Final Exam Biostats Methods II

James Cutler

Part I

See exam document.

Part II

See exam document.

Part III

I wasn't sure if you wanted me to copy and paste things to the word document, or if there was just rhetorical language carried over from the earlier parts of the exam, so in this part (and in part IV) I transitioned to having everything output in my own document.

1. Distribution of salary by sex:

		Sex		N		Mean	SD		Min	Q1	Median	Q3	Max
1	Salary	female		14		21357.14	6151.87		15000	16686	20495	24900	38045
1.1		male		38		24696.79	5646.41		16094	20525	24746	28200	36350

2. Distribution of academic rank by sex (Table 1):

Table 1: Academic rank by sex

	assistant (N=18)	associate (N=14)	full (N=20)	Total (N=52)	p value
Sex					0.109
female	8 (44.4%)	2 (14.3%)	4 (20.0%)	14 (26.9%)	
male	10 (55.6%)	12 (85.7%)	16 (80.0%)	38 (73.1%)	

3. Crude (unadjusted) difference in mean salary between male and female professors (see Figure 1).

Table 2: ANOVA two-way complete model for sex and rank

	Sum Sq	Df	F value	Pr(>F)
factor(Sex)	7074743	1	0.7590047	0.3881644
factor(Rank)	1239752324	2	66.5026156	0.0000000
factor(Sex):factor(Rank)	3101661	2	0.1663789	0.8472326
Residuals	428769653	46	NA	NA

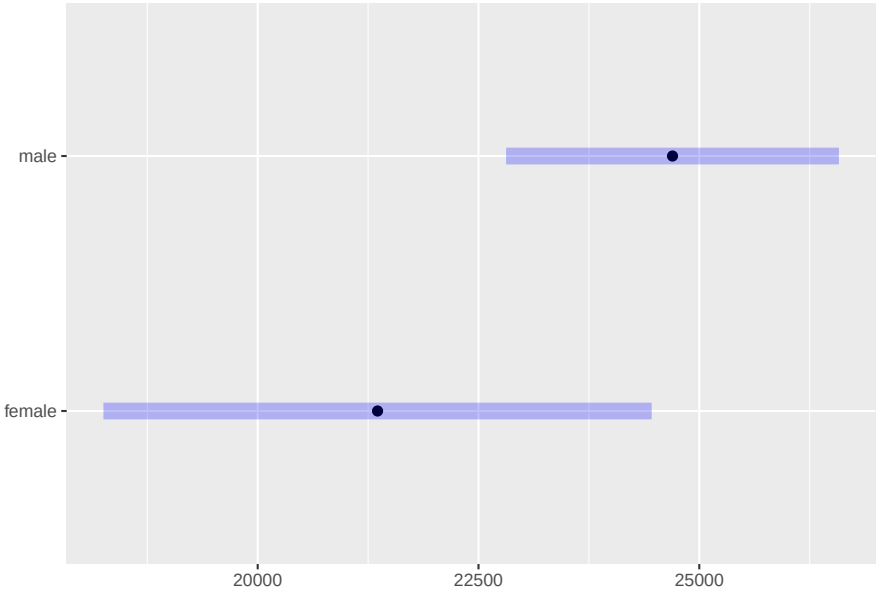


Figure 1: Means and 95% CIs for salary by sex (unadjusted)

4. An interaction is when the effect of one of the predictor variables on the outcome variable depends on the level of one of the other predictor variables. In this situation, the effect of the two predictor variables in question is not additive. My analysis of interaction between sex and academic rank with two-way complete ANOVA is shown in Table 2. This model does not show a significant relationship between rank and sex ($p = .847$). To aid the understanding of the reader is an interaction plot (Figure 2a), and a barplot of the academic ranks stacked by gender (Figure 2b).

