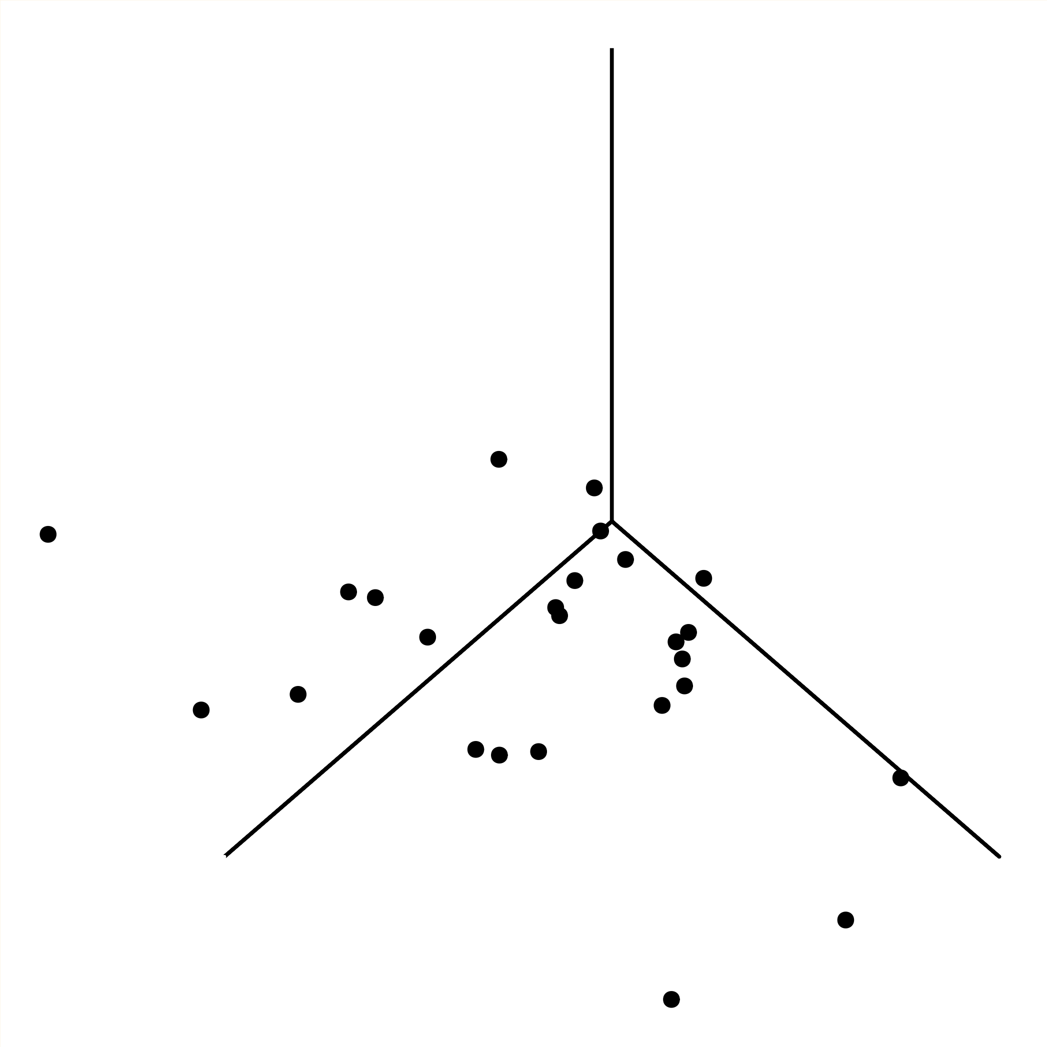
Multiple Linear Regression (MLR) Models

Stat 230

April 13 2022

Overview



Today:

- Multiple predictors
- Quadratic predictors
- Interactions of predictors
- Interpretation

MLR Variables

Y = quantitative response

- $x_1, \ldots, x_p$: p explanatory (predictor) variables
- x_j can be either quantitative or categorical
- we will cover categorical predictors in another lecture!

