disregard the covariance between observations within the same subject (slides 104 - 106) we see a drop in precision:

Response: lead

Df Sum Sq Mean Sq F value Pr(>F)

week 3 5104.4 1701.47 29.434 9.364e-16 \*\*\*

Residuals 196 11330.2 57.81

Denom. DF: 196

numDF F-value p-value

(Intercept) 1 1260.1142 <.0001

week 3 29.4336 <.0001

Parameter estimates: they are the same, but with less precision (larger SE’s)

Coefficients:

Value Std.Error t-value p-value

(Intercept) 20.762 1.075241 19.309159 0e+00

week0 5.778 1.520620 3.799765 2e-04

week1 -7.240 1.520620 -4.761214 0e+00

week4 -5.248 1.520620 -3.451223 7e-04

And the contrast, with the same caveat as mentioned above:

Linear Hypotheses:

Estimate Std. Error t value Pr(>|t|)

1 == 0 -5.778 1.521 -3.8 0.000193 \*\*\*

1. Using tlc data, analyze the same data with the different methods: My understanding of this problem is using that “same data” means analyzing the same data as before. So here it is
2. Mixed model (e.g., for longitudinal data): For this I fitted a random intercepts model

Random effects:

Formula: ~1 | ID

(Intercept) Residual

StdDev: 5.269501 5.480832

Fixed effects: lead ~ week

Value Std.Error DF t-value p-value

(Intercept) 20.762 1.075241 147 19.309159 0

week0 5.778 1.096166 147 5.271098 0

week1 -7.240 1.096166 147 -6.604837 0

week4 -5.248 1.096166 147 -4.787594 0

numDF denDF F-value p-value

(Intercept) 1 147 516.2184 <.0001

week 3 147 56.6411 <.0001

Of note is that the compound symmetry covariance structure assumed by this approach might not hold (and I believe it doesn’t, considering the covariance matrix parameter estimates from problem 1).

1. ANOVA (=one outcome at a time): don’t know if I got this right, but using the same data the only factor I can use is week, so here are the ANOVA results, which are the same as the ones from the linear regression model assuming uncorrelated error terms (except for the parameter estimates, with a post hoc test we could figure out which differences are significant).

week Residuals

Sum of Squares 5104.418 11330.204

Deg. of Freedom 3 196

Df Sum Sq Mean Sq F value Pr(>F)

week 3 5104.4 1701.47 29.434 9.364e-16 \*\*\*

Residuals 196 11330.2 57.81

Based on the (smaller) value of the F statistic, this type of analysis seems to be less precise (higher variance of the estimates, SSE) than what we would obtain if we assume the errors are correlated

1. repeated measures ANOVA

Error: id

Df Sum Sq Mean Sq F value Pr(>F)

Residuals 49 6914 141.1

Error: id:week

Df Sum Sq Mean Sq F value Pr(>F)

week 3 5104 1702 56.64 <2e-16 \*\*\*

Residuals 147 4416 30

By taking into account the correlation of repeated measurements within each subject (subject random effect or SS between), our test statistic is more precise (SSE is smaller, hence F is larger), because total variation (SST) is now divided into the SS between (subject effect), SS within (main effect, in this case week effect), and SS error (residual error). The week effect (SS week) remains virtually the same as in the previous analysis, but now SS error is smaller (because unexplained within individual error is now separated from between subject random effect), hence F is larger.

As for the mixed effect model, the compound symmetry covariance structure assumed by this approach might not hold. Still better than the simple ANOVA.

1. MANOVA (=ANOVA for multiple outcomes)

For this I compared all times versus baseline (3 parameter estimates), and did a multivariate analysis to see if these differences were significant (i.e. significantly greater than zero). Here are the summary results:

Df Hotelling-Lawley approx F num Df den Df Pr(>F)

(Intercept) 1 3.3412 52.345 3 47 5.091e-15 \*\*\*

Residuals 49

Which indicate a significant difference between all weeks and baseline. The approximate F statistic is very similar to the one obtained with repeated measures ANOVA. A difference though is that the degrees of freedom for the denominator have dropped dramatically, which suggests that as number of comparissons increase, this approach could be less efficient than repeated measures ANOVA. On the other hand, this approach does not assume sphericity, as repeated measures ANOVA does, which might not necessarily hold.

