Inference for Multivariate Regression

ANOVA for MLR

Grinnell College

December 9, 2024

Review

- Regression models a linear relationship between response variable y and explanatory variable X of the form

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

- We can expand this to include *combinations* of explanatory variables
- in fact we saw this earlier on in the semester too, albeit briefly

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. $y = \beta_0 + X_1\beta_1 + X_2\beta_2$

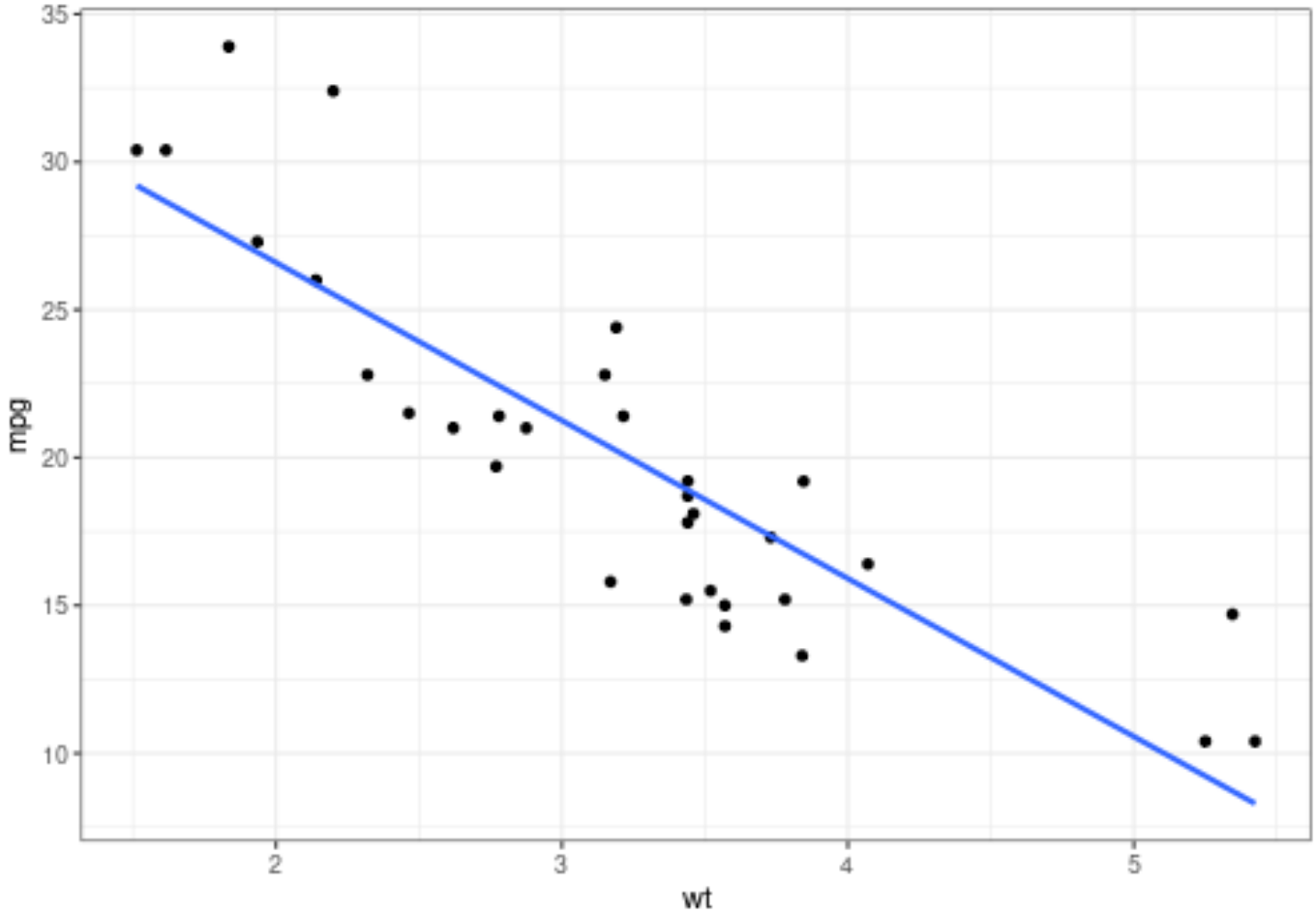
1. Simple linear, β_1 shows change in y given change in X
2. Simple categorical, reference variable and group means
3. Continuous and categorical, two regression lines with same slope but different intercept
4. Multiple categorical, combined reference variables
5. Multiple continuous, β_1 shows change in y given change in X_1, *assuming everything else held constant*

