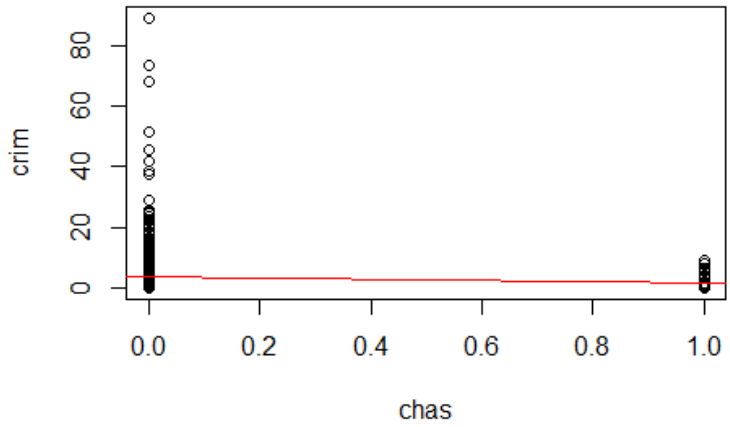
Neha Maddali

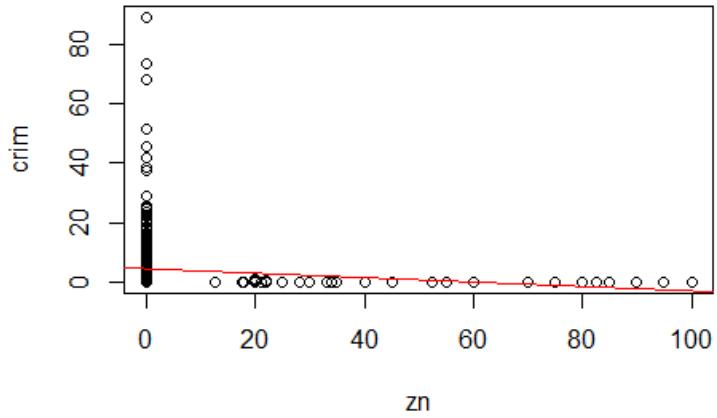
Question 1:

a. There are 506 rows and 13 variables in the data set Boston. The variable lstat represents the lower status of the population (percent).

b. $\hat{B}_0$ = -3.33054, standard error = 0.69376, p-value = 2.09e-06
$\hat{B}_1$ = 0.54880, standard error = 0.04776, p-value = <2e-16
So the crim rate is calculated by = (-3.3305) + 0.5488(lstat)

c. zn: if the proportion of residential land zoned for lots overs 25,000 sq.ft increases then the per capita crime rate decreases
$\hat{B}_0$ = 4.45369, $\hat{B}_1$ = -0.07393
indus: if the proportion of non-retail business acres per town increases, the per capita crime rate increases
$\hat{B}_0$ = -2.06374, $\hat{B}_1$ = 0.50978
chas: if the Charles River dummy variable increases, the per capita crime rate decreases
$\hat{B}_0$ = 3.7444, $\hat{B}_1$ = -1.8928
nox: if the nitrogen oxide concentration increases, the per capita crime rate increases
$\hat{B}_0$ = -13.720, $\hat{B}_1$ = 31.249
rm: if the average number of rooms per dwelling increases, the per capita crime rate decreases
$\hat{B}_0$ = 20.482, $\hat{B}_1$ = -2.684
age: if the proportion of owner-occupied units built prior to 1940 increases, the per capita crime rate decreases
$\hat{B}_0$ = -3.77791, $\hat{B}_1$ = 0.10779
dis: if the weighted mean of the distances to five Boston employment centers increases, the per capita crime rate decreases
$\hat{B}_0$ = 9.4993, $\hat{B}_1$ = -1.5509
rad: if the index of accessibility to radial highways increases, the per capita crime rate increases
$\hat{B}_0$ = -2.28716, $\hat{B}_1$ = 0.61791
tax: if the full-value property-tax rate per $10k increases, the per capita crime rate increases
$\hat{B}_0$ = -8.528369, $\hat{B}_1$ = 0.029742
ptratio: if the pupil-teacher ratio by town increases, the per capita crime rate increases
$\hat{B}_0$ = -17.6469, $\hat{B}_1$ = 1.1520
medv: if the median value of owner-occupied homes increases, the per capita crime rate decreases
$\hat{B}_0$ = 11.79654, $\hat{B}_1$ = -0.36316
We will need to test H0: B1 = 0 to tell which predictors are significant. It can be observed that all predictors have a p-value less than 0.05 except the predictor “chas.” Thus, there is a statistically significant association between each predictor and response except for the “chas” predictor. A scatter plot for the chas predictor and zn predictor are down below to show that chas does not have a statistically significant association with the response crim.





