CptS 575: Data Science, Fall 2018

Sukhjinder Singh

Assignment 4: Linear Regression

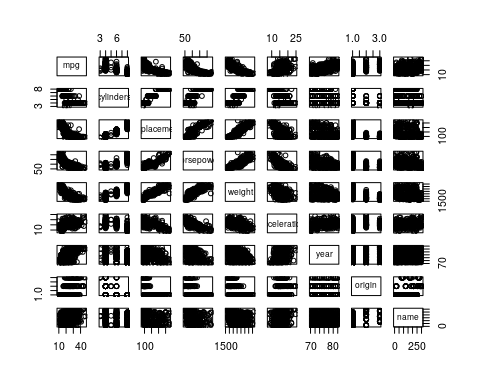
# Problem 1

This question involves the use of multiple linear regression on the Auto data set from the course webpage (<https://scads.eecs.wsu.edu/index.php/datasets/>). Ensure that you remove missing values from the dataframe, and that values are represented in the appropriate types (num or int for quantitative variables, factor, logi or str for qualitative).

Auto = read.csv("Auto.csv", header = T, na.strings = "?")

**a).Produce a scatterplot matrix which includes all of the variables in the data set.**

pairs(Auto)



Auto <- na.omit(Auto)

**b).Compute the matrix of correlations between the variables using the function cor(). You will need to exclude the name variable, which is qualitative.**

cor(Auto[sapply(Auto, function(x) !is.factor(x))])

## mpg cylinders displacement horsepower weight  
## mpg 1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442  
## cylinders -0.7776175 1.0000000 0.9508233 0.8429834 0.8975273  
## displacement -0.8051269 0.9508233 1.0000000 0.8972570 0.9329944  
## horsepower -0.7784268 0.8429834 0.8972570 1.0000000 0.8645377  
## weight -0.8322442 0.8975273 0.9329944 0.8645377 1.0000000  
## acceleration 0.4233285 -0.5046834 -0.5438005 -0.6891955 -0.4168392  
## year 0.5805410 -0.3456474 -0.3698552 -0.4163615 -0.3091199  
## origin 0.5652088 -0.5689316 -0.6145351 -0.4551715 -0.5850054  
## acceleration year origin  
## mpg 0.4233285 0.5805410 0.5652088  
## cylinders -0.5046834 -0.3456474 -0.5689316  
## displacement -0.5438005 -0.3698552 -0.6145351  
## horsepower -0.6891955 -0.4163615 -0.4551715  
## weight -0.4168392 -0.3091199 -0.5850054  
## acceleration 1.0000000 0.2903161 0.2127458  
## year 0.2903161 1.0000000 0.1815277  
## origin 0.2127458 0.1815277 1.0000000

**C).Use the lm() function to perform a multiple linear regression with mpg as the response and all other variables except name as the predictors. Use the summary() function to print the results. Comment on the output:**

model = lm(mpg ~. -name, data = Auto)  
summary(model)

##   
## Call:  
## lm(formula = mpg ~ . - name, data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.5903 -2.1565 -0.1169 1.8690 13.0604   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -17.218435 4.644294 -3.707 0.00024 \*\*\*  
## cylinders -0.493376 0.323282 -1.526 0.12780   
## displacement 0.019896 0.007515 2.647 0.00844 \*\*   
## horsepower -0.016951 0.013787 -1.230 0.21963   
## weight -0.006474 0.000652 -9.929 < 2e-16 \*\*\*  
## acceleration 0.080576 0.098845 0.815 0.41548   
## year 0.750773 0.050973 14.729 < 2e-16 \*\*\*  
## origin 1.426141 0.278136 5.127 4.67e-07 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.328 on 384 degrees of freedom  
## Multiple R-squared: 0.8215, Adjusted R-squared: 0.8182   
## F-statistic: 252.4 on 7 and 384 DF, p-value: < 2.2e-16

**i).Which predictors appear to have a statistically significant relationship to the response, and how do you determine this?**

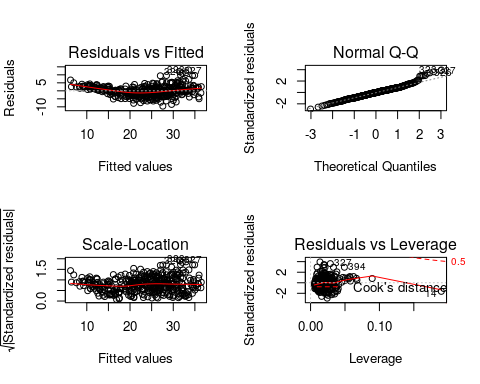
The low p-values for displacement, weight, year, and origin indicate a statistically significant relationship to mpg.The predictors that have the highest statstical sginficance are: weight, year and origin. This makes intuitive sense, as lighter cars would logically have better mpg and more modern cars employ better gas-saving technology. displacement and horsepower are also sgificant at the .05 level.

