FIT3152 Data analytics. Tutorial 04: Regression

Pre-tutorial Activity

The dataset “mpg" from the ggplot2 package contains fuel economy data for popular car models released between 1999 and 2008. City and highway fuel consumption in miles per gallon are given under the two variables “cty” and “hwy”. Nine other variables provide insight on the type and features of each car.

1. Fit separate linear regression models to predict city and highway fuel consumption based on the type and features of each car.

```
#Fit linear model to predict city fuel consumption
> cty_fit = lm(cty ~ manufacturer + model + displ + year + cyl + trans +
drv + fl + class, data = mpg)

#Fit linear model to predict highway fuel consumption
> hwy_fit = lm(hwy ~ manufacturer + model + displ + year + cyl + trans +
drv + fl + class, data = mpg)
```

2. Which three variables are the most significant predictors in each model? Are they the same in each case?

```
#Check hypothesis test output for significance in predictors using summary
for each of the models
>summary(cty_fit)
# model, year, cyl are the most significant predictors
>summary(hwy_fit)
#model, year and cyl remain equally important. Comparatively manufacturer
and fuel type also seems to be more significant in relation to hwy fuel
efficiency.
```

3. How good are the two models comparatively?

```
#Check R squared value, residual distributions and overall significance of
regression using summary for each of the models

>summary(cty_fit)
#Approximately symmetrical distribution of residuals with 93% coefficient
of determination and very low p value for significance in the overall
regression model.

>summary(hwy_fit)
#Approximately symmetrical distribution of residuals with 95% coefficient
of determination and very low p value for significance in the overall
regression model.
```

Tutorial Activities

4. The ‘diamonds’ data set comes packaged with ggplot2 and contains data about the price of diamonds as well as information on size as well as the 4 Cs affecting diamond price: carat (size), cut, colour and clarity.

(a) Taking a random sample using the code below, create a subset of the diamonds data set: ‘dsmall’ to use in the following analysis.

```
install.packages("ggplot2")
library(ggplot2)
set.seed(9999) # Random seed to make subset reproducible
dsmall <- diamonds[sample(nrow(diamonds), 1000), ] # sample of 1000 rows
```

