Response Surface Methodology, Stat 579 Homework 2 Name: Answer Key
Fall 2014 Prof. Erik B. Erhardt

Part I. (120 points) I recommend reading through all the parts of the HW (with my adjustments) before starting; this may save you some work.
MMA-RSM Chapter 2: 2.6, 2.12, 2.15, 2.16, 2.20, 2.25.

- Use externally studentized residuals in all residual plots.
- In 2.6 and 2.12 conduct lack-of-fit tests for both models.
- In 2.15 and 2.16 obtain a 95% prediction interval at the indicated values in part (b) also.
- In 2.25 fit the model to the original variables first. Then center the predictors before forming product and square terms and refit the model. What changes?

General: Try to do all calculations in R. All R code for the assignment should be included with the part of the problem it addresses (for code and output use a fixed-width font, such as Courier). Code is used to calculate result; text is used to report and interpret results – do not report or interpret results in the code.

(15pts) **1. 2.6** Heat treating is often used to carburize metal parts, such as gears. The thickness of the carburized layer is considered an important feature of the gear, and it contributes to the overall reliability of the part. Because of the critical nature of this feature, two different lab tests are performed on each furnace load. One test is run on a sample pin that accompanies each load. The other test is a destructive test, where an actual part is cross-sectioned. This test involved running a carbon analysis on the surface of both the gear pitch (top of the gear tooth) and the gear root (between the gear teeth). The data in Table E2.4 are the results of the pitch carbon analysis test for 32 parts.

(a) (5 pts) Fit a linear regression model relating the results of the pitch carbon analysis test (PITCH) to the five regressor variables.

Solution: Read data.

```
#### 2.6
fn.data <- "http://statacumen.com/teach/RSM/data/RSM_HW_02-06.txt"
df.2.6 <- read.table(fn.data, header=TRUE)
  # for 2.15, include a value to predict
  # TEMP= 1650, SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80
  df.2.6 <- rbind(df.2.6, c(1650, 1.00, 1.10, 1.00, 0.80, NA))
str(df.2.6)
```

```
## 'data.frame': 33 obs. of  6 variables:
##  $ TEMP    : num  1650 1650 1650 1650 1600 1600 1650 1650 1650 1650 ...
##  $ SOAKTIME: num  0.58 0.66 0.66 0.66 0.66 0.66 1 1.17 1.17 1.17 ...
##  $ SOAKPCT : num  1.1 1.1 1.1 1.1 1.15 1.15 1.1 1.1 1.1 1.1 ...
##  $ DIFFTIME: num  0.25 0.33 0.33 0.33 0.33 0.33 0.5 0.58 0.58 0.58 ...
##  $ DIFFPCT : num  0.9 0.9 0.9 0.95 1 1 0.8 0.8 0.8 0.8 ...
##  $ PITCH   : num  0.013 0.016 0.015 0.016 0.015 0.016 0.014 0.021 0.018 0.019 ...
```

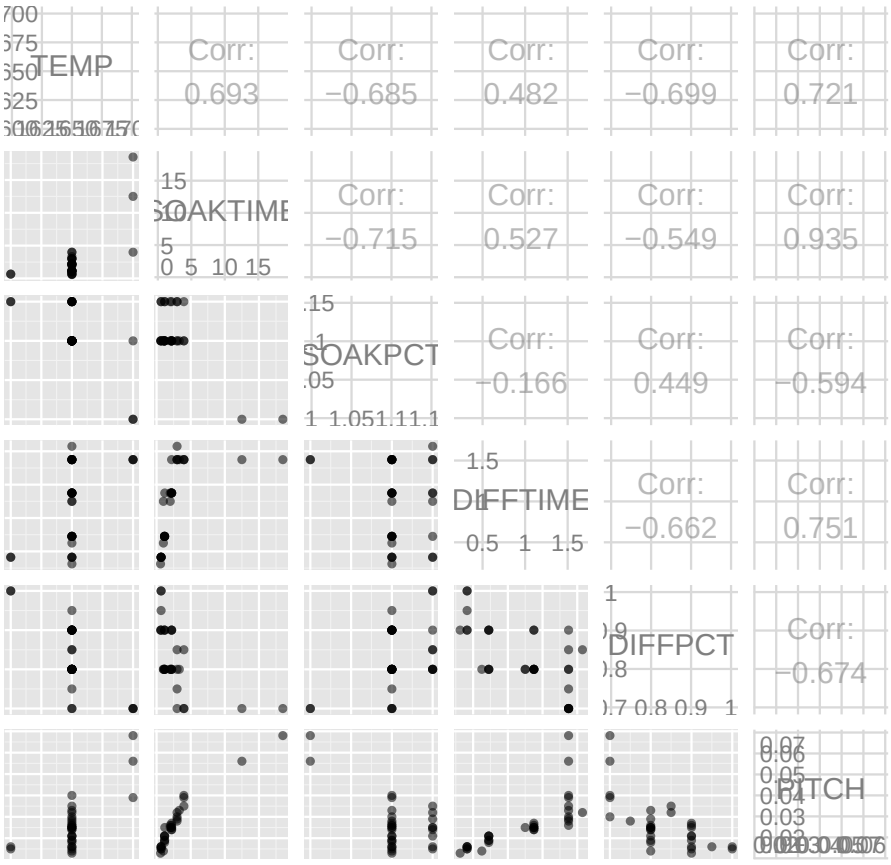
Scatterplot matrix shows some relationships between PITCH and other variables.

```
library(ggplot2)
suppressMessages(suppressWarnings(library(GGally)))
p <- ggpairs(df.2.6, alpha = 0.1)
# put scatterplots on top so y axis is vertical
#p <- ggpairs(df.2.6, upper = list(continuous = "points")
#                , lower = list(continuous = "cor")
#             )
print(p)
```

```
## Warning: Removing 1 row that contained a missing value
## Warning: Removing 1 row that contained a missing value
## Warning: Removing 1 row that contained a missing value
## Warning: Removing 1 row that contained a missing value
## Warning: Removing 1 row that contained a missing value
```

```
  # detach package after use so reshape2 works (old reshape (v.1) conflicts)
  detach("package:GGally", unload=TRUE)
  detach("package:reshape", unload=TRUE)
```

```
## Error: invalid 'name' argument
```



