Linear Regression

Categorical Predictors and Multiple Variables

Grinnell College

Review

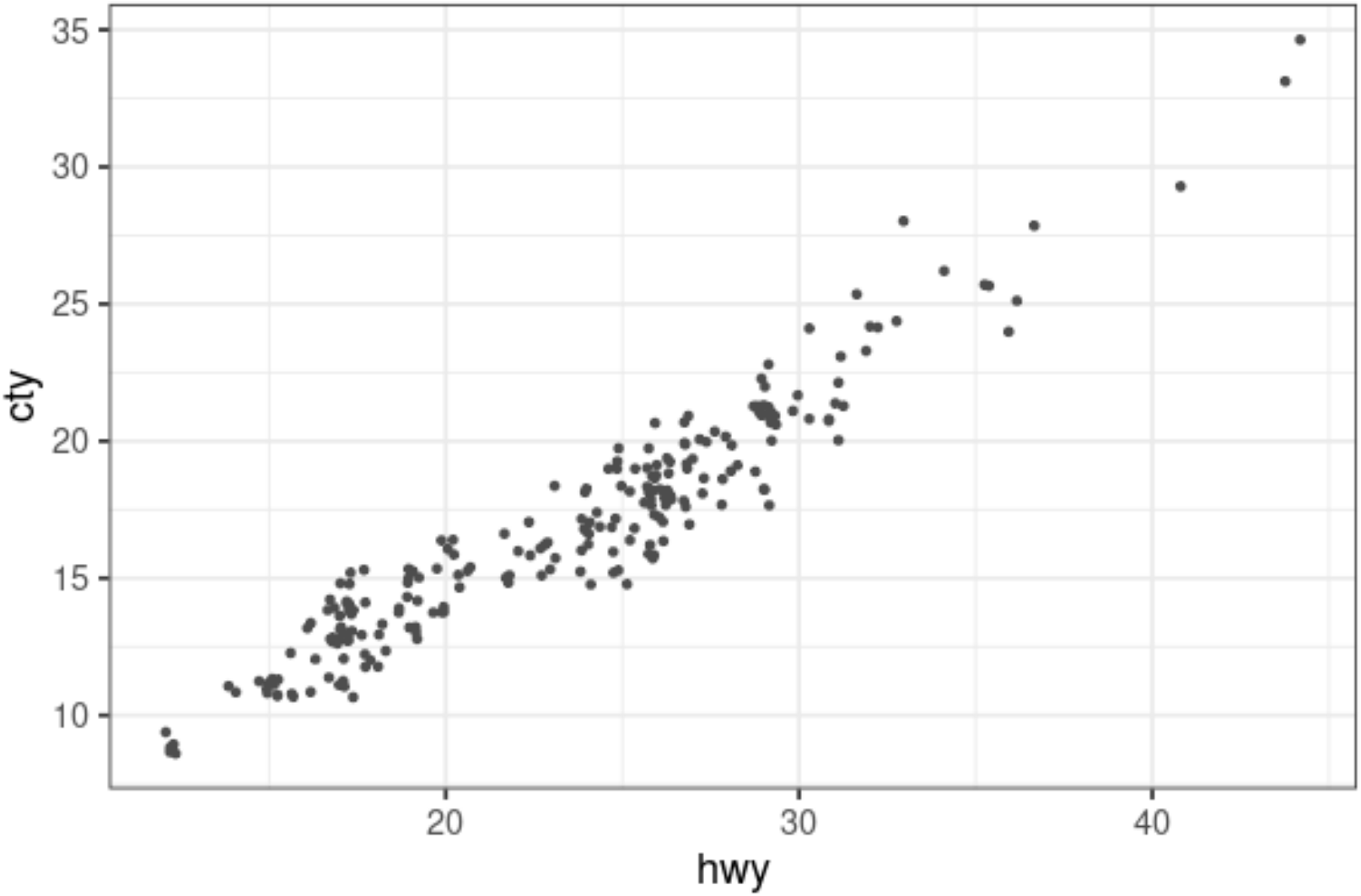
$$\hat{y} = b_0 + b_1 X$$

We spent some time talking about linear regression. Basically a fancy way of putting a line on a scatterplot to describe the relationship between variables.

