Assignment 5 Chunlei Zhou

Q1

(a)

Continuity of g(x) and continuity of the first derivative each represent a single linear constraint at each knot, so there are 6 linear constraints. These are

$$b_1 + 2c_1\eta_1 + 3d_1{\eta_1}^2 = 0$$

$$b_1 + 2c_1\eta_2 + 3d_1{\eta_2}^2 = b_2 + 2c_2\eta_2 + 3d_2{\eta_2}^2$$

$$b_2 + 2c_2\eta_3 + 3d_2{\eta_3}^2 = 0$$

for continuity of the first derivative, and

$$a_1 + b_1\eta_1 + c_1{\eta_1}^2 + d_1{\eta_1}^3 = a_0$$

$$a_1 + b_1\eta_2 + c_1{\eta_2}^2 + d_1{\eta_2}^3 = a_2 + b_2\eta_2 + c_2{\eta_2}^2 + d_2{\eta_2}^3$$

$$a_2 + b_2\eta_3 + c_2{\eta_3}^2 + d_2{\eta_3}^3 = a_3$$

for continuity.

(b)

Since the total order of all polynomials is 10, and the total constraints we have is 6, we can denote that the degrees of freedom equal to 10 – 6 = 4.

Q2

(a)

```
> table_a
             SSE       AIC       BIC DF
Model1 2.055594 -350.5192 -345.4540  2
Model2 1.671742 -367.7422 -360.1444  3
Model3 1.778792 -363.9699 -358.9047  2
Model4 1.722704 -366.9495 -361.8843  2
Model5 1.873968 -357.1224 -349.5246  3
Model6 1.638056 -367.6353 -357.5049  4
```

(b)

```
> table_b
             SSE        AIC       BIC DF
Model1 2.055594  -86.59667 -81.53147  2
Model2 1.671742 -103.81961 -96.22181  3
Model3 1.778792 -100.04730 -94.98210  2
Model4 1.722704 -103.02694 -97.96174  2
Model5 1.873968  -93.19980 -85.60200  3
Model6 1.638056 -103.71272 -93.58232  4
```

(c)

```
> table_c
             SSE        AIC       BIC DF
Model1 2.055594  -84.59667 -76.99887  2
Model2 1.671742 -101.81961 -91.68921  3
Model3 1.778792  -98.04730 -90.44950  2
Model4 1.722704 -101.02694 -93.42914  2
Model5 1.873968  -91.19980 -81.06940  3
Model6 1.638056 -101.71272 -89.04973  4
```

(d)

Standardized table for part(a):

```
> s_table_a
             SSE        AIC       BIC DF
Model1 2.055594 17.2229379 16.430268  2
Model2 1.671742  0.0000000  1.739929  3
Model3 1.778792  3.7723093  2.979639  2
Model4 1.722704  0.7926700  0.000000  2
Model5 1.873968 10.6198113 12.359741  3
Model6 1.638056  0.1068864  4.379415  4
```

