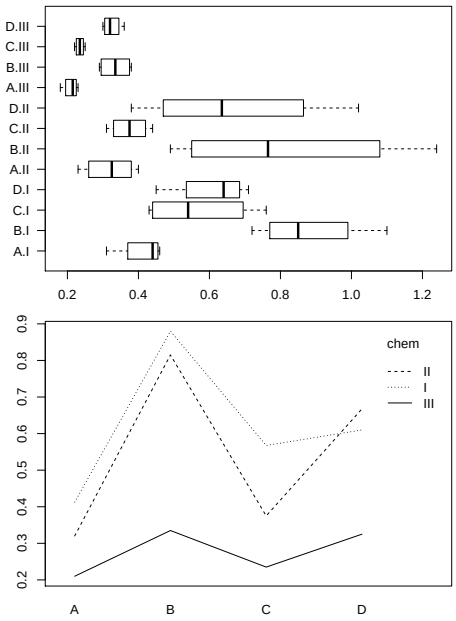
Stat 505 Assignment 3

Solutions

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? (4 pts)



The spreads vary widely from treatment:B, chemical:II with IQR = 0.42 to treatment:A, Chemical III with IQR = 0.20. Interactions seem to be weak, though lines are not always parallel. Both drug and treatment appear to have substantial effects on the means.

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. (3 pts)

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

The matrix has 8 columns, but it is only rank 6. (One for μ, three for treatment and 2 for chemical).

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 $\boldsymbol{X}_1$. (3 pts)

This matrix has rank 12. Depending on how you order the interactions, it's either an identity or a permutation of $\boldsymbol{I}$. We can build $\boldsymbol{X}_1$ from these columns. The vector of all ones is the sum of all these columns, then we sum three columns to get a treatment

	mu	t1	t2	t3	t4	c1	c2	c3
ChemA.TrtI	1	1	0	0	0	1	0	0
ChemB.TrtI	1	0	1	0	0	1	0	0
ChemC.TrtI	1	0	0	1	0	1	0	0
ChemD.TrtI	1	0	0	0	1	1	0	0
ChemA.TrtII	1	1	0	0	0	0	1	0
ChemB.TrtII	1	0	1	0	0	0	1	0
ChemC.TrtII	1	0	0	1	0	0	1	0
ChemD.TrtII	1	0	0	0	1	0	1	0
ChemA.TrtIII	1	1	0	0	0	0	0	1
ChemB.TrtIII	1	0	1	0	0	0	0	1
ChemC.TrtIII	1	0	0	1	0	0	0	1
ChemD.TrtIII	1	0	0	0	1	0	0	1

	1	2	3	4	5	6	7	8	9	10	11	12
ChemA.TrtI	1	0	0	0	0	0	0	0	0	0	0	0
ChemB.TrtI	0	1	0	0	0	0	0	0	0	0	0	0
ChemC.TrtI	0	0	1	0	0	0	0	0	0	0	0	0
ChemD.TrtI	0	0	0	1	0	0	0	0	0	0	0	0
ChemA.TrtII	0	0	0	0	1	0	0	0	0	0	0	0
ChemB.TrtII	0	0	0	0	0	1	0	0	0	0	0	0
ChemC.TrtII	0	0	0	0	0	0	1	0	0	0	0	0
ChemD.TrtII	0	0	0	0	0	0	0	1	0	0	0	0
ChemA.TrtIII	0	0	0	0	0	0	0	0	1	0	0	0
ChemB.TrtIII	0	0	0	0	0	0	0	0	0	1	0	0
ChemC.TrtIII	0	0	0	0	0	0	0	0	0	0	1	0
ChemD.TrtIII	0	0	0	0	0	0	0	0	0	0	0	1

column, and four columns to get a chemical vector. If you ordered columns of $\boldsymbol{X}_2$ into an identity, then $\boldsymbol{X}_2\boldsymbol{X}_1 = \boldsymbol{X}_1$. Otherwise, it's just a permutation of the rows of $\boldsymbol{X}_1$.

