INTRODUCTION TO STATISTICAL MODELING

Multiple explanatory variables

The `statisticalModeling` package

- To evaluate the model, need to set values for explanatory variables **Commonly use mean, median, or mode**
- To visualize the model, need to select several different levels of explanatory variables to include

```
# Load statisticalModeling package
> library(statisticalModeling)
```

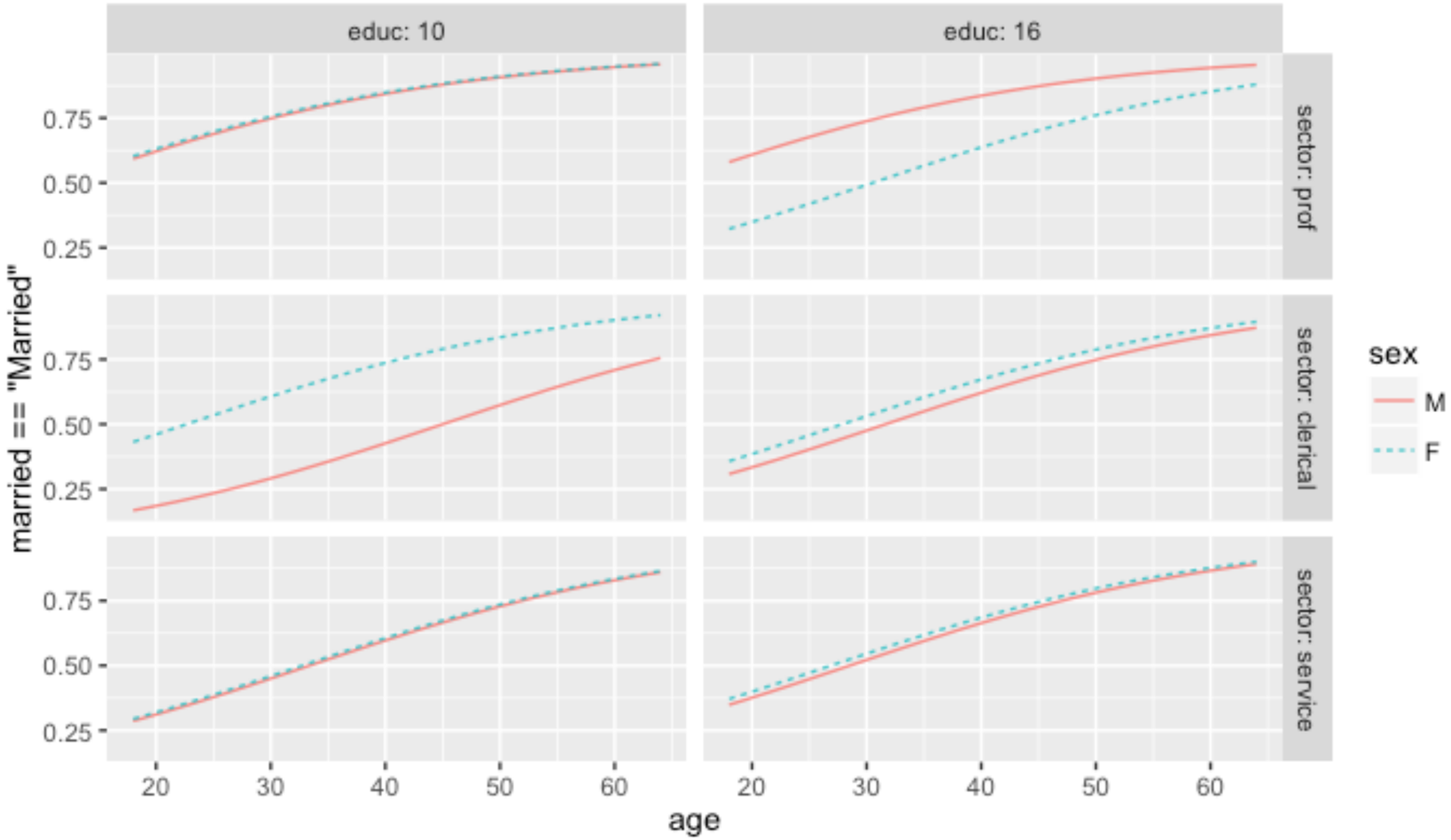
Using `effect_size()`

```
# Train model
> wage_model <- lm(wage ~ educ + sector + sex + exper,
                   data = CPS85)

# Effect size of education on wage: a slope
> effect_size(wage_model, ~ educ)
      slope educ  to:educ sector sex exper
1 0.7179628   12 14.61537   prof   M    15
```

