Simple Linear Regression – Categorical Predictors

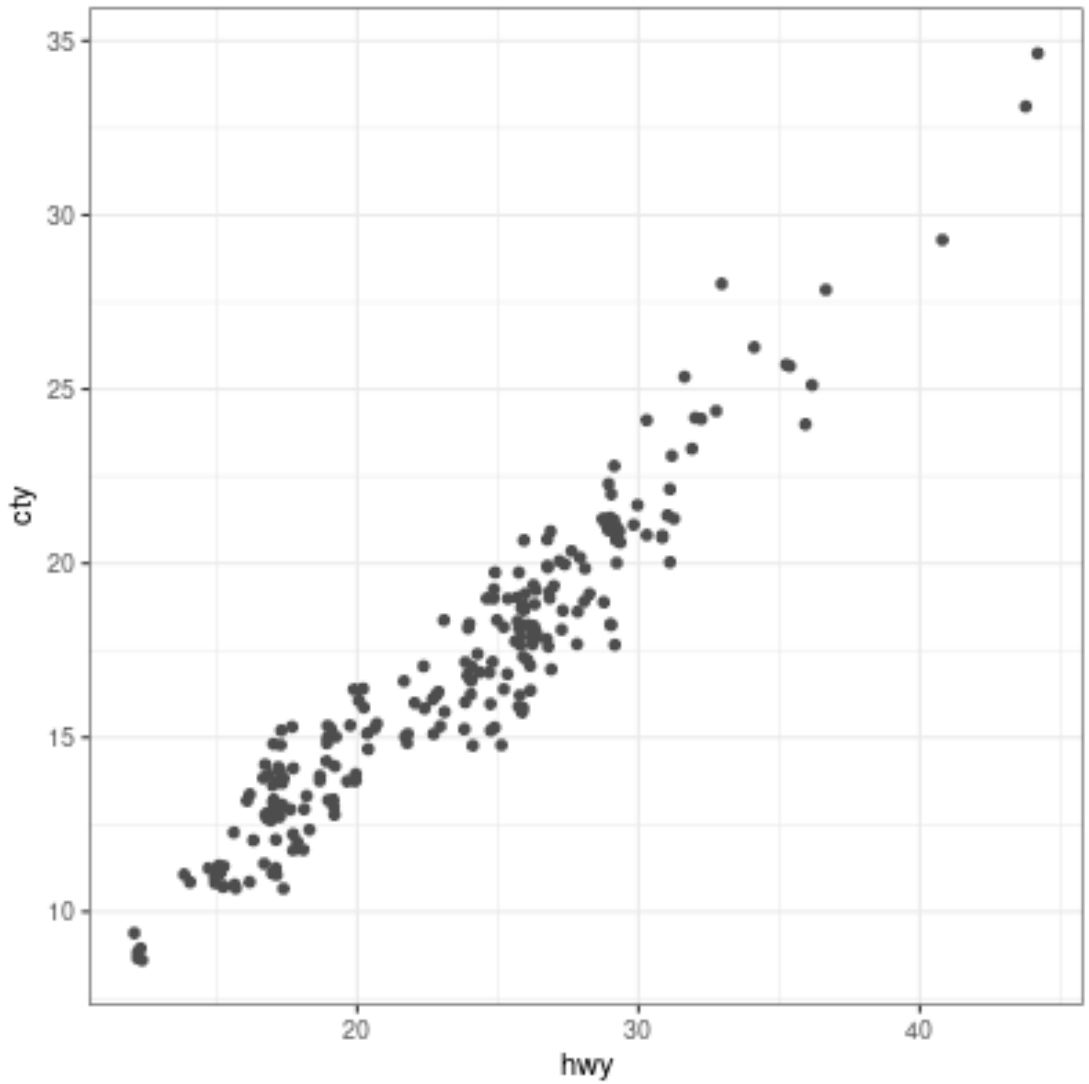
Grinnell College

February 16, 2024

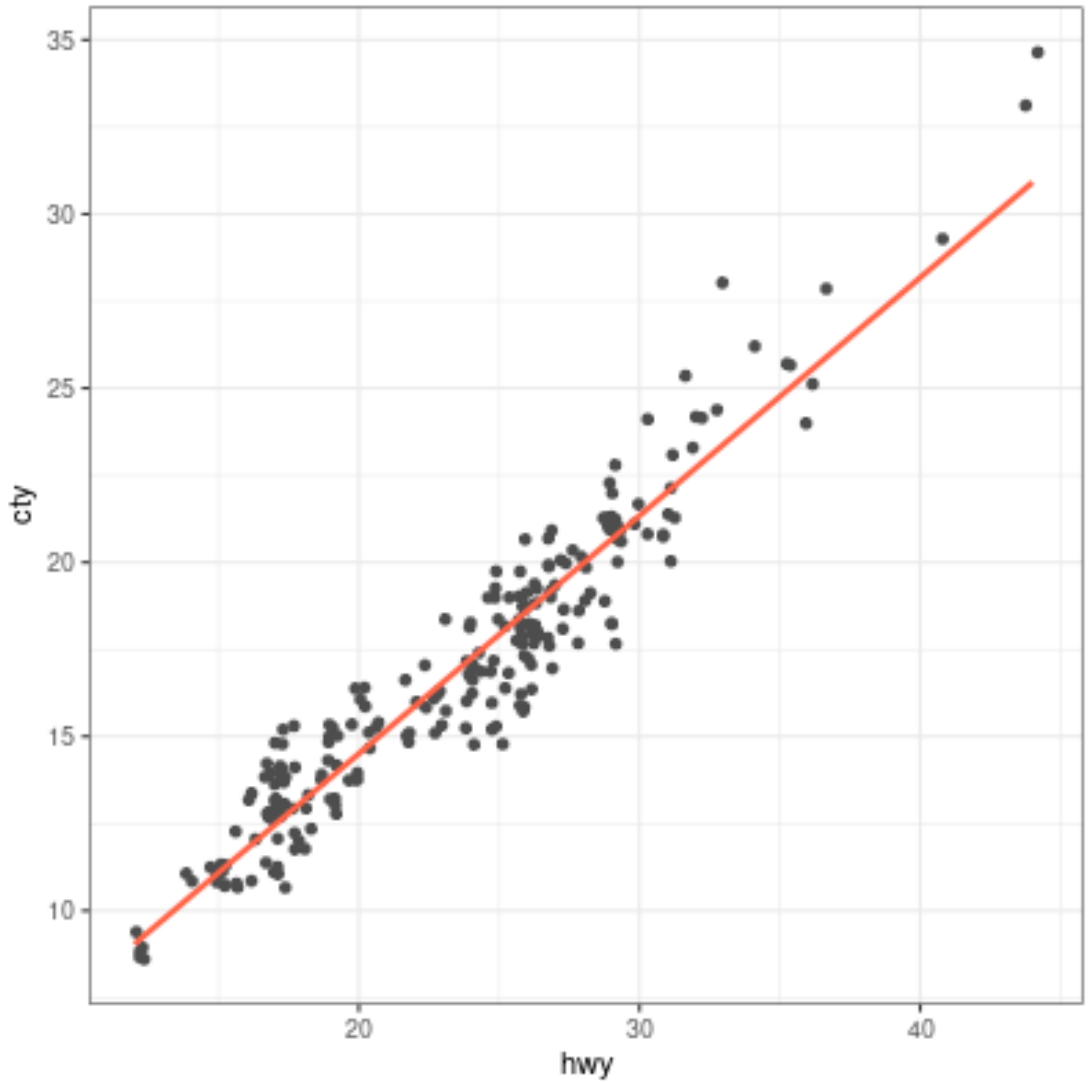
Review

$$\hat{y} = \hat{\beta}_0 + X \times \hat{\beta}_1$$

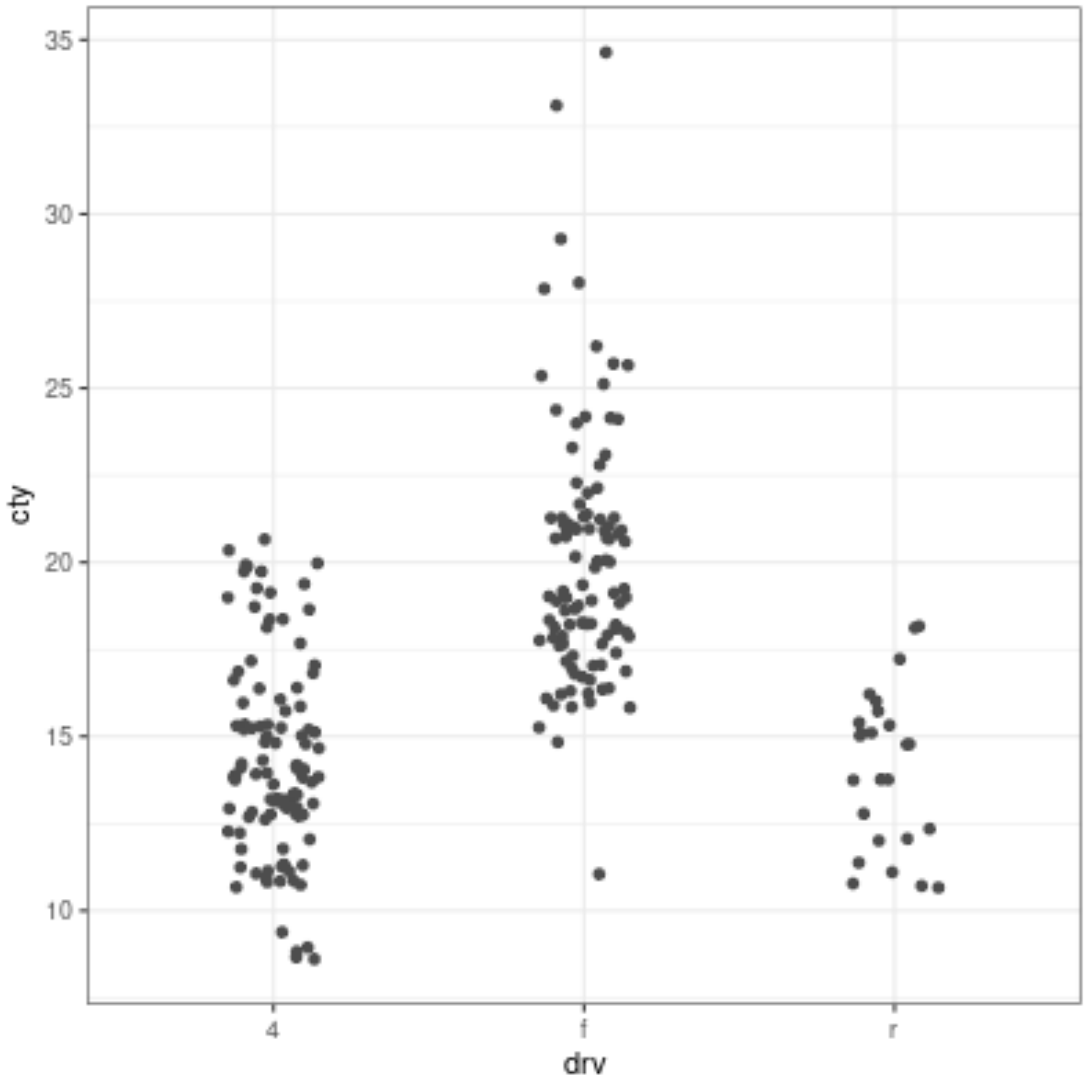
- Describe how correlation and regression related
- Why regression?
- Be able to predict an outcome, given a predictor
- Interpret the slope and intercept (if applicable)
- Assess the quality of a fitted line


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

Name	Private
Adrian College	Private
Alabama A&M	Public
Alfred University	Private
Beloit College	Private
Binghamton University	Public

How might these be used in regression?

Indicator Variables

Name	Private
Adrian College	Private
Alabama A&M	Public
Alfred University	Private
Beloit College	Private
Binghamton University	Public

Name	Public	Private
Adrian College	0	1
Alabama A&M	1	0
Alfred University	0	1
Beloit College	0	1
Binghamton University	1	0

