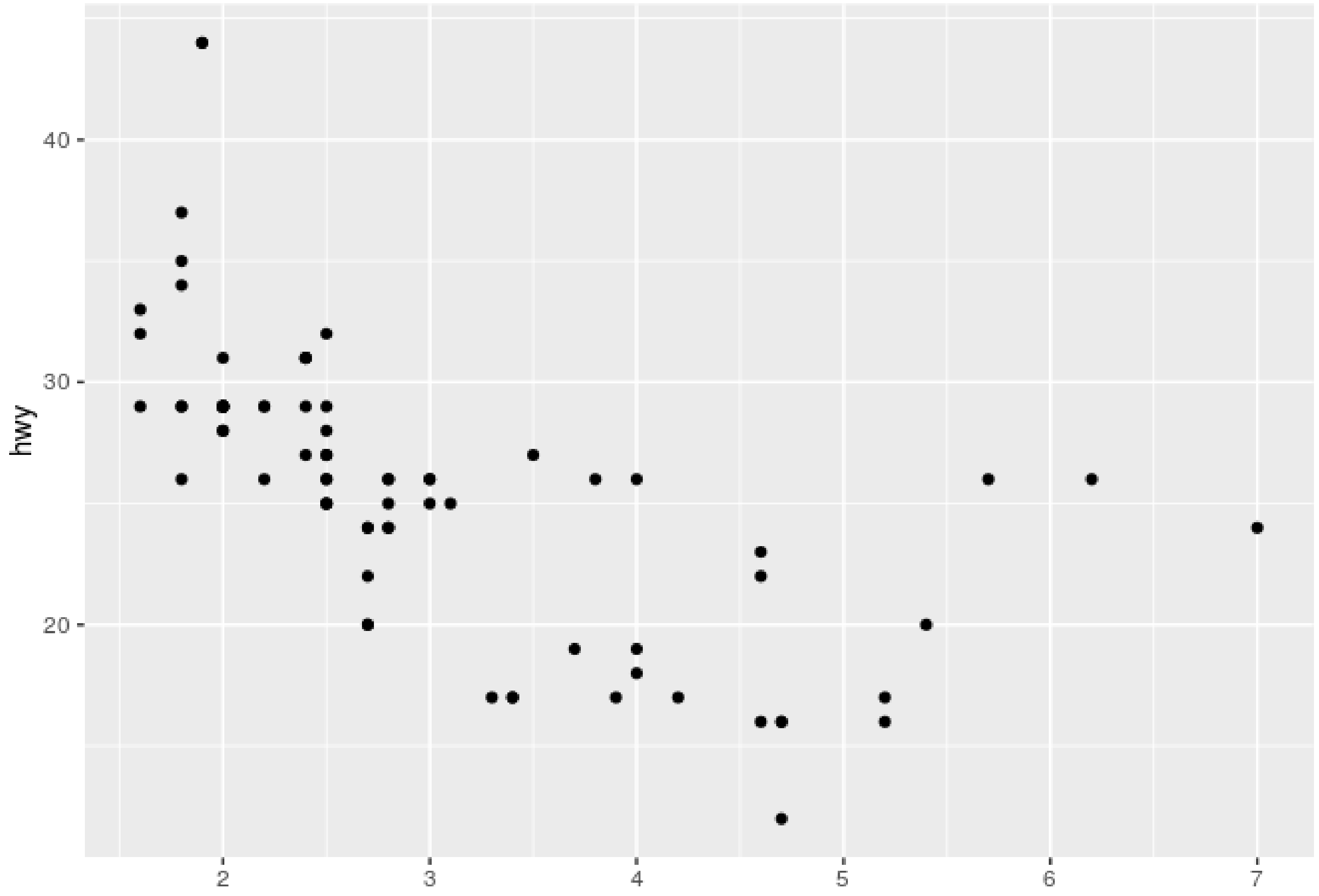
Multiple and logistic regression

MULTIPLE AND LOGISTIC REGRESSION

Ben Baumer
Instructor