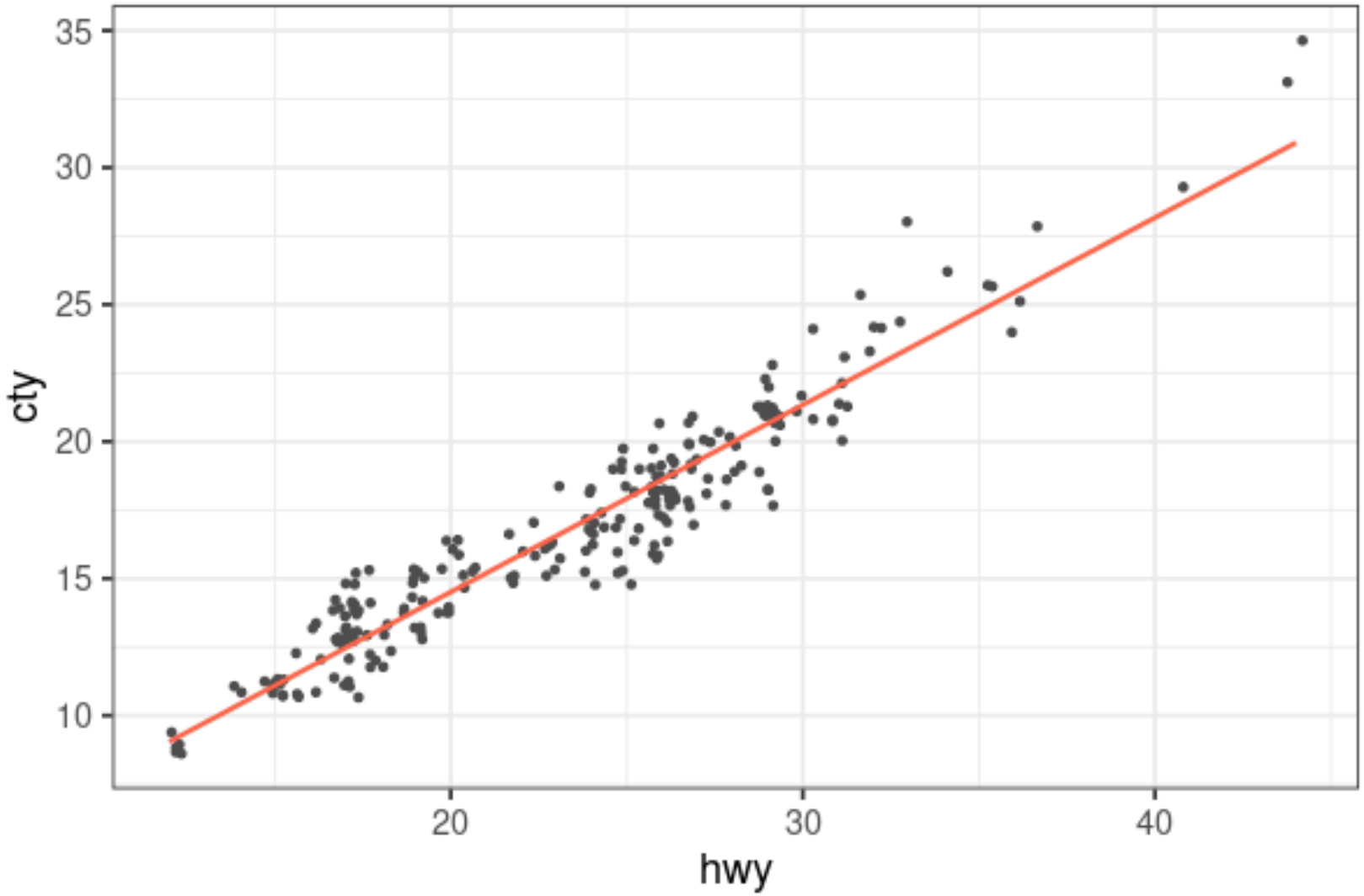
- Only works when there is a *linear* relationship
- There are formulas for slope and intercept
- Use line to make predictions
- Interpret the slope and intercept (if applicable)
- R^2 and r

