Project #2

Milica Cudina

2024-04-03

Problem #1 (55 points)

The `iris` data set is built-in in `R`. Start by studying the documentation of the data set, i.e., by entering `?iris` in the console. To familiarize yourselves with the architecture of an iris flower, go to:

US Forest Service

Your next step is exploratory data analysis.

(10 points) Which plot would you use to display pairwise associations between different measurements? How do you make sure that the different species are color-coded? Display the plot and write a few sentences about your conclusions.

Principal Component Analysis (PCA)

(20 **points)** Perform the PCA on the explanatory components of the above data, provide the report, and the relevant plots.

Principal Components Regression (PCR)

Your next task is to predict `Sepal.Length` from the other variables in the `iris` dataset.

(15 points) Run the PCR, provide an explanation for the output, and display the relevant plots (both validation and prediction).

(10 points) Split your dataset into training (4/5 of the data) and testing (1/5 of the data). Provide the mean squared error and an appropriate plot.

Problem #2 (20+5+10+10=45 points)

Solve **Problem 3.7.15** (page 128) from the textbook. \ \ 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.

```
Boston <- read.csv("Boston.csv")
```

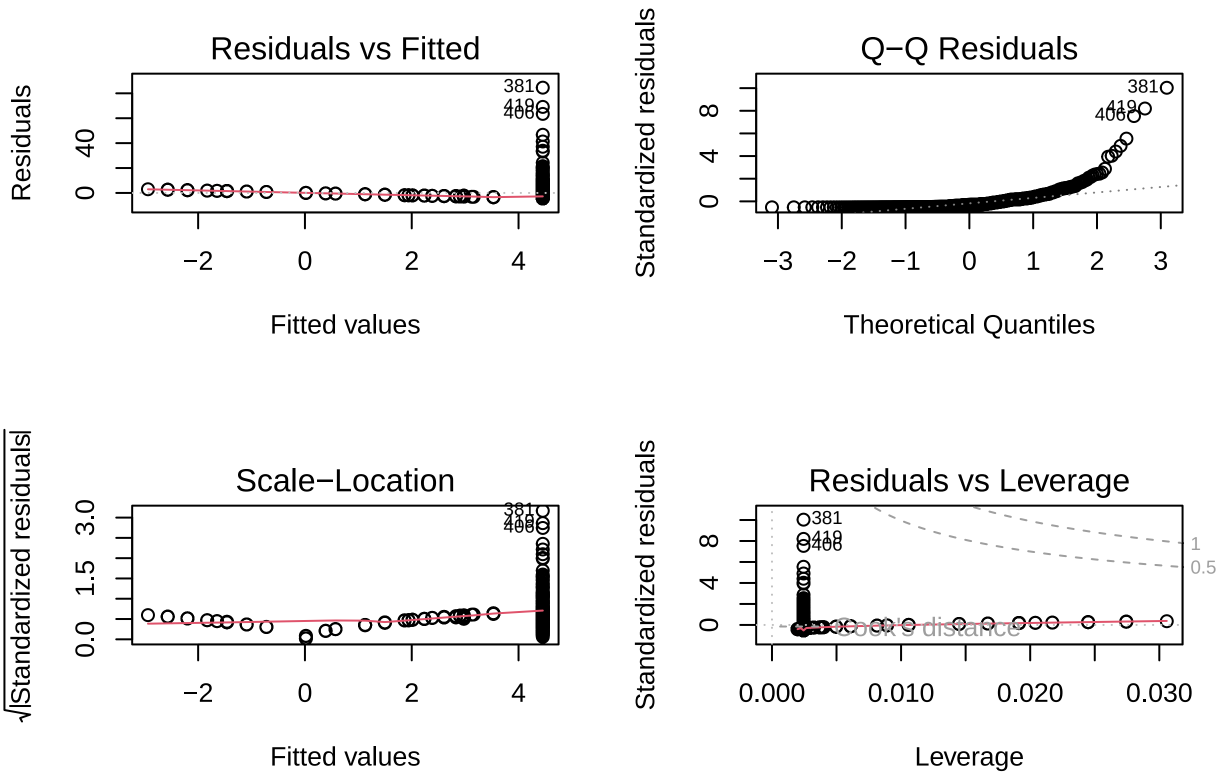
For each predictor, fit a simple linear regression model to predict the response. Describe your results. In which of the models is there a statistically significant association between the predictor and the response? Create some plots to back up your assertions.