Statistical Modeling

$$Y_i = \mu(Y \mid x) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma)$$

Simple Linear Regression model mean function:

$$\mu(Y \mid x) = \beta_0 + \beta_1 x$$

Multiple Linear Regression (MLR) model mean function:

$$\mu(Y \mid x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots$$

MLR: Basic model

$$\mu(Y \mid x) = \mu_{Y|x_1,...,x_p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots \beta_p x_p$$

- General: β_j is the change in the mean response for a one unit increase in x_j holding all other predictors fixed.
- β_0 : mean response when all predictor values are 0

MLR: Presence of logged variables

If Y is logged, then any changes in predictors result in a multiplicative change in the median of Y

- If x is unlogged, this interpretation is like the SLR **exponential model**
- If x is also logged, this interpretation is like the SLR **power model**

MLR: Model interpretation

- β_j have the same basic interpretation as in a SLR, holding all other predictors constant!
- E.g. Holding all other predictors fixed, increasing x_1 by 1 unit results in a β_1 change in the mean of Y

$$\begin{aligned}
\mu\left(y \mid x_1+1, x_2, \ldots, x_p\right) &= \beta_0+\beta_1\left(x_1+1\right)+\beta_2 x_2+\cdots+\beta_p x_p \\
&= \beta_0+\beta_1 x_1+\beta_1+\beta_2 x_2+\cdots+\beta_p x_p \\
&= \mu\left(y \mid x_1, x_2, \ldots, x_p\right)+\beta_1
\end{aligned}$$

EDA Tools

- Scatterplot matrix: a p by p matrix of scatterplots for all combos of (quantitative) variables in a data frame

```
pairs(my_data) # base-R
```

- `GGally` package (which is installed on Maize)
  - also includes correlation coefficient and density plots

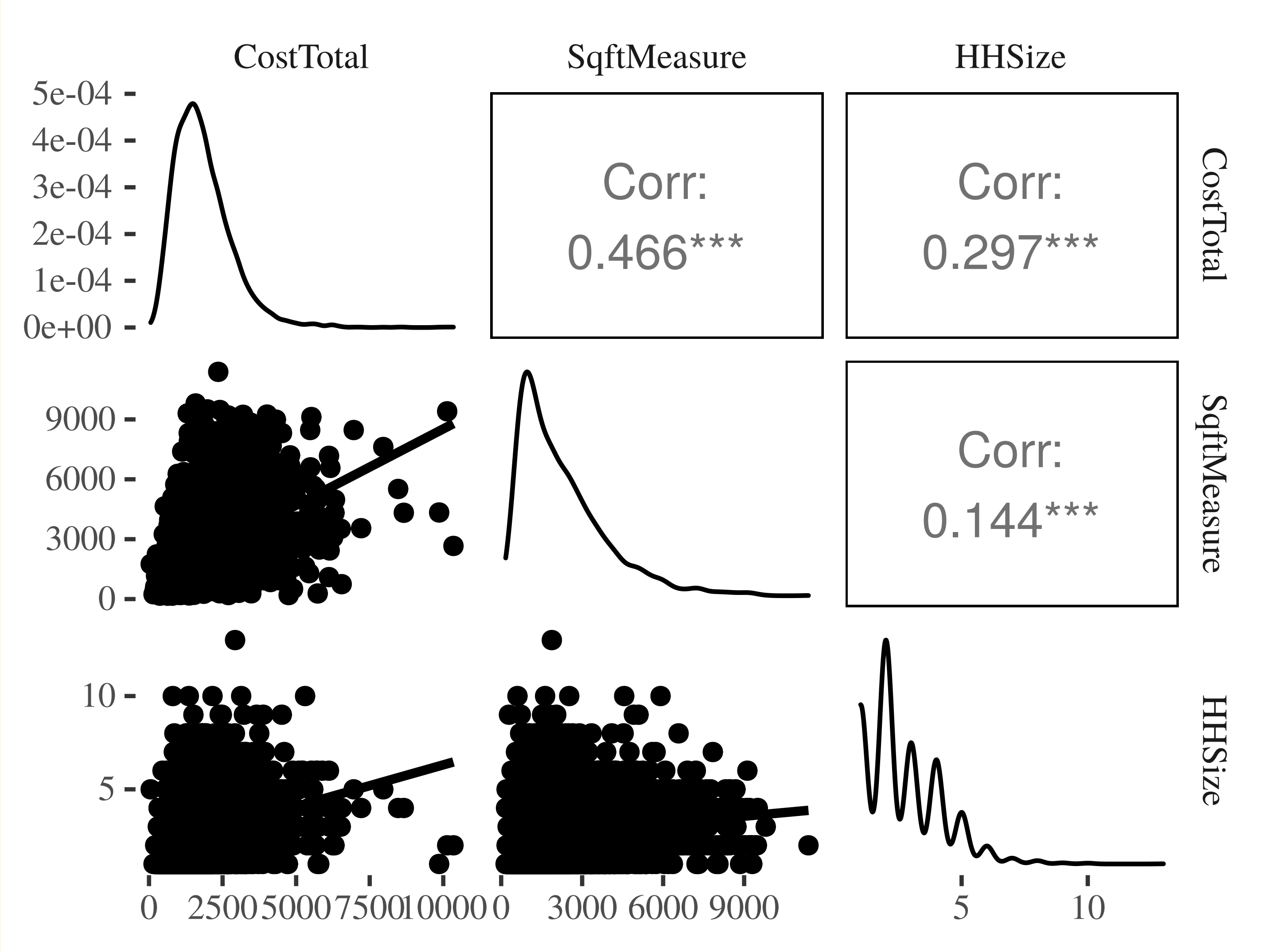
```
ggpairs(my_data) # includes all variables
```