4. What is the rank of combined matrix $\boldsymbol{X} = [\boldsymbol{X}_1 \;\; \boldsymbol{X}_2]$? (1 pt)

   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. (4 pts)

   I am now using the model:

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

   We have to remove 8 columns to get down to a square 12 by 12 matrix which could be invertible. We've seen that we could remove the first 8, because $\boldsymbol{X}_2$ is full column rank. This would be like constraining $\mu = 0 = \alpha_i = \beta_j$ for all $i = 1, 2, 3, 4$ and $j = 1, 2, 3$. Another possibility is the one SAS chooses to force $\alpha_4 = \beta_3 = 0 = \gamma_{13} = \gamma_{23} = \gamma_{33} = \gamma_{43} = \gamma_{41} = \gamma_{42}$. Each of these is an non–estimable parameter, so setting all to 0 provides a K which has rank 12. To find 12 columns of $\boldsymbol{X}$ with rank less than 12, I looked at it's first 12 columns, and found they only have rank 9. The 8 columns I removed have rank 8, but they are columns for some of the interactions, and setting them to zero constrains

an estimable function. In removing the last four columns, I set $\gamma_{D3}-\gamma_{D2}+\gamma_{C2}-\gamma_{C3}=0$, *when it really is estimable.* $E(\ \bar{y}_{D3}-\ \bar{y}_{D2}+\ \bar{y}_{C2}-\ \bar{y}_{C3})=\gamma_{D2}+\gamma_{C2}-\gamma_{C3}$, *and that linear combination of cell means is* -0.2025.

6. Use the Moore-Penrose generalized inverse to estimate cell means (you'll have to scale up to use all the data now). (3 pts)

	1	2	3	4	5	6	7	8	9	10	11	12
M-P est	0.29	0.21	0.15	-0.07	-0.05	0.22	0.01	0.11	-0.03	-0.06	0.04	0.17
DummyCoef	0.41	0.00	-0.09	-0.20	0.00	0.47	0.15	0.20	0.00	0.00	0.00	0.00

	13	14	15	16	17	18	19	20
M-P est	0.16	-0.11	0.07	-0.07	0.01	0.00	0.12	-0.01
DummyCoef	0.03	-0.34	0.00	-0.10	-0.13	0.00	0.15	-0.08

7. Demonstrate that this estimate differs from the usual interaction model fit by `lm`, and that cell mean estimates are the same. (3 pts)

```
## Error:  object 'dbetahat' not found
```

The coefficient vector from lm is only of length 12, and none of the Moore-Penrose values are zero. To show inequality, I printed them above and used the `which` *statement to find all "dummy" coefficients which agreed with the M-P estimates to tolerance of .001, and none did. To show equality, I found all fitted values using both methods, subtracted, and used summary. The differences are all less than* `1.332e-15` *which is zero to machine accuracy.*

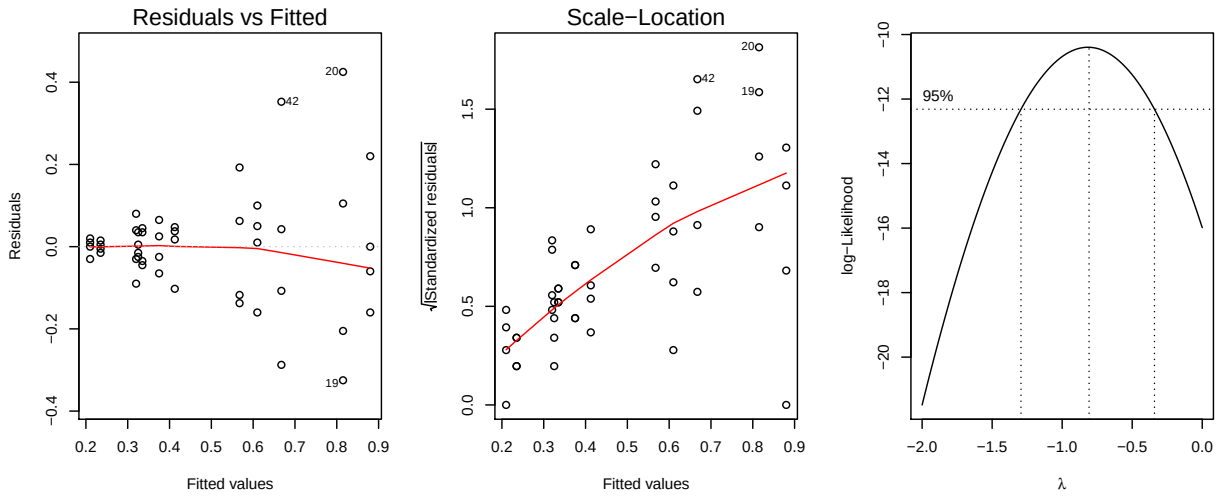
Report on Survival of Rats

(10 pts)

These data are a classic dataset used by Box and Cox in 1964 to demonstrate their transformation methods. Box refers to them in some of his design books, and tells us the response time is measured in 10 hour increments. I've not found which chemicals were used or what the treatments are.

