SLR – Categorical Predictors

Grinnell College

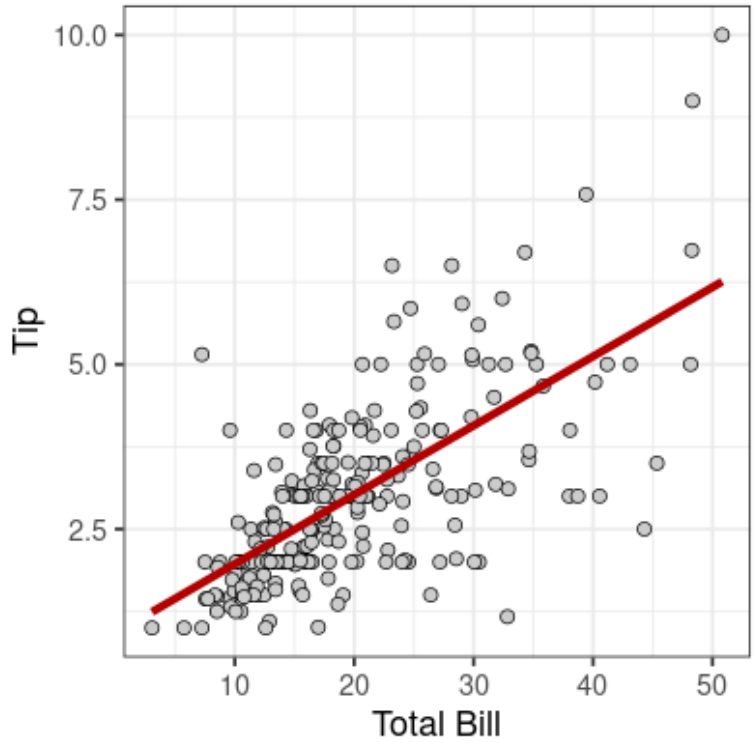
February 14, 2025

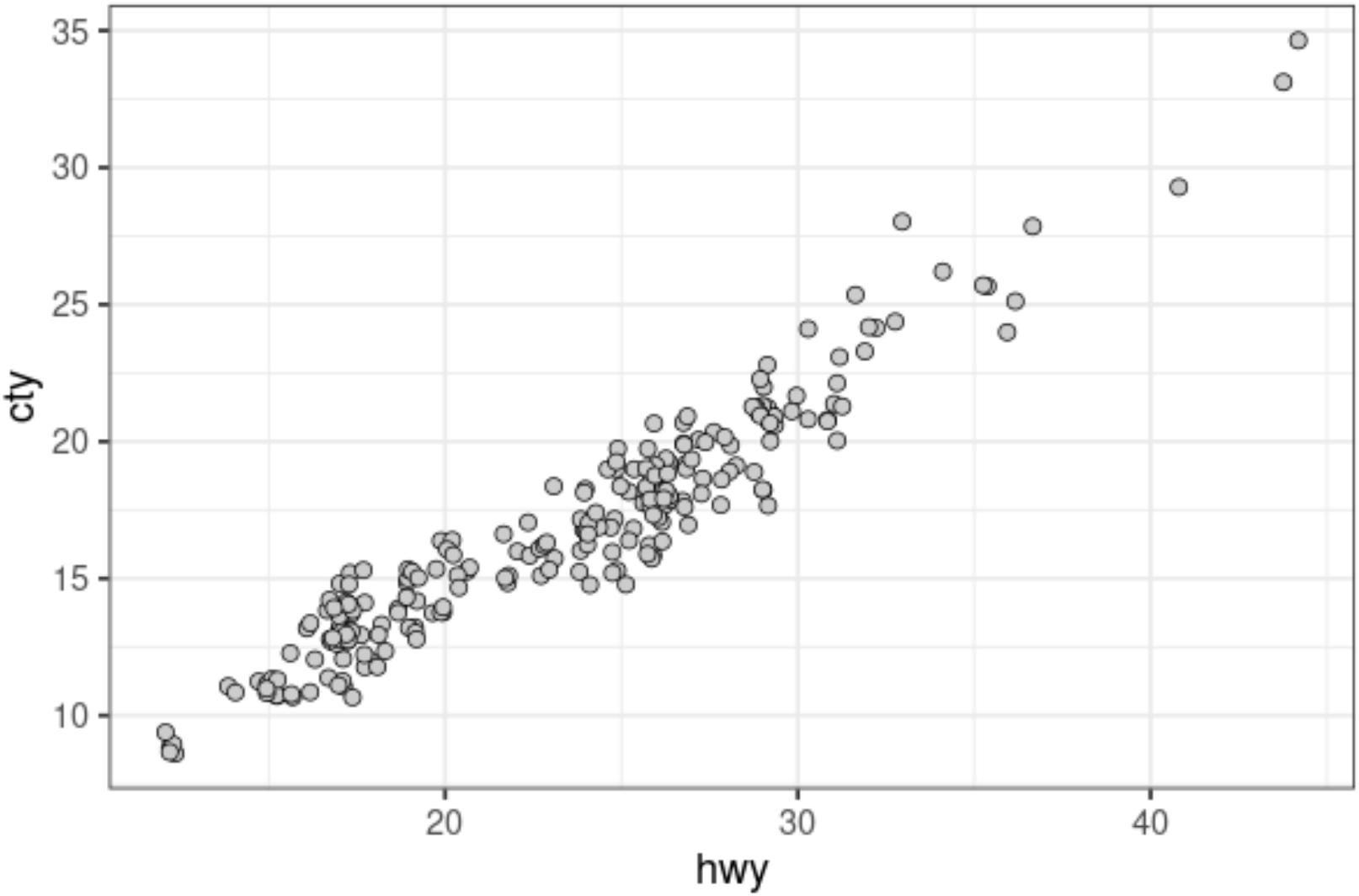
Review

