Linear Regression – Categorical Predictors

Grinnell College

October 6, 2025

Review

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

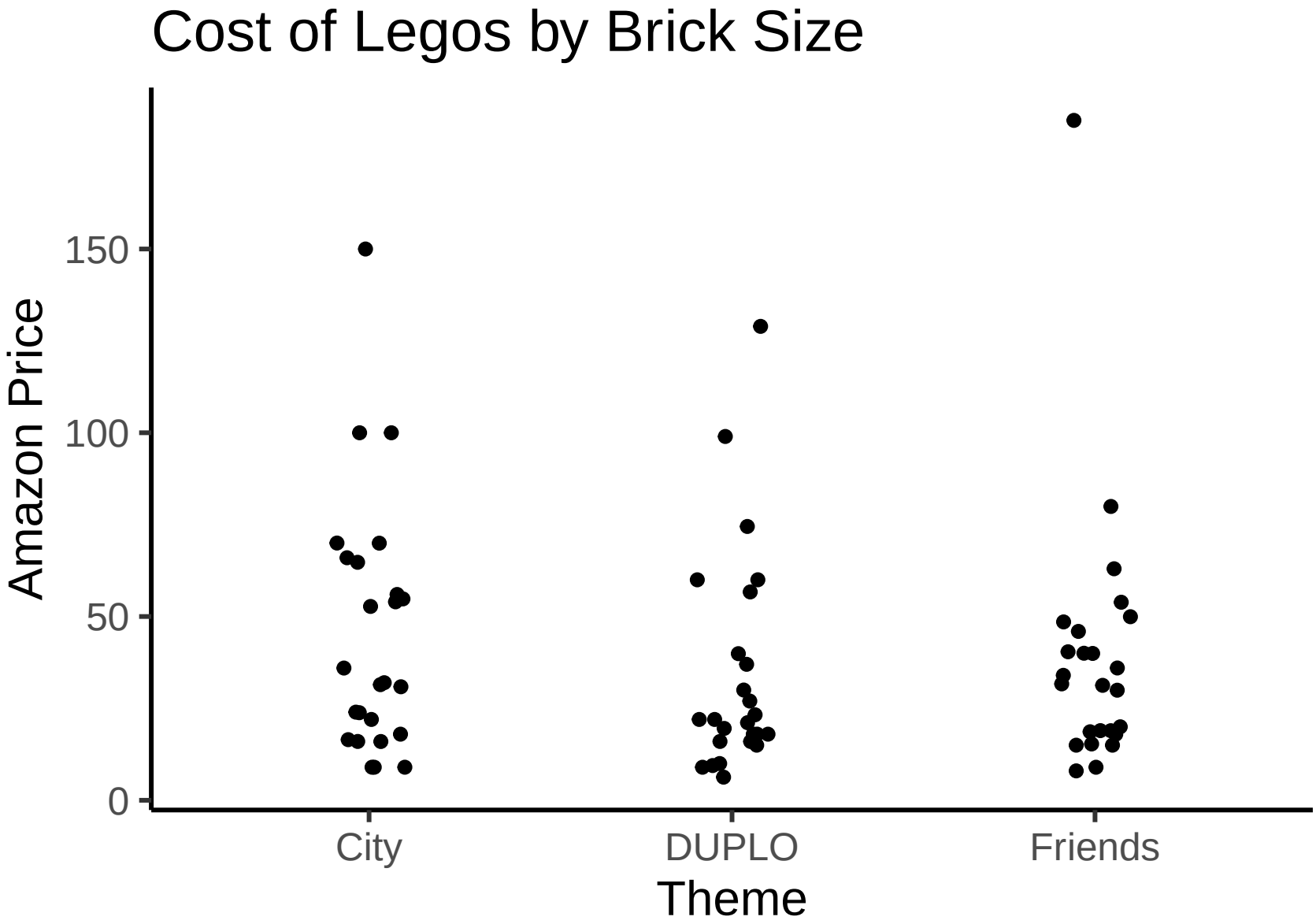
Linear Regression so far:

- 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 log(y))
- 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 just 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 new variable we make that indicates whether an observation belongs to a specific category or not

- Denoted with a $\mathbb{1}_{\text{CATEGORY HERE}}$
- Each category gets it's own indicator variable
- Sometimes called 'Dummy variables'
- "Hot-One" encoding in computer science and machine learning
- 1 indicates an obs. is in the category, 0 indicates otherwise

