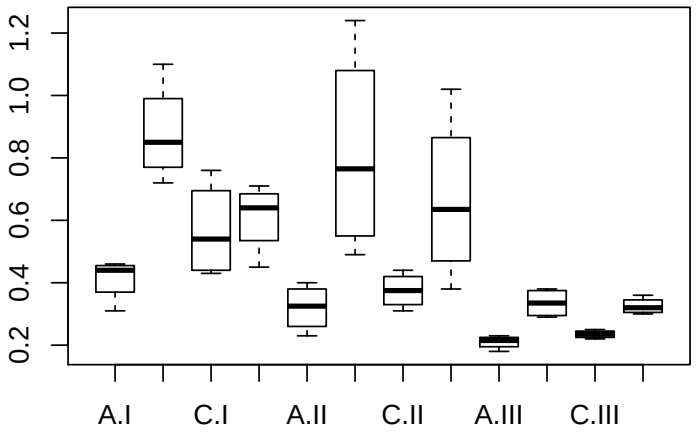
Stat 505 Assignment 3

Leslie Gains-Germain

We will use data from a designed experiment in which rats were exposed to one of three chemicals and given one of four treatments. The response is their survival time.

1. Use boxplots to view the data at each combination of treatment and chemical. Do there appear to be some interactions? Are spreads similar in each combination?



See the boxplot above. The spreads are not similar at each combination. In general, the spread is smaller for combinations with lower average survival times. There does appear to be an interaction between treatment and chemical because the treatment effect changes across levels of the chemical. When the chemical is at level 1*, the increase in survival time between treatments A and B is much larger than the increase in surival time between A and B when the chemical is at level* 3*.*

2. Without fitting a model, build the matrix $\boldsymbol{X}_1$ for a main effects model using just the 12 unique rows of data, and label each with chemical and treatment.

Give the rank of this matrix.

$$y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk},\ i = 1, \ldots, 4, j = 1, \ldots, 3,\ k = 1, \ldots, 4$$

The rank of the matrix below is 6.

```
##      chem2I chem2II chem2III trt2A trt2B trt2C trt2D
## 1A 1      1       0        0     1     0     0     0
## 1B 1      1       0        0     0     1     0     0
## 1C 1      1       0        0     0     0     1     0
## 1D 1      1       0        0     0     0     0     1
## 2A 1      0       1        0     1     0     0     0
## 2B 1      0       1        0     0     1     0     0
## 2C 1      0       1        0     0     0     1     0
## 2D 1      0       1        0     0     0     0     1
## 3A 1      0       0        1     1     0     0     0
## 3B 1      0       0        1     0     1     0     0
## 3C 1      0       0        1     0     0     1     0
## 3D 1      0       0        1     0     0     0     1
```

3. Again, without fitting, build the additional columns needed for a full interaction model. We'll call this one $\boldsymbol{X}_2$. Label each row and give its rank. Is $\boldsymbol{X}_1$ contained in the columnspace of $\boldsymbol{X}_2$? If not, demonstrate why not, if so, find a matrix to multiply by $\boldsymbol{X}_2$ to get $\boldsymbol{X}_1$.

   The rank of $\boldsymbol{X}_2$ *is* 12. $\boldsymbol{X}_1$ *is contained in the columnspace of* $\boldsymbol{X}_2$. *We simply multiply* $\boldsymbol{X}_2$ *by* $\boldsymbol{X}_1$ *and we get* $\boldsymbol{X}_1$ *back.*

```
##    [,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10] [,11] [,12]
## 1A    1    0    0    0    0    0    0    0    0     0     0     0
## 1B    0    1    0    0    0    0    0    0    0     0     0     0
## 1C    0    0    1    0    0    0    0    0    0     0     0     0
## 1D    0    0    0    1    0    0    0    0    0     0     0     0
## 2A    0    0    0    0    1    0    0    0    0     0     0     0
## 2B    0    0    0    0    0    1    0    0    0     0     0     0
## 2C    0    0    0    0    0    0    1    0    0     0     0     0
## 2D    0    0    0    0    0    0    0    1    0     0     0     0
## 3A    0    0    0    0    0    0    0    0    1     0     0     0
```

```
## 3B    0    0    0    0    0    0    0    0    0    1    0    0
## 3C    0    0    0    0    0    0    0    0    0    0    1    0
## 3D    0    0    0    0    0    0    0    0    0    0    0    1
```

4. What is the rank of combined matrix $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$?

   The rank of the combined matrix $\boldsymbol{X}$ *is* 12.

