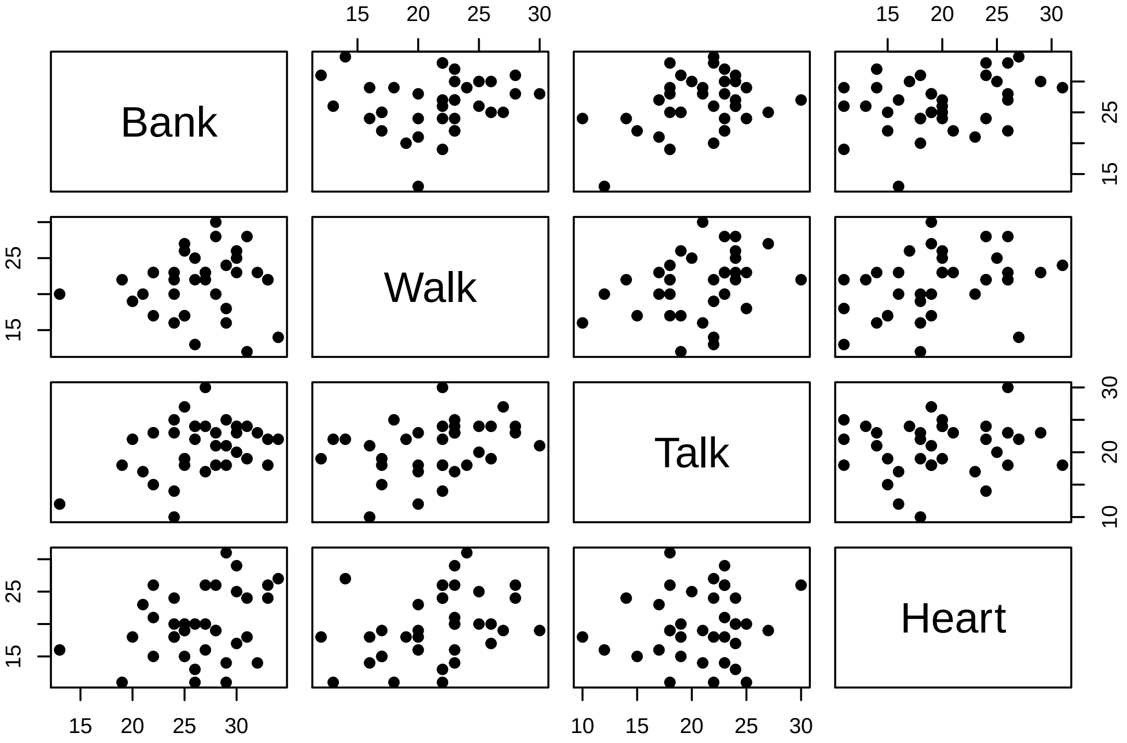
HW 10 S139

Callin Switzer

November 16, 2014

#2

(a)



(b)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	3.1787	6.3369	0.50	0.6194
Bank	0.4052	0.1971	2.06	0.0480
Walk	0.4516	0.2009	2.25	0.0316
Talk	-0.1796	0.2222	-0.81	0.4249

Table 1: Least Squares Fit for linear regression of heart on bank, walk, and talk

(a) *from hw assignment, rather than from book*

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-27.4291	35.7850	-0.77	0.4490
Walk	1.8487	1.7439	1.06	0.2970
Bank	1.3961	1.2649	1.10	0.2779
Walk:Bank	-0.0515	0.0616	-0.84	0.4091

Table 2: Least Squares Fit for linear regression of heart on bank, walk, and interaction, bank*walk

Here is the interpretation of the results

Heart: The age-adjusted death rates from ischemic heart disease Bank: The amount of time a sample of bank clerks takes to make change for two $20 bills or give two $20 bills for change Walk: The walking speed of pedestrians over 60 feet on a clear summer day in downtown.

β_0 is the Estimate for the (Intercept). This is the estimate for the mean Heart when Walk and Bank are both = 0.

β_2 is the Estimate for Bank (from the table above). This is the estimate for the mean change in Heart when Bank increases by one unit, when holding Walk constant. In other words, for each unit increase in Bank, we expect the mean of Heart to increase by 1.39 units, when Walk is contstant.

β_1 is the Estimate for Walk (from the table above). This is the estimate for the mean change in Heart when Walk increases by one unit, while accounting for Bank. In other words, we expect the mean of Heart to increase by β_1 units for every increase in one unit in Walk, while Bank = 0.

