CSC 423 - Chapter 4 & 5 Homework

Jasmine Dumas

October 17, 2015

Submit: page 184 #4.6

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/STREETVN.Rdata")
head(STREETVN, n= 15)
```

```
##    VENDOR EARNINGS AGE HOURS
## 1      21     2841  29    12
## 2      53     1876  21     8
## 3      60     2934  62    10
## 4     184     1552  18    10
## 5     263     3065  40    11
## 6     281     3670  50    11
## 7     354     2005  65     5
## 8     401     3215  44     8
## 9     515     1930  17     8
## 10    633     2010  70     6
## 11    677     3111  20     9
## 12    710     2882  29     9
## 13    800     1683  15     5
## 14    914     1817  14     7
## 15    997     4066  33    12
```

(a) First order model for a mean annual earnings, E(y), as a function of age(x1), and hours worked (x2):

y = β_0 + β_1 AGE + β_2 HOURS

(b) Least squares prediction on SAS output from textbook:

y = -20.35 + 13.35 AGE + 243.71 HOURS

(c) Interpret β coefficients from the model:

β_0 = -20.35 implies we are starting in the negative quadrant

β_1 implies for every 1-unit increase in annual earnings, AGE is multiplied by 13.35

β_2 implies for every 1-unit increase in annual earnings, HOURS is multiplied by 243.71

(d) Test of the global utility of the model (at $\alpha = 0.01$).

H_0: $\beta_1 = \beta_2 = 0$

(df) degrees of freedom in the numerator = k, `k = 2`; (df) degrees of freedom in the denominator = [n - (k + 1)], 12 (k is found in the original F-statistic equation where n is the sample size and k is the number of terms in the model).

H_A: At least one of the coefficients is nonzero

Test Statistic: F = Mean square (model) / Model square (error); F = `2509116 / 300016`, F = 8.363274

p-value = 0.0053

Since the $\alpha = 0.01$ is larger than the significance level of 0.0053, the data provides strong evidence that at least one of the model coefficients is nonzero. The overall model appears to be statistically useful for predicting annual earnings.

(e) Find and interpret the value of the adjusted multiple coefficient of determination $R_A^2 = 0.5126$ which indicates the model is not very adequate.

```
R.a.sq = 1 - ((15-1) / (15 - (2 + 1))) * (1 - 0.5823)
R.a.sq
```

```
## [1] 0.5126833
```

(f) Find and interpret s, the estimated standard deviation of the error term. We expect most (approximately 95%) of the observed y-values to lie within 2s of their respective least squares predicted values, y-hat. An error prediction of 547.7372 is undesirable since the annual earnings is small, `mean(STREETVN$EARNINGS)`, 2577.133.

```
s = sqrt(300016)
s
```

```
## [1] 547.7372
```

(g) Is age (x1) a statistically useful predictor of annual earnings? $\alpha = 0.01$ is less than the p-value of 0.1074 [Pr > |t|] so we can infer that age is not a statistically useful predictor of annual earnings.

(h) The 95% confidence interval for β_2 is 105.33 and 382.09. This interval is positive as the x2 is time in hours. Since β_2 represents the slope of the line relating y to x2 for a fixed x1 in a first-order model. β_2 measures the change in E(y) (annual earnings) for every 1-unit increase in x2 (HOURS) when the other independent variable in the model is held fixed (AGE).

page 186 #4.10

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/SNOWGEESE.Rdata")
head(SNOWGEESE, n=5)
```

```
##   TRIAL             DIET WTCHNG DIGEFF ADFIBER
## 1     1 Plants             -6.0    0.0    28.5
## 2     2 Plants             -5.0    2.5    27.5
## 3     3 Plants             -4.5    5.0    27.5
## 4     4 Plants              0.0    0.0    32.5
## 5     5 Plants              2.0    0.0    32.0
```

```
tail(SNOWGEESE, n=5)
```

```
##    TRIAL             DIET WTCHNG DIGEFF ADFIBER
## 38    38 Chow               9.0   59.0     8.5
## 39    39 Chow              12.0   52.5     8.0
## 40    40 Chow               8.5   75.0     6.0
## 41    41 Chow              10.5   72.5     6.5
## 42    42 Chow              14.0   69.0     7.0
```

(a) The least squares prediction equation for weight change:

y = 12.2 - 0.0265 DIGEST - 0.458 ADFIBRE

```
WTCHNG = SNOWGEESE$WTCHNG
DIGEFF = SNOWGEESE$DIGEFF
ADFIBER = SNOWGEESE$ADFIBER

model = lm(WTCHNG ~ DIGEFF + ADFIBER)
summary(model)
```

```
##
## Call:
## lm(formula = WTCHNG ~ DIGEFF + ADFIBER)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -7.0649 -2.0241  0.5645  2.4590  6.8556
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 12.18044    4.40236   2.767 0.008610 **
## DIGEFF      -0.02654    0.05349  -0.496 0.622555
## ADFIBER     -0.45783    0.12828  -3.569 0.000969 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.519 on 39 degrees of freedom
## Multiple R-squared:  0.5288, Adjusted R-squared:  0.5046
## F-statistic: 21.88 on 2 and 39 DF,  p-value: 4.25e-07
```

(b) Interpret β estimates in the equation from part a:

$\beta_0 = 12.2$

β_1 implies for every 1-unit increase in weight change, DIGEST is multiplied by -0.0265

β_2 implies for every 1-unit increase in weight change, ADFIBRE is multiplied -0.458

(c) Conduct an F-test for overall model adequacy using $\alpha = 0.01$. Since the $\alpha = 0.01$ is larger than the significance level of 0.000, the data provides strong evidence that at least one of the model coefficients is nonzero. The overall model appears to be statistically useful for predicting weight change.