Hint: The command `lapply` could be useful.

```r
predictors = colnames(Boston)[3:14]
simple_models <- list()
crime_zn_model <- lm(crim~zn, data = Boston)
summary(crime_zn_model )
##
## Call:
## lm(formula = crim ~ zn, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -4.429 -4.222 -2.620  1.250 84.523
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  4.45369    0.41722  10.675  < 2e-16 ***
## zn          -0.07393    0.01609  -4.594 5.51e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.435 on 504 degrees of freedom
## Multiple R-squared:  0.04019,    Adjusted R-squared:  0.03828
## F-statistic:  21.1 on 1 and 504 DF,  p-value: 5.506e-06
```

Crime and Zn We can notice that because we have a low p-value (5.506e-06 $<$ 0.05) and a F-statistic of 21.1 the probability of the results given the null hypothesis (no statistically significant association) is low. Therefore we can observe there is a statistically significant association between the predictor (zn) and response (crim)

```r
par(mfrow = c(2, 2))
plot(crime_zn_model)
```



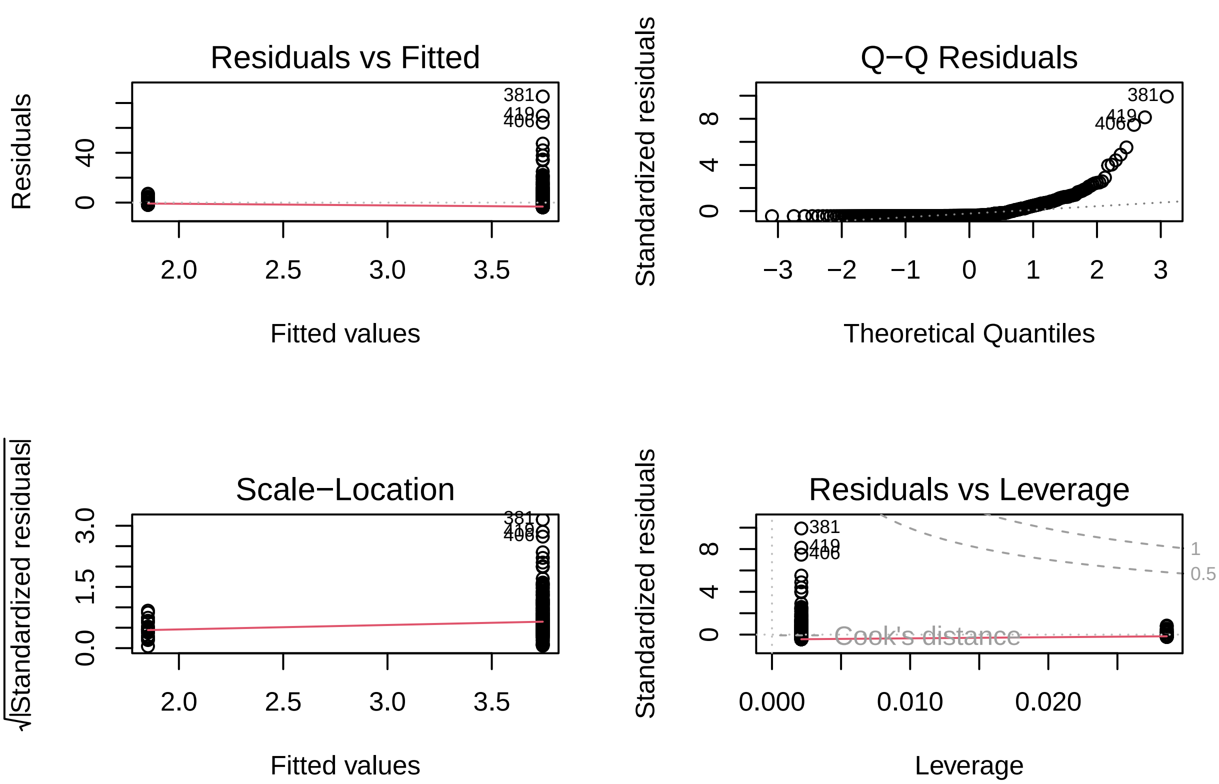

```
crime_indus_model <- lm(crim~indus, data = Boston)
summary(crime_indus_model)
##
## Call:
## lm(formula = crim ~ indus, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.972  -2.698  -0.736   0.712  81.813
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.06374    0.66723  -3.093  0.00209 **
## indus        0.50978    0.05102   9.991  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.866 on 504 degrees of freedom
## Multiple R-squared:  0.1653, Adjusted R-squared:  0.1637
## F-statistic: 99.82 on 1 and 504 DF,  p-value: < 2.2e-16
```

```
crime_chas_model <- lm(crim~chas, data = Boston)
summary(crime_chas_model)
##
## Call:
```

```
## lm(formula = crim ~ chas, data = Boston)
##
## 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 ***
## chas         -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
```