In comparing the treatments, I would guess that researchers want to know which treatment combinations kill rats more quickly. I will suppose that they are looking for a quick and humane way to sacrifice rats used in research. In the original scale, the research question might be, "Which mean survival times are shortest, and have smallest variance?" From the plots above, chemical III provides smallest mean survival times, and generally treatment A is better than treatments B, C, or D.

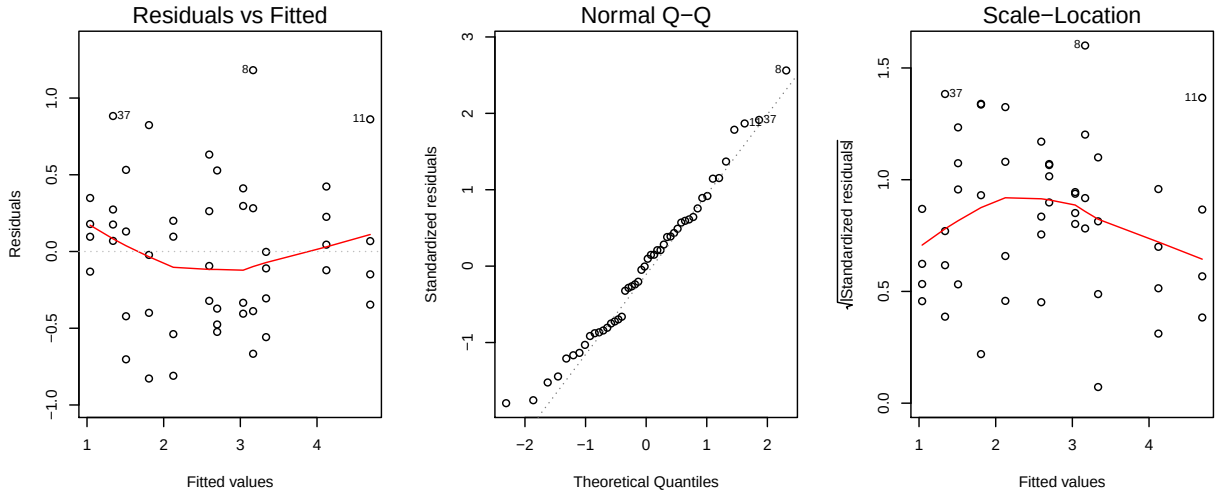
However, the problem of nonconstant variance we saw in the boxplots also exhibits itself in the residuals, so this is perfect data to illustrate the use of transformations. Here are two plots of the residuals versus fits which show increasing variance, and the Box–Cox plot which shows that a power of -1, or reciprocal transformation is approximately ideal for making residuals "normally distributed".



Applying this transformation, we get the following ANOVA table, which tells us that there is very strong evidence of a chemical effect and of a treatment effect, but evidence for an interaction is weak.

	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		

In order to simplify the report, I will omit the interaction terms. Due to the nicely balanced nature of these data, that does not change the sums of squares for the main effects, both of which are still quite strong. The interaction term is combined into the error line, giving it 6 more degrees of freedom.



The diagnostic plots indicate that the reciprocal transformation has done its job well. Variance is now stable across the fitted values, and residuals are nicely normal. (Leverage is

	Estimate	Std. Error	t value	Pr(>|t|)
chemI	2.6977	0.1744	15.47	0.0000
chemII	3.1663	0.1744	18.16	0.0000
chemIII	4.6941	0.1744	26.92	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

not of interest in a designed experiment, so I left off the 4th plot.)