Single Quantiative

```
> lm(mpg ~ wt, mtcars) %>% summary()


Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)   37.285      1.878   19.86 < 0.00000000002 ***
wt            -5.344      0.559   -9.56            0.000013 ***


Residual standard error: 3.05 on 30 degrees of freedom
Multiple R-squared:  0.753, Adjusted R-squared:  0.745
F-statistic: 91.4 on 1 and 30 DF,  p-value: 0.000000000129
```

Weight and MPG

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

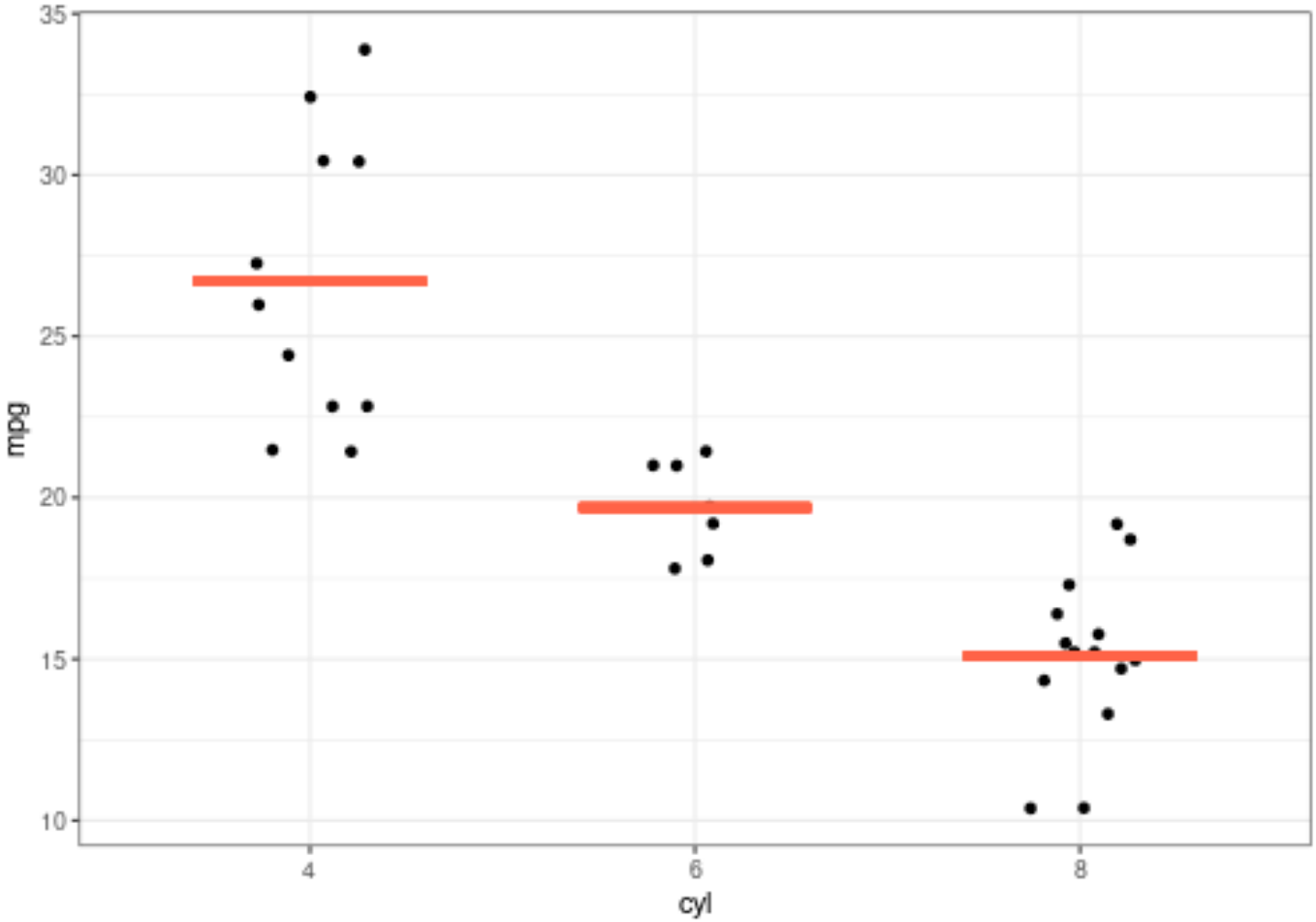


Single Categorical

```
> lm(mpg ~ cyl, mtcars) %>% summary()

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)    26.664      0.972   27.44 < 0.00000000002 ***
cyl6           -6.921      1.558   -4.44              0.00012 ***
cyl8          -11.564      1.299   -8.90    0.00000000086 ***


Residual standard error: 3.22 on 29 degrees of freedom
Multiple R-squared:  0.732, Adjusted R-squared:  0.714
F-statistic: 39.7 on 2 and 29 DF,  p-value: 0.00000000498
```