Figure 2a: Interaction plot for sex and rank

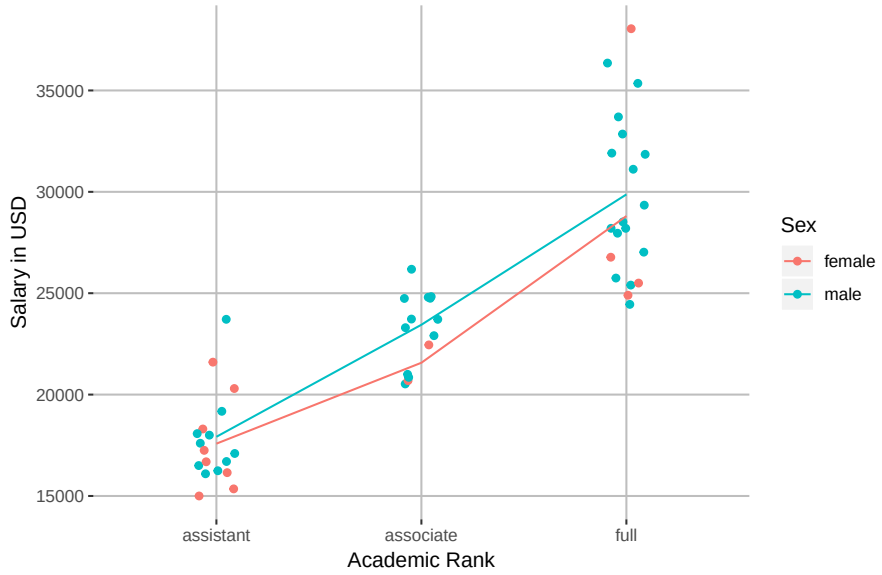
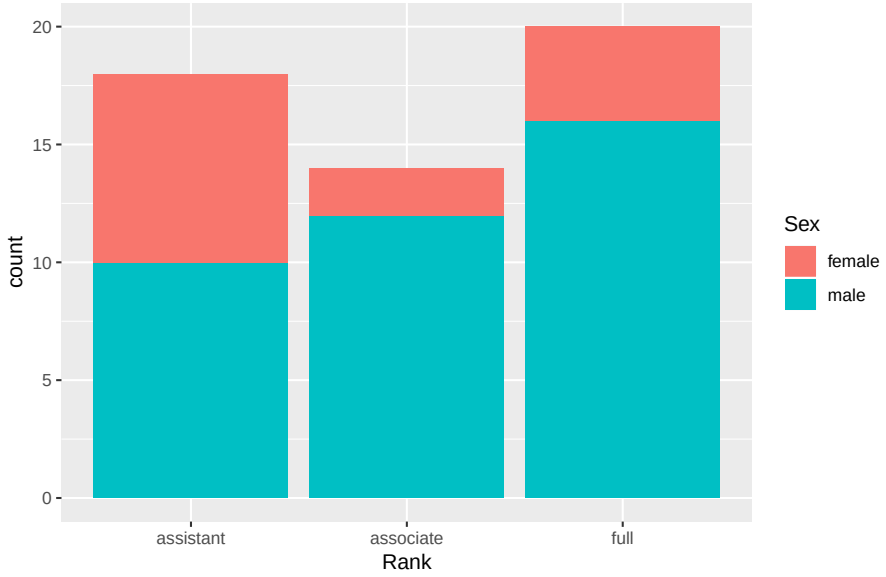


Figure 2b:Proportion of males and females in each rank



5. An appropriately adjusted estimate of the difference in mean salary between male and female professors, in my mind, should adjust for rank, years in rank, and years since degree (degree itself doesn't appear to have any significant effect on salary). In order to adjust for these variables, I decided to create arbitrary categorical variables out of years in rank and years since degree, after analyzing plots of these variables (see Figures 2a, 6, 8). The categories I created for years in rank are 0-4 years, 5-10 years, and over 10 years. The categories I created for years since degree are 0-9 years, 10-19 years, and 20 years and over. I found no interactions in a full/complete ANOVA model, so I have gone with a main effects ANOVA type II sums of squares model here (Table 3).

