Assignment17

John Bute

2024-10-15

R Markdown

```
library(MASS)

Boston$chas <- factor(Boston$chas, labels = c("N","Y"))
attach(Boston)
```

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.

```
library(ggplot2)
predictors <- names(Boston)[!names(Boston) %in% c("X", "crim")]


models <- lapply(predictors, function(predictor) {
  lm(as.formula(paste("crim ~", predictor)), data = Boston)
})




model_summaries <- lapply(models, summary)

summary_table <- data.frame(
  Predictor = predictors,
  Coefficient = sapply(model_summaries, function(model) coef(model)[2,1]),
  R_Squared = sapply(model_summaries, function(model) model$r.squared),

  P_Value = sapply(model_summaries, function(model) coef(model)[2,4])

)

print(summary_table)
```

```
##    Predictor Coefficient   R_Squared      P_Value
## 1         zn -0.07393498 0.040187908 5.506472e-06
## 2      indus  0.50977633 0.165310070 1.450349e-21
## 3       chas -1.89277655 0.003123869 2.094345e-01
## 4        nox 31.24853120 0.177217182 3.751739e-23
## 5         rm -2.68405122 0.048069117 6.346703e-07
## 6        age  0.10778623 0.124421452 2.854869e-16
## 7        dis -1.55090168 0.144149375 8.519949e-19
## 8        rad  0.61791093 0.391256687 2.693844e-56
## 9        tax  0.02974225 0.339614243 2.357127e-47
## 10   ptratio  1.15198279 0.084068439 2.942922e-11
## 11     black -0.03627964 0.148274239 2.487274e-19
```

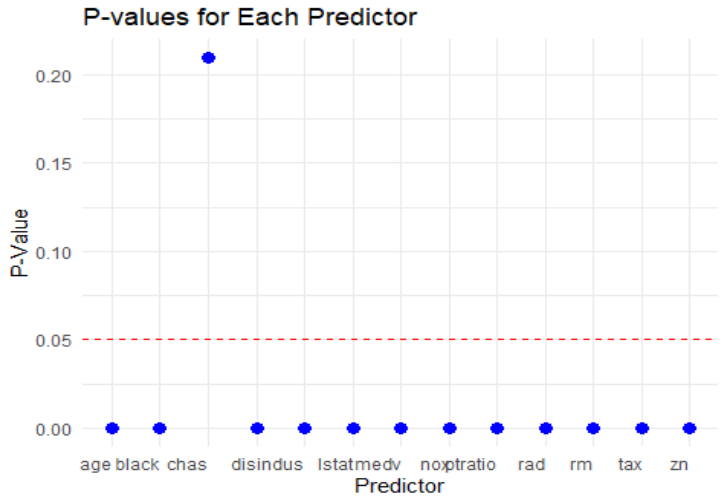
```
## 12     lstat  0.54880478 0.207590933 2.654277e-27
## 13      medv -0.36315992 0.150780469 1.173987e-19

significant_predictors <- subset(summary_table, P_Value < 0.05)

print(significant_predictors)

##    Predictor Coefficient  R_Squared      P_Value
## 1         zn -0.07393498 0.04018791 5.506472e-06
## 2      indus  0.50977633 0.16531007 1.450349e-21
## 4        nox 31.24853120 0.17721718 3.751739e-23
## 5         rm -2.68405122 0.04806912 6.346703e-07
## 6        age  0.10778623 0.12442145 2.854869e-16
## 7        dis -1.55090168 0.14414937 8.519949e-19
## 8        rad  0.61791093 0.39125669 2.693844e-56
## 9        tax  0.02974225 0.33961424 2.357127e-47
## 10   ptratio  1.15198279 0.08406844 2.942922e-11
## 11     black -0.03627964 0.14827424 2.487274e-19
## 12     lstat  0.54880478 0.20759093 2.654277e-27
## 13      medv -0.36315992 0.15078047 1.173987e-19

ggplot(summary_table, aes(x = Predictor, y = P_Value)) +
  geom_point(color = "blue", size = 3) +
  geom_hline(yintercept = 0.05, color = "red", linetype = "dashed") +   theme_minimal() +
  labs(title = "P-values for Each Predictor",
       x = "Predictor",
       y = "P-Value") +
  theme(axis.text.x = element_text(hjust = 1))
```



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.