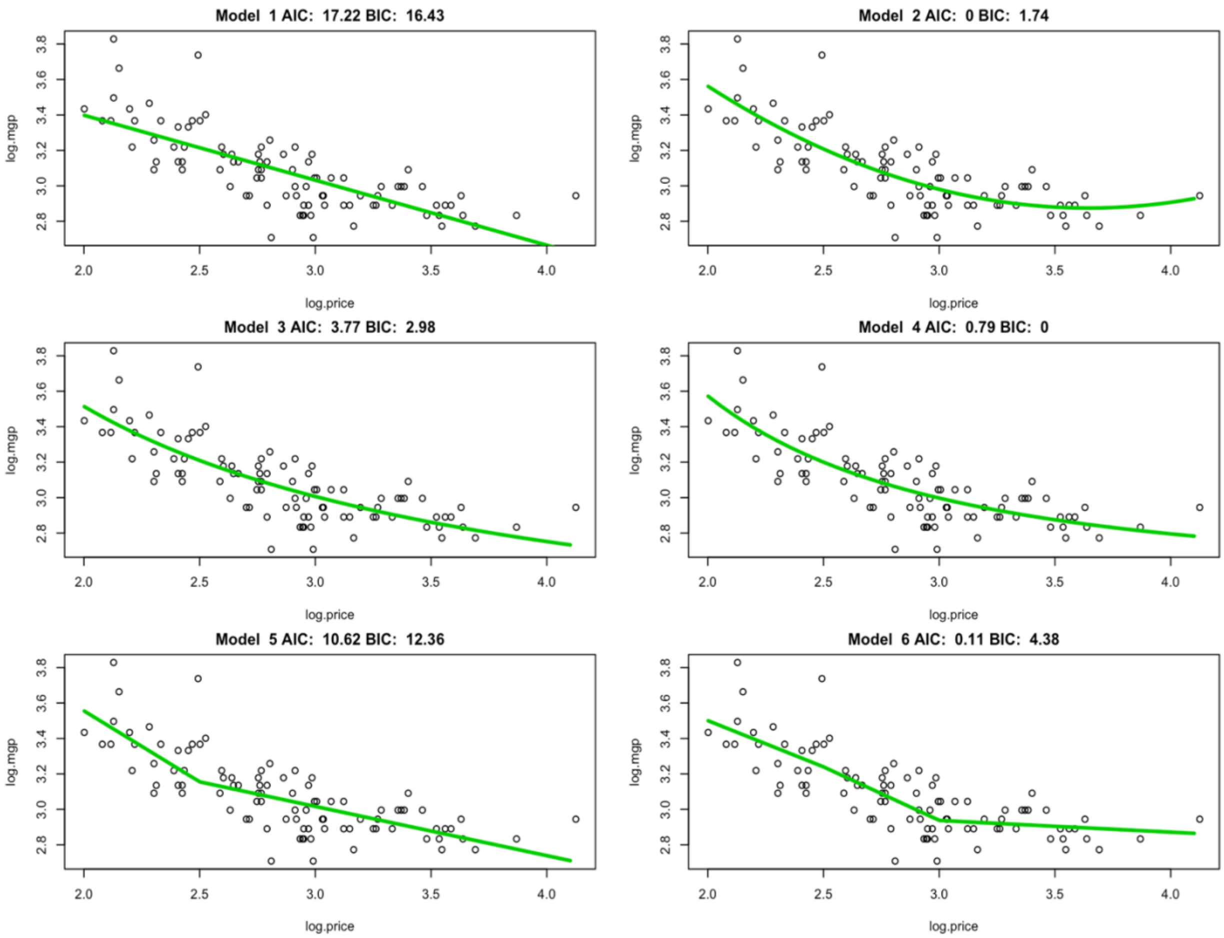
Standardized table for part(b):

```
> s_table_b
             SSE        AIC       BIC DF
Model1 2.055594 17.2229379 16.430268  2
Model2 1.671742  0.0000000  1.739929  3
Model3 1.778792  3.7723093  2.979639  2
Model4 1.722704  0.7926700  0.000000  2
Model5 1.873968 10.6198113 12.359741  3
Model6 1.638056  0.1068864  4.379415  4
```

Standardized table for part(c):

```
> s_table_c
             SSE        AIC       BIC DF
Model1 2.055594 17.2229379 16.430268  2
Model2 1.671742  0.0000000  1.739929  3
Model3 1.778792  3.7723093  2.979639  2
Model4 1.722704  0.7926700  0.000000  2
Model5 1.873968 10.6198113 12.359741  3
Model6 1.638056  0.1068864  4.379415  4
```

(e)



(f)

If the model selection application is based on AIC scores, then Model 2 has the optimal AIC and BIC scores (AIC = 0, which is the smallest among all six models). However, Model 2 is overfitting, because it fits the outlier. While Model 6 and Model 4, with the second and third smallest AIC scores respectively, are not overfitting, there is a distinct reason that models with the second or third lowest AIC might be used in place of the optimal AIC model.