β_3 is the Estimate for Bank:Walk (from the table above). This is the estimate interaction term to describe how the mean of Heart changes for different levels of Walk and Bank. For instance, if Bank = 1, then the slope of the regression line will be (β_2 + β_3) * Walk, and the intercept will be (β_0 + β_2). If Bank = 2, then the slope of the regression line will be ($\beta_2 + 2 * \beta_3$) * Walk, and the intercept will be ($\beta_0 + 2 * \beta_2$)

(b)

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	19.8694	0.8039	24.72	0.0000
Walk.Centered	0.4901	0.2189	2.24	0.0323
Bank.Centered	0.2923	0.1968	1.49	0.1472
Walk.Centered:Bank.Centered	-0.0515	0.0616	-0.84	0.4091

Table 3: Regression with centered Bank and Walk

Here is the interpretation of the results

Heart: The age-adjusted death rates from ischemic heart disease Bank: The amount of time a sample of bank clerks takes to make change for two $20 bills or give two $20 bills for change Walk: The walking speed of pedestrians over 60 feet on a clear summer day in downtown

β_0 is the Estimate for the (Intercept). This is the estimate for the mean Heart when Walk == mean(Walk) and Bank == mean(Bank).

β_2 is the Estimate for Bank.Centered (from the table above). This is the estimate for the mean change in Heart when Bank increases by one unit, when Walk == the mean of Walk. In other words, for each unit increase in Bank, we expect the mean of Heart to increase by 0.29 units, when Walk = mean(Walk).

β_1 is the Estimate for Walk.Centered (from the table above). This is the estimate for the mean change in Heart when Walk increases by one unit, while Bank = mean(Bank). In other words, we expect the mean of Heart to increase by β_1 units for every increase in one unit in Walk, while Bank = mean(Bank).

β_3 is the Estimate for the interaction of Bank.Centered*Walk.Centered. This term describes how the slope and intercept of Heart change as Bank.Centered and Walk.Centered both change away from their means. For* *instance, if Bank.Centered = 1, then the slope of the regression line will be (β_1 + β_3)* Walk.Centered, and the intercept will be (β_0 + β_2).

#2c on hand-written page

#3

(ch.10, ex. 11)

(a) The two-sided p-value for whether size of reserve has any effect on number of species, after accounting for the days of observation: 0.4879093

The one-sided p-value when alternative is that size has a positive effect: 0.2439546

This means that we don't have evidence to reject the hypothesis that lsize has no effect on number of species, when accounting for the number of days of observation.

If the researchers tended to spend much more time observing in the larger reserves, that could indicate that they simply don't have enough power to reject the null hypothesis. When interpreting the results, the researchers should be careful to mention that they didn't have enough evidence to reject the null hypothesis, but that the results may change if they had a more balanced sample from small and large reserves.

(b) The two-sided p-value for the test that the coefficient for lsize = 1 is 1.882618e-06.

(c) The confidence interval for β_1 is β_1 +- $t_{1-\alpha/2,13}$ * SE. This is [.0809 - 2.16 * .1131, .0809 + 2.16 * .1131] or [-0.1634, 0.3252].

(d) 100 - 11.41 = 88.59% of variation in log number of species remains unexplained by log size and days of observations.

- On the original scale, here are the interpretation of the coefficients.

β_0 is the log of the number of butterflies when lsize = 0 and days = 0. On the original scale, this means that the number of butterflies when lsize = 0 and days = 0 would be exp(3.775) or about 43 butterflies when size = 1 (logsize = 0).

β_1 is the effect of log of size of reserve on log of number of butterflies, while accounting for days of observation. On the original scale, this means that the median of {Number of butterflies|size} = exp(3.775)*size$^{0.0809}$, while accounting for number of day of observation.

Alternatively, a 1% increase in . . . .

β_2 is the effect of days of observation on log number of butterflies. This means that doubling the number of days of observation will be associated with a multiplicative change of 2$^{.0774} in the median of number butterflies, while accounting for the size of the reserve.