Cylinder and MPG

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

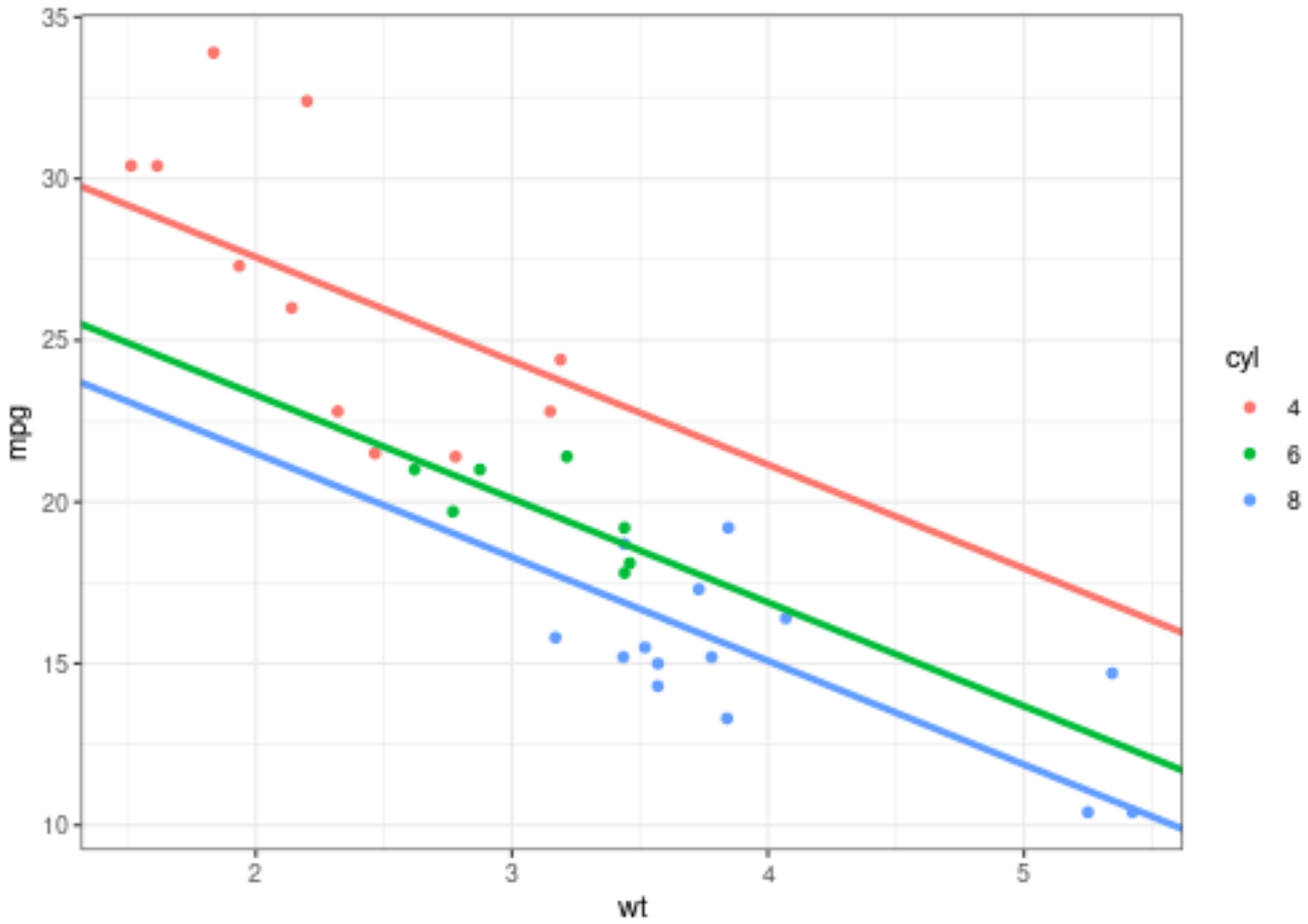


Categorical and Quantitative

```
> lm(mpg ~ wt + cyl, mtcars) %>% summary()

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)   33.991      1.888   18.01 < 0.00000000002 ***
wt            -3.206      0.754   -4.25             0.00021 ***
cyl6          -4.256      1.386   -3.07             0.00472 **
cyl8          -6.071      1.652   -3.67             0.00100 ***


Residual standard error: 2.56 on 28 degrees of freedom
Multiple R-squared:  0.837, Adjusted R-squared:  0.82
F-statistic: 48.1 on 3 and 28 DF,  p-value: 0.0000000000359
```