Suppose the relationship between tip amount and total bill can be described as

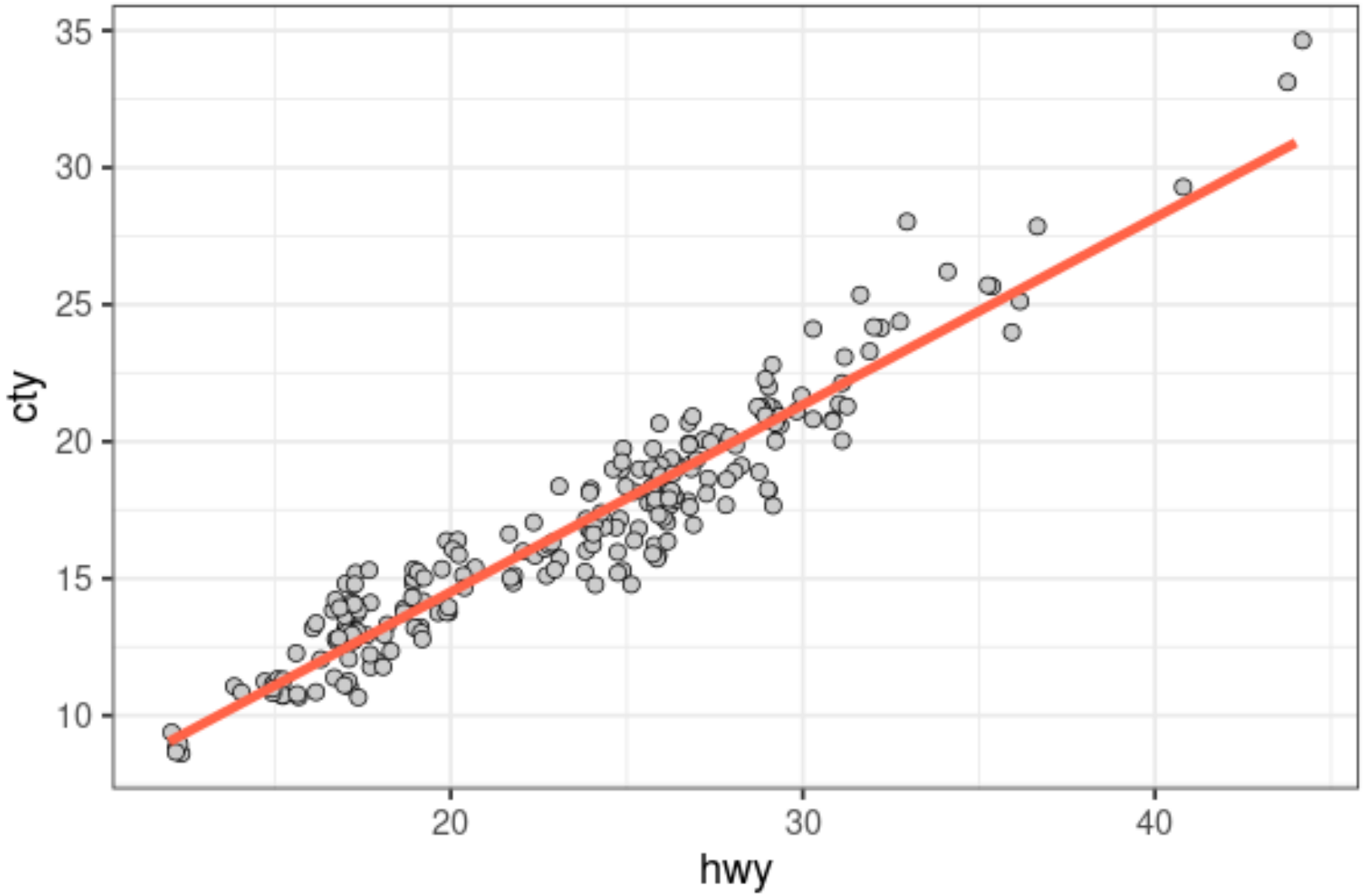
$$\widehat{\text{Tip}} = 0.92 + 0.105 \times \text{Total Bill}$$

