Linear models 2

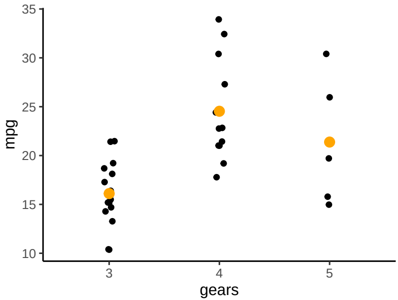
More bells and whistles

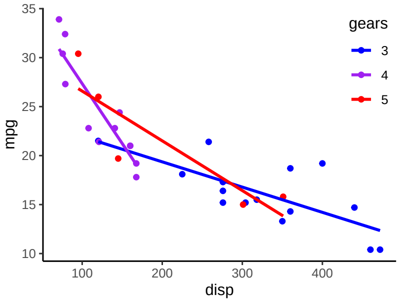
Samuel Robinson, Ph.D.

October 15, 2020

Motivation

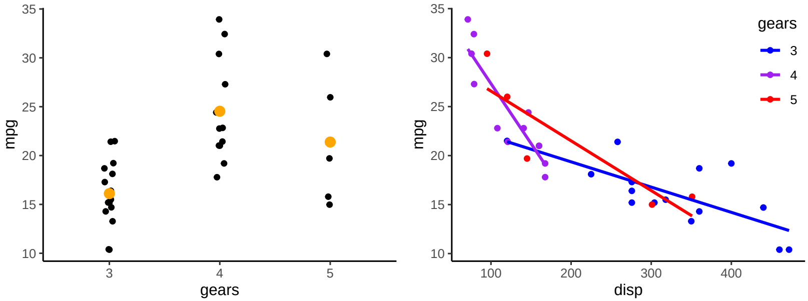
- *I have 2+ groups of data, and I want to know whether the means are different*



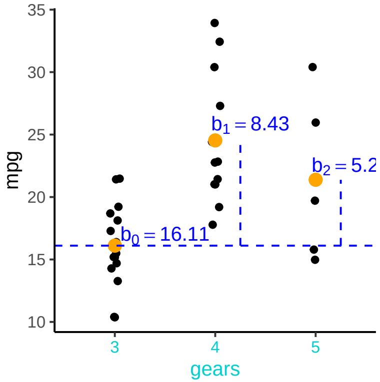


Motivation

- *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*



Categorial data, 3 categories



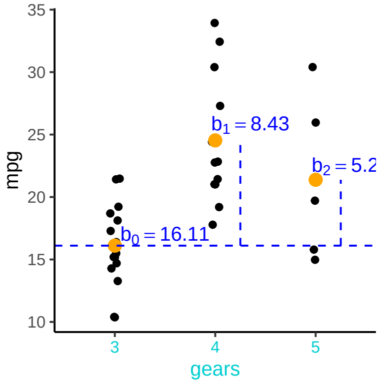
The more factor levels, the more coefficients:

- *mpg* is the thing you're interested in predicting

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

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

Categorial data, 3 categories



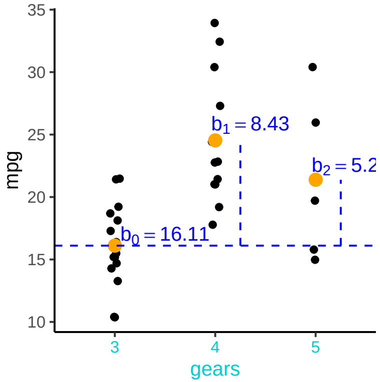
The more factor levels, the more coefficients:

- *mpg* is the thing you're interested in predicting
- $\hat{mpg}$ is the *predicted value* of *mpg*

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

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

Categorial data, 3 categories



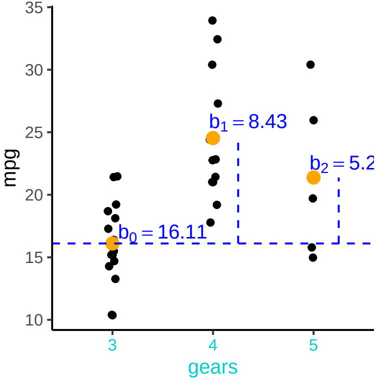
The more factor levels, the more coefficients:

- *mpg* is the thing you're interested in predicting
- $\hat{mpg}$ is the *predicted value* of *mpg*
- *gear* is the *predictor* of *mpg*

