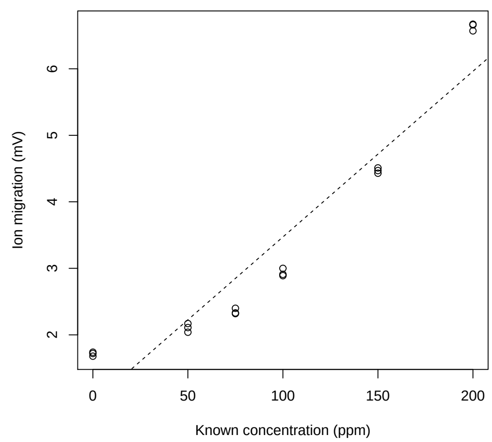
STAT 420 Spring 2014
Homework 7: Solutions

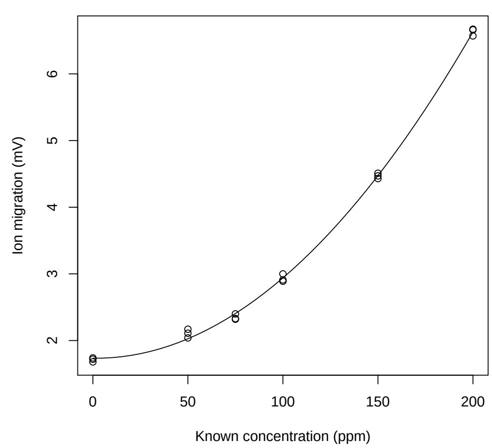
Exercise 1

(a) The scatterplot of the data (with best fitting linear regression line) is given below:



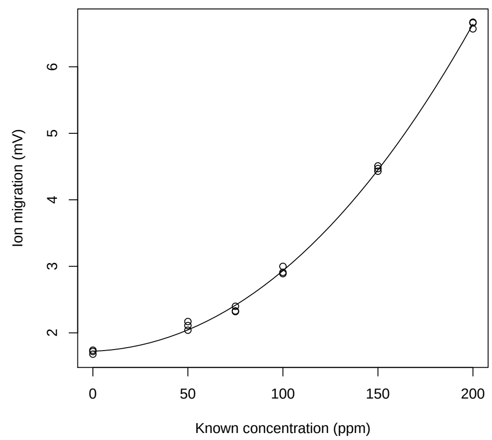
Note that the linear model does NOT seem appropriate here.

(b) The scatterplot of the data (with best fitting quadratic regression line) is given below:



Note that the quadratic model does seem appropriate here.

(c) The scatterplot of the data (with best fitting cubic regression line) is given below:



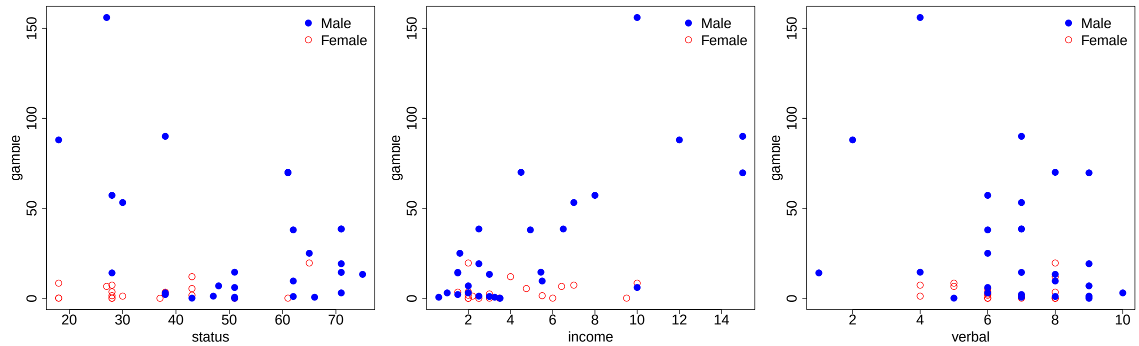
Note that the cubic effect is not significant, but the quadratic effect is significant.

```
> cmod=lm(mV~ppm+I(ppm^2)+I(ppm^3),data=ions)
> summary(cmod)

Call:
lm(formula = mV ~ ppm + I(ppm^2) + I(ppm^3), data = ions)

Residuals:
      Min        1Q    Median        3Q       Max
-0.095155 -0.039045 -0.004164  0.028073  0.125441

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.724e+00  3.631e-02  47.477  < 2e-16 ***
ppm         1.080e-03  1.668e-03   0.647 0.527929
I(ppm^2)    1.031e-04  2.171e-05   4.750 0.000311 ***
I(ppm^3)    7.188e-08  7.302e-08   0.984 0.341657
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 0.06347 on 14 degrees of freedom
Multiple R-squared:  0.9989,	Adjusted R-squared:  0.9987
F-statistic:  4325 on 3 and 14 DF,  p-value: < 2.2e-16
```