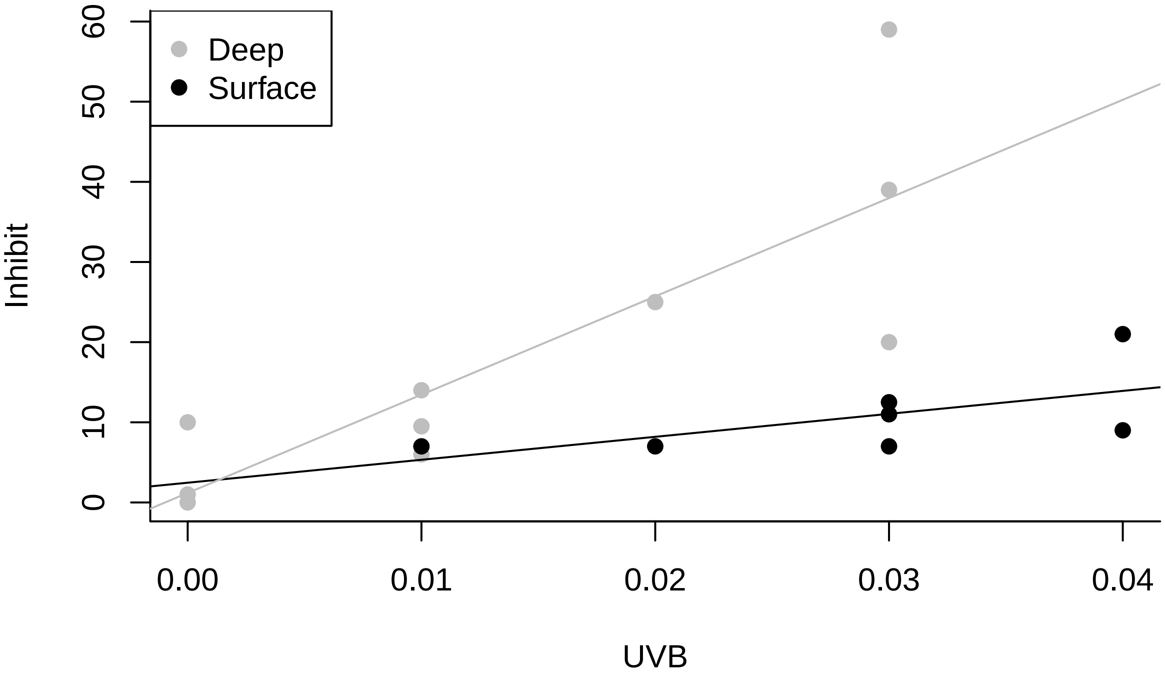
4a

- Does the effect of UVB exposure on the distribution of percentage inhibitation differ at the surface and in the deep?
  - Yes.

- How much difference is there?
  - β_3 is 1.278. This means that the slope of the line for Deep is 1.278 units larger than the slope of the line for Surface. In other words, the UVB inhibition increases to a greater degree for deep samples than surface samples.

Summary of Statistical Findings

Below is a plot showing the regression lines for the multiple regression, $\{\mu|uvb, DEPTH\} = \beta_0 + \beta_1 UVB +$ β_2 Surface + β_3 * Surface * UVB



	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1.1806	4.2921	0.28	0.7876
UVB	1226.3889	232.7730	5.27	0.0002
SurfaceSurface	1.2778	11.0659	0.12	0.9098
UVB:SurfaceSurface	-939.9306	409.8391	-2.29	0.0391

Table 4: Multiple regression for UVB and Depth vs. Inhibition

Details documenting statistical findings

From the plot, you can see that the slopes of the two depths do not seem to be the same. The table above supports this observation. From the table we see that p-value for the test where H_0: the interaction term , $(\beta_3) = 0$ is less than 0.05. We therefore reject the hypothesis that the slopes of the two lines are the same. This means that the effect of UVB exposure on the distribution of percentage inhibitation differs at the surface and in the deep. Furthermore, we can say that the slope of the Deep line is steeper than the Surface line.

Here are the estimated models for each level of depth:

At the Surface:

$\{\mu|uvb, DEPTH = Surface\}$ = (β_0 + β_2) + (β_1 + β_3*UVB)

or

$\{\mu|uvb, DEPTH = Surface\}$ = (2.459) + (286.458*UVB)

For Deep samples:

$\{\mu|uvb, DEPTH = Deep\}$ = β_0 + β_1 *UVB)