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

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

Categorial data, 3 categories



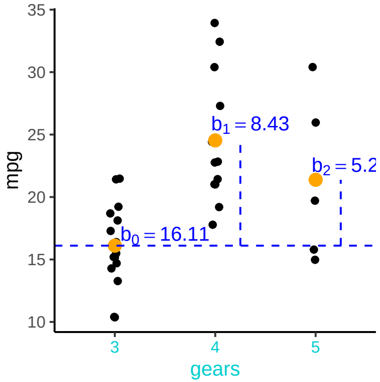
The more factor levels, the more coefficients:

- *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

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

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

Categorial data, 3 categories



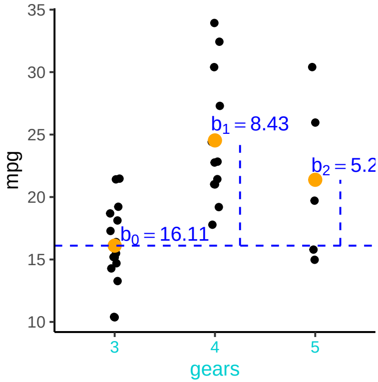
The more factor levels, the more coefficients:

- *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?"

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

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

Categorial data, 3 categories



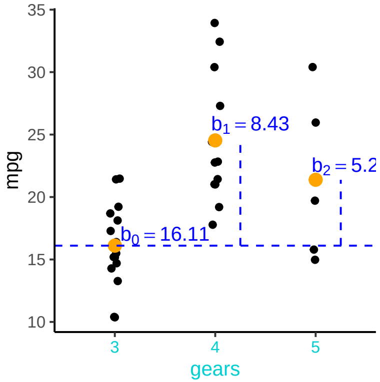
The more factor levels, the more coefficients:

- *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*

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

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

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)
```

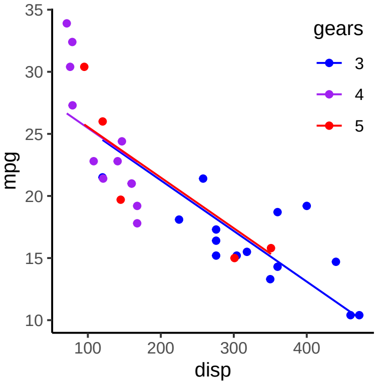
```
##
## Call:
## lm(formula = mpg ~ factor(gear), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -6.7333 -3.2333 -0.9067  2.8483  9.3667
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept)     16.107      1.216  13.250 7.87e-14 ***
## 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
```

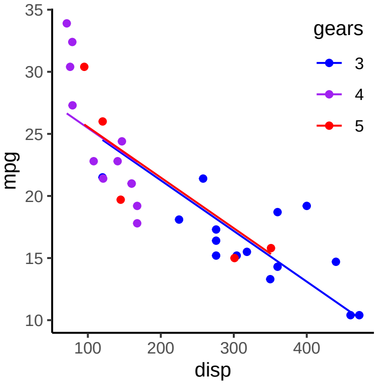
What about if 2 things are both important?



- Suppose that both *disp* and *gears* are important for predicting *mpg*?

$$\hat{mpg} = b_0 + b_1 disp$$
$$+ b_2 gears_4 + b_3 gears_5$$
$$mpg \sim Normal(\hat{mpg}, \sigma)$$

What about if 2 things are both important?

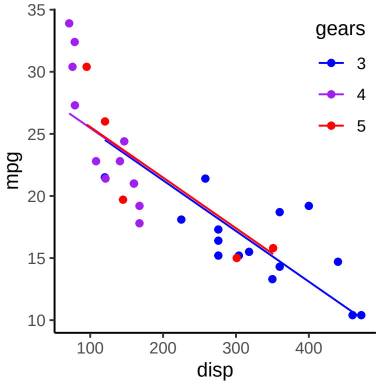


- Suppose that both *disp* and *gears* are important for predicting *mpg*?
- This is very similar to the last example, except that now we've added *disp*

$$\hat{mpg} = b_0 + b_1 disp + b_2 gears_4 + b_3 gears_5$$

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

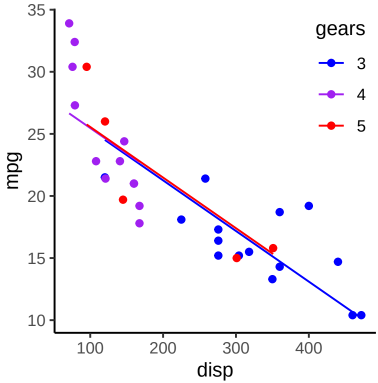
What about if 2 things are both important?



- Suppose that both *disp* and *gears* are important for predicting *mpg*?
- This is very similar to the last example, except that now we've added *disp*
- *gears* now changes the intercept, while *disp* changes the slope of all the lines

$$\hat{mpg} = b_0 + b_1 disp$$
$$+ b_2 gears_4 + b_3 gears_5$$
$$mpg \sim Normal(\hat{mpg}, \sigma)$$

What about if 2 things are both important?



- Suppose that both *disp* and *gears* are important for predicting *mpg*?
- This is very similar to the last example, except that now we've added *disp*
- *gears* now changes the intercept, while *disp* changes the slope of all the lines
- Does it look like *gear* is very important?

$$\hat{mpg} = b_0 + b_1 disp + b_2 gears_4 + b_3 gears_5$$

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

How do I get R to fit this model?

