R Lab #2 (Ch. 3) – Solutions

1. Section 3.6.2 Simple Linear Regression

Fit a simple linear regression model to the Boston data with medv as the response variable and lstat as the predictor variable. What are the bounds for the 95% prediction interval of medv associated with a lstat value of 20? (Round your answers to 1 decimal place.)

- Lower bound = ______
- Upper Bound = ______

Solution: 3.3, 27.8

```
lm.fit <- lm(medv ~ lstat, data = Boston)
predict(lm.fit, data.frame(lstat = 20), interval = "prediction")
#       fit      lwr      upr
# 1 15.55285 3.316021 27.78969
```

2. Section 3.6.3 Multiple Linear Regression

Fit a multiple linear regression model to the Boston data with medv as the response variable and all the other variables except age as the predictor variables (this is model lm.fit1 in the Lab). Which predictor has the highest p-value for this model fit?

a) crim
b) zn
c) indus
d) chas
e) tax

Solution: C

```
lm.fit1 <- lm(medv ~. -age, data = Boston)
summary(lm.fit1)
which.max(summary(lm.fit1)$coefficients[, 4])
#indus
#    4
```

3. Section 3.6.3 Multiple Linear Regression

Fit a multiple linear regression model to the Boston data with medv as the response variable and all the other variables except age as the predictor variables (this is model lm.fit1 in the Lab). The functions rstudent(), hatvalues(), and cooks.distance() return the studentized residuals, leverages, and Cook's distances (a combined measure of outlyingness and leverage for which values above 0.5 suggest possible influential observations and values above 1 suggest probable influential observations). Use these functions to find the index numbers of the observations with the highest absolute studentized residual, leverage, and Cook's distance, together with the values of these highest absolute studentized residual, leverage, and Cook's distance. (The index numbers will be whole numbers and you should round the values of the statistics to 2 decimal places).

- Highest absolute studentized residual: index number = _____, value = _____
- Highest leverage: index number = _____, value = _____
- Highest Cook's distance: index number = _____, value = _____

Solution: 369, 5.89, 381, 0.31, 369, 0.15

```
rstudent(lm.fit1)[which.max(abs(rstudent(lm.fit1)))]
#    369
#5.88543
hatvalues(lm.fit1)[which.max(hatvalues(lm.fit1))]
#      381
#0.3055797
cooks.distance(lm.fit1)[which.max(cooks.distance(lm.fit1))]
#      369
#0.1485981
```

4. Section 3.6.4 Interaction Terms

Fit the multiple linear regression model to the Boston dataset with medv as the response variable and lstat, age, and the interaction term lstat x age as predictors (call this model 1). Also fit the model with lstat, black, and the interaction term lstat x black as predictors (call this model 2). Based on the measures of model fit, RSE and R^2, is it true or false that model 1 "fits better" than model 2?

- True or False?

Solution: False

```
lm.model1 <- lm(medv ~ lstat*age, data = Boston)
lm.model2 <- lm(medv ~ lstat*black, data = Boston)
c(summary(lm.model1)$sigma, summary(lm.model1)$r.squared)
# [1] 6.1485133 0.5557265
c(summary(lm.model2)$sigma, summary(lm.model2)$r.squared)
# [1] 6.109289 0.561377
```

5. Section 3.6.5 Non-linear Transformations of the Predictors

Fit the simple linear regression model to the Boston dataset with medv as the response variable and log(rm) as the predictor variable. What is the coefficient estimate for the slope? (Round your answer to 1 decimal place.)

Solution: 54.1

```
summary(lm(medv ~ log(rm), data = Boston))$coefficients[2, 1]
# [1] 54.05457
```

6. Section 3.6.6 Qualitative Predictors

Fit the multiple linear regression model to the Carseats dataset with Sales as the response variable and all the other variables as well as the interaction terms, Income x Advertising and Price x Age, as predictors (this is the model fit in the Lab). Then fit the model again, but this time add the argument:

```
contrasts = list(ShelveLoc=contr.treatment(c("Bad", "Good", "Medium"), base = 3))
```

This changes the dummy variables so that the 3rd level in the ShelveLoc factor ("Medium") is used as the baseline level. You should find that the model fit is identical for both models (other than the intercept estimate and the coefficient estimates for the dummy variables). What are the two dummy variable coefficient estimates? (Round your answers to 2 decimal places.)

- ShelveLocBad: ______
- ShelveLocGood: ______

Solution: -1.95, 2.90

```
lm.fit.new <- lm(Sales ~ . + Income:Advertising + Price:Age, data = Carseats,
              contrasts = list(ShelveLoc=contr.treatment(c("Bad", "Good", "Medium"), base = 3)))
summary(lm.fit.new)$coefficients[7:8, 1]
#ShelveLocBad ShelveLocGood
#   -1.953262      2.895414
```