(d) Find and Interpret values for R^2 and R^2_A and determine which is the preferred measure of model fit: $R^2_A = 0.5046$. The R^2_A is the preferred measure for model fit.

(e) Conduct a test to determine if digestion efficiency, x1 is a useful linear predictor of weight change using $\alpha = 0.01$. $\alpha = 0.01$ is less than the p-value of 0.622555 [Pr > |t|] so we can infer that digestion efficiency is not a statistically useful predictor of weight change.

(f) Form a 99% confidence interval for β_2:

```
beta2 <- -0.45783
t_0.005 <- 9.925 # from page 758, df = 2
s_beta_i <- 0.12828

beta2 + t_0.005*s_beta_i # upper bounds
```

```
## [1] 0.815349
```

```
beta2 - t_0.005*s_beta_i # lower bounds
```

```
## [1] -1.731009
```

With the interval (0.815349, -1.731009) we are 99% confident that β_2 falls between the interval. Since β_2 is the slope of the line relating weight change (y) to acid detergent fiber percentage (x2) we can conclude that weight change increases 0.815349% to -1.731009% for every 1-unit increase in acid detergent fiber percentage (x2), holding the digestion efficiency is constant.

page 193 #4.22

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/QUASAR.Rdata")
head(QUASAR, n=5)
```

```
##   QUASAR REDSHIFT LINEFLUX LUMINOSITY AB1450 ABSMAG RFEWIDTH
## 1      1     2.81   -13.48      45.29  19.50 -26.27      117
## 2      2     3.07   -13.73      45.13  19.65 -26.26       82
## 3      3     3.45   -13.87      45.11  18.93 -27.17       33
## 4      4     3.19   -13.27      45.63  18.59 -27.39       92
## 5      5     3.07   -13.56      45.30  19.59 -26.32      114
```

```
REDSHIFT <- QUASAR$REDSHIFT # x1
LINEFLUX <- QUASAR$LINEFLUX # x2
LUMINOSITY <- QUASAR$LUMINOSITY # x3
AB1450 <- QUASAR$AB1450 # x4
RFEWIDTH <- QUASAR$RFEWIDTH # y

qu_model = lm(RFEWIDTH ~ REDSHIFT + LINEFLUX + LUMINOSITY + AB1450)
summary(qu_model)
```