(b) Using the data ‘dsmall’ calculate the regression of ln(price) on ln(carat) and each of the remaining categories (clarity, color and cut) separately. Which of, clarity, color or cut has the greatest effect on price? Which has the least? Justify your answer using regression output.

```
# See the structure of data
> str(dsmall)
Classes ‘tbl_df’, ‘tbl’ and 'data.frame': 1000 obs. of  10 variables:
 $ carat  : num  0.3 1.02 0.35 0.72 0.59 0.75 1.04 1.05 0.4 2.55 ...
$ cut    : Ord.factor w/ 5 levels "Fair"<"Good"<..: 4 4 4 3 5 4 3 2 5 ...
$ color  : Ord.factor w/ 7 levels "D"<"E"<"F"<"G"<..: 6 5 6 7 2 2 1 1 …
 $ clarity: Ord.factor w/ 8 levels "I1"<"SI2"<"SI1"<..: 3 5 5 5 6 2 2  ...
$ depth  : num  60.9 58.5 61.1 62.7 61.8 61.9 63.1 60 62.4 63.7 ...
$ table  : num  59 57 60 58 57 55 58 63 53 59 ...
$ price  : int  506 5569 552 1988 3026 2214 3780 4560 917 14775 ...
$ x      : num  4.36 6.65 4.52 5.67 5.35 5.89 6.39 6.59 4.72 8.66 ...
$ y      : num  4.34 6.61 4.58 5.72 5.4 5.84 6.33 6.64 4.76 8.57 ...
$ z      : num  2.65 3.88 2.78 3.57 3.32 3.63 4.01 3.97 2.96 5.49 ...

# Attach the data to call columns directly by name
attach(dsmall)

# Fit the linear model
fit <-lm(log(price) ~ log(carat) + color+ cut + clarity)
summary(fit)

Call:
lm(formula = log(price) ~ log(carat) + color + cut + clarity)

Residuals:
  Min      1Q  Median      3Q     Max
-0.3777 -0.0863 -0.0028  0.0825  0.4877

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.93521    0.04399  180.40  < 2e-16 ***
log(carat)   1.88238    0.00829  227.13  < 2e-16 ***
color2      -0.06630    0.01478   -4.49  8.1e-06 ***
color3      -0.11137    0.01499   -7.43  2.4e-13 ***
color4      -0.16237    0.01485  -10.93  < 2e-16 ***
color5      -0.24929    0.01622  -15.37  < 2e-16 ***
color6      -0.38923    0.01829  -21.29  < 2e-16 ***
color7      -0.52599    0.02095  -25.10  < 2e-16 ***
cut2         0.05223    0.02665    1.96     0.05 .
cut3         0.10750    0.02490    4.32  1.7e-05 ***
cut4         0.12407    0.02453    5.06  5.1e-07 ***
cut5         0.14464    0.02427    5.96  3.5e-09 ***
```

```
clarity2      0.36649    0.03760     9.75   < 2e-16 ***
clarity3      0.54418    0.03743    14.54   < 2e-16 ***
clarity4      0.68431    0.03758    18.21   < 2e-16 ***
clarity5      0.74038    0.03817    19.40   < 2e-16 ***
clarity6      0.88221    0.03938    22.40   < 2e-16 ***
clarity7      0.97215    0.04001    24.30   < 2e-16 ***
clarity8      1.04679    0.04239    24.70   < 2e-16 ***
—
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.13 on 981 degrees of freedom
Multiple R-squared:  0.983,   Adjusted R-squared:  0.983
F-statistic: 3.2e+03 on 18 and 981 DF,  p-value: <2e-16

# Creating contrast matrices for categorical factors
contrasts(clarity) = contr.treatment(8)
contrasts(color) = contr.treatment(7)
contrasts(cut) = contr.treatment(5)

# Print the contrast matrices to see which variable in the model
contrasts(color)
contrasts(cut)
contrasts(clarity)

> contrasts(colo .... [TRUNCATED]
  2 3 4 5 6 7
D 0 0 0 0 0 0
E 1 0 0 0 0 0
F 0 1 0 0 0 0
G 0 0 1 0 0 0
H 0 0 0 1 0 0
I 0 0 0 0 1 0
J 0 0 0 0 0 1

> contrasts(cut)
          2 3 4 5
Fair      0 0 0 0
Good      1 0 0 0
Very Good 0 1 0 0
Premium   0 0 1 0
Ideal     0 0 0 1

> contrasts(clarity)
     2 3 4 5 6 7 8
I1   0 0 0 0 0 0 0
SI2  1 0 0 0 0 0 0
SI1  0 1 0 0 0 0 0
VS2  0 0 1 0 0 0 0
VS1  0 0 0 1 0 0 0
VVS2 0 0 0 0 1 0 0
VVS1 0 0 0 0 0 1 0
IF   0 0 0 0 0 0 1

# We can see from the model summary that clarity is the most significant
variable as each category has a very low p-value. Second important one
is color then cut.

# As you can see level of cut-color-clarity are represented as a separate
# predictor in the model.
# Each category is a separate binary variable (either 0 or 1).
```

5. The file “body.dat.csv” contains data from a study on the relationship between body dimensions. The study measured 500+ active individuals.

The data was obtained from http://www.amstat.org/publications/jse/jse_data_archive.htm
A related article is http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html

(a) Test the hypothesis that men are taller than women on average. Assume a significance of 5%

(b) Test the hypothesis that men are heavier than women on average. Assume a significance of 1%

