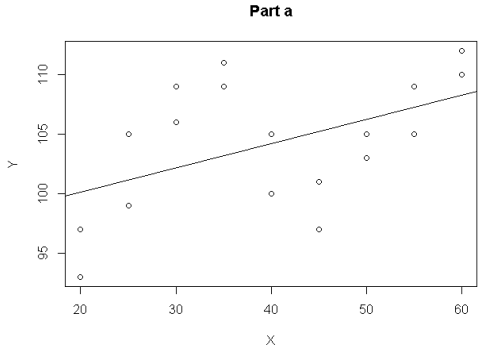
STAT 420 – Homework 8

1. Productivity Data

a. The *F*-test suggests that the model is significant. However, looking at the scatterplot and line, there is no way that we would want to use the simple linear model to fit this data. If you go to the next step and look at the residual plot, you'd see the model assumptions are violated.

```
> plot(X,Y, main="Part a")
> fit.a <- lm(Y ~ X)
> abline(fit.a)
```




```
> summary(fit.a)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  96.0889     3.7254   25.79  1.8e-14 ***
X             0.2033     0.0886    2.29    0.036 *
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.85 on 16 degrees of freedom
Multiple R-squared:  0.248,      Adjusted R-squared:   0.2
F-statistic: 5.26 on 1 and 16 DF,  p-value: 0.0357
```

b. The quadratic model has a *p*-value > 0.05 and fails the *F*-test. Further, neither of the model terms is significant when we look at the *t*-tests for the individual coefficients.

```
> summary(fit.b)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 92.11775   12.15652    7.58  1.7e-06 ***
X            0.42498    0.65033    0.65     0.52
I(X^2)      -0.00277    0.00805   -0.34     0.74
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.99 on 15 degrees of freedom
Multiple R-squared:  0.253,      Adjusted R-squared:  0.154
F-statistic: 2.55 on 2 and 15 DF,  p-value: 0.112
```

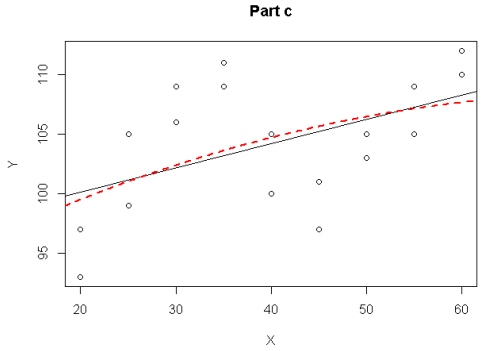
Also, the ANOVA test comparing the two models says that the simple linear model is better.

```
> anova(fit.a, fit.b)
Analysis of Variance Table
```

```
Model 1: Y ~ X
Model 2: Y ~ X + I(X^2)
  Res.Df RSS Df Sum of Sq    F Pr(>F)
1     16 377
2     15 374  1      2.96 0.12   0.74
```

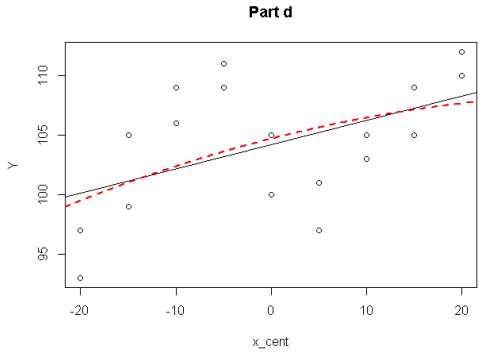
c. Over this domain the curvature of the quadratic model (red) isn't that much different from the slope of the simple linear model (black). This, coupled with the results of part b, suggest that there is collinearity between the two predictor terms.

```
> plot(X, Y, main="Part c")
> abline(fit.a)
> xplot <- seq(15,65,by=0.1)
> lines(xplot, predict(fit.b, newdata = data.frame(X = xplot)),
+       col="red", lty = 2, lwd=2)
```