5. How many columns must we remove from $\boldsymbol{X}$ to get a full column rank matrix? It does matter which columns we remove Explain at least two choices for removal which still allow us to estimate all cell means, and one which does not work. For the non-working one, what is the rank of the remaining columns? Explain how this relates to the information in the class notes on p 17 about non-estimable constraints.

   We must remove 8 *columns from* $\boldsymbol{X}$ *to get a full column rank matrix. It does matter which columns we remove. We could remove the "chem1" and "trtA" columns from* $\boldsymbol{X}_1$*, and the first six columns of* $\boldsymbol{X}_2$*. Or, we could remove the "chem3" and "trtD" columns for* $\boldsymbol{X}_1$*, and the last six columns of* $\boldsymbol{X}_2$*. It would not work to remove the "chem1" and "chem2" columns from* $\boldsymbol{X}_1$ *and the first six columns of* $\boldsymbol{X}_2$*. If we did that, the rank of the remaining columns would be* 11*. The whole reason we apply constraints, as you explain on page* 17 *of the notes, is so we can "obtain a reparametrized design matrix,* $\boldsymbol{X}^*$*, which is of full column rank". We apply constraints to make* $\mathbf{X}$ *full column rank so that* $\boldsymbol{X}^T\boldsymbol{X}$ *is invertible and we can find least squares estimates for* $\hat{\beta}$*. Although the constraints are not unique, we do need to choose our constraints appropriately so that the constrained* $\boldsymbol{X}$ *matrix is of full column rank.*

6. Use the Moore-Penrose generalized inverse to estimate cell means (you'll have to scale up to use all the data now).

```
##       [,1]  [,2]  [,3]    [,4]   [,5] [,6]    [,7]  [,8]    [,9]
## [1,] 0.288 0.206 0.148 -0.0666 -0.052 0.22 0.00675 0.113 -0.0295
##        [,10] [,11] [,12] [,13]  [,14]  [,15]   [,16]   [,17] [,18]
## [1,] -0.0635 0.041 0.166  0.16 -0.106 0.0668 -0.0673 0.00725 0.003
##      [,19]  [,20]
## [1,] 0.119 -0.009
```

The output above shows our parameter estimates. The output below shows the estimates of our cell means.

```
##  [1] 0.413 0.320 0.210 0.880 0.815 0.335 0.568 0.375 0.235 0.610 0.667
## [12] 0.325
```

7. Demonstrate that this estimate differs from the usual interaction model fit by `lm`, and that cell mean estimates are the same.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.4125	0.0746	5.53	0.0000
chemII	-0.0925	0.1055	-0.88	0.3862
chemIII	-0.2025	0.1055	-1.92	0.0628
trtB	0.4675	0.1055	4.43	0.0001
trtC	0.1550	0.1055	1.47	0.1503
trtD	0.1975	0.1055	1.87	0.0692
chemII:trtB	0.0275	0.1491	0.18	0.8547
chemIII:trtB	-0.3425	0.1491	-2.30	0.0276
chemII:trtC	-0.1000	0.1491	-0.67	0.5068
chemIII:trtC	-0.1300	0.1491	-0.87	0.3892
chemII:trtD	0.1500	0.1491	1.01	0.3212
chemIII:trtD	-0.0825	0.1491	-0.55	0.5836

Above are the parameter estimates from the interaction model fit with lm. We can see that the parameter estimates differ from those above. The cell means (printed below) are the same.

```
##     1     5     9    13    17    21    25    29    33    37    41    45
## 0.412 0.320 0.210 0.880 0.815 0.335 0.568 0.375 0.235 0.610 0.667 0.325
```