```
#mpg depends on disp and gears
mod2 <- lm(mpg ~ disp+factor(gear), data = mtcars)
summary(mod2)
```

```
##
## Call:
## lm(formula = mpg ~ disp + factor(gear), data = mtcars)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -4.9155 -2.1892 -0.9054  1.5790  7.2498
##
## Coefficients:
##                Estimate Std. Error t value Pr(>|t|)
## (Intercept)   29.411183   2.627966  11.192 7.58e-12 ***
## disp          -0.040774   0.007601  -5.364 1.03e-05 ***
## factor(gear)4  0.138017   2.021332   0.068    0.946
## factor(gear)5  0.224712   1.976090   0.114    0.910
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.365 on 28 degrees of freedom
## Multiple R-squared:  0.7185, Adjusted R-squared:  0.6883
## F-statistic: 23.82 on 3 and 28 DF,  p-value: 7.31e-08
```

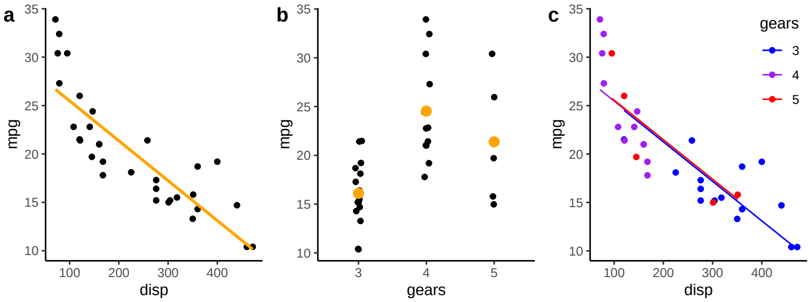
Dummy variables

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

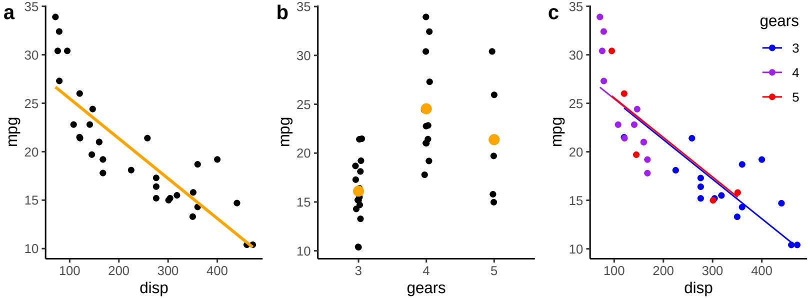
Interlude: problems with plotting raw data

- Say that I've fit the following model:
  `mpg ~ disp + factor(gear)`



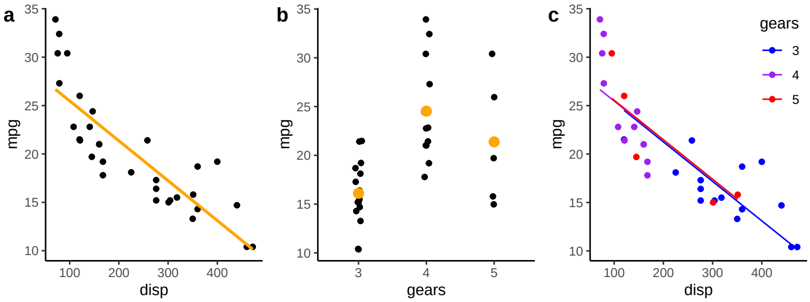
Interlude: problems with plotting raw data

- Say that I've fit the following model:
  `mpg ~ disp + factor(gear)`
