Project 6

Stat 217

Due Friday, April 24, 2015

The Federal Trade Commission annually rates varieties of domestic cigarettes according to their tar, nicotine, and carbon monoxide content. The United States Surgeon General considers each of these substances hazardous to a smoker's health. Past studies have shown that increases in the tar and nicotine content of a cigarette are accompanied by an increase in the carbon monoxide emitted from the cigarette smoke. This information motivates our goal of predicting the variable (carbon).

The dataset you are provided with contains measurements of weight, tar, nicotine, and carbon monoxide content for 60 cigarettes in a file called `cig.csv`. You should be able to import the data into RStudio the same way you have for the previous projects. The variables used in this dataset are below.

Variable	Description
tar	Tar Content (mg)
nicotine	Nicotine Content (mg)
weight	Weight (g)
carbon	Carbon monoxide content (mg)

Table 1: Description of Variables in Cigarette Data

The following R code produces a scatterplot matrix of the data and fits models that explain carbon using several different combinations of explanatory variables.

```
# make sure the data are imported as an object called cig
head(cig)
pairs(~tar + nicotine + weight + carbon, data = cig)
# the four models
m1 <- lm(carbon ~ tar, data = cig)
m2 <- lm(carbon ~ tar + nicotine, data = cig)
m3 <- lm(carbon ~ tar + weight, data = cig)
m4 <- lm(carbon ~ tar + nicotine + weight, data = cig)
```

Use the `summary` command to tell R to generate a summary of each fitted model. For example, to summarize the model which only uses `tar` as an explanatory variable

```
summary(m1)
```

```
Call:
lm(formula = carbon ~ tar, data = cig)

Residuals:
   Min     1Q Median     3Q    Max
-4.627 -1.173  0.118  0.820  5.002

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   1.9010     0.6492    2.93   0.0049
```

```
tar           0.8467     0.0513   16.49   <2e-16

Residual standard error: 1.88 on 58 degrees of freedom
Multiple R-squared:  0.824, Adjusted R-squared:  0.821
F-statistic:  272 on 1 and 58 DF,  p-value: <2e-16
```

Compare these 4 models, and decide which is best. Write a statistical report summarizing this model and explaining how you decided on this model. The "Writing a Statistical Report for STAT 217" suggestions will be useful, but make sure to include

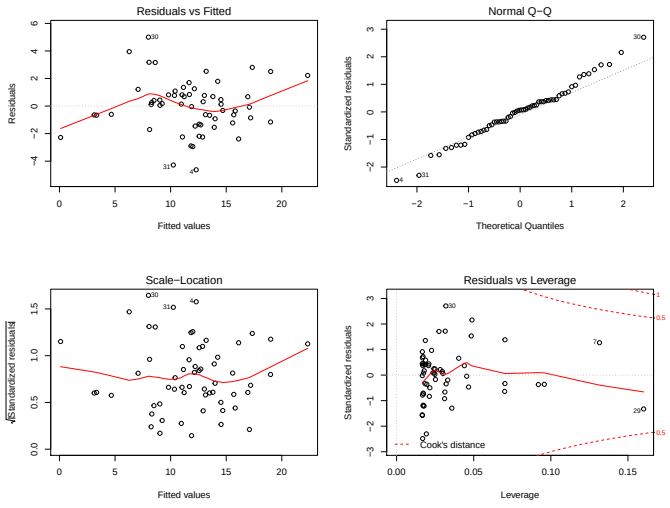
Introduction

- Provide a basic description of the data available and how they were collected
- The primary goal for this analysis is to identify which of the four models best explains carbon content in cigarettes

Statistical Procedures Used

- state the response variable and all potential explanatory variables
- use the scatterplot matrix to assess the strength, form, and direction of the relationship between the response and each explanatory variable
- state which models you compared and how you compared them
- assess the assumptions for the model that you chose to use. The assumptions must be met for this model. To assess these assumptions, you'll need to use the `plot` command like so

```
par(mfrow = c(2,2))    # generate two rows and two columns of plots
plot(m1)               # diagnostics plots for the model
par(mfrow = c(1,1))    # go back to one row and one column of plots
```



Summary of Statistical Findings

- for each comparison of two models, state which model is preferred and why
- write the estimated equation for your chosen model

- provide 95% confidence intervals for each slope coefficient in your model and interpret them in the context of the problem

Scope of Inference

As usual:

- is it permissible to infer these results to all possible cigarettes?
- can we say that changes in tar, nicotine, and weight cause changes in carbon?