Correlation matrix indicates all correlations with PITCH are different than zero.

```
# correlation matrix and associated p-values testing "H0: rho == 0"
library(Hmisc)
rcorr(as.matrix(df.2.6))
```

```
##           TEMP SOAKTIME SOAKPCT DIFFTIME DIFFPCT PITCH
## TEMP      1.00     0.69   -0.68     0.48   -0.70  0.72
## SOAKTIME  0.69     1.00   -0.71     0.53   -0.55  0.94
## SOAKPCT  -0.68    -0.71    1.00    -0.17    0.45 -0.59
## DIFFTIME  0.48     0.53   -0.17     1.00   -0.66  0.75
## DIFFPCT  -0.70    -0.55    0.45    -0.66    1.00 -0.67
## PITCH     0.72     0.94   -0.59     0.75   -0.67  1.00
##
## n
##          TEMP SOAKTIME SOAKPCT DIFFTIME DIFFPCT PITCH
## TEMP       33       33      33       33      33    32
## SOAKTIME   33       33      33       33      33    32
## SOAKPCT    33       33      33       33      33    32
## DIFFTIME   33       33      33       33      33    32
## DIFFPCT    33       33      33       33      33    32
## PITCH      32       32      32       32      32    33
##
## P
##          TEMP   SOAKTIME SOAKPCT DIFFTIME DIFFPCT PITCH
## TEMP            0.0000   0.0000  0.0045   0.0000  0.0000
## SOAKTIME 0.0000          0.0000  0.0016   0.0009  0.0000
## SOAKPCT  0.0000 0.0000           0.3554   0.0087  0.0003
## DIFFTIME 0.0045 0.0016   0.3554          0.0000  0.0000
## DIFFPCT  0.0000 0.0009   0.0087  0.0000          0.0000
## PITCH    0.0000 0.0000   0.0003  0.0000  0.0000
```

Fit first-order linear model.

```
library(rsm)
rsm.2.6.p.tssdd <- rsm(PITCH ~ FO(TEMP, SOAKTIME, SOAKPCT, DIFFTIME, DIFFPCT), data = df.2.6)
# externally Studentized residuals
rsm.2.6.p.tssdd$studres <- rstudent(rsm.2.6.p.tssdd)
summary(rsm.2.6.p.tssdd)
```

```
##
## Call:
## rsm(formula = PITCH ~ FO(TEMP, SOAKTIME, SOAKPCT, DIFFTIME, DIFFPCT),
##     data = df.2.6)
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.84e-02   7.36e-02   -1.06     0.30
## TEMP         4.39e-05   3.63e-05    1.21     0.24
## SOAKTIME     2.45e-03   2.08e-04   11.79  6.2e-12 ***
## SOAKPCT      1.83e-02   2.01e-02    0.91     0.37
## DIFFTIME     7.79e-03   1.35e-03    5.77  4.5e-06 ***
## DIFFPCT     -3.13e-03   8.05e-03   -0.39     0.70
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared:  0.969,Adjusted R-squared:  0.963
## F-statistic:  162 on 5 and 26 DF,  p-value: <2e-16
##
## Analysis of Variance Table
##
## Response: PITCH
##                                              Df  Sum Sq  Mean Sq F value Pr(>F)
## FO(TEMP, SOAKTIME, SOAKPCT, DIFFTIME, DIFFPCT)  5 0.00419 0.000838   162.4 <2e-16
## Residuals                                      26 0.00013 0.000005
## Lack of fit                                    19 0.00012 0.000007     4.5  0.025
## Pure error                                      7 0.00001 0.000001
##
## Direction of steepest ascent (at radius 1):
##      TEMP  SOAKTIME   SOAKPCT  DIFFTIME   DIFFPCT
##  0.002161  0.120758  0.902712  0.383081 -0.154182
##
## Corresponding increment in original units:
##      TEMP  SOAKTIME   SOAKPCT  DIFFTIME   DIFFPCT
##  0.002161  0.120758  0.902712  0.383081 -0.154182
```

From the parameter estimate table above, the fitted model is

$$\begin{aligned}\widehat{\text{PITCH}} &= -0.078 + 0.000044 \times \text{TEMP} + 0.0025 \times \text{SOAKTIME} \\ &\quad +0.018 \times \text{SOAKPCT} + 0.0078 \times \text{DIFFTIME} - 0.0031 \times \text{DIFFPCT}.\end{aligned}$$

Not all coefficients are significant.

(b) (5 pts) Test for significance of regression. Use $\alpha = 0.05$.
Solution: Regression is significant at the 0.05 level.

```
library(car)
Anova(rsm.2.6.p.tssdd, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: PITCH
##                                               Sum Sq Df F value Pr(>F)
## (Intercept)                                  0.00001  1    1.13    0.3
## FO(TEMP, SOAKTIME, SOAKPCT, DIFFTIME, DIFFPCT) 0.00419  5  162.43 <2e-16 ***
## Residuals                                    0.00013 26
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(c) (5 pts) Conduct a lack-of-fit test.

Solution: The lack-of-fit table is in the summary in part (a) and repeated immediately below.

