STAT 3008: Applied Linear Regression
2019-20 Term 2
Assignment #3 Solutions

Problem 1: (a) $\hat{\boldsymbol{\beta}} = (\mathbf{X'X})^{-1}\mathbf{X'Y} = \begin{pmatrix} 62.74663 \\ 10.54172 \\ -1.394128 \end{pmatrix}$

(b) $\text{RSS} = \mathbf{Y'Y} - \mathbf{Y'X}(\mathbf{X'X})^{-1}\mathbf{X'Y}$ =4107.409, $\hat{\sigma}^2 = \text{RSS}/(n-3) = 195.5909, \hat{\sigma} = 13.9854$ (12.65 ok)

(c) Optimal x = $-\hat{\beta}_1/(2\hat{\beta}_2) = 3.7808$.

(d) From part (b), SS_{res} = 4107.409. SS_{total} = $\mathbf{Y'Y} - n\bar{y}^2$ =89882.2642-24(56.6275)^2 = 12922.09.

The ANOVA Table is given by:

	df	SS	MS	F-stat	p-value
Regression	2	8814.68	4407.34	22.533	5.94E-06
Residuals	21	4107.41	195.591		
Total	23	12922.09			

(e) *(Section 4.3: x-values should be scattered like normal distribution in other to obtain a balance between goodness-of-fit and locating the center).*

Compared with linear regression, quadratic regression should rely on more data points on the two sides to provide better information about the curvature. The problem setup, however, have only one data point in the middle – which is difficult to locate the optimal value of *x* easily.

Suggestion: Allocate 1/4 to 1/3 of the data points in the middle of the *x* range [1,10], and the rest are evenly spread on the two sides.

Problem 2: (a) $\hat{Y} = 15952.1 + 244.5s + 409.9x + 4383.11U_2 + 8975.97U_3 - 1059.19U_2 s + 1582.95U_3 s$

(a) $\hat{\sigma}^2 \approx 2432^2$ = 5,914,624

```
fit0<-lm(Salary~ Sex +Year+ factor(Rank) + Sex:factor(Rank),data=salary); summary(fit0)
                     Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)          15952.10     855.91    18.638   < 2e-16 ***
Sex                    244.50    1159.16     0.211 0.833894
Year                   409.90      78.21     5.241 4.10e-06 ***
factor(Rank)2         4383.11    1063.99     4.119 0.000161 ***
factor(Rank)3         8975.97    1133.16     7.921 4.49e-10 ***
Sex:factor(Rank)2    -1059.19    2188.78    -0.484 0.630791
Sex:factor(Rank)3     1582.95    1836.99     0.862 0.393417
Residual standard error: 2432 on 45 degrees of freedom
Multiple R-squared:   0.8509,    Adjusted R-squared:   0.831
F-statistic:   42.8 on 6 and 45 DF,   p-value: < 2.2e-16
sum(fit0$res^2)
[1] 266244659
```

(b) Put U_2 = U_3 = x =0 and s=1 => $\hat{Y} = 15952.1 + 244.5(1) = \$16{,}196.6$

(c) RSS = 266,244,659 (or $2432^2(45)$=266,158,080)

(d)

	df	SS	MS	F-stat	p-value
Regression	4	642,448,811	160,612,203	27.146	1.708E-11
Residuals	45	266,244,659	5,916,548		
Total	49	908,693,470			

fit1<-lm(Salary~Sex+Year,data=salary)
anova(fit1,fit0)

```
Model 1: Salary ~ Sex + Year
Model 2: Salary ~ Sex + Year + factor(Rank) + Sex:factor(Rank)
  Res.Df       RSS    Df  Sum of Sq     F       Pr(>F)
1     49  908693470
2     45  266244659   4  642448811 27.146 1.708e-11 ***
```

(e) Since p-value = $1.708\times10^{-11} < \alpha = 0.05$, we reject H_o at α=0.05
We have sufficient evidence that rank is an important term to explain the salary.

(f) $E(Y \mid R=j, X=x) = \eta_0 + \beta x + \sum_{j=2}^{3} \eta_{0j} U_j$

(g) $E(Y \mid S=s, R=j, X=x) = \eta_0 + \eta_1 s + \beta x + \sum_{j=2}^{3} \left(\eta_{0j} U_j + \eta_{1j} U_j s \right)$

(h)

	df	SS	MS	F-stat	p-value
Regression	3	10,748,075	3,582,692	0.6055	0.6148
Residuals	45	266,244,659	5,916,548		
Total	48	276,992,734			

fit2<-lm(Salary~ Year+ factor(Rank),data=salary); summary(fit2)
anova(fit2,fit0)

```
Analysis of Variance Table
Model 1: Salary ~ Year + factor(Rank)
Model 2: Salary ~ Sex + Year + factor(Rank) + Sex:factor(Rank)
  Res.Df  RSS         Df  Sum of Sq   F      Pr(>F)
1     48  276992734
2     45  266244659   3   10748075 0.6055 0.6148
```