- What is my explanatory variable and what is my response?
- Which term best describes how Tip changes in response to Total Bill?
- Interpret the intercept. Is this meaningful?
- How would you describe the quality of the fit? What is the metric you would use to quantify it?

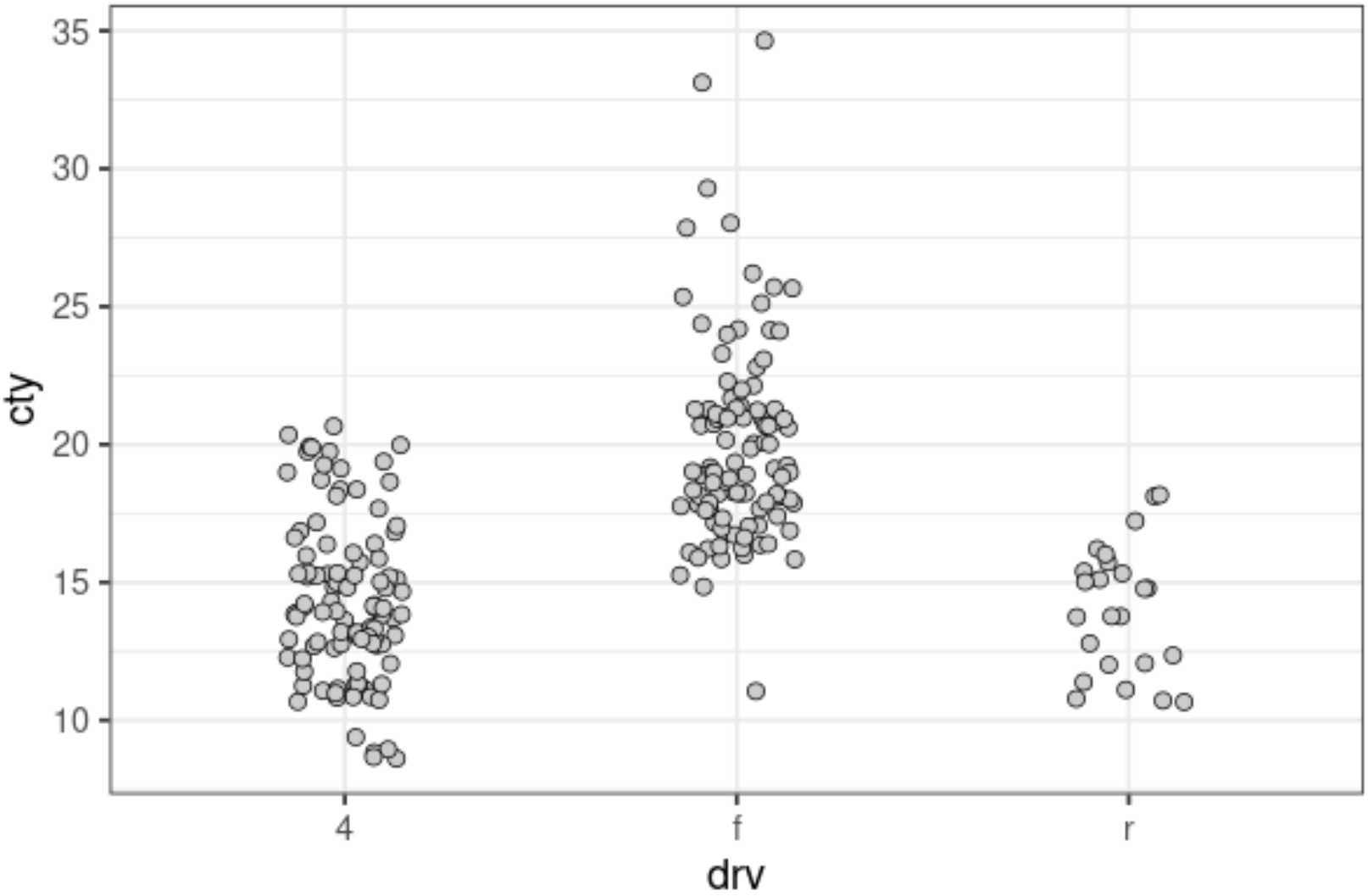



35
30
25
20
15
10
cty
20
30
40
hwy

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



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



Indicator Variables