d. The quadratic model with centralized X values does not provide a better fit than the quadratic model of part b. Why? They are essentially the same model, at least globally. Note that the F-test, R^2, RSE, and statistics for the x^2 term in the two models are all exactly the same as back in part b.

```
> x_cent <- X - mean(X)
> fit.d <- lm(Y ~ x_cent)
> fit2.d <- lm(Y ~ x_cent + I(x_cent^2))
> plot(x_cent, Y, main="Part d")
> abline(fit.d)
> xplot <- seq(-25,25,by=0.1)
> lines(xplot, predict(fit2.d, newdata = data.frame(x_cent = xplot)),
+ col="red", lty = 2, lwd=2)
```




```
> summary(fit2.d)
```

```
Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 104.68398    1.78470   58.66   <2e-16 ***
x_cent        0.20333    0.09118    2.23    0.041 *
I(x_cent^2)  -0.00277    0.00805   -0.34    0.735
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.99 on 15 degrees of freedom
Multiple R-squared:  0.253,      Adjusted R-squared:  0.154
F-statistic: 2.55 on 2 and 15 DF,  p-value: 0.112
```

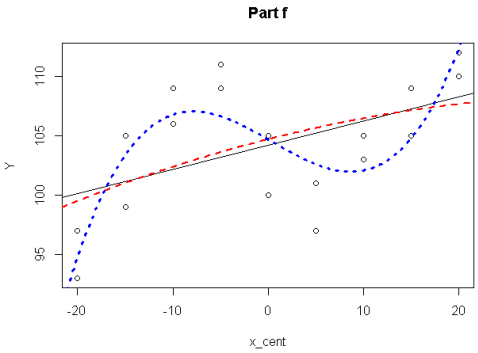
So why bother? Note that the significance of the x_cent term is now significant in the new model. We have removed the collinearity and should continue using the centralized values. But the overall fit is not good, so let's try introducing a higher-order term.

e. The cubic (third-order) model with *Y* as the response is much better than any seen thus far. We could continue to try to improve the fit by examining <u>odd</u> higher order models.

```
> fit3 <- lm(Y ~ x_cent + I(x_cent^2) + I(x_cent^3))
> summary(fit3)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.05e+02   1.17e+00   89.66  < 2e-16 ***
x_cent      -4.59e-01   1.56e-01   -2.94  0.01078 *
I(x_cent^2) -2.77e-03   5.26e-03   -0.53  0.60699
I(x_cent^3)  2.25e-03   4.89e-04    4.59  0.00042 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 3.27 on 14 degrees of freedom
Multiple R-squared:  0.702,      Adjusted R-squared:  0.638
F-statistic:   11 on 3 and 14 DF,  p-value: 0.000566

> anova(fit2.d, fit3)
Analysis of Variance Table

Model 1: Y ~ x_cent + I(x_cent^2)
Model 2: Y ~ x_cent + I(x_cent^2) + I(x_cent^3)
  Res.Df RSS Df Sum of Sq    F  Pr(>F)
1     15 374
2     14 149  1       225 21.1 0.00042 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

f. Add the best fit third-order line from the model in part e to the scatterplot.

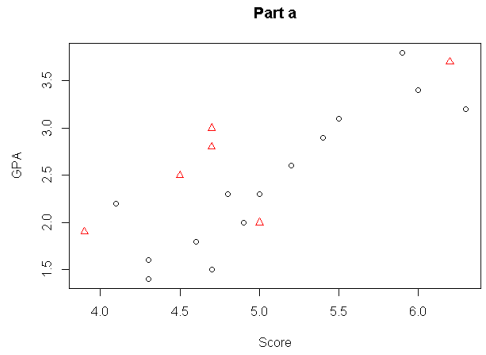
```
> lines(xplot, predict(fit3, newdata = data.frame(x_cent = xplot)),
+ col="blue", lty = 3, lwd=3)
```



2. Admissions Data

a. Overall, it seems that those who have declared a major tend to have higher first-year GPAs than those who do not declare. So it does seem Major is an important covariate.

```
> plot(Score, GPA, main="Part a", col=Major+1, pch=Major+1)
```



b. The R^2 value tells us that 72.9% of the variation in GPA is explained by this model.