Indicator Variables

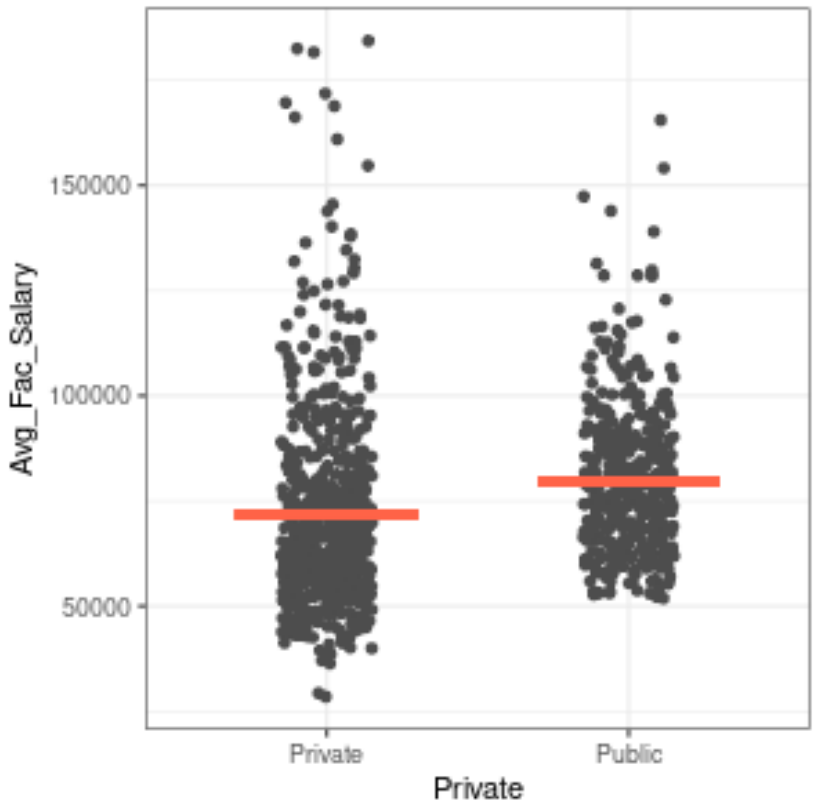
Name	Public	Private
Adrian College	0	1
Alabama A&M	1	0
Alfred University	0	1
Beloit College	0	1
Binghamton University	1	0

$$\mathbb{1}_{\text{Private}} = \begin{cases} 1 & \text{if Private} \\ 0 & \text{if Public} \end{cases}$$

$$\mathbb{1}_{\text{Public}} = \begin{cases} 1 & \text{if Public} \\ 0 & \text{if Private} \end{cases}$$

Indicator Variables

$$\widehat{\text{Avg Fac Sal}} = 71,836 \times \mathbb{1}_{\text{Private}} + 79,636 \times \mathbb{1}_{\text{Public}}$$



Name	Public	Private	Fac Sal
Adrian College	0	1	72873
Alabama A&M	1	0	63909
Alfred Uni	0	1	58410
Beloit College	0	1	63387
Binghamton Uni	1	0	88011
⋮	⋮	⋮	⋮

Private	Average Fac Salary
Private	71836
Public	79636

Linear Model in R

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

```
> lm(Avg_Fac_Salary ~ Private, college)

Coefficients:
  (Intercept)  PrivatePublic
        71836           7800
```

$$\widehat{\text{Avg Fac Sal}} = 71,836 + 7,800 \times \mathbb{1}_{\text{Public}}$$

