**Final Exam Question 1**  Thomas Dolan

1.a

> data1 <- read.table(file.choose())

> names(data1) = c("x1","x2","x3","y")

> x11 <- data1$x1 \* data1$x1

> x22 <- data1$x2 \* data1$x2

> x33 <- data1$x3 \* data1$x3

> fit <- lm(y~x1+x2+x3+x11+x22+x33,data1)

> summary(fit)

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

(Intercept) 1.22328 0.02578 47.446 5.91e-16 \*\*\*

x1 -0.03531 0.01711 -2.064 0.05957 .

x2 0.06253 0.01711 3.656 0.00291 \*\*

x3 0.05399 0.01711 3.156 0.00758 \*\*

x11 -0.05359 0.01665 -3.218 0.00673 \*\*

x22 -0.06296 0.01665 -3.781 0.00229 \*\*

x33 -0.06561 0.01665 -3.940 0.00169 \*\*

> x12 <- data1$x1 \* data1$x2

> x13 <- data1$x1 \* data1$x3

> x23 <- data1$x2 \* data1$x3

> fit2 <- lm(y~x1+x2+x3+x11+x22+x33+x12+x13+x23,data1)

> summary(fit2)

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

(Intercept) 1.223282 0.028689 42.640 1.21e-12 \*\*\*

x1 -0.035308 0.019035 -1.855 0.09329 .

x2 0.062532 0.019034 3.285 0.00822 \*\*

x3 0.053988 0.019034 2.836 0.01766 \*

x11 -0.053591 0.018530 -2.892 0.01605 \*

x22 -0.062956 0.018529 -3.398 0.00680 \*\*

x33 -0.065608 0.018529 -3.541 0.00535 \*\*

x12 -0.013125 0.024870 -0.528 0.60918

x13 0.004875 0.024870 0.196 0.84852

x23 -0.010625 0.024870 -0.427 0.67827

We can see in both summaries that the values for B11, B22, and B33 in both models are negative. For model 1, the standard deviations for B11, B22, and B33 are 0.01665 for them all. The values of B1, B2, and B3 are -0.03531, 0.06253, and 0.05399. For model 2, the standard deviations for B11, B22, and B33 are 0.018530, 0.018529, and 0.018529. The values of B1, B2, and B3 are -0.035308, 0.062532, and 0.053988. I would prefer model 1 in these circumstances, as standard deviation in general is lower, giving us a closer fit.

1.b

Here, we need to maximize expected response for model 1. In order to do this, we need to maximize the individual expected response. The data is already centered since the mean for x1, x2, and x3 are all 0. We should look at the correlation matrix for the data:

> data2 <- transform(data1, x1=x1,x2=x2,x3=x3,x11=x1^2,x22=x2^2,x33=x3^2,y=y)

> cor(data2)

x1 x2 x3 y x11 x22 x33

x1 1.000000e+00 0.0000000 0.0000000 -0.2396597 4.960383e-05 -4.341219e-06 -4.341219e-06

x2 0.000000e+00 1.0000000 0.0000000 0.4244329 0.000000e+00 0.000000e+00 0.000000e+00

x3 0.000000e+00 0.0000000 1.0000000 0.3664419 0.000000e+00 0.000000e+00 0.000000e+00

y -2.396597e-01 0.4244329 0.3664419 1.0000000 -2.953555e-01 -3.671966e-01 -3.875382e-01

x11 4.960383e-05 0.0000000 0.0000000 -0.2953555 1.000000e+00 -9.031417e-02 -9.031417e-02

x22 -4.341219e-06 0.0000000 0.0000000 -0.3671966 -9.031417e-02 1.000000e+00 -9.032425e-02

x33 -4.341219e-06 0.0000000 0.0000000 -0.3875382 -9.031417e-02 -9.032425e-02 1.000000e+00

We can see that the interaction between x1&x11, x2&x22, and x3&x33 is not very strong, thus reducing the possibility of multicollinearity. This is to be expected since the data is centered. In order to determine the maximums, we must consider where each beta value is maximized. We can do this by setting of the variables x2 and x3 equal to 0, x1 and x3 equal to 0, & x1 and x2 equal to 0.