To determine if there is a statistically significant association between the predictor and the response, we must look at the null hypothesis that the p-value of the predictor is less than 0.05, for the null hypothesis that its coefficient is 0. If it is higher than 0.05, then it is not statisically significant and we cannot reject the null hypothesis. The only p-value we can reject is chas, as the rest of the predictors have a significantly small p-value that is way below the 0.05 threshold.

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 H0 : βj = 0?

For a multiple linear regression model, we use the following format:

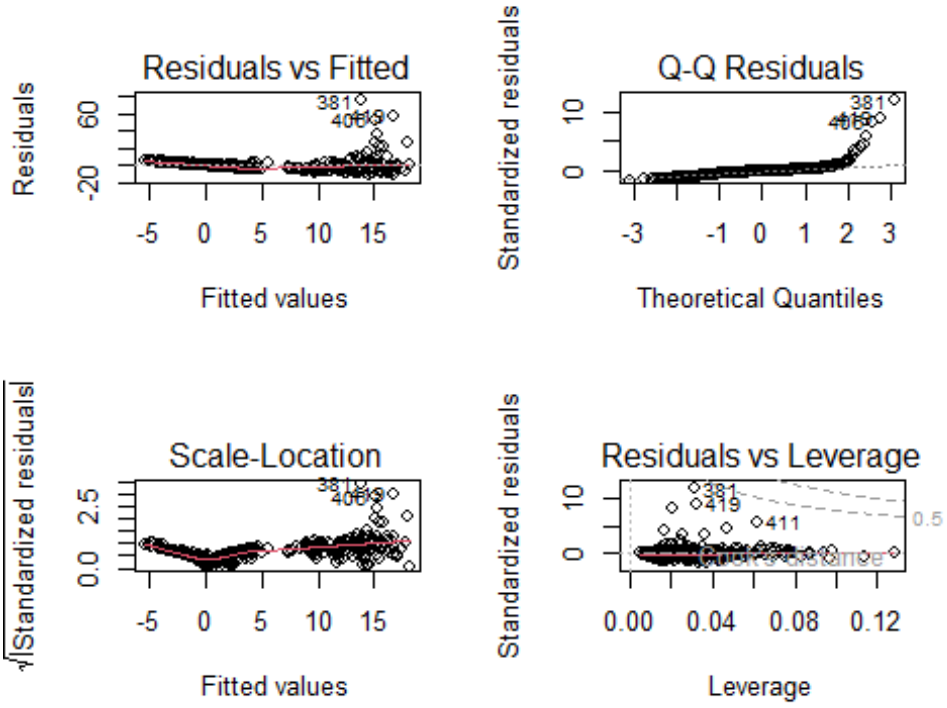
$$\hat{y} = \beta_0 + \sum_{i=1}^{13} \beta_i X_i$$

```
lm_multiple <- lm(crim ~ ., data = Boston[, !names(Boston) %in% c("X")])

summary(lm_multiple)

##
## Call:
## lm(formula = crim ~ ., data = Boston[, !names(Boston) %in% c("X")])
##
## 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
## chasY        -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

y=coefficients(lm_multiple)[2:14]
par(mfrow=c(2,2))
plot(lm_multiple)
```



We reject the null hypothesis for only zn, dis, black, rad, and medv, meaning that these predictors are the only ones with significant statistical association. Looking at the plots, we notice that there are not any high-leverage and outlier points, but there are a few outliers, but there are not enough to change the fit of the regression.

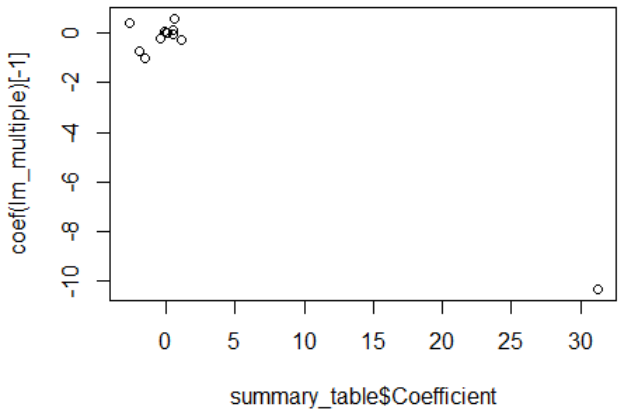
In terms of the residuals vs fitted plot, it seems that it is linear for the most part, except on the right, where it is scattered towards the top, which suggests that the data might not be linear or homoscedastic.

When it comes to the Q-Q plot, there is a massive deviation on the right, and deviate from the line, suggesting there are outliers in the upper tail.