Consider how data is stored in our data frames in R

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

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

An **indiator variable**, $\mathbb{1}_A$, is a variable that, for each observation, is equal to 1 if the observation is in group A and 0 otherwise

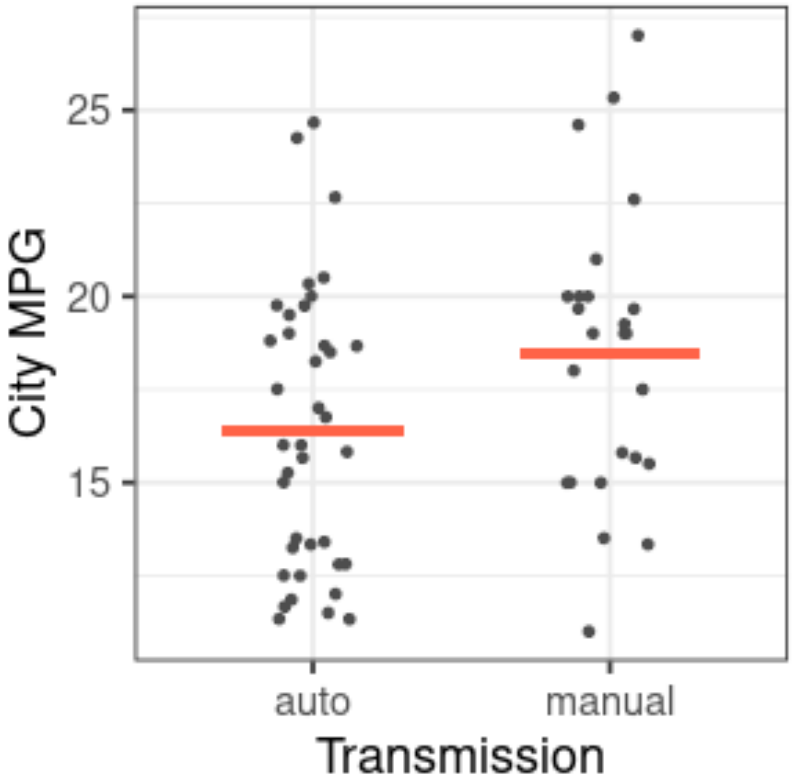
Model	Manual	Auto
audi a4	0	1
audi a4	1	0
chevrolet c1500	0	1
dodge pickup 4wd	0	1
ford explorer 4wd	1	0
hyundai sonata	0	1

