

# Inference for Simple Linear Regression

Grinnell College

May 1, 2024

- ▶ Regression posits linear relationship between dependent variable  $y$  and independent variable  $X$  of the form

$$y = \beta_0 + \beta_1 X + \epsilon$$

- ▶ Expand this to include combinations of independent variables
- ▶ We will talk about the error term on Friday

# Cases

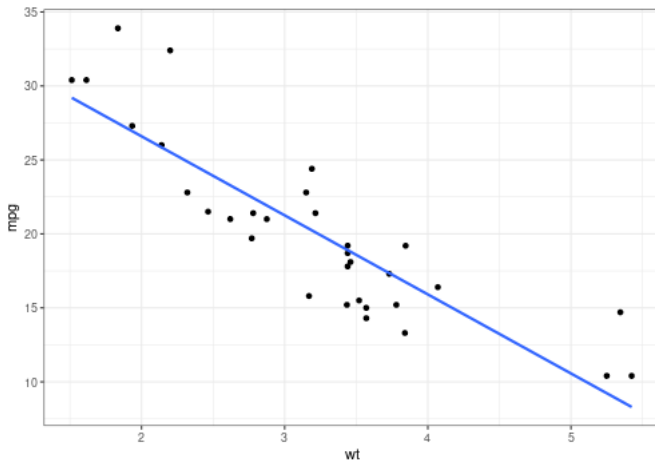
1.  $y = \beta_0 + X\beta_1$
  2.  $y = \beta_0 + \mathbb{1}_A\beta_1$
  3.  $y = \beta_0 + \mathbb{1}_A\beta_1 + X\beta_2$
  4.  $y = \beta_0 + \mathbb{1}_A\beta_1 + \mathbb{1}_B\beta_2$
  5.  $y = \beta_0 + X_1\beta_1 + X_2\beta_2$
1. Simple linear,  $\beta_1$  shows change in  $y$  given change in  $X$
  2. Simple categorical, reference variable and group means
  3. Continuous and categorical, two regression lines with same slope but different intercept
  4. Multiple categorical, combined reference variables
  5. Multiple continuous,  $\beta_1$  shows change in  $y$  given change in  $X_1$ , *assuming everything else held constant*

# Single Quantitative

```
1 > lm(mpg ~ wt, mtcars) %>% summary()
2
3
4 Coefficients:
5             Estimate Std. Error t value      Pr(>|t|)
6 (Intercept)   37.285      1.878   19.86 < 0.00000000002 ***
7 wt           -5.344      0.559   -9.56    0.000013 ***
8
9
10 Residual standard error: 3.05 on 30 degrees of freedom
11 Multiple R-squared:  0.753, Adjusted R-squared:  0.745
12 F-statistic: 91.4 on 1 and 30 DF,  p-value: 0.000000000129
```

# Weight and MPG

$$\hat{y} = 37.285 - 5.34 \times \text{Weight}$$

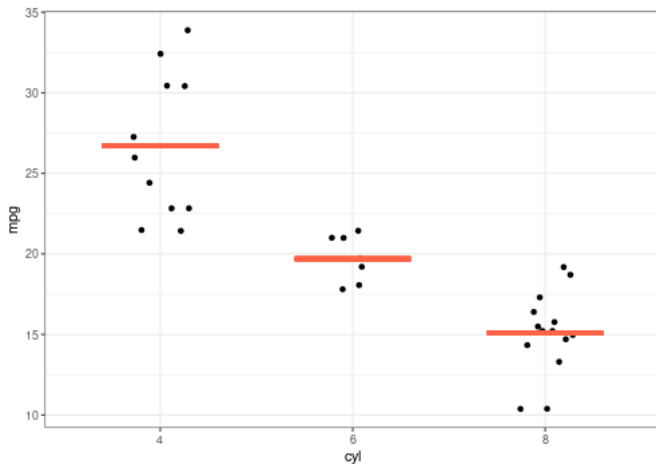


# Single Categorical

```
1 > lm(mpg ~ cyl, mtcars) %>% summary()
2
3 Coefficients:
4             Estimate Std. Error t value      Pr(>|t|)
5 (Intercept)   26.664      0.972   27.44 < 0.00000000002 ***
6 cyl6          -6.921      1.558   -4.44     0.00012 ***
7 cyl8         -11.564      1.299   -8.90     0.00000000086 ***
8
9
10 Residual standard error: 3.22 on 29 degrees of freedom
11 Multiple R-squared:  0.732, Adjusted R-squared:  0.714
12 F-statistic: 39.7 on 2 and 29 DF,  p-value: 0.00000000498
```

# Cylinder and MPG

$$\hat{y} = 26.66 - 6.92 \times \mathbb{1}_{6\text{cyl}} - 11.564 \times \mathbb{1}_{8\text{cyl}}$$