(i) Since p-value =0.6148 > 0.05 , we do not reject H_o at α=0.05.
We do have sufficient evidence that the salary for male and female are different for some of the 3 ranks.
(We do not have sufficient evidence that sex is important to explain the annual salary)

Problem 3:

(a) Based on Step i = 1,2,3 and 4 below, the terms added from the intercept model are in the sequence of x_1, x_2, x_4 and x_3. Therefore, parsimonious model is

Model 16: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + e$

Step		Start				
i=1	Models	**y~1**	**y~x1**	y~x2	y~x3	y~x4
	AIC	**-68.1**	**-151**	-121.8	-148.8	-66.7
i=2	Models	**y~x1**	**y~x1+x2**	y~x1+x3	y~x1+x4	
	AIC	**-151**	**-609.1**	-149.3	-150.9	
i=3	Models	**y~x1+x2**	y~x1+x2+x3	**y~x1+x2+x4**		
	AIC	**-609.1**	-608.2	**-7317.1**		
i=4	Models	**y~x1+x2+x4**	**y~x1+x2+x4+x3**			
	AIC	**-7317.1**	**-7317.6**			

(b) Based on Step i =1 and 2 below, the only term being removed from the full model is x_3. Therefore, parsimonious model is Model 12: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_4 x_4 + e$

Step		Start				
i=1	Models	**y~x1+x2+x3+x4**	y~x1+x2+x3	**y~x1+x2+x4**	y~x1+x3+x4	y~x2+x3+x4
	BIC	**-7304.6**	-597.8	**-7306.7**	-138.6	-449.1
i=2	Models	**y~x1+x2+x4**	y~x1+x2	y~x1+x4	y~x2+x4	
	BIC	**-7306.7**	-601.3	-143.1	-112.5	

(c) Let p_c to be the # of parameters in Model 16.

Based on the AIC and BIC of Model 16, p_c(log(n)-2) = -7304.6-(-7317.6) = 13.0

Based on the AIC and BIC of Model 12, (p_c-1)(log(n)-2) = -7306.7-(-7317.1) = 10.4

Hence log(n)-2 = 13.0-10.4 =2.6 => n = exp(4.6) = 99.484 => n = 99 or 100

(d) Yes, x_3 is likely to be highly collinear with x_1 because the AIC for (i) Model #2 and #6 are similar, but (ii) AIC for Model #5 is comparable to the AIC for Model #2.

Problem 4: (a) Based on the R outputs, the parsimonious models for both forward and backward selection methods are y ~ x4 + x3 + x6

```
> AIC.F<-stepAIC(fit0,scope=list(lower=fit0, upper=fit1),direction="forward",trace=1) # forward selection
Start:  AIC=253.58
y ~ 1
       Df Sum of Sq     RSS    AIC
+ x4    1   1661.66  884.73 181.57
+ x2    1   1120.45 1425.95 214.99
+ x1    1    504.47 2041.92 240.12
+ x3    1    462.22 2084.18 241.55
+ x6    1    360.22 2186.17 244.90
+ x5    1    281.47 2264.92 247.38
<none>              2546.39 253.58

Step:  AIC=181.58
y ~ x4
       Df Sum of Sq    RSS    AIC
+ x3    1   141.523 743.21 171.37
+ x5    1    90.016 794.72 176.06
+ x6    1    49.592 835.14 179.54
+ x2    1    25.046 859.69 181.56
<none>              884.73 181.57
+ x1    1     8.291 876.44 182.92

Step:  AIC=171.37
y ~ x4 + x3
       Df Sum of Sq    RSS    AIC
+ x6    1    45.431 697.78 168.96
+ x2    1    21.519 721.69 171.32
<none>              743.21 171.37
+ x1    1     8.535 734.68 172.56
+ x5    1     1.637 741.57 173.22
Step:  AIC=168.96
y ~ x4 + x3 + x6
       Df Sum of Sq    RSS    AIC
<none>              697.78 168.96
+ x1    1   15.1423 682.64 169.42
+ x2    1   13.2307 684.55 169.62
+ x5    1    4.2635 693.52 170.53
> AIC.B<-stepAIC(fit1,scope=list(lower=fit0, upper=fit1),direction="backward",trace=1) # backward selection
Start:  AIC=172.67
y ~ x1 + x2 + x3 + x4 + x5 + x6
       Df Sum of Sq     RSS    AIC
- x2    1      2.93  678.29 170.98
- x5    1      2.95  678.31 170.98
- x1    1      7.01  682.38 171.40
<none>               675.36 172.67
- x6    1     44.82  720.18 175.17
- x3    1     59.18  734.54 176.55
- x4    1    646.12 1321.48 217.66
Step:  AIC=170.98
y ~ x1 + x3 + x4 + x5 + x6
       Df Sum of Sq     RSS    AIC
- x5    1      4.34  682.64 169.42
- x1    1     15.22  693.52 170.53
<none>               678.29 170.98
- x6    1     54.83  733.12 174.42
- x3    1     74.01  752.30 176.22
- x4    1   1190.09 1868.39 239.90
Step:  AIC=169.42
y ~ x1 + x3 + x4 + x6
       Df Sum of Sq     RSS    AIC
- x1    1     15.14  697.78 168.96
<none>               682.64 169.42
- x6    1     52.04  734.68 172.56
- x3    1    148.70  831.34 181.22
- x4    1   1222.35 1904.99 239.26
Step:  AIC=168.96
y ~ x3 + x4 + x6
       Df Sum of Sq     RSS    AIC
<none>               697.78 168.96
- x6    1     45.43  743.21 171.37
- x3    1    137.36  835.14 179.54
- x4    1   1278.97 1976.75 239.85
```