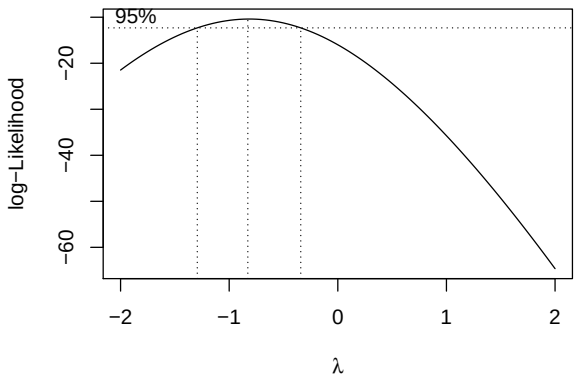
Report on Survival of Rats

Introduction

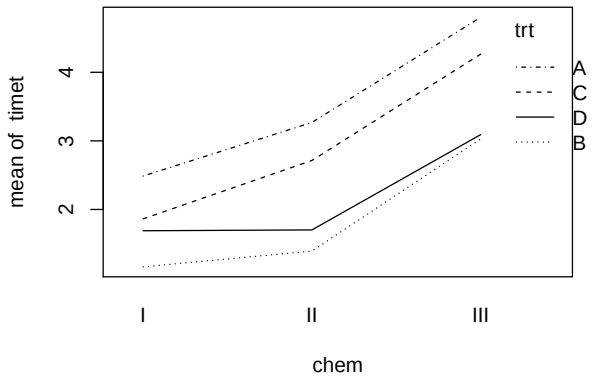
Rats were exposed to one of three chemicals and one of four treatments. Their survival times were recorded.

Statistical Procedures

I first looked at the boxplots of survival rates for each treatment combination. As I mentioned above, I saw that those groups with smaller average survival rates also had smaller spread. This pattern indicates that a transformation of the response variable may be a appropriate, so I did a boxcox procedure and chose to model the reciprocals of the survival times.

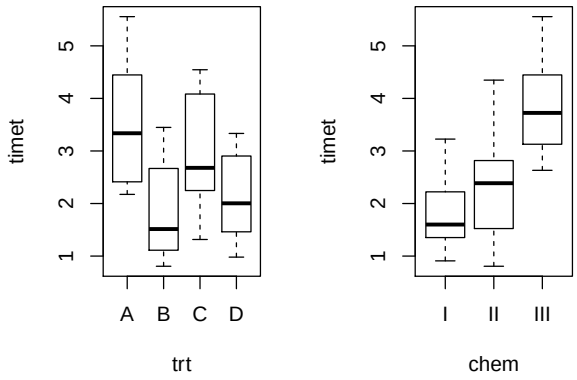


I then looked at the interaction plot of mean reciprocal survival times. The lines in the interaction plot are mostly parallel, so it appears that the change in the mean reciprocal survival times between treatments does not depend on the level of the chemical. Just to be sure, however, I fit the interaction model. Before using this model for inference, I checked the diagnostic plots and found no problems with the assumptions. As I suspected, I found no evidence of a two-factor interaction between treatment and chemical (p-value= .3867 *from F-stat=*1.09 *on* 6 *and* 36 *df).*

