This problem focuses on the collinearity problem

(a) Write out the form of the linear model. What are the regression coefficients?

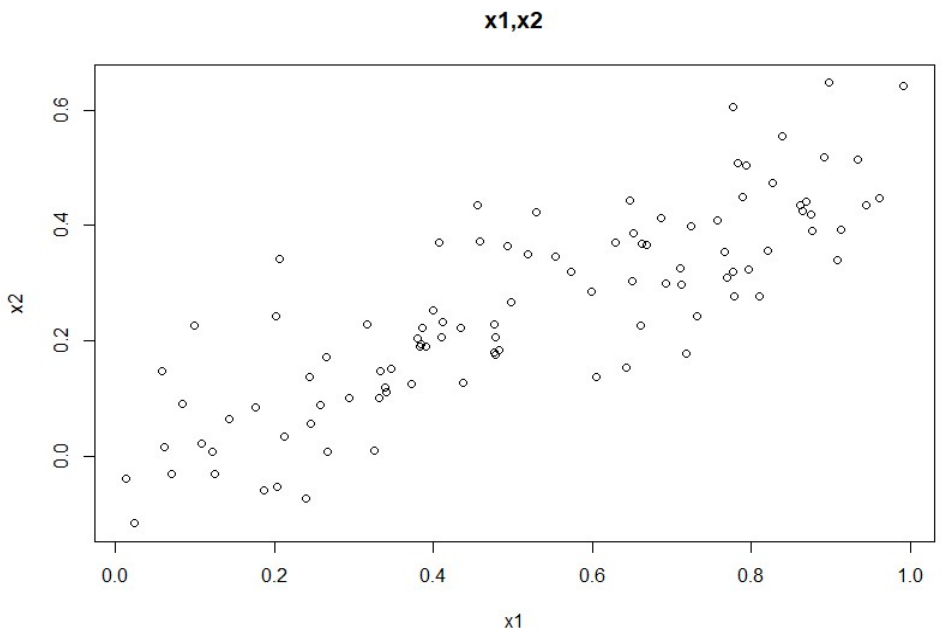
The model will be:

$$Y = \beta_0 + \beta_1 x, + \beta_2 x_2 + \varepsilon$$

$$\beta_0 = 2, \quad \beta_1 = 2, \quad \beta_2 = 0.3$$

(b) What is the correlation between x1 and x2? Create a scatterplot displaying the relationship between the variables.

The correlation is 0.7392279, which is more than 0.7. means x1 x2 is highly correlated



(c) Using this data, fit a least squares regression to predict y using x1 and x2. Describe the results obtained. What are βˆ 0, βˆ 1, and βˆ 2? How do these relate to the true β0, β1, and β2? Can you reject the null hypothesis H0 : β1 = 0? How about the null hypothesis H0 : β2 = 0?

The estimated function is $\hat{y} = 2.1305 + 1.4396x, +1.0097x_2,$

estimated coefficient is $\hat{\beta}_0 = 2.1305,\ \hat{\beta}_1 = 1.4396, \hat{\beta}_2 = 1.0097$

The $\hat{\beta}_0$ is closed to the true β_0 ,but all three predict coefficient are within the 95% interval as following:

	2.5 %	97.5 %
(Intercept)	1.670278673	2.590721
x1	0.008213776	2.870897
x2	1.240451256	3.259800

By the hypothesis testing we can said, we have enough evidence to support that Y is linearly related to x1($\beta 1 \neq 0$),and don't have enough evidence to support that Y is linearly related to x2(do not reject H0: $\beta 2=0$)

```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.1305     0.2319   9.188 7.61e-15 ***
x1            1.4396     0.7212   1.996   0.0487 *  
x2            1.0097     1.1337   0.891   0.3754    
---
```

(d) Now fit a least squares regression to predict y using only x1. Comment on your results. Can you reject the null hypothesis H0 : $\beta 1 = 0$?

```
Call:
lm(formula = y ~ x1)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.89495 -0.66874 -0.07785  0.59221  2.45560 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.1124     0.2307   9.155 8.27e-15 ***
x1            1.9759     0.3963   4.986 2.66e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.055 on 98 degrees of freedom
Multiple R-squared:  0.2024,	Adjusted R-squared:  0.1942 
F-statistic: 24.86 on 1 and 98 DF,  p-value: 2.661e-06
```

```
> anova(lm.fitd)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)    
x1         1  27.676 27.6758  24.863 2.661e-06 ***
Residuals 98 109.089  1.1132                      
```

We can see the coefficient is much closer than the result in (c), and we reject the null hypothesis H0 : $\beta 1 = 0$, that is we have enough evidence to support that Y is linearly related to x1.