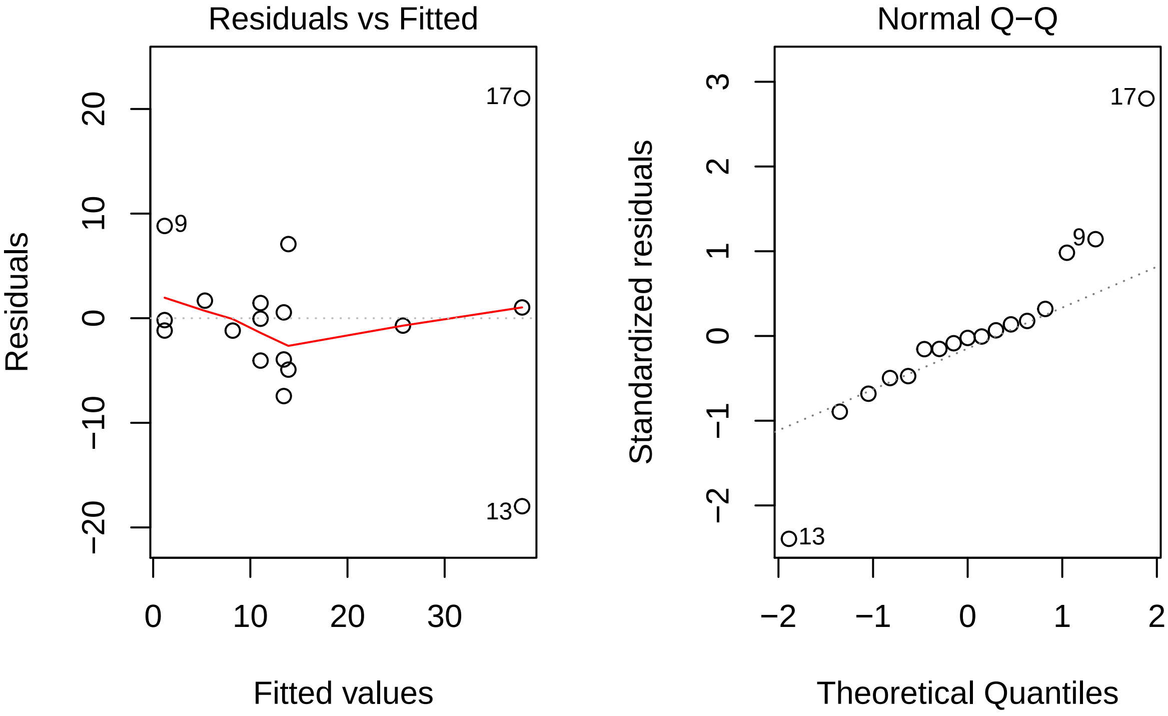
or

$\{\mu|uvb, DEPTH = Deep\}$ = 1.181 + 1226.389*UVB

#4b

I graphed the data above. I also plotted the residual plot and Q-Q plot below. Here are the assumptions for multiple regression: 1. Linearity 2. Constant variance along the line(s) 3. Normality of each subpopulation of responses 4. Independence:Location in relation to the mean cannot be predicted with knowledge of other responses 5. Random sample

From the plot of the original data, I'm not really worried about the linearity assumption. From the residual plot, I see that the data may not have equal variance along the line. Also, the data may not be completely independent, because the locations may make one point similar to another (i.e. if we have information about one location, and another location is close, then we may have a better idea about the other location's values). The normality assumption may not hold. The Q-Q plot shows that the distribution of residuals is not exactly normal. Also, we don't know if there was a random sample



#4c

I found the four top residuals to be rows 17, 13, 9, and 3 from the data. The table below shows that they all come from the "deep" group. If the residuals for one group have a higher standard deviation, then it doesn't make sense to assume that the variances are equal. Thus, you could fit two separate models – one for surface and one for deep.

	number
Deep	4
Surface	0

#4d

[1] "Here is the design matrix"

	x0	xi1	xi2	xi3
1	1.00	0.00	0.00	0.00
2	1.00	0.00	0.00	0.00
3	1.00	0.01	0.00	0.00
4	1.00	0.01	1.00	0.01
5	1.00	0.02	1.00	0.02
6	1.00	0.03	1.00	0.03
7	1.00	0.04	1.00	0.04
8	1.00	0.01	0.00	0.00
9	1.00	0.00	0.00	0.00
10	1.00	0.03	1.00	0.03
11	1.00	0.03	1.00	0.03
12	1.00	0.01	0.00	0.00
13	1.00	0.03	0.00	0.00
14	1.00	0.04	1.00	0.04
15	1.00	0.02	0.00	0.00
16	1.00	0.03	0.00	0.00
17	1.00	0.03	0.00	0.00