Model	Trans
audi a4	auto
audi a4	manual
chevrolet c1500	auto
dodge pickup 4wd	auto
ford explorer 4wd	manual
hyundai sonata	auto

Model	Manual	Auto
audi a4	0	1
audi a4	1	0
chevrolet c1500	0	1
dodge pickup 4wd	0	1
ford explorer 4wd	1	0
hyundai sonata	0	1

Indicator Variables

Indicator Variables are often denoted with a stylistic "1" and a subscript to denote the category

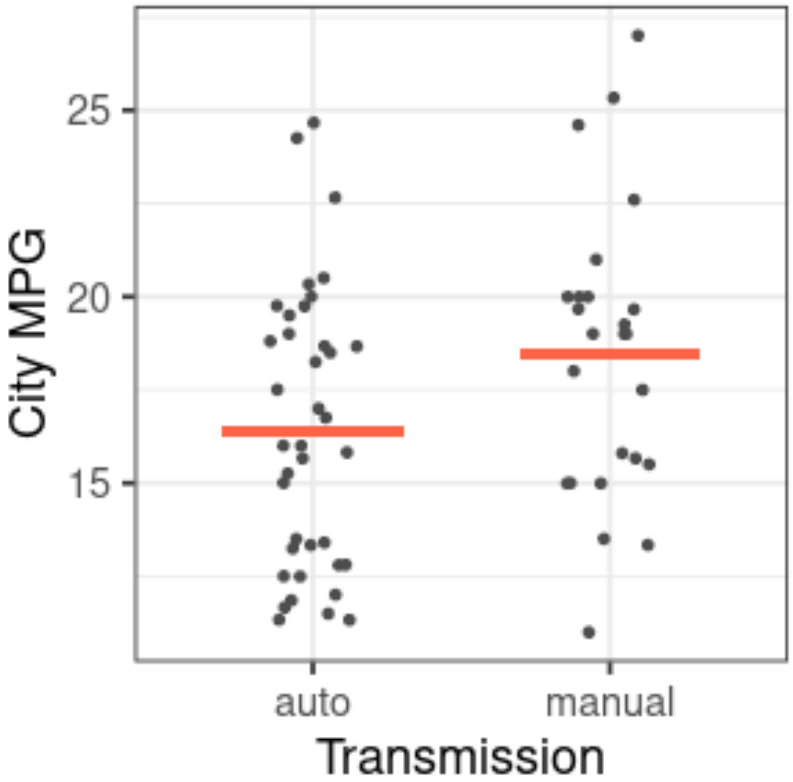
Model	Manual	Auto
audi a4	0	1
audi a4	1	0
chevrolet c1500	0	1
dodge pickup 4wd	0	1
ford explorer 4wd	1	0
hyundai sonata	0	1

$$\mathbb{1}_{\text{Manual}} = \begin{cases} 1 & \text{if Manual} \\ 0 & \text{if Automatic} \end{cases}$$

$$\mathbb{1}_{\text{Automatic}} = \begin{cases} 1 & \text{if Automatic} \\ 0 & \text{if Manual} \end{cases}$$

Indicator Variables

Maybe we can make predictions for groups using their averages?



Model	Manual	Auto	cty
audi a4	0	1	18.250
audi a4	1	0	19.667
chevy c1500	0	1	12.800
dodge pickup	0	1	12.500
ford explorer	1	0	15.000
hyundai sonata	0	1	19.000

Transmission	Average City MPG
auto	16.370
manual	18.457

$$\widehat{\text{City mpg}} = 16.370 \times \mathbb{1}_{\text{Automatic}} + 18.457 \times \mathbb{1}_{\text{Manual}}$$

Linear Model in R

By default, the first indicator will be absorbed into the intercept, making it the *reference variable*

```
> lm(cty ~ trans, mpg2)

Coefficients:
(Intercept)  transmanual
      16.37          2.09
```

Compare equations:

$$\widehat{\text{City mpg}} = 16.37 \times \mathbb{1}_{\text{Automatic}} + 18.457 \times \mathbb{1}_{\text{Manual}}$$

$$\widehat{\text{City mpg}} = 16.37 + 2.09 \times \mathbb{1}_{\text{Manual}}$$

Practice

More than 2 categories?!
What are my indicator variables going to look like?

model	cty	drv
new beetle	21	f
gti	19	f
mustang	18	r
grand cherokee 4wd	11	4
sonata	21	f
civic	24	f
toyota tacoma 4wd	15	4

Categories of 'drv': 4-wheel drive (4), rear-wheel drive (r), front-wheel drive (f)