Cylinder, weight and MPG

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

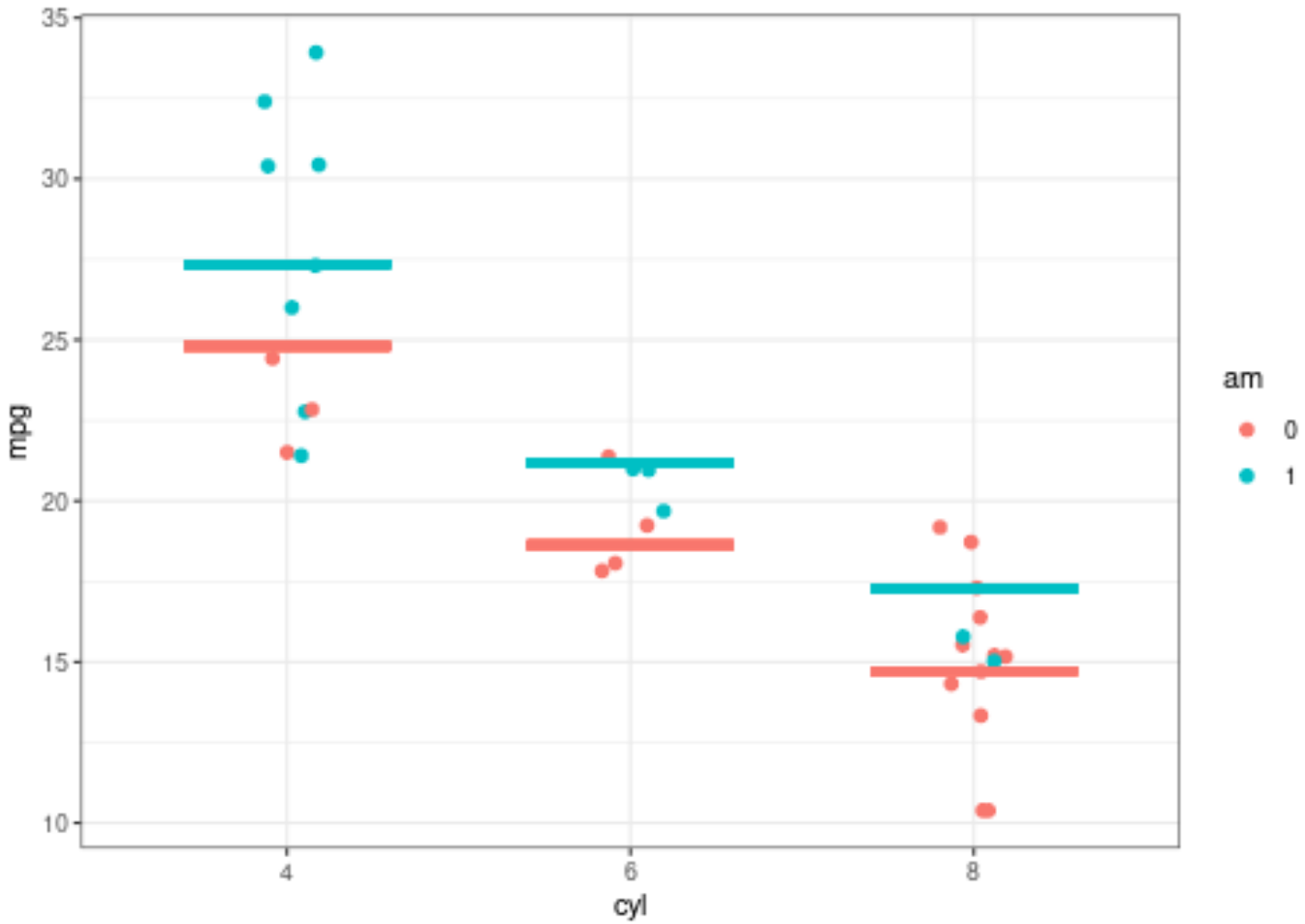


Multiple Categorical

```
> lm(mpg ~ cyl + am, mtcars) %>% summary()

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
(Intercept)    24.80       1.32   18.75 < 0.00000000002 ***
cyl6           -6.16       1.54   -4.01             0.00041 ***
cyl8          -10.07       1.45   -6.93          0.00000015 ***
am1             2.56       1.30    1.97             0.05846 .


Residual standard error: 3.07 on 28 degrees of freedom
Multiple R-squared:  0.765, Adjusted R-squared:  0.74
F-statistic: 30.4 on 3 and 28 DF,  p-value: 0.00000000596
```