d. We can reject the null hypothesis for zn, dis, rad, and medv because they all have values less than 0.05 (α = 0.05)

```
Call:
lm(formula = crim ~ ., data = Boston)

Residuals:
   Min     1Q Median     3Q    Max
-8.534 -2.248 -0.348  1.087 73.923

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.7783938  7.0818258   1.946 0.052271 .
zn           0.0457100  0.0187903   2.433 0.015344 *
indus       -0.0583501  0.0836351  -0.698 0.485709
chas        -0.8253776  1.1833963  -0.697 0.485841
nox         -9.9575865  5.2898242  -1.882 0.060370 .
rm           0.6289107  0.6070924   1.036 0.300738
age         -0.0008483  0.0179482  -0.047 0.962323
dis         -1.0122467  0.2824676  -3.584 0.000373 ***
rad          0.6124653  0.0875358   6.997 8.59e-12 ***
tax         -0.0037756  0.0051723  -0.730 0.465757
ptratio     -0.3040728  0.1863598  -1.632 0.103393
lstat        0.1388006  0.0757213   1.833 0.067398 .
medv        -0.2200564  0.0598240  -3.678 0.000261 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.46 on 493 degrees of freedom
Multiple R-squared:  0.4493,    Adjusted R-squared:  0.4359
F-statistic: 33.52 on 12 and 493 DF,  p-value: < 2.2e-16
```

e. The table below shows that using multiple simple linear regression models is not sufficient compared to a multiple linear regression model. This is because in a simple

regression model, all other predictors are ignored while in a multiple regression model, the other predictors are fixed.

	simple	multiple
zn	-0.07393498	0.0457100386
indus	0.50977633	-0.0583501107
chas	-1.89277655	-0.8253775522
nox	31.24853120	-9.9575865471
rm	-2.68405122	0.6289106622
age	0.10778623	-0.0008482791
dis	-1.55090168	-1.0122467382
rad	0.61791093	0.6124653115
tax	0.02974225	-0.0037756465
ptratio	1.15198279	-0.3040727572
lstat	0.54880478	0.1388005968
medv	-0.36315992	-0.2200563590

f. Training MSE = 42.49345
Test MSE = 41.19923

g. Training MSE = 43.97466
Test MSE = 39.62763
We were using less predictors in this model so I expected the training MSE to be higher and the test MSE to be lower than results in part f.

h. H_0: $B_j = 0$, H_1: $B_j \neq 0$
Test statistic: 3.264
Null distribution: t-distribution with 495 degrees of freedom n-p+1 = 506-12+1=495
p-value: 0.001173
The null hypothesis is rejected and the results are statistically significant. B_j is significantly different from 0 at significance level 0.05. We have to assume H_0 is true for this to work

i. H_0: $B_j = 5$, H_1: $B_j \neq 5$
Test statistic: (2.839993 - 5) / 0.870007 = -2.4827
Null distribution: t-distribution with 495 degrees of freedom n-p+1 = 506-12+1=495
p-value: 0.01336767
The null hypothesis is rejected and the results are statistically significant. B_j is significantly different from 0 at significance level 0.05. We have to assume H_0 is true for this to work

Question 2:

a. 1. There is a linear relationship between X1, X2…, Xp and Y
2. $E(\epsilon i) = 0$
3. $Var(\epsilon i) = \sigma^2$
4. ϵi's are uncorrelated

b. No, this is not a true population regression model. The student used $E(Y_i)$. They should've used Y_i. Basically, instead of using a particular Y_i, they used the true mean of Y_i in their model.

c. This is false. The student should just be using B_j but they used $\hat{B}_j$ which is an estimator of the parameter B_j. We do not need to test an estimate.

d. This is true. The training MSE should be smaller than the test MSE because the training MSE is calculated from data used to build the model while test MSE is calculated using data the model has not seen yet.

Problem 3:

B0 = 0, B1 = 1, B2 = 2, B3 = 3, B4 = 4, B5 = 5

There are 19 individual t-tests that are significant at α = 0.05. Here, there should only be 5 predictors that are significant but the individual t-tests tell us that there are 19. We should not depend on these individual t-tests to tell us whether there is a relationship between at least one of the predictors and Y. Doing these individual t-tests could result in some small p-values. And if we do a lot of individual tests, there might be a chance of type 1 errors. For example, if we are predicting salary based on job, factors like eye color, and height are unrelated predictors. But we might see that these unrelated predictors are significant due to the multiple testing problem – this basically amplifies the probability of a false-positive result. Using f-tests can solve the issue and make sure that at least one predictor is significant.

```
set.seed(2)
x = matrix(NA,1000,200)
n=1000
beta_0 = 0
beta_1 = 1
beta_2 = 2
beta_3 = 3
beta_4 = 4
beta_5 = 5
error = rnorm(n,0,1)

for(i in 1:200){
  x[,i] = rnorm(1000)
}
Y = beta_0 + beta_1*x[,1] + beta_2*x[,2] + beta_3*x[,3] + beta_4*x[,4] + beta_5*x[,5] + error
data = as.data.frame(cbind(Y,x))

fit = lm(Y~.,data=data)
summary(fit)
p_values = summary(fit)$coefficients[,4]
length(which(p_values<0.05))
```