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. [*Answer:* $\texttt{quality} = 122.45 + 0.12\texttt{agencyB} + 0.15\texttt{agencyC} - 0.57\texttt{agencyD} - 0.96\texttt{feemed} - 31.16\texttt{feelow} - 10.95\texttt{suptrav} - 5.44\texttt{scopesubcon} - 13.84\texttt{suptrav} : \texttt{scopesubcon}$]

(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. [*Answer:* $H_0 : \beta_1 = \beta_2 = \beta_3 = 0$ *versus* H_1 *: at least one of the three is nonzero.* $P = .8982 > 5\%$ *so we cannot reject* H_0*. We have no evidence to say that there is a difference in quality between agencies after for controlling for fee, scope and supervision.*]

```
> drop1(fit, test="F")
Model: quality ~ agency + fee + sup * scope
          Df Sum of Sq     RSS     AIC  F value    Pr(>F)
<none>                   268.4 100.624
agency     3       4.1   272.5  95.344   0.1964    0.8982
fee        2   10044.3 10312.7 271.757 729.7061 < 2.2e-16 ***
sup:scope  1     574.8   843.2 153.568  83.5137 3.077e-11 ***

> summary(fit)
                           Estimate Std. Error t value Pr(>|t|)
(Intercept)                122.4542     1.1360 107.796  < 2e-16 ***
```

```
agencyB                      0.1250     1.0710   0.117    0.908
agencyC                      0.1500     1.0710   0.140    0.889
agencyD                     -0.5667     1.0710  -0.529    0.600
feemed                      -0.9625     0.9275  -1.038    0.306
feelow                     -31.1563     0.9275 -33.591  < 2e-16 ***
suptravel                  -10.9500     1.0710 -10.224 1.36e-12 ***
scopesubcontract            -5.4417     1.0710  -5.081 9.70e-06 ***
suptravel:scopesubcontract -13.8417     1.5146  -9.139 3.08e-11 ***
```

(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. [*Answer:* $H_0 : \beta_4 = \beta_5 = 0$ *versus* H_1 *: at least one of the two is nonzero.* $P = 2.2 \times 10^{-16} < .05$ *(from the above output), so reject* H_0*. Based on an inspection of the coefficients, low fee produces substantially lower quality, but there is little difference in quality between medium and high fee.*]

(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. [*Answer: The estimate* -0.9625 *tells the expected difference in quality between medium and high fee, after controlling for the other factors.* $H_0 : \beta_4 = 0$ *versus* $H_1 : \beta_4 \neq 0$*. We cannot reject* H_0 *(*$P = .306$*, from above output), and so we do not have evidence to conclude that high fee produces a different level of quality than medium fee.*]

(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). [*Answer:* $H_0 : \beta_8 = 0$ *versus* $H_1 : \beta_8 \neq 0$*.* $P = 3.08 \times 10^{-11} < 5\%$*, so we reject* H_0*.*]

(f) Construct and interaction plot for `sup` and `scope`. Write one sentence summarizing what the interaction plot tells you. [*Answer:* *`interaction.plot(mrcontract$sup, mrcontract$scope, mrcontract$quality, col=1:2)`. There is a substantial drop in quality when the supervision travels and the scope is subcontracted, but the other three combinations have roughly equal average quality levels.*]

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. [*Answer: Either of these count*]

```
with(dat, interaction.plot(type, temp, y, col=1:3))
with(dat, interaction.plot(temp, type, y, col=1:3))
```

(b) Fit a model with main effects and an interaction term. Hint: `y ~ type * factor(temp)`. [*Answer: The interaction is significant*]

```
Single term deletions

Model: y ~ type * factor(temp)
                  Df Sum of Sq    RSS    AIC F value    Pr(>F)
<none>                           6579 166.39
type:factor(temp)  4    290552 297131 261.26  198.73 1.254e-14 ***

Call: lm(formula = y ~ type * factor(temp), data = dat)

                       Estimate Std. Error t value Pr(>|t|)
(Intercept)            572.6667    11.0381  51.881  < 2e-16 ***
typeB                  -19.6667    15.6102  -1.260   0.2238
typeC                    0.6667    15.6102   0.043   0.9664
factor(temp)125        514.6667    15.6102  32.970  < 2e-16 ***
factor(temp)150        813.3333    15.6102  52.103  < 2e-16 ***
typeB:factor(temp)125  -32.6667    22.0762  -1.480   0.1562
typeC:factor(temp)125  -33.3333    22.0762  -1.510   0.1484
typeB:factor(temp)150  -53.3333    22.0762  -2.416   0.0265 *
typeC:factor(temp)150 -500.0000    22.0762 -22.649 1.11e-14 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 19.12 on 18 degrees of freedom
Multiple R-squared:  0.9973,Adjusted R-squared:  0.9961
F-statistic: 824.8 on 8 and 18 DF,  p-value: < 2.2e-16
```

(c) Test whether the overall model is significant by stating the null an alternative hypothesis, P-value and decision. Use $\alpha = 0.05$. [*Answer:* $H_0 : \beta_1 = \cdots = \beta_8 = 0$ *versus* H_1 *: at least one* $\beta_j \neq 0$ *for* $j = 1, \ldots, 8$. $P = 2.2 \times 10^{-16} < .05$ *so reject* H_0.]