```
##
## Call:
## lm(formula = RFEWIDTH ~ REDSHIFT + LINEFLUX + LUMINOSITY + AB1450)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -19.757  -9.039  -2.250   1.756  48.628
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 21087.951  18553.161   1.137   0.2691
## REDSHIFT      108.451     88.740   1.222   0.2359
## LINEFLUX      557.910    315.990   1.766   0.0927 .
## LUMINOSITY   -340.166    320.763  -1.060   0.3016
## AB1450         85.681      6.273  13.658 1.34e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 15.42 on 20 degrees of freedom
## Multiple R-squared:  0.9118, Adjusted R-squared:  0.8942
## F-statistic: 51.72 on 4 and 20 DF,  p-value: 2.867e-10
```

The 95% prediction interval for the fifth observation of the QUASAR data set is (90.69, 158.57), which implies that the RFEWIDTH (y) would fall between that interval for the fifth observation, as it does with RFEWIDTH = 114 according to the SPSS output on 194.

page 199 #4.28

(a) least squares prediction equation

y = 1041.89 - 13.23 AGE + 103.30 HOURS + 3.62 AGEHRS

(b) Estimated slope relating to annual earnings (y) to age (x1) when number of hours worked (x2) is 10: **22.97**

Estimated x1 slope = β_1 + β_3 * x_2 represents the change in E(y) for every 1-unit increase in, x_1, holding x_2 fixed

Estimated x1 slope = - 13.23 + (3.62 * 10)

(c) Estimated slope relating annual earnings (y) to hours worked (x2) when age (x1) is 40: **248.1**

Estimated x2 slope = β_2 + β_3 * x_1 represents the change in E(y) for every 1-unit increase in, x_2, holding x_1 fixed

Estimated x2 slope = 103.30 + (3.62 * 40)

(d) Null hypothesis for whether age (x1) and hours worked (x2) interact:

H_0: $\beta_1 = \beta_2 = \beta_3 = 0$

(e) p-value from part d: 0.0124 from text book page 199, SAS printout

(f) An $\alpha = 0.05$ does exceed the p-value, so this model is statistically useful predictor of annual earnings.

page 207 #4.40

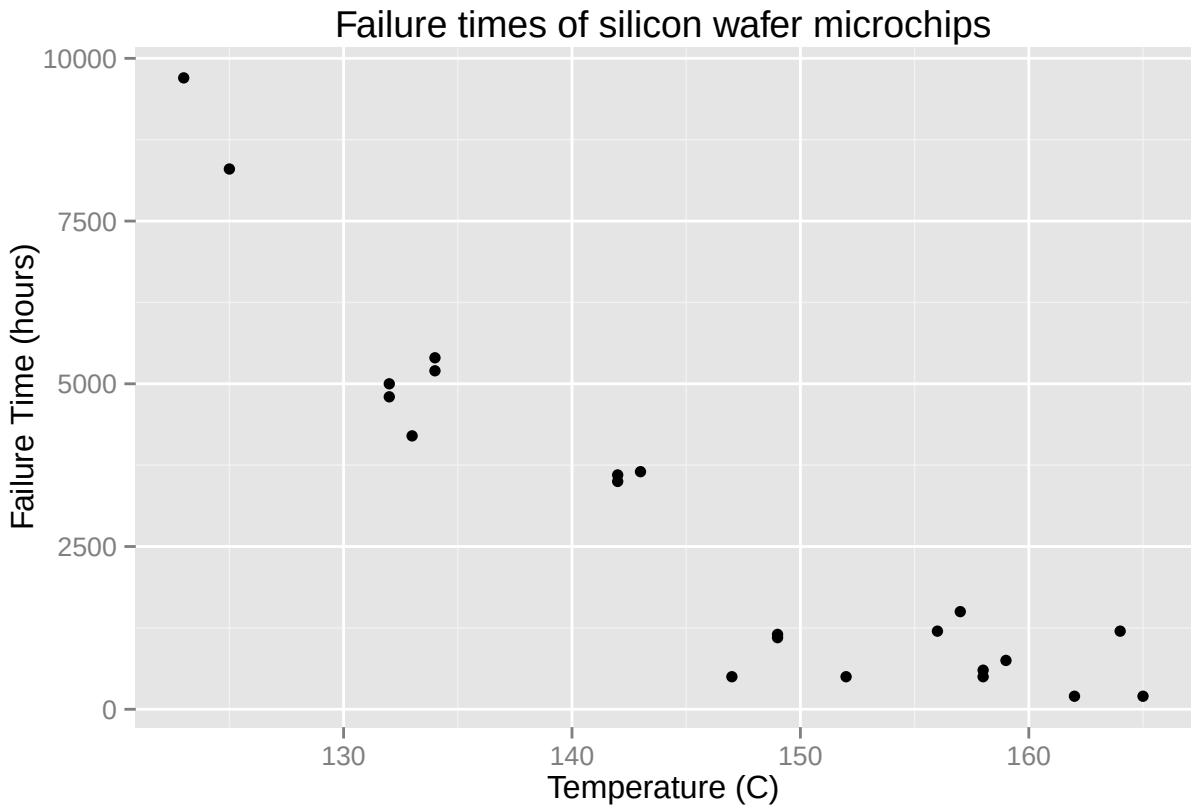
```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/WAFER.Rdata")
head(WAFER)
```

```
##   TEMP FAILTIME
## 1  165      200
## 2  162      200
## 3  164     1200
## 4  158      500
## 5  158      600
## 6  159      750
```

(a) construct a scatter plot: The relationship between temperature and failure time appears curvilinear

```
library(ggplot2)
TEMP.x <- WAFER$TEMP
FAILTIME.y <- WAFER$FAILTIME
wafer.df = data.frame(TEMP.x, FAILTIME.y)
wafer.plot <- ggplot(wafer.df, aes(TEMP.x, FAILTIME.y)) +
               geom_point() + # scatterplot
               labs(title = "Failure times of silicon wafer microchips",
                    x = "Temperature (C)", y = "Failure Time (hours)")

print(wafer.plot)
```



(b) fit model and give least squares prediction equation:

y = 154242.914 - 1908.850x + 5.929x^2

```
temp_sq <- TEMP.x^2 # to get the second-order term
wf_model = lm(FAILTIME.y ~ TEMP.x + temp_sq)
summary(wf_model)
```