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

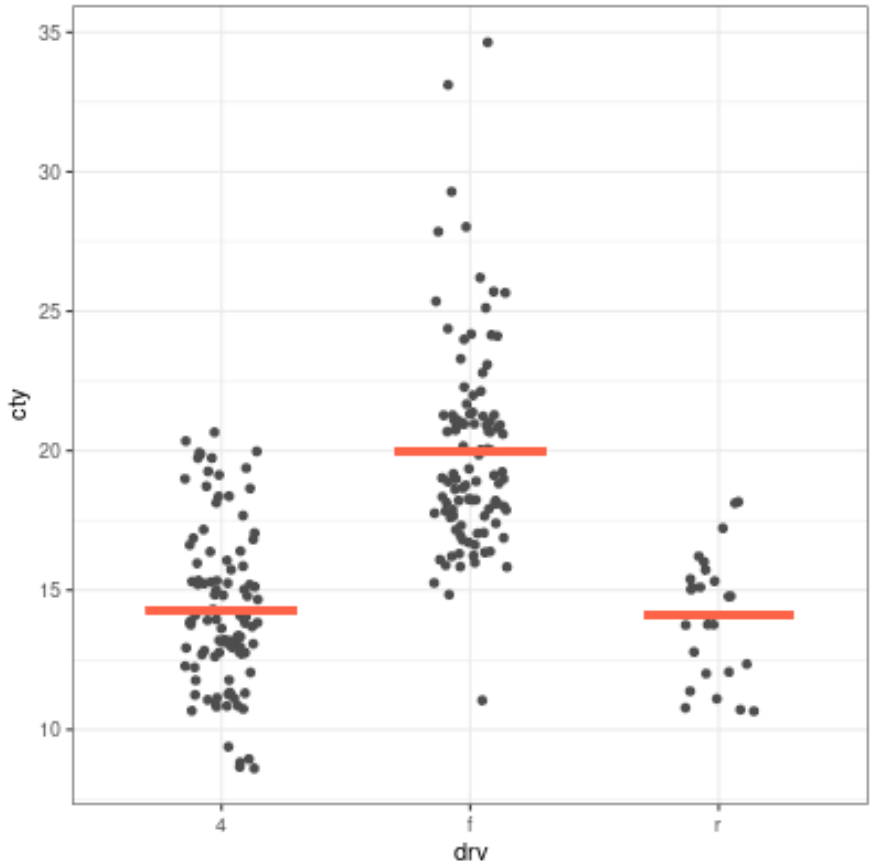
Practice

```
> 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
```



Categorical and Quantiative Predictor

Often times where we have multiple predictors in linear regression model

Consider an example where we investigate odontoblasts measured in 60 guinea pigs, each receiving three doses of vitamin C a day (0.5, 1, 2mg/day) by one of two methods (orange juice (OJ) or ascorbic acid (VC))

We are interested in finding out two things:

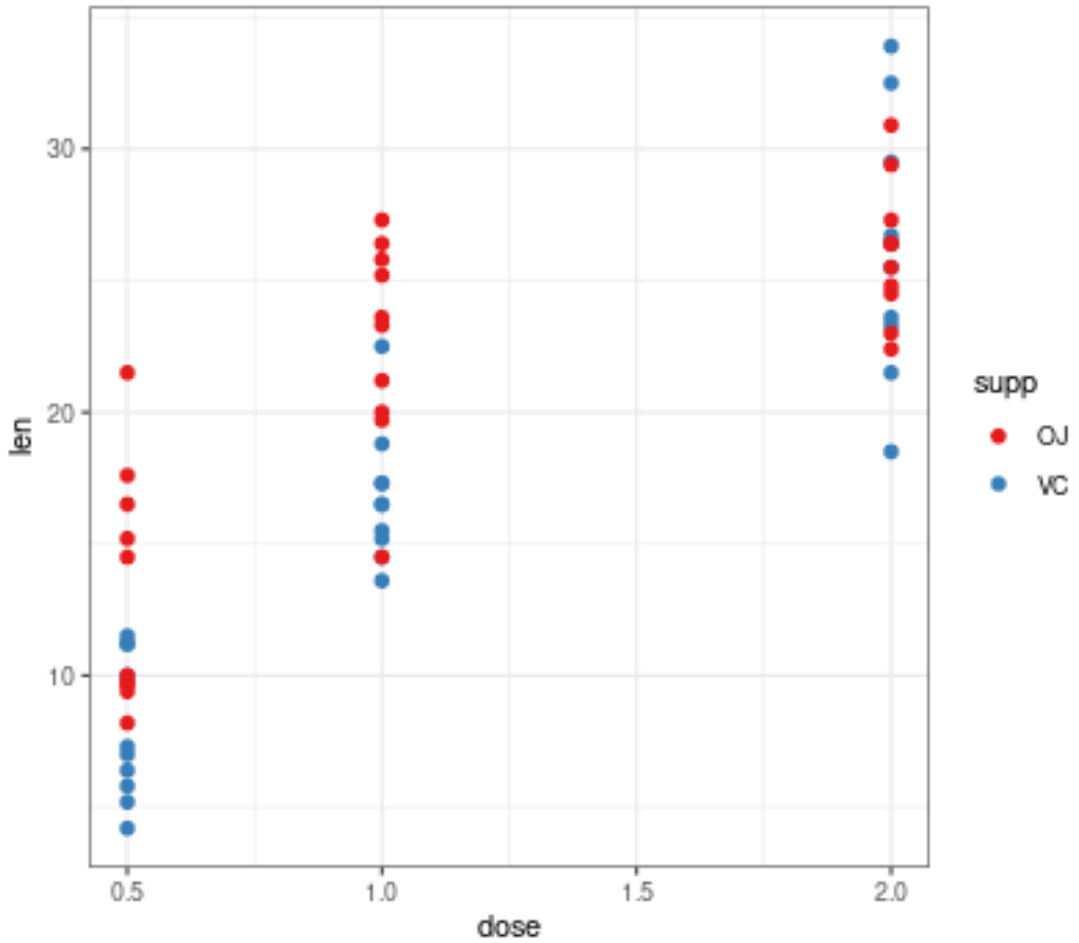
1. Is more vitamin C associated with greater odontoblast growth?
2. Does the ROA of vitamin C influence odontoblast growth?