(c) BMI is calculated as $\frac{weight\ (kg)}{(height\ (m))^2}$. Test the hypothesis that men have a higher BMI than women on average

```
#(a)
# Load & attach the dataset
body.dat <-read.csv("body.dat.csv")
attach(body.dat)

# Apply 2 sample hypothesis test
t.test(Height[Gender == "Male"], Height[Gender == "Female"], "greater",
      conf.level = 0.95)

      Welch Two Sample t-test

data:  Height[Gender == "Male"] and Height[Gender == "Female"]
t = 21, df = 495, p-value <2e-16
alternative hypothesis: true difference in means is greater than 0
95 percent confidence interval:
 11.9  Inf
sample estimates:
mean of x mean of y
      178       165

# p-value is smaller than 0.05 therefore we can reject the null hypothesis
# that is, there is no difference between the heights of men and women
# and conclude that men are taller than women.

#(b)
# Apply 2 sample hypothesis test
t.test(Weight[Gender == "Male"], Weight[Gender == "Female"], "greater",
      conf.level = 0.99)

      Welch Two Sample t-test

data:  Weight[Gender == "Male"] and Weight[Gender == "Female"]
t = 20, df = 495, p-value <2e-16
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
       15.5  Inf
      sample estimates:
      mean of x mean of y
           78.1      60.6

# p-value is smaller than 0.01 therefore we can reject the null hypothesis
# that is, there is no difference between the weights of men and women
# and conclude that men are heavier than women.
```

```
#(c)
# Create a BMI column then re-attach the dataframe
body.dat$BMI <- body.dat$Weight / (body.dat$Height*body.dat$Height)
attach(body.dat)

# Apply 2 sample hypothesis test
t.test(BMI[Gender == "Male"], BMI[Gender == "Female"], "greater",
      conf.level = 0.99)

      Welch Two Sample t-test

data:  BMI[Gender == "Male"] and BMI[Gender == "Female"]
t = 9, df = 502, p-value <2e-16
alternative hypothesis: true difference in means is greater than 0
99 percent confidence interval:
 0.00018     Inf
sample estimates:
mean of x mean of y
  0.00247   0.00223

# p-value is smaller than 0.01 therefore we can reject the null hypothesis
# that is, there is no difference between the BMI of men and women
# and conclude that men have a higher BMI than women.
```

(d) Calculate the regression of Height on the other body measurements for men and women separately. Which measurements are the most significant predictors of height for each gender?

```
#(d)
body.dat <-read.csv("body.dat.csv")
# Create a dataframe for males  only
body.male <-body.dat[which(body.dat$Gender == "Male"),]

# Remove the gender column as all are male.
body.male$Gender <-NULL

# Create a linear regression model to check significant predictors in the
      model summary.
male.fit <-lm(Height ~ ., data = body.male)

summary(male.fit)

Call:
lm(formula = Height ~ ., data = body.male)

Residuals:
    Min       1Q  Median      3Q     Max
-11.713  -2.841  -0.102   2.461  12.668

Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)   200.08048   13.48850   14.83  < 2e-16 ***
ShoulderWidth   0.39604    0.17002    2.33   0.0207 *
Pelvis          0.21254    0.20076    1.06   0.2909
Hips            0.71103    0.28823    2.47   0.0144 *
ChestDepth      0.14811    0.21383    0.69   0.4893
ChestDiam      -0.19514    0.23711   -0.82   0.4114
ElbowDiam       0.73736    0.49910    1.48   0.1410
WristDiam       0.24705    0.62822    0.39   0.6945
KneeDiam       -1.08693    0.37765   -2.88   0.0044 **
AnkleDiam       0.60736    0.43226    1.41   0.1614
```

```
ShoulderGirth  -0.11466    0.08964   -1.28   0.2022
Chest          -0.10679    0.10375   -1.03   0.3044
Waist          -0.48646    0.09088   -5.35  2.1e-07 ***
Abdomen        -0.14433    0.08288   -1.74   0.0830 .
HipGirth       -0.21055    0.13411   -1.57   0.1178
ThighGirth     -0.38345    0.15184   -2.53   0.0123 *
Bicep          -0.28485    0.24129   -1.18   0.2390
Forearm        -0.77910    0.36615   -2.13   0.0344 *
KneeGirth       0.32502    0.22134    1.47   0.1434
CalfGirth      -0.79929    0.18798   -4.25  3.1e-05 ***
AnkleGirth      0.27659    0.27090    1.02   0.3084
WristGirth      0.38288    0.54884    0.70   0.4861
Age             0.00517    0.03715    0.14   0.8895
Weight          1.15596    0.09818   11.77  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.1 on 223 degrees of freedom
Multiple R-squared:  0.698,   Adjusted R-squared:  0.667
F-statistic: 22.5 on 23 and 223 DF,  p-value: <2e-16

# For males, we can see that Weight, Waist and CalfGirth are the most
# significant predictors with lowest p-values.

# Create a dataframe for females only
body.female <-body.dat[which(body.dat$Gender == "Female"),]

# Remove the gender column as all are Female.
body.female$Gender <-NULL

# Create a linear regression model to check significant predictors in the
      model summary
female.fit <-lm(Height ~ ., data = body.female)

summary(female.fit)

Call:
lm(formula = Height ~ ., data = body.female)

Residuals:
    Min      1Q  Median      3Q     Max
-14.197  -2.390  -0.133   2.664  12.775

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept)   191.0434    12.8550   14.86   <2e-16 ***
ShoulderWidth   0.6941     0.2221    3.12    0.002 **
Pelvis          0.3990     0.1886    2.12    0.035 *
Hips           -0.0520     0.2542   -0.20    0.838
ChestDepth     -0.2912     0.2026   -1.44    0.152
ChestDiam      -0.0552     0.2516   -0.22    0.827
ElbowDiam       0.6122     0.5641    1.09    0.279
WristDiam      -0.1208     0.6985   -0.17    0.863
KneeDiam       -1.1241     0.4533   -2.48    0.014 *
AnkleDiam       0.7992     0.5092    1.57    0.118
ShoulderGirth   0.0448     0.0928    0.48    0.629
Chest          -0.2965     0.1311   -2.26    0.025 *
Waist          -0.5936     0.1000   -5.93    1e-08 ***
Abdomen         0.0839     0.0711    1.18    0.239
HipGirth       -0.1641     0.1467   -1.12    0.265
ThighGirth     -0.3553     0.1682   -2.11    0.036 *
Bicep          -0.3923     0.2609   -1.50    0.134
Forearm        -0.9869     0.4384   -2.25    0.025 *
```

```
KneeGirth        0.1456     0.2311    0.63    0.529
CalfGirth       -0.4669     0.2193   -2.13    0.034 *
AnkleGirth      -0.1853     0.3136   -0.59    0.555
WristGirth       1.0100     0.6780    1.49    0.138
Age             -0.0252     0.0383   -0.66    0.512
Weight           1.2516     0.1266    9.88   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.2 on 236 degrees of freedom
Multiple R-squared:  0.62,    Adjusted R-squared:  0.582
F-statistic: 16.7 on 23 and 236 DF,  p-value: <2e-16

# For females, we can see that Weight, Waist  are the most significant
# predictors with lowest p-values. Waist is an important predictor for
# female as well, although not as significant as in males.
```