```
summary(rsm.2.6.p.tssdd)$lof
```

```
## Analysis of Variance Table
##
## Response: PITCH
##                                            Df  Sum Sq  Mean Sq F value Pr(>F)
## FO(TEMP, SOAKTIME, SOAKPCT, DIFFTIME, DIFFPCT)  5 0.00419 0.000838   162.4 <2e-16 ***
## Residuals                                  26 0.00013 0.000005
## Lack of fit                                19 0.00012 0.000007     4.5  0.025 *
## Pure error                                  7 0.00001 0.000001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(model, SSE, df): (Full, 0.00001017, 7) and (first-order, 0.00013421, 26). Therefore:

$$\begin{aligned} F_0 &= \frac{\frac{0.00013421 - 0.00001017}{26-7}}{\frac{0.00001017}{7}} \\ &= 4.493505 \end{aligned}$$

The p-value from an F-distribution with 19 and 7 degrees-of-freedom is 0.02483785. Therefore, at the 0.05 level, we reject the null hypothesis of no lack-of-fit in favor of the alternative hypothesis, concluding that this model (in part (a)) does not adequately fit the observed data.

(35 pts) **2. 2.12** Exercise 2.6 presents data on heat treating gears.

(a) (5 pts) Estimate σ^2 for the model.

Solution:

```
summary(rsm.2.6.p.tssdd)$sigma
```

```
## [1] 0.002272
```

From 2.6a, estimated MSE is $\hat{\sigma}^2 = 0.002272^2 = 0.00000516$.

(b) (5 pts) Find the standard errors of the regression coefficients.

Solution: From the parameter estimate table above, these are the standard errors.

```
summary(rsm.2.6.p.tssdd)$coefficients[,"Std. Error"]
```

```
## (Intercept)        TEMP    SOAKTIME     SOAKPCT    DIFFTIME     DIFFPCT
##   7.360e-02   3.627e-05   2.082e-04   2.014e-02   1.349e-03   8.053e-03
```

(c) (5 pts) Evaluate the contribution of each regressor to the model using the t-test with $\alpha = 0.05$.

Solution: From the parameter estimate table above, these are the p-values for the t-tests.

```
summary(rsm.2.6.p.tssdd)$coefficients[,"Pr(>|t|)"]
```

```
## (Intercept)        TEMP    SOAKTIME     SOAKPCT    DIFFTIME     DIFFPCT
##   2.967e-01   2.369e-01   6.223e-12   3.707e-01   4.462e-06   7.003e-01
```

Only SOAKTIME and DIFFTIME are (highly) significant, conditional on the other terms in the model.

(d) (5 pts) Fit a new model to the response PITCH using new regressors $x_1 = \text{SOAKTIME} \times \text{SOAKPCT}$ and $x_2 = \text{DIFFTIME} \times \text{DIFFPCT}$.

Solution:

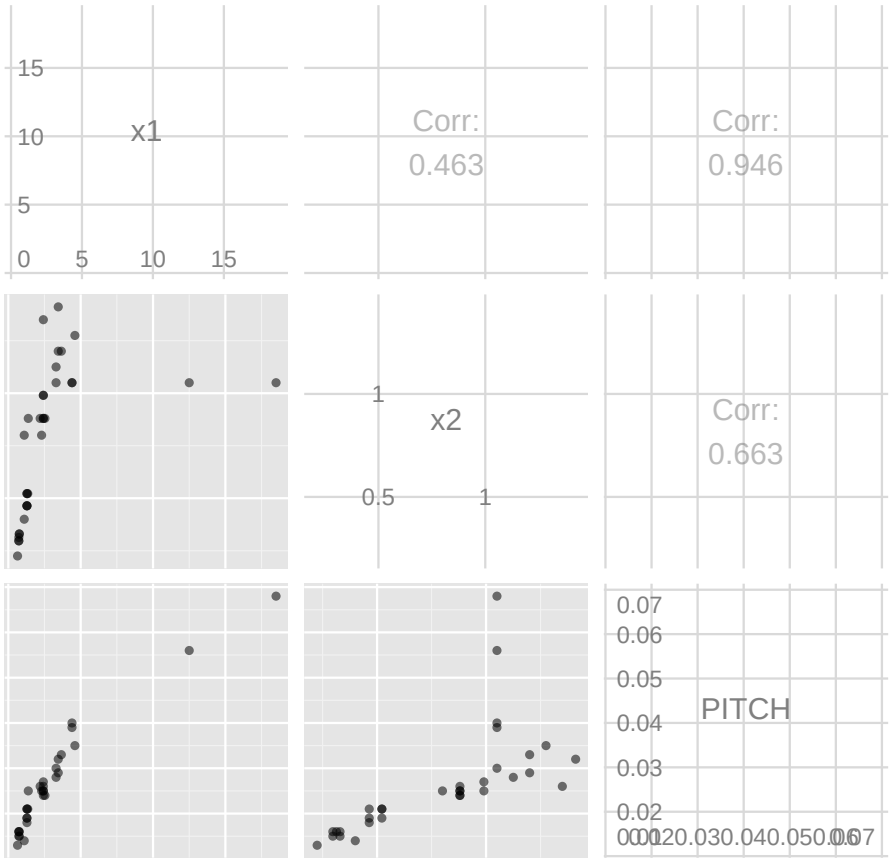
```
# create new variables
df.2.6$x1 <- df.2.6$SOAKTIME  * df.2.6$SOAKPCT
df.2.6$x2 <- df.2.6$DIFFTIME * df.2.6$DIFFPCT
```

```
library(ggplot2)
suppressMessages(suppressWarnings(library(GGally)))
p <- ggpairs(subset(df.2.6, select = c("x1", "x2", "PITCH")), alpha = 0.1)
# put scatterplots on top so y axis is vertical
#p <- ggpairs(df.2.6, upper = list(continuous = "points")
#                   , lower = list(continuous = "cor")
#             )
print(p)
```

```
## Warning:  Removing 1 row that contained a missing value
## Warning:  Removing 1 row that contained a missing value
```

```
  # detach package after use so reshape2 works (old reshape (v.1) conflicts)
  detach("package:GGally", unload=TRUE)
  detach("package:reshape", unload=TRUE)
```

```
## Error:  invalid 'name' argument
```