Using fmodel()

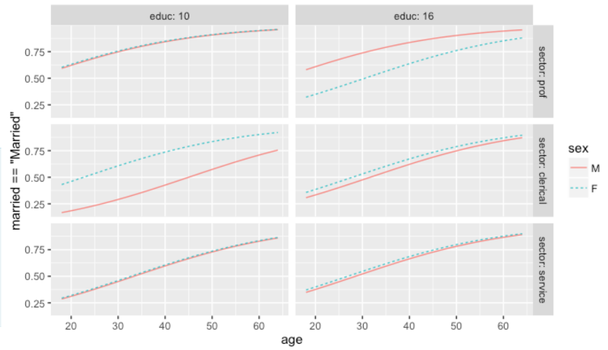
```
# A model of the probability of being married
> married_model <- glm(married == "Married" ~ educ * sector * sex + age,
                       data = CPS85, family = "binomial")

> fmodel(married_model, ~ age + sex + sector + educ, data = CPS85,
         type = "response", educ = c(10, 16))
```

Using `fmodel()`



Designing graphs of models

```
> fmodel(married_model, ~ age + sex + sector + educ, data = CPS85,
         type = "response", educ = c(10, 16))
```



1. Response variable always on y-axis
2. Explanatory variables of primary interest on x-axis
3. Choose one, two, or three variables you want in display
4. If others, choose a fixed value that's of interest

`fmodel()` does (2) - (3) automatically
and (4) either automatically or manually

INTRODUCTION TO STATISTICAL MODELING

Let’s practice!

INTRODUCTION TO STATISTICAL MODELING

Categorical response variables

The question at hand

- For a quantitative response variable and a...
  - Quantitative explanatory variable → **Effect size is a rate**
  - Categorical explanatory variable → **Effect size is a difference**

But what happens when the response variable is categorical?

Model output for categorical response

Two ways to frame the output:

- As categories or *classes*
- As probabilities

Example: marital status

```
# Create model and set inputs
> married_model <- rpart(married ~ educ + sex + age,
                         data = CPS85, cp = 0.005)

# Output as a category (i.e. class)
> evaluate_model(married_model, type = "class", age = c(25, 30),
                 educ = 12, sex = "F")
  educ sex age model_output
1   12   F  25      Married
2   12   F  30      Married

# Output as a probability
> evaluate_model(married_model, type = "prob", age = c(25, 30),
                 educ = 12, sex = "F")
  educ sex age model_output.Married model_output.Single
1   12   F  25            0.6333333           0.3666667
2   12   F  30            0.7425743           0.2574257
```

Extra 5 years of age associated with 11% increase in probability of being married

INTRODUCTION TO STATISTICAL MODELING

Let’s practice!