Review

('mpg' dataset)
Highway miles per gallon vs. City miles per gallon for vehicles



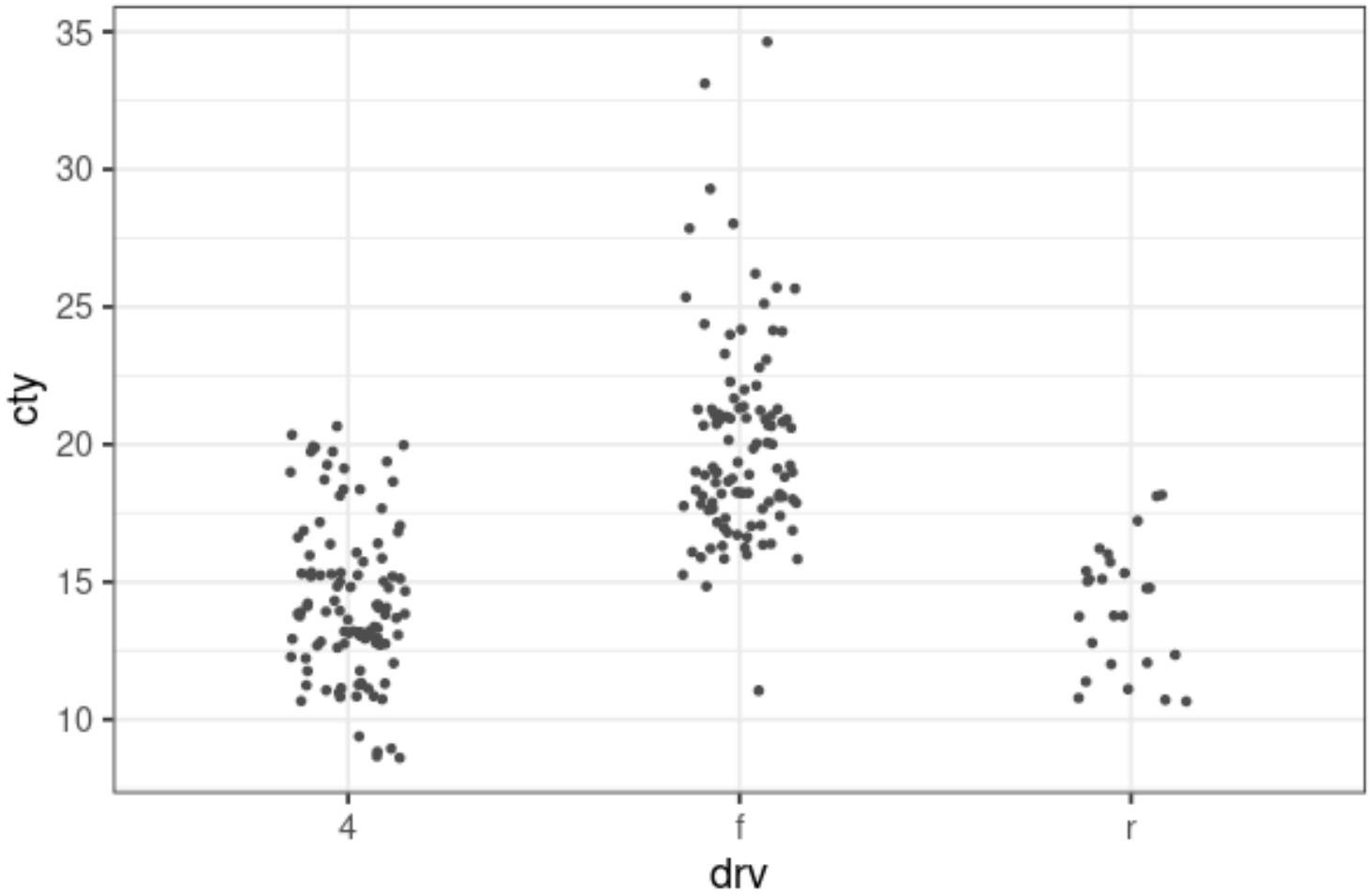
$$\widehat{\text{City mpg}} = 0.844 + (0.683 \times \text{Highway mpg})$$