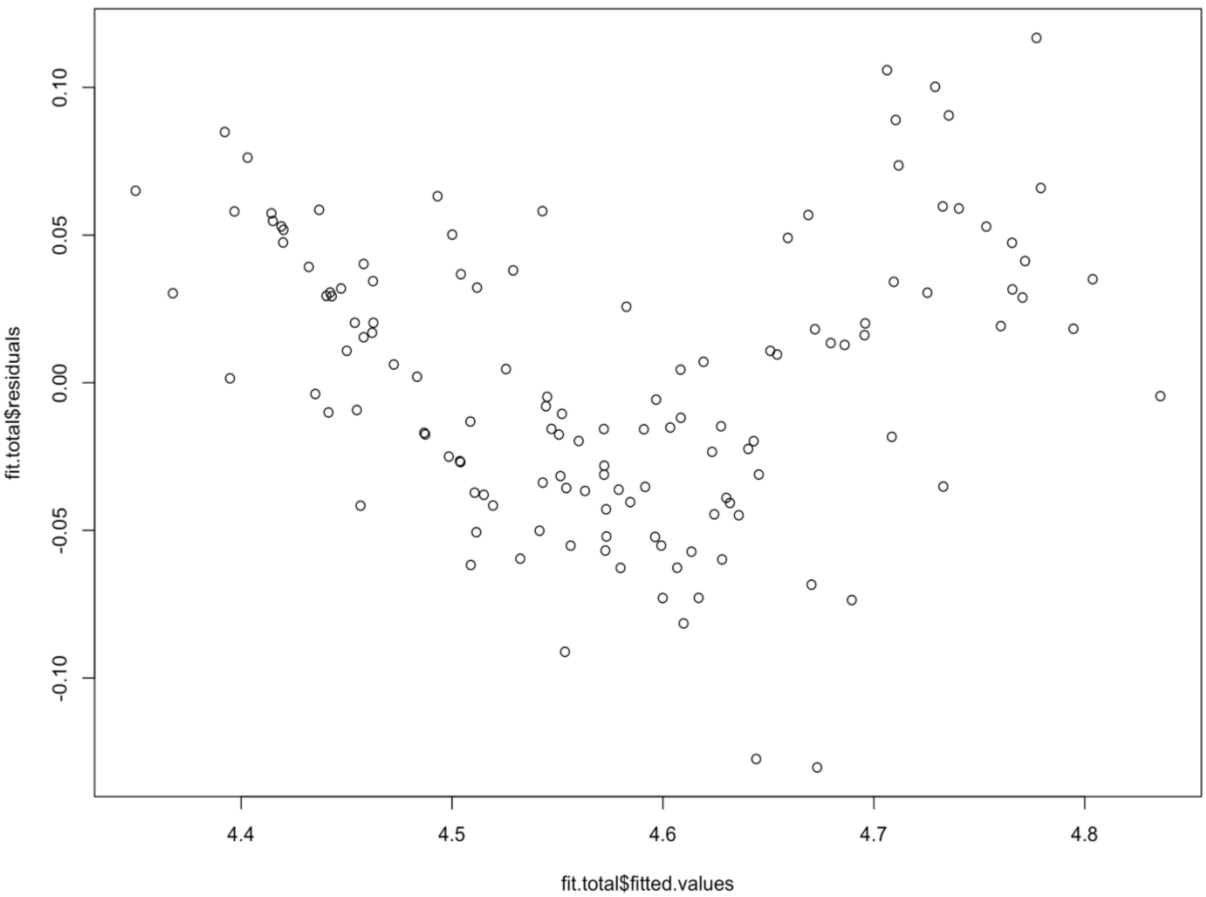
Q3

(a)

R^2 for each predictor:

[1] "R-square for predicor = rank is 0.906048203507413"
[1] "R-square for predicor = LSAT is 0.617296896570578"
[1] "R-square for predicor = GPA is 0.586749347746426"
[1] "R-square for predicor = faculty is 0.128602021730388"
[1] "R-square for predicor = clsize is 0.0629362367654693"
[1] "R-square for predicor = studfac is 0.0377271898430339"
[1] "R-square for predicor = llibvol is 0.520850141125811"
[1] "R-square for predicor = lcost is 0.253040457876905"

Predictor rank appears to be most informative of salary.

