Linear models 2

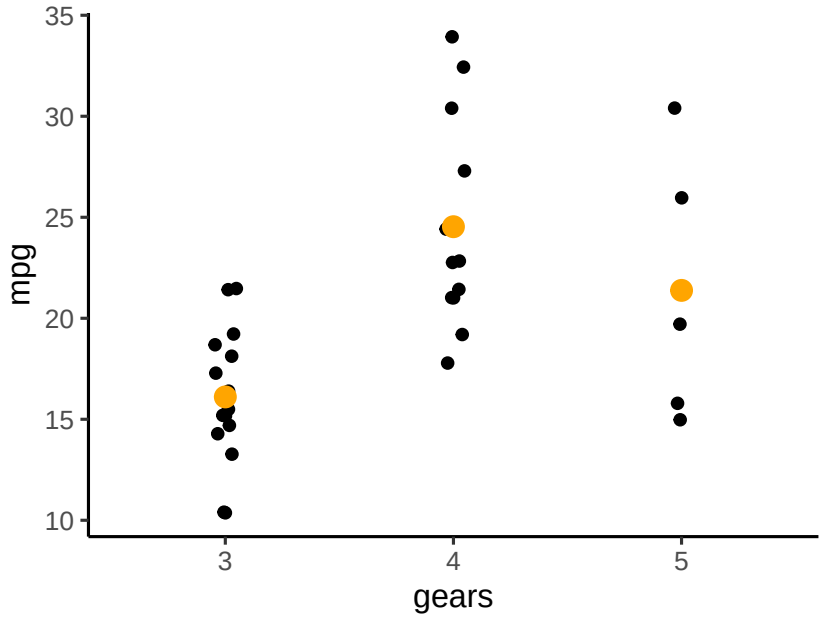
More bells and whistles

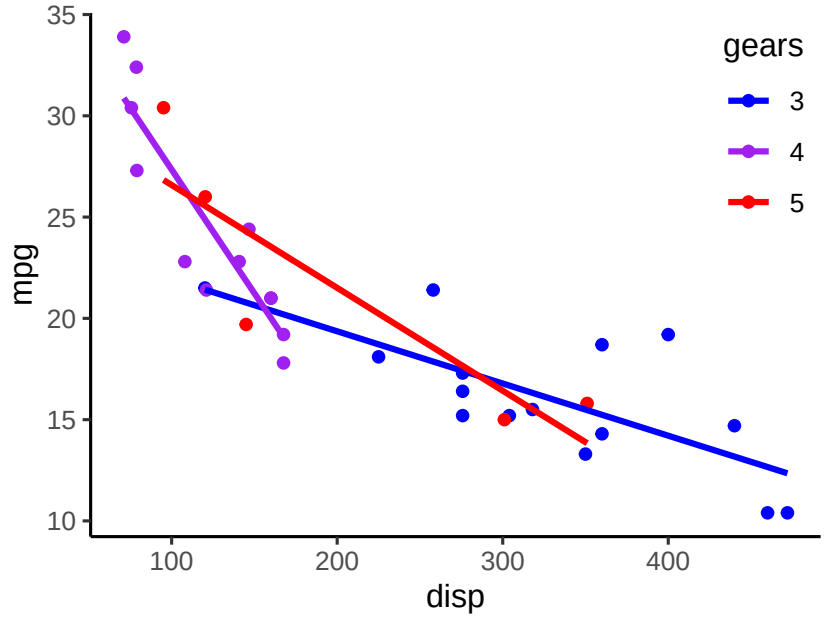
Samuel Robinson, Ph.D.

