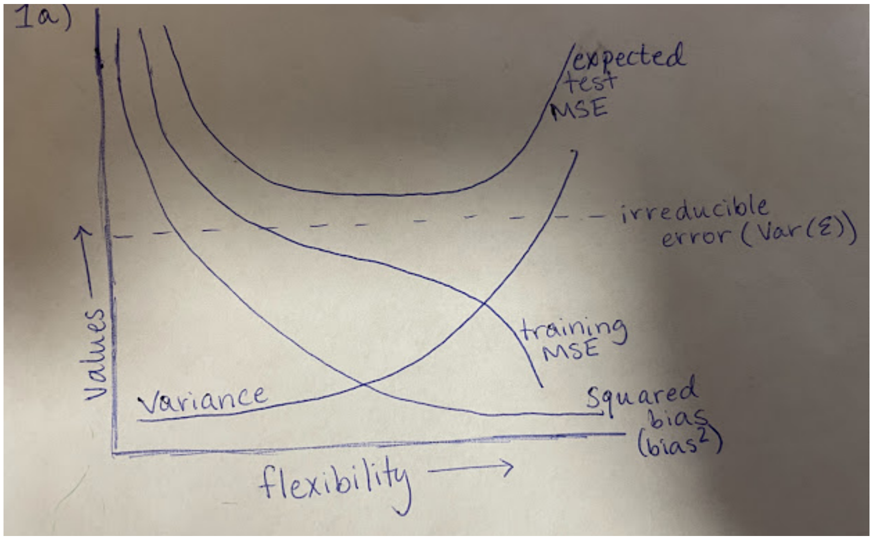
Neha Maddali HW1

Problem 1:

a. 

b. Expected test MSE is the expected average error if we were able to use the entire population of data. Training MSE is the average error able to be calculated using a sample of the entire population data which is basically an estimate of the expected test MSE. Bias is the tendency in which there are differences between the facts and results which is also just the difference between the actual value and the prediction from our model. Variance is how much the estimate may differ if different data was used to train the model. Irreducible error simply stated is that error can't be reduced by creating good models; it is caused by random things outside our control.

c. - The irreducible error is constant and it is a parallel line. This "curve" lies below the test MSE curve since the expected test MSE will always be greater than the Var(E) which is the irreducible error.
   - The training MSE decreases as flexibility increases. Because as flexibility increases, the function overfits the data and at the end, the error shows a minimum constant value.
   - The test MSE decreases until it comes to a point where flexibility starts to overfit the training data which is when the error starts to increase.
   - The squared bias decreases as the flexibility increases. Bias basically refers to the error that is introduced by approximating a real-life problem. For example, it's unlikely a real-life problem would have a simple linear relationship. So using a simple model like the linear regression line will result in some bias in the estimate of the function.
   - The variance increases as the flexibility increases. The variance is the amount of which the function hat would change if it was estimated using a different training data set. So if you were to change any point, it might cause the function hat to change which will lead to some variance.

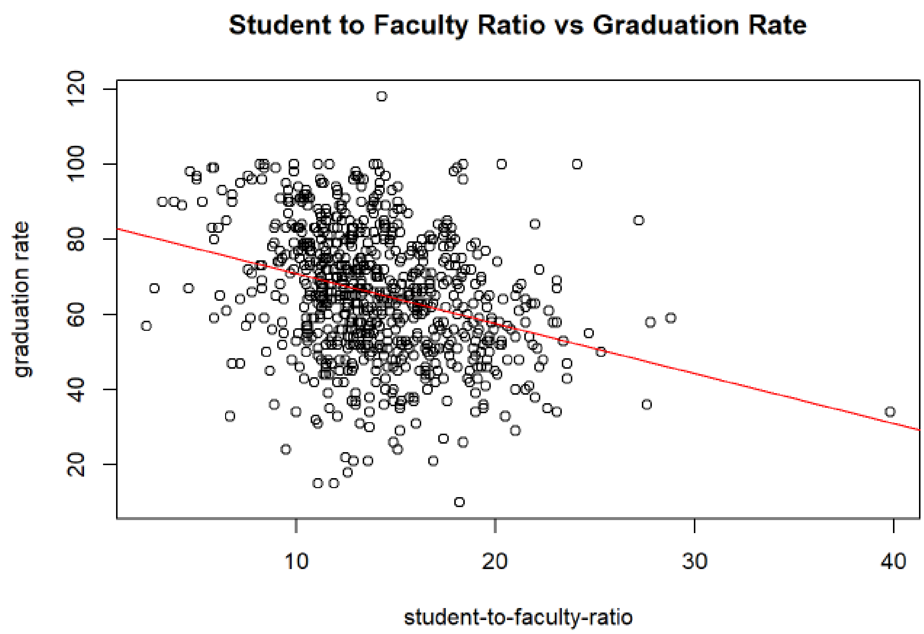
d. Some of the advantages of a flexible approach is that it might have a better fit for non-linear models and decrease the bias. Some of the disadvantages of a flexible approach is that it could possibly overfit the training data, have a higher variance and make it harder to interpret.A more flexible approach would probably be preferred when we are interested in prediction and not interpretability of results. Possibly also when the system is under fitted. A less flexible approach would probably be preferred when we are interested in the interpretability of the results.