Practice

Categories of 'drv': 4-wheel drive (4), rear-wheel drive (r), front-wheel drive (f)

What are my indicator variables going to look like?

model	cty	drv
new beetle	21	f
gti	19	f
mustang	18	r
grand cherokee	11	4
sonata	21	f
civic	24	f
toyota tacoma	15	4

model	cty	drvf	drvr	drv4
new beetle	21	1	0	0
gti	19	1	0	0
mustang	18	0	1	0
grand cherokee	11	0	0	1
sonata	21	1	0	0
civic	24	1	0	0
toyota tacoma	15	0	0	1

Practice

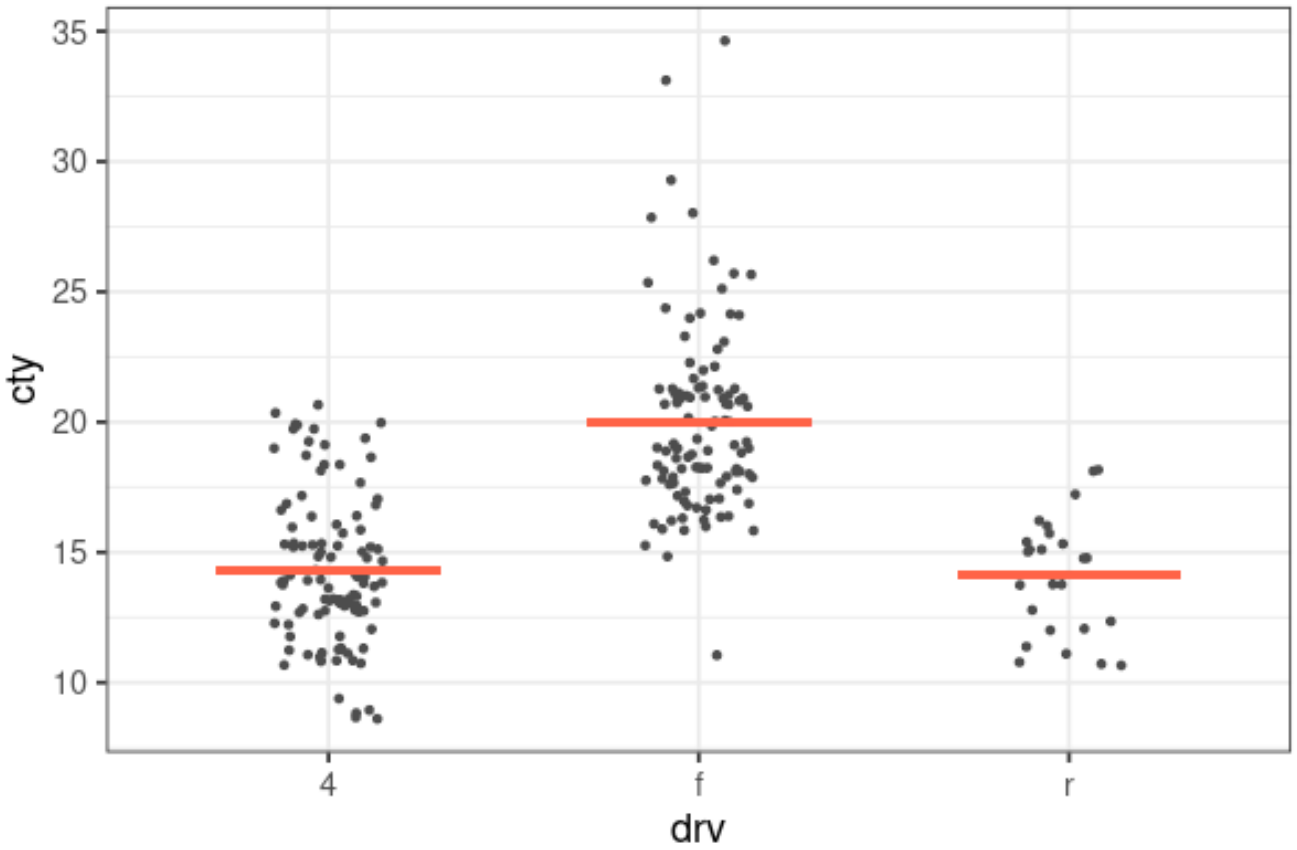
Categories of 'drv': 4-wheel drive (4), rear-wheel drive (r), front-wheel drive (f)