October 15, 2020

Motivation

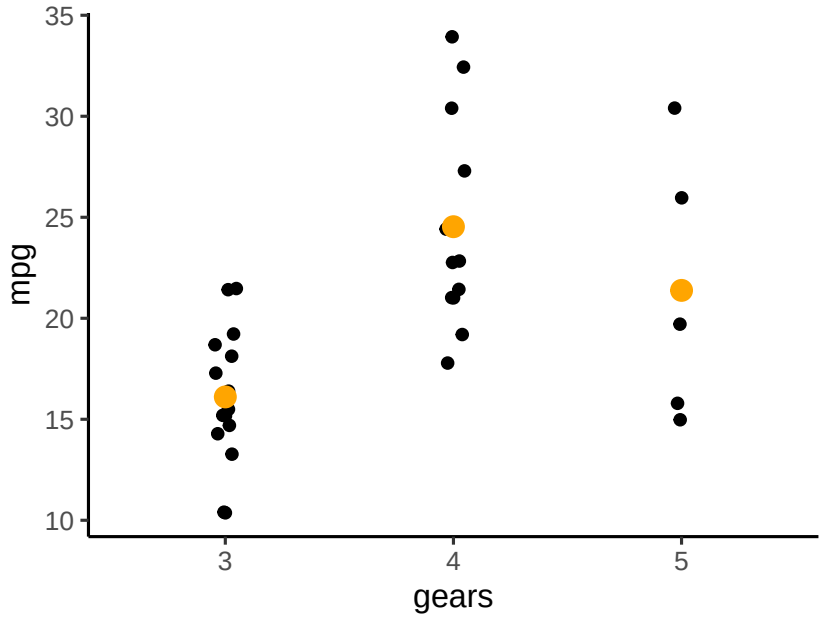
- *I have 2+ groups of data, and I want to know whether the means are different*

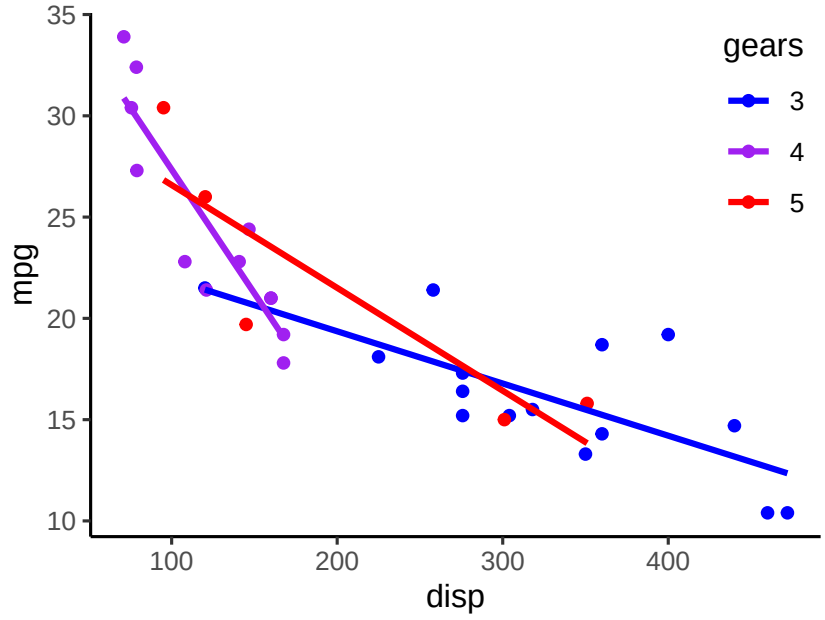




Motivation

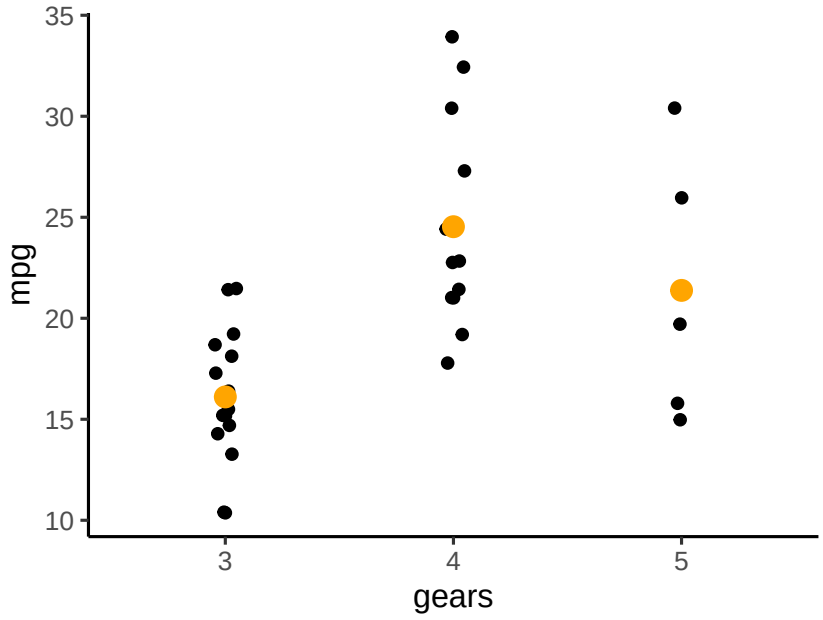
- *I have 2+ groups of data, and I want to know whether the means are different*
- *I have 2+ groups of bivariate data, and I want to know whether the relationships differ between groups*