INTRODUCTION TO STATISTICAL MODELING

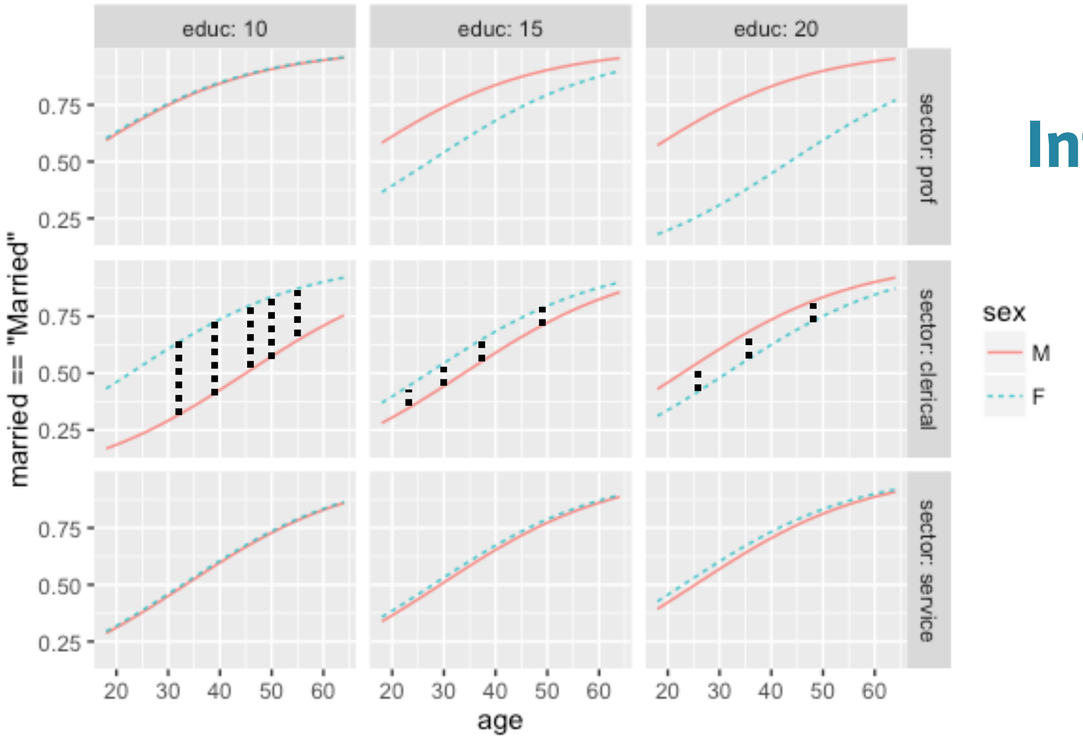
Interactions among explanatory variables

Interaction

Effect size of one variable may change with the other explanatory variables

Probability of being married

```
> married_model <- glm(married == "Married" ~ educ * sector * sex + age,
                       data = CPS85, family = "binomial")
> fmodel(married_model, ~ age + sex + sector + educ, data = CPS85, type = "response")
```