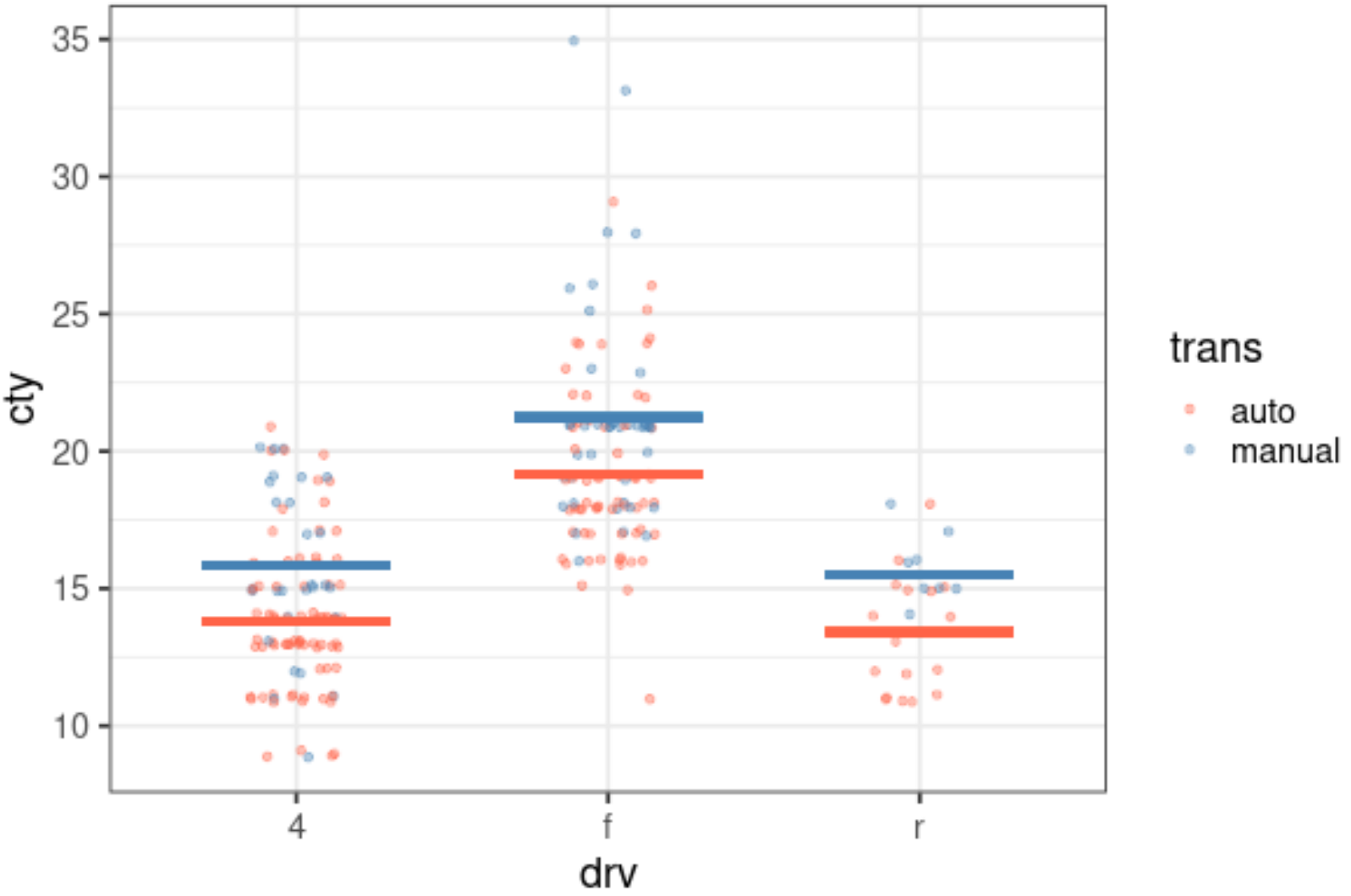
```
> lm(cty ~ drv, mpg)

Coefficients:
(Intercept)         drvf         drvr
      14.33         5.64        -0.25
```

- What is the *reference variable*
- Equation for line?
- Interpretation of intercept? Slope?
- What is the average city mileage for:
  - 4-wheel drive?
  - Front-wheel drive?
  - Rear-wheel drive?

Practice

```
> lm(cty ~ drv, mpg)

Coefficients:
(Intercept)          drvf          drvr
      14.33          5.64         -0.25
```




35
30
25
20
15
10
cty
4
f
r
drv
trans
auto
manual