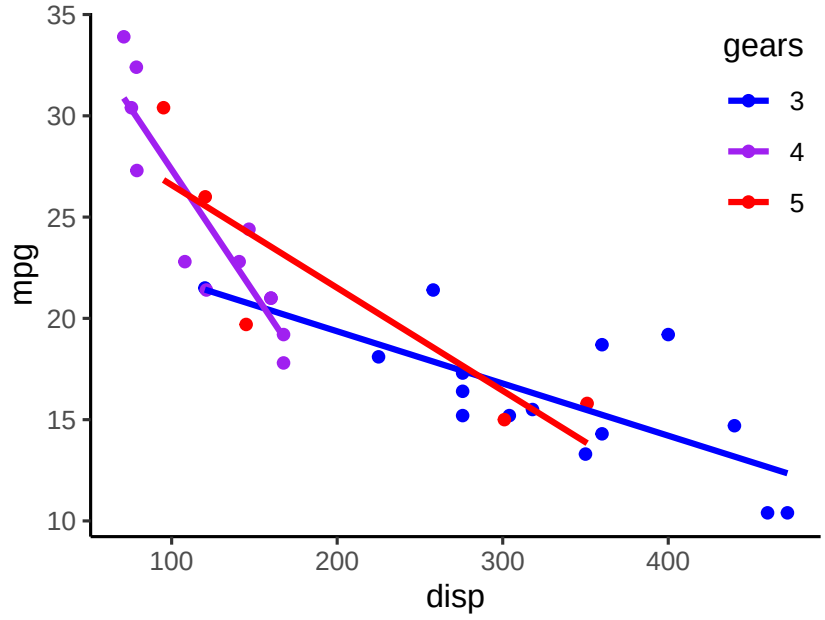




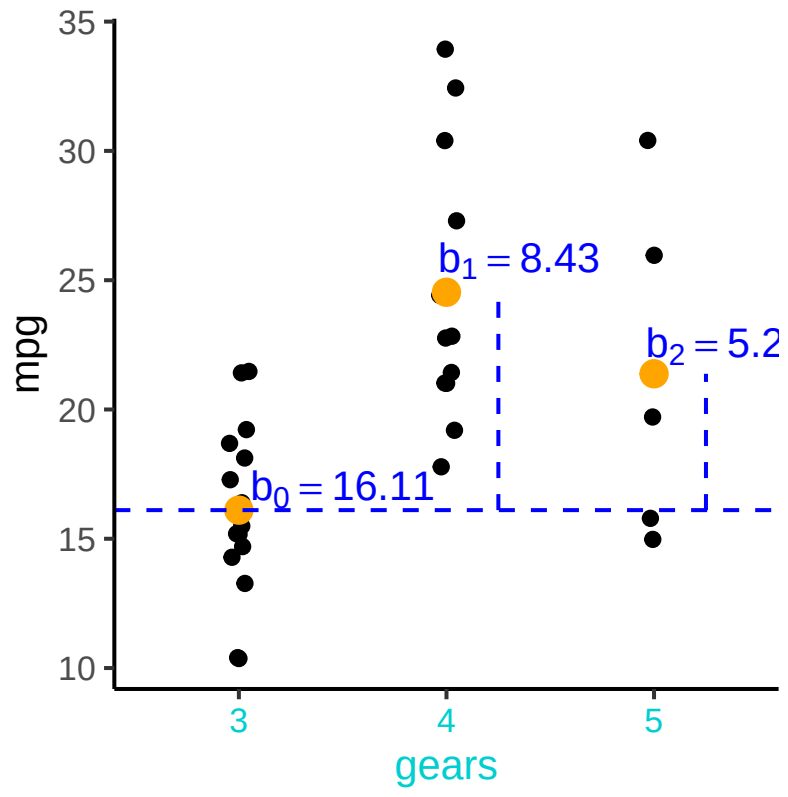
Motivation

- *I have 2+ groups of data, and I want to know whether the means are different*
- *I have 2+ groups of bivariate data, and I want to know whether the relationships differ between groups*
- **How do we know if any of this matters?**