(e) Now fit a least squares regression to predict y using only x2. Comment on your results. Can you reject the null hypothesis H0 : β1 = 0?

```
Call:
lm(formula = y ~ x2)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.62687 -0.75156 -0.03598  0.72383  2.44890 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.3899     0.1949   12.26  < 2e-16 ***
x2            2.8996     0.6330    4.58 1.37e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.072 on 98 degrees of freedom
Multiple R-squared:  0.1763,    Adjusted R-squared:  0.1679 
F-statistic: 20.98 on 1 and 98 DF,  p-value: 1.366e-05
```

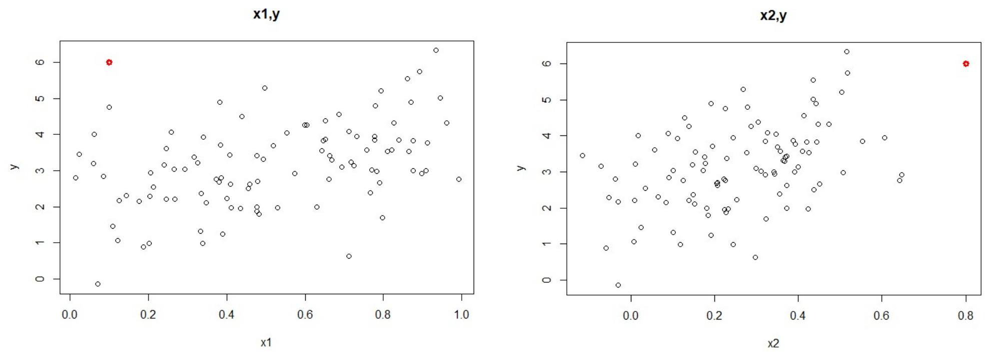
```
> anova(lm.fite)
Analysis of Variance Table

Response: y
          Df  Sum Sq Mean Sq F value    Pr(>F)    
x2         1  24.116 24.1159   20.98 1.366e-05 ***
Residuals 98 112.649  1.1495
```

Both the regression sum of square and the R-squared is less than in part(d) so Y~x1 is better than Y~x2, but we can still reach the conclusion that we should reject null hypothesis H0 : β1 = 0.

(f) Do the results obtained in (c)–(e) contradict each other? Explain your answer.

Yes, in part (c) while we do the multiple linear regression, we can only said β1is not equal to 0. However, in part (d) and (e), we have enough evidence that both β1 and β2 is not equals to 0. I think that because there is highly correlated and x1 covered the x2 effect on y.

(g) Now suppose we obtain one additional observation, which was unfortunately mismeasured. Re-fit the linear models from (c) to (e) using this new data. What effect does this new observation have on the each of the models? In each model, is this observation an outlier? A high-leverage point? Both? Explain your answers.



The scatter plot of (x1,y), (x2,y) is shown above, it seems that the new observation(red point) is a outlier on x1, and high leverage point on x2.(questions below is for further discussion)

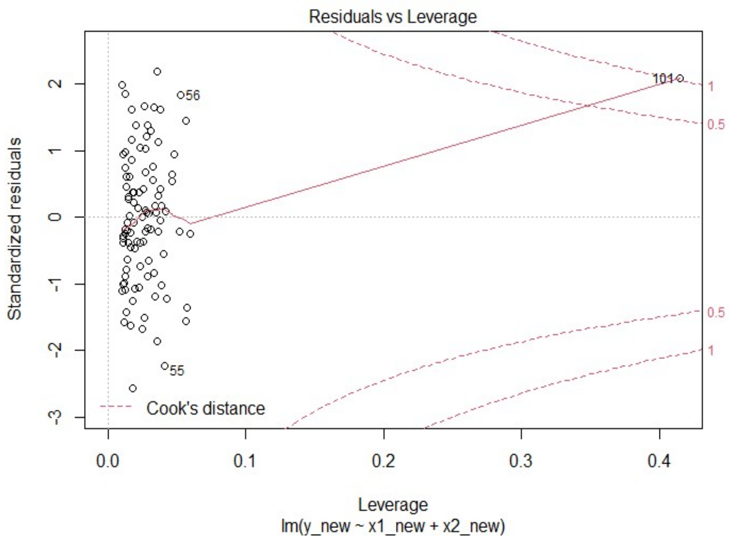
C.

