INFO370 Lab: Categorical variables

February 12, 2022

Instructions

This lab is about interpreting categorical variables. In particular you will look at diamond prices and how different cuts affect the price. You will also predict the price based on your model.

The dataset here is diamonds data (extracted from R ggplot2 package). It contains the following variables:

carat diamond's mass, in carats (=0.2 g)

cut shape of the diamond. Certain shapes are considered more valuable than others, "Ideal" is the best (consult Blue nile website for more explanations about diamonds cut, color and other characteristics).

color color of diamonds, D (colorless) is the best.

clarity transparency of diamonds, FL (flawless) is the best.

depth diamonds relative height (see brilliance.com for explanations)

table diamonds width, related to *depth*

price in dollars

x, y, z dimension, mm

1 Diamonds data

Our first task is to take a look at diamonds data.

1. Load the data and perform basic checks. How many diamonds do we have?
2. Inspect the variable *cut*. What kind of different cuts are there? How frequent are those?

   Hint: remember the method `value_counts`?

---

2 Regression analysis

Now it is time for regression analysis

1. Perform a linear regression analysis where you include two variables: mass of diamonds (*carat*) and *cut*. Estimate model of a form

$$\text{price}_i = \beta_0 + \beta_1 \text{carat}_i + \boldsymbol{\beta}_2' \cdot \mathbf{cut}_i + \epsilon_i \tag{1}$$

   Show the estimation results (the table).

   Note: in the notation above $\boldsymbol{\beta}_2$ is a vector of cut-related coefficients and **cut** is a vector of dummies. *smf* will do this automatically for you.

2. What is the reference category for *cut*?

3. Interpret the following coefficients:

   (a) What is "carat" (correct value should be 7871)?

   (b) What is "cut[T:Ideal]" (correct value should be 1800)?

   (c) How much more expensive are ideal-cut diamonds compared to very good-cut diamonds (in average) given they weight the same?

   (d) What does your model predict–what is be the (average) price for 1ct premium cut diamond?

   Note: if you are wondering what does negative intercept mean here, check out the first challenge question below.

---

3 Challenge (not graded)

If you have time and interest, then consider also doing the following tasks:

1. Plot the price versus mass and add the regression line to it. Can you explain why do we have negative intercept on this plot?

2. Experiment with log scale for a) price, b) carat, c) both. Which plot does look the best?

3. Add the corresponding log-transform to your model. You can run log-transformed models just like `smf.ols("np.log(price)~np.log(carat)", ...)`.

4. What does R^2 suggest: is log-transformed model better than non-log transformed?

5. Add `cut` to this model and interpret the coefficients.

   Hint: see https://faculty.washington.edu/otoomet/machineLearning.pdf Section 4.1.8 *Feature Transformations*.