```
##
## Call:
## lm(formula = FAILTIME.y ~ TEMP.x + temp_sq)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1260.49  -475.70   -15.57   528.45  1131.69
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 154242.914  21868.474   7.053 1.03e-06 ***
## TEMP.x       -1908.850    303.664  -6.286 4.92e-06 ***
## temp_sq          5.929      1.048   5.659 1.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 688.1 on 19 degrees of freedom
## Multiple R-squared:  0.9415, Adjusted R-squared:  0.9354
## F-statistic: 152.9 on 2 and 19 DF,  p-value: 1.937e-12
```

(c) Test to determine if there is an upward curvature in the relationship between failure time and solder temperature. $\alpha = 0.05$: The figure shows an upward curvature in the relationship between failure time and temperature. To test this we can state:

H_0: $\beta_2 = 0$ (no curvature in response curve)

H_a: $\beta_2 > 0$ (upward concavity exist in the response curve)

The test statistic for β_2 is t = 5.659 and the associated two-tailed p-value is 1.86e-05; where the one-tail appropriate p-value is `1.86e-05 / 2`. The alpha exceeds this p-value, there is strong evidence of upward curvature in the population, implying Failure time increases more faster per unit than increase in Temperature.

page 226 #4.59

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/WHEATRNA.Rdata")
head(WHEATRNA)
```

```
##    RNA MNSOD PLD   X1   Y X2   X1SQ X1X2 X1SQX2
## 1 0.00   401  80 0.00 401  1 0.0000 0.00 0.0000
## 2 0.00   336  83 0.00 336  1 0.0000 0.00 0.0000
## 3 0.00   337  75 0.00 337  1 0.0000 0.00 0.0000
## 4 0.33   711 132 0.33 711  1 0.1089 0.33 0.1089
## 5 0.33   637 148 0.33 637  1 0.1089 0.33 0.1089
## 6 0.33   602 115 0.33 602  1 0.1089 0.33 0.1089
```

(a) least square prediction equation from MINITAB printout on page 226:

y = 80.2 + 156x_1 - 42x_1^2 + 273x_2 + 760 x_1x_2 + 47 $x_1x_2^2$

(b) determine model usefulness: F = 417.05, p-value = 0.000, since $\alpha = 0.01$ is greater than the p-value, this model is statistically useful for predicting transcript copy number (y)

(c) Transcript copy number (y) could be curvilinear related to proportion of RNA to x_1, because x_1 is positive.

Submit: page 271 #5.8

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/GASTURBINE.Rdata")
head(GASTURBINE, n=5)
```

```
##        ENGINE SHAFTS   RPM CPRATIO INLETTEMP EXHTEMP AIRFLOW POWER
## 1 Traditional      1 27245     9.2      1134     602       7  1630
## 2 Traditional      1 14000    12.2       950     446      15  2726
## 3 Traditional      1 17384    14.8      1149     537      20  5247
## 4 Traditional      1 11085    11.8      1024     478      27  6726
## 5 Traditional      1 14045    13.2      1149     553      29  7726
##   HEATRATE  LHV ISOWORK
## 1    14622 24.6  232.86
## 2    13196 27.3  181.73
## 3    11948 30.1  262.35
## 4    11289 31.9  249.11
## 5    11964 30.1  266.41
```

```
tail(GASTURBINE, n=5)
```

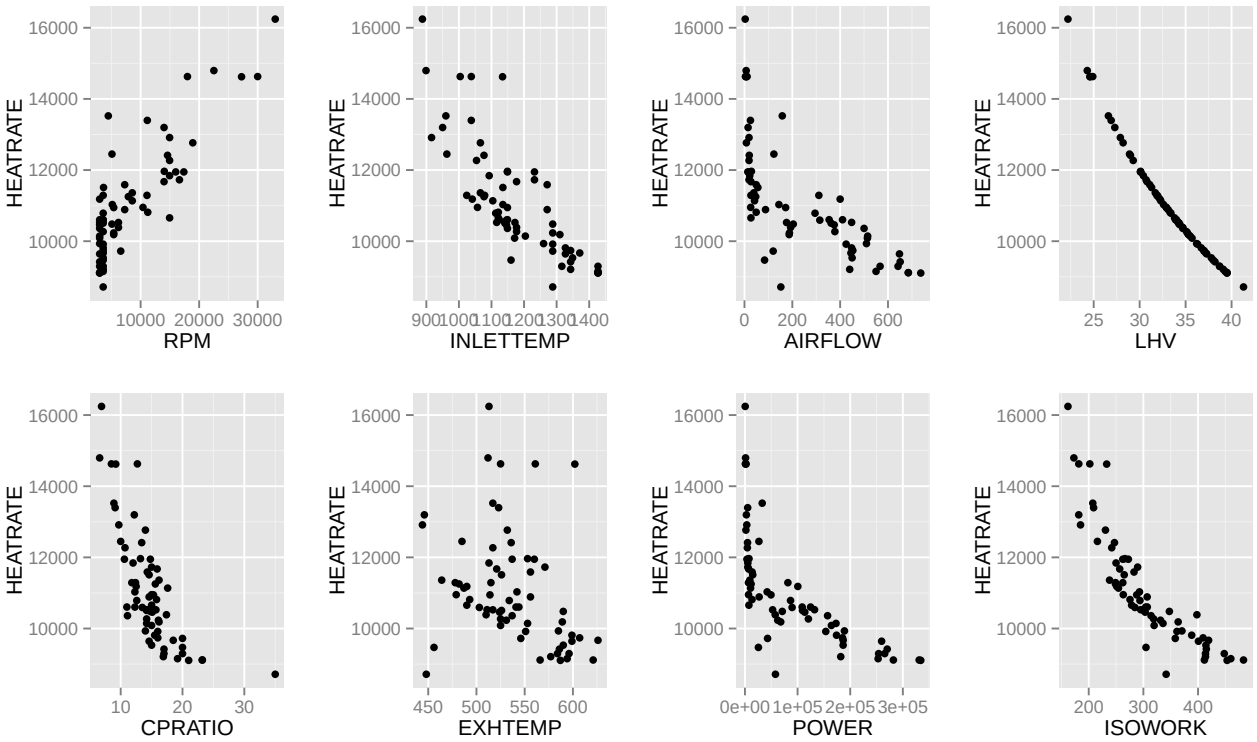
```
##         ENGINE SHAFTS   RPM CPRATIO INLETTEMP EXHTEMP AIRFLOW POWER
## 63 Aeroderiv        2 18910    14.0      1066     532       8  1845
## 64 Aeroderiv        3  3600    35.0      1288     448     152 57930
## 65 Aeroderiv        3  3600    20.0      1160     456      84 25600
## 66 Aeroderiv        2 16000    10.6      1232     560      14  3815
## 67 Aeroderiv        1 14600    13.4      1077     536      20  4942
##    HEATRATE  LHV ISOWORK
## 63    12766 28.2  230.63
## 64     8714 41.3  341.64
## 65     9469 38.0  304.76
## 66    11948 30.1  272.50
## 67    12414 29.0  247.10
```

Scatter plots relating heat rate (y) to each of the independent variables:

```
library(ggplot2)

rpm <- ggplot(GASTURBINE, aes(y = HEATRATE, x = RPM)) + geom_point()
cpratio <- ggplot(GASTURBINE, aes(y = HEATRATE, x = CPRATIO)) + geom_point()
inlettemp <- ggplot(GASTURBINE, aes(y = HEATRATE, x = INLETTEMP)) + geom_point()
exhtemp <- ggplot(GASTURBINE, aes(y = HEATRATE, x = EXHTEMP)) + geom_point()
airflow <- ggplot(GASTURBINE, aes(y = HEATRATE, x = AIRFLOW)) + geom_point()
power <- ggplot(GASTURBINE, aes(y = HEATRATE, x = POWER)) + geom_point()
lhv <- ggplot(GASTURBINE, aes(y = HEATRATE, x = LHV)) + geom_point()
isowork <- ggplot(GASTURBINE, aes(y = HEATRATE, x = ISOWORK)) + geom_point()

# function credit: www.cookbook-r.com/Graphs/Multiple_graphs_on_one_page_(ggplot2)/
source('~/Desktop/depaul/CSC423/multiplot.R')
multiplot(rpm, cpratio, inlettemp, exhtemp, airflow, power, lhv, isowork, cols = 4)
```



Hypothesize a polynomial model relating y to each independent variable for each plot:

1. Heat rate vs RPM: positive linear relationship
2. Heat rate vs INLETTEMP: negative linear relationship
3. Heat rate vs AIRFLOW: quadratic second-order model relationship
4. Heat rate vs LHV: quadratic second-order model relationship
5. Heat rate vs CPRATIO: quadratic second-order model relationship
6. Heat rate vs EXHTEMP: negative linear relationship
7. Heat rate vs POWER: quadratic second-order model relationship
8. Heat rate vs ISOWORK: quadratic second-order model relationship

page 272 #5.10

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/TIRES2.Rdata")
head(TIRES2, n=7)
```

```
##   X_PSI Y_THOUS
## 1    30      29
## 2    31      32
## 3    32      36
## 4    33      38
## 5    34      37
## 6    35      33
## 7    36      26
```

(a) Scatter plot of data:

```
rough <- ggplot(TIRES2, aes(y = Y_THOUS, x = X_PSI)) +
              geom_point(colour = "darkblue") +
              labs(title = " Tire wear and pressure",
                   x ="Pressure (lbs per square inch)", y = "Mileage (Thousands)")
print(rough)
```

(b) for x = 30, 31, 32, 33 the type of model I would suggest would be a positive linear relationship, a simple first-order model. For x = 33, 34, 35, 36 the type of model I would suggest would be curvilinear, negative second-order model.

page 281 #5.17

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/QUASAR.Rdata")
head(QUASAR)
```

```
##   QUASAR REDSHIFT LINEFLUX LUMINOSITY AB1450 ABSMAG RFEWIDTH
## 1      1     2.81   -13.48      45.29  19.50 -26.27      117
## 2      2     3.07   -13.73      45.13  19.65 -26.26       82
## 3      3     3.45   -13.87      45.11  18.93 -27.17       33
## 4      4     3.19   -13.27      45.63  18.59 -27.39       92
## 5      5     3.07   -13.56      45.30  19.59 -26.32      114
## 6      6     4.15   -13.95      45.20  19.42 -26.97       50
```

(a) Complete second-order model with: REDSHIFT = x1 LINEFLUX = x2 AB1450 = x4

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + \beta_4 x_1 x_2 + \beta_5 x_1 x_4 + \beta_6 x_2 x_4 + \beta_7 x_1^2 + \beta_8 x_2^2 + \beta_9 x_4^2$$

(b) Fit the model. Since with an assumed $\alpha = 0.05$ is larger than the p-value of 2.2e-16, the model is statistically useful as a predictor of rest frame width (y)