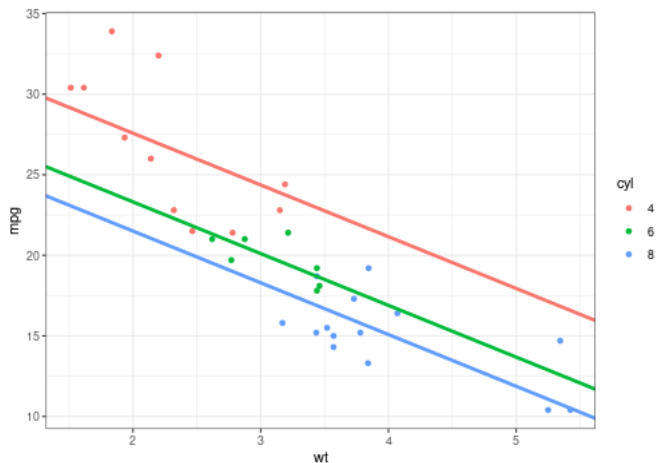
# Categorical and Quantitative

```
1 > lm(mpg ~ wt + cyl, mtcars) %>% summary()
2
3 Coefficients:
4             Estimate Std. Error t value      Pr(>|t|)
5 (Intercept)   33.991     1.888   18.01 < 0.00000000002 ***
6 wt           -3.206     0.754    -4.25     0.00021 ***
7 cyl          -4.256     1.386    -3.07     0.00472 **
8 cyl8         -6.071     1.652    -3.67     0.00100 ***
9
10
11 Residual standard error: 2.56 on 28 degrees of freedom
12 Multiple R-squared:  0.837, Adjusted R-squared:  0.82
13 F-statistic: 48.1 on 3 and 28 DF,  p-value: 0.0000000000359
```



# Cylinder, weight and MPG

$$\hat{y} = 33.99 - 3.21 \times \text{weight} - 4.26 \times \mathbb{1}_{6\text{cyl}} - 6.07 \times \mathbb{1}_{8\text{cyl}}$$

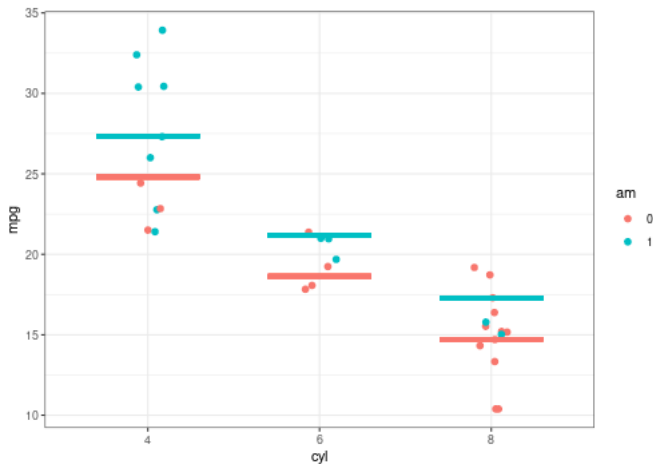


# Multiple Categorical

```
1 > lm(mpg ~ cyl + am, mtcars) %>% summary()
2
3 Coefficients:
4             Estimate Std. Error t value      Pr(>|t|)
5 (Intercept)    24.80        1.32   18.75 < 0.00000000002 ***
6 cyl6           -6.16        1.54   -4.01    0.00041 ***
7 cyl8          -10.07        1.45   -6.93    0.000000015 ***
8 am1             2.56        1.30    1.97    0.05846 .
9
10
11 Residual standard error: 3.07 on 28 degrees of freedom
12 Multiple R-squared:  0.765, Adjusted R-squared:  0.74
13 F-statistic: 30.4 on 3 and 28 DF,  p-value: 0.00000000596
```

# Cylinder, transmission and MPG

$$\hat{y} = 24.8 - 6.16 \times \mathbb{1}_{6\text{cyl}} - 10.07 \times \mathbb{1}_{8\text{cyl}} + 2.56 \times \mathbb{1}_{\text{Manual}}$$

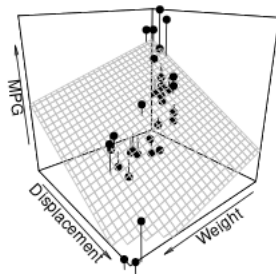
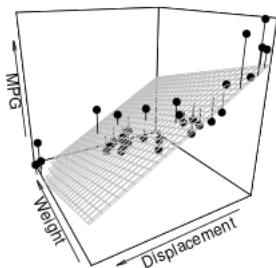


# Multiple Quantitative

```
1 > lm(mpg ~ wt + disp, mtcars) %>% summary()
2
3 Coefficients:
4             Estimate Std. Error t value      Pr(>|t|)
5 (Intercept)  34.96055     2.16454   16.15 0.0000000049 ***
6 wt          -3.35083     1.16413    -2.8  0.0074 **
7 disp        -0.01772     0.00919    -1.93  0.0636 .
8
9
10 Residual standard error: 2.92 on 29 degrees of freedom
11 Multiple R-squared:  0.781, Adjusted R-squared:  0.766
12 F-statistic: 51.7 on 2 and 29 DF, p-value: 0.000000000274
```

# Cylinder, transmission and MPG

$$\hat{y} = 34.96 - 3.35 \times \text{weight} - 0.017 \times \text{displacement}$$

