Linear Regression – Categorical Predictors

Grinnell College

Feburary, 2026

Review

$$\hat{y} = \beta_0 + \beta_1 X$$

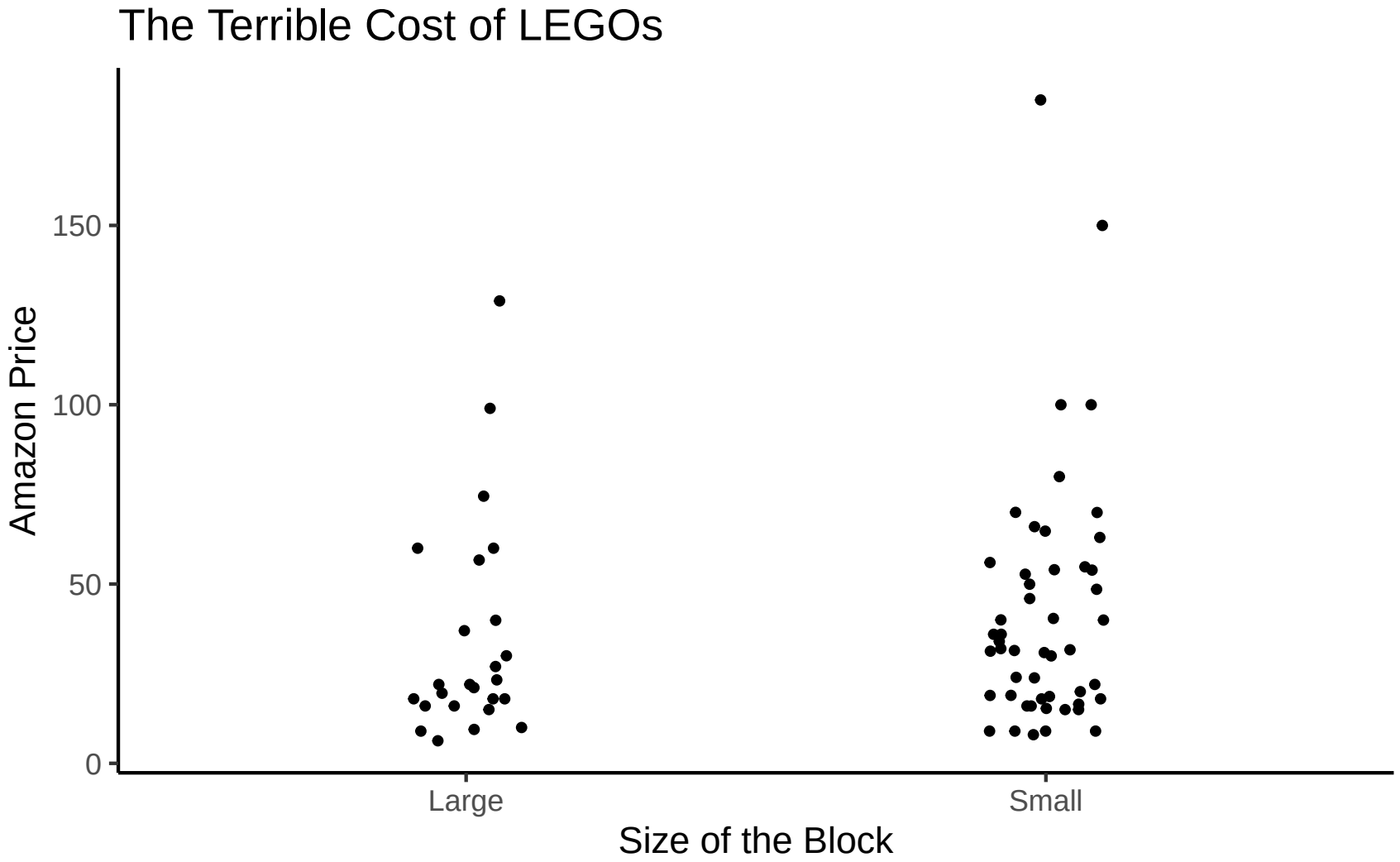
Linear Regression so far:

- We replace both β's with $\hat{\beta}$
- Both response and explanatory variable have been numeric
- Only works when there is a *linear* relationship
- There are formulas for slope and intercept (use R!)
- Use line to make predictions
- Interpret the slope and intercept (if applicable)
- R^2 and r

What if my explanatory variable was categorical?

How would you make a guess for a category?

What if my explanatory variable was categorical? How would you make a guess for a category?



Goals

What we want in our model:

- Each category gets a mean (or median for a log transformed response)
- WE HAVE ALREADY DONE THIS!!
  - Aggregate() to find the mean of a response variable for each category
- Need way to do that but "math-y" with equations and stuff
- We need to transform the categories into numbers somehow
  - We can't say category A is 1, category B is 2, etc....
  - That implies B is twice as much as A
- Indicator Variables to the rescue!