Cylinder, transmission and MPG

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

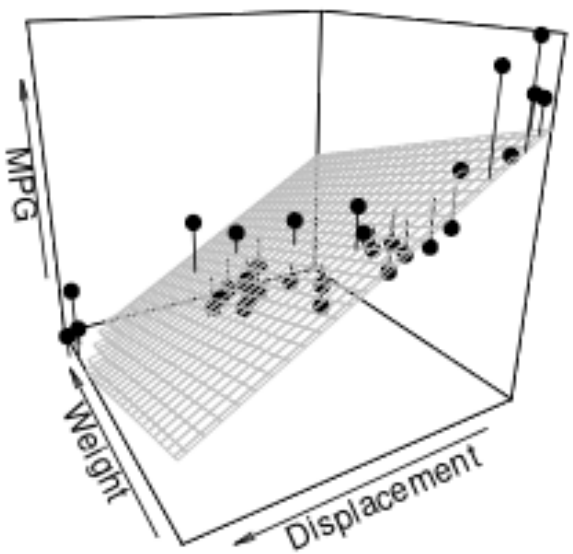


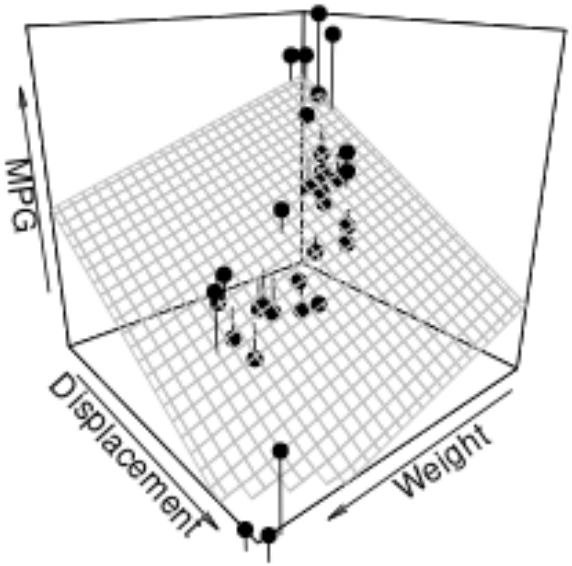
Multiple Quantitative

```
> lm(mpg ~ wt + disp, mtcars) %>% summary()

Coefficients:
             Estimate Std. Error t value          Pr(>|t|)
(Intercept) 34.96055    2.16454   16.15  0.000000049 ***
wt          -3.35083    1.16413   -2.8         0.0074 **
disp        -0.01772    0.00919   -1.93        0.0636 .


Residual standard error: 2.92 on 29 degrees of freedom
Multiple R-squared:  0.781, Adjusted R-squared:  0.766
F-statistic: 51.7 on 2 and 29 DF,  p-value: 0.000000000274
```

