CS100 Homework #3, Part 1

Linear Regression

Part 1 Due: November 14, 2019, at 11:59 pm

Instructions

Please submit your solutions to Canvas, as an R markdown (.Rmd) file. Please also knit your R markdown file, and submit the resulting .HTML file as well.

Be sure to follow the CS100 course collaboration policy as you work on this and all CS100 assignments.

Overview

The topic of this homework assignment is supervised learning. The first half is concerned with linear regression, and the second half, classification.

Part 1: Pricing Monet Paintings

![Monet, Grainstack](image)

Figure 1: Monet, Grainstack
Claude Monet (1840-1926) was one of the founders of French Impressionism, and his artwork remains highly prized today. Monet’s paintings have sold for record amounts; for example, ‘Grainstack’ (picture above) was sold at auction for $81.4 Million in 2016.

The goal of this problem is for you to develop an interpretable multiple regression model that predicts the price (dependent variable) of a Monet painting from a few of its features (independent variables). The data are located here.

Load the data into R, and examine the features. WIDTH and HEIGHT represent the width and height of the painting. The SIGNED feature indicates whether or not Monet’s signature appears on the painting: 1 means a painting is signed, and 0, unsigned. PICTURE is an ID for the painting, and HOUSE indicates the auction house at which the painting was sold.

Using the `pairs` and `cor` functions, explore the relationship between different the various features in the data set. Which features exhibit the highest correlation with one another? And which features are most highly correlated with PRICE?

**Simple Linear Regression**

Two of the features of these data that are correlated with PRICE are WIDTH and HEIGHT. However, these two features are also highly correlated with each other! Including correlated features in a linear regression can make the model difficult to interpret, because it is difficult to understand the effect of a change in any one predictor on the prediction, when changing one inadvertently changes another. In such cases, it may make sense to replace the correlated features with a single feature that represents their interaction. Mutate the data to create a new feature SIZE that accounts for both HEIGHT and WIDTH.

Find the correlation between PRICE and SIZE. Transform the variables to see if there might be a stronger correlation in another space. Try various combinations of `sqrt` and `log` transformations. Create scatter plots of these combinations to visualize their correlations.

Take a close look at the plots where you do not transform the predictor. What do you notice about these plots? We noticed a clustering effect. It seems as though paintings whose SIZE is less than about 3000 might demand a different pricing model than paintings whose SIZE is greater. Your goal in this problem is to build a model of smaller paintings only. (It would be challenging to build a credible model of bigger paintings from the very few data available.)

Filter from the data all paintings whose size is greater than, say, 2500. Then revisit (i.e., redo) the transformations to identify the best one, and create a simple linear regression model in the best space. *(Hint: Use `lm`.*) Depict your linear model on a scatter plot using `abline`. *(Hint: If your model is called `model`, the line of code `abline(model, col = "red")` adds a red line corresponding to your model to a plot.)*
Next, plot the residuals. A residual plot depicts fitted (i.e., predicted) values on the 

\[ x \]

axis and the residuals on the 

\[ y \]

axis. Note that residual plots do not show observations on the 

\[ x \]

axis, because they are correlated with the residuals, making such plots harder to interpret.

*Hint:* The `fitted` function extracts the fitted values from a model, and the `residuals` function extracts the residuals: i.e., the difference between the observations and the fitted values.

*Another Hint:* If your model is called `model`, you can view the residual plots, and a few other diagnostic plots, using `plot(model)`. To view only the residual plot, use `plot(model, which = 1)`.

Comment on the goodness of fit of your model in terms of the relevant statistics (e.g., its \( p \)-values, its \( R^2 \)) and anything else of note.

Beyond the aforementioned regression diagnostics, it can also be useful to visualize the curve that your linear model represents. To do so, you should add this curve to a scatter plot of your data, in this case PRICE vs. SIZE.

To add a curve to a plot, you can use (you guessed it!) the `curve` function, which takes as input a function of \( x \) (i.e., \( x \times 2 \)) describing a curve to add to a plot. You should input to the curve function the inverse of the transformation you applied: i.e., where you took a log, you should exponentiate.