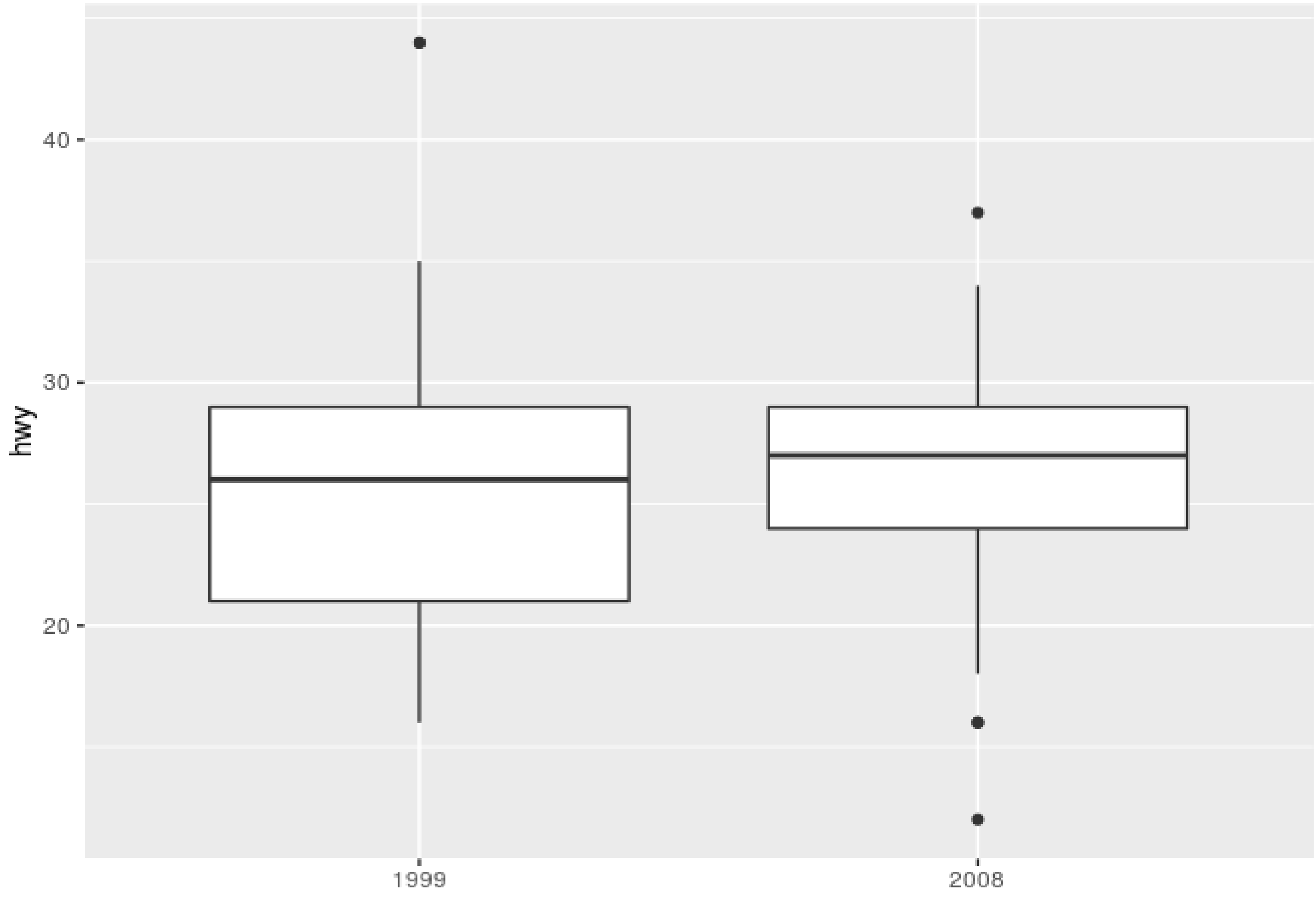
Fuel efficiency by engine size

```
ggplot(data = mpg_manuals, aes(x = displ, y = hwy)) +
  geom_point()
```