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

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

Indicator Variables

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



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

Clarification

To be clear on terms:

- $\mathbb{1}_A$ is a variable in the same way that X is a variable (i.e., height, enrollment, flipper length). That is, its value describes an *attribute* of an observation
- The values in front of an indicator variable are called *coefficients* and function in the exact same way as $X\beta$
- The difference between an indicator $\mathbb{1}_A$ and a quantitiative variable X is that for the indicator, the coefficient β represents a *change in intercept* while for X, β represents a change in slope.

Linear Model in R

By default, the first indicator will be absorbed into an intercept, making it the **reference variable**. The value of all other coefficients (i.e., β) will be *in reference* to this

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

Coefficients:
(Intercept)  transmanual
      16.37         2.09
```

Compare:

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

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

Practice

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

Practice

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

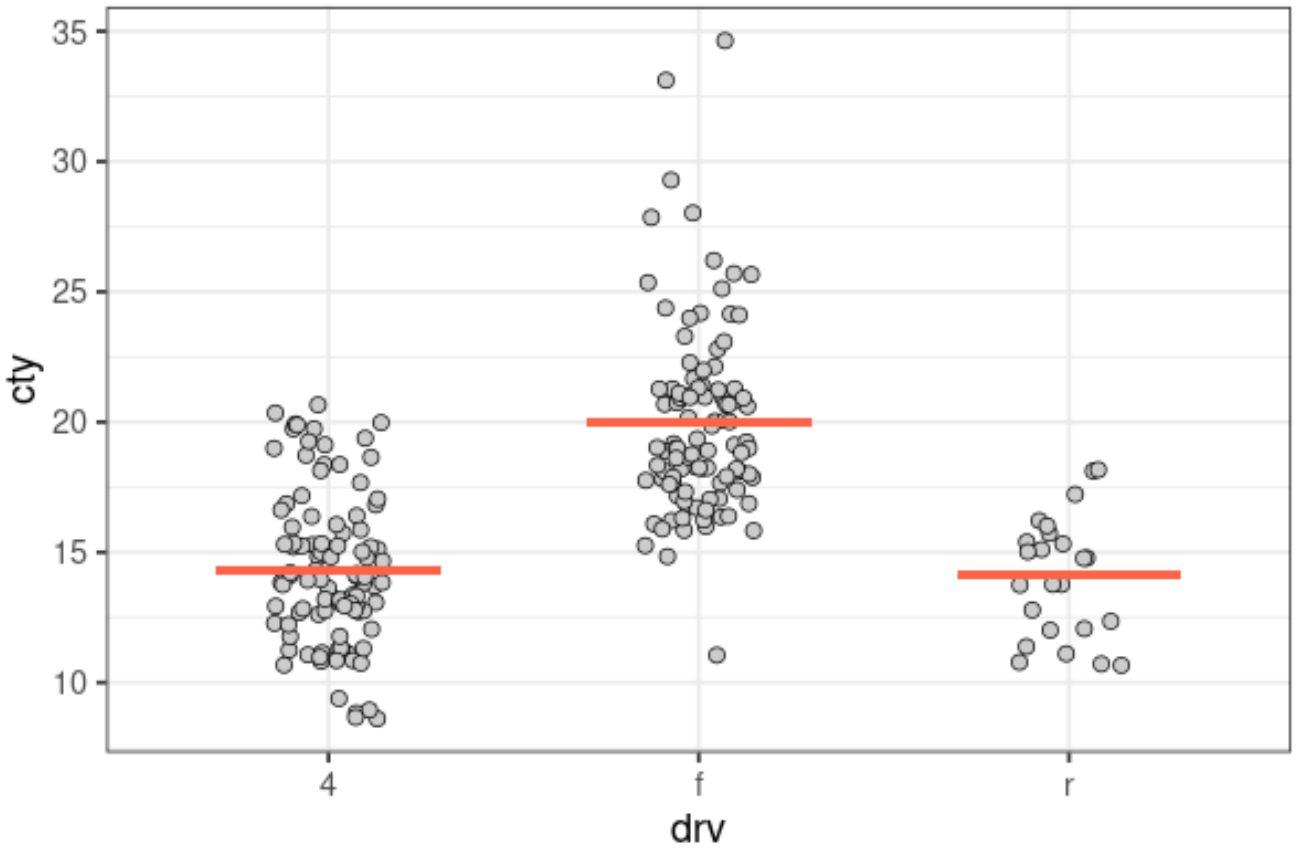
Practice