- select variables (and order)

```
ggpairs(my_data,
 columns = c("y", "x1", "x2")) # include only y, x1, x2 variables
```

Example: RECS MLR

```
library(GGally)
ggpairs(energy,
columns = c("CostTotal", "SqftMeasur
lower = list(continuous = wrap("smoo
```



Example: RECS

Regression of log of energy cost against log of square footage and household size

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	4.367	0.071	61.386	0	4.227	4.506
logSqft	0.372	0.010	38.750	0	0.353	0.391
HHSize	0.084	0.005	18.191	0	0.075	0.093

$$\hat{\mu}(\log(\text{ Cost }) \mid x) = 4.3667 + 0.3722 \log(\text{ Sqft }) + 0.0839 \text{ HHSize}$$

In the original scale of the response

$$\begin{aligned} \hat{\text{med}}(\text{ Cost } \mid x) &= e^{4.3667 + 0.3722 \log(\text{ Sqft }) + 0.0839 HH \text{ Size}} \\ &= e^{4.3667} \times (Sqft)^{0.3722} \times e^{0.0839 HH \text{ Size}} \end{aligned}$$

Example: RECS

$$\hat{\mu}(\log(\text{ Cost }) \mid x) = 4.3667 + 0.3722 \log(\text{ Sqft }) + 0.0839 \text{ HHSize}$$

$$\widehat{\text{med}}(\text{ Cost } \mid x) = e^{4.3667} \times (Sqft)^{0.3722} \times e^{0.0839 HH \text{ Size}}$$

- $\hat{\beta}_1 = 0.3722$: An increase in log of square footage of 1 unit is associated with an estimated 0.3722 unit increase in the mean of the log of cost, holding household size constant.
- **Power model** effect $2^{0.3722} = 1.29$: A doubling of square footage is associated with an estimated 29% increase in the median energy cost, holding household size constant.

