Q17

Rahul Atre

2023-10-17

Q17 [Regression Modeling]

This problem involves the Boston dataset, 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 dataset. In other words, per capita crime rate is the response, and the other variables are the predictors.

1. 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.

First we must load the Boston dataset (it is part of the MASS library in R).

```r
data("Boston")
```

```r
fit1 = lm(Boston$crim ~ Boston$zn)
fit2 = lm(Boston$crim ~ Boston$indus)
fit3 = lm(Boston$crim ~ Boston$chas)
fit4 = lm(Boston$crim ~ Boston$nox)
fit5 = lm(Boston$crim ~ Boston$rm)
fit6 = lm(Boston$crim ~ Boston$age)
fit7 = lm(Boston$crim ~ Boston$dis)
fit8 = lm(Boston$crim ~ Boston$rad)
fit9 = lm(Boston$crim ~ Boston$tax)
fit10 = lm(Boston$crim ~ Boston$ptratio)
fit11 = lm(Boston$crim ~ Boston$black)
fit12 = lm(Boston$crim ~ Boston$lstat)
fit13 = lm(Boston$crim ~ Boston$medv)
```

Let's now generate the summary tables to obtain the parameters of each fitted model (using crime as response and the 13 other variables as predictor). The summary table's have been hidden to keep the page limit under 6.

$Y = \beta_0 + \beta_1 X$, where Y rep. the crim and X rep. one out of the 13 predictor variables.

```r
summary(fit1)
summary(fit2)
summary(fit3)
summary(fit4)
summary(fit5)
summary(fit6)
summary(fit7)
summary(fit8)
summary(fit9)
summary(fit10)
summary(fit11)
summary(fit12)
summary(fit13)
```

Based on the summary table, we can obtain the estimated linear regression line, and p-value, which will let us know which parameters are statistically significant.

- Crim (per capita crime rate) and zn: Y = 4.45369 - 0.07393X, **statistically significant**
- Crim and indus: Y = -2.06374 + 0.50978X, **statistically significant**
- Crim and chas: Y = 3.7444 - 1.8928X
- Crim and nox: Y = -13.720 + 31.249X, **statistically significant**
- Crim and rm: Y = 20.482 - 2.684X, **statistically significant**
- Crim and age: Y = -3.77791 + 0.10779X, **statistically significant**
- Crim and dis: Y = 9.4993 - 1.5509X, **statistically significant**
- Crim and rad: Y = -2.28716 + 0.61791X, **statistically significant**
- Crim and tax: Y = -8.528369 + 0.029742X, **statistically significant**
- Crim and ptratio: Y = -17.6469 + 1.1520X, **statistically significant**
- Crim and black: Y = 16.553529 - 0.036280X, **statistically significant**
- Crim and lstat: Y = -3.33054 + 0.53880X, **statistically significant**
- Crim and medv: Y = 11.79654 - 0.36316X, **statistically significant**

We can see that every predictor is statistically significant except chas, which is the dummy variable.

2. 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 $H_0 : \beta_j = 0$?

```r
multiple_model = lm(crim ~ ., data = Boston)
summary(multiple_model)
```

```
##
## 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
```

The following predictors are statistically significant, and have a low p-value (less than 0.05): zn, indus, dis, rad, black, medv. Thus, for these predictors, we can reject their respective null hypothesis (i.e. $H_0 : \beta_j = 0$).