Adjusting for rank: Figure 3 (for why it might be reasonable to adjust for rank, refer back to Figure 2a). These estimates of the adjusted means are somewhat suspicious to me, since men

Table 3: Main effects ANOVA model of salary by three variables: rank, years in rank, and years since degree

	Sum Sq	Df	F value	Pr(>F)
Sex	187189.3	1	0.0251234	0.8747853
Rank	426746925.8	2	28.6376284	0.0000000
yr.rank.cat	1174999.0	2	0.0788504	0.9243084
yr.dg.cat	71958187.8	2	4.8288850	0.0127069
Residuals	327835540.1	44	NA	NA

don't appear to be earning ever so slight less in these categories, based on the plot. I'm not sure why the estimates turned out the way they do in Figure 3. This was not expected.

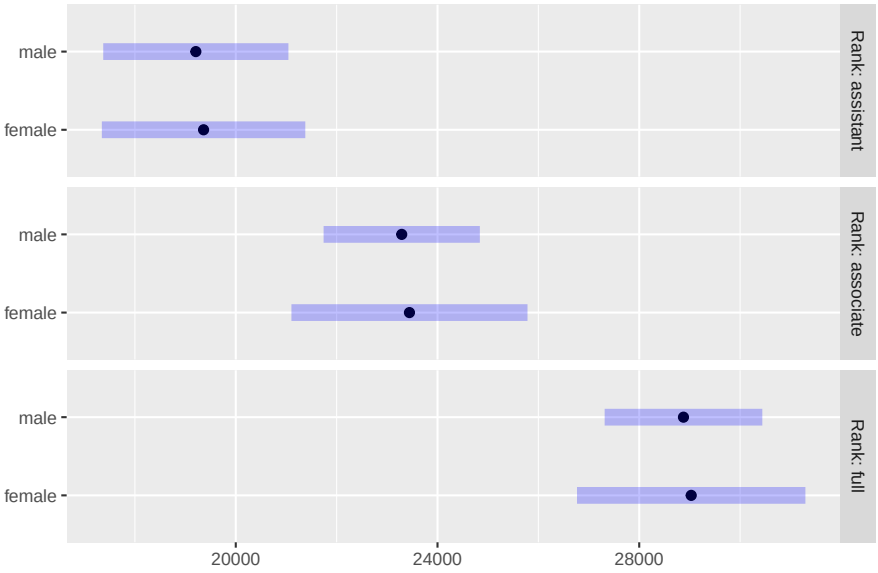


Figure 3: Means and 95% CIs for salary by sex, adjusted for rank

Adjusting for years in rank: Figure 5 (for why it might seem reasonable to adjust for years in rank, see Figure 6). As before, I am still not sure why men are shown as earning very slightly less than women.