Cylinder, transmission and MPG

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



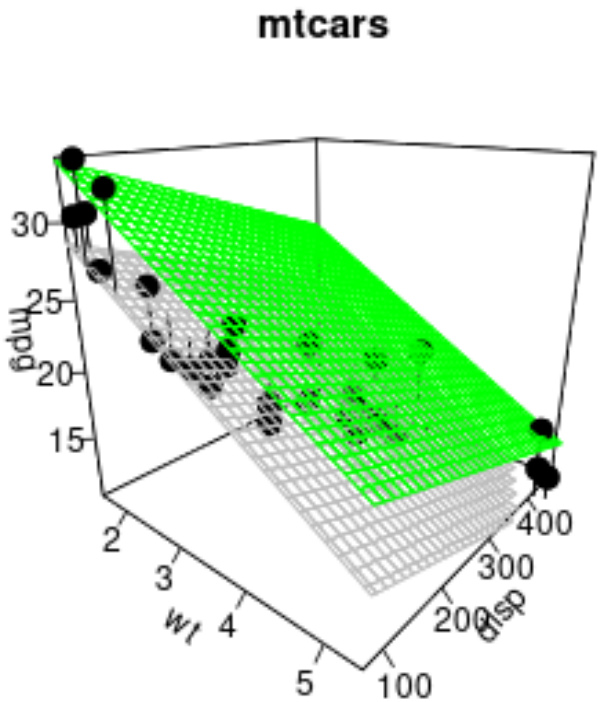


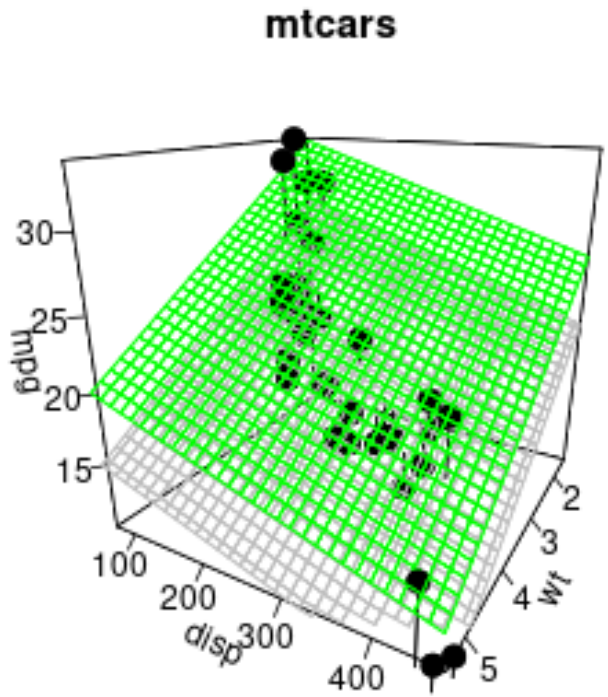
Multiple Quantiative and categorical

```
> lm(mpg ~ wt + disp + am,  mtcars) %>% summary()

Coefficients:
            Estimate Std. Error t value       Pr(>|t|)
(Intercept) 34.67591    3.24061   10.70 0.000000000021 ***
wt          -3.27904    1.32751   -2.47            0.020 *
disp        -0.01780    0.00937   -1.90            0.068 .
am1          0.17772    1.48432    0.12            0.906


Residual standard error: 2.97 on 28 degrees of freedom
Multiple R-squared:  0.781, Adjusted R-squared:  0.758
F-statistic: 33.3 on 3 and 28 DF,  p-value: 0.00000000225
```

Multiple quantiative with categorical

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





Correlated Explanatory Variables

We have been using the fact that there is correlation between the response and explanatory variables