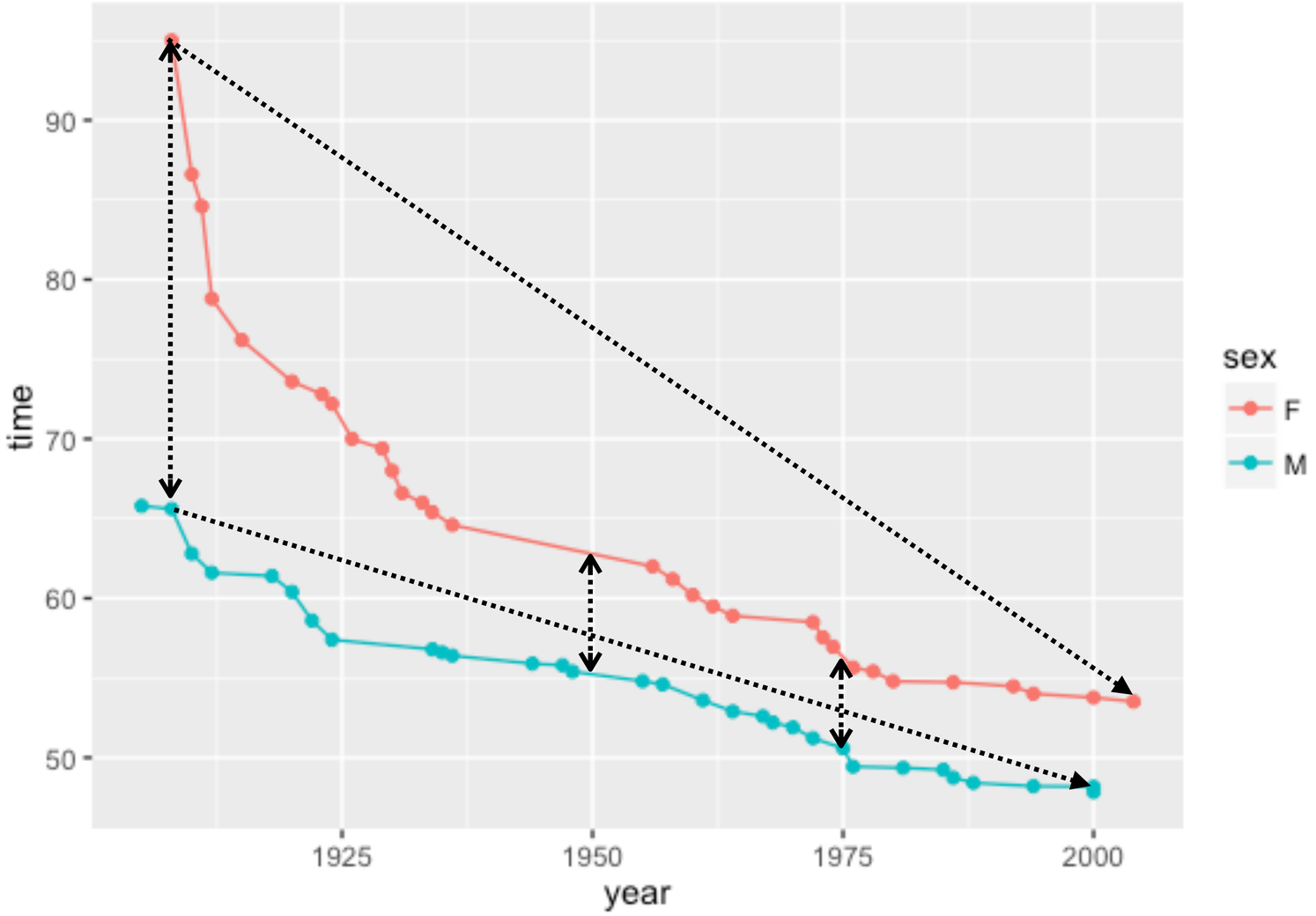


Interaction effect

Interactions and model architecture

- `lm()` includes interactions only if you ask for them
- `rpart()` has interactions built into the method