Correlation matrix indicates all correlations with PITCH are different than zero.

```
# correlation matrix and associated p-values testing "H0: rho == 0"
library(Hmisc)
rcorr(as.matrix(subset(df.2.6, select = c("x1", "x2", "PITCH"))))
```

```
##         x1   x2 PITCH
## x1    1.00 0.46  0.95
## x2    0.46 1.00  0.66
## PITCH 0.95 0.66  1.00
##
## n
##       x1 x2 PITCH
## x1    33 33    32
## x2    33 33    32
```

```
## PITCH 32 32    33
##
## P
##       x1     x2     PITCH
## x1           0.0067 0.0000
## x2    0.0067        0.0000
## PITCH 0.0000 0.0000
```

Fit first-order linear model.

```
library(rsm)
rsm.2.6.p.x1x2 <- rsm(PITCH ~ FO(x1, x2), data = df.2.6)
# externally Studentized residuals
rsm.2.6.p.x1x2$studres <- rstudent(rsm.2.6.p.x1x2)
summary(rsm.2.6.p.x1x2)
```

```
##
## Call:
## rsm(formula = PITCH ~ FO(x1, x2), data = df.2.6)
##
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.010918   0.001082   10.09  5.4e-11 ***
## x1          0.002691   0.000142   18.97  < 2e-16 ***
## x2          0.009356   0.001414    6.62  3.0e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared:  0.958,Adjusted R-squared:  0.955
## F-statistic:  332 on 2 and 29 DF,  p-value: <2e-16
##
## Analysis of Variance Table
##
## Response: PITCH
##             Df  Sum Sq  Mean Sq F value Pr(>F)
## FO(x1, x2)   2 0.00415 0.002073  332.22 <2e-16
## Residuals   29 0.00018 0.000006
## Lack of fit 21 0.00017 0.000008    6.08 0.0065
## Pure error   8 0.00001 0.000001
##
## Direction of steepest ascent (at radius 1):
##     x1     x2
## 0.2764 0.9610
##
## Corresponding increment in original units:
##     x1     x2
## 0.2764 0.9610
```

Fitted model is

$$\widehat{\text{PITCH}} = 0.011 + 0.0027 \times x_1 + 0.0094 \times x_2$$

where $x_1 = \text{SOAKTIME} \times \text{SOAKPCT}$ and $x_2 = \text{DIFFTIME} \times \text{DIFFPCT}$.

(e) (5 pts) Test the model in part (d) for significance of regression using $\alpha = 0.05$. Also calculate the t-test for each regressor and draw conclusions.

Solution:

```
library(car)
Anova(rsm.2.6.p.x1x2, type=3)
```

```
## Anova Table (Type III tests)
##
## Response: PITCH
##              Sum Sq Df F value  Pr(>F)
## (Intercept) 0.00064  1     102 5.4e-11 ***
## FO(x1, x2)  0.00415  2     332 < 2e-16 ***
## Residuals   0.00018 29
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

In the ANOVA table above, the $F_0 = 332.22$ and the associated p-value is < 0.0001, so the model is highly significant. The parameter estimate table indicates both regressors are highly significant.

(f) (5 pts) Estimate σ^2 for the model from part (d), and compare this with the estimate of σ^2 obtained in part (b) above. Which estimate is smaller? Does this offer any insight regarding which model might be preferable?

Solution:

```
summary(rsm.2.6.p.tssdd)$sigma^2
## [1] 5.162e-06
summary(rsm.2.6.p.x1x2)$sigma^2
## [1] 6.239e-06
```

These two variances are not very different; the variance from part (b) is slightly smaller than from part (d). While we wish to reduce error variance, in part (b) it is at the cost of estimating unnecessary parameters.

(g) (5 pts) Conduct a lack-of-fit test.

Solution:

```
summary(rsm.2.6.p.x1x2)$lof
## Analysis of Variance Table
##
## Response: PITCH
##             Df  Sum Sq  Mean Sq F value Pr(>F)
## FO(x1, x2)   2 0.00415 0.002073  332.22 <2e-16 ***
## Residuals   29 0.00018 0.000006
## Lack of fit 21 0.00017 0.000008    6.08 0.0065 **
## Pure error   8 0.00001 0.000001
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

(model, SSE, df): (Full, 0.00001067, 8) and (first-order, 0.00018093, 29). Therefore:

$$\begin{aligned} F_0 &= \frac{\frac{0.00018093 - 0.00001067}{29-8}}{\frac{0.00001067}{8}} \\ &= 6.078815 \end{aligned}$$

The p-value from an F-distribution with 21 and 8 degrees-of-freedom is 0.006516022. Therefore, at the 0.05 level, we reject the null hypothesis of no lack-of-fit in favor of the alternative hypothesis, concluding that this model (in part (d)) does not adequately fit the observed data.

(15 pts) **3. 2.15** Consider the heat-treating data from Exercise 2.6.

(a) (5 pts) Find 95% confidence intervals on the regression coefficients.

Solution: Using $t_{0.975,26} = 2.055529$ we have

```
confint(rsm.2.6.p.tssdd)
##                                                          2.5 %    97.5 %
## (Intercept)                                          -2.297e-01 0.0729117
## FO(TEMP, SOAKTIME, SOAKPCT, DIFFTIME, DIFFPCT)TEMP     -3.064e-05 0.0001185
## FO(TEMP, SOAKTIME, SOAKPCT, DIFFTIME, DIFFPCT)SOAKTIME  2.026e-03 0.0028824
## FO(TEMP, SOAKTIME, SOAKPCT, DIFFTIME, DIFFPCT)SOAKPCT  -2.305e-02 0.0597478
## FO(TEMP, SOAKTIME, SOAKPCT, DIFFTIME, DIFFPCT)DIFFTIME  5.012e-03 0.0105597
## FO(TEMP, SOAKTIME, SOAKPCT, DIFFTIME, DIFFPCT)DIFFPCT  -1.969e-02 0.0134189
```

(b) (10 pts) Find a 95% interval on mean PITCH on TEMP= 1650, SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80. Also, find a 95% prediction interval.