- correlation coefficient (r) measured this for SLR
- larger $|r|$ value $\rightarrow$ larger correlation $\rightarrow$ better predictions

It actually turns out there can be correlation between the explanatory variables too

- this is actually *not* good (bad)
- intuition: if there is correlation between explanatory variables then each explanatory variable tells us info about the other
- this actually means we are effectively using *less* than 2 explanatory variables because the variables have overlapping info about response y

Correlated Explanatory Variables

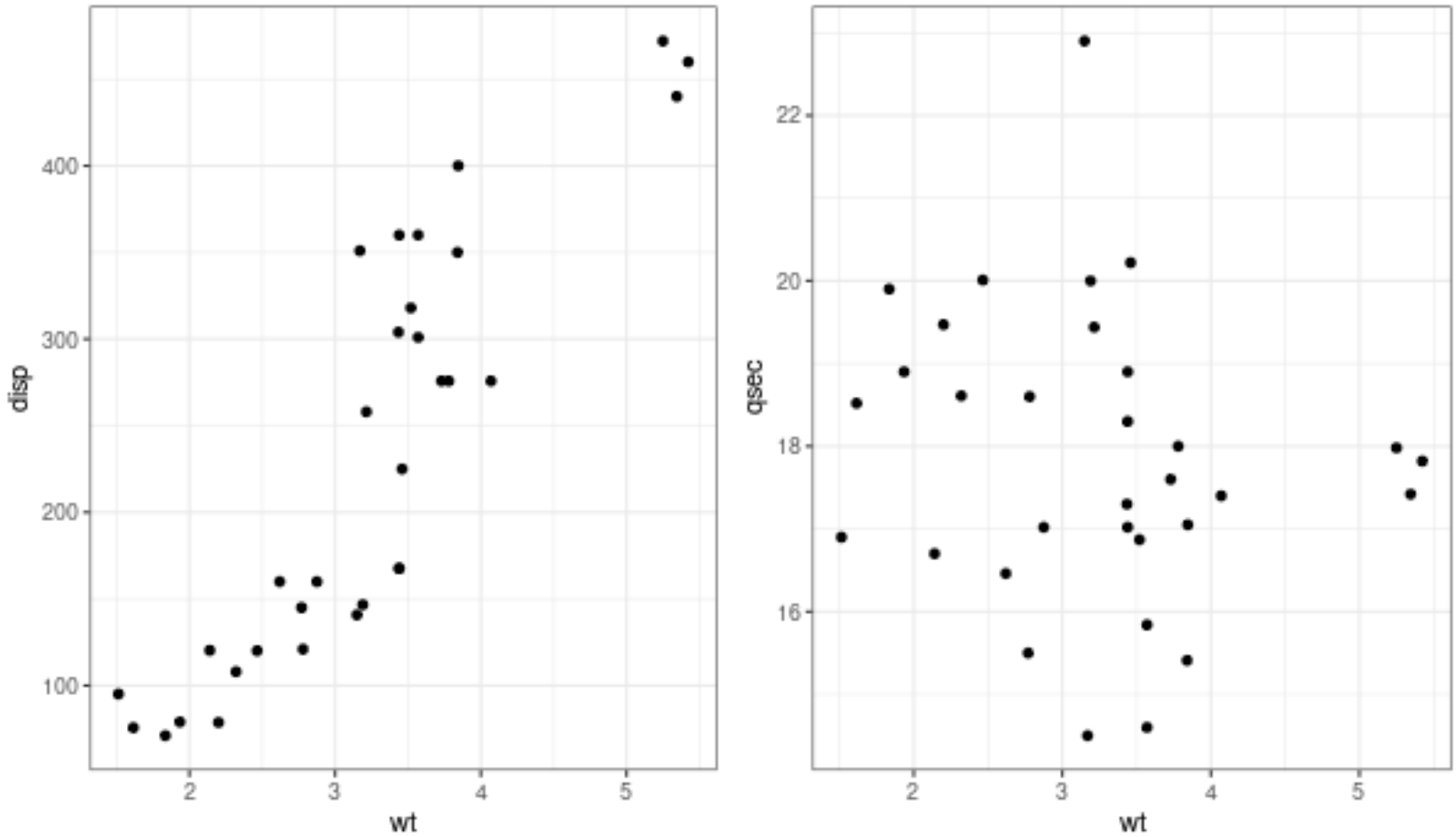
Regression: predicting MPG with Weight

```
> lm(mpg ~ wt, mtcars) %>% summary()


Coefficients:
            Estimate Std. Error t value     Pr(>|t|)
(Intercept)   37.285      1.878   19.86    < 0.000002 ***
wt            -5.344      0.559   -9.56      0.000013 ***
```