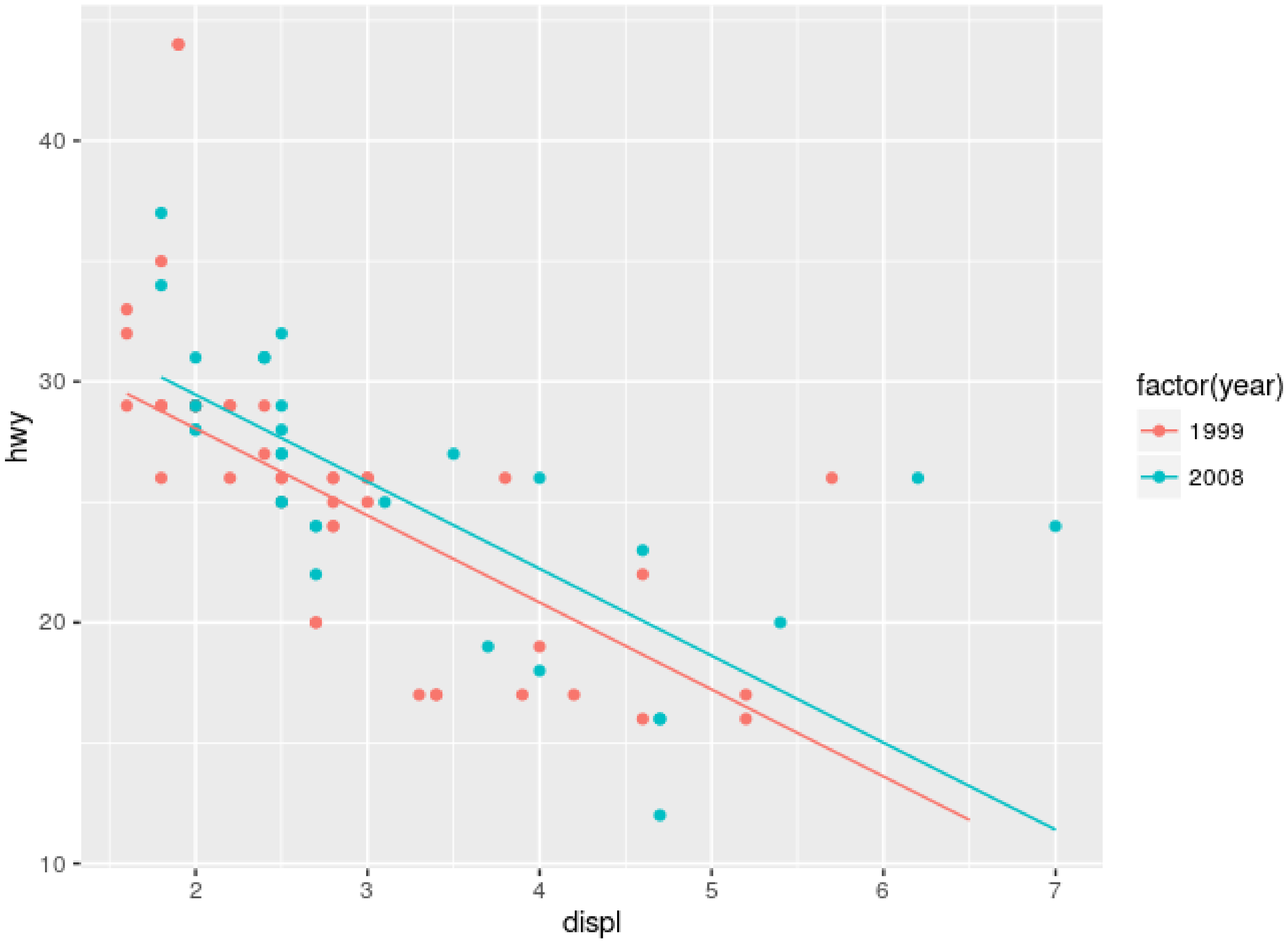


Fuel efficiency over time

```
ggplot(data = mpg_manuals, aes(x = factor(year), y = hwy)) +
  geom_boxplot()
```



A parallel slopes model



Adding a new variable

Consider:

$$hwy = \beta_0 + \beta_1 \cdot displ + \beta_2 \cdot year + \epsilon$$

Adding a new variable in R

```
lm(hwy ~ displ + factor(year), data = mpg)
```

Let's practice!

