FORE224/STAT202 Assignment 8

Modeling with both continuous and categorical predictors, and linear transformations of variables

Due at 12 noon, 13 October 2021

As usual, create a new project in its own folder for this assignment.

[1] Load the `tidyverse` library and the `GGAlly` library.

Download the file diamonds.csv from Learn and save it in your project folder. The file contains data for 660 diamonds. Three variables are recorded:

- the continuous variable price, which is the diamond’s price in US$.
- the continuous variable carat, which is a measure of the weight of the diamond (1 carat = 0.2 g). This is a cheapskate dataset - the diamonds are all less than 1 carat :)
- the continuous variable x, which is the diamond’s width in mm.
- the categorical variable cut, with values “Good”, “Very Good”, “Premium”

You will use this data to create a model for diamond price.

Read the dataset from diamonds.csv and make sure that the cut variable is treated as a factor with levels “Good”, “Very Good”, “Premium” where “Good” is the baseline group.

Use `set.seed` to set the random seed to your student ID number and make a subset called my_diamonds of 640 cases from this diamonds data using `sample_n` as in previous labs. Use this subset of the data for the rest of the assignment.

[2] Use `ggpairs` from the `GGally` library to create a matrix of plots and correlations between the continuous variables. Optional: include `aes(colour = cut)` inside the `ggpairs` function to get the points coloured by cut.

Optional: If you want to you can try using the `scatterplot3d` library, as used in the some of the code posted for lectures (or any other 3-d plotting package of your choice) to create a 3-d scatterplot showing the 3 continuous variables price, carat, and x. The following code will create a basic scatterplot3d with points coloured by cut.

```
library(scatterplot3d)

s3d <- scatterplot3d(my_diamonds[, c(2, 3, 1)], # get price on z axis
  pch = 16,
  color = as.numeric(my_diamonds$cut)
)
```

Alternatively you could try `plot3d` from the `rgl` library, which makes a rotatable plot in a separate window:

```
library(rgl)
# make rotatable 3-d plot in separate window
plot3d(my_diamonds[, c(2, 3, 1)], col = as.numeric(my_diamonds$cut))
```

[3] Use `lm` to create a model called m1 for price using the continuous predictors carat and x and the categorical predictor cut, including both continuous-categorical interactions. Specify the model formula like this:

```
price ~ carat + x + cut + x:cut + carat:cut
```

Show the code used to create model m1 in your report and also show the analysis of variance for this model given by R's `anova` function.

Now create an equivalent model (same response variable and same predictors) called m2 but specify the model formula like this:

```
price ~ carat + x + cut + carat:cut + x:cut
```

Show the code used to create model m2 in your report and also show the analysis of variance for this model given by R's `anova` function.

Very briefly (about 20 words) point out the differences in the `anova` output for the two models.

[4] Create a model called m3 for response price using the continuous predictors carat and x and the categorical predictor cut, including the interaction between carat & cut but without the interaction between x & cut. Your choice on the order of the terms in the formula!

Use R's `anova` function to test the full model m1 against the reduced model m3. Show the code you used to perform the test and the output in your report.

Similarly, test the full model m2 against the reduced model m3 and show the code you used to perform the test and the output in your report.

Compare the `anova` output from these two tests. Is it different? What do you conclude about using output from `anova` with a single model compared to using output from a 'nested model' test with `anova` to assess the statistical significance of the interaction between x & cut? Answer in about 80-100 words.

[5] Is the interaction between x and cut statistically significant? Give the evidence that you are using and your conclusion (about 40 words in total).

[6] Create model m4 that removes the interaction between carat and cut from model m3 (ie, uses carat, x and cut as predictors but with no interactions at all). Perform a nested model F-test of the full model (m1 or m2) against m4 using `anova`. Show the code and the `anova` output.

Explain in about 40 words why m4 is compared to the full model (m1 or m2) rather than being compared to m3 in this test.

Give give your conclusions for this test (about 20 words).

Is statistical significance the only thing we would consider when deciding whether to keep the interaction in the model? Briefly (50-60 words) outline other consideration(s), if any, that might be relevant.

[7] Describe model m3 geometrically in about 30 words. (Eg, an equation *y* = *mx* + *c* represents a single straight line; what does model m3 represent?)

Create the regression diagnostic plots for model m3 (at least the usual 4 plots and any others you want to add). Include the plots in your report. Comment in 40-50 words on anything that causes you concern.

Optional: you can also investigate the distribution of the residuals using a density plot, for example using the following code (there are also other ways to get the residuals and plot them and you can choose what to use if you want to take this option):

```
library(broom)
my_diamonds_augmented <- augment(m3)
my_diamonds_augmented %>% ggplot(aes(x = .resid, color = cut)) +
  geom_density()
```

[8] Show the model summary output for model m3.

Create a new variable in the data called price_NZ000 which converts the US$ prices in the original data to NZ$ and also gives the values in units of NZ$1000. 1.45 is a reasonable exchange rate to use for the currency conversion. Show the code used to create the new variable in your report.

Create a model m5 which is the equivalent of model m3 but where the response is the variable price_NZ000. Show the code used to create the model and the model summary output in your report.

Explain any differences between the summary output for models m3 and m5 and comment on whether using the new predictor variable has changed the model fit (about 80 words in total).

[9] Create a tibble of new data to use for prediction. The new data should have carat values 0.25, 0.4, 0.6, 0.9, x values all 4.7 and all Premium cut.

Use R's `predict` function with model m5 and the new data with `interval = "confidence"` and then with `interval = "prediction"`.

Explain in about 80-100 words what the output from `predict` shows when `interval = "confidence"` and when `interval = "prediction"`.

Include all the answers and code used to obtain them in a Word document and submit it through Learn. Remember to use your own words when answering the questions.