**ii).What does the coefficient for the cylinders variable suggest, in simple terms?**

The coefficient ot the “cylinder” variable suggests that the average effect of an increase of 1 year is an decrease of 0.493376 in “mpg” (all other predictors remaining constant).

**d).Use the plot() function to produce diagnostic plots of the linear regression fit. Comment on any problems you see with the fit. Do the residual plots suggest any unusually large outliers? Does the leverage plot identify any observations with unusually high leverage?**

par(mfrow=c(2,2))  
plot(model)



I have come up with diifferent comments based on these graphs.

1). On-linearity of response-predictors values  
 There does not seems to be any pattern for Residuals vs Fitted graph, so it points no strong evidence of non-linearity.  
  
2). Non-constant Variance of Error Terms  
 There is a bit of funnel shape(assume) for the Residuals vs Fitted graph, so it presents a bit of heteroscedasticity.  
  
3).High Leverage Points  
 Specifically the observation 14 is a highly leverage point as shown in Residuals vc Leverage graph.

**e).Use the ‘\*’ and : symbols to fit linear regression models with interaction effects. Do any interactions appear to be statistically significant?**

autolm3 = lm(mpg ~ (.-name)\*(.-name), data = Auto)  
summary(autolm3)

##   
## Call:  
## lm(formula = mpg ~ (. - name) \* (. - name), data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -7.6303 -1.4481 0.0596 1.2739 11.1386   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.548e+01 5.314e+01 0.668 0.50475   
## cylinders 6.989e+00 8.248e+00 0.847 0.39738   
## displacement -4.785e-01 1.894e-01 -2.527 0.01192 \*   
## horsepower 5.034e-01 3.470e-01 1.451 0.14769   
## weight 4.133e-03 1.759e-02 0.235 0.81442   
## acceleration -5.859e+00 2.174e+00 -2.696 0.00735 \*\*  
## year 6.974e-01 6.097e-01 1.144 0.25340   
## origin -2.090e+01 7.097e+00 -2.944 0.00345 \*\*  
## cylinders:displacement -3.383e-03 6.455e-03 -0.524 0.60051   
## cylinders:horsepower 1.161e-02 2.420e-02 0.480 0.63157   
## cylinders:weight 3.575e-04 8.955e-04 0.399 0.69000   
## cylinders:acceleration 2.779e-01 1.664e-01 1.670 0.09584 .   
## cylinders:year -1.741e-01 9.714e-02 -1.793 0.07389 .   
## cylinders:origin 4.022e-01 4.926e-01 0.816 0.41482   
## displacement:horsepower -8.491e-05 2.885e-04 -0.294 0.76867   
## displacement:weight 2.472e-05 1.470e-05 1.682 0.09342 .   
## displacement:acceleration -3.479e-03 3.342e-03 -1.041 0.29853   
## displacement:year 5.934e-03 2.391e-03 2.482 0.01352 \*   
## displacement:origin 2.398e-02 1.947e-02 1.232 0.21875   
## horsepower:weight -1.968e-05 2.924e-05 -0.673 0.50124   
## horsepower:acceleration -7.213e-03 3.719e-03 -1.939 0.05325 .   
## horsepower:year -5.838e-03 3.938e-03 -1.482 0.13916   
## horsepower:origin 2.233e-03 2.930e-02 0.076 0.93931   
## weight:acceleration 2.346e-04 2.289e-04 1.025 0.30596   
## weight:year -2.245e-04 2.127e-04 -1.056 0.29182   
## weight:origin -5.789e-04 1.591e-03 -0.364 0.71623   
## acceleration:year 5.562e-02 2.558e-02 2.174 0.03033 \*   
## acceleration:origin 4.583e-01 1.567e-01 2.926 0.00365 \*\*  
## year:origin 1.393e-01 7.399e-02 1.882 0.06062 .   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.695 on 363 degrees of freedom  
## Multiple R-squared: 0.8893, Adjusted R-squared: 0.8808   
## F-statistic: 104.2 on 28 and 363 DF, p-value: < 2.2e-16

