STOR 565 Spring 2018 Homework 2

Due on 01/31/2018 in Class

YOUR NAME

Remark. This homework aims to help you go through the necessary preliminary from linear regression. Credits for **Theoretical Part** and **Computational Part** are in total 100 pt. For **Computational Part**, please complete your answer in the **RMarkdown** file and summit your printed PDF homework created by it.

Computational Part

1. (*35 pt*) Consider the dataset "Boston" in predicting the crime rate at Boston with associated covariates.

```
head(Boston)
```

```
##      crim zn indus chas   nox    rm  age    dis rad tax ptratio  black
## 1 0.00632 18  2.31    0 0.538 6.575 65.2 4.0900   1 296    15.3 396.90
## 2 0.02731  0  7.07    0 0.469 6.421 78.9 4.9671   2 242    17.8 396.90
## 3 0.02729  0  7.07    0 0.469 7.185 61.1 4.9671   2 242    17.8 392.83
## 4 0.03237  0  2.18    0 0.458 6.998 45.8 6.0622   3 222    18.7 394.63
## 5 0.06905  0  2.18    0 0.458 7.147 54.2 6.0622   3 222    18.7 396.90
## 6 0.02985  0  2.18    0 0.458 6.430 58.7 6.0622   3 222    18.7 394.12
##   lstat medv
## 1  4.98 24.0
## 2  9.14 21.6
## 3  4.03 34.7
## 4  2.94 33.4
## 5  5.33 36.2
## 6  5.21 28.7
```

Suppose you would like to predict the crime rate with explantory variables

- `medv` - Median value of owner-occupied homes
- `dis` - Weighted mean of distances to employement centers
- `indus` - Proportion of non-retail business acres

Run with the linear model

```
mod1 <- lm(crim ~ medv + dis + indus, data = Boston)
summary(mod1)
```

```
##
## Call:
## lm(formula = crim ~ medv + dis + indus, data = Boston)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -11.625  -3.345  -1.242   1.608  78.994
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 11.67738    2.12190   5.503 5.95e-08 ***
## medv        -0.26061    0.04204  -6.199 1.19e-09 ***
## dis         -0.96320    0.22758  -4.232 2.75e-05 ***
## indus        0.13145    0.07728   1.701   0.0896 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.519 on 502 degrees of freedom
## Multiple R-squared:  0.2404, Adjusted R-squared:  0.2358
## F-statistic: 52.95 on 3 and 502 DF,  p-value: < 2.2e-16
```

Answer the following questions.

(i) What do the following quantities that appear in the above output mean in the linear model? Provide a breif description.

- `t value` and `Pr(>|t|)` of `medv`

Answer: YOUR ANSWER.

- `Multiple R-squared`

Answer: YOUR ANSWER.

- `F-statistic`, `DF` and corresponding `p-value`

Answer: YOUR ANSWER.

(ii) Are the following sentences True of False? Briefly justify your answer.

- `indus` is not a significant predictor of crim, and we can drop this from the model.

Answer: YOUR ANSWER.

- `Multiple R-squared` is preferred to `Adjusted R-squared` as it takes into account all the variables.

Answer: YOUR ANSWER.

- `medv` has a negative effect on the response.

Answer: YOUR ANSWER.

- Our model residuals appear to be normally distributed.

Hint. You need to access to the model residuals in justifying the last sentence. The following commands might help.

```r
# Obtain the residuals
res1 <- residuals(mod1)

# Normal QQ-plot of residuals
plot(mod1, 2)

# Conduct a Normality test via Shapiro-Wilk and Kolmogorov-Smirnov test
shapiro.test(res1)
ks.test(res1, "pnorm")
```

Answer: YOUR ANSWER.

2. (*35 pt*, Textbook Exercises 3.10) This question should be answered using the `Carseats` data set.

```r
head(Carseats)
```

```
##   Sales CompPrice Income Advertising Population Price ShelveLoc Age
## 1  9.50       138     73          11        276   120       Bad  42
## 2 11.22       111     48          16        260    83      Good  65
## 3 10.06       113     35          10        269    80    Medium  59
## 4  7.40       117    100           4        466    97    Medium  55
## 5  4.15       141     64           3        340   128       Bad  38
## 6 10.81       124    113          13        501    72       Bad  78
##   Education Urban  US
## 1        17   Yes Yes
## 2        10   Yes Yes
## 3        12   Yes Yes
## 4        14   Yes Yes
## 5        13   Yes  No
## 6        16    No Yes
```

(a) Fit a multiple regression model to predict `Sales` using `Price`, `Urban`, and `US`.

Answer: YOUR ANSWER.

(b) Provide an interpretation of each coefficient in the model. Be careful—some of the variables in the model are qualitative!

Answer: YOUR ANSWER.

(c) Write out the model in equation form, being careful to handle the qualitative variables properly.

Answer: YOUR ANSWER.

(d) For which of the predictors can you reject the null hypothesis $H_0 : \beta_j = 0$?

Answer: YOUR ANSWER.

(e) On the basis of your response to the previous question, fit a smaller model that only uses the predictors for which there is evidence of association with the outcome.

Answer: YOUR ANSWER.

(f) How well do the models in (a) and (e) fit the data?

Answer: YOUR ANSWER.

(g) Using the model from (e), obtain 95% confidence intervals for the coefficient(s).

Answer: YOUR ANSWER.

(h) Using the leave-one-out cross-validation and 5-fold cross-validation techniques to compare the performance of models in (a) and (e). What can you tell from (f) and (h)?

Hint. Functions `update` (with option `subset`) and `predict`.

Answer: YOUR ANSWER.