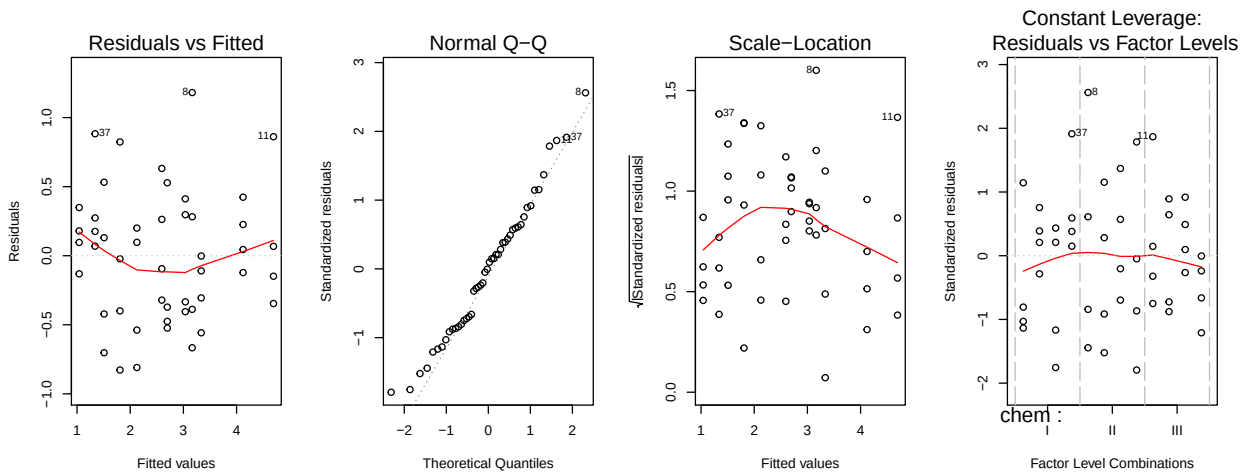


	Df	Sum Sq	Mean Sq	F value	Pr(>F)
chem	2	34.88	17.44	72.63	0.0000
trt	3	20.41	6.80	28.34	0.0000
chem:trt	6	1.57	0.26	1.09	0.3867
Residuals	36	8.64	0.24		

Note that in the plots below I do not see a linear relationship in the reciprocal survival times across levels of either explanatory variable. For this reason, I am choosing to treat chemical and treatment both as factors.



I chose to use the additive model for inference. The diagnostic plots are shown below. Again, I see no problems with the model assumptions.



Below is the additive model summary and the ANOVA.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.6977	0.1744	15.47	0.0000
chemII	0.4686	0.1744	2.69	0.0103
chemIII	1.9964	0.1744	11.45	0.0000
trtB	-1.6574	0.2013	-8.23	0.0000
trtC	-0.5721	0.2013	-2.84	0.0069
trtD	-1.3583	0.2013	-6.75	0.0000

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
chem	2	34.88	17.44	71.71	0.0000
trt	3	20.41	6.80	27.98	0.0000
Residuals	42	10.21	0.24		

Summary of Findings

There is strong evidence that the mean reciprocal survival time of these rats depends on type of chemical (p-value$<$ 0.0001 *from F-stat*=71.71 *on* 2 *and* 42 *df). The mean reciprocal survival time for rats receiving chemical I and treatment A is estimated to be* 2.70 *with an estimated* 95% *confidence interval from* 2.35 *to* 3.05*. At a fixed treatment level, the mean reciprocal survival time is estimated to increase by* 0.4686 *for a change from chemical I to chemical II, with an associated* 95% *confidence interval from* 0.117 *to* 0.820*. The mean reciprocal survival time is estimated to increase by* 2.0 *for a change from chemical I to chemical III, with a* 95% *confidence interval from* 1.64 *to* 2.35*.*

There is also strong evidence that the mean reciprocal survival time of these rats depends on treatment (p-value$<$ 0.0001 *from F-stat*=27.98 *on* 3 *and* 42 *df). At a fixed chemical level,*

the mean reciprocal survival time is estimated to decrease by 1.66 *for a change in treatment from A to B, with a* 95% *confidence interval from a* 2.06 *to a* 1.25 *decrease. The mean reciprocal survival time is estimated to decrease by* 0.57 *for a change in treatment from A to C, with a* 95% *confidence interval from a* 0.98 *to a* 0.16 *decrease. The mean reciprocal survival time is estimated to decrease by* 1.36 *for a change in treatment from A to D, with a* 95% *confidence interval from a* 1.76 *to a* 0.95 *decrease.*

Note that the mean reciprocal survival time can be thought of as a speed to death so that the smaller the mean reciprocal survival time (speed to death), the longer the survival time.