If we look into the model then we had an improvement in R2 from 0.82 to almost 0.89, maybe it can be overfitting, though the interactive term most significant was acceleration:origin with a good coefficient in comparison with the main terms and a small p-value, validating thecoefficient.If we check the interactions between displacement and year, acceleration and year, and acceleration and origin all have low p values that indicate significance.

**f).Try transformations of the variables with X3 and log(X). Comment on your findings.**

par(mfrow = c(2, 2))  
autolmx2 <- lm(mpg ~ (horsepower)^3 + (weight)^3 + (acceleration)^3, data = Auto)  
summary(autolmx2)

##   
## Call:  
## lm(formula = mpg ~ (horsepower)^3 + (weight)^3 + (acceleration)^3,   
## data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -11.079 -2.736 -0.331 2.170 16.262   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 45.6782929 2.4085431 18.965 < 2e-16 \*\*\*  
## horsepower -0.0474956 0.0159891 -2.970 0.00316 \*\*   
## weight -0.0057894 0.0005776 -10.024 < 2e-16 \*\*\*  
## acceleration -0.0020657 0.1233378 -0.017 0.98665   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 4.246 on 388 degrees of freedom  
## Multiple R-squared: 0.7064, Adjusted R-squared: 0.7041   
## F-statistic: 311.1 on 3 and 388 DF, p-value: < 2.2e-16

autolmlog <- lm(mpg ~ log(horsepower) + log(weight) + log(acceleration), data = Auto)  
summary(autolmlog)

##   
## Call:  
## lm(formula = mpg ~ log(horsepower) + log(weight) + log(acceleration),   
## data = Auto)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -10.8237 -2.5240 -0.2389 2.0105 15.3681   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 190.152 8.255 23.035 < 2e-16 \*\*\*  
## log(horsepower) -11.799 1.933 -6.103 2.53e-09 \*\*\*  
## log(weight) -12.306 1.820 -6.762 5.03e-11 \*\*\*  
## log(acceleration) -5.363 1.970 -2.723 0.00677 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 3.961 on 388 degrees of freedom  
## Multiple R-squared: 0.7445, Adjusted R-squared: 0.7425   
## F-statistic: 376.8 on 3 and 388 DF, p-value: < 2.2e-16

After applying the log function to each of the variables which resulted into the highest R2 value and F-statistic. It also provided the lowest individual p-values for horsepower and acceleration while squaring the weight variable resulted in the lowest p-value.

# Problem 2

This problem involves the Boston data set, which we saw in the lab. 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.

library(MASS)  
attach(Boston)  
summary(Boston)

## crim zn indus chas   
## Min. : 0.00632 Min. : 0.00 Min. : 0.46 Min. :0.00000   
## 1st Qu.: 0.08204 1st Qu.: 0.00 1st Qu.: 5.19 1st Qu.:0.00000   
## Median : 0.25651 Median : 0.00 Median : 9.69 Median :0.00000   
## Mean : 3.61352 Mean : 11.36 Mean :11.14 Mean :0.06917   
## 3rd Qu.: 3.67708 3rd Qu.: 12.50 3rd Qu.:18.10 3rd Qu.:0.00000   
## Max. :88.97620 Max. :100.00 Max. :27.74 Max. :1.00000   
## nox rm age dis   
## Min. :0.3850 Min. :3.561 Min. : 2.90 Min. : 1.130   
## 1st Qu.:0.4490 1st Qu.:5.886 1st Qu.: 45.02 1st Qu.: 2.100   
## Median :0.5380 Median :6.208 Median : 77.50 Median : 3.207   
## Mean :0.5547 Mean :6.285 Mean : 68.57 Mean : 3.795   
## 3rd Qu.:0.6240 3rd Qu.:6.623 3rd Qu.: 94.08 3rd Qu.: 5.188   
## Max. :0.8710 Max. :8.780 Max. :100.00 Max. :12.127   
## rad tax ptratio black   
## Min. : 1.000 Min. :187.0 Min. :12.60 Min. : 0.32   
## 1st Qu.: 4.000 1st Qu.:279.0 1st Qu.:17.40 1st Qu.:375.38   
## Median : 5.000 Median :330.0 Median :19.05 Median :391.44   
## Mean : 9.549 Mean :408.2 Mean :18.46 Mean :356.67   
## 3rd Qu.:24.000 3rd Qu.:666.0 3rd Qu.:20.20 3rd Qu.:396.23   
## Max. :24.000 Max. :711.0 Max. :22.00 Max. :396.90   
## lstat medv   
## Min. : 1.73 Min. : 5.00   
## 1st Qu.: 6.95 1st Qu.:17.02   
## Median :11.36 Median :21.20   
## Mean :12.65 Mean :22.53   
## 3rd Qu.:16.95 3rd Qu.:25.00   
## Max. :37.97 Max. :50.00