6. The data file "Dunnhumby1-20.csv" is a cut down and modified set of test data from the Kaggle competition to predict when consumers would next visit a Dunnhumby supermarket and how much they would spend. See: http://www.kaggle.com/c/dunnhumbychallenge for more information. The current modified data set contains the customer ID, Date of visit, Days since last visit (Delta), and Spend for 20 customers from the test set.

   Calculate the regression of Spend vs Delta for each customer and summarize the results in a data frame similar to that below. *Hint: try using "by" function or "plyr" package and dlply function.*

CustomerID	RegIntercept	RegSlope

```
# using plyr - for the iris data
library(plyr)
models <- dlply(iris, "Species", function(df) lm(Sepal.Length ~
      Sepal.Width, data = df))
as.data.frame(ldply(models, coef))

DH <- read.csv("Dunhumby1-20.csv")
options(digits = 3)
models <- dlply(DH, "customer_id", function(df) lm(visit_spend ~
      visit_delta, data = df))
options(digits = 3)
as.data.frame(ldply(models, coef))
```

7. Using the data from the UCI Machine Learning Repository comment on the factors affecting red wine quality. Data site is: http://archive.ics.uci.edu/ml/datasets/Wine+Quality The file name is: winequality-red.csv.

```
rwine <- read.csv("winequality-red.csv")
attach(rwine)
wine.model <- lm(quality ~ ., data = rwine)
summary(wine.model)

Call:
lm(formula = quality ~ ., data = rwine)
```

```
Residuals:
   Min     1Q Median     3Q    Max
-2.689 -0.366 -0.047  0.452  2.025

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)           2.20e+01   2.12e+01    1.04    0.300
fixed.acidity         2.50e-02   2.59e-02    0.96    0.336
volatile.acidity     -1.08e+00   1.21e-01   -8.95  < 2e-16 ***
citric.acid          -1.83e-01   1.47e-01   -1.24    0.215
residual.sugar        1.63e-02   1.50e-02    1.09    0.276
chlorides            -1.87e+00   4.19e-01   -4.47  8.4e-06 ***
free.sulfur.dioxide   4.36e-03   2.17e-03    2.01    0.045 *
total.sulfur.dioxide -3.27e-03   7.29e-04   -4.48  8.0e-06 ***
density              -1.79e+01   2.16e+01   -0.83    0.409
pH                   -4.14e-01   1.92e-01   -2.16    0.031 *
sulphates             9.16e-01   1.14e-01    8.01  2.1e-15 ***
alcohol               2.76e-01   2.65e-02   10.43  < 2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.648 on 1587 degrees of freedom
Multiple R-squared:  0.361,   Adjusted R-squared:  0.356
F-statistic: 81.3 on 11 and 1587 DF,  p-value: <2e-16

# you can see that alcohol, acidity and sulphates are the greatest
# determinants of quality.
```

