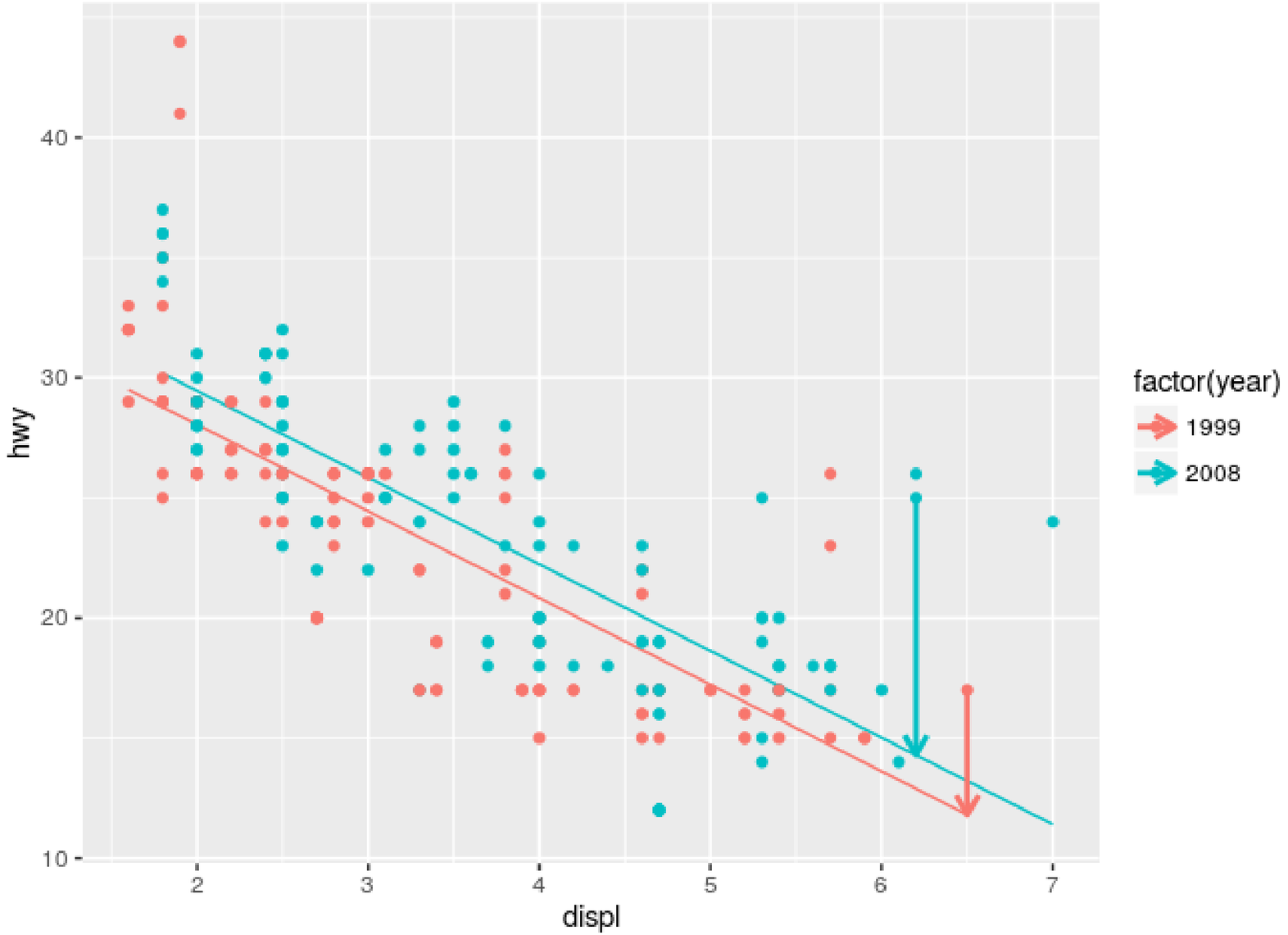
Model fit, residuals, and prediction

MULTIPLE AND LOGISTIC REGRESSION IN R

Ben Baumer
Instructor


40
30
20
10
hwy
2
3
4
5
6
7
displ
factor(year)
1999
2008

Model Fit

- Recall: $R^2 = 1 - \frac{SSE}{SST}$
- SSE get smaller $\Rightarrow$ R^2 increases
- As p (number of explanatory variables) increases...
- Solution: $R^2_{adj} = 1 - \frac{SSE}{SST} \cdot \frac{n-1}{n-p-1}$