Categorical predictor?

What if my explanatory variable was categorical? Can we use linear regression?

$$\hat{y} = \ldots$$



Indicator Variables

Let's look at how the data is stored in the data frame

Model	Transmission
audi a4	auto
audi a4	manual
chevrolet c1500 suburban 2wd	auto
dodge dakota pickup 4wd	auto
ford explorer 4wd	manual
hyundai sonata	auto

How might these be used in regression?

Indicator Variables

Indicator Variables: are a new variable we make that **indicates** whether an observation belongs to a specific category or not

- sometimes called 'Dummy variables' / "One-Hot encoding"
- 1 indicates an observation is in the category
- 0 indicates an observation is **not** in the category

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 original variable name

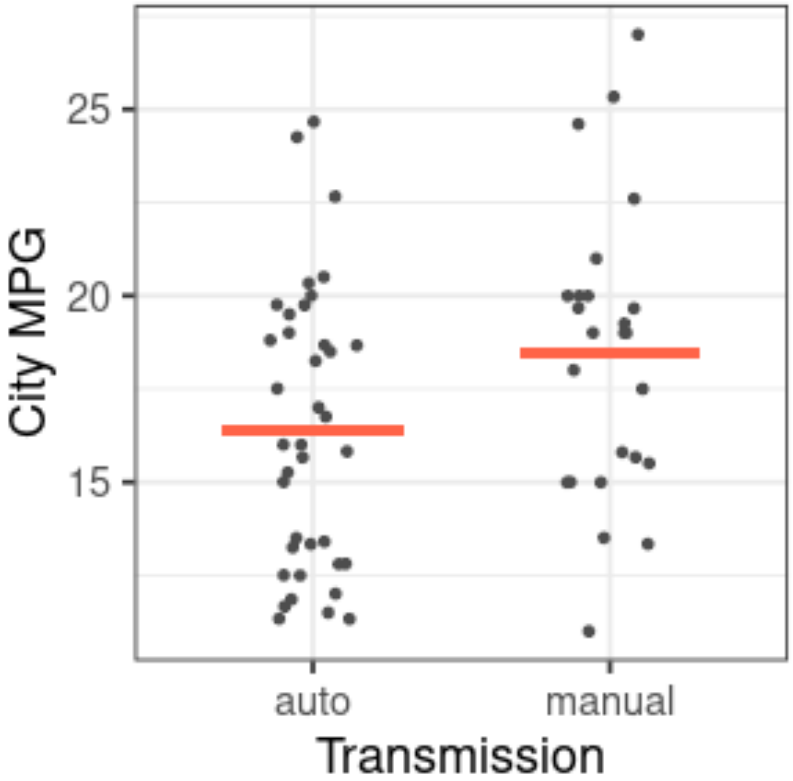
Model	Manual	Auto
audi a4	0	1
audi a4	1	0
chevrolet c1500	0	1
dodge pickup 4wd	0	1
ford explorer 4wd	1	0
hyundai sonata	0	1

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

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

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

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

Reference Group

Some statistical software will try to make the equation look similar to what we've seen before by only wanting to use 1 variable (not both)

- the slope for one group turns into the intercept (reference group)
- the *difference* between the variables will become the slope

Compare equations:

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

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

$$\widehat{\text{City mpg}} = 18.457 - 2.09 \times \mathbb{1}_{\text{Automatic}}$$

Multiple Groups

We can apply this to more than 2 categories as well.