Building the model

```
> lm(len ~ supp + dose, ToothGrowth)

Coefficients:
(Intercept)         suppVC           dose
       9.27          -3.70           9.76
```

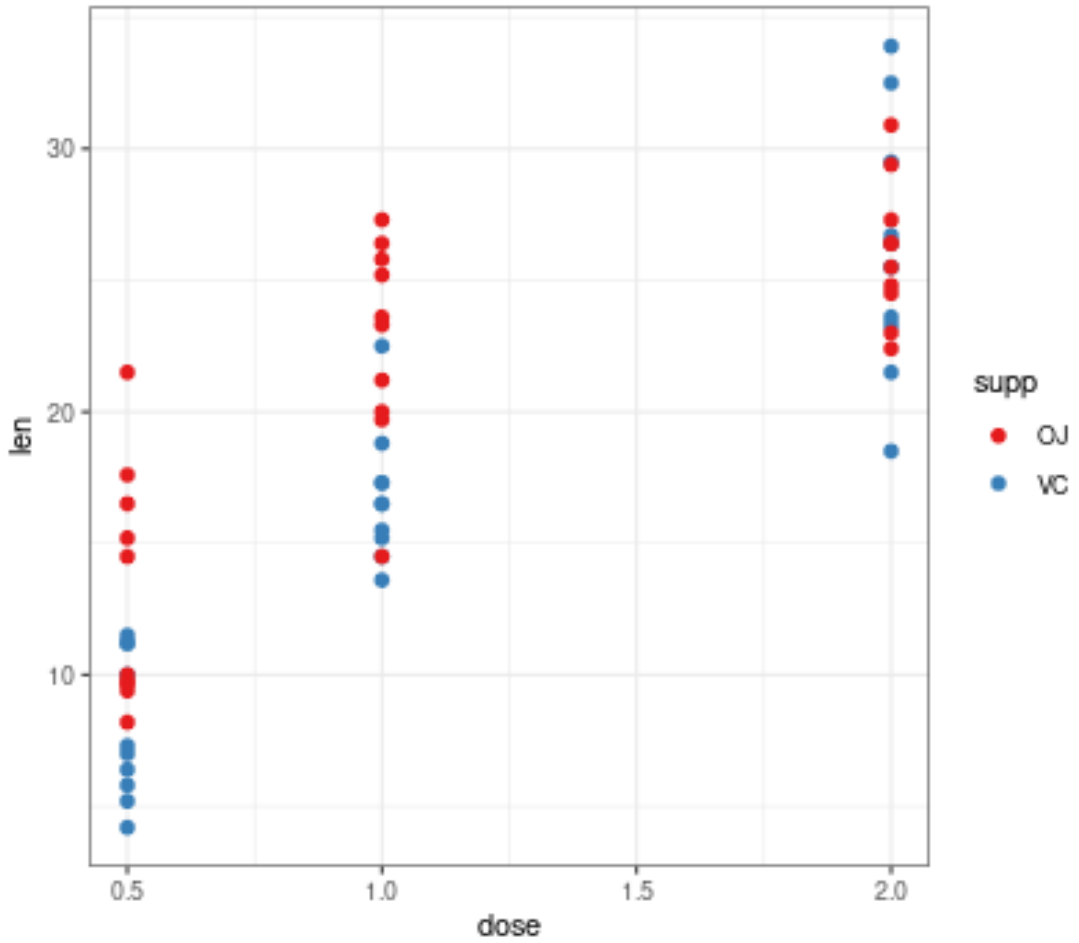
- What is the formula for this regression line?
- What is my reference variable?
- How does ascorbic acid compare with orange juice?
- Is more vitamin C associated with greater tooth growth?

