MA679-Homework 1

Jinfei Xue

1/28/2019

3.1

Describe the null hypotheses to which the p-values given in Table 3.4 correspond. Explain what conclusions you can draw based on these p-values. Your explanation should be phrased in terms of sales, TV, radio, and newspaper, rather than in terms of the coefficients of the linear model.

The null hypotheses associated with table 3.4 are that advertising budgets of TV, radio or newspaper do not have an effect on sales. That is to say, the null hypothesis is that any of the coefficients of TV, radio or newspaper is equal to 0.

From the table 3.4, we can see that the corresponding p-values are highly significant for TV and radio but not significant for newspaper. Therefore, we can conclude that the newspaper advertising budget is not associated with sales.

3.2

Carefully explain the differences between the KNN classifier and KNN regression methods.

3.11

```
#generate a predictor x and a response y
set.seed (1)
x=rnorm (100)
y=2*x+rnorm (100)
```

(a) regression without intercept (y onto x)

```
r11_a=lm(y~x+0)
summary(r11_a)
```

```
##
## Call:
## lm(formula = y ~ x + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -1.9154 -0.6472 -0.1771  0.5056  2.3109
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## x   1.9939     0.1065   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.9586 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

From the summary above, we can see that the coefficient of x is 1.9939, which means if x increases by 1 unit, then the value of y will approximately increase by 1.9939 units.

the t-value for null hypothesis is 18.73 and the corresponding p-value is much smaller than 0.05. Therefore, we can reject the null hypothesis and indicate that x is associated with y.

R^2=0.7798 is large enough to show that a large proportion of the variability in the response has been explained by the regression.

(b) regression without intercept (x onto y)

```
r11_b <- lm(x ~ y + 0)
summary(r11_b)
```

```
##
## Call:
## lm(formula = x ~ y + 0)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8699 -0.2368  0.1030  0.2858  0.8938
##
## Coefficients:
##   Estimate Std. Error t value Pr(>|t|)
## y  0.39111    0.02089   18.73   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4246 on 99 degrees of freedom
## Multiple R-squared:  0.7798, Adjusted R-squared:  0.7776
## F-statistic: 350.7 on 1 and 99 DF,  p-value: < 2.2e-16
```

From the summary above, we can see that the coefficient of y is 0.39111, which means if y increases by 1 unit, then the value of x will approximately increase by 0.39111 units.

the t-value for null hypothesis is 18.73 (which is equal to the value in (a)) and the corresponding p-value is much smaller than 0.05. Therefore, we can reject the null hypothesis and indicate that y is associated with x.

R^2=0.7798 (which is equal to the value in (a)) is large enough to show that a large proportion of the variability in the response has been explained by the regression.

(c) What is the relationship between the results obtained in (a) and (b)?

We obtain the same value for the t statistics and consequently the same value for the corresponding p-value, and R-squared. Both results in (a) and (b) reflect the same line created in (a).

(d) Show the t-statistic formula algebraically and confirm it numerically in R;

```
n <- length(x)
t <- sqrt(n - 1)*(x %*% y)/sqrt(sum(x^2) * sum(y^2) - (x %*% y)^2)
as.numeric(t)
```

```
## [1] 18.72593
```

We can see that the t value above is exactly same as the t-statistic given in the summary of "r11_a".

(e) Using the results from (d), argue that the t-statistic for the regression of y onto x is the same as the t-statistic for the regression of x onto y.

From the formula in (d), if we replace x_i with y_i and replace y_i with x_i in the formula for the t-statistic, the result would be the same.