MULTIPLE AND LOGISTIC REGRESSION

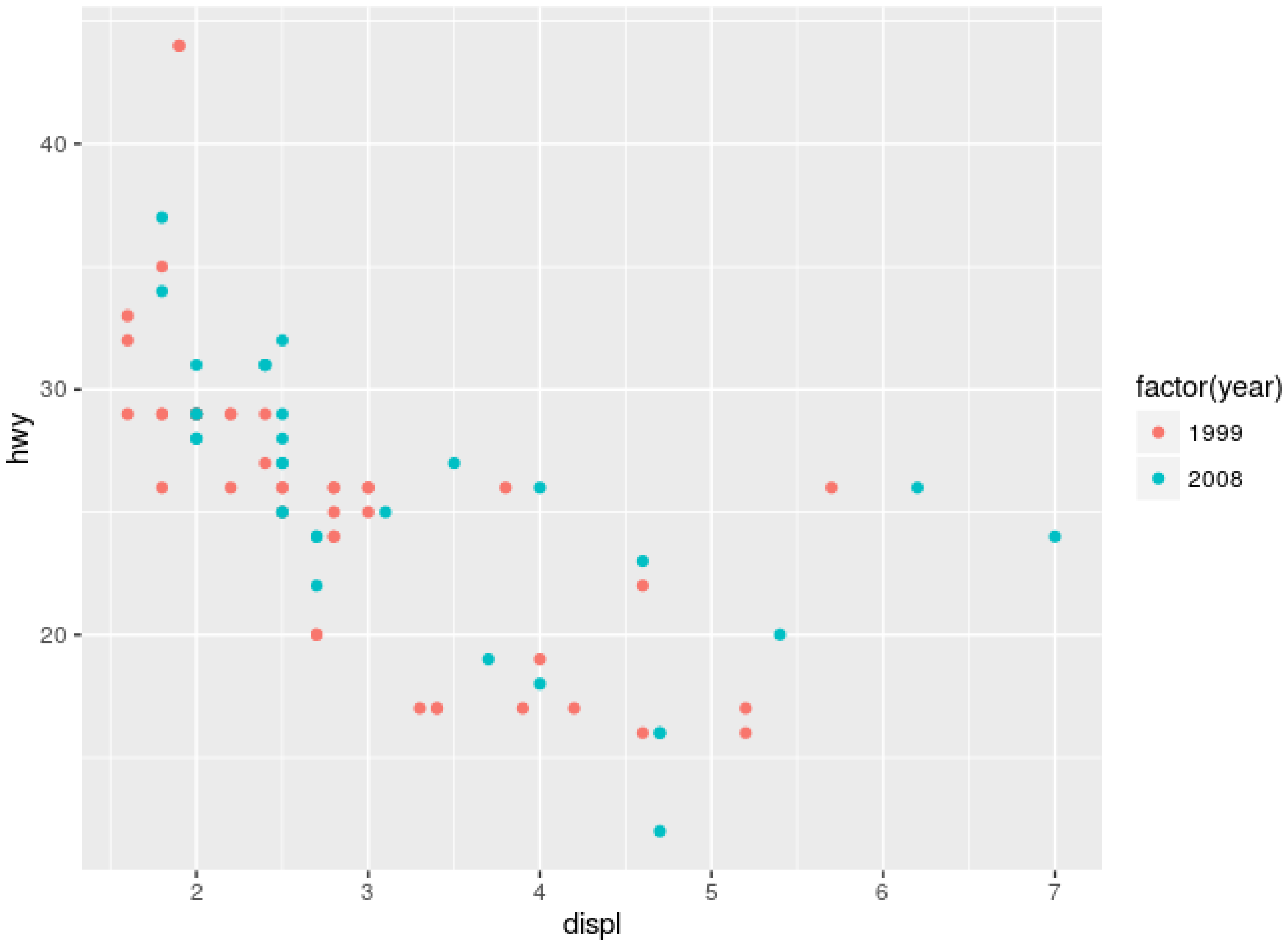
Visualizing parallel slopes models

MULTIPLE AND LOGISTIC REGRESSION

Ben Baumer
Instructor

Three variables, one plot

```
data_space
```



Setting up the model

- Define

$$newer = \begin{cases} 1 & \text{if } year = 2008, \\ 0 & \text{if } year = 1999 \end{cases}$$