8. Install the "ISLR" library. Using the "Carseats" data, calculate the regression equation predicting Sales (child car seat sales) as a function of the input variables. Which variables are significant predictors?

```
install.packages("ISLR")
library(ISLR)

str(Carseats)

'data.frame':     400 obs. of  11 variables:
 $ Sales      : num  9.5 11.22 10.06 7.4 4.15 ...
 $ CompPrice  : num  138 111 113 117 141 124 115 136 132 132 ...
 $ Income     : num  73 48 35 100 64 113 105 81 110 113 ...
 $ Advertising: num  11 16 10 4 3 13 0 15 0 0 ...
 $ Population : num  276 260 269 466 340 501 45 425 108 131 ...
 $ Price      : num  120 83 80 97 128 72 108 120 124 124 ...
 $ ShelveLoc  : Factor w/ 3 levels "Bad","Good","Medium": 1 2 3 3 1 1 ...
 $ Age        : num  42 65 59 55 38 78 71 67 76 76 ...
 $ Education  : num  17 10 12 14 13 16 15 10 10 17 ...
 $ Urban      : Factor w/ 2 levels "No","Yes": 2 2 2 2 2 1 2 2 1 1 ...
 $ US         : Factor w/ 2 levels "No","Yes": 2 2 2 2 1 2 1 2 1 2 ...

attach(Carseats)
contrasts(ShelveLoc) = contr.treatment(3)
contrasts(Urban) = contr.treatment(2)

contrasts(US) = contr.treatment(2)
CS.model = lm(Sales ~ ., data = Carseats)
summary(CS.model)

Call:
lm(formula = Sales ~ ., data = Carseats)

Residuals:
```

```
   Min      1Q Median      3Q     Max
-2.869 -0.691  0.021  0.664  3.411

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)      5.660623   0.603449    9.38  < 2e-16 ***
CompPrice        0.092815   0.004148   22.38  < 2e-16 ***
Income           0.015803   0.001845    8.56  2.6e-16 ***
Advertising      0.123095   0.011124   11.07  < 2e-16 ***
Population       0.000208   0.000370    0.56     0.58
Price           -0.095358   0.002671  -35.70  < 2e-16 ***
ShelveLocGood    4.850183   0.153110   31.68  < 2e-16 ***
ShelveLocMedium  1.956715   0.126106   15.52  < 2e-16 ***
Age             -0.046045   0.003182  -14.47  < 2e-16 ***
Education       -0.021102   0.019720   -1.07     0.29
UrbanYes         0.122886   0.112976    1.09     0.28
USYes           -0.184093   0.149842   -1.23     0.22
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.02 on 388 degrees of freedom
Multiple R-squared:  0.873,   Adjusted R-squared:  0.87
F-statistic:  243 on 11 and 388 DF,  p-value: <2e-16

# Price, shelf location, competitor price, age, income and advertising
      have a major effect.
```

9. The text, G. James et al., An Introduction to Statistical Learning: with Applications in R (ISLR) uses the "Advertising" data set to illustrate a number of different learning models. A description of the data (p15) follows: "The Advertising data set consists of the sales of that product in 200 different markets, along with advertising budgets for the product in each of those markets for three different media: TV, radio, and newspaper. A copy of the data was downloaded from: https://www.kaggle.com/ashydv/advertising-dataset and is on Moodle.