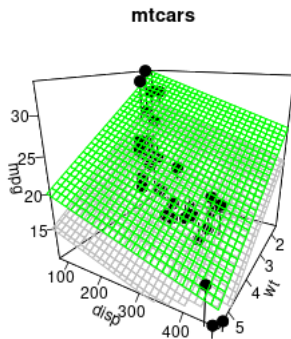
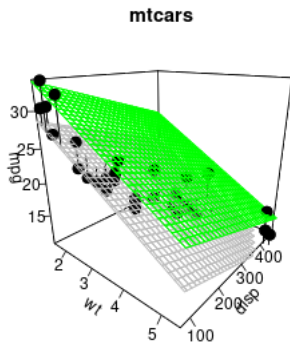


# Multiple Quantitative and categorical

```
1 > lm(mpg ~ wt + disp + am, mtcars) %>% summary()
2
3 Coefficients:
4             Estimate Std. Error t value      Pr(>|t|)
5 (Intercept) 34.67591    3.24061   10.70 0.000000000021 ***
6 wt          -3.27904    1.32751    -2.47    0.020 *
7 disp        -0.01780    0.00937    -1.90    0.068 .
8 am           0.17772    1.48432     0.12    0.906
9
10
11 Residual standard error: 2.97 on 28 degrees of freedom
12 Multiple R-squared:  0.781, Adjusted R-squared:  0.758
13 F-statistic: 33.3 on 3 and 28 DF,  p-value: 0.00000000225
```

# Multiple quantiative with categorical

$$\hat{y} = 34.67 - 3.27 \times \text{weight} - 0.018 \times \text{displacement} + 0.17 \times \mathbb{1}_{\text{Manual}}$$

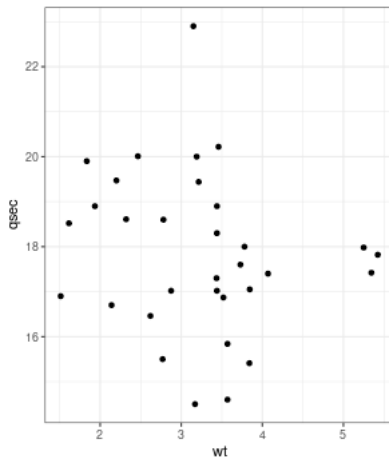
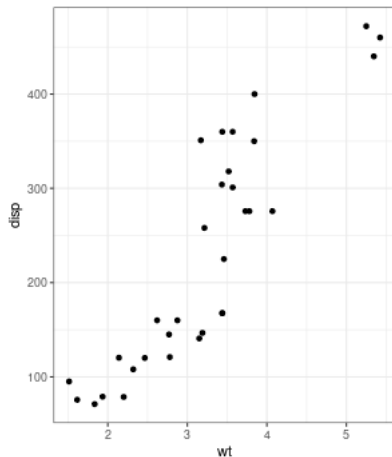


# Correlated Covariates

```
1 > lm(mpg ~ wt, mtcars) %>% summary()
2
3
4 Coefficients:
5             Estimate Std. Error t value      Pr(>|t|)
6 (Intercept)   37.285      1.878    19.86 < 0.000002 ***
7 wt           -5.344      0.559    -9.56  0.000013 ***
8
9
10 > lm(mpg ~ wt + disp, mtcars) %>% summary()
11
12 Coefficients:
13             Estimate Std. Error t value      Pr(>|t|)
14 (Intercept)  34.96055      2.16454    16.15 0.0000000049 ***
15 wt          -3.35083      1.16413     -2.8   0.0074 **
16 disp        -0.01772      0.00919     -1.93  0.0636 .
```



# Correlated Covariates



# Correlated Covariates

```
1 > lm(mpg ~ wt, mtcars) %>% summary()
2
3           Estimate Std. Error t value      Pr(>|t|)
4 (Intercept)   37.285     1.878   19.86 < 0.000002 ***
5 wt           -5.344     0.559   -9.56  0.000013 ***
6 R-squared = 0.75
7
8 > lm(mpg ~ wt + disp, mtcars) %>% summary()
9
10          Estimate Std. Error t value      Pr(>|t|)
11 (Intercept) 34.96055     2.16454   16.15 0.000000049 ***
12 wt          -3.35083     1.16413    -2.8   0.0074 **
13 disp        -0.01772     0.00919    -1.93  0.0636 .
14 R-squared = 0.78
15
16 > lm(mpg ~ wt + qsec, mtcars) %>% summary()
17
18          Estimate Std. Error t value      Pr(>|t|)
19 (Intercept)  19.746     5.252     3.76   0.00077 ***
20 wt          -5.048     0.484   -10.43 0.000000000025 ***
21 qsec         0.929     0.265     3.51   0.00150 **
22 R-squared = 0.82
```

# Key Takeaways

- ▶ Quantitative variables represent slopes (changes in  $X$  lead to  $\beta$  changes in  $y$ )
- ▶ Categorical variables represent horizontal shifts
- ▶ Any number of categorical or quantitative variables can be added to model
- ▶ Lookout for correlated variables
- ▶ Always interpret regression coefficients as *everything else being fixed*