The significant differences hold for all comparisons, as showed bellow

Week 1 vs baseline:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -13.018 1.031 -12.63 <2e-16 \*\*\*

Week 4 vs baseline:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -11.026 1.064 -10.36 6.06e-14 \*\*\*

Week 6 vs baseline:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) -5.778 1.138 -5.078 5.92e-06 \*\*\*

Of all, I would stick to mixed model.

1. With the first 2 visits in tlc data, analyze the data with different methods for pre vs. post data:

I think I’m going to move here to the tlc-data. Just realized I might have had it all wrong by using the lead-data on #2…I think is too late for that now.

Here it is:

1. ANOVA

Df Sum Sq Mean Sq F value Pr(>F)

group 1 3101 3101.4 69.94 4.24e-13 \*\*\*

Residuals 98 4346 44.3

It shows that the mean lead level is significantly lower for the succimer group at week 1 (visit 2).

1. Change analysis (i.e., change as outcome): For this I run a regression model where the outcome was the difference baseline - week1, and the predictor was group (placebo or succimer)

Df Sum Sq Mean Sq F value Pr(>F)

group 1 3252.4 3252.4 103.72 < 2.2e-16 \*\*\*

Residuals 98 3072.9 31.4

We see the F value is higher than the ANOVA, because in a sense, by removing the baseline value, we are removing the error associated with it (less noise in the data).

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 13.0180 0.7919 16.44 <2e-16 \*\*\*

groupP -11.4060 1.1199 -10.18 <2e-16 \*\*\*

From this model we also see that the succimer group (intercept) has an average drop of 13 units from baseline measurement (and this is significantly different from zero), while the placebo has an average drop of about 1.6. We don’t get to see though in this analysis if the drop in the placebo group is significantly different from zero.

1. ANCOVA

Df Sum Sq Mean Sq F value Pr(>F)

group 1 3101 3101.4 102.8 < 2e-16 \*\*\*

lead.base 1 1418 1418.4 47.0 6.61e-10 \*\*\*

We see from the above table the F statistic is larger compared to the ANOVA. This is because we have taken from the error term the variability due to the lead levels at baseline (similar to what we did before with the change analysis). Still the F statistic is a bit smaller (less precise) than change analysis. This is because the values of lead at baseline and week 1 are correlated (0.42, which is less than 0.50 threshold in the paper mentioned).

Since this is a randomized trial with a balanced design and with no attrition, results from all the three above mentioned methods are valid estimates of the average causal effect of treatment.

1. Paired t-test: here I tested the difference before baseline and week 1 separated for each treatment group (the null for both was that E(difference) = 0). So I did two paired t-tests: one for the succimer group and one for the placebo

For the succimer group:

t = 12.627, df = 49, p-value < 2.2e-16

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

10.94617 15.08983

sample estimates:

mean of x

13.018

This is mostly in agreement with change analysis (except perhaps the t-value, here slightly lower)

For the control group:

t = 3.6852, df = 49, p-value = 0.0005711

alternative hypothesis: true mean is not equal to 0

95 percent confidence interval:

0.7329538 2.4910462

sample estimates:

mean of x

1.612

Here we see that the difference for the control group, although smaller than for the treatment group, is still significantly different from zero. A drawback from this approach is that we have too many groups we will run into the problem of multiple testing. For this we could do a Bonferroni adjustment.

1. How to estimate and interpret “Adjusted means” (=LS means) from #3?

Since I use R, I don’t have adjusted means in my output for ANCOVA.

What I found is that adjusted means or predicted marginal means are values of the predicted least squares line taken at fixed value of the covariate. This way we can obtain estimates of the (predicted) mean at levels of the variate of main interest (factor) that are not affected by imbalances in the values of the covariate amongst groups.

I fail to see how to get a predicted least squares line from an ANCOVA (I can clearly see that from a regression though).

Anyway, I found a package in R to do the adjusted means from an ANCOVA, and here are the results:

group lsmean SE df lower.CL upper.CL

A 13.4205 0.7770662 97 11.87824 14.96276

P 24.7615 0.7770662 97 23.21924 26.30376

These means are not quite different from the raw means (13.52 for A and 24.66 for P). This is because this was a randomized clinical trial, where baseline values are set to be equivalent by design.