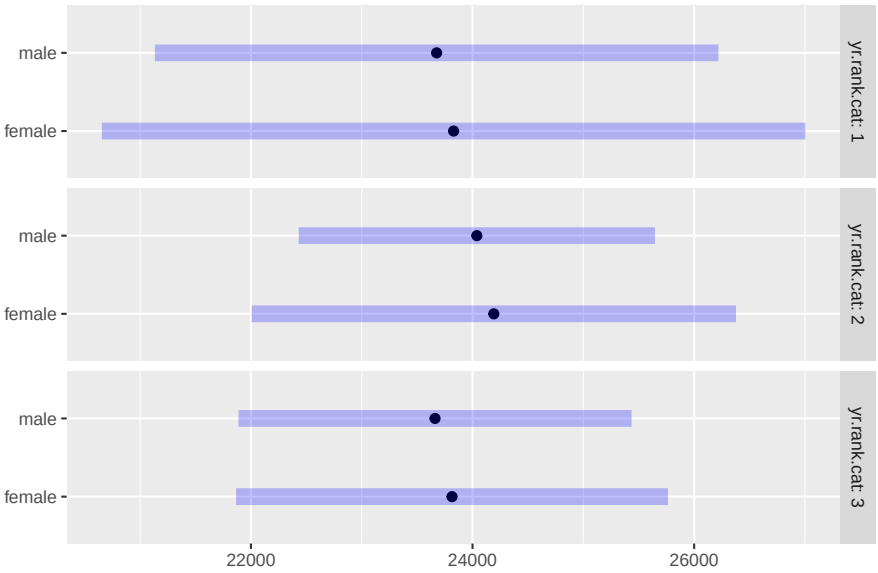
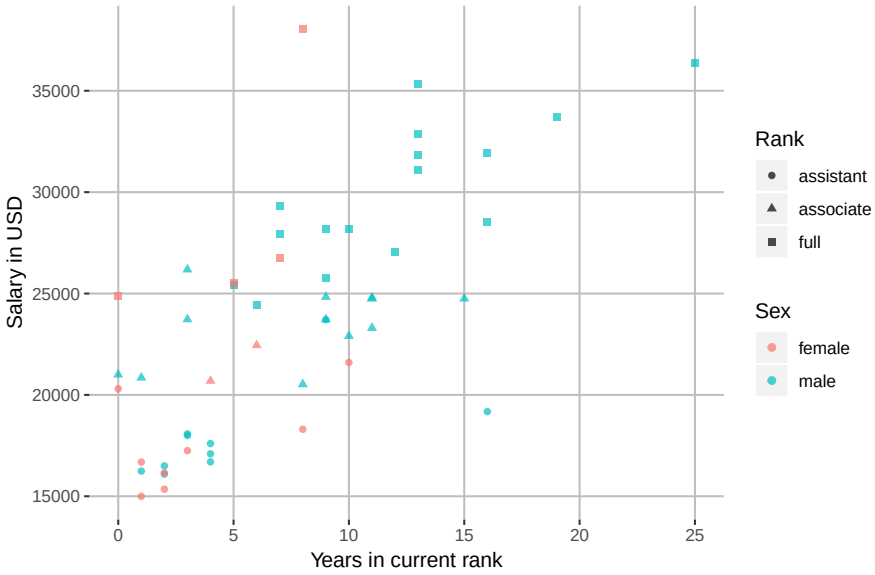


Figure 5: Means and 95% CIs for salary by sex, adjusted for years in rank

Figure 6: Does gender vary with years in rank?



And adjusting for years since degree: Figure 7 (for why it might be reasonable to adjust for years since degree, see Figure 8). As with the past two adjusted means estimates, I also can't say I know why these are similarly seemingly backwards, with males' salaries being very slightly lower.