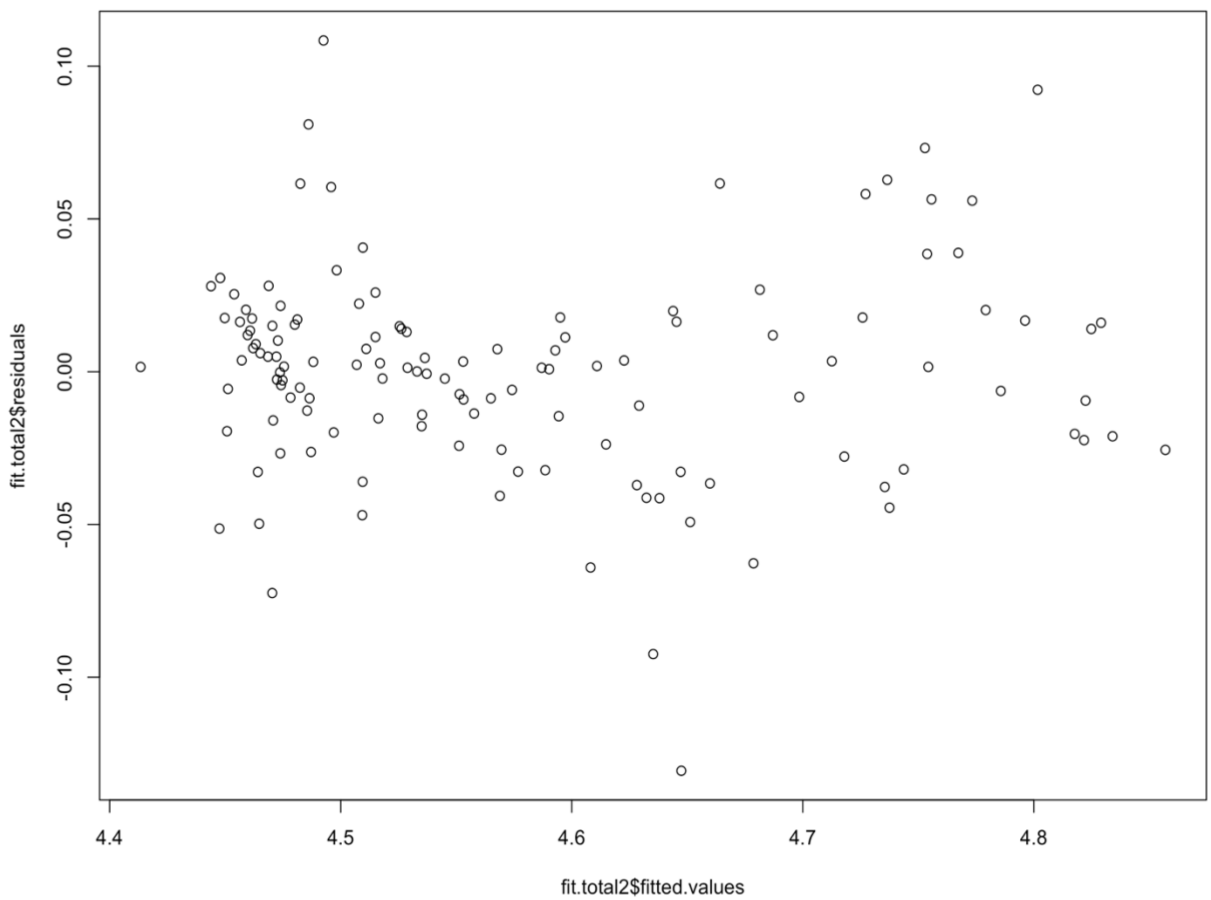


I noticed that this is a megaphone shaped scatter plot, which indicates that the model might not be suitable.

(b)

```
> apply(xs, 2, var)
       rank       rank2        LSAT       LSAT2         GPA        GPA2 
0.007692308 0.007692308 0.007692308 0.007692308 0.007692308 0.007692308 
    faculty    faculty2      clsize     clsize2     studfac    studfac2 
0.007692308 0.007692308 0.007692308 0.007692308 0.007692308 0.007692308 
    llibvol    llibvol2       lcost      lcosr2 
0.007692308 0.007692308 0.007692308 0.007692308 
```

It is obvious that all column variances are equal.



It seems that variables are much more random compared to those in part (a).

(c)

```
> coef.1se.lasso[,1][which(coef.1se.lasso[,1] !=0)]
(Intercept)        rank       rank2        LSAT         GPA     llibvol 
 4.57751884 -0.98047207  0.27602144  0.10377027  0.02742180  0.01957966 
```

The above listed variables are included in the 1se solution.