1. How and why to use “Contrast” statement in SAS/R?

Why: You want to use a contrast to test hypotheses that are not directly implied by the model specification, e.g. comparing regression coefficients of levels of a categorical variable that do not involve the baseline level between each other, or comparing if the value at baseline is greater than the average of follow up values.

How: writing an adequate contrast matrix or specifying it in a function for the respective package. For example, for a regression model where a categorical variable has 4 levels, and the reference one is the first one, the appropriate matrix for a linear contrast for the hypothesis that coefficients for levels 3 and 4 are different would be (0 -1 1 0), and for testing differences between coefficients for levels 3 and 4 it would be (0 0 -1 1). Example of the code is my HW code.

1. After reading Tomasetti and Vogestein (Science 2015), aka, “bad luck cancer paper, a renowned statistician stated “*The R2 value doesn't matter, because it is not cor(Y, X). Instead, it is cor(Ybar, X). (Ybar is substituted for Y in their analysis at each of 31 values of X, that is, cancer incidences are averages).Their R2 could be 0.9999 while cor(Y,X) could be 0.0001.*

*Also, you can have R2=1 while the slope is barely off zero.*”

Please justify if this statement is correct. [For example, you can simulate a toy data. Other

options are permitted.]

I simulated a dataset where there were 6 levels of the independent variable X, namely 0, 10, 20, 30 ,40, 50, 60, with 10 repeated observations within each level (e.g. ten 0’s. ten 10’, etc), and Y had a relation with x of the form Y = b0 + b1\*X + error, where the error was normal (0, 300), and b0=b1=20.

I ran a regression model on this simulated dataset, with the following results:

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 53.253 66.641 0.799 0.427

x 18.647 2.201 8.472 9.97e-12 \*\*\*

And an R2 = 0.5531

Then I substituted Ybar for Y for each level of X, and calculated the cor(X, Ybar) and squared this value to get R2= 0.96. This a much larger value than cor(X,Y).

With this in mind, I would agree with the author on the part regarding “*Their R2 could be 0.9999 while cor(Y,X) could be 0.0001”*, as in my simulated data the R2 obtained from using Ybar was much larger the one obtained using Y.

I would presume that full sentence would be then: “*in this particular paper* *the R2 value doesn't matter…*”, because the matter of the fact is that it does (if not using Ybar). Again, I’m maybe reading between the lines here, but when I first read that part, the message I got was that the statistician was saying we shouldn’t worry about R2 because they weren’t expressing the cor(Y,X).

Regarding this part now: *“Also, you can have R2=1 while the slope is barely off zero”*

Here’s the (simulation) prove to that:

I simulated a dataset where the independent variable was Normal(10, 0.5) and dependent was determined by the formula Y = b0 + b1\*X + error, where b0=10, b1=0.01 (virtually zero), and error=0. When Y was regressed on X I got and R2=1, but with a slope of only 0.01, as showed below:

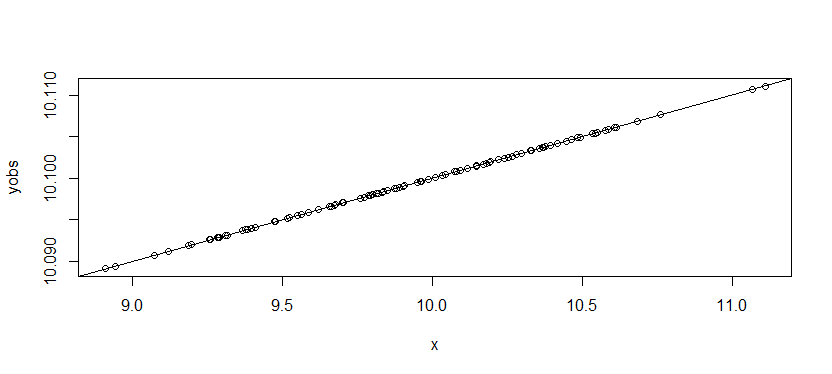
Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 1.000e+01 2.006e-14 4.985e+14 <2e-16 \*\*\*

x 1.000e-02 2.018e-15 4.956e+12 <2e-16 \*\*\*

R-squared: 1



So, the slope could be very low, but still have a large R2 (even 1 as above), provided the variance of Y conditional on X is very small, or zero as above (i.e. the distribution of Y is completely specified by the distribution of X). Likewise, large slopes can come with low R2, provided the variance of Y given X is large.