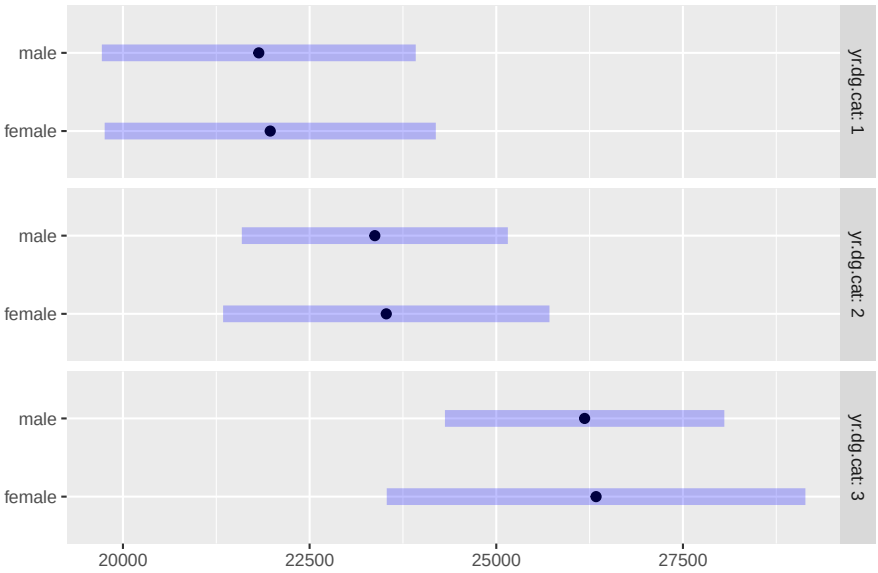
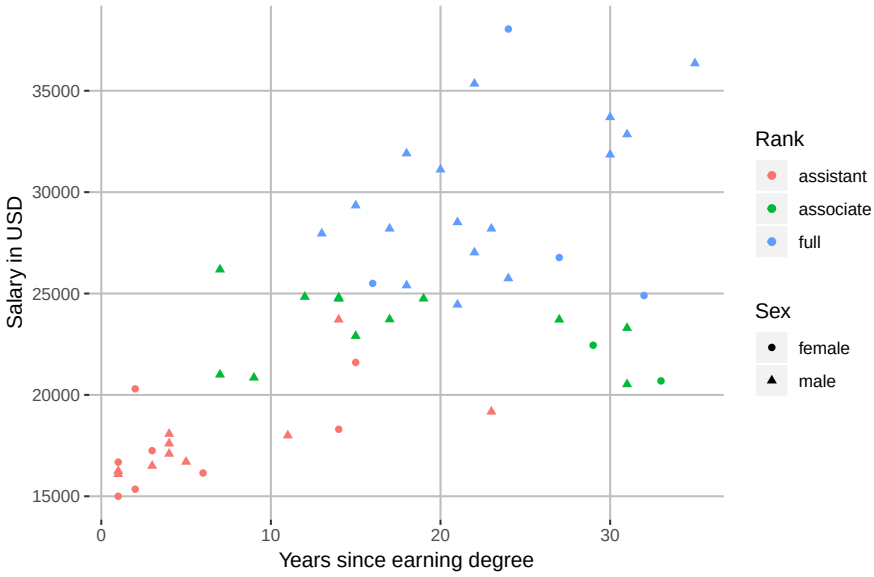


Figure 7: Means and 95% CIs for salary by sex, adjusted for years since degree

Figure 8: Do rank and gender vary with years since degree?



Part IV

1. If we treat time as a continuous variable, then it will be a straightforward linear regression model, and we will get regression coefficients for time and for group (one for each group). If, however, time is treated as a categorical variable, then we'll be able to see individual estimates of change by time (=2, =3, =4), as well as an estimate for group 2 in relation to the intercept. So, basically, these choices will lead to somewhat different model outputs. As an aside, ANOVA does not seem justifiable in this case since regarding the time intervals as individual groups would violate the assumption of independence of the subjects in each group from those in the other groups.

2. If, over time, we see that one group (say, the treatment group) does rise in hormone concentration significantly, while the other group does not, then that would indicate an

interaction between time and type of agent administered.

3. Time is a repeated factor/variable in this analysis because I imagine the intent is to see if the hormone levels change between before and after administration of the treatment, and to capture multiple snapshots following treatment, in case the earliest one doesn't capture the effect, or perhaps in case there is a dropoff in hormone concentration following an initial rise. Either way, repeated measures at different intervals after treatment will be more likely to pick that up. Measuring at a before time and an after time will make it at least possible to tell if there was a change. Recording hormone levels at only one time would leave you without a way to tell if the treatment made a difference or not compared to the placebo because it could be that the placebo group was already low, and the treatment was higher, to start out with, and the treatment could have done nothing and it would still give the appearance of working in such a scenario.

4. As long as animals are randomly assigned to either the placebo or the treatment group, then this between subject factor should still enable us to tell if the treatment works or not. Any differences among the animals to start out with would ideally be randomly distributed between the groups so as to neutralize the effects of each other in comparison with the other group, leaving only the treatment and placebo to be reasonably responsible for any difference in outcome. This would justify using a between subject factor.

5. PROC MIXED advantages over PROC GLM: Since I do not use SAS for this class, I would bet that the most important advantage to gain in choice of methodology in this case would be to choose a method that would allow for repeated measures analysis, or mixed models.

6. Statistical model (with time as a categorical variable) for testing whether the pattern of change over time differs between the two groups (see Table 4).