Example: RECS

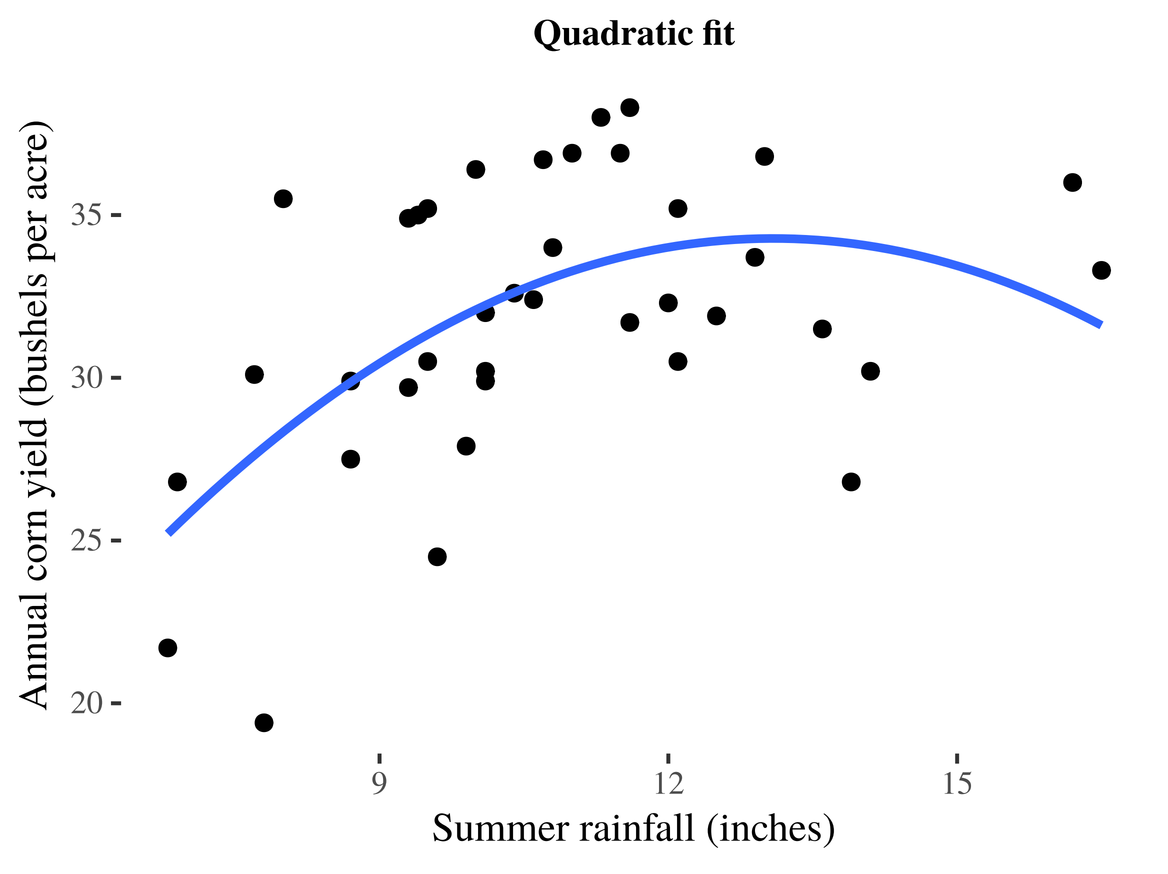
$$\hat{\mu}(\log(\text{ Cost }) \mid x) = 4.3667 + 0.3722 \log(\text{ Sqft }) + 0.0839 \text{ HHSize}$$

$$\widehat{\text{med}}(\text{ Cost } \mid x) = e^{4.3667} \times (Sqft)^{0.3722} \times e^{0.0839 HH \text{ Size}}$$

- $\hat{\beta}_2 = 0.0839$: An increase in household size of 1 person is associated with an estimated 0.0839 unit increase in the mean of the log of cost, holding square footage constant.
- **Exponential model** effect $e^{0.0839} = 1.09$: An increase in household size of 1 person is associated with an estimated 9% increase in the median energy cost, holding square footage constant.

Example: Corn Yield (Textbook Ex. 9.15)

```
ggplot(ex0915, aes(Rainfall, Yield)) +
  geom_point() +
  geom_smooth(method = "lm", formula = y ~ x + I(x^2), se =
  labs(x='Summer rainfall (inches)',
       y='Annual corn yield (bushels per acre)',
       title='Quadratic fit') +
  theme(plot.title = element_text(hjust=0.5, size=9, face='b
```



MLR: Quadratic model

$$\mu_{y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_3 x_2$$

- Quadratic with respect to x_1
- What happens when we change x_1 by one unit (holding x_2 constant)?

$$\begin{aligned}
\mu\left(y \mid x_1 + 1, x_2\right) &= \beta_0 + \beta_1\left(x_1 + 1\right) + \beta_2 (x_1 + 1)^2 + \beta_3 x_2 \\
&= \beta_0 + \beta_1 x_1 + \beta_1 + \beta_2 x_1^2 + \beta_2 2 x_1 + \beta_2 + \beta_3 x_2 \\
&= \mu\left(y \mid x_1, x_2\right) + \beta_1 + \beta_2\left(2x_1 + 1\right)
\end{aligned}$$

MLR: Quadratic model

$$\mu\left(y \mid x_1 + 1, x_2\right) = \mu\left(y \mid x_1, x_2\right) + \beta_1 + \beta_2\left(2x_1 + 1\right)$$

- x_1 effect: a 1 unit increase in x_1 is associated with a $\beta_1 + \beta_2\left(2x_1 + 1\right)$ change in the mean response holding all other predictors fixed.
- Because of the nonlinear association, the change in y depends on the value of x_1.
- For example, if x_1 moves from 1 to 2 units, the mean change is $\beta_1 + 3\beta_2$.