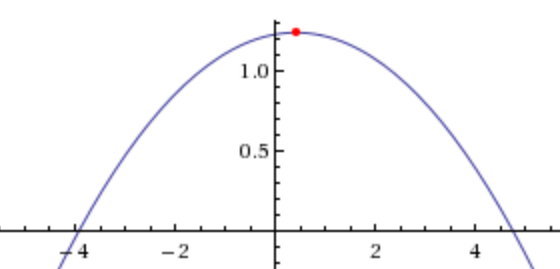
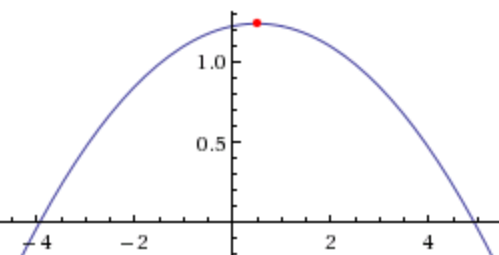
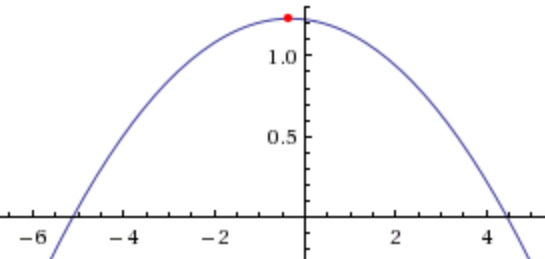
For X2 = X3 = 0: Y= 1.2233 – 0.0353 B1 – 0.0536 B11 = 1.2233 – 0.0353 B1 – 0.0536B1\*B1

For X1 = X3 = 0: Y= 1.2233 + 0.0625 B2 – 0.06296 B22 = 1.2233 + 0.0625 B2 – 0.06296 B2\*B2

For X1 = X2 = 0: Y= 1.2233 + 0.05399 B3 – 0.0656 B33 = 1.2233 + 0.05399 B3 – 0.0656 B3\*B3

When these functions are graphed, we see:

x1 x2 x3



We get the max values of each parabola as estimates: x1\* = -0.3293, x2\* = 0.4963, and x3\* = 0.4115

1.c

> rstandard(fit)

> rstudent(fit)

> fit$resid

> hatvalues(fit)

> cooks.distance(fit)

> vif(fit)

> dffits(fit)

Running these diagnostics, we find outliers from using the studentized values and the dffits values. The results are:

> rstudent(fit)

1 2 3 4 5 6 7

-0.62365053 -1.35870007 -0.23543415 -2.18021726 0.67675739 0.17620969 -0.09711196

8 9 10 11 12 13 14

-1.23636503 3.45720820 -0.37382676 0.12011682 2.36856196 -0.48786994 3.79458297

15 16 17 18 19 20

0.36349101 0.14524161 -0.18807348 -0.37304237 -0.02134180 -0.17136063

> dffits(fit)

1 2 3 4 5 6

-0.403201128 -0.878423693 -0.152212352 -1.409549122 0.437529556 0.113921100

7 8 9 10 11 12

-0.062783732 -0.799320777 4.298930426 -0.464894426 0.149375336 2.945505268

13 14 15 16 17 18

-0.606707150 4.718881895 0.162366914 0.064877622 -0.084010083 -0.166633384

19 20

-0.009533115 -0.076544662

From these values, we can spot 3 noticeable outliers: rows 9, 12, and 14. If we remove these from the data set, we get a new fit:

> data2 <- data1[-c(9,12,14),]

> x11 <- data2$x1 \* data2$x1

> x22 <- data2$x2 \* data2$x2

> x33 <- data2$x3 \* data2$x3

> fit2 <- lm(y~x1+x2+x3+x11+x22+x33,data2)

> summary(fit2)

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

(Intercept) 1.220514 0.008559 142.607 < 2e-16 \*\*\*

x1 -0.013219 0.007152 -1.848 0.09431 .