7. Estimates for the mean concentration at each time point for both groups are in Tables 5 and 6.

8. Using my model's results to comment on the group's comparability prior to the intervention, the estimate (with 95% CI) for the means of group 1 and 2 at time 1 are 26.36 mg/dl (20.35, 29.44) and 28.40 mg/dl (25.23, 33.89), respectively. Looking at the CIs here, there is a good deal of overlap (this will be clear in the plot below).

9. Which is the first time at which the two groups of animals differ in the mean hormone concentration? What is the between-group difference? The answer is that at no time do the CIs of the two groups ever distance themselves from each other widely enough to leave no overlap between them. There is a point at which the CIs of neither group overlap the means of the other: That's at time 4. The difference between the two means is 9.61. First, I would like to show a graph of the data without CIs, then CIs below it, to reinforce what I'm saying:

Table 4: Linear model of hormone levels with time, by group

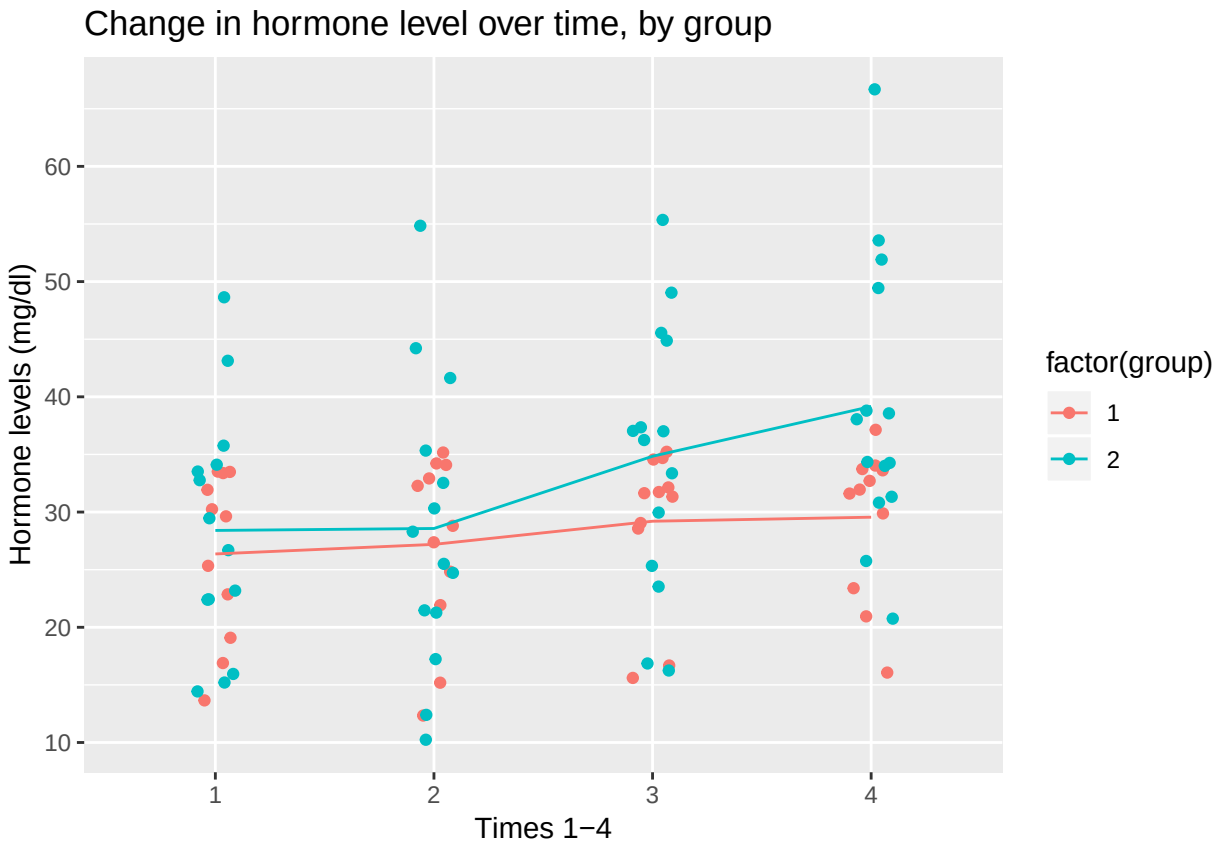
	Dependent variable:
	conc
factor(time)2	0.459
	(2.819)
factor(time)3	4.853*
	(2.819)
factor(time)4	7.428***
	(2.819)
factor(group)2	4.668**
	(2.008)
Constant	24.893***
	(2.289)
Observations	100
R^2	0.137
Adjusted R^2	0.100
Residual Std. Error	9.967 (df = 95)
F Statistic	3.765*** (df = 4; 95)
Note:	*p<0.1; **p<0.05; ***p<0.01