- Our model is:

$$\hat{hwy} = \hat{\beta}_0 + \hat{\beta}_1 \cdot displ + \hat{\beta}_2 \cdot newer$$

Two vintages of cars

```
mod <- lm(hwy ~ displ + factor(year), data = mpg)
mod
```

```
## Coefficients:
##      (Intercept)              displ   factor(year)2008
##           35.276             -3.611              1.402
```

- For `year = 2008`, we have

$$\hat{hwy} = 35.276 - 3.611 \cdot displ + 1.402 \cdot (1) =$$

$$= (35.276 + 1.402) - 3.611 \cdot displ$$

- For `year = 1999`, we have

$$\hat{hwy} = 35.276 - 3.611 \cdot displ + 1.402 \cdot (0) =$$

$$= 35.276 - 3.611 \cdot displ$$

Two parallel lines

$$\begin{aligned} \hat{hwy} &= (\hat{\beta}_0 + \hat{\beta}_2) + \hat{\beta}_1 \cdot displ \\ &= (35.276 + 1.402) - 3.611 \cdot displ \\ &= \overset{intercept}{36.678} - \overset{slope}{3.611} \cdot displ \end{aligned}$$

$$\begin{aligned} \hat{hwy} &= \hat{\beta}_0 + \hat{\beta}_1 \cdot displ \\ &= 35.276 - 3.611 \cdot displ \\ &= \overset{intercept}{35.276} - \overset{slope}{3.611} \cdot displ \end{aligned}$$