Indicator Variables

Indicator Variables: are a variable we create that indicates whether an observation belongs to a specific category (1) or not (0)

- Each category gets it's own indicator variable
- Sometimes called 'Dummy variables'; Machine Learning (AI?) call it "One Hot" encoding
- 1 indicates an obs. is in the category, 0 indicates otherwise

Set Name	Block Size	Large Block	Small Block
Farmer's Market	Large	1	0
Puppy Playground	Small	0	1
Police Monster Truck Heist	Small	0	1
Baby Animals	Large	1	0

(The two indicators are on the right hand side)

Indicator Variables

Indicator Variables are often denoted with a stylistic "1" and a subscript to denote the category, $\mathbb{1}_{\text{CATEGORY HERE}}$.

Set Name	Block Size	Large Block	Small Block
Farmer's Market	Large	1	0
Puppy Playground	Small	0	1
Police Monster Truck Heist	Small	0	1
Baby Animals	Large	1	0

$$\mathbb{1}_{\text{Large}} = \begin{cases} 1 & \text{if Large} \\ 0 & \text{if Small} \end{cases}$$

$$\mathbb{1}_{\text{Small}} = \begin{cases} 0 & \text{if Large} \\ 1 & \text{if Small} \end{cases}$$

For like 3 years I thought it was an uppercase "i" (for indicator)

So how would we leverage indicator variables to change our linear regression equation?

So how would we leverage indicator variables to write out our linear regression equation? 2 ways!

$$y = \alpha_0 * \mathbb{1}_{Large} + \alpha_1 * \mathbb{1}_{Small} + e \quad (1)$$

The α are the (population) means for their respective categories

---

$$y = \beta_0 + \beta_1 * \mathbb{1}_{Small} + e \quad (2)$$

- β_0 is the mean of our **baseline** or reference variable.
  - Easiest way to identify is it's the category not mentioned
- β_1 is the difference between the means of the indicator's group/category and the baseline's group/category
- R uses this one, first level/category is the baseline by default

One to the Other

The first model is intuitive to understand, the coefficient of the indicator is the group's mean. Second is more nuanced. Below is the relevant relationships.

$$\begin{aligned} \beta_0 &= \alpha_0 \\ \beta_1 &= \alpha_1 - \alpha_0 \end{aligned}$$

$$\begin{aligned} \alpha_0 &= \beta_0 \\ \alpha_1 &= \beta_0 + \beta_1 \end{aligned}$$

As almost always we don't know these actual values so we use our estimated values instead, again using the ^ symbol

Estimated Linear Model via R

```
> my_mod_size <- lm(amazon_price ~ Size, data = legos)
> summary(my_mod_size)

Call:
lm(formula = amazon_price ~ Size, data = legos)

Residuals:
   Min     1Q Median     3Q    Max
-33.97 -23.02 -10.69  11.98 143.03

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)   34.262      6.665   5.140 2.21e-06 ***
SizeSmall      7.697      8.163   0.943    0.349
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.33 on 73 degrees of freedom
Multiple R-squared:  0.01203,   Adjusted R-squared:  -0.001502
F-statistic: 0.889 on 1 and 73 DF,  p-value: 0.3488
```

$$\widehat{Amazon\ Price} = 34.262 + 7.697 * \mathbb{I}_{Small}$$

Practice

$$\widehat{Amazon\ Price} = 34.262 \ + \ 7.697 * \mathbb{I}_{Small}$$

- 34.262 is the estimated mean of the Large group
  - $\hat{\beta}_0$ $(= \hat{\alpha_0})$
- 7.697 is the estimated difference in the means from the Small to the Large group
  - $\hat{\beta}_1$ $(= \hat{\alpha_1}$ - $\hat{\alpha_0})$
- Predict the price of an amazon LEGO set with small blocks

Practice

$$\widehat{Amazon\ Price} = 34.262 \ + \ 7.697 * \mathbb{I}_{Small}$$

- 34.262 is the estimated mean of the Large group
  - $\hat{\beta}_0$ $(= \hat{\alpha}_0)$
- 7.697 is the estimated difference in the means from the Small to the Large group
  - $\hat{\beta}_1$ $(= \hat{\alpha}_1$ - $\hat{\alpha}_0)$
- Predict the price of an amazon LEGO set with small blocks
  - $\hat{\beta}_0 + \hat{\beta}_1$ *1
  - 34.262 + 7.697
  - 41.959

What about more than 2 categories?