```
> fit.b <- lm(GPA ~ Score + Major)
> summary(fit.b)

Call:
lm(formula = GPA ~ Score + Major)

Residuals:
    Min      1Q  Median      3Q     Max
-0.7978 -0.2343  0.0639  0.1968  0.6297

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.060      0.682   -3.02   0.0077 **
Score          0.887      0.133    6.67  3.9e-06 ***
Major          0.425      0.196    2.17   0.0443 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.396 on 17 degrees of freedom
Multiple R-squared:  0.729,      Adjusted R-squared:  0.697
F-statistic: 22.9 on 2 and 17 DF,  p-value: 1.51e-05
```

c. The estimated coefficient for the Score term tells us that for each additional point on the admissions test, a student's first-year GPA is expected to increase by 0.887 points on average. The estimated coefficient for the Major term tells us that those who have declared a major can expect to see a 0.425 point (additive) increase to their GPA on average as compared to those who do not declare a major.

d. The R^2 value tells us that 74.6% of the variation in GPA is explained by this model. (That's a very marginal increase in comparison to the model in part b.)

```
> fit.d <- lm(GPA ~ Score + Major + Score:Major)
> summary(fit.d)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -2.547      0.827   -3.08   0.0072 **
Score          0.983      0.162    6.07  1.6e-05 ***
Major          1.862      1.401    1.33   0.2023
Score:Major   -0.293      0.282   -1.04   0.3156
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.395 on 16 degrees of freedom
Multiple R-squared:  0.746,      Adjusted R-squared:  0.698
F-statistic: 15.7 on 3 and 16 DF,  p-value: 5.08e-05
```

e. The *p*-value is rather large, so we fail to reject the null hypothesis (smaller model). Thus, the additive model (non-interaction model) from part b is preferred.

```
> anova(fit.b, fit.d)
Analysis of Variance Table

Model 1: GPA ~ Score + Major
Model 2: GPA ~ Score + Major + Score:Major
  Res.Df  RSS Df Sum of Sq    F Pr(>F)
1     17 2.67
2     16 2.50  1     0.168 1.07   0.32
```

3. Prostate Data

a. First, we fit the full model.

```
> fit.a1 <- lm(lpsa ~ ., prostate)
> summary(fit.a1)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.66934    1.29639    0.52   0.6069
lcavol       0.58702    0.08792    6.68  2.1e-09 ***
lweight      0.45447    0.17001    2.67   0.0090 **
age         -0.01964    0.01117   -1.76   0.0823 .
lbph         0.10705    0.05845    1.83   0.0704 .
svi          0.76616    0.24431    3.14   0.0023 **
lcp         -0.10547    0.09101   -1.16   0.2496
gleason      0.04514    0.15746    0.29   0.7750
pgg45        0.00453    0.00442    1.02   0.3089
```

Since `gleason` has the highest *p*-value above $\alpha = 0.10$ among the predictors, that is the first variable to be eliminated. Run a new model with `gleason` removed.

```
> fit.a2 <- lm(lpsa ~ . - gleason, prostate)
> summary(fit.a2)


Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.95393    0.82944    1.15   0.2532
lcavol       0.59161    0.08600    6.88  8.1e-10 ***
lweight      0.44829    0.16777    2.67   0.0090 **
age         -0.01934    0.01107   -1.75   0.0840 .
lbph         0.10767    0.05811    1.85   0.0672 .
svi          0.75773    0.24128    3.14   0.0023 **
lcp         -0.10448    0.09048   -1.15   0.2513
pgg45        0.00532    0.00343    1.55   0.1249
```

Now remove `lcp`.

```
> fit.a3 <- lm(lpsa ~ . - gleason - lcp, prostate)
> summary(fit.a3)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.98008    0.83066    1.18   0.2412
lcavol       0.54577    0.07643    7.14  2.3e-10 ***
lweight      0.44945    0.16808    2.67   0.0089 **
age         -0.01747    0.01097   -1.59   0.1147
lbph         0.10576    0.05819    1.82   0.0725 .
svi          0.64167    0.21976    2.92   0.0044 **
pgg45        0.00353    0.00307    1.15   0.2533
```

Eliminate `pgg45`.