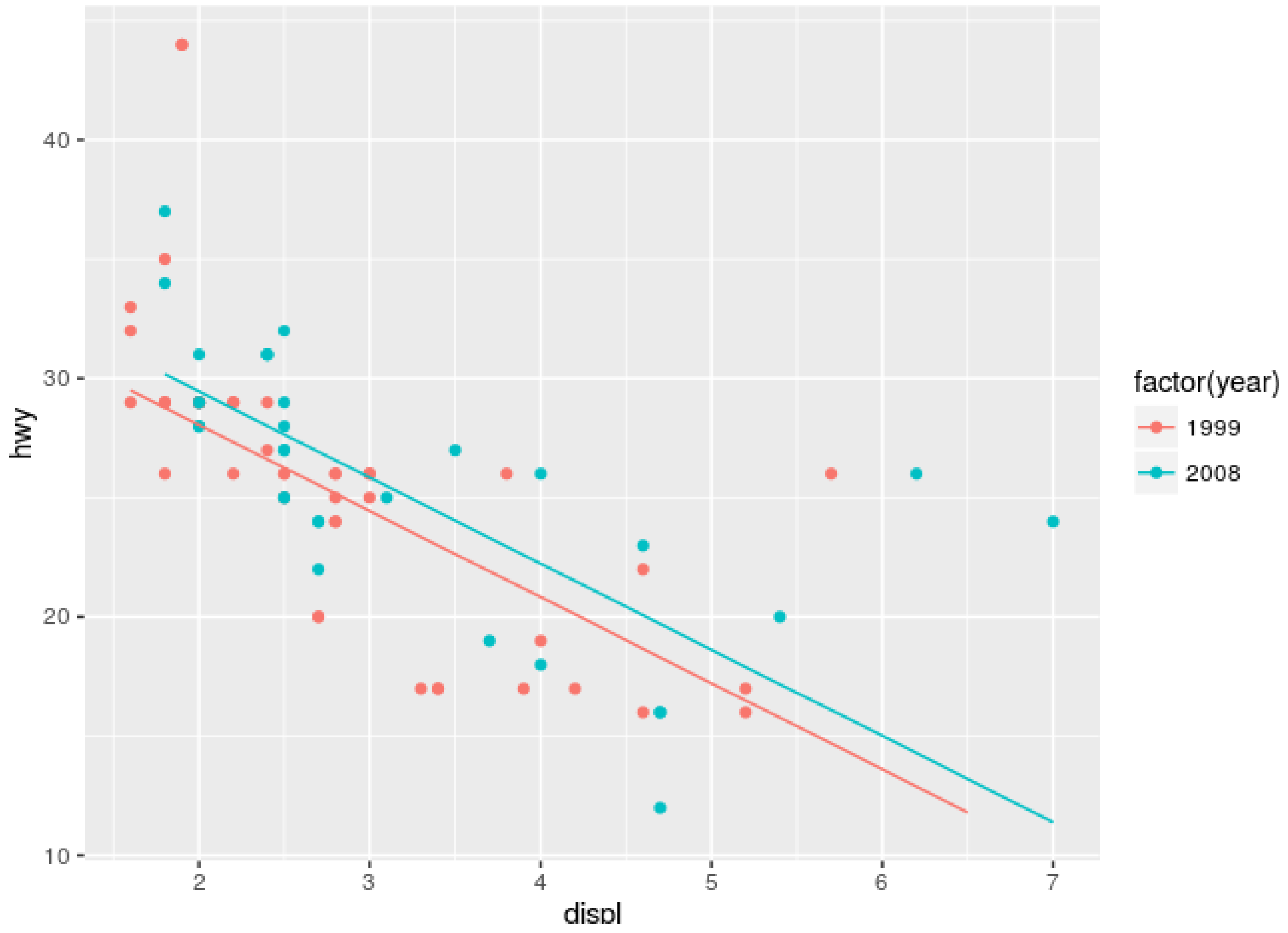
Retrieving the coefficients

```
library(broom)
augment(mod)
```

```
##       hwy displ factor.year.  .fitted    .se.fit       .resid         .hat
## 1      29   1.8         1999 28.77593 0.4522966  0.22406921 0.014314273
## 2      29   1.8         1999 28.77593 0.4522966  0.22406921 0.014314273
## 3      31   2.0         2008 29.45587 0.4753645  1.54412984 0.015811613
## 4      30   2.0         2008 29.45587 0.4753645  0.54412984 0.015811613
## 5      26   2.8         1999 25.16494 0.3617297  0.83505537 0.009155689
## 6      26   2.8         1999 25.16494 0.3617297  0.83505537 0.009155689
## 7      27   3.1         2008 25.48379 0.3661035  1.51621462 0.009378436
## 8      26   1.8         1999 28.77593 0.4522966 -2.77593079 0.014314273
## 9      25   1.8         1999 28.77593 0.4522966 -3.77593079 0.014314273
## 10     28   2.0         2008 29.45587 0.4753645 -1.45587016 0.015811613
```