Categorial data, 3 categories



$$\hat{mpg} = b_0 + b_1 gears_4 + b_2 gears_5$$

$$mpg \sim Normal(\hat{mpg}, \sigma)$$

The more factor levels, the more coefficients:

- *mpg* is the thing you're interested in predicting
- $\hat{mpg}$ is the *predicted value* of *mpg*
- *gear* is the *predictor* of *mpg*
  - set of 0s and 1s
  - $gears_4$ = "is this data point from a 4-gear car?"
- b_0 = *intercept*
- $[b_1, b_2]$ = are *coefficients* for *gears*

How do I get R to fit this model?

```r
#Formula structure: y ~ x
mod1 <- lm(mpg ~ factor(gear), #mpg depends on gears
           data = mtcars) #Name of the dataframe containing mpg & gears
summary(mod1)
```

```
##
## Call:
## lm(formula = mpg ~ factor(gear), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7333 -3.2333 -0.9067  2.8483  9.3667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    16.107      1.216  13.250 7.87e-14 ***
## factor(gear)4   8.427      1.823   4.621 7.26e-05 ***
## factor(gear)5   5.273      2.431   2.169   0.0384 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.708 on 29 degrees of freedom
## Multiple R-squared:  0.4292, Adjusted R-squared:  0.3898
## F-statistic:  10.9 on 2 and 29 DF,  p-value: 0.0002948
```

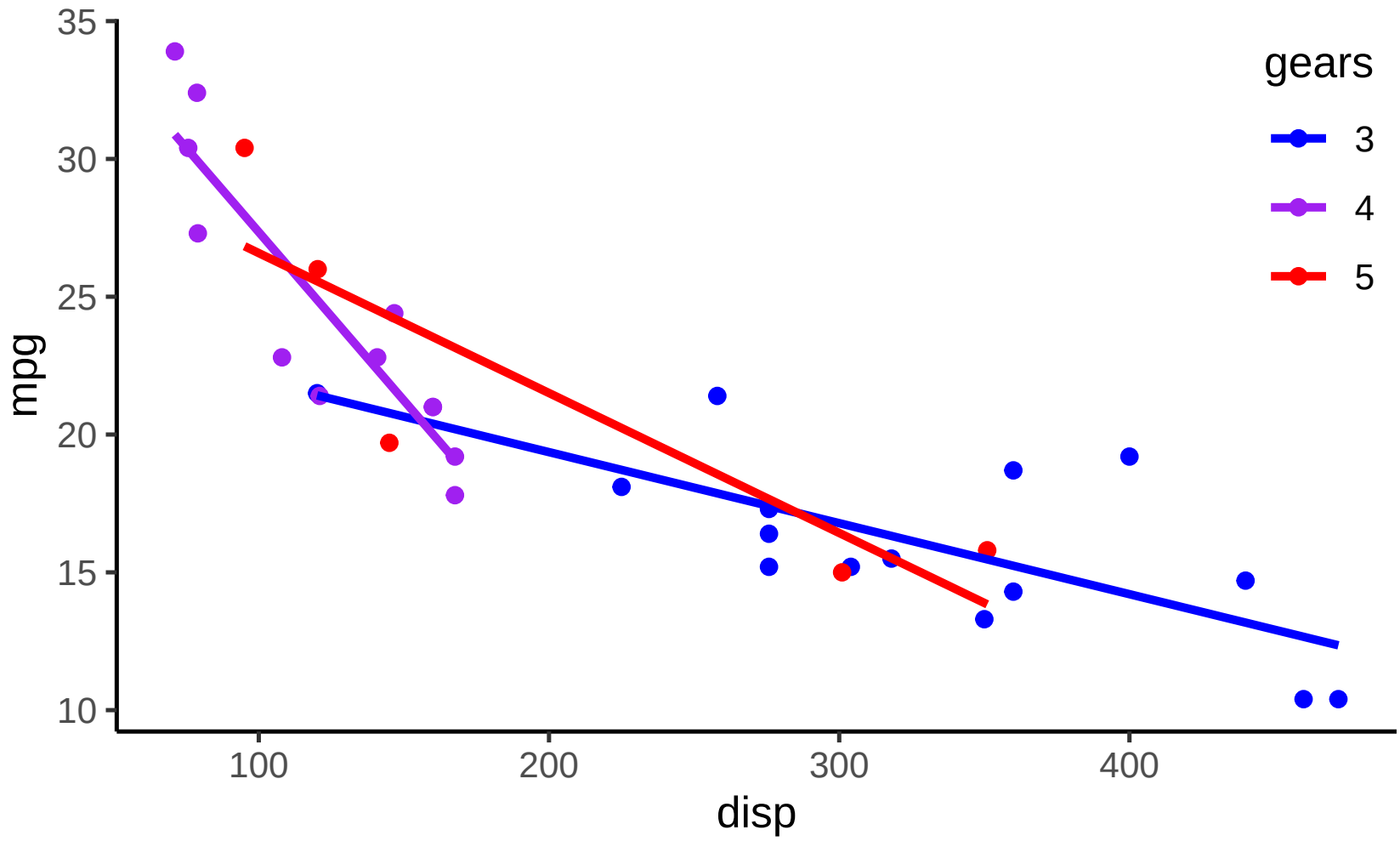
Dummy variables

```
mod1Matrix <- model.matrix(mod1) #Get model matrix (columns used to predict mpg)
head(mod1Matrix,28) #Show first 28 rows of model matrix
```