You should choose the quadratic model.

Exercise 2

(a) The plots of `gamble` versus the other predictors are given below:



All three plots suggest the need for the interaction term between `sex` and the other three predictors, since the rates of the relationships between `gamble` and `status`, `income`, and `verbal` are different for `sex=0` and `sex=1`.

(b)
```
> gmod=lm(gamble~status*sex+income*sex+verbal*sex,data=teengamb)
> summary(gmod)

Call:
lm(formula = gamble ~ status * sex + income * sex + verbal * sex, data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-56.654  -7.589  -1.016   3.323  83.903

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  27.6354    17.6218   1.568   0.1249
status       -0.1456     0.3316  -0.439   0.6631
sex         -33.0132    35.0530  -0.942   0.3521
income        6.0291     1.0538   5.721 1.26e-06 ***
verbal       -2.9748     2.4265  -1.226   0.2276
status:sex    0.3529     0.5492   0.643   0.5243
sex:income   -5.3478     2.4244  -2.206   0.0334 *
sex:verbal    2.8355     4.5973   0.617   0.5410
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 20.98 on 39 degrees of freedom
```

```
Multiple R-squared:  0.6243,        Adjusted R-squared:  0.5569
F-statistic:  9.26 on 7 and 39 DF,  p-value: 1.06e-06
```

Dropping the non-significant terms produces the model:

```
> amod=lm(gamble~income*sex,data=teengamb)
> summary(amod)

Call:
lm(formula = gamble ~ income * sex, data = teengamb)

Residuals:
    Min      1Q  Median      3Q     Max
-56.522  -4.860  -1.790   6.273  93.478

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  -2.6596     6.3164  -0.421  0.67580
income        6.5181     0.9881   6.597 4.95e-08 ***
sex           5.7996    11.2003   0.518  0.60724
income:sex   -6.3432     2.1446  -2.958  0.00502 **
---
Signif. codes:  0 `***' 0.001 `**' 0.01 `*' 0.05 `.' 0.1 ` ' 1

Residual standard error: 20.98 on 43 degrees of freedom
Multiple R-squared:  0.5857,        Adjusted R-squared:  0.5568
F-statistic: 20.26 on 3 and 43 DF,  p-value: 2.451e-08
```

Exercise 3

Suppose a complete second-order model

$$y_i = b_0 + b_1 x_{i1} + b_2 x_{i2} + b_3 x_{i1} x_{i2} + b_4 x_{i1}^2 + b_5 x_{i2}^2 + e_i$$

was fit to $n = 24$ observations.

```
> sum( lm( y ~ 1 )$residuals^2 )
[1] 360
> sum( lm(y ~ x1+x2)$residuals^2 )
[1] 126
> sum( lm(y ~ x1+x2+I(x1*x2))$residuals^2 )
```

```
[1] 100
> sum( lm(y ~ x1+x2+I(x1*x2)+I(x1^2)+I(x2^2))$residuals^2 )
[1] 72
```

(a) From the given information, we know that SSE = 72 and SST = 360; this implies that SSR = SST − SSE = 360 − 72 = 288. Next note that $df_{\text{SST}} = 23$ (because $n = 24$), and $df_{\text{SSE}} = 18$ and $df_{\text{SSR}} = 5$ (because $n = 24$ and $p = 5$).