Example: Corn Yield

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	-5.015	11.442	-0.438	0.664	-28.242	18.213
Rainfall	6.004	2.039	2.945	0.006	1.865	10.144
I(Rainfall^2)	-0.229	0.089	-2.588	0.014	-0.409	-0.049

$$\hat{\mu}_{\text{yield} \mid \text{rain}} = -5.015 + 6.004(\text{ rain }) - 0.229(\text{ rain })^2$$

Example: Corn Yield

- An increase from 9 to 10 inches of rainfall is associated with a mean yield increase of 1.646 bushels per acre.

$$6.004 - 0.229(2 \times 9 + 1) = 1.646$$

- An increase from 14 to 15 inches of rainfall is associated with a mean yield decrease of 0.648 bushels per acre.

$$6.004 - 0.229(2 \times 14 + 1) = -0.648$$

Example: Perch interaction

```
library(Stat2Data)
data("Perch")
glimpse(Perch)
ggplot(Perch, aes(x = Width, y = Weight)) +
 geom_point() +
 geom_smooth(method = "lm", se = FALSE) +
facet_wrap(~ ntile(Length, n = 4))
```

```
Rows: 56
Columns: 4
$ Obs    <int> 104, 105, 106, 107, 108, 109, 110, 111, 112, 113, 114, 115, 116…
$ Weight <dbl> 5.9, 32.0, 40.0, 51.5, 70.0, 100.0, 78.0, 80.0, 85.0, 85.0, 110…
$ Length <dbl> 8.8, 14.7, 16.0, 17.2, 18.5, 19.2, 19.4, 20.2, 20.8, 21.0, 22.5…
$ Width  <dbl> 1.4, 2.0, 2.4, 2.6, 2.9, 3.3, 3.1, 3.1, 3.0, 2.8, 3.6, 3.3, 3.7…
```

MLR: Interaction model

$$\mu_{y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2$$

- What happens when we change x_1 by one unit (holding x_2 constant)?

$$\begin{aligned}
\mu\left(y \mid x_1 + 1, x_2\right) &= \beta_0 + \beta_1\left(x_1 + 1\right) + \beta_2 x_2 + \beta_3\left(x_1 + 1\right) x_2 \\
&= \beta_0 + \beta_1 x_1 + \beta_1 + \beta_2 x_2 + \beta_3 x_1 x_2 + \beta_3 x_2 \\
&= \mu\left(y \mid x_1, x_2\right) + \beta_1 + \beta_3 x_2
\end{aligned}$$

- The effect of x_1 on the response, depends on the value of x_2 !

Example: Perch

term	estimate	std_error	statistic	p_value	lower_ci	upper_ci
intercept	113.935	58.784	1.938	0.058	-4.025	231.894
Length	-3.483	3.152	-1.105	0.274	-9.808	2.842
Width	-94.631	22.295	-4.244	0.000	-139.370	-49.892
Length:Width	5.241	0.413	12.687	0.000	4.412	6.070

$$\hat{\mu}_{\text{Weight } | x} = 113.93 - 3.48 \text{ Length } - 94.63 \text{ Width } + 5.24 \text{ Length } \times \text{ Width}$$

Example: Perch

$$\hat{\mu}_{\text{Weight}\ |x} = 113.93 - 3.48\ \text{Length}\ - 94.63\ \text{Width}\ + 5.24\ \text{Length}\ \times\ \text{Width}$$

- How does width affect mean weight?
  - it depends on the length of the fish in a model with a length and width interaction
- Holding length fixed, a 1 unit increase in width is associated with an estimated mean change in weight of

$$\hat{\beta}_{\text{Width}} + \hat{\beta}_{\text{Width:Length}}\ \text{Length}\ = -94.63 + 5.24\ \text{Length}$$