Scope of Inference

Rats were not randomly selected from a larger population of rats, so inference does not extend beyond the rats in the study.

As far as we know, rats were not randomly assigned to type of chemical or treatment level, so we cannot establish a causal relationship between treatments and survival times.

R Code

```
survival <- read.csv("http://www.math.montana.edu/~jimrc/classes/stat505/data/ratSurvival.csv")
#names(survival)
boxplot(survival$sTime~survival$trt+survival$chem)
```

```
chem2 <- c(rep("I",4), rep("II",4), rep("III",4))
trt2 <- c("A", "B", "C", "D", "A", "B", "C", "D", "A", "B", "C", "D")
Time <- survival$sTime[1:12]
lm.1 <- lm(Time~chem2+0)
lm.2 <- lm(Time~trt2+0)
#model.matrix(lm.1)
x1 <- cbind(1,model.matrix(lm.1),model.matrix(lm.2))
rownames(x1) <- c("1A", "1B", "1C", "1D", "2A", "2B", "2C", "2D", "3A", "3B", "3C", "3D")
x1
```

```
x2 <- diag(12)
rownames(x2) <- c("1A", "1B", "1C", "1D", "2A", "2B", "2C", "2D", "3A", "3B", "3C", "3D")
x2
```

```
lm.3 <- lm(survival$sTime~survival$chem+0)
lm.4 <- lm(survival$sTime~survival$trt+0)
#model.matrix(lm.1)
x3 <- cbind(1,model.matrix(lm.3),model.matrix(lm.4))
a1 <- c(rep(1,4),rep(0,44))
a2 <- c(rep(0,4),rep(1,4),rep(0,40))
a3 <- c(rep(0,8),rep(1,4),rep(0,36))
a4 <- c(rep(0,12),rep(1,4),rep(0,32))
a5 <- c(rep(0,16),rep(1,4),rep(0,28))
```

```
a6 <- c(rep(0,20),rep(1,4),rep(0,24))
a7 <- c(rep(0,24),rep(1,4),rep(0,20))
a8 <- c(rep(0,28),rep(1,4),rep(0,16))
a9 <- c(rep(0,32),rep(1,4),rep(0,12))
a10 <- c(rep(0,36),rep(1,4),rep(0,8))
a11 <- c(rep(0,40),rep(1,4),rep(0,4))
a12 <- c(rep(0,44),rep(1,4))
a <- cbind(a1,a2,a3,a4,a5,a6,a7,a8,a9,a10,a11,a12)
x6 <- cbind(x3,a)
#lm.fit1 <- lm(sTime~chem*trt+0, data=survival)
#m <- cbind(1,model.matrix(lm.fit1))
require(MASS)
paramestimates <- ginv(crossprod(x6))%*%t(x6)%*%survival$sTime
cellmeans <- x6%*%ginv(crossprod(x6))%*%t(x6)%*%survival$sTime
t(paramestimates)
```

```
cellmeans[seq(from=1,to=48,by=4)]
```

```
lm.int <- lm(sTime~chem*trt, data=survival)
require(xtable)
xtable(summary(lm.int))
```

```
fitted(lm.int)[seq(1,48,4)]
```

```
require(MASS)
fit.boxcox <- lm(sTime ~ chem * trt, data = survival)
boxcox(fit.boxcox)
timet <- survival$sTime^-1
fit.trans <- lm(timet ~ chem*trt, data=survival)
fit.add <- lm(timet~chem+trt, data=survival)
```

```
par(mfrow=c(1,2))
plot(timet~trt, data=survival)
plot(timet~chem, data=survival)
```

```
require(xtable)
xtable(anova(fit.trans))
```

```
par(mfrow=c(1,2))
plot(timet~trt, data=survival)
plot(timet~chem, data=survival)
```

```
par(mfrow=c(1,4))
plot(fit.add)
```

```
require(xtable)
xtable(summary(fit.add))
xtable(anova(fit.add))
```