- All of the plots below are using raw data, but which one is "telling the truth"?



Interlude: problems with plotting raw data

- Say that I've fit the following model:
  `mpg ~ disp + factor(gear)`
- All of the plots below are using raw data, but which one is "telling the truth"?
- Answer: **c**. *a* and *b* are hiding the effect of the other variable



How do I plot these model results?

Rule for plotting model results:

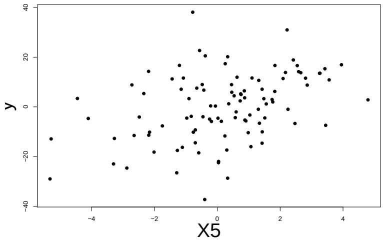
1. If the model uses `N` variables, you should show all `N` effects *simultaneously*
2. If this is impractical, you should use a **partial effects plot**

Other names for partial effects:

- *counterfactual* plot, *predictor effect* plot, *leverage* plot
- Try using `effects` or `ggeffects`. Requires the `effects` and `ggeffect` packages

Incorrect example, using raw data:

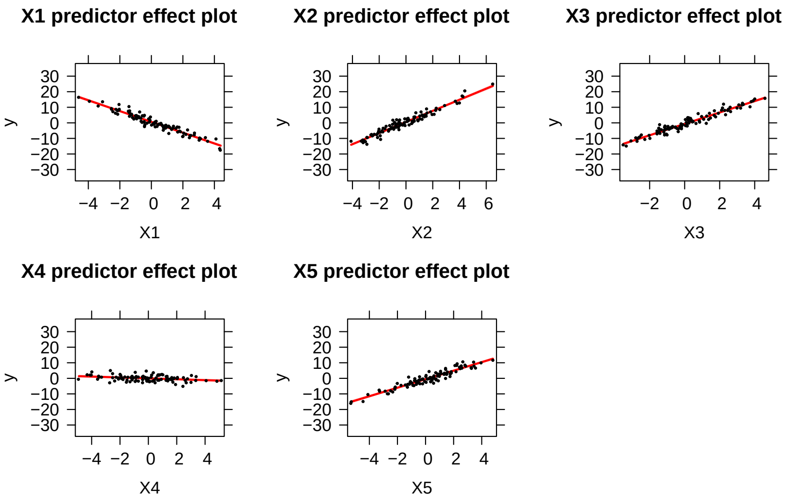
```
#Fit model with 5 variables (all important)
simMod <- lm(y~X1+X2+X3+X4+X5,data=pred)
#Incorrect way, using raw data
plot(y~X5,data=pred,pch=19,cex.lab=3)
```



The effect of *X5* is actually **very** strong ($p > 0.0001$), but it doesn't look like it from this plot!

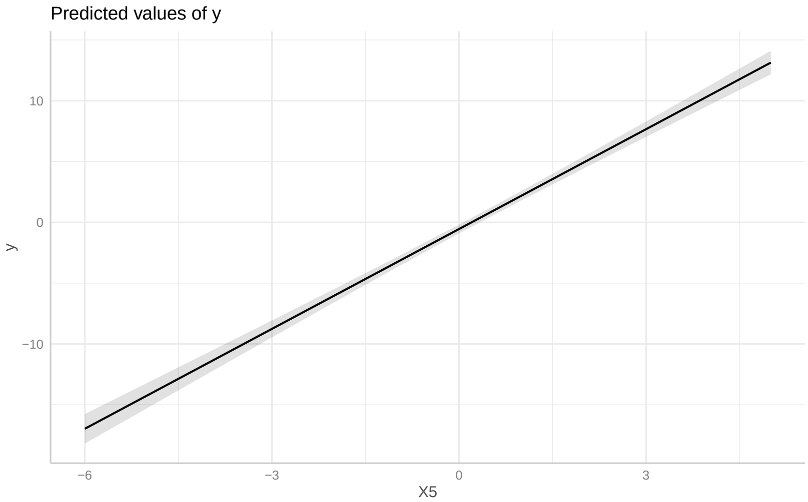
Partial effects plots - using *effects*