(d)

Ranking is from high to low with 1 representing the largest value.

```
> abs.coef.1se.ridge = abs(coef.1se.ridge[,1][which(coef.1se.ridge[,1] !=0)])
> rank(-abs.coef.1se.ridge)
(Intercept)        rank       rank2        LSAT       LSAT2         GPA 
          1           2           4           5          14           3 
       GPA2     faculty    faculty2      clsize     clsize2     studfac 
          8          10           9          12          15          16 
   studfac2     llibvol    llibvol2       lcost      lcosr2 
         11           6          17           7          13 
> abs.coef.1se.lasso = abs(coef.1se.lasso[,1][which(coef.1se.lasso[,1] !=0)])
> rank(-abs.coef.1se.lasso)
(Intercept)        rank       rank2        LSAT         GPA     llibvol 
          1           2           3           4           5           6 
```

The ranks of those coefficients selected by the LASSO model of Part (c) is:

Variable name	Ranks
rank	1
rank2	3
LSAT	4
GPA	2
llibvol	5

(e)

For forward model, the included variables, their values, and their ranks are as bellow:
(Ranking is from high to low with 1 representing the largest absolute value.)

```
> coef(fit.stepaicf)[which(coef(fit.stepaicf) !=0)]
 (Intercept)         rank        rank2         LSAT        LSAT2 
 4.577518837 -1.117582256  0.404255775  0.148780040 -0.033026242 
         GPA         GPA2      faculty     faculty2       clsize 
 0.078167265  0.062324420  0.138273116  0.032697006  0.002092815 
     clsize2      studfac     studfac2      llibvol     llibvol2 
 0.029783741  0.072780553 -0.045937763 -0.045803915  0.001250647 
       lcost       lcosr2 
-0.060717584  0.003852739 
> rank(-coef.stepaicf)
(Intercept)        rank       rank2        LSAT       LSAT2         GPA 
          1           2           3           4          12           6 
       GPA2     faculty    faculty2      clsize     clsize2     studfac 
          8           5          13          16          14           7 
   studfac2     llibvol    llibvol2       lcost      lcosr2 
         10          11          17           9          15 
```

For backward model, the included variables, their values, and their ranks are as bellow: (Ranking is from high to low with 1 representing the largest absolute value.)

```
> coef(fit.stepaicb)[which(coef(fit.stepaicb) !=0)]
(Intercept)        rank       rank2        LSAT     faculty    faculty2 
 4.57751884 -1.11854248  0.42132919  0.15337353  0.06796724  0.07139841 
> rank(-coef.stepaicb)
(Intercept)        rank       rank2        LSAT     faculty    faculty2 
          1           2           3           4           6           5 
```

For all subset model, the included variables, their values, and their ranks are as bellow: (Ranking is from high to low with 1 representing the largest absolute value.)

```
> coef(allsub.bestmodelbyaic)[which(coef(allsub.bestmodelbyaic) !=0)]
(Intercept)        rank       rank2        LSAT    faculty2      clsize 
 4.57751884 -1.10979722  0.41860566  0.17069993  0.07579634  0.08026534 
> rank(-coef.allsub)
(Intercept)        rank       rank2        LSAT    faculty2      clsize 
          1           2           3           4           6           5 
```