```
y = QUASAR$RFEWIDTH
x1 = QUASAR$REDSHIFT
x2 = QUASAR$LINEFLUX
x4 = QUASAR$AB1450

qu.model <- lm(y ~ x1 + x2 + x4 + I(x1*x2) + I(x1*x4) + I(x2*x4) + I(x1^2) + I(x2^2) + I(x4^2))

summary(qu.model)
```

```
##
## Call:
## lm(formula = y ~ x1 + x2 + x4 + I(x1 * x2) + I(x1 * x4) + I(x2 *
##     x4) + I(x1^2) + I(x2^2) + I(x4^2))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.5046 -0.7861  0.4722  1.4742  3.9819
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 1.028e+04  2.336e+03   4.403 0.000514 ***
## x1          2.768e+02  1.470e+02   1.882 0.079345 .
## x2          3.325e+03  4.765e+02   6.978 4.44e-06 ***
## x4          1.301e+03  1.152e+02  11.294 9.85e-09 ***
## I(x1 * x2)  4.198e+01  1.433e+01   2.930 0.010344 *
## I(x1 * x4)  1.598e+01  4.237e+00   3.771 0.001848 **
## I(x2 * x4)  2.074e+02  1.183e+01  17.533 2.11e-11 ***
## I(x1^2)     9.895e-01  4.302e+00   0.230 0.821199
## I(x2^2)     2.666e+02  2.484e+01  10.731 1.96e-08 ***
## I(x4^2)     4.022e+01  2.428e+00  16.563 4.75e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Residual standard error: 3.121 on 15 degrees of freedom
## Multiple R-squared:  0.9973, Adjusted R-squared:  0.9957
## F-statistic: 613.3 on 9 and 15 DF,  p-value: < 2.2e-16
```

(c) The Pr(<|t|) value for x2 and x4 are significantly lower than the assumed $\alpha = 0.05$, which means that LINEFLUX and AB1450 are statistically useful predictors for rest frame width (y)

page 287 #5.22

Potential for extreme round-off errors in the parameter estimates for the model: because the independent variable, Temperature (x) has a narrow range. [123 to 165]

Propose and fit another alternative model: The correlation value of the independent variable is highly correlated (r = 0.999273) with the original model. Once the x variable has been coded and the model re-ran the new value is much lower (r = -0.2389866), implying the round-off error problem has been fixed.

```
# get the correlation value r to see that the values are highly correlated
cor(TEMP.x, temp_sq)
```

```
## [1] 0.999273
```

```
# function to code data from x to u (z score)
coded <- function(x){
  (x - mean(WAFER$TEMP)) / sd(WAFER$TEMP)
}
# lapply the function over the Temperature variable
coded.val = lapply(WAFER$TEMP, coded)
# add the new set of data to the WAFER data set and change into a numeric
WAFER$TEMPCODED <- as.numeric(coded.val)
# second order of the new coded temperature varaible
WAFER$TEMPCODEDsq <- (WAFER$TEMPCODED)^2
# new fited model
wf_model_coded = lm(FAILTIME ~ TEMPCODED + TEMPCODEDsq, data = WAFER)
summary(wf_model_coded)
```

```
##
## Call:
## lm(formula = FAILTIME ~ TEMPCODED + TEMPCODEDsq, data = WAFER)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -1260.49  -475.70   -15.57   528.45  1131.69
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   1900.1      223.2   8.512 6.60e-08 ***
## TEMPCODED    -2275.7      154.6 -14.716 7.70e-12 ***
## TEMPCODEDsq    997.6      176.3   5.659 1.86e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 688.1 on 19 degrees of freedom
## Multiple R-squared:  0.9415, Adjusted R-squared:  0.9354
## F-statistic: 152.9 on 2 and 19 DF,  p-value: 1.937e-12
```

```
# get the correlation value r to see if the rounding error has been decreased
cor(WAFER$TEMPCODED, WAFER$TEMPCODEDsq)
```

```
## [1] -0.2389866
```

page 302 #5.27

(a) Interaction model relating wine quality to grape-picking method and soil type:

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_1 x_2 + \beta_5 x_1 x_3$

Main effects model:
$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$

where:

$x_1 = \{1$ if soil type gravel and 0 if not

$x_2 = \{1$ if soil type clay and 0 if not

$x_3 = \{1$ if manual was used and 0 if automated

(b) The β_0 in the model is the mean value of y at the base levels of "Automated" and "sand"

(c) For "manual" and "clay":

$x_1 = 0\ x_2 = 1\ x_3 = 1$

$E(y) = \beta_0 + \beta_1\ (0) + \beta_2\ (1) + \beta_3\ (1)$

$E(y) = \beta_0 + \beta_2 + \beta_3$

(d) The difference between the E(y), mean wine quality is β_3 for the following conditions:

"manual" and "sand": $x_1 = 0\ x_2 = 0\ x_3 = 1$

$E(y) = \beta_0 + \beta_3$

and

"automated" and "sand": $x_1 = 0\ x_2 = 0\ x_3 = 0$

$E(y) = \beta_0$

page 303 #5.30

The flaw in the model is that x = 0 for lecturer would equal the same as the β_0 since the β_1 (0) would drop out of the model. An alternative curvilinear model with a squared term could suffice, since there would be an eventual plateau in salary from the increase in professor rank. The second order model would result in the administration's objective.

page 320 #5.42

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/LASERS.Rdata")
LASERS
```

```
##   CURRENT WAVEGUIDE
## 1     273      0.15
## 2     175      0.20
## 3     146      0.25
## 4     166      0.30
## 5     162      0.35
## 6     165      0.40
## 7     245      0.50
## 8     314      0.60
```