In terms of the scale-location plot, it suggests heteroscedasticity as there is some fanning out at the end. In total, I do not think that we should continue with the assumption of linearity and homoscedacity, as there are multiple cases of deviation from the norm, heterodascity, and outliers.

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 regressionmodel is shown on the y−axis.

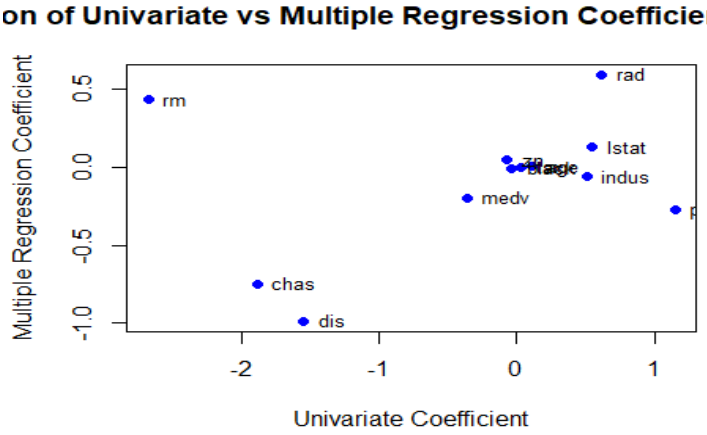
```
par(mfrow = c(1,1))
plot(summary_table$Coefficient, coef(lm_multiple)[-1])
```



Most coefficients for the predictors have been close to the same value, with the exception of nox, which we can see at the bottom right. There is small variations, but overall it seems that both models have predicter values close to each other. Let us further investigate

```
simple_coefficients <- summary_table$Coefficient[summary_table$Predictor != "nox"]
multiple_coefficients <- coef(lm_multiple)[-1][names(coef(lm_multiple)[-1]) != "nox"]
predictor_names <- summary_table$Predictor[summary_table$Predictor != "nox"]

par(mfrow = c(1,1))
plot(simple_coefficients, multiple_coefficients,
     xlab = "Univariate Coefficient",
     ylab = "Multiple Regression Coefficient",
     main = "Comparison of Univariate vs Multiple Regression Coefficients (without 'nox')",
     pch = 16, col = "blue")
text(simple_coefficients, multiple_coefficients, labels = predictor_names, pos = 4, cex = 0.8)
```



There is still some variation, with rm, pt, dis, chas, and rad.

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 = β0 + β1X1 + β2X2 + β3X3

For this question, we will fit a cubic regression model for each predictor y^=β0,i+β1,iXi +β2,iXi2+β3,iXi3, We will remove Chas from the predictor list as it only has 2 possible values. Thus fitting a polynomial of degree 3 will cause an error

```
predictors <- predictors[predictors != "chas"]
fit_models <- function(predictor) {
  formula <- as.formula(paste("crim ~ poly(", predictor, ", 3)"))
  model <- lm(formula, data=Boston)
  return(summary(model)$coefficients[3:4, 4])
}

print(predictors)

##  [1] "zn"      "indus"   "nox"     "rm"      "age"     "dis"     "rad"
##  [8] "tax"     "ptratio" "black"   "lstat"   "medv"
models_summary <- lapply(predictors, fit_models)

p_values <- do.call(rbind, models_summary)

colnames(p_values) <- c("Quadratic", "Cubic")

rownames(p_values) <- predictors

p_values

##            Quadratic        Cubic
## zn      4.420507e-03 2.295386e-01
## indus   1.086057e-03 1.196405e-12
## nox     7.736755e-05 6.961110e-16
## rm      1.508545e-03 5.085751e-01
## age     2.291156e-06 6.679915e-03
## dis     7.869767e-14 1.088832e-08
## rad     9.120558e-03 4.823138e-01
## tax     3.665348e-06 2.438507e-01
## ptratio 2.405468e-03 6.300514e-03
## black   4.566044e-01 5.436172e-01
## lstat   3.780418e-02 1.298906e-01
## medv    2.928577e-35 1.046510e-12

rowSums(p_values< 0.05)

##      zn   indus     nox      rm     age     dis     rad     tax ptratio   black
##       1       2       2       1       2       2       1       1       2       0
##   lstat    medv
##       1       2
```

Most models have evidence of non-linearity, as those predictors with only 1 term with a p-value less than 0.05 are in fact fitted quadratically, those with 2 are better fit for cubic regression. However, for black, it seems that we were right to assume linearity.