Fitted values

```
# returns a vector
predict(mod)

# returns a data.frame
augment(mod)
```

Predictions

```
new_obs <- data.frame(displ = 1.8, year = 2008)

# returns a vector
predict(mod, newdata = new_obs)
```

```
##        1
## 30.17807
```

```
# returns a data.frame
augment(mod, newdata = new_obs)
```

```
##   displ year  .fitted   .se.fit
## 1   1.8 2008 30.17807 0.5024495
```

Let's practice!

MULTIPLE AND LOGISTIC REGRESSION IN R

Understanding interaction

MULTIPLE AND LOGISTIC REGRESSION IN R

Ben Baumer
Instructor

Interaction

```
ggplot(data = mpg, aes(x = displ, y = hwy, color = factor(year))) +
  geom_point() +
  geom_smooth(method = "lm", se = 0)
```

Adding interaction terms

$$\hat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 \cdot displ + \hat{\beta}_2 \cdot is\_newer + \hat{\beta}_3 \cdot displ \cdot is\_newer$$

- For older cars,

$$\hat{mpg} = \hat{\beta}_0 + \hat{\beta}_1 \cdot displ$$

- For newer cars,

$$\hat{mpg} = (\hat{\beta}_0 + \hat{\beta}_2) + (\hat{\beta}_1 + \hat{\beta}_3) \cdot displ$$

Interaction syntax

```
# add interaction term manually
lm(hwy ~ displ + factor(year) + displ:factor(year), data = mpg)
```

Reasoning about interaction

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

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

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

```
Coefficients:
          (Intercept)                 displ
              35.7922               -3.7684
     factor(year)2008  displ:factor(year)2008
               0.3445                0.3052
```

Let's practice!