(a) The equation for the coded variable, u to waveguide, x:

(x - mean(LASERS $ WAVEGUIDE)) / sd(LASERS $ WAVEGUIDE)

(b)

```
# function to code data from x to u (z score)
coded.laser <- function(x){
  (x - mean(LASERS$WAVEGUIDE)) / sd(LASERS$WAVEGUIDE)
}
# lapply the function over the waveguide variable
coded.laser.sharks = lapply(LASERS$WAVEGUIDE, coded.laser)
print(coded.laser.sharks)
```

```
## [[1]]
## [1] -1.272864
##
## [[2]]
## [1] -0.944383
##
## [[3]]
## [1] -0.615902
##
## [[4]]
## [1] -0.2874209
##
## [[5]]
## [1] 0.04106013
##
## [[6]]
## [1] 0.3695412
##
## [[7]]
## [1] 1.026503
##
## [[8]]
## [1] 1.683465
```

(c) Coefficient of correlation between x and x^2: r = 0.9839865

```
LASERS$WAVEGUIDESQ = (LASERS$WAVEGUIDE)^2
cor(LASERS$WAVEGUIDE, LASERS$WAVEGUIDESQ)
```

```
## [1] 0.9839865
```

(d) Coefficient of correlation between u and u^2: r = 0.4030942

```
LASERS$waveguide.coded = as.numeric(coded.laser.sharks)
LASERS$waveguide.coded.sq = (LASERS$waveguide.coded)^2
cor(LASERS$WAVEGUIDE, LASERS$waveguide.coded.sq)
```

```
## [1] 0.4030942
```

The two values are different in that WAVEGUIDE variable was highly correlated as it was a narrow range.

(d) Model fitted with u and u^2:

```
solar.model <- lm(CURRENT ~ waveguide.coded + waveguide.coded.sq, data = LASERS)
summary(solar.model)
```

```
##
## Call:
## lm(formula = CURRENT ~ waveguide.coded + waveguide.coded.sq,
##     data = LASERS)
##
## Residuals:
##         1         2         3         4         5         6         7
##  31.02977 -26.54641 -27.63398   7.76708   6.65674   0.03503  23.25744
##         8
## -14.56567
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)         155.020     13.162  11.778 7.76e-05 ***
## waveguide.coded       5.486     10.592   0.518  0.62660
## waveguide.coded.sq   57.977     10.903   5.318  0.00315 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 25.65 on 5 degrees of freedom
## Multiple R-squared:  0.8802, Adjusted R-squared:  0.8323
## F-statistic: 18.37 on 2 and 5 DF,  p-value: 0.004968
```

The p-value is lower than the $\alpha = 0.05$ which means that this model is a statistically useful for determining Threshold Current.

page 321 #5.44

(a) Interaction model

$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_1 x_2 + \beta_7 x_1 x_2 + \beta_7 x_1 x_3 + \beta_8 x_1 x_4 + \beta_9 x_1 x_5 + \beta_{10} x_2 x_3 + \beta_{11} x_2 x_4 + \beta_{12} x_2 x_5 + \beta_{13} x_2 x_4 + \beta_{14} x_3 x_4 + \beta_{14} x_3 x_5 + \beta_{15} x_4 x_5$

(b) β_1 is the college level Business Administration

(c) β_2 is the college level Engineering

(d) β_3 is the college level Liberal Arts & Sciences

(e) β_4 is the college level Journalism

(f) β_5 is the Gender

(g) The inteaction terms test against each college configuration and the gender to determine the mean starting salary. To test this more specifically you could drop all the beta-term that do not include the x_5 then there would a simplified model testing the gender against each college level (ie)

$E(y) = E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_9 x_1 x_5 + \beta_{12} x_2 x_5 + \beta_{14} x_3 x_5 + \beta_{15} x_4 x_5$

page 323 #5.51 (include graphs of interaction terms)

```
load("~/Desktop/depaul/CSC423/rdata/R/Exercises&Examples/SYNFUELS.Rdata")
head(SYNFUELS)
```

```
##   BrakePow FuelType X1 X2 BurnRate
## 1        4      DF2  1  0     13.2
## 2        4      BLN  0  1     17.5
## 3        4      ADV  0  0     17.5
## 4        6      DF2  1  0     26.1
## 5        6      BLN  0  1     32.7
## 6        6      ADV  0  0     43.5
```