x2 0.048969 0.007152 6.847 4.48e-05 \*\*\*

x3 0.029969 0.007152 4.190 0.00186 \*\*

x11 -0.067222 0.007928 -8.479 7.05e-06 \*\*\*

x22 -0.064349 0.007928 -8.116 1.04e-05 \*\*\*

x33 -0.082011 0.007928 -10.344 1.16e-06 \*\*\*

The interaction terms B11, B22, and B33 are all still negative. If we calculate new maximum point estimates, we find (as per our method in part b), we get:

x1: 1.2205 -0.01322 B1 –0.06722 B1^2 => x1\* = -0.09833

x2: 1.2205 + 0.04897 B2 –0.06435 B2^2 => x2\* = 0.3805

x3: 1.2205 -0.01322 B3 –0.06722 B3^2 => x3\* = 0.1827

Removing those outliers greatly influenced what our estimated maximums are.

**Final Exam Question 2**

2.a

The matrices will be designed as so:

> X

[,1] [,2] [,3] [,4] [,5] [,6] [,7] [,8] [,9] [,10]

[1,] 1 0 0 0 0 0 0 0 0 0

[2,] 1 0 0 0 1 0 0 0 0 0

[3,] 1 1 0 0 0 0 0 0 0 0

[4,] 1 1 0 0 1 0 0 1 0 0

[5,] 1 0 1 0 0 0 0 0 0 0

[6,] 1 0 1 0 1 0 0 0 1 0

[7,] 1 1 1 0 0 1 0 0 0 0

[8,] 1 1 1 0 1 1 0 1 1 0

[9,] 1 0 0 1 0 0 0 0 0 0

[10,] 1 0 0 1 1 0 0 0 0 1

[11,] 1 1 0 1 0 0 1 0 0 0

[12,] 1 1 0 1 1 0 1 1 0 1

> beta

[,1]

[1,] "u"

[2,] "a2"

[3,] "B2"

[4,] "B3"

[5,] "G2"

[6,] "S22"

[7,] "S23"

[8,] "C22"

[9,] "E22"

[10,] "E32"

> Y

[,1]

[1,] "y111"

[2,] "y112"

[3,] "y211"

[4,] "y212"

[5,] "y121"

[6,] "y122"

[7,] "y221"

[8,] "y222"

[9,] "y131"

[10,] "y132"

[11,] "y231"

[12,] "y232"

2.b

> names(data1) <- ("number","temp",SO2","citric","ACY",TP")

> fit <- lm(ACY ~ temp + SO2 + citric + temp\*SO2 +temp\*citric + SO2\*citric, data= data1)

> summary(fit)

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

(Intercept) 10.23611 53.31430 0.192 0.855

temp 0.59278 0.79069 0.750 0.487

SO2 0.86950 0.70785 1.228 0.274

citric 9.87778 58.48694 0.169 0.873

temp:SO2 -0.00530 0.01035 -0.512 0.630

temp:citric 0.05778 0.84529 0.068 0.948

SO2:citric -0.18200 0.31058 -0.586 0.583

Doing the same for square root, we get:

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

(Intercept) 3.7766766 3.2498530 1.162 0.298

temp 0.0463633 0.0481979 0.962 0.380

SO2 0.0625439 0.0431480 1.450 0.207

citric 1.0189877 3.5651590 0.286 0.786

temp:SO2 -0.0004402 0.0006311 -0.698 0.517

temp:citric -0.0004375 0.0515258 -0.008 0.994

SO2:citric -0.0129276 0.0189318 -0.683 0.525

And for the log transformation:

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

(Intercept) 2.8241706 0.8197299 3.445 0.0183 \*

temp 0.0146657 0.0121573 1.206 0.2816

SO2 0.0183235 0.0108835 1.684 0.1531

citric 0.3903275 0.8992614 0.434 0.6823

temp:SO2 -0.0001435 0.0001592 -0.902 0.4086

temp:citric -0.0014328 0.0129967 -0.110 0.9165

SO2:citric -0.0038007 0.0047753 -0.796 0.4622

If we use a factor transformation:

> fit <- lm(ACY ~ factor(temp) + factor(SO2) + factor(citric) + factor(temp\*SO2) + factor(temp\*citric) + factor(SO2\*citric), data= data1)

> summary(fit)