Parallel lines on the scatterplot

```
data_space +
  geom_line(data = augment(mod), aes(y = .fitted, color = factor.yea
```



Let's practice!

MULTIPLE AND LOGISTIC REGRESSION

Interpreting parallel slopes coefficients

MULTIPLE AND LOGISTIC REGRESSION

Ben Baumer
Instructor

Intercept interpretation

```
lm(hwy ~ displ + factor(year), data = mpg)
```

```
##
## Call:
## lm(formula = hwy ~ displ + factor(year), data = mpg)
##
## Coefficients:
##      (Intercept)             displ  factor(year)2008
##           35.276            -3.611             1.402
```

Slope interpretation

```
lm(hwy ~ displ + factor(year), data = mpg)
```

```
##
## Call:
## lm(formula = hwy ~ displ + factor(year), data = mpg)
##
## Coefficients:
##      (Intercept)              displ  factor(year)2008
##           35.276             -3.611             1.402
```

Avoiding misunderstandings

- There is only *one* slope
- Which is the reference level?
- What are the units?
- After controlling for...

Let's practice!

MULTIPLE AND LOGISTIC REGRESSION

Three ways to describe a model

MULTIPLE AND LOGISTIC REGRESSION

Ben Baumer
Instructor

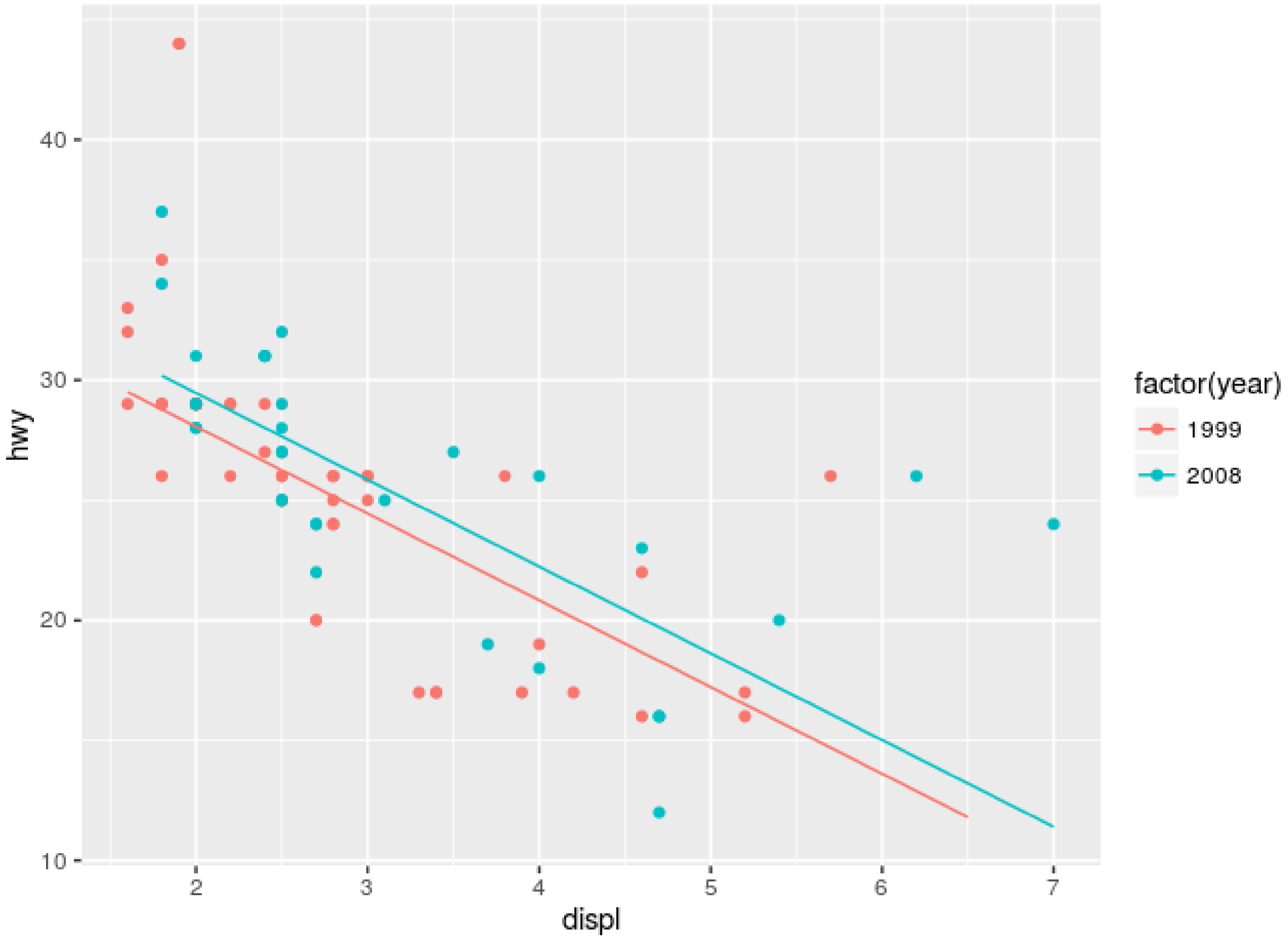
Three ways to describe a model

- Mathematical
- Geometric
- Syntactic

Mathematical

- Equation: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \epsilon$
- Residuals: $\epsilon \sim N(0, \sigma_\epsilon)$
- Coefficients: $\beta_0, \beta_1, \beta_2$

Geometric



Syntactic

```
lm(hwy ~ displ + factor(year), data = mpg)
```

```
##
## Call:
## lm(formula = hwy ~ displ + factor(year), data = mpg)
##
## Coefficients:
##      (Intercept)             displ  factor(year)2008
##           35.276            -3.611             1.402
```

Multiple regression

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \epsilon$
- `y ~ x1 + x2 + x3`
- one line becomes multiple lines or a plane, or even multiple planes

Let's practice!

MULTIPLE AND LOGISTIC REGRESSION