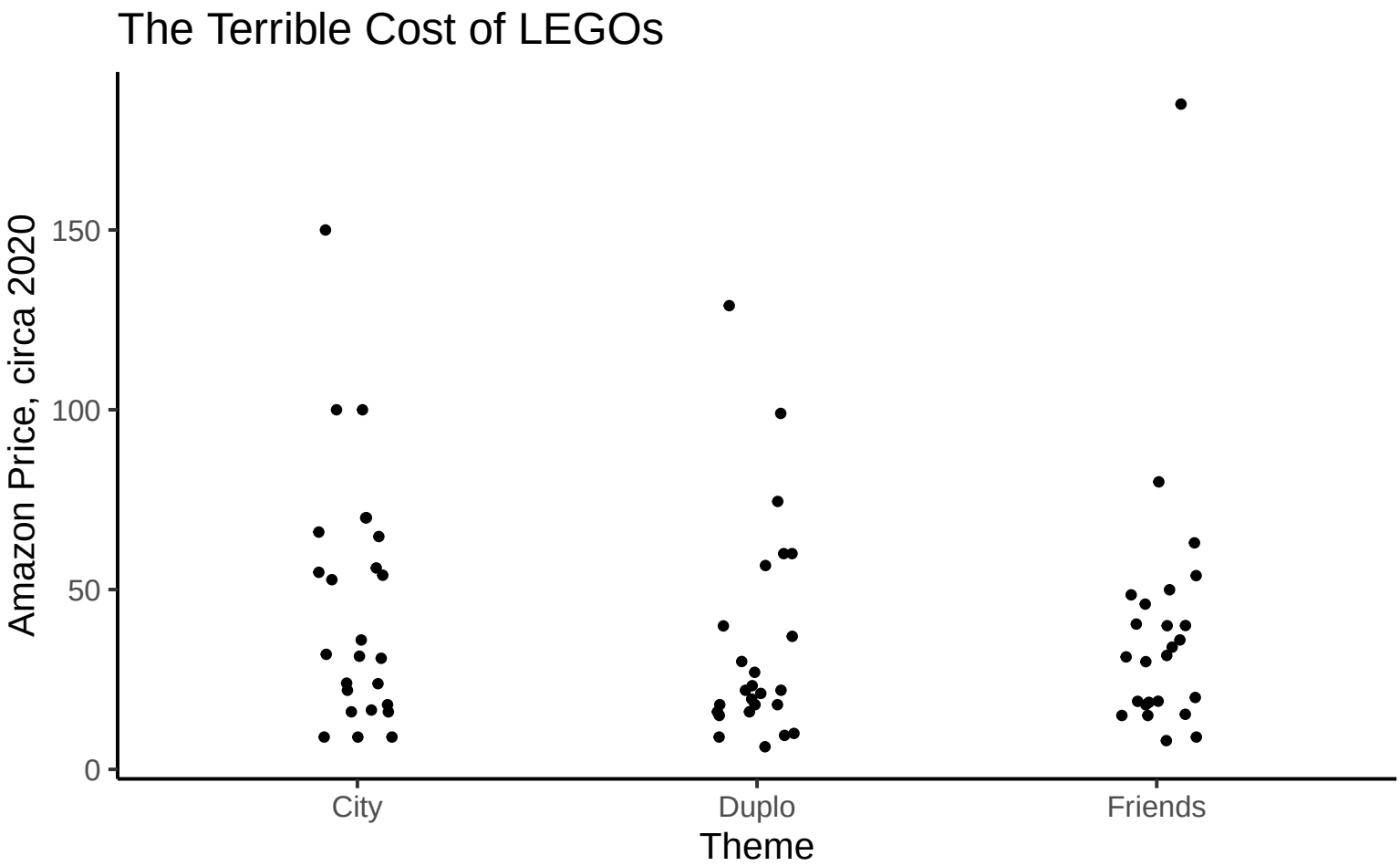
Set Name	Theme
Farmer's Market	Duplo
Puppy Playground	Friends
Police Monster Truck Heist	City
Baby Animals	Duplo

What about more than 2 categories?

Hadley would view indicator variables as making the data table "wide"

Set Name	Theme	Duplo	Friends	City
Farmer's Market	Duplo	1	0	0
Puppy Playground	Friends	0	1	0
Police Monster Truck Heist	City	0	0	1
Baby Animals	Duplo	1	0	0

And again, we would still guess the means.


The Terrible Cost of LEGOs
Amazon Price, circa 2020
150
100
50
0
City
Duplo
Friends
Theme

Model

```
> my_mod <- lm(amazon_price ~ Theme, data = legos)
> summary(my_mod)

Call:
lm(formula = amazon_price ~ Theme, data = legos)

Residuals:
   Min     1Q Median     3Q    Max
-36.28 -20.98 -10.99  10.31 146.34

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    45.270      6.689   6.768 2.95e-09 ***
ThemeDuplo    -11.007      9.459  -1.164    0.248
ThemeFriends   -6.620      9.459  -0.700    0.486
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 33.44 on 72 degrees of freedom
Multiple R-squared:  0.01871,   Adjusted R-squared:  -0.00855
F-statistic: 0.6863 on 2 and 72 DF,  p-value: 0.5067
```

$$\text{Predicted Amazon Price} = 45.27 - 11.007\mathbb{1}_{DUPLO} - 6.620\mathbb{1}_{Friends}$$