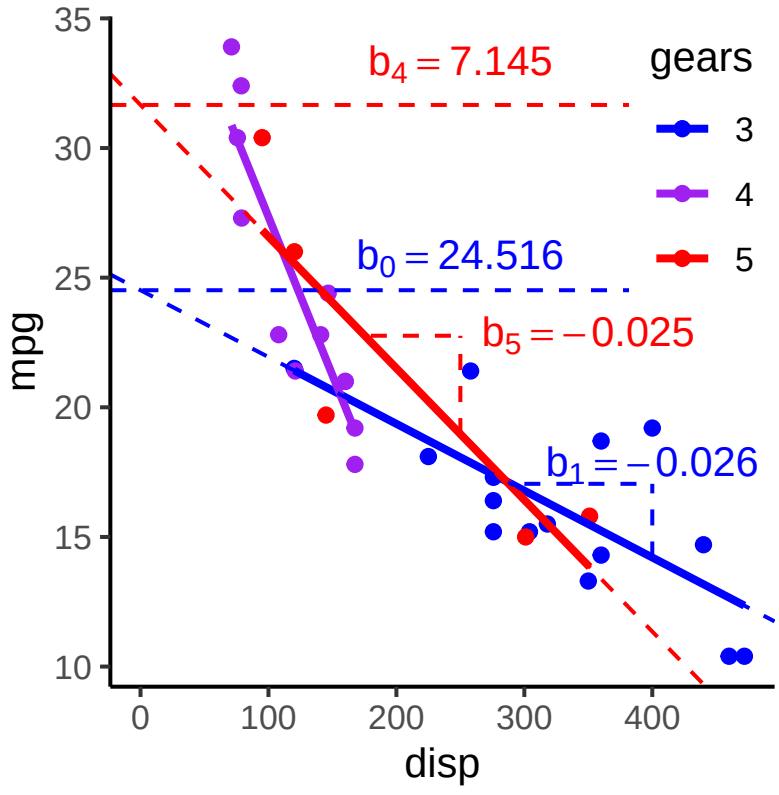
```
##                   (Intercept) factor(gear)4 factor(gear)5
## Mazda RX4                   1              1              0
## Mazda RX4 Wag               1              1              0
## Datsun 710                  1              1              0
## Hornet 4 Drive              1              0              0
## Hornet Sportabout           1              0              0
## Valiant                     1              0              0
## Duster 360                  1              0              0
## Merc 240D                   1              1              0
## Merc 230                    1              1              0
## Merc 280                    1              1              0
## Merc 280C                   1              1              0
## Merc 450SE                  1              0              0
## Merc 450SL                  1              0              0
## Merc 450SLC                 1              0              0
## Cadillac Fleetwood          1              0              0
## Lincoln Continental         1              0              0
## Chrysler Imperial           1              0              0
## Fiat 128                    1              1              0
## Honda Civic                 1              1              0
## Toyota Corolla              1              1              0
## Toyota Corona               1              0              0
## Dodge Challenger            1              0              0
## AMC Javelin                 1              0              0
## Camaro Z28                  1              0              0
## Pontiac Firebird            1              0              0
## Fiat X1-9                   1              1              0
## Porsche 914-2               1              0              1
## Lotus Europa                1              0              1
```

Interactions

What if the slopes *and* intercepts differ between groups?