```
> 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

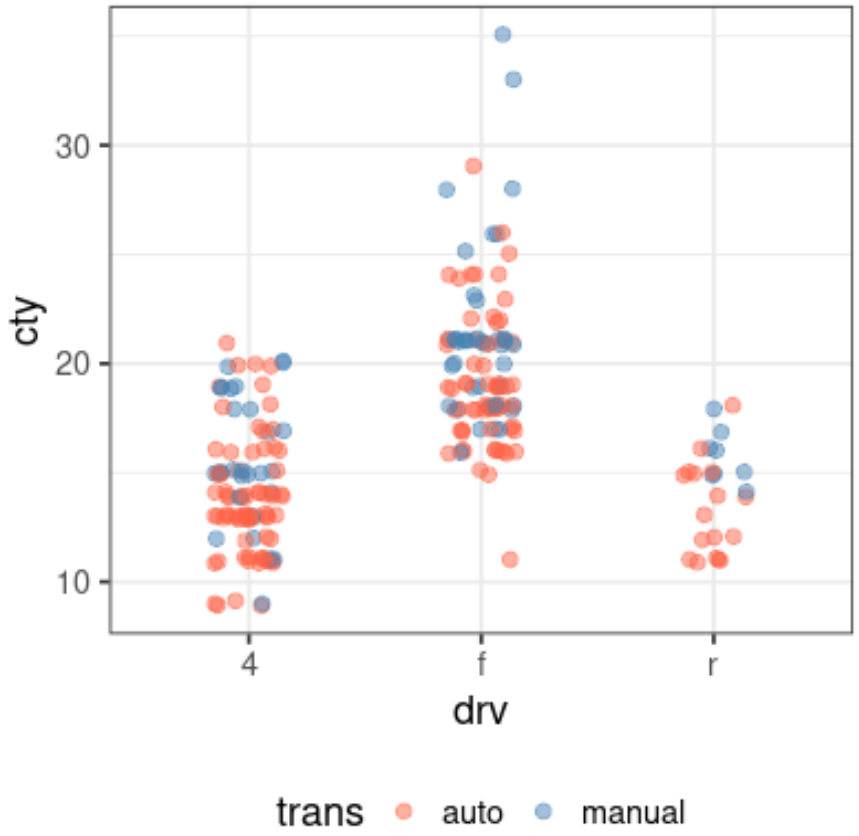
```
> lm(cty ~ drv, mpg)

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



Extending to Multiple Variables

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

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



Extending to Multiple Variables

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

Coefficients:
(Intercept)          drvf          drvr  transmanual
      13.77          5.40         -0.35          2.07
```