Solution:

```
predict(rsm.2.6.p.tssdd, df.2.6[dim(df.2.6)[1],], interval = "confidence")
```

```
##        fit     lwr     upr
## 33 0.02201 0.02063 0.02339
```

```
predict(rsm.2.6.p.tssdd, df.2.6[dim(df.2.6)[1],], interval = "prediction")
```

```
##        fit     lwr     upr
## 33 0.02201 0.01714 0.02688
```

The fitted value is: 0.022009
95% CI: $(0.020629, 0.023389)$ with length 0.0028
95% PI: $(0.017139, 0.026878)$ with length 0.0098 (almost 4 times as wide).

(10 pts) **4. 2.16** Reconsider the heat treating in Exercise 2.6 and 2.12, where we fit a model to PITCH using regressors $x_1 =$ SOAKTIME $\times$ SOAKPCT and $x_2 =$ DIFFTIME $\times$ DIFFPCT.

(a) (5 pts) Using the model with regressors x_1 and x_2, find a 95% confidence interval on mean PITCH when SOAKTIME = 1.00, SOAKPCT = 1.10, DIFFTIME = 1.00, and DIFFPCT = 0.80.

Solution:

```
predict(rsm.2.6.p.x1x2, df.2.6[dim(df.2.6)[1],], interval = "confidence")
```

```
##        fit    lwr     upr
## 33 0.02136 0.0203 0.02243
```

```
predict(rsm.2.6.p.x1x2, df.2.6[dim(df.2.6)[1],], interval = "prediction")
```

```
##        fit     lwr     upr
## 33 0.02136 0.01614 0.02658
```

The fitted value is: 0.021362
95% CI: $(0.020297, 0.022428)$ with length 0.0021
95% PI: $(0.016144, 0.026581)$ with length 0.0105 (5 times as wide).

(b) (5 pts) Compare the length of this confidence interval with the length of the confidence interval on mean PITCH at the same point from Exercise 2.15 part (b), where an additive model in SOAKTIME, SOAKPCT, DIFFTIME, and DIFFPCT was used. Which confidence interval is shorter? Does this tell you anything about which model is preferable?
Repeat with the prediction interval.

Solution: The CI length is shorter for the x_1, x_2 model, but the PI length is longer. If there is any meaning here it may only apply for the specific $\underset{\sim}{x}$ values in question.

(15 pts) **5. 2.20** In Exercise 2.12 we fitted a model to the response PITCH in the heat treating data of Exercise 2.6 using new regressors $x_1 =$ SOAKTIME $\times$ SOAKPCT and $x_2 =$ DIFFTIME $\times$ DIFFPCT.

(a) (5 pts) Calculate the R^2 for this model, and compare it with the value of R^2 from the original model in Exercise 2.6. Does this provide some information about which model is preferable?

Solution:

```
# R2
summary(rsm.2.6.p.tssdd)$r.squared
```

```
## [1] 0.969
```

```
summary(rsm.2.6.p.x1x2)$r.squared
```

```
## [1] 0.9582
```

```
# Adj-R2
summary(rsm.2.6.p.tssdd)$adj.r.squared
```

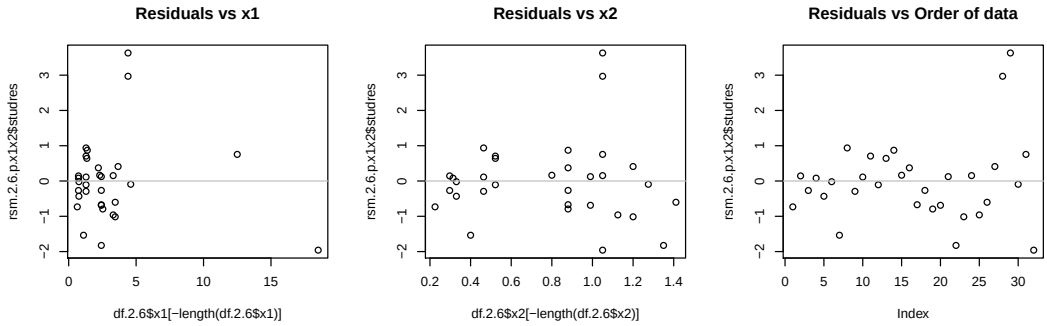
```
## [1] 0.963
summary(rsm.2.6.p.x1x2)$adj.r.squared
## [1] 0.9553
```

$R^2 = 0.968979$ for original model.
$R^2 = 0.958180$ for x_1, x_2 model.
The first model has a slightly larger R^2 but also has 5 regressors instead of 2, though the first model also has a slightly larger Adjusted-R^2.

(b) (5 pts) Plot the residuals from this model versus y and on a normal probability scale. Comment on model adequacy.

Solution: The residual plot indicates no striking departure from model assumptions (no structure, constant variance), except for two outliers (obs 28 and 29).

```
# plot diagnostics
par(mfrow=c(1,3))

plot(df.2.6$x1[-length(df.2.6$x1)], rsm.2.6.p.x1x2$studres, main="Residuals vs x1")
  # horizontal line at zero
  abline(h = 0, col = "gray75")

plot(df.2.6$x2[-length(df.2.6$x2)], rsm.2.6.p.x1x2$studres, main="Residuals vs x2")
  # horizontal line at zero
  abline(h = 0, col = "gray75")

# residuals vs order of data
plot(rsm.2.6.p.x1x2$studres, main="Residuals vs Order of data")
  # horizontal line at zero
  abline(h = 0, col = "gray75")
```

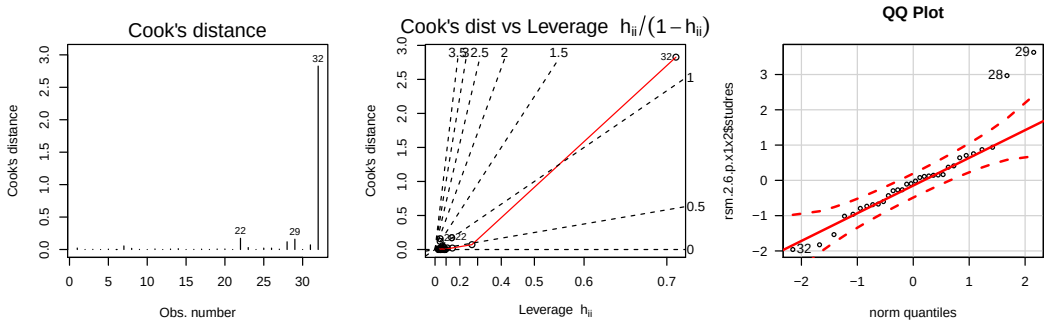


(c) (5 pts) Find the values of Cook's distance measure. Are any observations unusually influential?