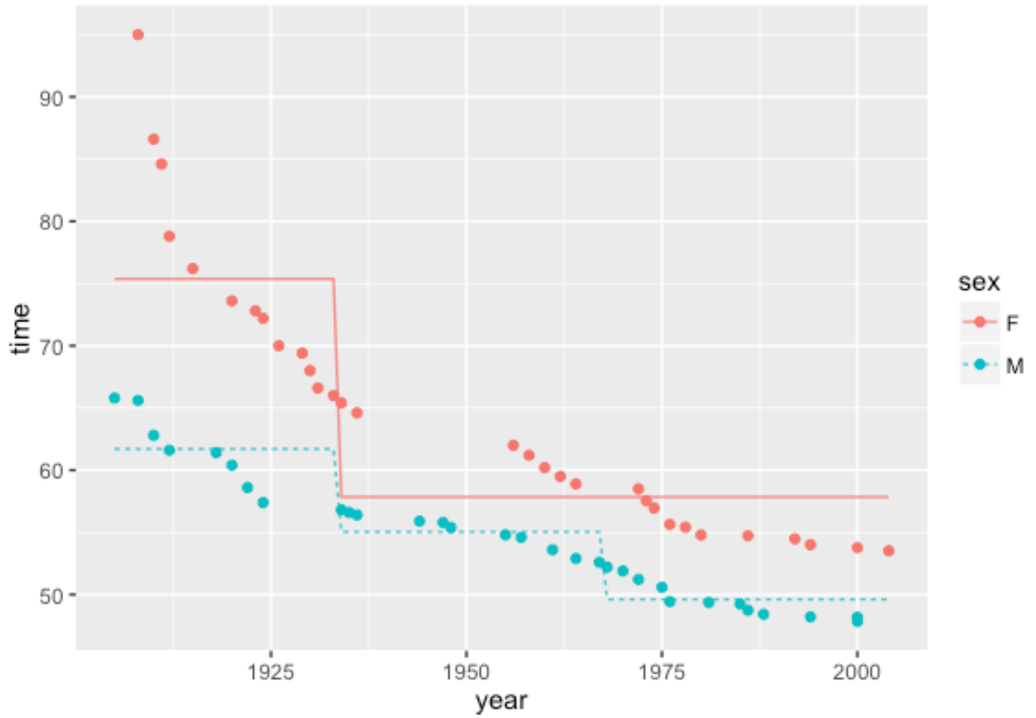
World swimming records

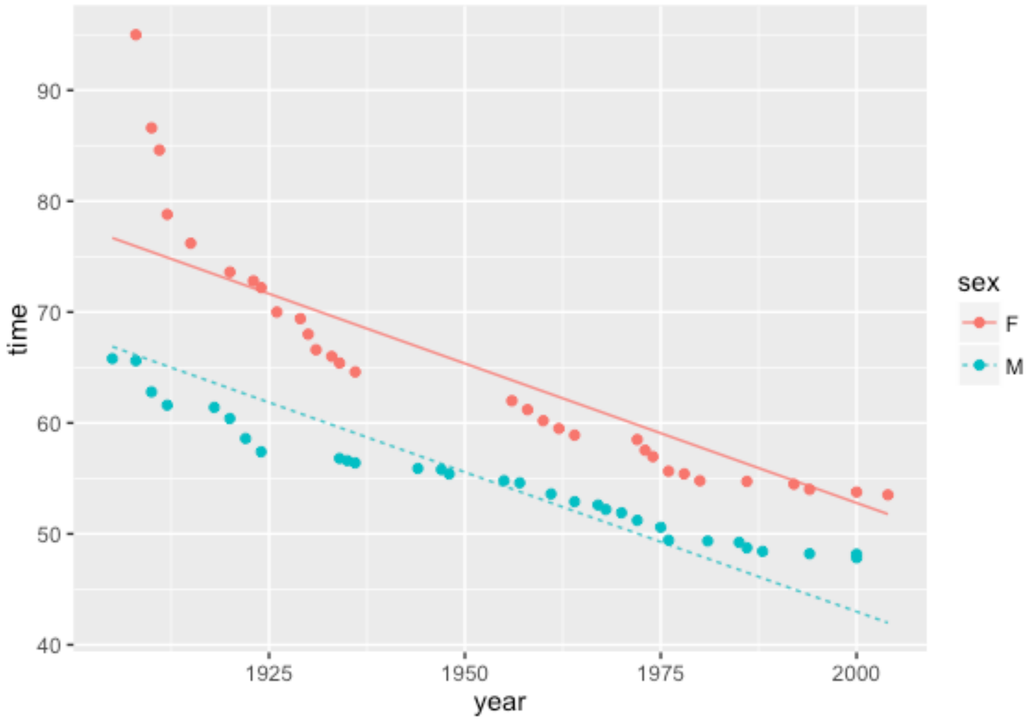


World swimming records

```
mod1 <- rpart(time ~ sex + year, data = SwimRecords)
```