(a) Test if the brake power and fuel interact with $\alpha = 0.01$. Since the p-value is lower than the α with an F-statistic of 25.65 so that concludes that brake power and fuel type contribute to the prediction of Burn Rate (y).

```
power = SYNFUELS$BrakePow # x1
x2 =  SYNFUELS$X1 # x2
x3 = SYNFUELS$X2 # x3
burn = SYNFUELS$BurnRate # y

diesel <- lm(burn ~ power + x2 + x3 + I(power*x2) + I(power*x3))
summary(diesel)
```

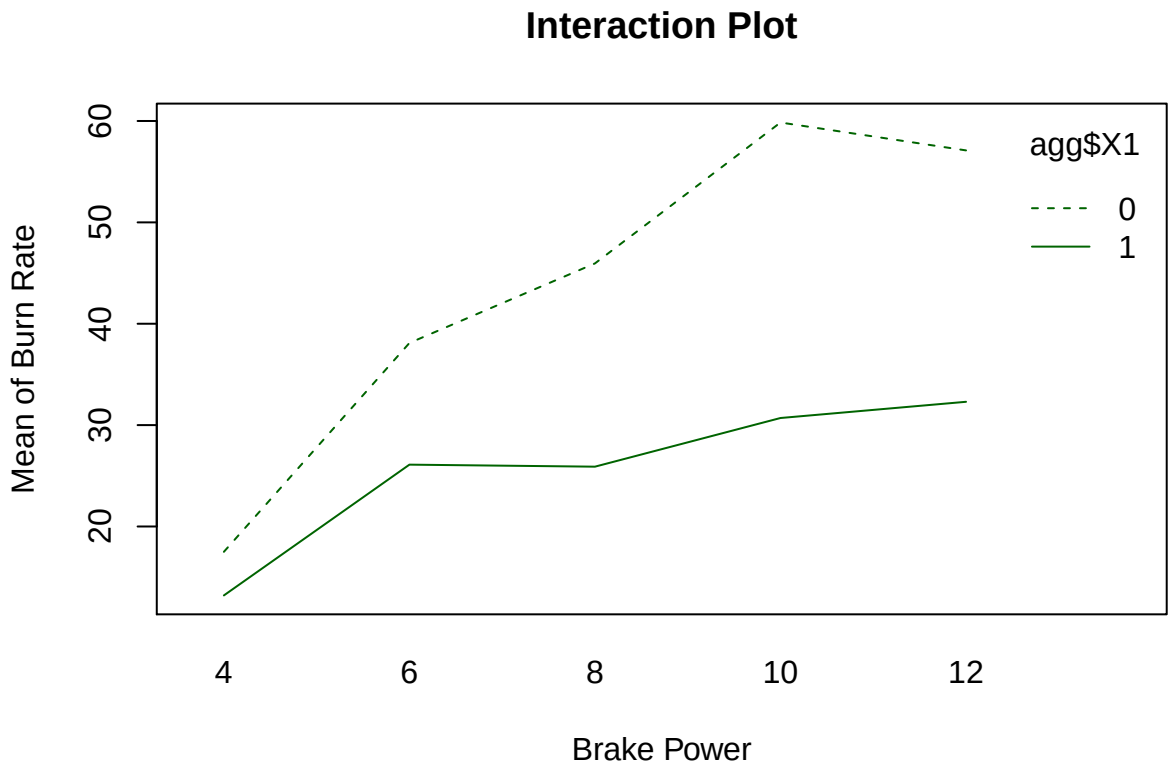
```
##
## Call:
## lm(formula = burn ~ power + x2 + x3 + I(power * x2) + I(power *
##     x3))
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.090 -3.163  0.225  1.573  7.440
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)    -10.830      8.277  -1.308  0.22708
## power            7.815      1.126   6.938  0.00012 ***
## x2              19.350     10.686   1.811  0.10777
## x3              12.790     10.686   1.197  0.26561
## I(power * x2)   -5.675      1.380  -4.114  0.00337 **
## I(power * x3)   -2.950      1.380  -2.138  0.06494 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.037 on 8 degrees of freedom
## Multiple R-squared:  0.9413, Adjusted R-squared:  0.9046
## F-statistic: 25.65 on 5 and 8 DF,  p-value: 9.984e-05
```

(b) Interaction graphs + slope

```
agg = aggregate(BurnRate ~ BrakePow + X1 + X2, data = SYNFUELS,  mean)
agg
```

```
##    BrakePow X1 X2 BurnRate
## 1         4  0  0     17.5
## 2         6  0  0     43.5
## 3         8  0  0     45.6
## 4        10  0  0     68.9
## 5         4  1  0     13.2
## 6         6  1  0     26.1
## 7         8  1  0     25.9
## 8        10  1  0     30.7
## 9        12  1  0     32.3
## 10        4  0  1     17.5
## 11        6  0  1     32.7
## 12        8  0  1     46.3
## 13       10  0  1     50.8
## 14       12  0  1     57.1
```

```
interaction.plot(agg$BrakePow, agg$X1, response = agg$BurnRate, main="Interaction Plot",xlab="Brake Pow
```



```r
interaction.plot(agg$BrakePow, agg$X2, response = agg$BurnRate, main="Interaction Plot",xlab="Brake Pow
```