# lego_land_lab

2025-10-17

Today's lab is going to be a little frustrating most likely. The point of this lab is to let you explore the concepts on your own and try to understand what is happening before being formally taught it.

And actually, I don't plan to teach this material in class nor test over it. I want to give you some introduction to MLR just so you are aware it exists.

Briefly, we will explore only legos of the small size variety/not Duplo theme'd. First we will read in the data, subset out Duplo and then look at the first few rows of data.

```
blocks <- read.csv('https://vinnys-classes.github.io/data/legos_data.csv')

blocks <- subset(blocks,
                 Theme != 'Duplo')
#The use of != means "not equal to" in R


head(blocks)
```

```
##    Item_Number                    Set_Name   Theme Pieces Year Pages
## 26       41330 Stephanie's Soccer Practice Friends    119 2018    48
## 27       41333    Olivia's Mission Vehicle Friends    223 2018    84
## 28       41335            Mia's Tree House Friends    351 2018   120
## 29       41340            Friendship House Friends    722 2018   164
## 30       41353     Friends Advent Calendar Friends    500 2018     4
## 31       41356       Stephanie's Heart Box Friends     85 2019    32
##    Minifigures Packaging Unique_Pieces  Size amazon_price age
## 26           1       Box            78 Small        40.40   6
## 27           1       Box           106 Small        45.95   6
## 28           2       Box           151 Small        53.88   6
## 29           3       Box           309 Small       184.99   6
## 30          NA       Box           202 Small        34.00   6
## 31           1      <NA>            36 Small        14.99   6
```

## The Problem

The following few questions build a linear model and does a deeper dive into checking residuals.

### Question 1

Using ggplot, please create a scatterplot with Pieces on the x-axis and amazon_price on the y-axis. Add a best-fit-line using geom_smooth.

**Question 2**

Copy your graph from question 1 down here. Color and shape the points by Theme.

**Question 3**

Comment on the differences between the the city and friends group of legos.

**Question 4**

Fit a linear model using the lm() function with amazon_price as the response and Pieces as the explanatory variable.

**Question 5**

Plot the residuals from question 3. Color and shape the points by Theme.

**Question 6**

Do you believe the normality and homoskedasticity assumptions are met? Do you think all points are “identically distributed” (eg no pattern regardless of theme or location on the graph)

**Question 7**

Copy down your answer to question 2. Add a best fit line via geom_smooth. Also, add an aes() within geom_smooth() like you do for geom_point() or ggplot(). Set the parameter called “group” equal to Theme. That is…

```
#geom_smooth(method = "lm",
#            aes(group = Theme))
```

SIDE NOTE: You can actually color and shape your lines as well in the aes() function here using color = Theme and linetype = Theme. Try it to see what happens.

**Question 8**

Calculate the mean amazon_price for both Themes by using the aggregate() function. See the below hashtaged code for the format of the function

```
#aggregate(Response  ~ Explanatory_var,
 #         data = your_data_goes_here,
#          FUN = the_function_you_want)
```

## A Step Up

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.

$\hat{y}_{city} = 9.44 + .13$ * Number of Pieces $\hat{y}_{friends} = 3.40 + .13$ * Number of Pieces

### Question 9

Find the difference between the two y-intercepts. What does this number represent?

### Question 10

Think about how you could use the Theme variable to put the two linear regressions together into a single model (formula). Try to write it out using the estimates given right before question 9.

HINT: You'll need to use an indicator variable to do this

### Question 11

In R we can represent the equation that you should hopefully have found in question 10 by putting into the lm() function 2 different explanatory variables.

Do this by replacing Expalantory_var_1 and 2 in the below code with Pieces and Theme. Also replace the response with amazon_price and the data with whatever you named your data set (blocks most likely). Remove the hashtags as well.

```
#new_mod <- lm(RESPONSE  ~ EXPLANATORY_VAR_1 +
#                          EXPLANATORY_VAR_2,
#               data = YOUR_DATA)
```

### Question 12

Using the summary() function, write down the estimated regression equation

### Question 13

What does the -7.32 next to ThemeFriends mean in the Coefficients table?

## Many Steps Up

We are now reading in our data again but this time NOT removing the duplo sets. Note the data is now being saved as "legos". Also, we are no long focusing on the difference between the three themes and are instead focusing on the size of the lego bricks.

```
legos <- read.csv('https://vinnys-classes.github.io/data/legos_data.csv')

head(legos)
```

```
##   Item_Number                   Set_Name Theme Pieces Year Pages Minifigures
## 1       10859          My First Ladybird Duplo      6 2018     9          NA
## 2       10860          My First Race Car Duplo      6 2018     9          NA
## 3       10862       My First Celebration Duplo     41 2018     9          NA
## 4       10864 Large Playground Brick Box Duplo     71 2018    32           2
## 5       10867            Farmers' Market Duplo     26 2018     9           3
## 6       10870               Farm Animals Duplo     16 2018     8          NA
##     Packaging Unique_Pieces  Size amazon_price age
## 1         Box            5 Large        16.00   1
## 2         Box            6 Large         9.45   1
## 3         Box           18 Large        39.89   1
## 4 Plastic box           49 Large        56.69   2
## 5         Box           18 Large        36.99   2
## 6         Box           13 Large         9.99   2
```

### Question 14

Plot the amazon_price by Pieces and color by the Size. Include two best-fit-lines like in question 7.

### Question 15

Do the slopes of the two lines look similar? If not, how so?

### Question 16

Based on your previous answers, would it be enough to add an indicator variable like we did in Question 11? Why or why not?

### Question 17

Use the subset() function to split our data set into two. One will be large bricks and the other will be small bricks. To do this you will need to create two new data sets. I did the first for you, now make the second for Small bricks.

```
data_large <- subset(legos,
                 Size == "Large")
```

### Question 18

Run a linear model using the lm() function for both data sets with amazon_price as the response and Pieces as the explanatory variable. Write out the two estimated linear regression equations (one for large bricks, and one for small bricks).

### Question 19

Using indicator variables, how could you write out a linear model that would allow you to account for two different slopes? Use $\hat{\beta}$'s and not numbers. HINT: You will again need an indicator variable for this.

### Question 20

Run one more linear model similar to question 11. This time run the code similar to before where between the two explanatory variables there is a * and not a +

Write out the estimated linear equation

```
#my_newest_mod <- lm(RESPONSE  ~ EXPLANATORY_VAR_1 *
#                           EXPLANATORY_VAR_2,
#              data = YOUR_DATA)
```

### Question 21

Predict the price for a lego set with 55 pieces using your above equation

### Question 22

Locate the Monster Truck lego set in the data (row 71, you can use the function tail(legos) to see it actually). Using your rpediction in question 21 and the price of this lego set calculate a residual.