MULTIPLE AND LOGISTIC REGRESSION IN R

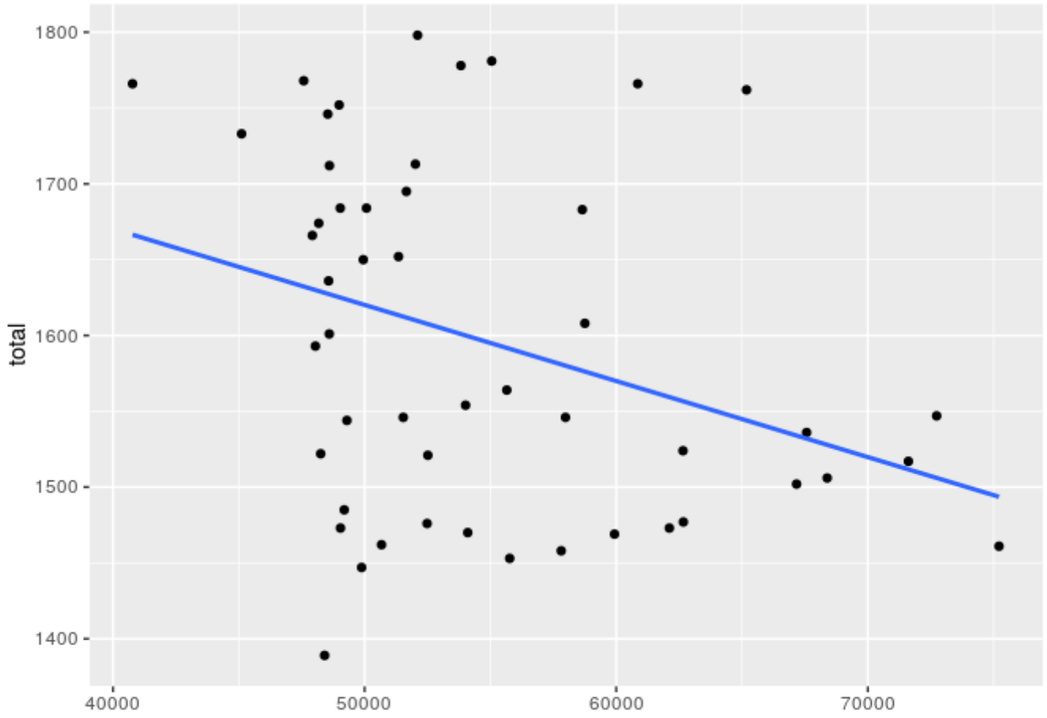
Simpson's Paradox

MULTIPLE AND LOGISTIC REGRESSION IN R

Ben Baumer
Instructor

SAT scores and teacher salary

```
ggplot(data = SAT, aes(x = salary, y = total)) +
  geom_point() +
  geom_smooth(method = "lm", se = 0)
```



Percentage taking the SAT

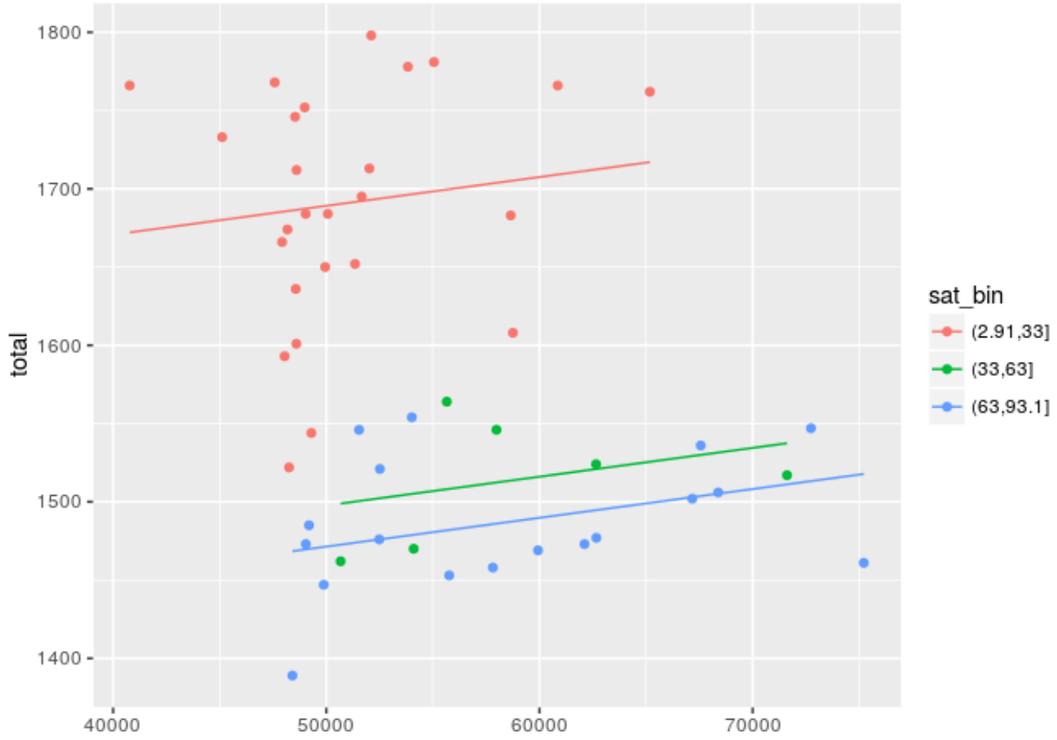
```
SAT_wbin <- SAT %>%
  mutate(sat_bin = cut(sat_pct, 3))
mod <- lm(formula = total ~ salary + sat_bin, data = SAT_wbin)

mod
```

```
## Coefficients:
##      (Intercept)            salary    sat_bin(33,63]  sat_bin(63,93.1]
##       1597.10773           0.00184        -191.45221        -217.73480
```

Simpson's paradox

```
ggplot(data = SAT_wbin, aes(x = salary, y = total, color = sat_bin))
  geom_point() +
  geom_line(data = broom::augment(mod), aes(y = .fitted))
```



Let's practice!

MULTIPLE AND LOGISTIC REGRESSION IN R