MLDS 401: Homework 4
Due: Oct 30, 15:00
Professor Malthouse

1. A marketing research consultant evaluated the effects of the fee schedule, scope of work, and type of supervisory control on the quality of work performed under contract by independent marketing research agencies. The quality of work performed was measured by an index taking into account several characteristics of quality. Four agencies were chosen for each factor level combination and the quality of their work evaluated.

```
mrcontract = expand.grid(agency=LETTERS[1:4], sup=c("local","travel"),
  scope=c("in-house", "subcontract"), fee=c("high","med","low"))
mrcontract$quality=c(124.3,120.6,120.7,122.6,112.7,110.2,113.5,108.6,115.1,
  119.9,115.4,117.3,88.2,96,96.4,90.1,119.3,118.9,125.3,121.4,113.6,109.1,
  108.9,112.3,117.2,114.4,113.4,120,92.7,91.1,90.7,87.9,90.9,95.3,88.8,
  92,78.6,80.6,83.5,77.1,89.9,83,86.5,82.7,58.6,63.5,59.8,62.3)
```

   (a) Regress `quality` on `agency`, `fee` and an interaction between `sup` and `scope`. State the estimated regression equation and use drop1 to test which terms are significant.

   (b) Are there differences in quality between the agencies? To receive full credit state the null and alternative hypotheses, find the P value, state you decision (reject or not), and summarize your conclusion.

   (c) Are there differences in quality between the fee values? To receive full credit state the null and alternative hypotheses, find the P value, state you decision (reject or not), and summarize your conclusion.

   (d) What does the coefficient for `feemed` tell you? Test whether it is different from 0 and discuss what the results of this tell you from a managerial perspective.

   (e) Is the interaction between `sup` and `scope` significant? To receive full credit state the null and alternative hypotheses, find the P value, and state you decision (reject or not).

   (f) Construct and interaction plot for `sup` and `scope`. Write one sentence summarizing what the interaction plot tells you.

2. An experiment is conducted to study the influence of operating temperature and three types of face-plate glass in the light output of an oscilloscope tube.

```
dat = data.frame(type=c(rep("A",9), rep("B",9), rep("C",9)),
  temp=rep(c(100,125,150), 9),
  y=c(580,1090,1392,568,1087,1380,570,1085,1386,550,1070,1328,530,1035,1312,
    579,1000,1299,546,1045,867,575,1053,904,599,1066,889))
```

(a) Generate an interaction plot.

(b) Fit a model with main effects and an interaction term. Hint: `y ~ type * factor(temp)`.

(c) Test whether the overall model is significant by stating the null an alternative hypothesis, P-value and decision. Use $\alpha = 0.05$.

(d) Test whether the interaction is significant by stating the null an alternative hypothesis, P-value and decision. Use $\alpha = 0.05$.

(e) Write a few sentences interpreting the results (tell the story).

(f) Would it be appropriate to treat temperature as a numerical variable with the model `y ~ type * temp`. Explain why or why not.

3. Fit the following model using the auto data.

```
auto$origin = factor(auto$origin, 1:3, c("US", "Europe", "Japan"))
fit = lm(log(mpg) ~ origin*log(displacement) + year, data=auto)
```

(a) State the model that is being estimated in terms of the parameters (β_j) rather than the estimates. This part establishes the notation.

(b) Compute the partial (`drop1`) sums of squares and F tests. What null and alternative hypotheses are being tested by the `origin:log(displacement)` line? State them using the notation from part (a) and say what they mean in English.

(c) Show where each of the numbers in the `origin:log(displacement)` come from with the exception of AIC, which has not been covered yet. What do you conclude from the test?

(d) Examine the `summary` and interpret the `originJapan:log(displacement)` line using both symbols from part (a) and in simple English.

(e) For each value of `origin` write out the equation for how `log(displacement)` is associated with (mean) `mpg`, controlling for (year).

(f) Draw a graph showing how `log(displacement)` is associated with `mpg` with three lines, one for each `origin`. Encode `origin` using color and/or line type (e.g., solid, dashed, dotted, etc.). One way to think of this is for fixed year, let some constant a = intercept + year*(year slope estimate). The three lines are all relative to a.

4. Build a model to predict Divvy demand as a function of the predictors you have been given. You will not be able to include all of the variables in a model. Start by examining a correlation matrix and scatterplots. Ultimately, I want a model that is both interesting and correct. The conclusions should also be robust to small changes in the specification. Submit your model, VIFs, and a written summary of your conclusions. Here are a few hints:

    - Think about why some crimes should be positively associated with trips, why others would be negatively associated, and why some should not affect trips at all. Test whether your explanations are correct. Having a good reason why is critical to making the results interesting.
    - You may want to form composite variables. For example, you could form a new variable that measures the extent that a station is located in a "central business district" (CBD). Many variables indicate a CBD, such as having many businesses and train stations, while residential neighborhoods will have fewer businesses. For each observation, you could average variables that you have to measure CBD. You will also want to group crime variables into types of crimes.
    - Come to class on Monday prepared to talk about your model!