Solution: The plot shows that the last observation (32) has a high Cook's D 2.827 >> 1, because of high leverage, and is the only indicated influential observation.

```
# plot diagnostics
par(mfrow=c(1,3))
plot(rsm.2.6.p.x1x2, which = c(4,6))

# Normality of Residuals
library(car)
qqPlot(rsm.2.6.p.x1x2$studres, las = 1, id.n = 3, main="QQ Plot")
## 29 28 32
## 32 31  1
cooks.distance(rsm.2.6.p.x1x2)
##          1          2          3          4          5          6          7          8          9
## 2.256e-02 7.222e-04 2.494e-03 2.138e-04 5.823e-03 1.206e-05 5.539e-02 1.789e-02 1.798e-03
##         10         11         12         13         14         15         16         17         18
## 2.663e-04 8.796e-03 2.048e-04 7.246e-03 1.257e-02 3.155e-04 1.907e-03 5.840e-03 9.280e-04
```

```
##         19         20         21         22         23         24         25         26         27
## 7.945e-03 8.296e-03 2.749e-04 1.727e-01 3.028e-02 4.378e-04 2.152e-02 2.152e-02 4.982e-03
##         28         29         30         31         32
## 1.190e-01 1.590e-01 3.037e-04 7.155e-02 2.827e+00
```



(30pts) **6. 2.25**

An article in the Journal of Pharmaceuticals Sciences (vol. 80, 1991, pp. 971–977) presents data on the observed mole fraction solubility of a solute at a constant temperature to the dispersion, dipolar, and hydrogen bonding Hansen partial solubility parameters. The data are in Table E2.5, where y is the negative logarithm of the mole fraction solubility, x_1 is the dispersion Hansen partial solubility, x_2 is the dipolar partial solubility, and x_3 is the hydrogen bonding partial solubility.

(a) (15 pts) Fit the full second-order model with two-way interactions of y vs x_1, x_2, and x_3. Fit the model to the original variables first. Then center the predictors before forming product and square terms and refit the model. What changes?

Solution: Read data.

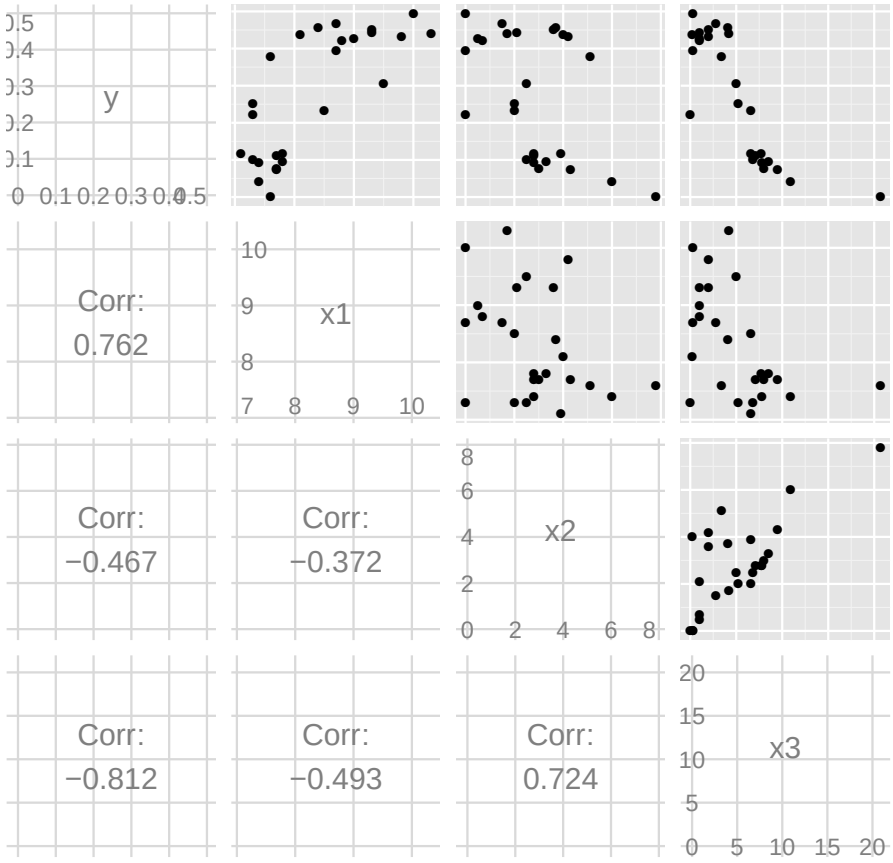
```
#### 2.25
fn.data <- "http://statacumen.com/teach/RSM/data/RSM_HW_02-25.txt"
df.2.25 <- read.table(fn.data, header=TRUE)
str(df.2.25)
```

```
## 'data.frame': 26 obs. of  4 variables:
##  $ y : num  0.222 0.395 0.422 0.437 0.428 0.467 0.444 0.378 0.494 0.456 ...
##  $ x1: num  7.3 8.7 8.8 8.1 9 8.7 9.3 7.6 10 8.4 ...
##  $ x2: num  0 0 0.7 4 0.5 1.5 2.1 5.1 0 3.7 ...
##  $ x3: num  0 0.3 1 0.2 1 2.8 1 3.4 0.3 4.1 ...
```