Regression: predicting MPG with Weight and Displacement

```
> lm(mpg ~ wt + disp, mtcars) %>% summary()

Coefficients:
            Estimate Std. Error t value      Pr(>|t|)
(Intercept) 34.96055    2.16454   16.15  0.000000049 ***
wt          -3.35083    1.16413   -2.8         0.0074 **
disp        -0.01772    0.00919   -1.93        0.0636 .

> cor(mtcars$wt, mtcars$disp)
0.8879799
```

Correlated Explanatory Variables



Correlated Explanatory Variables

```
> lm(mpg ~ wt, mtcars) %>% summary()

            Estimate Std. Error t value     Pr(>|t|)
(Intercept)   37.285      1.878   19.86   < 0.000002 ***
wt            -5.344      0.559   -9.56     0.000013 ***
R-squared = 0.75

> lm(mpg ~ wt + disp, mtcars) %>% summary()

            Estimate Std. Error t value      Pr(>|t|)
(Intercept) 34.96055    2.16454   16.15  0.000000049 ***
wt          -3.35083    1.16413   -2.8         0.0074 **
disp        -0.01772    0.00919   -1.93        0.0636 .
R-squared = 0.78

> lm(mpg ~ wt + qsec, mtcars) %>% summary()

            Estimate Std. Error t value       Pr(>|t|)
(Intercept)   19.746      5.252    3.76        0.00077 ***
wt            -5.048      0.484  -10.43 0.000000000025 ***
qsec           0.929      0.265    3.51        0.00150 **
R-squared = 0.82
```

Key Takeaways

- Quantitative variables represent slopes (changes in X lead to β changes in y)
- Categorical variables represent horizontal shifts
- Any number of categorical or quantiative variables can be added to model
- Always interpret regression coefficients as *everything else being fixed*
- Look out for correlated variables
  - makes models harder to interpret and usually don't improve prediction much