Model Interpretation



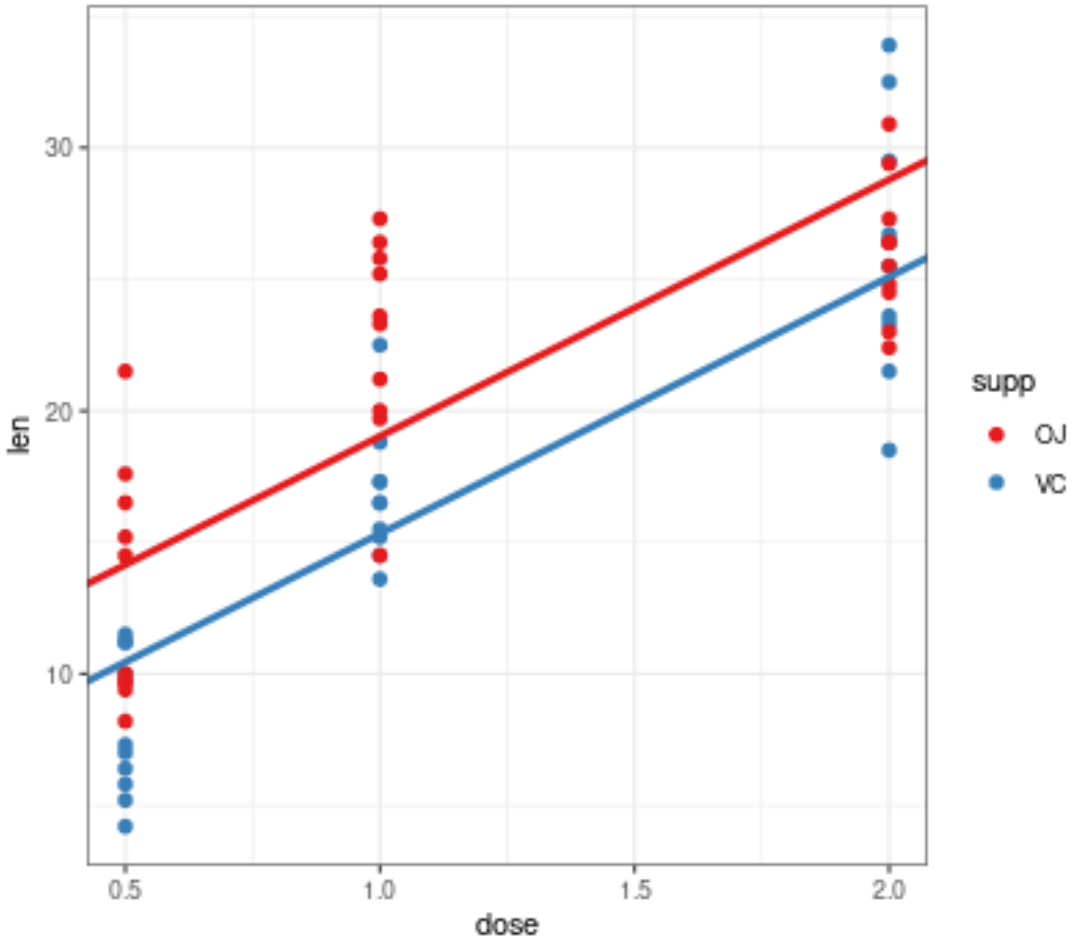
Model Interpretation

$$\widehat{\text{Tooth Growth}} = 9.27 + -3.70 \times \mathbb{1}_{\text{VC}} + 9.76 \times \text{Dose}$$



Model Interpretation

$$\widehat{\text{Tooth Growth}} = \begin{cases} 9.27 + 9.76 \times \text{Dose} & \text{if OJ} \\ 5.57 + 9.76 \times \text{Dose} & \text{if VC} \end{cases}$$



Review

- Categorical variables can be used in regression
- Indicator and reference variables
- Interpretation using only one categorical
- Interpretation using categorical and continuous