```
library(effects) #Load effects package
simModEff <- predictorEffects(simMod,partial.residuals=TRUE) #Calculate partial effects
#Plot partial effects
plot(simModEff,lines=list(col='red'), partial.residuals=list(pch=19,col='black',cex=0.25))
```



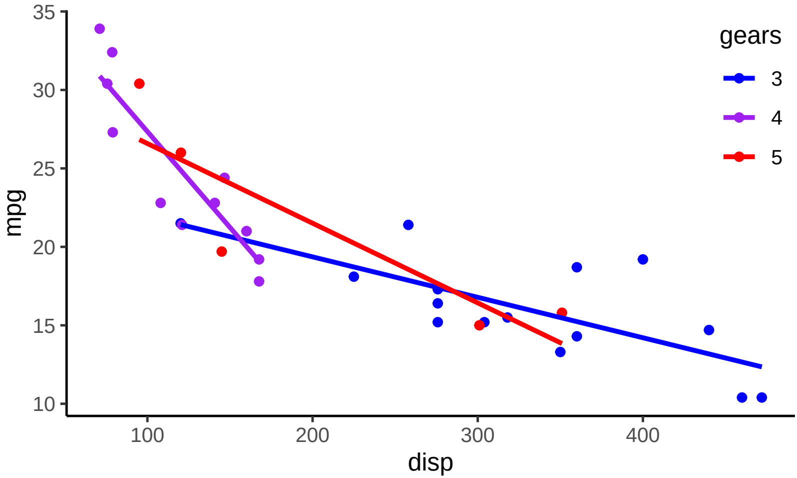
Partial effects plots - using *ggpredict*

```
library(ggeffects) #Load ggeffects package
simModEff2 <- ggeffect(simMod,terms=c('X5')) #Calculate partial effects for X5
plot(simModEff2) #Plot effect of X5
```

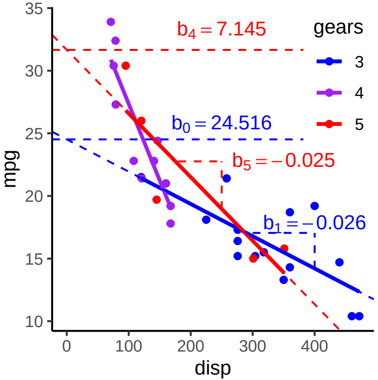


Interactions

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



Interactions

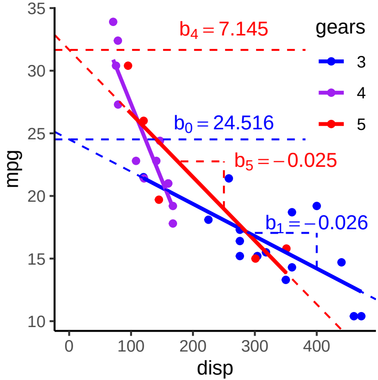


$$
\begin{aligned}
\hat{mpg} &= b_0 + b_1 disp \\
&\quad + b_2 gears_4 + b_3 gears_5 \\
&\quad + b_4 (disp \times gears_4) \\
&\quad + b_5 (disp \times gears_5)
\end{aligned}
$$

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

- Interactions occur when predictors are *multiplied*

Interactions

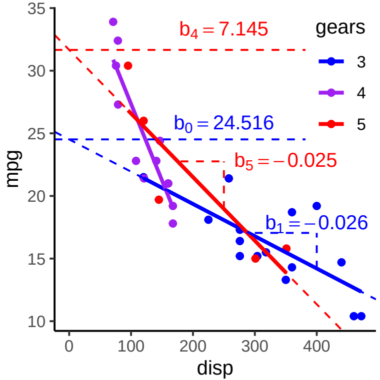


$$\begin{aligned}\hat{mpg} &= b_0 + b_1 disp \\ &\quad + b_2 gears_4 + b_3 gears_5 \\ &\quad + b_4 (disp \times gears_4) \\ &\quad + b_5 (disp \times gears_5)\end{aligned}$$

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

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

Interactions



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

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

- Interactions occur when predictors are *multiplied*
- In this case, *disp* is multiplied by $gears_4$ and $gears_5$
- *gears* now changes the intercept and the slope of the relationship between *mpg* and *disp*

How do I get R to fit this model?

```
#mpg depends on disp interacted (*) with gears
mod2 <- lm(mpg ~ disp*factor(gear), data = mtcars)
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
```

Beware of fitting too many interactions, or else the *Bilbo effect* occurs!

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
```

A challenger approaches!

- Since you're all bat folks, here's some bat data!
  - `batDat.csv`
- Data: 100 bat weights from 2 cities, recorded along with sex and age
- How do these variables affect bat weight?
  - Think about how these variables might be related to weight using your `brain`
  - Fit a model using `lm`
  - Make some plots, using `effects` or `ggeffects`