(b) Based on the R outputs, the parsimonious models for both forward and backward selection methods are y ~ x4 + x3 + x6, which is the same as that for part (a).

```
> BIC.F<-stepAIC(fit0,scope=list(lower=fit0, upper=fit1),direction="forward",trace=1,k=log(n)) # forward selection
Start:  AIC=254.06
y ~ 1
       Df Sum of Sq     RSS    AIC
+ x4    1   1661.66  884.73 182.54
+ x2    1   1120.45 1425.95 215.96
+ x1    1    504.47 2041.92 241.09
+ x3    1    462.22 2084.18 242.52
+ x6    1    360.22 2186.17 245.87
+ x5    1    281.47 2264.92 248.35
<none>              2546.39 254.06

Step:  AIC=182.55
y ~ x4
       Df Sum of Sq    RSS    AIC
+ x3    1   141.523 743.21 172.83
+ x5    1    90.016 794.72 177.52
+ x6    1    49.592 835.14 180.99
<none>              884.73 182.54
+ x2    1    25.046 859.69 183.02
+ x1    1     8.291 876.44 184.37
Step:  AIC=172.83
y ~ x4 + x3
       Df Sum of Sq    RSS    AIC
+ x6    1    45.431 697.78 170.90
<none>              743.21 172.83
+ x2    1    21.519 721.69 173.26
+ x1    1     8.535 734.68 174.50
+ x5    1     1.637 741.57 175.16
Step:  AIC=170.9
y ~ x4 + x3 + x6
       Df Sum of Sq    RSS    AIC
<none>              697.78 170.90
+ x1    1   15.1423 682.64 171.85
+ x2    1   13.2307 684.55 172.04
+ x5    1    4.2635 693.52 172.95
```

> BIC.B<-stepAIC(fit1,scope=list(lower=fit0, upper=fit1),direction="backward",trace=1,k=log(n)) # backward selection

```
Start:  AIC=176.07
y ~ x1 + x2 + x3 + x4 + x5 + x6
       Df Sum of Sq     RSS    AIC
- x2    1      2.93  678.29 173.88
- x5    1      2.95  678.31 173.89
- x1    1      7.01  682.38 174.31
<none>               675.36 176.07
- x6    1     44.82  720.18 178.08
- x3    1     59.18  734.54 179.46
- x4    1    646.12 1321.48 220.57
Step:  AIC=173.89
y ~ x1 + x3 + x4 + x5 + x6
       Df Sum of Sq     RSS    AIC
- x5    1      4.34  682.64 171.85
- x1    1     15.22  693.52 172.95
<none>               678.29 173.88
- x6    1     54.83  733.12 176.84
- x3    1     74.01  752.30 178.65
- x4    1   1190.09 1868.39 242.33
Step:  AIC=171.85
y ~ x1 + x3 + x4 + x6

       Df Sum of Sq     RSS    AIC
- x1    1     15.14  697.78 170.90
<none>               682.64 171.85
- x6    1     52.04  734.68 174.50
- x3    1    148.70  831.34 183.16
- x4    1   1222.35 1904.99 241.20
Step:  AIC=170.9
y ~ x3 + x4 + x6
       Df Sum of Sq     RSS    AIC
<none>               697.78 170.90
- x6    1     45.43  743.21 172.83
- x3    1    137.36  835.14 180.99
- x4    1   1278.97 1976.75 241.31
```

(c) From the R-output, $VIR_5 = 1/(1-R^2_{-5}) = 1/(1-0.8343) = 6.0346$

> fitx5<-lm(x5~x1+x2+x3+x4+x6);   1/(1-summary(fitx5)$r.squared)

[1] 6.034562