Crime and Chas We can notice that because we have a high p-value ($0.20294 < 0.05$) and a F-statistic of 1.579 the probability of the results given the null hypothesis (no statistically significant association) is not low. Therefore we can observe there is not a statistically significant association between the predictor (chas) and response (crim)

```
par(mfrow = c(2, 2))
plot(crime_chas_model)
```



```r
crime_nox_model <- lm(crim~nox, data = Boston)
summary(crime_nox_model)
##
## Call:
## lm(formula = crim ~ nox, data = Boston)
##
## 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
```

```r
crime_rm_model <- lm(crim~rm, data = Boston)
summary(crime_rm_model)
##
## Call:
## lm(formula = crim ~ rm, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.604 -3.952 -2.654  0.989 87.197
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   20.482      3.365   6.088 2.27e-09 ***
## rm            -2.684      0.532  -5.045 6.35e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.401 on 504 degrees of freedom
## Multiple R-squared:  0.04807,    Adjusted R-squared:  0.04618
## F-statistic: 25.45 on 1 and 504 DF,  p-value: 6.347e-07
```

```r
crime_age_model <- lm(crim~age, data = Boston)
summary(crime_age_model)
##
## Call:
## lm(formula = crim ~ age, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -6.789 -4.257 -1.230  1.527 82.849
##
## Coefficients:
```

```
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.77791    0.94398  -4.002 7.22e-05 ***
## age          0.10779    0.01274   8.463 2.85e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.057 on 504 degrees of freedom
## Multiple R-squared:  0.1244, Adjusted R-squared:  0.1227
## F-statistic: 71.62 on 1 and 504 DF,  p-value: 2.855e-16
```

```
crime_dis_model <- lm(crim~dis, data = Boston)
summary(crime_dis_model)
##
## Call:
## lm(formula = crim ~ dis, data = Boston)
##
## 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
```

```
crime_tax_model <- lm(crim~tax, data = Boston)
summary(crime_tax_model)
##
## Call:
## lm(formula = crim ~ tax, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -12.513  -2.738  -0.194   1.065  77.696
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -8.528369   0.815809  -10.45   <2e-16 ***
## tax          0.029742   0.001847   16.10   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.997 on 504 degrees of freedom
## Multiple R-squared:  0.3396, Adjusted R-squared:  0.3383
## F-statistic: 259.2 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
crime_ptratio_model <- lm(crim~ptratio, data = Boston)
summary(crime_ptratio_model)
##
## Call:
## lm(formula = crim ~ ptratio, data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -7.654 -3.985 -1.912  1.825 83.353
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -17.6469     3.1473  -5.607 3.40e-08 ***
## ptratio       1.1520     0.1694   6.801 2.94e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 8.24 on 504 degrees of freedom
## Multiple R-squared:  0.08407,    Adjusted R-squared:  0.08225
## F-statistic: 46.26 on 1 and 504 DF,  p-value: 2.943e-11
```

```r
crime_lstat_model <- lm(crim~lstat, data = Boston)
summary(crime_lstat_model)
##
## Call:
## lm(formula = crim ~ lstat, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -13.925  -2.822  -0.664   1.079  82.862
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.33054    0.69376  -4.801 2.09e-06 ***
## lstat        0.54880    0.04776  11.491  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.664 on 504 degrees of freedom
## Multiple R-squared:  0.2076, Adjusted R-squared:  0.206
## F-statistic:   132 on 1 and 504 DF,  p-value: < 2.2e-16
```

```r
crime_medv_model <- lm(crim~medv, data = Boston)
summary(crime_medv_model)
##
## Call:
## lm(formula = crim ~ medv, data = Boston)
##
## 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
```

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?

```
crime_model_all <- lm(crim ~ ., data = Boston)
summary(crime_model_all)
##
## Call:
## lm(formula = crim ~ ., data = Boston)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -8.403 -2.319 -0.363  1.006 73.805
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) 13.870138   7.087527   1.957 0.050915 .
## X           -0.001814   0.002837  -0.640 0.522787
## zn           0.046925   0.018897   2.483 0.013355 *
## indus       -0.058749   0.083688  -0.702 0.483010
## chas        -0.805138   1.184529  -0.680 0.497007
## nox         -9.829024   5.296814  -1.856 0.064101 .
## rm           0.656326   0.608967   1.078 0.281665
## age         -0.002719   0.018196  -0.149 0.881266
## dis         -1.027203   0.283603  -3.622 0.000323 ***
## rad          0.626037   0.090123   6.946 1.19e-11 ***
## tax         -0.003270   0.005235  -0.625 0.532531
## ptratio     -0.302240   0.186494  -1.621 0.105735
## lstat        0.136453   0.075856   1.799 0.072654 .
## medv        -0.221891   0.059929  -3.703 0.000238 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.464 on 492 degrees of freedom
## Multiple R-squared:  0.4498, Adjusted R-squared:  0.4353
## F-statistic: 30.94 on 13 and 492 DF,  p-value: < 2.2e-16
```

As we see above, zn,dis, rad, and medv have p values less than 0.05, and are therefore we can reject the null hypothesis for those predictors