a). For each predictor, fit a simple linear regression model to predict the response. Include the code, but not the output for all models in your solution.

**Include the code for all the model**

#First Model  
fitresultzn <- lm(crim ~ zn)  
#Second model  
fitresultindus <- lm(crim ~ indus)  
#Third Model  
chas <- as.factor(chas)  
#Fourth Model  
fitresultchas <- lm(crim ~ chas)  
#Fifth Model  
fitresultnox <- lm(crim ~ nox)  
#6th Model  
fitresultrm <- lm(crim ~ rm)  
#7th model  
fitresultage <- lm(crim ~ age)  
#8th Model  
fitresultdis <- lm(crim ~ dis)  
#9th Model  
fitresultrad <- lm(crim ~ rad)  
#10th Model  
fitresulttax <- lm(crim ~ tax)  
#11th Model  
fitresultmedv <- lm(crim ~ medv)  
#12thmodel  
fitresultptratio = lm(crim ~ ptratio)  
#13th model  
fitresultblack = lm(crim ~ black)  
#14th model  
fitresultlstat = lm(crim ~ lstat)

**In which of the models is there a statistically significant association between the predictor and the response?**

To find which model has significant association between the predictor and the response, we have to test H0:β1=0. All predictors have a p-value less than 0.05 except “chas”, so we may conclude that there is a statistically significant association between each predictor and the response except for the “chas” predictor.

**Considering the meaning of each variable, discuss the relationship between crim and nox, chas, medv and dis in particular. How do these relationships differ?**

I have considered following relationships when I saw the dataset and run the linear model to check the relationships.

1).We can see that there is a strong correlation between the predictor and the response for every variable apart from the Charles River Dummy.   
2). Linear regression with the response variables vs crime in simple scatter-plots gives us a better prediction of crime than just using the mean of crime.   
3).The low R(squared) indicates that the level of the variation in the response described by these predictors is also very low.   
4).When looking at the response variables and crime in simple scatter plots, one can see how a general linear regression with these variables would allow for a better prediction of crime than simply using the mean of crime. That is, the data seems to have some slight shape sloping up or down, and isn’t a random cloud of data. That being said, while almost every variable is statistically significant, R-squared is very low, and so these predictors only describe a small amount of the variation in the response.

fitresultnox <- lm(crim ~ nox)  
summary(fitresultnox)

##   
## Call:  
## lm(formula = crim ~ nox)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -12.371 -2.738 -0.974 0.559 81.728   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -13.720 1.699 -8.073 5.08e-15 \*\*\*  
## nox 31.249 2.999 10.419 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.81 on 504 degrees of freedom  
## Multiple R-squared: 0.1772, Adjusted R-squared: 0.1756   
## F-statistic: 108.6 on 1 and 504 DF, p-value: < 2.2e-16

chas <- as.factor(chas)  
fitresultchas <- lm(crim ~ chas)  
summary(fitresultchas)

##   
## Call:  
## lm(formula = crim ~ chas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -3.738 -3.661 -3.435 0.018 85.232   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 3.7444 0.3961 9.453 <2e-16 \*\*\*  
## chas1 -1.8928 1.5061 -1.257 0.209   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 8.597 on 504 degrees of freedom  
## Multiple R-squared: 0.003124, Adjusted R-squared: 0.001146   
## F-statistic: 1.579 on 1 and 504 DF, p-value: 0.2094

fit.medv <- lm(crim ~ medv)  
summary(fitresultmedv)

