Building Your Multiple Regression Model with LEGO Bricks-Activity

Building Bricks: For this lab, you will be working with data about a popular brand of building bricks, specifically Lego bricks. A random sample of Lego sets was selected and data was collected from LEGO.com, Brickset.com, and BrickInstructions.com on September 18th, 2020. We would like to predict the price of Lego sets.

1. **Price of Sets vs Number of Pieces:** We will start by working with the Lego City and Friends dataset. Lego City and Lego Friends are two Lego set themes, both set themes may be combined and the pieces fit together. Two variables in the dataset include the number of pieces in the set and the Amazon price per set. For this portion of the lab, you will examine the relationship between Amazon price and number of pieces per set.

   Research Question 1: Is there a relationship between the Amazon price versus the number of pieces per set for Lego City and Lego Friends sets in our sample?

   (a) What are the explanatory and response variables? Explain. You may consider reading this post on how the prices of Lego sets are determined: `https://www.lego.com/en-us/service/help/products/themes-sets/how-we-decide-the-prices-of-lego-sets-408100000008322`.

   (b) Create a scatterplot comparing Amazon price versus number of pieces. Describe the relationship between price and number of pieces based on this scatterplot.

   (c) Run a simple linear regression analysis using the fit Y by X utility in JMP. What is the least squares regression equation?

   (d) What is the value of the slope? Interpret the value of the slope in context.

   (e) Provide an answer to Research Question 1.

2. **Lego City vs Lego Friends:** For this portion of the lab we will continue to work with the Lego City and Friends dataset. Lego City sets are marketed more towards boys (`https://www.lego.com/en-us/themes/city/products`) and Lego Friends are marketed more towards girls (`https://www.lego.com/en-us/themes/friends/products`).

An article was published that suggested that Lego Friends sets are less complex than other comparable Lego sets (`https://momsla.com/why-my-daughters-wont-be-playing-with-lego-friends/`). They claimed that "it's clear to see how these Friends are dumbed down." In this activity, we will use the number of pieces per set as a measure of complexity in order to investigate their claim.

Research Question 2: Is there a difference in Lego Friends sets and Lego City sets regarding the Amazon price and the complexity in our sample?

Scatterplot with Several Regression Lines: We would like to create a scatterplot showing the least squares regression lines between price and number of pieces split by the Lego set theme.

- Click on *Graph*, *Graph Builder*.
- Drag your quantitative explanatory variable to the x-axis and the quantitative response variable to the y-axis.
- Drag your categorical grouping variable to the *Overlay* box on the top right.
- Click the small icon at the top that shows a small scatterplot with a line on it.
- Right click on the scatterplot. Go to *Line of Fit* and select *Equation*.

(a) Describe the relationship between the variables. Does one Lego theme tend to cost more than the other?

(b) Is the relationship for the two set themes similar between price and number of pieces (i.e., are the slopes of the lines similar)?

(c) Provide an answer to Research Question 2.

Consider a situation where the estimated slopes for Lego City and Lego Friends were identical, 0.13, the y-intercept for Lego City was 9.44, and the y-intercept for Lego Friends was 3.40. In this case we have the following two estimated models.

$\hat{y}_{City} = 9.44 + 0.13(\text{number of pieces})$
$\hat{y}_{Friends} = 3.40 + 0.13(\text{number of pieces})$

(d) Quantify the difference between the y-intercepts for the two lines. What does this difference tell you?

We would like to collapse these two estimated simple linear regression models into a single multiple linear regression model. For simplicity, assume the slopes from both models are the same.

We need to represent the categorical variable Theme as a number so it can be included in a mathematical equation. We will use the following coding scheme:

Theme 2 is coded as a $\underline{1}$ for a City set and coded as a $\underline{0}$ for a Friends set.

(e) Consider the coding scheme for Theme 2 and the difference in the y-intercepts. Think about how you could use the Theme 2 variable to combine the two simple linear regression models into a single model. Start by writing out the estimated Friends model shown above. Then, add to your model that includes a difference in y-intercepts that depends on Theme using the Theme 2 coding scheme.

Numerical Code of a Categorical Variable in JMP: To implement this model in JMP, we need to create the Theme 2 variable. Remember, this is the numeric representation of the Theme variable. To do this:

- Highlight the Column Theme.
- Click on *Cols*, *Recode*.
- Enter the coding scheme in the *New Values (2)* column.
- Click *Recode*. JMP will label this new column as Theme 2.
- Right click on your new column, Theme 2, and click *Column Info...*.
- Change *Data Type* to *Numeric* and *Modeling Type* to *Continuous*. Click *OK*.