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

Multiple Groups

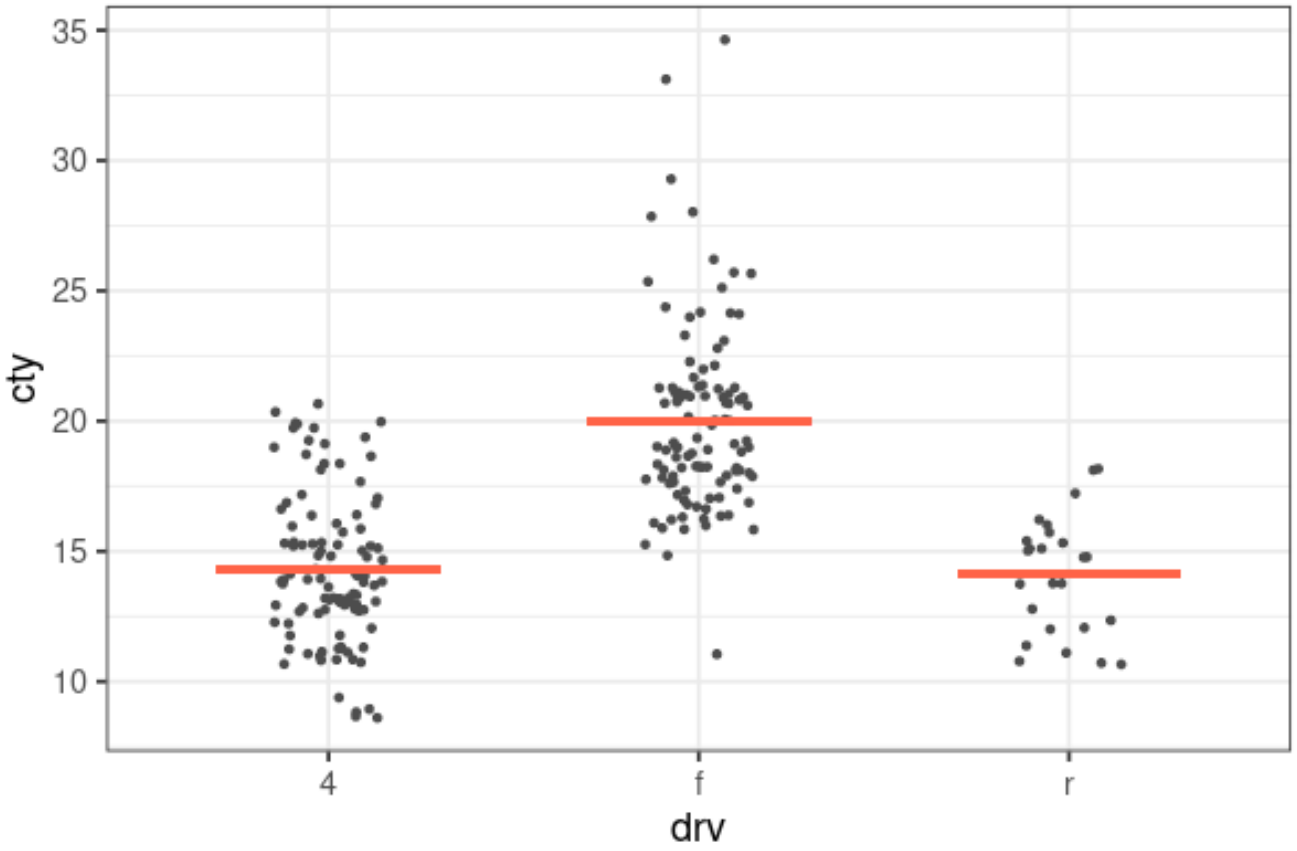
Categories of 'drv': 4-wheel drive (4), rear-wheel drive (r), front-wheel drive (f)

$$\widehat{\text{City mpg}} = 14.33 + 5.64 \times \mathbb{1}_{\text{Front}} - .25 \times \mathbb{1}_{\text{Rear}}$$

- What is the *reference group*?
- Interpretation of intercept?
- Interpretation of slopes?
- What is the average city mileage for:
  - 4-wheel drive?
  - Front-wheel drive?
  - Rear-wheel drive?