##   
## Call:  
## lm(formula = crim ~ medv)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -9.071 -4.022 -2.343 1.298 80.957   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 11.79654 0.93419 12.63 <2e-16 \*\*\*  
## medv -0.36316 0.03839 -9.46 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.934 on 504 degrees of freedom  
## Multiple R-squared: 0.1508, Adjusted R-squared: 0.1491   
## F-statistic: 89.49 on 1 and 504 DF, p-value: < 2.2e-16

fitresultdis <- lm(crim ~ dis)  
summary(fitresultdis)

##   
## Call:  
## lm(formula = crim ~ dis)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -6.708 -4.134 -1.527 1.516 81.674   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 9.4993 0.7304 13.006 <2e-16 \*\*\*  
## dis -1.5509 0.1683 -9.213 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 7.965 on 504 degrees of freedom  
## Multiple R-squared: 0.1441, Adjusted R-squared: 0.1425   
## F-statistic: 84.89 on 1 and 504 DF, p-value: < 2.2e-16

**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?**

fit.all <- lm(crim ~ ., data = Boston)  
summary(fit.all)

##   
## 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

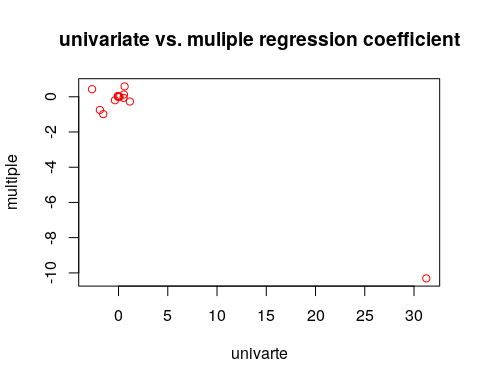
As we fit the multiple regression model, very few variables appear to be statistically significant at the following levels:  
dis- .001, rad- .001, medv – .01, black – .05 and zn -.05. In this case R squared is significantly higher than either of the predictors.  
For every other variable, The Null Hypothesis cannot be rejected for all other variables. R-squared is also much higher using a multiple regression model than any of the predictors on their own, meaning we better explain more of the variance in the outcome.

**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. What does this plot tell you about the various predictors?**

univcoof=c(coefficients(fitresultzn)[2],  
 coefficients(fitresultindus)[2],  
 coefficients(fitresultchas)[2],  
 coefficients(fitresultnox)[2],  
 coefficients(fitresultrm)[2],  
 coefficients(fitresultage)[2],  
 coefficients(fitresultdis)[2],  
 coefficients(fitresultrad)[2],  
 coefficients(fitresulttax)[2],  
 coefficients(fitresultptratio)[2],  
 coefficients(fitresultblack)[2],  
 coefficients(fitresultlstat)[2],  
 coefficients(fitresultmedv)[2])  
  
fooBoston <- (lm(crim ~., data = Boston))  
  
fooBoston$coefficients[2:14]

## zn indus chas nox rm   
## 0.044855215 -0.063854824 -0.749133611 -10.313534912 0.430130506   
## age dis rad tax ptratio   
## 0.001451643 -0.987175726 0.588208591 -0.003780016 -0.271080558   
## black lstat medv   
## -0.007537505 0.126211376 -0.198886821

plot(univcoof,fooBoston$coefficients[2:14],main = "univariate vs. muliple regression coefficient",xlab = "univarte",ylab = "multiple",col="red")



If we look into the plots, then there is a difference between the simple and multiple regression coefficients. This difference is due to the fact that in the simple regression case, the slope term represents the average effect of an increase in the predictor, ignoring other predictors.   
  