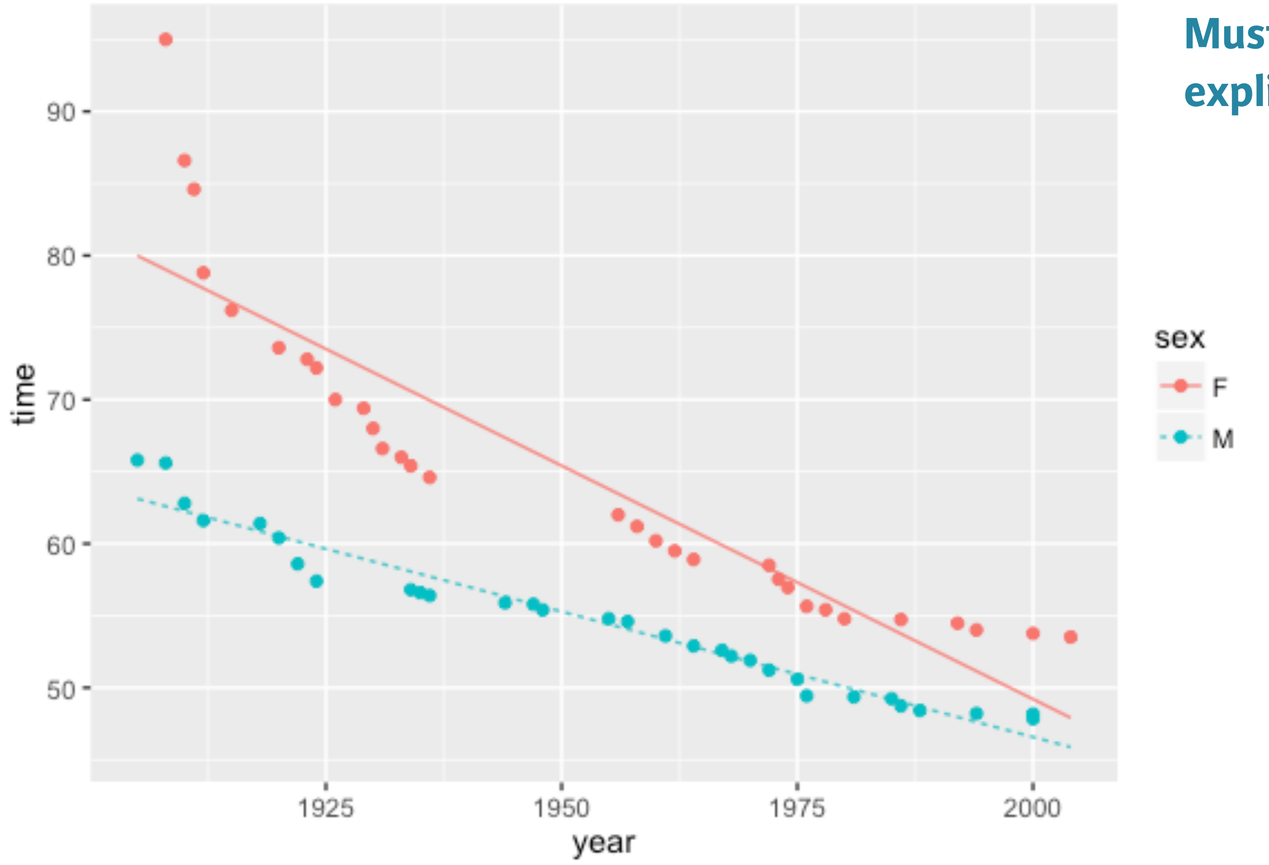
```
mod2 <- lm(time ~ sex + year, data = SwimRecords)
```





Formulas with interactions

```
mod3 <- lm(time ~ sex * year, data = SwimRecords)
```



Must specify interaction explicitly in `lm()`

Does an interaction improve a model?

Use cross validation to see which is better:

- `mod2: ~ year + sex`
- `mod3: ~ year * sex`

```
> t.test(mse ~ model, data = cv_pred_error(mod2, mod3))
data:  mse by model
t = 20, df = 18, p-value = 1.323e-13
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 4.2 5.2
sample estimates:
mean in group mod2 mean in group mod3
                17                 12
```

Error for mod3 < Error for mod2

INTRODUCTION TO STATISTICAL MODELING

Let’s practice!