[1] "Here is the Y vector"

	Y
1	0.00
2	1.00
3	6.00
4	7.00
5	7.00
6	7.00
7	9.00
8	9.50
9	10.00
10	11.00
11	12.50
12	14.00
13	20.00
14	21.00
15	25.00
16	39.00
17	59.00

[1] "Here are my results for $(X^TX)^{-1}X^TY$. These results are the same as the estimates above in Table 4"

	Y
x0	1.18
xi1	1226.39
xi2	1.28
xi3	-939.93

4e

Here are the square roots of the diagonal entries for the matrix for $\hat{\sigma}^2(X^TX)^{-1}$. They are the same as the SE's for the linear regression above.

	sqrt.diag.ses..
x0	4.29
xi1	232.77
xi2	11.07
xi3	409.84

#5

(a) Anova table for regression model

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TIME	1	886.95	886.95	21.38	0.0001
Intensity	1	2579.75	2579.75	62.18	0.0000
Residuals	21	871.24	41.49		

(b) Regression model with light as indicator

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
TIME	1	886.95	886.95	16.23	0.0017
INTENSITY	5	2683.51	536.70	9.82	0.0006
TIME:INTENSITY	5	111.55	22.31	0.41	0.8342
Residuals	12	655.92	54.66		

(c) Here is the extra sum-of-squares F-Test

H_0: The reduced model is the best model

H_A: The full model is significantly better at explaining the data

From the table below, we fail to reject the null hypothesis that the reduced model is the best one.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	21	871.24				
2	12	655.92	9	215.31	0.44	0.8894