In contrast, in the multiple regression case, the slope term represents the average effect of an increase in the predictor, while holding other predictors fixed. It does make sense for the multiple regression to suggest no relationship between the response and some of the predictors while the simple linear regression implies the opposite because the correlation between the predictors show some strong relationships between some of the predictors.

cor(Boston[-c(1, 4)])

## zn indus nox rm age dis  
## zn 1.0000000 -0.5338282 -0.5166037 0.3119906 -0.5695373 0.6644082  
## indus -0.5338282 1.0000000 0.7636514 -0.3916759 0.6447785 -0.7080270  
## nox -0.5166037 0.7636514 1.0000000 -0.3021882 0.7314701 -0.7692301  
## rm 0.3119906 -0.3916759 -0.3021882 1.0000000 -0.2402649 0.2052462  
## age -0.5695373 0.6447785 0.7314701 -0.2402649 1.0000000 -0.7478805  
## dis 0.6644082 -0.7080270 -0.7692301 0.2052462 -0.7478805 1.0000000  
## rad -0.3119478 0.5951293 0.6114406 -0.2098467 0.4560225 -0.4945879  
## tax -0.3145633 0.7207602 0.6680232 -0.2920478 0.5064556 -0.5344316  
## ptratio -0.3916785 0.3832476 0.1889327 -0.3555015 0.2615150 -0.2324705  
## black 0.1755203 -0.3569765 -0.3800506 0.1280686 -0.2735340 0.2915117  
## lstat -0.4129946 0.6037997 0.5908789 -0.6138083 0.6023385 -0.4969958  
## medv 0.3604453 -0.4837252 -0.4273208 0.6953599 -0.3769546 0.2499287  
## rad tax ptratio black lstat medv  
## zn -0.3119478 -0.3145633 -0.3916785 0.1755203 -0.4129946 0.3604453  
## indus 0.5951293 0.7207602 0.3832476 -0.3569765 0.6037997 -0.4837252  
## nox 0.6114406 0.6680232 0.1889327 -0.3800506 0.5908789 -0.4273208  
## rm -0.2098467 -0.2920478 -0.3555015 0.1280686 -0.6138083 0.6953599  
## age 0.4560225 0.5064556 0.2615150 -0.2735340 0.6023385 -0.3769546  
## dis -0.4945879 -0.5344316 -0.2324705 0.2915117 -0.4969958 0.2499287  
## rad 1.0000000 0.9102282 0.4647412 -0.4444128 0.4886763 -0.3816262  
## tax 0.9102282 1.0000000 0.4608530 -0.4418080 0.5439934 -0.4685359  
## ptratio 0.4647412 0.4608530 1.0000000 -0.1773833 0.3740443 -0.5077867  
## black -0.4444128 -0.4418080 -0.1773833 1.0000000 -0.3660869 0.3334608  
## lstat 0.4886763 0.5439934 0.3740443 -0.3660869 1.0000000 -0.7376627  
## medv -0.3816262 -0.4685359 -0.5077867 0.3334608 -0.7376627 1.0000000

**For example,**

when “age” is high there is a tendency in “dis” to be low, hence we can say in simple linear regression which only examines “crim” versus “age”, we can observe that there is higher values of “age” are associated with higher values of “crim”, even though “age” does not actually affect “crim”.So “age” is a surrogate for “dis”; “age” gets credit for the effect of “dis” on “crim”.

**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 = β0 + β1X + β2X2 + β3X3+ ε**

y=lm(crim ~ zn + I(zn^2) + I(zn^3), data = Boston)  
y=lm(crim ~ indus + I(indus^2) + I(indus^3), data = Boston)  
y=lm(crim ~ chas + I(chas^2) + I(chas^3), data = Boston)  
y=lm(crim ~ nox + I(nox^2) + I(nox^3), data = Boston)  
y=lm(crim ~ rm + I(rm^2) + I(rm^3), data = Boston)  
y=lm(crim ~ age + I(age^2) + I(age^3), data = Boston)  
y=lm(crim ~ dis + I(dis^2) + I(dis^3), data = Boston)  
y=lm(crim ~ rad + I(rad^2) + I(rad^3), data = Boston)  
y=lm(crim ~ tax + I(tax^2) + I(tax^3), data = Boston)  
y=lm(crim ~ ptratio + I(ptratio^2) + I(ptratio^3), data = Boston)  
y=lm(crim ~ black + I(black^2) + I(black^3), data = Boston)  
y=lm(crim ~ lstat + I(lstat^2) + I(lstat^3), data = Boston)  
y=lm(crim ~ medv + I(medv^2) + I(medv^3), data = Boston)

Observing the results,

The first thing to note is the chas variable, we get NA values for the squared and cubed term. This makes sense as chas is a dummy variable, composed of only 0s and 1s, and these values will not change if they are squared or cubed.I come up with some analysis results such as For “zn”, “rm”, “rad”, “tax” and “lstat” as predictor, the p-values suggest that the cubic coefficient is not statistically significant; however,for “indus”, “nox”, “age”, “dis”, “ptratio” and “medv” as predictor, the p-values suggest the adequacy of the cubic fit;Similarly, for “black” as predictor, the p-values suggest that the quandratic and cubic coefficients are not statistically significant, so in this latter case no non-linear effect is visible.