The estimated function is $\hat{y} = 2.2267 + 0.5694x, +2.25146x_2,$

estimated coefficient is $\hat{\beta}_0 = 2.2267,\ \hat{\beta}_1 = 0.5694, \hat{\beta}_2$=2.25146

This time we reject H0 : β2 = 0, do not reject H0 : β1 = 0. The residual mean square increased from 1.1155 to 1.1550.

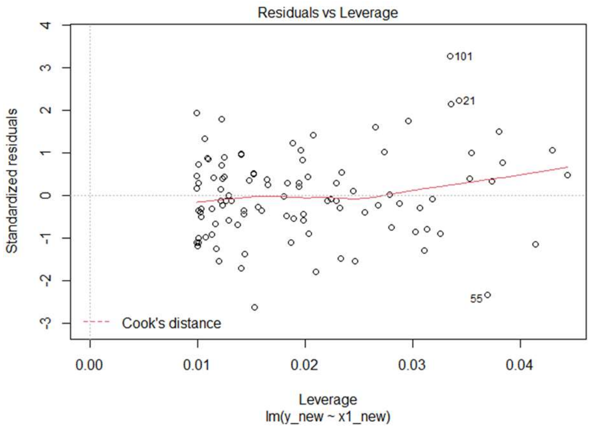
The Residuals-leverage plot shows the new observation (101) is in the top-right position, indicate that it is not only outliers but also leverage point.

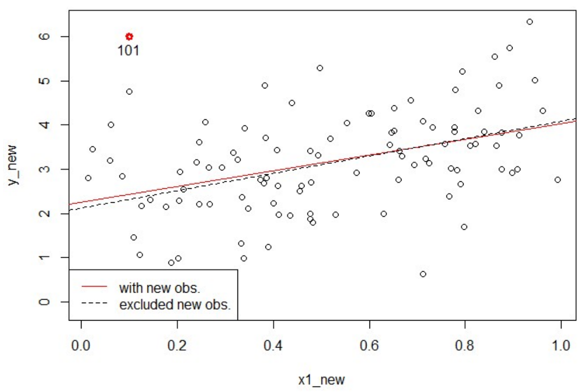


D.

```
Call:
lm(formula = y ~ x1)

Residuals:
    Min      1Q  Median      3Q     Max 
-2.8897 -0.6556 -0.0909  0.5682  3.5665 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.2569     0.2390   9.445 1.78e-15 ***
x1            1.7657     0.4124   4.282 4.29e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.111 on 99 degrees of freedom
Multiple R-squared:  0.1562,    Adjusted R-squared:  0.1477 
F-statistic: 18.33 on 1 and 99 DF,  p-value: 4.295e-05
```

We can still reject H0(H0 : β1 = 0) that there is a linear relationship between Y and x1.



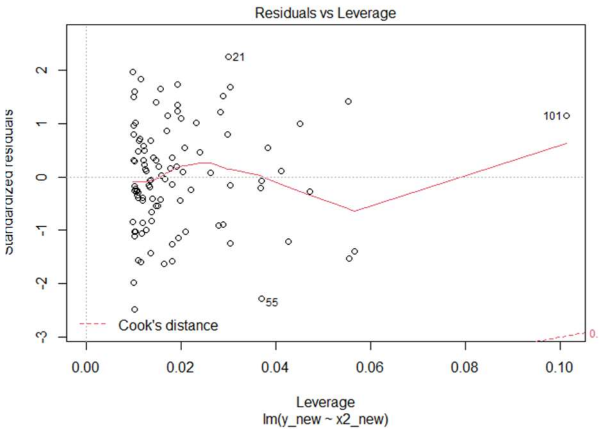


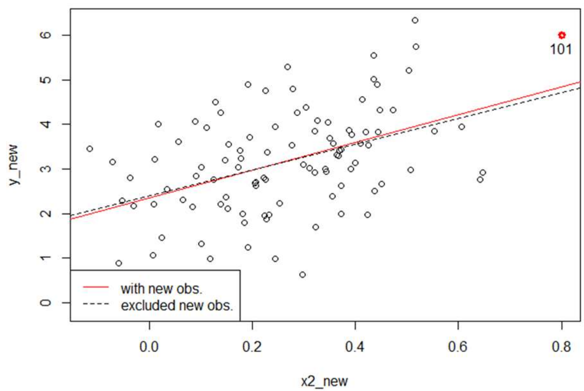
The left plot shows the new observation is an outlier (residual bigger than 3), but not the leverage points. The right plot shows how new observation effect the regression model.