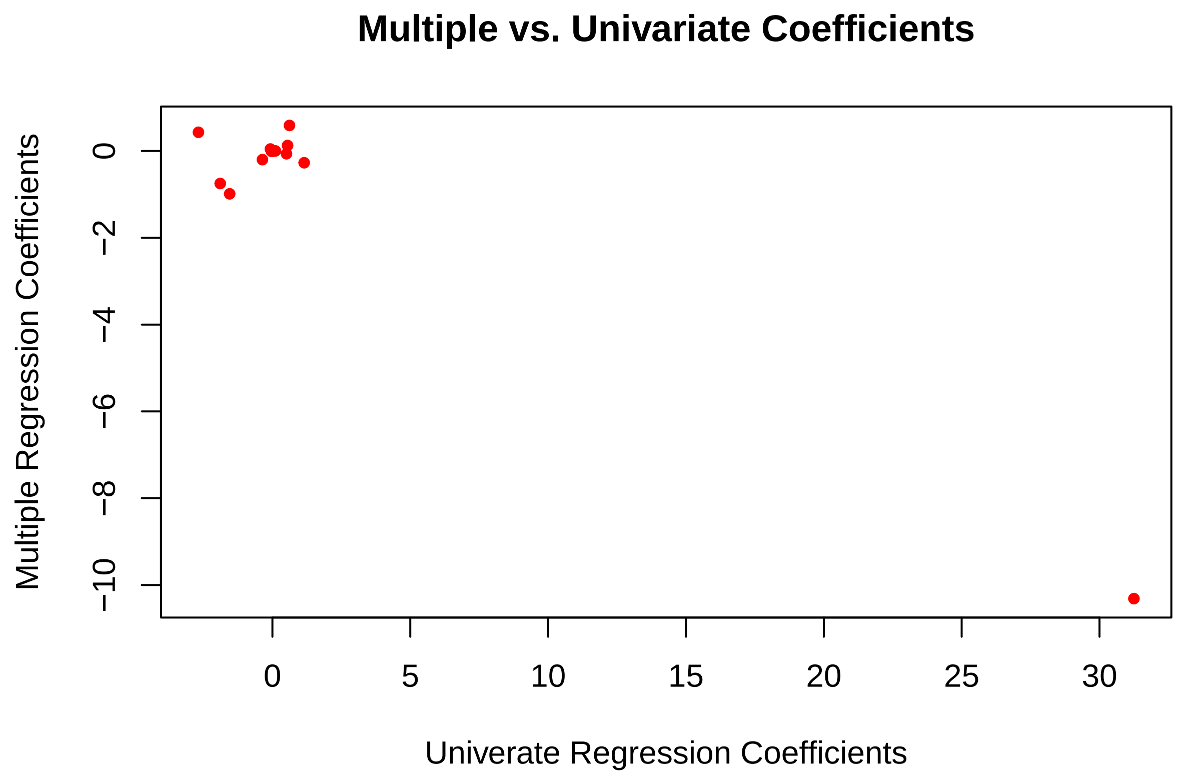
3. How do your results from 1. compare to your results from 2.? Create a plot displaying the univariate regression coefficients from 1. on the x-axis, and the multiple regression coefficients from 2. on the y-axis. That is, each predictor is displayed as a single point in the plot. Its coefficient in 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 results in part (a) are quite different from part (b), in the sense that, when we fit each parameter individually in (a), they were bound to be statistically significant, whereas in the multiple linear fit, only a few predictors were statistically significant (since they were compared against each other on the same fit).

```r
#For each fit, we obtain the beta 1 parameter and add it to our universate coefficient
univerate_regression = vector("numeric", 0)
univerate_regression = c(univerate_regression, fit1$coefficients[2])
univerate_regression = c(univerate_regression, fit2$coefficients[2])
univerate_regression = c(univerate_regression, fit3$coefficients[2])
univerate_regression = c(univerate_regression, fit4$coefficients[2])
univerate_regression = c(univerate_regression, fit5$coefficients[2])
univerate_regression = c(univerate_regression, fit6$coefficients[2])
univerate_regression = c(univerate_regression, fit7$coefficients[2])
univerate_regression = c(univerate_regression, fit8$coefficients[2])
univerate_regression = c(univerate_regression, fit9$coefficients[2])
univerate_regression = c(univerate_regression, fit10$coefficients[2])
univerate_regression = c(univerate_regression, fit11$coefficients[2])
univerate_regression = c(univerate_regression, fit12$coefficients[2])
univerate_regression = c(univerate_regression, fit13$coefficients[2])

multiple_regression = vector("numeric", 0)
multiple_regression = c(multiple_regression, multiple_model$coefficients)
multiple_regression = multiple_regression[-1] #Remove coefficient of crim
```

```r
plot(univerate_regression, multiple_regression, xlab = "Univerate Regression Coefficients", ylab = "Mul
```



4. 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 $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon$.

```r
cubic1 = lm(crim ~ zn + I(zn^2) +I(zn^3), data = Boston)
cubic2 = lm(crim ~ indus + I(indus^2) +I(indus^3), data = Boston)
cubic3 = lm(crim ~ chas + I(chas^2) +I(chas^3), data = Boston)
cubic4 = lm(crim ~ nox + I(nox^2) +I(nox^3), data = Boston)
cubic5 = lm(crim ~ rm + I(rm^2) +I(rm^3), data = Boston)
cubic6 = lm(crim ~ age + I(age) +I(age^3), data = Boston)
cubic7 = lm(crim ~ dis + I(dis^2) +I(dis^3), data = Boston)
cubic8 = lm(crim ~ rad + I(rad^2) +I(rad^3), data = Boston)
cubic9 = lm(crim ~ tax + I(tax^2) +I(tax^3), data = Boston)
cubic10 = lm(crim ~ ptratio + I(ptratio^2) +I(ptratio^3), data = Boston)
cubic11 = lm(crim ~ black + I(black^2) +I(black^3), data = Boston)
cubic12 = lm(crim ~ lstat + I(lstat) +I(lstat^3), data = Boston)
cubic13 = lm(crim ~ medv + I(medv^2) +I(medv), data = Boston)
```

```r
summary(cubic1)
summary(cubic2)
summary(cubic3)
summary(cubic4)
summary(cubic5)
summary(cubic6)
summary(cubic7)
summary(cubic8)
summary(cubic9)
summary(cubic10)
summary(cubic11)
summary(cubic12)
summary(cubic13)
```

From the above summary calls, we can check which models are non-linear by checking the associated p-values. Though they have been hidden for page limit purposes, the following models were seen to be statistically significant, i.e. with evidence of non-linear association: **indus, nox, age, dis, ptratio, medv**.