# Problem 3

An important assumption of the linear regression model is that the error terms are uncorrelated (independent). But error terms can sometimes be correlated, especially in time-series data.

**a).What are the issues that could arise in using linear regression (via least squares estimates) when error terms are correlated? Comment in particular with respect to i) regression coefficients ii) the standard error of regression coefficients iii) confidence intervals**

Solution:

**(i) Regression Coefficients**

Regression coefficients represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. This statistical control that regression provides is important because it isolates the role of one variable from all of the others in the model  
  
Multicolinearity is often at the source of the problem when a positive simple correlation with the dependent variable leads to a negative regression coefficient in multiple regression. Some regression techniques may help there : ridge regression, partial least square regression. Start by finding out which variable(s) are causing the colinearity (i.e with the inflation or the distortion factor). Remove them or attenuate the correlation with the ridge coefficient.

**(ii) Standard error of regression coefficients**

The sample standard deviation of the errors is a downward-biased estimate of the size of the true unexplained deviations in Y because it does not adjust for the additional "degree of freedom" used up by estimating the slope coefficient. An unbiased estimate of the standard deviation of the true errors is given by the standard error of the regression, denoted by s. In the special case of a simple regression model, it is:

**Standard error of regression coefficient = STDEV.S(errors) x SQRT((n-1)/(n-2))**

The sum of squared errors is divided by n-2 in this calculation rather than n-1 because an additional degree of freedom for error has been used up by estimating two parameters (a slope and an intercept) rather than only one (the mean) in fitting the model to the data. The standard error of the regression is an unbiased estimate of the standard deviation of the noise in the data, i.e., the variations in Y that are not explained by the model.

**For Example :**

When multicollinearity occurs, the least-squares estimates are still unbiased and efficient. The problem is that the estimated standard errors of the coefficients tend to be inflated. That is, the standard error tends to be larger than it would be in the absence of multicollinearity because the estimates are very sensitive to changes in the sample observations or in the model specification. In other words, including or excluding a particular variable or certain observations may greatly change the estimated coefficients.

**(iii) Confidence Intervals**

Confidence intervals for the mean and for the forecast are equal to the point estimate plus-or-minus the appropriate standard error multiplied by the appropriate 2-tailed critical value of the t distribution. The critical value that should be used depends on the number of degrees of freedom for error (the number data points minus number of parameters estimated, which is n-1 for this model) and the desired level of confidence. So, for example, a 95% confidence interval for the forecast is given by   
  
Bo=(plus-or-minus) SEfcst + T.INV.2T(0.05,n-1)  
  
In general, T.INV.2T(0.05, n-1) is fairly close to 2 except for very small samples, i.e., a 95% confidence interval for the forecast is roughly equal to the forecast plus-or-minus two standard errors.

**For Example:**

The issue that depends on the correctness of the model and the representativeness of the data set, particularly in the case of time series data. If the model is not correct or there are unusual patterns in the data, then if the confidence interval for one period's forecast fails to cover the true value, it is relatively more likely that the confidence interval for a neighboring period's forecast will also fail to cover the true value, because the model may have a tendency to make the same error for several periods in a row.

**b).What methods can be applied to deal with correlated errors? Mention at least one method.**

As I mentioned if there’s multicollinearity problem occurs in linear regression model or VIF for a factor is near or above 5.We can apply following methods.

1). Remove highly correlated predictors from the model- If we have two or more factor with a high variance inflation factor, remove one of the model.Because they supply redundant information, removing one of the correlated factors usually doesn't drastically reduce the R-squared. We can consider using stepwise regression, best subsets regression, or specialized knowledge of the data set to remove these variables.We can select the model that has the highest R-squared value.  
  
2).We can use Partial Least Squares Regression (PLS) or Principal Components Analysis (PCA), regression methods that cut the number of predictors to a smaller set of uncorrelated components.