# INFO370 Problem Set 6: Multiple Regression, Categorical Variables

November 17, 2020

## Instructions

This PS has three goals:

1. learn to use and interpret multiple regression models.
2. learn how to handle categorical (non-numeric) data;
3. learn about log-transformed data.
4. (the extra credit task also asks you to implement and interpret interaction effects)

As an extra bonus, you learn a little bit about basketball ☺

## 1 How Is Basketball Game Score Calculated? (40pt)

In this section you will work with basketball data. Basketball is a big business, and there is a lot of analytics collected about high-profile games. Game score is one of the popular measures of player's performance in game. But how is it calculated?

Here we look at one particular dataset about James Harden's (see photo) 2018-2019 season. We recommend you to be familiarize yourself with the basics of basketball, including what are field goals, turnovers, and personal fouls (wikipedia is a good source).

The dataset contains 30 variables, including field goals, field goal attempts, 3-point field goals, rebounds and personal fouls. See the included readme file for more information.

The central variable in current context is *GmSc*, the game score. It is a summary performance score for the player (given he played in the game).

Here are the tasks:

1. (2pt) Load the data (*harden-18-19.csv*). Do basic sanity checks.
2. (2pt) How many games did James Harden play in the 2018-2019 season?

   Note: the data also includes games where he did not play.
3. (4pt) Clean the data and ensure the relevant variables are of numeric type so we can use those in the regression models. It is your task to find what is wrong with the data in its present form (it is downloaded directly from basketball-reference.com), and fix these issues.

   Hint: a good way to transform text to number is `pd.to_numeric`.


James Harden playing for Rockets in 2017. Keith Allison from Hanover, MD, USA, CC BY-SA 2.0 https://creativecommons.org/licenses/by-sa/2.0, via Wikimedia Commons.

4. (4pt) Analyze the game score *GmSc*. What is its distribution? Range? Mean? Standard deviation?

5. (5pt) First, let's run a simple regression model explaining game score *GmSc* by field goal attempts *FGA*:

$$GmSc_{\mathsf{g}} = \beta_0 + \beta_1 \cdot FGA_{\mathsf{g}} + \epsilon_{\mathsf{g}}$$

   where $\mathsf{g}$ indexes games. (Call this Model 5).

   Display the results and answer the following questions:

   (a) What is the interpretation of *Intercept* ($\beta_0$)?

   (b) What is the interpretation of *FGA* ($\beta_1$)? Is it statistically significant?

6. (7pt) Next, let's analyse how is game score related to field goals (*FG*) and field goal attempts (*FGA*). Estimate the model

$$GmSc_{\mathsf{g}} = \beta_0 + \beta_1 \cdot FG_{\mathsf{g}} + \beta_2 \cdot FGA_{\mathsf{g}} + \epsilon_{\mathsf{g}}.$$

(Call this Model 6).

If done correctly, you should see results approximately 6.9, 3.4, -0.7.

Answer the following questions:

(a) What is the interpretation of $FG$? Is it statistically significant?

(b) What is the interpretation of $FGA$ ($\beta_2$)? Is it statistically significant?

(c) How do you explain the fact that model 5 shows positive and model 6 shows a negative estimate for $FGA$?

(d) What is the $R^2$ of the model? How does it compare to the model 5?

7. (5pt) Now let's include personal fouls ($PF$) to the previous model 6. Estimate the new model and answer the following questions:

   (a) Interpret the effect of $PF$.

   (b) Does adding fouls change the estimates for $FG$ and $FGA$ in any major way?

   (c) What is $R^2$ of the model? How does it compare to the model 6?

8. (7pt) Now include all the independent numerical variables, i.e. $FG$, $FGA$, $3P$, $3PA$, $FT$, $FTA$, $ORB$, $DRB$, $AST$, $STL$, $BLK$, $TOV$, $PF$ into the model. Estimate it, and discuss the results.

   Answer the following questions:

   (a) How do standard errors and `t`-values look like in this model?

   (b) What is $R^2$ of this model? What does it tell you about how game score is calculated?

   (c) What do the results tell about turnover ($TOV$)? Is it good or bad for the team?

   Suggestion: check out patsy `Q()` quoting to include non-valid variable names.

9. (4pt) Finally, consult the game score explanation at https://www.nbastuffer.com/analytics101/game-score/. Did you recover the same formula?

## 2 Model AirBnB Price (60pt)

Your next task is to analyze the Beijing AirBnB listing price (variable *price*). It is downloaded from Inside Airbnb. You have to work with several sorts of categorical variables, including those that contain way too many too small categories. You are also asked to do log-transforms and interpret the results.

1. (4pt) Load the data. Select only relevant variables you need below. Even better, check out the `usecols` argument for `read_csv`. Do basic sanity checks.

2. (10pt) Do the basic data cleaning:

   (a) convert *price* to numeric.

   (b) remove entries with missing or invalid price, bedrooms, and other variables you need below

3. (6pt) Analyze the distribution of *price*. Does it look like normal? Does it look like something else? Does it suggest you should do a log-transformation?

   Hint: consult lecture notes Section 4.1.6 *Interactions and Feature Transformations*.

4. (12pt) Convert the number of bedrooms into another variable with a limited number of categories only, such as 1, 2, 3, 4+; and now convert these into dummies.

   Hint: consult the python companion for lecture notes `https://otoomet.bitbucket.io/machinelearning-py` `html#cleaning_data`.

5. (14pt) Run an OLS where you explain the price with number of bedrooms where bedrooms uses these four categories. Interpret the results, including $R^2$.

   Hint: if you choose 0-BR as the reference category, the effect for 1BR should be -12.62.

6. (13pt) Now repeat the process with the model where you analyze log price instead of price. Interpret the results. Which model behaves better in the sense of $R^2$?

   For the following tasks use either *log(price)* or *price*, depending on your answer here.

7. (11pt) Finally we just add three more variables to the model: *room type*, *accommodates*, and *bathrooms*. While room type only contains three values, the other two contain many different categories. Recode these as

   - accommodates: "1", "2", "3", "4 and more"
   - bathrooms: "0", "1", "2", "3 and more", where the 0.5 is rounded up to the next integer, e.g. 0.5 becomes 1 and 1.5 becomes 2.

   Run this model. Interpret and comment the more interesting/important results. Do not forget to explain what are the relevant reference categories and $R^2$.

## 3 Extra credit (10pt = 1EC)

Here you have to introduce interaction terms: superhost×room type, and interpret the results

1. (1pt) add variable *host is superhost* to your data.

2. (1pt) clean it in a way that only the observations with valid values ($t$ and $f$) are included (drop the other instances).

3. (2pt) introduce interaction effects between superhost and room type.

   Hint: as there are 6 combinations of superhost and room type, your model should include 5 corresponding parameters.

   Hint2: consult James *et al.* (2015), section 3.3.2 (Extensions of the Linear model) and Lecture Notes Section 4.1.6 *Interactions and Feature Transformations*.

4. (6pt) interpret the interaction effect results.

## References

James, G., Witten, D., Hastie, T. and Tibshirani, R. (2015) *An Introduction to Statistical Learning with Applications in R*, Springer.