Scatterplot matrix shows some relationships between y and other variables.

```
library(ggplot2)
suppressMessages(suppressWarnings(library(GGally)))
#p <- ggpairs(df.2.25, alpha = 0.1)
# put scatterplots on top so y axis is vertical
p <- ggpairs(df.2.25, upper = list(continuous = "points")
             , lower = list(continuous = "cor")
          )
print(p)

  # detach package after use so reshape2 works (old reshape (v.1) conflicts)
  detach("package:GGally", unload=TRUE)
  detach("package:reshape", unload=TRUE)
```

```
## Error:  invalid 'name' argument
```



Correlation matrix indicates all correlations with y are different than zero.

```
# correlation matrix and associated p-values testing "H0: rho == 0"
library(Hmisc)
rcorr(as.matrix(df.2.25))
```

```
##        y    x1    x2    x3
## y   1.00  0.76 -0.47 -0.81
## x1  0.76  1.00 -0.37 -0.49
## x2 -0.47 -0.37  1.00  0.72
## x3 -0.81 -0.49  0.72  1.00
##
## n= 26
##
##
## P
##    y      x1     x2     x3
## y         0.0000 0.0161 0.0000
## x1 0.0000        0.0615 0.0106
## x2 0.0161 0.0615        0.0000
## x3 0.0000 0.0106 0.0000
```

Fit second-order linear model, uncentered variables.

```
# uncentered
library(rsm)
rsm.2.25.y.SOx123 <- rsm(y ~ SO(x1, x2, x3), data = df.2.25)
# externally Studentized residuals
rsm.2.25.y.SOx123$studres <- rstudent(rsm.2.25.y.SOx123)
summary(rsm.2.25.y.SOx123)
```

```
##
## Call:
```

```
## rsm(formula = y ~ SO(x1, x2, x3), data = df.2.25)
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.769364   1.286976   -1.37    0.188
## x1           0.420798   0.294173    1.43    0.172
## x2           0.222453   0.130742    1.70    0.108
## x3          -0.127995   0.070245   -1.82    0.087 .
## x1:x2       -0.019876   0.012037   -1.65    0.118
## x1:x3        0.009151   0.007621    1.20    0.247
## x2:x3        0.002576   0.007039    0.37    0.719
## x1^2        -0.019325   0.016797   -1.15    0.267
## x2^2        -0.007449   0.012048   -0.62    0.545
## x3^2         0.000824   0.001441    0.57    0.575
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared:  0.917,Adjusted R-squared:  0.87
## F-statistic: 19.6 on 9 and 16 DF,  p-value: 5.05e-07
##
## Analysis of Variance Table
##
## Response: y
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## FO(x1, x2, x3)   3  0.620  0.2066   55.66 1.1e-08
## TWI(x1, x2, x3)  3  0.024  0.0079    2.14    0.14
## PQ(x1, x2, x3)   3  0.012  0.0040    1.08    0.38
## Residuals       16  0.059  0.0037
## Lack of fit     16  0.059  0.0037
## Pure error       0  0.000
##
## Stationary point of response surface:
##     x1     x2     x3
## 13.285 -1.669  6.501
##
## Eigenanalysis:
## $values
## [1]  0.001959 -0.002142 -0.025766
##
## $vectors
##       [,1]    [,2]    [,3]
## x1  0.2816 -0.4238  0.8609
## x2 -0.1682  0.8615  0.4791
## x3  0.9447  0.2797 -0.1713
```

Fit second-order linear model, centered variables.

```
# center variables by coding them
cd.2.25 <- coded.data(df.2.25, c1 ~ (x1 - mean(df.2.25$x1)) / 1
                             , c2 ~ (x2 - mean(df.2.25$x2)) / 1
                             , c3 ~ (x3 - mean(df.2.25$x3)) / 1
                      )
head(as.data.frame(cd.2.25))
```

```
##       y     c1   c2     c3
## 1 0.222 -1.012 -2.8 -5.104
## 2 0.395  0.388 -2.8 -4.804
## 3 0.422  0.488 -2.1 -4.104
## 4 0.437 -0.212  1.2 -4.904
## 5 0.428  0.688 -2.3 -4.104
## 6 0.467  0.388 -1.3 -2.304
```

```
library(rsm)
rsm.2.25.y.SOc123 <- rsm(y ~ SO(c1, c2, c3), data = cd.2.25)
# externally Studentized residuals
rsm.2.25.y.SOc123$studres <- rstudent(rsm.2.25.y.SOc123)
summary(rsm.2.25.y.SOc123)
```

```
##
## Call:
## rsm(formula = y ~ SO(c1, c2, c3), data = cd.2.25)
```

```
##
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.288307   0.027448   10.50  1.4e-08 ***
## c1           0.090602   0.022533    4.02  0.00099 ***
## c2           0.028677   0.017213    1.67  0.11517
## c3          -0.036304   0.008663   -4.19  0.00069 ***
## c1:c2       -0.019876   0.012037   -1.65  0.11818
## c1:c3        0.009151   0.007621    1.20  0.24731
## c2:c3        0.002576   0.007039    0.37  0.71918
## c1^2        -0.019325   0.016797   -1.15  0.26685
## c2^2        -0.007449   0.012048   -0.62  0.54511
## c3^2         0.000824   0.001441    0.57  0.57543
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Multiple R-squared:  0.917,Adjusted R-squared:  0.87
## F-statistic: 19.6 on 9 and 16 DF,  p-value: 5.05e-07
##
## Analysis of Variance Table
##
## Response: y
##                 Df Sum Sq Mean Sq F value  Pr(>F)
## FO(c1, c2, c3)   3  0.620  0.2066   55.66 1.1e-08
## TWI(c1, c2, c3)  3  0.024  0.0079    2.14    0.14
## PQ(c1, c2, c3)   3  0.012  0.0040    1.08    0.38
## Residuals       16  0.059  0.0037
## Lack of fit     16  0.059  0.0037
## Pure error       0  0.000
##
## Stationary point of response surface:
##     c1     c2     c3
##  4.973 -4.469  1.397
##
## Stationary point in original units:
##     x1     x2     x3
## 13.285 -1.669  6.501
##
## Eigenanalysis:
## $values
## [1]  0.001959 -0.002142 -0.025766
##
## $vectors
##       [,1]    [,2]    [,3]
## c1  0.2816 -0.4238  0.8609
## c2 -0.1682  0.8615  0.4791
## c3  0.9447  0.2797 -0.1713
```

The significance of the models based on the F-statistic is the same, however there are differences in the parameter estimates, the standard errors, t-values, and p-values.
No regressor is significant in the uncentered case! In the centered case, x_1 and x_3 are significant.