Problem 2:

a. 777 observations, 18 variables
b. Statistics for a large number of US Colleges from the 1995 issue of US News and World Report.
c. row 278 shows Iowa State University

```
                      Private Apps Accept Enroll Top10perc Top25perc F.Undergrad P.Undergrad
Iowa State University      No 8427   7424   3441        26        59       18676        1715
                      Outstate Room.Board Books Personal PhD Terminal S.F.Ratio perc.alumni Expend
Iowa State University     7550       3224   640     2055  81       88      19.2          22   8420
                      Grad.Rate
Iowa State University        65
```

d. Average graduation rate across all colleges: 65.46332
Iowa State University's graduation rate is 65% which is just below the average
e. $\hat{B}_0$ = 84.2168, standard error = 2.1713, p value = <2e-16
$\hat{B}_1$ = -1.3310, standard error = 0.1484, p value = <2e-16
So graduation rate is calculated by = 84.2168 + (-1.3310)(S.F.Ratio)
Given these statistics, $\hat{B}_1$ = -1.3310 is the average change in graduation rate as there is a one unit change in the student to faculty ratio
f.



g. Residuals:

```
Abilene Christian University            Adelphi University              Adrian College
                  -0.1255965                   -11.9785259                 -13.0468224
         Agnes Scott College      Alaska Pacific University
                 -14.9680482                   -53.3778274
```

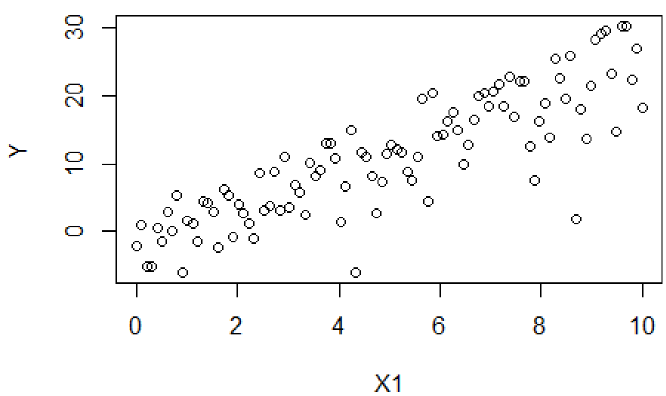
Fitted values:

```
Abilene Christian University            Adelphi University              Adrian College
                    60.12560                      67.97853                    67.04682
         Agnes Scott College      Alaska Pacific University
                    73.96805                      68.37783
```