Using the Advertising data, answer the following questions (taken from pp59-60 ISLR):

(a) Is there a relationship between advertising budget and sales?

```
# Load the dataset
df = read.csv("advertising.csv")
# Create a linear model and see the summary
model = lm(Sales~., data=df)
summary(model)

Call:
lm(formula = Sales ~ ., data = df)

Residuals:
   Min      1Q Median      3Q     Max
-7.303 -0.824 -0.001  0.898  3.747

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.625124   0.307501   15.04   <2e-16 ***
TV          0.054446   0.001375   39.59   <2e-16 ***
Radio       0.107001   0.008490   12.60   <2e-16 ***
Newspaper   0.000336   0.005788    0.06     0.95
```

```
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.66 on 196 degrees of freedom
Multiple R-squared:  0.903,   Adjusted R-squared:  0.901
F-statistic:  605 on 3 and 196 DF,  p-value: <2e-16

# p-value overall is very small, hence we can conclude that at least
# one of the predictors is significant. Therefore we can conclude that
# there is a relationship between advertising budget and sales.
```

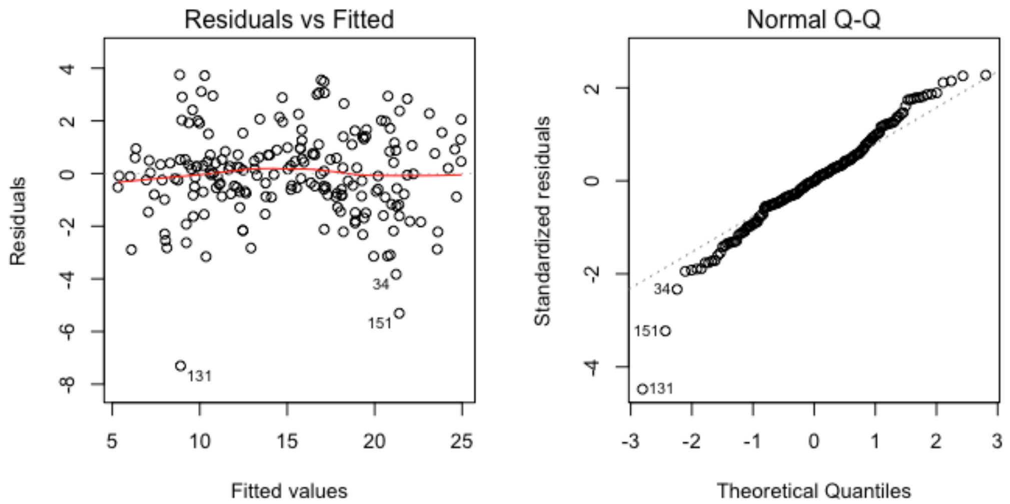
(b) How strong is this relationship?

```
# The strength of the relationship is given by R^2, which in this case is
    # approx. 0.9, meaning that 90% of the variability in the target
    variable # is explained by the predictors of the model.
```

(c) Is the relationship linear?

```
# The plots of residuals show they are approximately normal (q-q) and that
    # they are more or less uncorrelated (resid vs fitted). therefore we
# assume the relationship is linear.
```

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



(d) Which media contribute to sales?

```
# TV and Radio has very low p-values i.e <2e-16 therefore we can conclude
# they contribute to sales.
# Newspaper has a very high p-value, therefore we cannot conclude it has
# a significant contribution to sales.
```

(e) How accurately can we estimate the effect of each medium on sales?

```
# Standard Error is the estimated standard deviation of the error of
# the estimate. This indicates the relative accuracy of each predictor.
```

(f) (Extension) Is there synergy (interactions) among the advertising media?

```
# Let's add the interaction of Radio*TV to our model
model2 = lm(Sales~.+TV*Radio, data=df)
summary(model2)
```

```
Call:
lm(formula = Sales ~ . + TV * Radio, data = df)

Residuals:
   Min     1Q Median     3Q    Max
-6.269 -0.877 -0.048  0.934  3.652

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 6.172270   0.419243   14.72  < 2e-16 ***
TV          0.043546   0.002498   17.43  < 2e-16 ***
Radio       0.041458   0.015129    2.74   0.0067 **
Newspaper   0.001349   0.005454    0.25   0.8049
TV:Radio    0.000444   0.000087    5.10  7.9e-07 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.56 on 195 degrees of freedom
Multiple R-squared:  0.914,   Adjusted R-squared:  0.912
F-statistic:  519 on 4 and 195 DF,  p-value: <2e-16

# Looking at summary output, we can see that p-value of interaction term
      is quite low suggesting it is an important variable for our model.
# We can also observe that R^2 value is increased around 1%.
```

Potential ways of addressing these questions using regression models and extensive discussion of regression can be found on pages 59-82 of ISLR.