Interactions



$$\hat{mpg} = b_0 + b_1 disp + b_2 gears_4 + b_3 gears_5 + b_4(disp \times gears_4) + b_5(disp \times gears_5)$$

$$mpg \sim Normal(\hat{mpg}, \sigma)$$

- Interactions occur when predictors are *multiplied*
- In this case, $disp$ is multiplied by $gears_4$ and $gears_5$

How do I get R to fit this model?

```r
#Formula structure: y ~ x
mod2 <- lm(mpg ~ disp*factor(gear), #mpg depends on disp interacted with gears
           data = mtcars) #Name of the dataframe
summary(mod2)
```

```
##
## Call:
## lm(formula = mpg ~ disp * factor(gear), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.5986 -1.5990 -0.0143  1.6329  4.9926
##
## Coefficients:
##                      Estimate Std. Error t value Pr(>|t|)
## (Intercept)        24.515566   2.462431   9.956 2.32e-10 ***
## disp               -0.025770   0.007265  -3.547 0.001505 **
## factor(gear)4      15.051963   3.558043   4.230 0.000256 ***
## factor(gear)5       7.145380   3.535913   2.021 0.053711 .
## disp:factor(gear)4 -0.096442   0.021261  -4.536 0.000114 ***
## disp:factor(gear)5 -0.025005   0.013320  -1.877 0.071742 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2.579 on 26 degrees of freedom
## Multiple R-squared:  0.8465, Adjusted R-squared:  0.817
## F-statistic: 28.67 on 5 and 26 DF,  p-value: 8.452e-10
```

Dummy variables

```
mod2Matrix <- model.matrix(mod2) #Get model matrix (columns used to predict mpg)
colnames(mod2Matrix) <- gsub('factor\\(gear\\)','gear',colnames(mod2Matrix)) #Shorten colnames
head(mod2Matrix,28) #Show first 28 rows of model matrix
```

```
##                   (Intercept)  disp gear4 gear5 disp:gear4 disp:gear5
## Mazda RX4                   1 160.0     1     0      160.0        0.0
## Mazda RX4 Wag               1 160.0     1     0      160.0        0.0
## Datsun 710                  1 108.0     1     0      108.0        0.0
## Hornet 4 Drive              1 258.0     0     0        0.0        0.0
## Hornet Sportabout           1 360.0     0     0        0.0        0.0
## Valiant                     1 225.0     0     0        0.0        0.0
## Duster 360                  1 360.0     0     0        0.0        0.0
## Merc 240D                   1 146.7     1     0      146.7        0.0
## Merc 230                    1 140.8     1     0      140.8        0.0
## Merc 280                    1 167.6     1     0      167.6        0.0
## Merc 280C                   1 167.6     1     0      167.6        0.0
## Merc 450SE                  1 275.8     0     0        0.0        0.0
## Merc 450SL                  1 275.8     0     0        0.0        0.0
## Merc 450SLC                 1 275.8     0     0        0.0        0.0
## Cadillac Fleetwood          1 472.0     0     0        0.0        0.0
## Lincoln Continental         1 460.0     0     0        0.0        0.0
## Chrysler Imperial           1 440.0     0     0        0.0        0.0
## Fiat 128                    1  78.7     1     0       78.7        0.0
## Honda Civic                 1  75.7     1     0       75.7        0.0
## Toyota Corolla              1  71.1     1     0       71.1        0.0
## Toyota Corona               1 120.1     0     0        0.0        0.0
## Dodge Challenger            1 318.0     0     0        0.0        0.0
## AMC Javelin                 1 304.0     0     0        0.0        0.0
## Camaro Z28                  1 350.0     0     0        0.0        0.0
## Pontiac Firebird            1 400.0     0     0        0.0        0.0
## Fiat X1-9                   1  79.0     1     0       79.0        0.0
## Porsche 914-2               1 120.3     0     1        0.0      120.3
## Lotus Europa                1  95.1     0     1        0.0       95.1
```

How do I know if any of this matters?