e.

```
Call:
lm(formula = y_new ~ x2_new)

Residuals:
     Min       1Q   Median       3Q      Max 
-2.64729 -0.71021 -0.06899  0.72699  2.38074 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)   2.3451     0.1912  12.264  < 2e-16 ***
x2_new        3.1190     0.6040   5.164 1.25e-06 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 1.074 on 99 degrees of freedom
Multiple R-squared:  0.2122,	Adjusted R-squared:  0.2042 
F-statistic: 26.66 on 1 and 99 DF,  p-value: 1.253e-06
```

We also have same conclusion if we add new observation. Reject H0: β1 = 0, x2 and y have linear relationship.



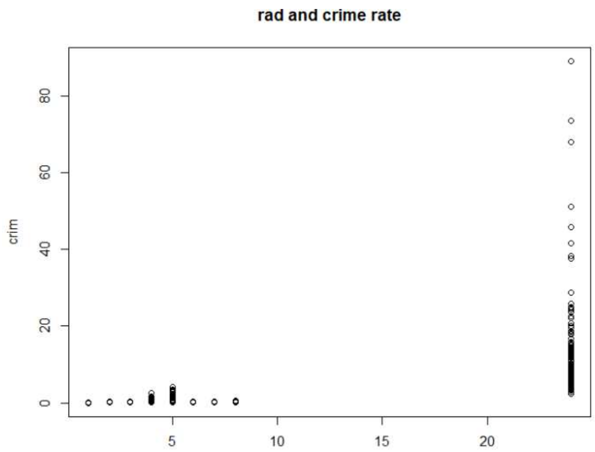


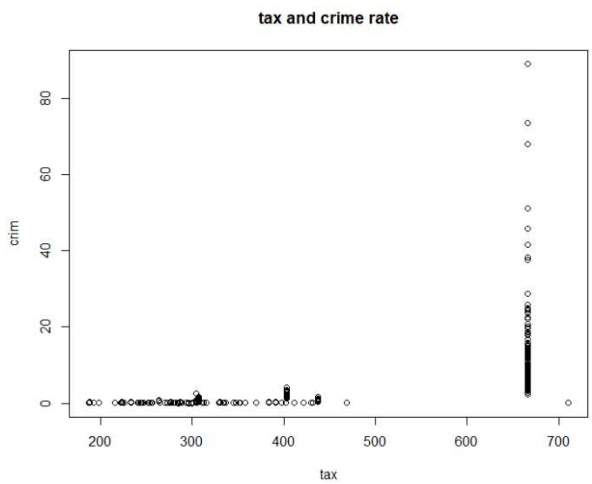
Form the plots above, it says the new observation have large leverage value, but standardized residual is less than 3 so it is a leverage point but not a outlier, The right plot shows how new observation effect the regression model.

This problem involves the Boston data set, which we saw in the lab for this chapter. We will now try to predict per capita crime rate using the other variables in this data set. In other words, per capita crime rate is the response, and the other variables are the predictors.

(a) For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

almost every model(but *lm(crim~chas)*) is statistically significant (reject H0: β1 = 0), but some model have higher R-squared values *lm(crim ~ rad),* and *lm(crim ~ tax)* ,the R-squared is 0.3913 and 0.3396, respectively. The graphs below show the higher tax and "rad", the higher the crime rate .