In reality, sample slopes will not be exactly identical for each group. In this section we will use software to force identical slopes for the relationship between the Amazon price and the number of pieces for each Lego set theme. JMP will utilize formulas to optimize over all choices of coefficients to minimize the overall sum of squared errors.

Multiple Regression Model in JMP: We would like to create an estimated model using number of pieces per set and Lego set theme (City or Friends) to predict Amazon price. Here we assume identical slopes for each Lego set theme.

- Click on *Analyze*, *Fit Model*.
- Drag your response variable to the *Y* box. This should be the Price.
- Drag your explanatory variables to the *Construct Model Effects* box. These should include Number of Pieces and Theme 2.
- Click on *Run* on the right.
- Click on the red arrow next to *Response* on the top left, go to *Estimates*, and select *Show Prediction Expression*.

(f) What is the least squares regression equation?

(g) What do you think the coefficient in front of the variable Theme 2 tells us?

3. **Lego Sets of Different Sizes:** For this portion of the lab we will work with a dataset including both Lego sets with large bricks and Lego sets with small bricks. There are multiple Lego set themes and Lego brick sizes. Lego City and Lego Friends sets contain smaller bricks while Lego Duplo sets contain larger bricks (`https://www.lego.com/en-us/themes/duplo/products#section-3`). The Duplo sets are geared toward 1.5-5 year old children while the Lego City and Lego Friends sets are typically made for children 5 - 12 years old. The Lego City and Lego Friends sets have coordinating bricks that work across different sets. The variable labeled "Size" is recorded as "Large" for Duplo sets while Lego City and Lego Friends were recorded as "Small."

   Research Question 3: Does the relationship between the Amazon price and the number of pieces differ between small and large bricks in our sample?

   (a) Create a scatterplot displaying the relationship between price and number of pieces with regression lines for both sizes. Does it appear that one size of brick tends to cost more than the other? Explain.

   (b) Is the relationship similar between price and number of pieces (i.e., are the slopes of the lines similar)?

   (c) Based on your answer to the previous question, do you think it would be appropriate to fit a model like you fit to the Lego City and Lego Friends data? Explain.

   (d) What is the least squares regression equation for Small bricks?

   (e) What is the least squares regression equation for Large bricks?

Previously you created a coding scheme for City and Friends to “indicate” if the set was City or Friends. Now, let’s create an “indicator” variable for size of bricks. Create a new variable labeled Size 2, that will take the value of 1 if the set contains Large Lego bricks and the value of 0 if the set contains smaller Lego bricks.

We will now use JMP to create a single estimated model that will allow different slopes for Large brick and Small brick sets.

Modeling in JMP with Multiple Slopes:

- Create your indicator variable as defined above making sure to change the *Data Type* to *Numeric* and *Modeling Type* to *Continuous*.
- Click on *Analyze*, *Fit Model*.
- Put your quantitative response variable in the Y box at the top middle.
- *Add* both your quantitative explanatory variable and your categorical indicator variable (coded as 0’s and 1’s) in the *Construct Model Effects* box on the bottom.
- Click on your quantitative explanatory variable in the list of variables on the left side of the window. Hold down the Ctrl key and click on your categorical indicator variable (coded as 0’s and 1’s) so both variables are highlighted. Click on the *Cross* button under *Construct Model Effects*. You should now see three things in your *Construct Model Effects* box, each variable and the cross product between them.
- Click on the red arrow next to *Model Specification* on the top left and un-select *Center Polynomials*.
- Click on *Run* on the right.
- Click on the red arrow next to *Response* on the top left, go to *Estimates*, and select *Show Prediction Expression*.

(f) What is your estimated model? Write down the prediction expression.

(g) Suppose we use our new estimated model to describe the relationship between price and number of pieces for Small bricks. Simplify your estimated model by putting in 0 for your indicator variable. Make sure to show your work. How does this compare to your previous model for Small bricks in question 3d?

(h) Suppose we to use our new estimated model to describe the relationship between price and number of pieces for Large bricks. Simplify your estimated model by putting in 1 for your indicator variable. Make sure to show your work. How does this compare to your previous estimated model for Large bricks in question 3e?

(i) Consider the slope for Size 2 in the estimated model for question 3f. What is the purpose of the slope for Size 2? You may refer to the computations you went through in question 3h to answer this question.

(j) Consider the slope for (Pieces)(Size 2) in the estimated model for question 3f. What is the purpose of the slope for (Pieces)(Size 2)? You may refer to the computations you went through in question 3h to answer this question.

(k) Use your estimated model to predict the sale price of a set of large bricks that has 25 pieces.

(l) Provide an answer to Research Question 3.