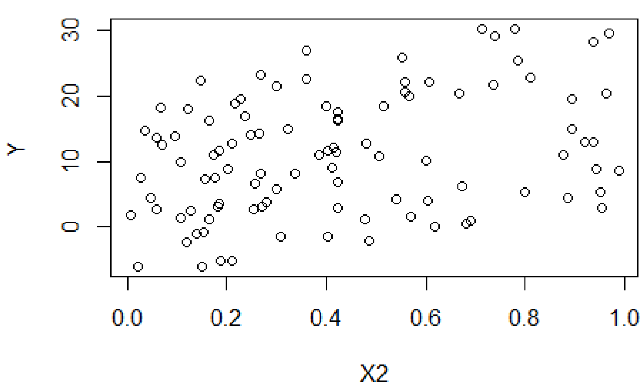
h. predicted graduation rate when the student-to-faculty ratio is 10 = 70.90674
i. A test MSE of the model is required to be calculated to understand the prediction accuracy of the model on data that was not included in the model developed. Like stated before, the test MSE is what we could expect the average error to be on the new data the model has not yet seen. To do this, split the data into a training and testing set. Train the model using the training data and then use this model on the test data. Then the MSE for the test data can be calculated.
    - Expected test MSE = 295.8065

Problem 3:

a. $B_0 = 2$, $B_1 = 3$, $B_2 = 5$
b.
```
 [1] -2.1921818  1.0513192 -5.2340411 -5.2527506  0.6130698 -1.3275483  3.0231447  0.2095631
 [9]  5.3691176 -6.0070956  1.5486576  1.2272484 -1.5214411  4.5327420  4.3160611  2.9698086
[17] -2.2023579  6.2622636  5.3479473 -0.8227962  3.9624270  2.8104741  1.1405682 -0.9413192
[25]  8.5987375  3.1469534  3.8824735  8.9387477  3.2629220 10.9802664  3.6369071  6.7871781
[33]  5.6833423  2.4277324 10.1071745  8.1188537  9.0541070 13.0552466 12.9685007 10.7307054
[41]  1.3162208  6.6933637 15.0283071 -5.9812614 11.7488068 10.9663141  8.2874999  2.6263054
[49]  7.4071819 11.4446759 12.7936996 12.0439339 11.6482617  8.8523216  7.5346995 11.0337281
[57] 19.5942452  4.5032082 20.4689567 14.0500062 14.2995218 16.3340760 17.5659908 15.0263008
[65]  9.8428349 12.8395858 16.4120028 19.9774494 20.4942753 18.5145262 20.7158312 21.7848897
[73] 18.5126209 22.7701192 16.9338435 22.2707934 22.1745199 12.6690765  7.4424298 16.3522494
[81] 18.8488076 13.8966091 25.4766260 22.5636627 19.5619779 26.0131639  1.8184290 17.9869321
[89] 13.6733143 21.5684086 28.4326513 29.1750410 29.5668087 23.2434093 14.8205705 30.2984978
[97] 30.3245585 22.4497786 26.9079566 18.1962604
```
c. Scatter plot of X1 and Y