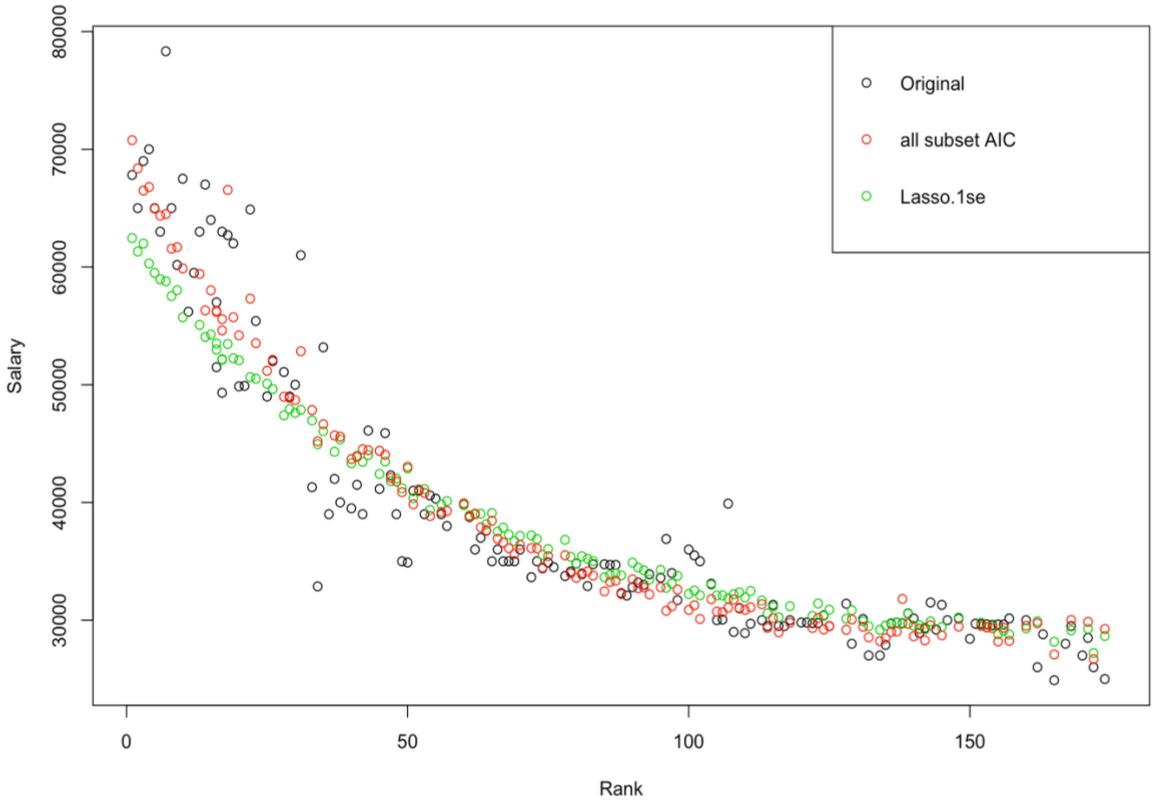
The computation time of AIC forward (st1), AIC backward (st1), all subset model (st3), and LASSO model (st4) are listed in the photo below:

```
> st1
   user  system elapsed 
  0.003   0.000   0.002 
> st2
   user  system elapsed 
  0.081   0.003   0.082 
> st3
   user  system elapsed 
 77.176   1.548  79.703 
> st4
   user  system elapsed 
  0.053   0.000   0.053 
```

Considering the elapsed time, AIC forward model is the fastest while all subset model is the slowest. The LASSO model is faster than AIC backward model.

Variables rank, rank2 and LAST are included in all four models. This makes sense considering the R^2 in part(a).

(f)



It seems that all subset model has some shrinkage effect.

Q4

The log likelihood function l(θ;ξ) after some math is:

$$l(\theta;\xi) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}{\epsilon_i}^2$$

$$= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}SSE$$

$$-2 \times l(\theta;\xi) = nlog(2\pi) + n\log(\sigma^2) + \frac{1}{\sigma^2}SSE$$

$$since\ we\ have\ \sigma_{MLE}^2 = \frac{SSE}{n}, we\ can\ denote\ that$$

$$-2 \times l(\theta;\xi) = nlog(2\pi) + n\log\left(\frac{SSE}{n}\right) + n$$

$$nlog\left(\frac{SSE}{n}\right) - \left(-2 \times l(\theta;\xi)\right)$$

$$= nlog\left(\frac{SSE}{n}\right) - nlog(2\pi) - n\log\left(\frac{SSE}{n}\right) - n$$

$$= nlog(2\pi) - n$$

It is obvious that the for a Gaussian multiple regression model the quantities $nlog\left(\frac{SSE}{n}\right)$ and $-2 \times l(\theta;\xi)$ differ by a constant $nlog(2\pi) - n$ which depends only on n. Therefore, the AIC and BIC evaluation methods of Parts (a) and (b) in Q2 are equivalent from the point of view of model selection.