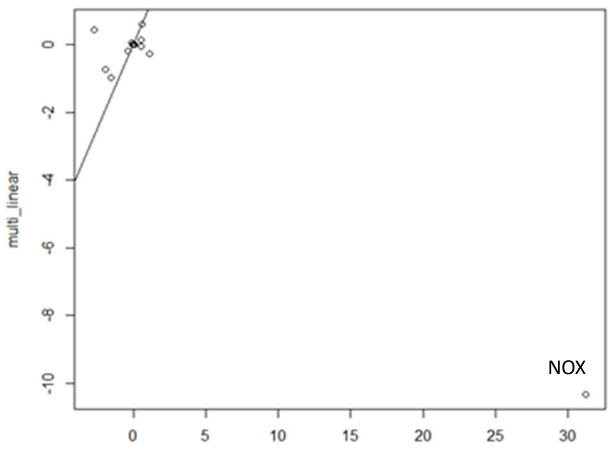


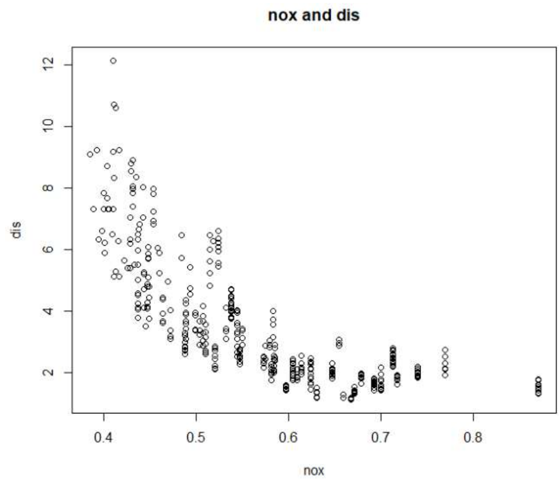
(b) Fit a multiple regression model to predict the response using all of the predictors. Describe your results. For which predictors can we reject the null hypothesis H0 : βj = 0?

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

Residuals:
   Min     1Q Median     3Q    Max 
-9.924 -2.120 -0.353  1.019 75.051 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  17.033228   7.234903   2.354 0.018949 *  
zn            0.044855   0.018734   2.394 0.017025 *  
indus        -0.063855   0.083407  -0.766 0.444294    
chas         -0.749134   1.180147  -0.635 0.525867    
nox         -10.313535   5.275536  -1.955 0.051152 .  
rm            0.430131   0.612830   0.702 0.483089    
age           0.001452   0.017925   0.081 0.935488    
dis          -0.987176   0.281817  -3.503 0.000502 ***
rad           0.588209   0.088049   6.680 6.46e-11 ***
tax          -0.003780   0.005156  -0.733 0.463793    
ptratio      -0.271081   0.186450  -1.454 0.146611    
black        -0.007538   0.003673  -2.052 0.040702 *  
lstat         0.126211   0.075725   1.667 0.096208 .  
medv         -0.198887   0.060516  -3.287 0.001087 ** 
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.439 on 492 degrees of freedom
Multiple R-squared:  0.454,	Adjusted R-squared:  0.4396 
F-statistic: 31.47 on 13 and 492 DF,  p-value: < 2.2e-16
```

Under the significant level at 0.05, we can reject H0 : βj = 0 in variable zn, dis, rad, black, medv. Under the multiple linear regression, we get higher R-squared values 0.454.

(c) How do your results from (a) compare to your results from (b)? Create a plot displaying the univariate regression coefficients from (a) on the x-axis, and the multiple regression coefficients from (b) on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in a simple linear regression model is shown on the x-axis, and its coefficient estimate in the multiple linear regression model is shown on the y-axis.





The plot is shown in left, most of the coefficient is close to zero. The black line is x=y, Means having same coefficient in both single and multiple linear regression. Only some variables are on the line. Noticed that the variable “NOX” have extremely high value in simple regression but become low in the multiple regression. This may because it has collinearity with other variables. The scatterplot on the right shows there might be a negative relationship between “nox” and “dis”.

(d) Is there evidence of non-linear association between any of the predictors and the response? To answer this question, for each predictor X, fit a model of the form
$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3$

R report an error while doing *lm(crim~poly(chas,3))* because data "chas" only include 0 and 1, the degree of poly. should not be greater than the unique points of predictor.

Most of model indicate every coefficient aren't equals to zero.

	Zn	Indus	Chas	Nox	Rm	Age	Dis	Rad	Tax	Ptratio	Black	Lstat	medv
X^1	***	***	Na	***	***	***	***	***	***	***	***	***	***
X^2	<.005	<.005	Na	***	<.005	***	***	<0.01	***	<.005	.45	<.05	***
X^3	.22	***	Na	***	.5	<0.01	***	.48	.24	<.01	.54	.13	***

*** for p-value <0.001, blue color means we don't reject H0 in hypothesis testing

We can see the "nox", "dis", "medv" have is significant not equal to zero, furthermore, while applying the polynomial regression, the R-squared value increase. For example, "medv" R-squared increase form 0.15 to 0.42 and "nox" R-squared value increase form 0.177 to 0.3. The plot below shows there may be some kind of polynomial relationship.