Example: Perch

$$\hat{\mu}_{\text{Weight }|x} = 113.93 - 3.48 \text{ Length } - 94.63 \text{ Width } + 5.24 \text{ Length } \times \text{ Width}$$

- Holding length fixed at 20 cm, a 1 cm increase in width is associated with an estimated mean increase in weight of

$$\hat{\beta}_{\text{Width}} + \hat{\beta}_{\text{Width:Length}}\, 20 = -94.63 + 5.24(20) = 10.194 \text{grams}$$

- Holding length fixed at 40 cm, a 1 cm increase in width is associated with an estimated mean increase in weight of

$$\hat{\beta}_{\text{Width}} + \hat{\beta}_{\text{Width:Length}}\, 20 = -94.63 + 5.24(40) = 115.02 \text{ grams}$$

- The positive interaction parameter estimate means the effect of width on weight is greater for larger values of length
- same is true for the effect of length on weight

Fitting MLR in R

- Planar

```
lm(y~x1 + x2 + x3, data = )
```

- Quadratic

```
lm(y ~ x1 + I(x1^2) + x2, data = )
```

- Interaction

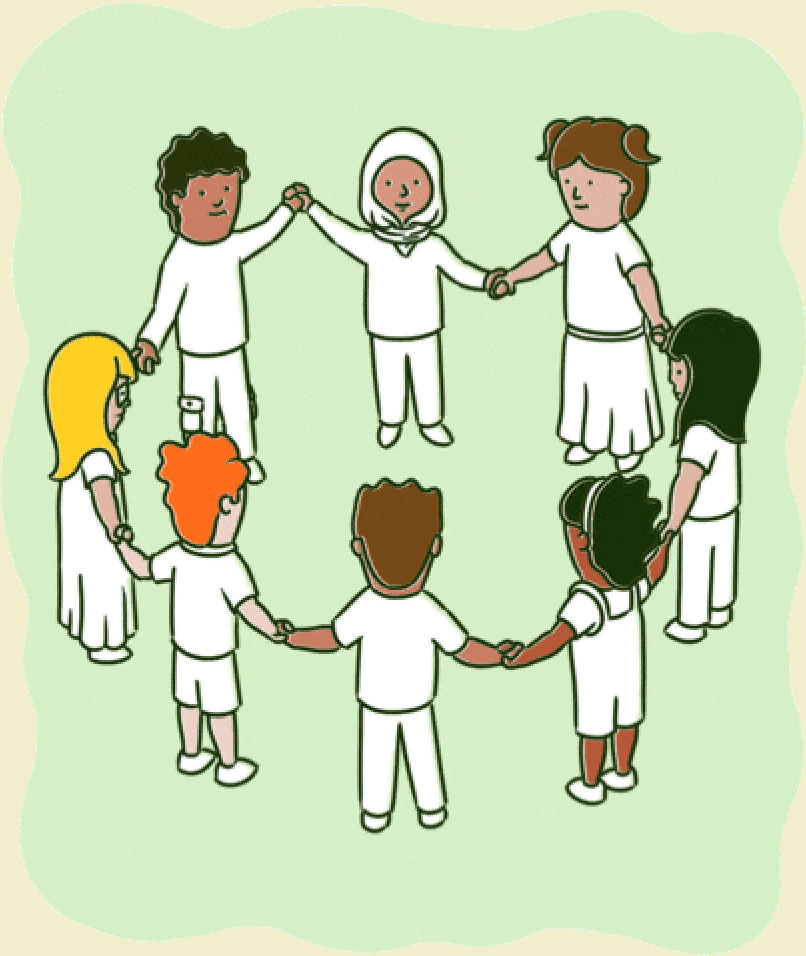
```
lm(y ~ x1 + x2 + x1:x2, data = ) # explicitly add interaction
lm(y ~ x1*x2, data = ) # equals x1 + x2 + x1:x2
```

- Updating an existing model

```
my_lm <- lm(y ~ x1, data = ) # initial model
new_lm <- update(my_lm, . ~ . + x2 + x3) # equals y ~ x1 + x2 + x3
```

Your Turn 1

05:00



- Get the in class activity file from moodle
- We will further practice the concepts seen in the slides