- 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

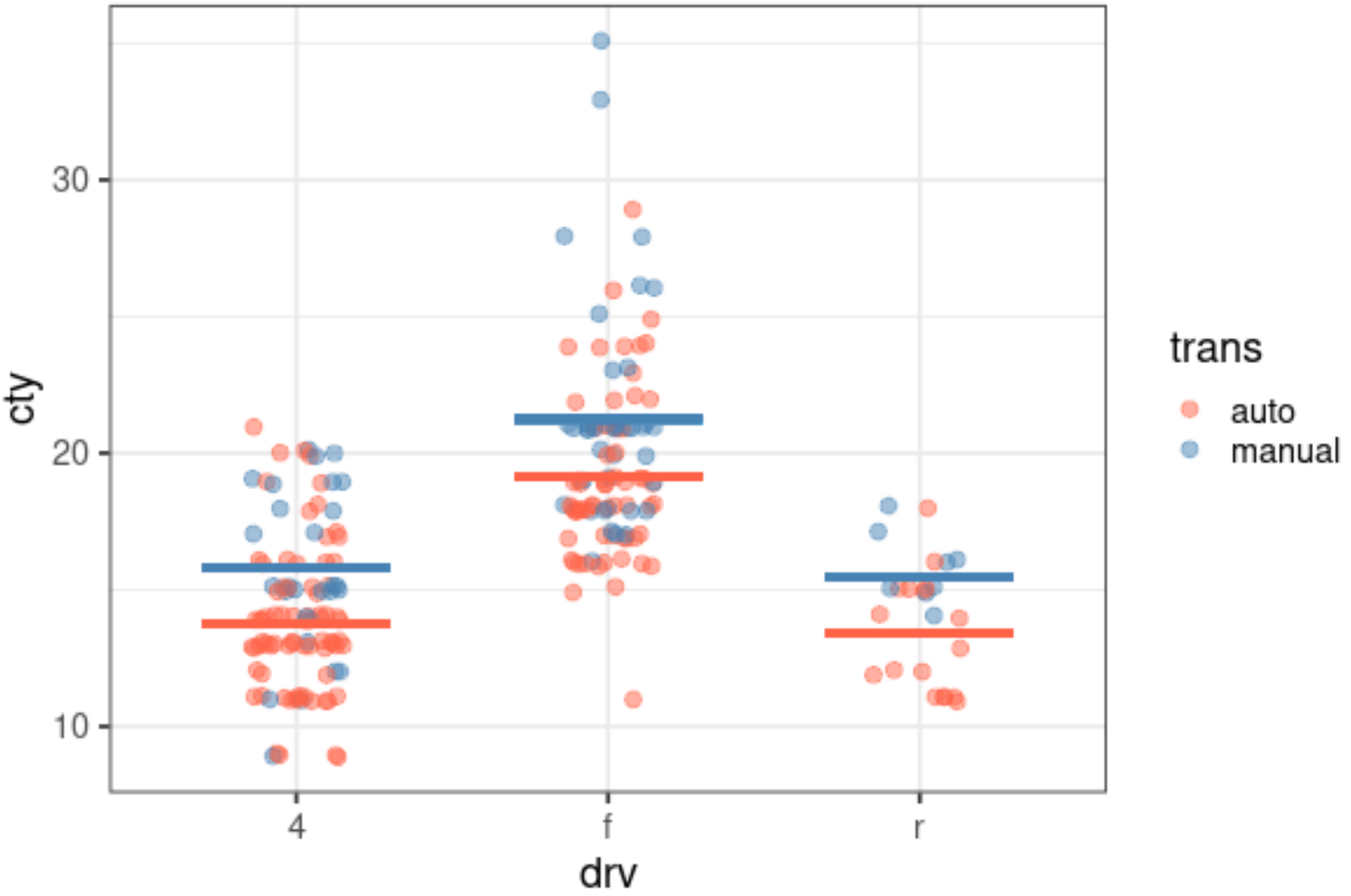
```
> lm(cty ~ drv + trans, mpg)

Coefficients:
(Intercept)          drvf          drvr  transmanual
      13.77          5.40         -0.35          2.07
```

Observed:

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

Predicted:

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


cty
30
20
10
4
f
r
drv
trans
auto
manual