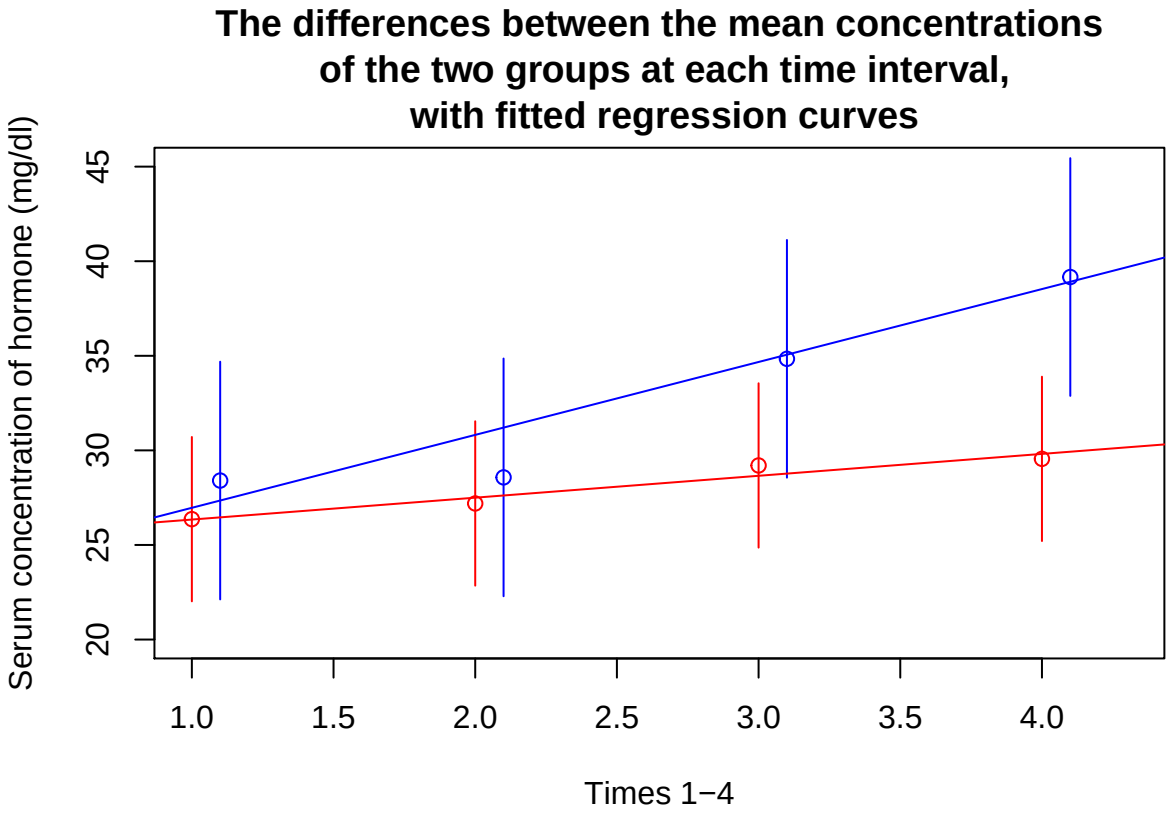
Table 5: LS means group 1 with 95 percent CI

time	lsmean	SE	df	lower.CL	upper.CL
1	26.36399	2.149613	40	22.01946	30.70852
2	27.19336	2.149613	40	22.84883	31.53789
3	29.20308	2.149613	40	24.85855	33.54761
4	29.55180	2.149613	40	25.20727	33.89633