So, the overall F test is given by

$$F^* = \frac{\text{SSR}/df_{\text{SSR}}}{\text{SSE}/df_{\text{SSE}}} = \frac{\text{MSR}}{\text{MSE}} = \frac{288/5}{72/18} = \frac{57.6}{4} = 14.4$$

and we know that $F^* \sim F_{5,18}$, so the p-value is given by $P(F_{5,18} > F^*) \approx 0$; in this case, we reject the null hypothesis $H_0 : b_1 = b_2 = b_3 = b_4 = b_5 = 0$ at $\alpha = 0.05$.

(b) To test the second order terms, the full model is the model we used in part (a), i.e., $\text{SSE}_F = 72$ and $df_{\text{SSE}_F} = 18$. The reduced model only contains the additive effects of x_1 and x_2. From the given information, the reduced model has $\text{SSE}_R = 126$ and $df_{\text{SSE}_R} = 21$ (because $n = 24$ and $p = 2$ now).

So, the necessary F test is given by

$$F^* = \frac{(\text{SSE}_R - \text{SSE}_F)/(df_{\text{SSE}_R} - df_{\text{SSE}_F})}{\text{SSE}_F/df_{\text{SSE}_F}} = \frac{(126-72)/(21-18)}{72/18} = 4.5$$

and we know that $F^* \sim F_{3,18}$, so the p-value is given by $P(F_{3,18} > F^*) = 0.01593309$; in this case, we reject the null hypothesis $H_0 : b_3 = b_4 = b_5 = 0$ at $\alpha = 0.05$.

Exercise 4

(a) The R code to read-in the data and fit the interaction model is given below:

```
> ceo=read.csv("/Users/Nate/Desktop/ceo.txt",header=TRUE)
> pmod=lm(profit~income*stock,data=ceo)
> summary(pmod)

Call:
lm(formula = profit ~ income * stock, data = ceo)

Residuals:
    Min      1Q  Median      3Q     Max
```

```
-3674.4  -621.1  -476.8   175.8  3938.4

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)  1160.50587  983.14706   1.180   0.2717
income          0.12176    0.04234   2.876   0.0206 *
stock           6.02726   61.19247   0.098   0.9240
income:stock   -0.03528    0.01168  -3.021   0.0165 *
---
Signif. codes:  0 *** 0.001 ** 0.01 * 0.05 . 0.1   1

Residual standard error: 2311 on 8 degrees of freedom
Multiple R-squared:  0.5704,        Adjusted R-squared:  0.4093
F-statistic: 3.541 on 3 and 8 DF,  p-value: 0.0678
```

The overall F test is significant at $\alpha = 0.01$: $F^* = 3.54 \sim F_{3,8}$ ($p = 0.0678$), so the overall model is statistically useful for predicting company profit at $\alpha = 0.10$.

(b) Yes, there is evidence that CEO income (x_1) and stock percentage (x_2) interact. From the above coefficient summary table, we see that we reject $H_0 : b_3 = 0$ at $\alpha = 0.05$; the test statistic is $t^* = -3.021 \sim t_8$ (and the p-value is $p = 0.0165$).

(c) Note that the term $b_1 + b_3x_2$ represents the change in $E(Y)$ for every 1-unit increase in x_1, while holding x_2 fixed. So, given a stock percentage of $x_2 = 2$, we would expect the change in profit to be

$$\hat{b}_1 + \hat{b}_3(2) = 0.12176 + (-0.03528)(2) = 0.0512 \text{ million dollars}$$

for every one thousand dollar increase in a CEO's income; note that 0.0512 million dollars is \$51,200.

