Doug Woodward

CS613 HW 4

|  |  |
| --- | --- |
| # | Answer |
| 1 | The null hypothesis in question is if b1 - bn or, rather, binterrcept, btv bradio and bnewspaper are all 0.The corresponding p values show us that the newspaper is an ineffective predictor of sales, while the tv and radio are both likely to be excellent predictors of sales. |
| 2 | A KNN classifier and a KNN regressor will both begin by identifying the nearest K to the target point. Then, the classifier will make a majority rules decision to determine what the target value should be whereas the regressor will be take the average of the nearest K points for the target value. |
| 3 | Y = 50 + gpa\*20 + IQ\*0.07 + gender\*35 + gpaIq \* 0.01 + gpaGender \* -10   1. 3 iis correct. The b3 coefficient of 35 shows that males earn more given that their gpa is high enough to overcome the 35 coefficient towards females. 2. Y = 4.0\*20 + 110\*0.07 + 35 + (4\*110)\*0.01 + (4\*1)\*-10 = $87000 3. True, because the term is near 0, the effect of the interaction has little bearing on the outcome |
| 4 | 1. It would be expected that the RSS for the linear regression would be smaller, as that would be a better fit to the actual model. 2. In the case of test, it may be difficult to tell as n = 100 is not a particularly large sample size, perhaps the larger population is closer to a cubic model. 3. It would depend on the actual data, say that the mid point between a cubic and linear model is a squared model, then if the model is between squared and cubic, we expect the rss to be larger for the linear model and vice versa. 4. Same answer as b + d. |
| 5 | Using the first equation and then substituting in the value given in the second, we do  Yhati = xi Bhat , Bhat = Sum(xiyi)/Sum(xi^2)  Yhati = xi\*(Sum(xiprime\*yiprime)/Sum(xi^2)  Yhati = Sum(xiprime\*xi/Sum(xi^2))yiprime  -> aiprime = (xiprime\*xi)/(sum(xi^2)) |
| 6 | From 3.5 -> y = b0 + bix  With 3.4 -> b0 = yhat - b1\*xhat  If xhat, yhat is on the line, then  0 = b0 + b1\*xhat - yhat  Sub b0 for above from 3.4  0 = (yhat - b1\*xhat) + b1\*xhat - yhat  0 = 0 |
| 7 | 3.7.jpg |
| 8 | * 1. Yes - large F statistic + small p value of that f statistic mean there is a relationship   2. This is fairly strong relationship as F is large   3. The relationship is negative   4. hp = 98 -> mpg = 24.47,   Fit lwr upr  1 24.46708 23.97308 24.96108   fit lwr upr  1 24.46708 14.8094 34.12476  b)    c) Residuals indicate a possibility of non linearity |
| 9 | 2. mpg cylinders displacement horsepower weight acceleration year origin   mpg 1.0000000 -0.7776175 -0.8051269 -0.7784268 -0.8322442 0.4233285 0.5805410 0.5652088  cylinders -0.7776175 1.0000000 0.9508233 0.8429834 0.8975273 -0.5046834 -0.3456474 -0.5689316  displacement -0.8051269 0.9508233 1.0000000 0.8972570 0.9329944 -0.5438005 -0.3698552 -0.6145351  horsepower -0.7784268 0.8429834 0.8972570 1.0000000 0.8645377 -0.6891955 -0.4163615 -0.4551715  weight -0.8322442 0.8975273 0.9329944 0.8645377 1.0000000 -0.4168392 -0.3091199 -0.5850054  acceleration 0.4233285 -0.5046834 -0.5438005 -0.6891955 -0.4168392 1.0000000 0.2903161 0.2127458  year 0.5805410 -0.3456474 -0.3698552 -0.4163615 -0.3091199 0.2903161 1.0000000 0.1815277  origin 0.5652088 -0.5689316 -0.6145351 -0.4551715 -0.5850054 0.2127458 0.1815277 1.0000000  c)  i) Yes, the F statistic is fairly high with a small p value.  ii) Because weight, year, and origin have the smallest p values, we can say these are the most influential factors.  iii) The coefficents of 0.75 indicate that the mpg goes up by about 0.75 each year.  d)  Our residuals plot has a noticeable curve, this indicates a poor fit. Additionally, point 14 seems to have high leverage, and there are multiple possible outliers as shown in the QQ plot.  e) Adding cylinder\*displacement and displacement\*weight terms(take from checking the earlier correlation matrix), we find that the displacement\*weight term is significant, while the other is not.  f) Adding a horsepower^2 term to the model leads to a worse fit model as the anova functions shows F stat of 89 with a near zero p value in favor of the original model.  Transforming acceleration to log(acceleration) gives a slightly better model though. |
|  |  |

8 R code

library(ISLR)

attach(Auto)

lm.fit = lm(mpg~horsepower)

summary(lm.fit)

coef(lm.fit)

predict(lm.fit,data.frame(horsepower=c(98)), interval="confidence")

predict(lm.fit,data.frame(horsepower=c(98)), interval="prediction")

plot(horsepower, mpg)

abline(lm.fit)

par(mfrow=c(2,2))

plot(lm.fit)

9 R Code

library(ISLR)

pairs(Auto)

attach(Auto)

subAuto = subset(Auto, select=-name)

detach(Auto)

cor(subAuto)

attach(subAuto)

lm.fit = lm(mpg~.,data=subAuto)

summary(lm.fit)

coefficients(lm.fit)["year"]

par(mfrow=c(2,2))

plot(lm.fit)

lm.fitI = lm(mpg~cylinders\*displacement+displacement\*weight)

summary(lm.fitI)

lm.fit2 = lm(mpg~cylinders+displacement+weight+acceleration+year+origin+horsepower+I(horsepower^2))

summary(lm.fit2)

plot(lm.fit2)

anova(lm.fit2, lm.fit)

lm.fit3 = lm(mpg~cylinders+displacement+weight+acceleration+log(acceleration)+year+origin+horsepower)

summary(lm.fit3)

plot(lm.fit3)

anova(lm.fit, lm.fit3)