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

(Intercept) 28.658 2.476 11.575 0.00738 \*\*

factor(temp)80 20.383 3.132 6.509 0.02280 \*

factor(SO2)50 48.125 3.322 14.488 0.00473 \*\*

factor(SO2)100 44.550 3.322 13.412 0.00551 \*\*

factor(citric)1 14.517 3.132 4.635 0.04353 \*

factor(temp \* SO2)2500 15.750 3.836 4.106 0.05450 .

factor(temp \* SO2)4000 NA NA NA NA

factor(temp \* SO2)5000 15.900 3.836 4.145 0.05356 .

factor(temp \* SO2)8000 NA NA NA NA

factor(temp \* citric)50 -1.733 3.132 -0.553 0.63555

factor(temp \* citric)80 NA NA NA NA

factor(SO2 \* citric)50 -9.150 3.836 -2.386 0.13980

factor(SO2 \* citric)100 -18.200 3.836 -4.745 0.04166 \*

We see the factor transformation provides the closest fit, given the lower p-values.

2.c

Using backwards elimination we find:

> MSE=(summary(fit)$sigma)^2

> step(fit, scale=MSE, direction="backward")

AIC = 10. We see that given their high values, the interactions between SO2&citric, between temp&citric, and temp&SO2 were all significant. We get final coefficients of:

Df Sum of Sq RSS Cp

<none> 14.7 10.000

- factor(SO2 \* citric) 2 165.6 180.3 28.516

- factor(temp \* citric) 2 188.6 203.3 31.636

- factor(temp) 1 311.6 326.3 50.362

- factor(temp \* SO2) 4 4103.3 4118.0 559.829

Coefficients:

(Intercept) factor(temp)80 factor(temp \* SO2)2500

28.66 20.38 63.88

factor(temp \* SO2)4000 factor(temp \* SO2)5000 factor(temp \* SO2)8000

48.12 60.45 44.55

factor(temp \* citric)50 factor(temp \* citric)80 factor(SO2 \* citric)50

12.78 14.52 -9.15

factor(SO2 \* citric)100

-18.20

2.d

Running through the same transformations again, we see that the factor transformation again is the most significant:

> fit <- lm(TP ~ factor(temp) + factor(SO2) + factor(citric) + factor(temp\*SO2) + factor(temp\*citric) + factor(SO2\*citric), data= data1)

> summary(fit)

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

(Intercept) 60.442 25.607 2.360 0.14218

factor(temp)80 6.417 32.390 0.198 0.86127

factor(SO2)50 342.550 34.355 9.971 0.00991 \*\*

factor(SO2)100 589.875 34.355 17.170 0.00337 \*\*

factor(citric)1 26.483 32.390 0.818 0.49948

factor(temp \* SO2)2500 -84.500 39.670 -2.130 0.16689

factor(temp \* SO2)4000 NA NA NA NA

factor(temp \* SO2)5000 -280.050 39.670 -7.060 0.01948 \*

factor(temp \* SO2)8000 NA NA NA NA

factor(temp \* citric)50 -28.867 32.390 -0.891 0.46685

factor(temp \* citric)80 NA NA NA NA

factor(SO2 \* citric)50 -10.900 39.670 -0.275 0.80927

factor(SO2 \* citric)100 -14.650 39.670 -0.369 0.74734

Doing backwards elimination again, we get:

> MSE=(summary(fit)$sigma)^2

> step(fit, scale=MSE, direction="backward")

Df Sum of Sq RSS Cp

<none> 2787 1.5416

- factor(temp \* SO2) ` 4 507737 510523 638.8306

Coefficients:

(Intercept) factor(temp \* SO2)2500 factor(temp \* SO2)4000

69.68 242.17 347.52

factor(temp \* SO2)5000 factor(temp \* SO2)8000

292.07 592.97

We observe that for TP, the most significant factor is the interaction between temp&SO2.

2.e

It appears that the best setting for the company is to set temp and SO2 to the 2 setting, and citric to the 1 setting. Citric did not end up being significant in determining the value of TP, hence it need not be present, reducing cost to the company. Temp and SO2 should be at the respective 2 settings since they were both relevant to determining ACY and TP value. 2 is the max setting for temp and correlated to higher values. 2 is the middle setting for SO2, maximizing value while minimizing cost as much as possible.