Exercise 5

Suppose the interaction model

$$y_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i1}x_{i2} + e_i$$

was fit to n = 20 data points, and the following results were obtained:

```
> sum( lm( y ~ 1 )$residuals^2 )
[1] 57
> sum( lm( y ~ x1 )$residuals^2 )
[1] 40
```

```
> sum( lm( y ~ x2 )$residuals^2 )
[1] 45
> sum( lm( y ~ x1 + x2 )$residuals^2 )
[1] 36
> sum( lm( y ~ x1 + x2 + I(x1*x2) )$residuals^2 )
[1] 30
> lm( y ~ x1 + x2 + I(x1*x2) )$coefficients
(Intercept)      x1      x2     I(x1 * x2)
         10       5      -2              3
```

(a) From the given information, we know that SSE $= 30$ and SST $= 57$; this implies that SSR $=$ SST $-$ SSE $= 57 - 30 = 27$. Next note that $df_{\text{SST}} = 19$ (because $n = 20$), and $df_{\text{SSE}} = 16$ and $df_{\text{SSR}} = 3$ (because $n = 20$ and $p = 3$).

So, the overall F test is given by

$$F^* = \frac{\text{SSR}/df_{\text{SSR}}}{\text{SSE}/df_{\text{SSE}}} = \frac{\text{MSR}}{\text{MSE}} = \frac{27/3}{30/16} = \frac{9}{1.875} = 4.8$$

and we know that $F^* \sim F_{3,16}$, so the p-value is given by $P(F_{3,16} > F^*) = 0.01432341$; in this case, we reject the null hypothesis $H_0 : b_1 = b_2 = b_3 = 0$ at $\alpha = 0.05$.

(b) To test the interaction, the full model is the model we used in part (a), i.e., $\text{SSE}_F = 30$ and $df_{\text{SSE}_F} = 16$. From the given information, the reduced model has $\text{SSE}_R = 36$ and $df_{\text{SSE}_R} = 17$ (because $n = 20$ and $p = 2$ now).

So, the necessary F test is given by

$$F^* = \frac{(\text{SSE}_R - \text{SSE}_F)/(df_{\text{SSE}_R} - df_{\text{SSE}_F})}{\text{SSE}_F/df_{\text{SSE}_F}} = \frac{(36-30)/(17-16)}{30/16} = 3.2$$

and we know that $F^* \sim F_{1,16}$, so the p-value is given by $P(F_{1,16} > F^*) = 0.09258566$; in this case, we retain the null hypothesis $H_0 : b_3 = 0$ at $\alpha = 0.05$.

(c) To test the influence of x_2, the full model is the model we used in part (a), i.e., $\text{SSE}_F = 30$ and $df_{\text{SSE}_F} = 16$. The reduced model is the model containing only x_1, i.e., we are testing $H_0 : b_2 = b_3 = 0$ versus $H_1 :$ at least one $b_k \neq 0$ for $k \in \{2, 3\}$. From the given information, the reduced model has $\text{SSE}_R = 40$ and $df_{\text{SSE}_R} = 18$ (because $n = 20$ and $p = 1$ now).

So, the necessary F test is given by

$$F^* = \frac{(\text{SSE}_R - \text{SSE}_F)/(df_{\text{SSE}_R} - df_{\text{SSE}_F})}{\text{SSE}_F/df_{\text{SSE}_F}} = \frac{(40-30)/(18-16)}{30/16} = 2.666667$$

and we know that $F^* \sim F_{2,16}$, so the p-value is given by $P(F_{2,16} > F^*) = 0.1001129$; in this case, we retain the null hypothesis $H_0 : b_2 = b_3 = 0$ at $\alpha = 0.05$.

(d) Note that the term $b_1 + b_3 x_2$ represents the change in $E(Y)$ for every 1-unit increase in x_1, while holding x_2 fixed. So, given a fixed value of $x_2 = 2$, we would expect the change in the response to be

$$\hat{b}_1 + \hat{b}_3(2) = 5 + 3(2) = 11$$

for every 1-unit increase in x_1.

(e) Note that the term $b_2 + b_3 x_1$ represents the change in $E(Y)$ for every 1-unit increase in x_2, while holding x_1 fixed. So, given a fixed value of $x_1 = 3$, we would expect the change in the response to be

$$\hat{b}_2 + \hat{b}_3(3) = -2 + 3(3) = 7$$

for every 1-unit increase in x_2.