(d) Test whether the interaction is significant by stating the null an alternative hypothesis, P-value and decision. Use $\alpha = 0.05$. [*Answer:* $H_0 : \beta_5 = \cdots = \beta_8 = 0$ *versus* H_1 *: at least one* $\beta_j \neq 0$ *for* $j = 5, \ldots, 8$. $P = 1.25 \times 10^{-14} < .05$ *so reject* H_0.]

(e) Write a few sentences interpreting the results (tell the story). [*Answer: I assume we want high y. Type A or B at a temperature of 150 gives the largest values of y. If you must use C then set the temperature to 125.*]

(f) Would it be appropriate to treat temperature as a numerical variable with the model `y` ∼ `type * temp`. Explain why or why not. [*Answer: No. If we treat it as linear then we assume the effect of temperature on y is linear, possibly with different slopes across types. The effect of temperature is not linear for type C, which has more of an inverted U shape based on the interaction plot.*]

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)
```

```
> fit = lm(log(mpg) ~ origin*log(displacement) + year, data=auto)
> drop1(fit, test="F")
Single term deletions

Model:
log(mpg) ~ origin * log(displacement) + year
                         Df Sum of Sq    RSS     AIC  F value    Pr(>F)
<none>                                 7.207 -1577.5
year                      1    4.1374 11.344 -1399.4 223.8920 < 2.2e-16 ***
origin:log(displacement)  2    0.2270  7.434 -1569.2   6.1422  0.002364 **

> summary(fit)

Call: lm(log(mpg) ~ origin * log(displacement) + year, data = auto)
Coefficients:
                                Estimate Std. Error t value Pr(>|t|)
(Intercept)                     3.241432   0.222751  14.552  < 2e-16 ***
originEurope                    1.276599   0.409746   3.116  0.00197 **
originJapan                    -0.416641   0.353779  -1.178  0.23964
log(displacement)              -0.480425   0.020793 -23.105  < 2e-16 ***
year                            0.030578   0.002044  14.963  < 2e-16 ***
originEurope:log(displacement) -0.276179   0.086682  -3.186  0.00156 **
originJapan:log(displacement)   0.090394   0.075651   1.195  0.23286
---
Residual standard error: 0.1359 on 390 degrees of freedom
Multiple R-squared:  0.8426,Adjusted R-squared:  0.8402
F-statistic: 348.1 on 6 and 390 DF,  p-value: < 2.2e-16
```

(a) State the model that is being estimated in terms of the parameters (β_j) rather than the estimates. This part establishes the notation. [*Answer:* $\log(\texttt{mpg} + \beta_0 + \beta_1\texttt{Eur} + \beta_2\texttt{Japan} + \beta_3\log(\texttt{disp}) + \beta_4\texttt{year} + \beta_5\texttt{Eur}\log(\texttt{disp}) + \beta_6\texttt{Japan}\log(\texttt{disp})$]

(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. [*Answer: $H_0 : \beta_5 = \beta_6 = 0$ versus $H_1 : \beta_5 \neq 0$ and/or $\beta_6 \neq 0$. In English, H_0 says that the slopes of log(displace) for Europe and Japan are the same as for the US; H_1 says that at least one of the other regions have a different slope for log(displace).*]

(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?

```
> fit2 = lm(log(mpg) ~ origin+log(displacement) + year, data=auto)
                         Df Sum of Sq    RSS     AIC  F value    Pr(>F)
<none>                                 7.207 -1577.5
origin:log(displacement)  2    0.2270  7.434 -1569.2   6.1422  0.002364 **
> deviance(fit) # RSS for full model
[1] 7.206968
> deviance(fit2)  # RSS for reduced model
[1] 7.433975
> deviance(fit2) - deviance(fit) # difference sum of squares
[1] 0.2270073
> (0.2270073/2) / (7.206968/390) # F
[1] 6.14217
> 1-pf(6.1422, 2, 390)   # P(>F)
[1] 0.002363883
```

[*Answer: The 2df refer to constraining $\beta_5 = \beta_6 = 0$. Under the constraint we have only main effects for origin and log(displacement) and* $7.433975 - 7.206968 = 0.2270$ *is how much the sums of squares increase. See my comments to the R code for how the other values are computed.*]

(d) Examine the `summary` and interpret the `originJapan:log(displacement)` line using both symbols from part (a) and in simple English. [*Answer: It tests $H_0 : \beta_6 = 0$ versus $H_1 : \beta_6 \neq 0$. In English, the slope for Japan of log(displace) equals the slope for the base category (US). $P = 0.2929 > 5\%$ indicating that we do not have evidence to say that the Japan slope is different from the US slope.*]

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

$$\text{US}: \log(\texttt{mpg}) = 3.24 - 0.480\log(\texttt{displace}) + 0.0306\texttt{year}$$

$$\text{Europe}: \log(\texttt{mpg}) = (3.24+1.28) + \underbrace{(-0.480 - 0.276)}_{-0.756}\log(\texttt{displace}) + 0.0306\texttt{year}$$

$$\text{Japan}: \log(\texttt{mpg}) = (3.24-0.417) + \underbrace{(-0.480 + 0.0904)}_{-0.390}\log(\texttt{displace}) + 0.0306\texttt{year}$$

]

(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. [*Answer: The three lines are US:* $\log(\texttt{mpg}) = a - 0.480\log(\texttt{displace})$. *Europe:* $\log(\texttt{mpg}) = (a+1.28) - 0.756\log(\texttt{displace})$. *Japan:* $\log(\texttt{mpg}) = (a-0.417) - 0.390\log(\texttt{displace})$.]

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!