Scatter plot of X2 and Y

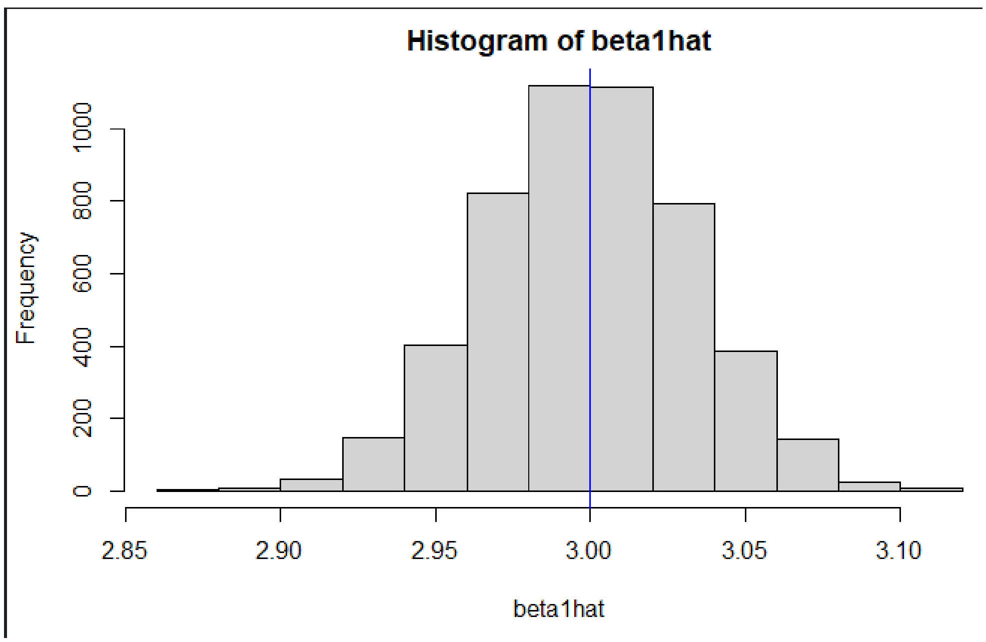


The scatter plot of X1 compared to Y shows a strong positive linear correlation. The scatter plot of X2 compared to Y shows a weak positive linear correlation d.

```
B = 5000
beta0hat = beta1hat = beta2hat = rep(NA,B)
Yhat = rep(NA,B)
for(i in 1:B){
  error = rnorm(n,0,1)
  Y = beta_0 + beta_1*X1 + beta_2*log(X2) + error
  fit = lm(Y~X1 + log(X2))
  beta1hat[i] = fit$coefficients[[2]]
}
beta1hat[i]
```

Mean = 3.008005 which means B1hat is an unbiased estimator of B1

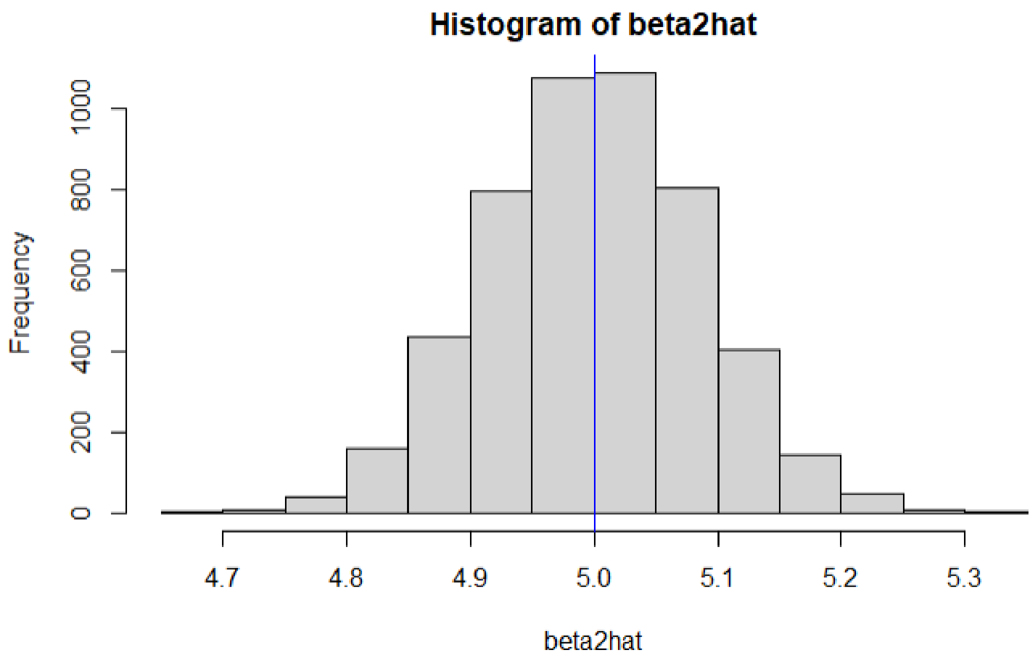
e.



f.

```
B = 5000
beta0hat = beta1hat = beta2hat = rep(NA,B)
Yhat = rep(NA,B)
for(i in 1:B){
  error = rnorm(n,0,1)
  Y = beta_0 + beta_1*X1 + beta_2*log(X2) + error
  fit = lm(Y~X1 + log(X2))
  beta2hat[i] = fit$coefficients[[3]]
}
beta2hat[i]
```

Mean = 5.009441 so B2hat is an unbiased estimator of B2 g.


Histogram of beta2hat
Frequency
0
200
400
600
800
1000
4.7
4.8
4.9
5.0
5.1
5.2
5.3
beta2hat