Because I have opted for a reciprocal transformation, the question of interest is now "Do mean rates differ, and if so, which mean rates of death are greatest?". The means differ strongly by chemical ($F_{3,42} = 500$, p–value $< .0001$) and by treatment ($F_{3,42} = 27.98$, p–value $< .0001$). To look for the best combination, the coefficient table shows that Treatment A with Chemical III has the greatest mean at 4.69 rats killed in 10 hours (SE = 0.17) and next best is Treatment C with chemical III which is slightly slower by 0.57 rats per 10 hours (SE of the difference = .20). The second best chemical is II which reduces mean rate to 3.16 (SE = 0.17) with treatment A.

The "Sleuth" recommendations for a stat report say that we should back transform to the original scale. I think that's not necessary here because rates of death are readily interpretable, but it is easy to do the back transform. An approximate 95% CI for survival time under treatment A, chemical III is $(1.98, 2.3)$ hours.

The treatments were applied at random, so we can say that the chemicals and treatments "caused" the observed effects. I would guess that the rats are a convenience sample of available lab rats, which does limit the scope of inference to this sample. However, chemistry is pretty universal, and what kills one mammal is very likely to kill another in the same way, so I'm open to hearing arguments from experts that would suggest that these inferences are extensible to a larger population of rats.

R Code

```
survival <- read.csv("http://www.math.montana.edu/~jimrc/classes/stat505/data/ratSurviva
par(mfrow = c(2, 1), mar = c(2, 3, 1, 1) + 0.1)
boxplot(sTime ~ trt * chem, survival, horizontal = TRUE, yaxt = "n")
axis(side = 2, at = 1:12, paste(LETTERS[1:4], rep(c("I", "II", "III"), each = 4),
    sep = "."), las = 2)
with(survival, interaction.plot(trt, chem, sTime))
```

```
Xone <- data.frame(mu = rep(1, 12), t1 = rep(1:0, c(1, 3)))
Xone <- cbind(Xone, t2 = Xone$t1[c(12, 1:11)], t3 = Xone$t1[c(11:12, 1:10)],
    t4 = Xone$t1[c(10:12, 1:9)], c1 = rep(1:0, c(4, 8)))
Xone <- cbind(Xone, c2 = Xone$c1[c(9:12, 1:8)], c3 = Xone$c1[c(5:12, 1:4)])
rownames(Xone) <- paste("Chem", LETTERS[rep(1:4, 3)], ".Trt", rep(c("I",
    "II", "III"), each = 4), sep = "")
xtable(Xone, digits = 0)
```

```
Xtwo <- NULL
for (j in 6:8) {
    for (i in 2:5) {
        Xtwo <- cbind(Xtwo, Xone[, i] * Xone[, j])
    }
}
rownames(Xtwo) <- rownames(Xone)
xtable(Xtwo)
```

```
Xone <- as.matrix(Xone)
Xtwo <- as.matrix(Xtwo)
all.equal(Xone, Xtwo %*% Xone)
```

```
## [1] TRUE
```

```
bigX <- cbind(Xone, Xtwo)
```

```
bigX <- cbind(1, model.matrix(sTime ~ chem + 0, survival), model.matrix(sTime ~
    trt + 0, survival), diag(12)[rep(1:12, each = 4), ])
betahat <- cbind(MASS::ginv(crossprod(bigX)) %*% crossprod(bigX, survival$sTime),
    unlist(dummy.coef(lm(sTime ~ chem * trt, survival))))
dimnames(betahat) <- list(1:20, c("M-P est", "DummyCoef"))
xtable(round(t(betahat[1:12, ]), 2))
tbhat <- round(t(betahat[13:20, ]), 2)
colnames(tbhat) <- 13:20
xtable(tbhat)
```

```
which(abs(betahat - dbetahat) < 1e-04)
summary(fitted(lm(sTime ~ .^2, survival)) - bigX %*% betahat)
```

```
par(mfrow = c(1, 3))
rawfit <- lm(sTime ~ chem * trt, survival)
plot(rawfit, which = c(1, 3))
MASS::boxcox(rawfit, lambda = seq(-2, 0, 0.1))
```

```
recipFit <- lm(I(1/sTime) ~ .^2, survival)
xtable(anova(recipFit))
```

```
survival$rate <- 1/survival$sTime
recipFit2 <- lm(rate ~ chem + trt + 0, survival)
coef2 <- coef(recipFit2)
par(mfrow = c(1, 3))
plot(recipFit2, which = 1:3)
xtable(recipFit2)
```