Table 6: LS means group 2 with 95 percent CI

time	lsmean	SE	df	lower.CL	upper.CL
1	28.40486	3.13126	52	22.12153	34.68820
2	28.57297	3.13126	52	22.28963	34.85630
3	34.84041	3.13126	52	28.55707	41.12374
4	39.16414	3.13126	52	32.88081	45.44748





10. In group 1, did the hormone levels change between baseline and any of the other three

times of measurement? Judging by the linear model in Table 7, there does not appear to be any significant change in hormone levels after baseline.

Table 7: Linear model of group 1 hormone levels by time

	Dependent variable:
	conc
factor(time)2	0.829
	(3.040)
factor(time)3	2.839
	(3.040)
factor(time)4	3.188
	(3.040)
Constant	26.364***
	(2.150)
Observations	44
R^2	0.037
Adjusted R^2	−0.035
Residual Std. Error	7.129 (df = 40)
F Statistic	0.516 (df = 3; 40)
Note:	*p<0.1; **p<0.05; ***p<0.01

11. A) What is the advantage of using an unstructured covariance matrix? I believe the advantage is that you are not imposing any constraints on the values in your matrix. In our case, where the variances are not really that equal, this is a benefit and should give us the best model fit, because each of the values in the matrix will be closer to the data.

B) The value in the first row and first column of the [covariance] (I'm not sure which the R matrix is) matrix: 1.2626. The population parameter it estimates is the variance in time.

```
##           time     group       conc
## time  1.262626 0.0000000   3.368376
## group 0.000000 0.2488889   1.161698
## conc  3.368376 1.1616982 110.427721
```

C) The value in the third row and third column is 110.4277, and estimates the variance in hormon concentration, in mg/dl.

D) The value in the first row and third column is 3.3684, and estimates the covariance

between time and concentration. That is, with every increasing interval in time, there is a 3.37 mg/dl increase in hormone.

E) The appearance of the covariance matrix I'm seeing is symmetric in that it has the same headings going left to right across the top as it does top to bottom down the left side. This means some of the values–the ones situated diagonally from each other in a bottom left to upper right, rather than a lower right to upper left fashion–are identical to one another.

Bonus question: Response to "The question is, can these two deaths really be treated as independent? Two SIDS deaths occurred in the same home under the supervision of the same two parents. How likely is it that those are unrelated events?"

My response is that these two SIDS events probably should not be treated as independent. If an asteroid had struck the Clarks' home, and they had rebuilt, and then years later, another asteroid had struck their new home, then that would be a case of independent events. But there is a very important causal variable linking the two SIDS deaths–both babies come from the same parents. They inherit much of the same health/disease propensities, and they also are under the supervision of the same parents. Many "SIDS" deaths are only reported as such in order to allay the guilt of the parents that they might not have been practicing safe sleep guidelines as well as they could have. Many parents don't even know what those guidelines are. Often, SIDS could just be suffocation because the baby wasn't positioned on their back on a flat bare surface. This is why the shared supervision is an important element militating against these deaths being unrelated.

The consequence of misusing statistics can be severe, as illustrated by this story. It can affect everything from the general quality of medical literature (see John Ioannidis, for example), to being misled by the statistics people sometimes deliberately use to distort another's perception of reality, either to take risks he or she would not take in their better judgment, or to reach any unsound conclusion about something that could end up harming others. That is why I love statistics. It is vital to our ability to avoid unnecessary waste and harm, as well as to our ability to move closer towards optimization in so many areas of life and society.

I have adhered to the University's standards on academic integrity, have neither given nor received assistance during this examination, and have used no materials that the instructor did not specifically permit. - James Cutler