For example, if you took the log of PRICE, so that your linear model is

\[ \log(Y) = a + bX \]

, where

\[ Y \]

is PRICE, then your curve is

\[ e^{a+bX} \]

. You can can plot this curve as follows:

```r
curve(exp(a + b * x), col = "red", add = TRUE)
```
Alternatively, if you took the log of both PRICE and SIZE, so that your linear model is
\[ \log(Y) = a + b \log(X) \]
, then your curve is
\[ e^a X^b \]
You can plot this curve as follows:
\[ \text{curve}(\exp(a) \times (x ** b), \col = "red", \text{add} = \text{TRUE}) \]
Finally, if you took the log of PRICE and the sqrt of SIZE, so that your linear model is
\[ \log(Y) = a + b \sqrt{X} \]
, then your curve is
\[ e^{a+b\sqrt{X}} \]
You can plot this curve as follows:
\[ \text{curve}(\exp(a + b \times \sqrt{x}), \col = "red", \text{add} = \text{TRUE}) \]
You can extract the relevant coefficients from your model as follows:
\begin{verbatim}
a <- model$coefficients[1]
b <- model$coefficients[2]
\end{verbatim}

**Multiple Linear Regression**

There are other variables beyond the size of a painting that influence its price. Intuitively, what other variables might matter? Both SIGNED and HOUSE are good contenders. You are going to explore the effect of SIGNED in the second part of this analysis. (You should continue to work with only small paintings.)

SIGNED is a binary value; its values are either 0 or 1. Create a box plot of prices across the SIGNED category. What is the average price of signed vs. unsigned paintings? Do the data for signed and unsigned paintings seem to be drawn from the same distribution?

Create a scatter plot of PRICE vs. SIZE, and color the points in your plot based on the SIGNED feature.

*Hint:* To do so, call the `plot` function with the `col` argument, and set its value to `SIGNED + 1` (or plus 2 or 3 or 4, if you don’t like red and black).

Then, continuing with your preferred transformation, use `lm` to create a multiple linear regression model for the transformed data.

*Hint:* In R, creating a multiple linear regression model is nearly as simple as creating a simple linear regression model: use the `lm` function as follows: `lm(PRICE ~ SIZE + SIGNED)`.

As above, summarize the model. Then add an `abline` to your plot (in the transformed space) for signed paintings, and a second one for unsigned paintings.
Hint: Recall that SIGNED is a binary, or dummy, variable. Consequently, the linear equation \( \text{PRICE} = a + b \times \text{SIZE} + \text{dummy} \times \text{SIGNED} \) reduces to either \( \text{PRICE} = a + b \times \text{SIZE} \) or \( \text{PRICE} = a + b \times \text{SIZE} + \text{dummy} \), depending on the value of dummy. In other words, your linear model implicitly encodes not one, but two, lines.

Hint: Earlier, you simply called the abline function with your linear model as an argument, and R did the hard work of extracting the relevant coefficients from your model to draw the requisite line. As we would like you to extract two lines from your multiple regression model, you should call abline twice, with two different values of the \( a \) (intercept) and \( b \) (slope) parameters that depend on the coefficients of your model, namely:

\[
\begin{align*}
a &\leftarrow \text{model}$$coefficients[1] \\
b &\leftarrow \text{model}$$coefficients[2] \\
dummy &\leftarrow \text{model}$$coefficients[3]
\end{align*}
\]

Finally, plot the residuals, comment on the goodness-of-fit statistics of your model, and plot the two curves that correspond to your two lines in linear space.

If you are unhappy with the fit your model obtained, repeat some of the earlier steps (e.g., try a different transformation) until you are satisfied. But don’t forget to change the inputs to the curve function based on the transformations you apply!

Which transformation seems to create the best model for Monet’s small paintings? Try to provide an intuitive explanation of what this transformation says about the price of Monet’s (smaller) paintings?