Practice

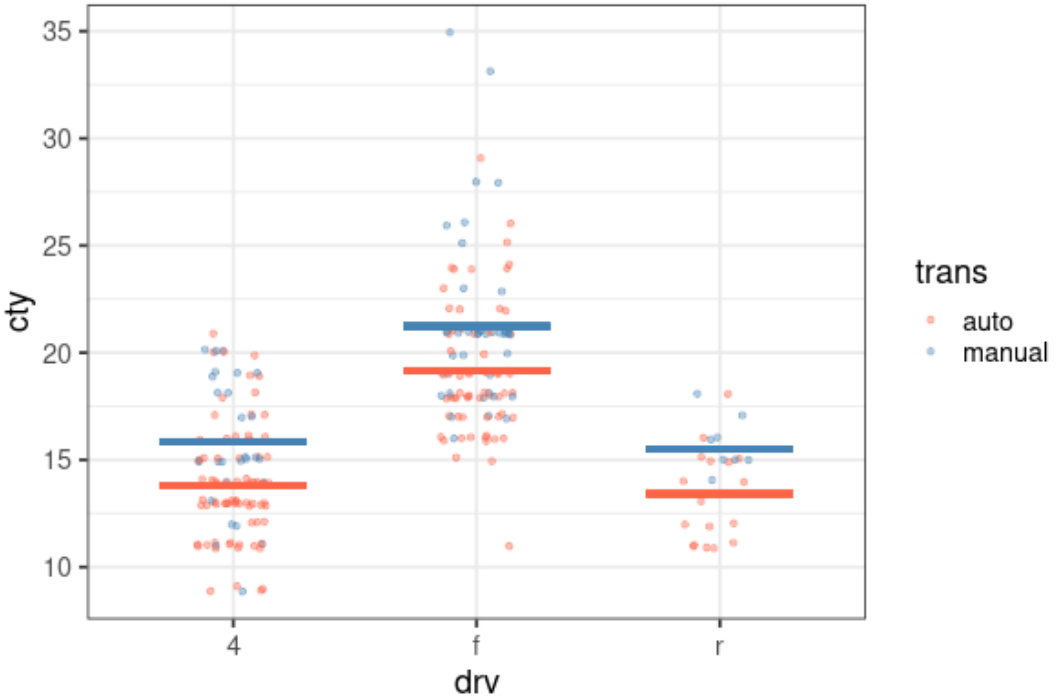
$$\widehat{\text{City mpg}} = 14.33 + 5.64 \times \mathbb{1}_{\text{Front}} - .25 \times \mathbb{1}_{\text{Rear}}$$



Extending to Multiple Variables

Here we have the average city miles per gallon for each combination of drive train and transmission

Type	4wd	fwd	rwd
Automatic	13.85	19.11	13.29
Manual	15.61	21.34	15.75



Multiple Variables

$$\widehat{\text{City mpg}} = 13.77 + 5.40 \times \mathbb{1}_{\text{Front}} - .35 \times \mathbb{1}_{\text{Rear}} + 2.07 \times \mathbb{1}_{\text{Manual}}$$

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

Observed vs Predicted Means

When we have more than 2 variables, the regression forces the slope for manual transmission to be the same for all groups and so predictions are no longer exactly the same as group means

Observed group means:

Transmission	4wd	fwd	rwd
Automatic	13.85	19.11	13.29
Manual	15.61	21.34	15.75

Predicted mileage for each group:

Transmission	4wd	fwd	rwd
Automatic	13.76	19.17	13.42
Manual	15.83	21.24	15.49