```
> fit.a4 <- lm(lpsa ~ . - gleason - lcp - pgg45, prostate)
> summary(fit.a4)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.9510     0.8317    1.14  0.25588
lcavol        0.5656     0.0746    7.58  2.8e-11 ***
lweight       0.4237     0.1669    2.54  0.01281 *
age          -0.0149     0.0108   -1.38  0.16953
lbph          0.1118     0.0581    1.93  0.05716 .
svi           0.7210     0.2090    3.45  0.00085 ***
```

Eliminate `age`.

```
> fit.a5 <- lm(lpsa ~ . - gleason - lcp - pgg45 - age, prostate)
> summary(fit.a5)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   0.1455     0.5975    0.24    0.808
lcavol        0.5496     0.0741    7.42  5.6e-11 ***
lweight       0.3909     0.1660    2.35    0.021 *
lbph          0.0901     0.0562    1.60    0.112
svi           0.7117     0.2100    3.39    0.001 **
```

Eliminate lbph.

```
> fit.a6 <- lm(lpsa ~ . - gleason - lcp - pgg45 - age - lbph, prostate)
> summary(fit.a6)

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  -0.2681     0.5435   -0.49    0.623
lcavol        0.5516     0.0747    7.39  6.3e-11 ***
lweight       0.5085     0.1502    3.39    0.001 **
svi           0.6662     0.2098    3.18    0.002 **
```

And finally we have found a model where all predictors have an individual significance to the model below $\alpha = 0.10$. The final model utilizes lcavol, lweight, and svi as predictors.

b. Using backward variable selection with AIC as the selection criterion, the predictors gleason, lcp, and pgg45 are eliminated as in part a, but it stops there. Here, the final model utilizes lcavol, lweight, age, lbph, and svi as predictors

```
> fit.b <- step(fit.a1, direction="backward")
Start:  AIC=-58.32
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + gleason +
    pgg45

           Df Sum of Sq  RSS   AIC
- gleason  1      0.04 44.2 -60.2
- pgg45    1      0.53 44.7 -59.2
- lcp      1      0.67 44.8 -58.9
<none>                 44.2 -58.3
- age      1      1.55 45.7 -57.0
- lbph     1      1.68 45.8 -56.7
- lweight  1      3.59 47.7 -52.7
- svi      1      4.94 49.1 -50.0
- lcavol   1     22.37 66.5 -20.6

Step:  AIC=-60.23
lpsa ~ lcavol + lweight + age + lbph + svi + lcp + pgg45

           Df Sum of Sq  RSS   AIC
- lcp      1      0.66 44.9 -60.8
<none>                 44.2 -60.2
- pgg45    1      1.19 45.4 -59.7
- age      1      1.52 45.7 -59.0
- lbph     1      1.71 45.9 -58.6
- lweight  1      3.55 47.8 -54.7
- svi      1      4.90 49.1 -52.0
- lcavol   1     23.50 67.7 -20.9

Step:  AIC=-60.79
lpsa ~ lcavol + lweight + age + lbph + svi + pgg45

           Df Sum of Sq  RSS   AIC
- pgg45    1      0.66 45.5 -61.4
<none>                 44.9 -60.8
- age      1      1.26 46.1 -60.1
- lbph     1      1.65 46.5 -59.3
- lweight  1      3.56 48.4 -55.4
- svi      1      4.25 49.1 -54.0
- lcavol   1     25.42 70.3 -19.2

Step:  AIC=-61.37
```

```
lpsa ~ lcavol + lweight + age + lbph + svi

          Df Sum of Sq  RSS   AIC
<none>                 45.5 -61.4
- age      1      0.96 46.5 -61.4
- lbph     1      1.86 47.4 -59.5
- lweight  1      3.23 48.8 -56.7
- svi      1      5.95 51.5 -51.5
- lcavol   1     28.77 74.3 -15.9
```

c. We want a large value for R^2. Of course the smallest model – with only three predictors – from part a will have the lowest R^2 value (0.6264). But because its value isn't that far off from the other two models, we're okay with sacrificing a little R^2 in order to have a simpler model. By this rationale, the "best" model from part a is considered the best among these three models.

```
> summary(fit.a1)$r.squared   # full model
[1] 0.6548
> summary(fit.a6)$r.squared   # best from part a
[1] 0.6264
> summary(fit.b)$r.squared    # best from part b
[1] 0.6441
```