Code

```
##########1
x <- matrix(c(rep(1, 10), 1:10), ncol = 2)
x
t(x)%*%x
sum((x[,2])^2)

10*mean(x[,2])^2
xtxInv <- solve(t(x)%*%x)

sum(x[,2]^2) / (10*sum(x[,2]^2) - 10^2*mean(x[,2])^2)


55*55
sum(x[,2]*x[,2])

sum(x[,2]^2)
10*mean(x[,2]^2)
mean(x[,2]^2)

y <- 1:10 * 2 + rnorm(10)
summary(lm(y~x[,2]))
plot(y~x[,2])
xty <- t(x)%*%y
10*mean(y)
sum(x[,2]*y)

solve(t(x)%*%x)%*%xty

(10*mean(y)*sum(x[,2]^2) + sum(x[,2]*y)*-mean(x[,2])*10)/(10*sum(x[,2]^2) - 10^2 * mean(x)^2)

xty[1,1]* xtxInv[1,1] + xty[2,1]* xtxInv[1,2]

n <- 10
ybar <- mean(y)
xbar <- mean(x[,2])

(ybar * sum(x[,2]^2) - xbar*sum(x[,2]*y))/(sum(x[,2]^2) - n * xbar^2)


(-n*ybar*xbar + sum(x[,2]*y)) / (sum(x[,2]^2) - n * xbar^2)

var(x[,2])
sum((x[,2]-xbar)^2)/(10-1)

sum(x[,2]^2)/10

sum(x[,2]^2)
xbar^2

sum(x[,2]^2) / (10 * sum((x[,2] - xbar)^2))

1/10 + xbar^2/((10-1)*(var(x[,2])))
```

```
##########2a
options(xtable.comment = FALSE)
library(xtable)
pace <- read.csv("data/ex0914.csv")
library(ggplot2)
pairs(pace, pch = 19)
##########2b
pmod <- lm(Heart~Bank + Walk + Talk, data = pace)
px <- xtable(pmod, caption = "Least Squares Fit for linear regression of
             heart on bank, walk, and talk")
print(px)
##########2ab
pmoda <- lm(Heart~Walk*Bank, data = pace)
px <- xtable(pmoda, caption = "Least Squares Fit for linear regression of
             heart on bank, walk, and interaction, bank*walk")
print(px)
##########2b
Heart <- pace$Heart
Bank.Centered <- pace$Bank - mean(pace$Bank)
Walk.Centered <- pace$Walk - mean(pace$Walk)

pmodc <- lm(Heart~Walk.Centered*Bank.Centered)
#summary(pmodc)
print(xtable(pmodc, caption = "Regression with centered Bank and Walk"))


mean(pace$Walk)
mean(pace$Bank)
-27.4291 + 1.8487*mean(pace$Walk) + 1.3961 *mean(pace$Bank) + -0.0515*mean(pace$Walk)*mean(pace$Bank)

1.8487*mean(pace$Walk) - -0.0515*mean(pace$Walk)*mean(pace$Bank)

-27.4291 +1.84*mean(pace$Walk) + 1.39 * mean(pace$Bank) + -0.0515*mean(pace$Walk)*mean(pace$Bank)

1.8487 + -0.0515*mean(pace$Bank)

1.39 + -0.0515*mean(pace$Walk)
##########3a
stuff <- 2*pt(q = (0.0809 - 1)/.1131, df = 13, lower.tail = TRUE)
round(stuff, 7)

# slide 45, lecture 21
# slide 3, lecture 19
#2*pt(q = .7139, df = 13, lower.tail = FALSE)
#pt(q= 0.7139, df = 13, lower.tail = FALSE)
#2*pt(q = (0.0809 - 1)/.1131, df = 13, lower.tail = TRUE)
#qt(0.975, df = 13)

##########4a
sea <- read.csv("data/ex1026.csv")
#head(sea)
modS <- lm(Inhibit~UVB + Surface + UVB:Surface, dat = sea)
```

```
plot(Inhibit~UVB, col = c("grey", "black")[Surface], pch = 19, dat = sea, bty = "l")
legend("topleft", legend = c("Deep","Surface"), col = c("grey", "black"), pch = 19)
# line for the surface
abline(a = 1.181 + 1.278, b = 1226.389 + -939.931, col = "black")

# line for the deep
abline(a = 1.181, b = 1226.389, col = "grey")
#points(sea$Inhibit[ sea$Surface == "Surface"]~sea$UVB[sea$Surface == "Surface"], col = "red", pch = 19

#ggplot(sea, aes(x = UVB, y =Inhibit)) +
#     geom_point(aes(color = Surface))

print(xtable(modS, caption = "Multiple regression for UVB and Depth vs. Inhibition"))

##########4b
par(mfrow = c(1,2))
plot(modS, which = 1:2)
par(mfrow = c(1,1))
##########4c
soredRes <- sort(abs(modS$residuals), decreasing = T)
top4 <- as.numeric(names(soredRes[1:4]))
atb <- xtable(table(number = sea$Surface[top4]), caption =
                  "Number of residuals in each group, out of the top four")
print(atb, floating = F)
##########4d
print("Here is the design matrix")
X <- data.frame(x0 = rep(1, nrow(sea)), xi1 = sea$UVB, xi2 = -1 + as.numeric(sea$Surface), xi3 = I((-1
# note: deep = 0, and surface = 1
#X
print(xtable(as.matrix(X), caption = "Design matrix for Ozone Layer Study"), floating = F)
Y <- data.frame(Y = sea$Inhibit)
print("Here is the Y vector")
print(xtable(Y), floating = F)

X <- as.matrix(X)
Y <- as.matrix(Y)

print("Here are my results for $(X^{T} X)^{-1} X^{T} Y$. These results are the same as the estimates ab
print(xtable(solve(t(X)%*%X)%*%t(X)%*%Y), floating = F)
##########4e
ses <- (sum(modS$residuals^2)/(length(Y) - 4))*solve(t(X)%*%X)
print(xtable(data.frame(sqrt(diag(ses)))), floating = F)

##########5a
foam <- read.csv("data/case0901.csv")
foam$TIME <- as.factor(foam$Time)
#plot(foam$Flowers~foam$Intensity, col = foam$TIME)
modF <- lm(Flowers~TIME+Intensity, data = foam)
#summary(modF)
AF <- anova(modF)
print(xtable(AF))
##########5b
```

```
#summary(aov(modF))

INTENSITY <- as.factor(foam$Intensity)
modF1 <- lm(Flowers~TIME*INTENSITY, data = foam)
#summary(modF1)
AF1 <- anova(modF1)
print(xtable(AF1))

########## 5c
SSRRed <- sum(modF$residuals^2)
dfRed <- 21
sigmaHat2Red <- 41.49

SSRFull <- sum(modF1$residuals^2)
dfFull <- 12
sigmaHat2Full <- 7.393^2

FStat <- ((SSRRed - SSRFull)/ (12 - 3)) / sigmaHat2Full

#pf(FStat, df1 = 12-3, df2 = dfFull, lower.tail = F)

print(xtable(anova(modF, modF1)))
```