- drop-1 (Type III) ANOVA for *entire factors*
  - e.g. "Does adding gear matter?"
- Wald t-scores/Z-scores for *levels of factors*
  - e.g. "Is gear3 different from gear4?"
- **p-values are only meaningful if the model assumptions are valid**

drop-1 ANOVA

```r
#mpg depends on gears
mod1 <- lm(mpg ~ factor(gear), data = mtcars)
drop1(mod1,test='F') #Effect of gears is very strong
```

```
## Single term deletions
##
## Model:
## mpg ~ factor(gear)
##              Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                     642.8 102.00
## factor(gear)  2    483.24 1126.0 115.94  10.901 0.0002948 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#mpg depends on disp
mod2 <- lm(mpg ~ disp, data = mtcars)
drop1(mod2,test='F') #Effect of disp is also very strong
```

```
## Single term deletions
##
## Model:
## mpg ~ disp
##        Df Sum of Sq     RSS     AIC F value   Pr(>F)
## <none>               317.16  77.397
## disp    1    808.89 1126.05 115.943  76.513 9.38e-10 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

drop-1 ANOVA

```r
#mpg depends on disp and gear
mod3 <- lm(mpg ~ disp + factor(gear), data = mtcars)
drop1(mod3,test='F') #Effect of disp is very strong, and erases the effect of gear
```

```
## Single term deletions
##
## Model:
## mpg ~ disp + factor(gear)
##              Df Sum of Sq    RSS     AIC F value    Pr(>F)
## <none>                    317.01  81.383
## disp          1    325.79 642.80 102.003 28.7755 1.025e-05 ***
## factor(gear)  2      0.15 317.16  77.397  0.0065    0.9935
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
#mpg depends on disp interacted with gear
mod4 <- lm(mpg ~ disp*factor(gear), data = mtcars)
drop1(mod4,test='F') #Interaction effect is strong. Why are disp and gear not shown?
```

```
## Single term deletions
##
## Model:
## mpg ~ disp * factor(gear)
##                   Df Sum of Sq    RSS    AIC F value    Pr(>F)
## <none>                         172.87 65.978
## disp:factor(gear)  2    144.14 317.01 81.383  10.839 0.0003771 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Wald t-scores

- Wald t-scores are shown in model `summary`
- t-score (aka Z-score) = mean÷SD
- p-value comes from Student's t-distribution (similar to Normal, but has longer tails depending on sample size)

```
summary(mod1)
```

```
##
## Call:
## lm(formula = mpg ~ factor(gear), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7333 -3.2333 -0.9067  2.8483  9.3667
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)    16.107      1.216  13.250 7.87e-14 ***
## factor(gear)4   8.427      1.823   4.621 7.26e-05 ***
## factor(gear)5   5.273      2.431   2.169   0.0384 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 4.708 on 29 degrees of freedom
## Multiple R-squared:  0.4292, Adjusted R-squared:  0.3898
## F-statistic:  10.9 on 2 and 29 DF,  p-value: 0.0002948
```

Comparing between intercepts

- If you're comparing between many intercepts, you need to account for *multiple comparisons*
- One common method: Tukey's Honestly Significant Difference (HSD)

```
library(multcomp) #Loads the multcomp package (needs to be installed first)
mod1Comp <- glht(mod1, linfct = mcp('factor(gear)'='Tukey')) #Fits multcomp object using gear
summary(mod1Comp)
```

```
##
##   Simultaneous Tests for General Linear Hypotheses
##
## Multiple Comparisons of Means: Tukey Contrasts
##
##
## Fit: lm(formula = mpg ~ factor(gear), data = mtcars)
##
## Linear Hypotheses:
##             Estimate Std. Error t value Pr(>|t|)
## 4 - 3 == 0    8.427      1.823   4.621   <0.001 ***
## 5 - 3 == 0    5.273      2.431   2.169   0.0919 .
## 5 - 4 == 0   -3.153      2.506  -1.258   0.4255
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
## (Adjusted p values reported -- single-step method)
```