(b) (5 pts) Test for significance of regression, using $\alpha = 0.05$.

Solution: Both models have $F = 19.63$ with an associated p-value$< .0001$, which is significant.

(c) (5 pts) Plot the residuals, and comment on model adequacy.

Solution: The residual plots (first row) indicates no striking departure from model assumptions (no structure, constant variance).
However, there is one high-leverage point (obs 26) which has a very large Cook's D, and the normality assumption does not appear to be met.

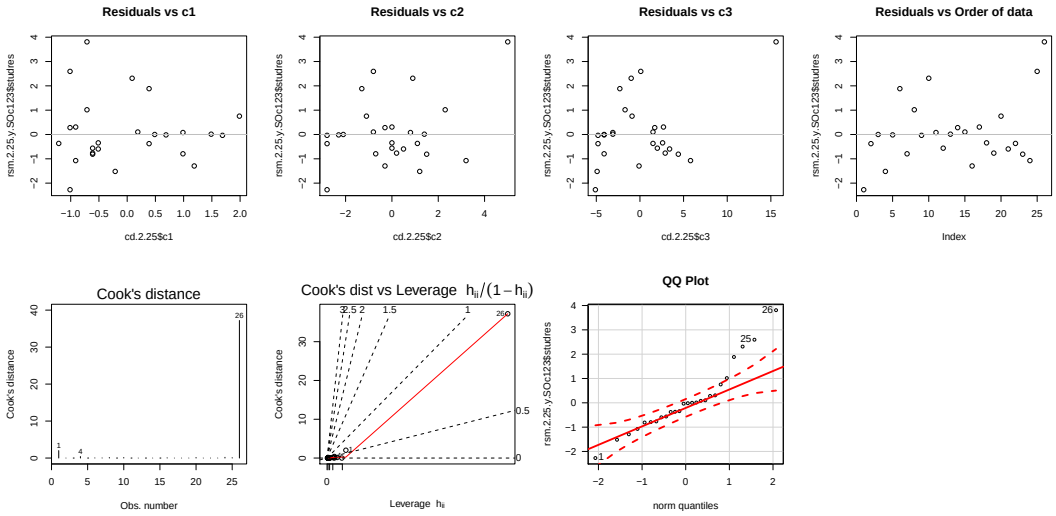
```
# plot diagnistics
par(mfrow=c(2,4))

plot(cd.2.25$c1, rsm.2.25.y.SOc123$studres, main="Residuals vs c1")
  # horizontal line at zero
```

```
  abline(h = 0, col = "gray75")

plot(cd.2.25$c2, rsm.2.25.y.SOc123$studres, main="Residuals vs c2")
  # horizontal line at zero
  abline(h = 0, col = "gray75")

plot(cd.2.25$c3, rsm.2.25.y.SOc123$studres, main="Residuals vs c3")
  # horizontal line at zero
  abline(h = 0, col = "gray75")

# residuals vs order of data
plot(rsm.2.25.y.SOc123$studres, main="Residuals vs Order of data")
  # horizontal line at zero
  abline(h = 0, col = "gray75")

plot(rsm.2.25.y.SOc123, which = c(4,6))

# Normality of Residuals
library(car)
qqPlot(rsm.2.25.y.SOc123$studres, las = 1, id.n = 3, main="QQ Plot")
```

```
## 26 25  1
## 26 25  1
```

```
cooks.distance(rsm.2.25.y.SOc123)
```

```
##           1           2           3           4           5           6           7           8           9
## 2.014e+00 6.990e-03 1.072e-08 4.062e-01 6.804e-06 5.099e-02 4.166e-02 1.481e-01 2.410e-04
##          10          11          12          13          14          15          16          17          18
## 1.470e-01 2.409e-04 3.597e-03 6.506e-05 2.271e-03 7.077e-04 7.932e-02 2.714e-03 1.993e-03
##          19          20          21          22          23          24          25          26
## 8.694e-03 1.543e-01 5.181e-03 1.097e-02 1.105e-02 2.108e-01 1.116e-01 3.720e+01
```



(d) (5 pts) Use the extra sum of squares method to test the contribution of the second-order terms, using $\alpha = 0.05$.
Solution:

```
# fit the first-order model
rsm.2.25.y.FOc123 <- rsm(y ~ FO(c1, c2, c3) + TWI(c1, c2, c3), data = cd.2.25)
# compare the reduced first-order model to the full second-order model
anova(rsm.2.25.y.FOc123, rsm.2.25.y.SOc123)
```

```
## Analysis of Variance Table
##
## Model 1: y ~ FO(c1, c2, c3) + TWI(c1, c2, c3)
## Model 2: y ~ FO(c1, c2, c3) + TWI(c1, c2, c3) + PQ(c1, c2, c3)
```

```
##   Res.Df    RSS Df Sum of Sq    F Pr(>F)
## 1     19 0.0715
## 2     16 0.0594  3    0.0121 1.08   0.38
```

The test of quadratic terms gives $F_0 = 1.08$ (also available from the PQ line in the original ANOVA table) with an associated p-value of 0.38, thus we fail to reject H_0 that all these terms are equal to zero. This implies that a first-order with two-way interaction model is sufficient for these data.