STAT 213
The Modeling Process, Part II

Colin Reimer Dawson

Oberlin College

June 2, 2021

Outline

Outline

A Process of Statistical Modeling

1. **CHOOSE** — Pick a form (or forms) for the model (or models)
2. **FIT** — Estimate parameters (if any)
3. **ASSESS** — Is the model adequate? Does it strike a good balance between simplicity and fidelity? Are conditions met? (If no go back to 1 and consider more candidates)
4. **USE** — Answer the question of interest (e.g., draw conclusions, make predictions)

Outline

Example: Sleep and Caffeine

- A **sample of 24 adults** are randomly divided equally into **two groups** and given a list of 24 words to memorize.
- During a break, one group takes a **90-minute nap** while another group is given a **100mg caffeine** pill.
- The response variable of interest is the **number of words participants are able to recall** following the break.
- **Is there is a difference in the average number of words** a person can recall depending on their "treatment"?

Sleep vs. Caffeine: Four Step Process

1. **CHOOSE**: Write down candidate models

Model 1: No predictors

$$Y_i = c + \varepsilon_i$$

Model 2: Predictor based on group

$$Y_i = c_{X_i} + \varepsilon_i$$

Sleep vs. Caffeine: Four Step Process

2. **FIT**: Find estimates for parameters that best fit the dataset

Model 1: No predictors

$$Y_i = c + \varepsilon_i$$

Might choose $\hat{c} = \bar{Y}$, the overall sample mean for both groups combined.

Model 2: Predictor based on group

$$Y_i = c_{X_i} + \varepsilon_i$$

Might choose $\hat{c}_{\text{sleep}} = \bar{Y}_{\text{sleep}}$ and $\hat{c}_{\text{caffeine}} = \bar{Y}_{\text{caffeine}}$, the sample means for each group

Sleep vs. Caffeine: Four-Step Process

3. **ASSESS:** How to decide which model suits our purposes better?

Model 1: No predictors

$$Y_i = c + \varepsilon_i$$

Model 2: Predictor based on group

$$Y_i = c_{X_i} + \varepsilon_i$$

Outline

How can we decide between two models?

Model 0: No predictors

$$Y_i = c + \varepsilon_i$$

Model 1: Predictor based on group

$$Y_i = c_{X_i} + \varepsilon_i$$

How would you decide which model is better? (ASSESS step)

Simplicity vs. Fit

- The more complex model is guaranteed to be able to give "predictions" **as close or closer** to the data we have, compared to the simpler one (Why?)

Simplicity vs. Fit

- The more complex model is guaranteed to be able to give "predictions" **as close or closer** to the data we have, compared to the simpler one (Why?)
- Need to balance **fit** to the dataset with **simplicity**.

Simplicity vs. Fit

- The more complex model is guaranteed to be able to give "predictions" **as close or closer** to the data we have, compared to the simpler one (Why?)
- Need to balance **fit** to the dataset with **simplicity**.
- Occam's Razor: **"All else equal", prefer the simpler model**

Simplicity vs. Fit

- The more complex model is guaranteed to be able to give "predictions" **as close or closer** to the data we have, compared to the simpler one (Why?)
- Need to balance **fit** to the dataset with **simplicity**.
- Occam's Razor: **"All else equal", prefer the simpler model**
  - Practically: More complexity → more **sensitivity** to noise (less **robust**) → potentially worse **generalization**

Simplicity vs. Fit

- The more complex model is guaranteed to be able to give "predictions" **as close or closer** to the data we have, compared to the simpler one (Why?)
- Need to balance **fit** to the dataset with **simplicity**.
- Occam's Razor: **"All else equal", prefer the simpler model**
  - Practically: More complexity → more **sensitivity** to noise (less **robust**) → potentially worse **generalization**
- But **what counts as "all else equal"**? Exactly equal only?

Simplicity vs. Fit

- The more complex model is guaranteed to be able to give "predictions" **as close or closer** to the data we have, compared to the simpler one (Why?)
- Need to balance **fit** to the dataset with **simplicity**.
- Occam's Razor: **"All else equal", prefer the simpler model**
  - Practically: More complexity → more **sensitivity** to noise (less **robust**) → potentially worse **generalization**
- But **what counts as "all else equal"**? Exactly equal only?
  - Need to determine whether the improvement in fit is **large enough to overcome the sensitivity** to noise

Hypothesis Testing as Model Selection

Null Hypothesis Testing approach: Adopt the simpler model "by default", and see **whether the data is convincing enough for a skeptic** to embrace the more complex one.

$$H_0 : c_{\text{Sleep}} = c_{\text{Caffeine}}$$
$$H_1 : c_{\text{Sleep}} \neq c_{\text{Caffeine}}$$

Hypothesis Testing as Model Selection

Null Hypothesis Testing approach: Adopt the simpler model "by default", and see **whether the data is convincing enough for a skeptic** to embrace the more complex one.

$$H_0 : c_{\text{Sleep}} = c_{\text{Caffeine}}$$
$$H_1 : c_{\text{Sleep}} \neq c_{\text{Caffeine}}$$

$H_0 \Leftrightarrow$ Model 0
$H_1 \Leftrightarrow$ Model 1

ASSESS/TEST: Select Among Competing Models

Hypothesis testing logic:

- $\bar{Y}_{\text{Sleep}}$ and $\bar{Y}_{\text{Caffeine}}$ **will differ** even when μ_{Sleep} and μ_{Caffeine} do not.

ASSESS/TEST: Select Among Competing Models

Hypothesis testing logic:

- $\bar{Y}_{\text{Sleep}}$ and $\bar{Y}_{\text{Caffeine}}$ **will differ** even when μ_{Sleep} and μ_{Caffeine} do not.
- P-**value**: What is the likelihood **in a world** where $\mu_{\text{Sleep}} = \mu_{\text{Caffeine}}$) that $\bar{Y}s$ would differ as much as they do?

ASSESS/TEST: Select Among Competing Models

Hypothesis testing logic:

- $\bar{Y}_{\text{Sleep}}$ and $\bar{Y}_{\text{Caffeine}}$ **will differ** even when μ_{Sleep} and μ_{Caffeine} do not.
- P-**value**: What is the likelihood **in a world** where $\mu_{\text{Sleep}} = \mu_{\text{Caffeine}}$) that $\bar{Y}s$ would differ as much as they do?
- If P small, **the skeptic is surprised**; can reject H_0 and conclude **we need the more complex model**.

Sleep vs. Caffeine: Model Comparison

```
library(mosaic); library(Lock5Data); data(SleepCaffeine)
t.test(Words ~ Group, data = SleepCaffeine)


     Welch Two Sample t-test

    data:  Words by Group
    t = -2.1438, df = 21.894, p-value = 0.04342
    alternative hypothesis: true difference in means is not equal to 0
    95 percent confidence interval:
     -5.90300367 -0.09699633
    sample estimates:
    mean in group Caffeine    mean in group Sleep
                     12.25                  15.25
```

Conclusion: The difference we see in the sample would be **suprising** if all sample variability is attributable to **random chance** / individual differences; there is **sufficient evidence** that the extra complexity of Model 2 is justified.

USE and Interpretation

- We might be wrong: unlikely things happen sometimes

USE and Interpretation

- We might be wrong: unlikely things happen sometimes
- If we're not and the long run / population means really would differ, **can we say the difference is due to the difference in treatments?**

USE and Interpretation

- We might be wrong: unlikely things happen sometimes
- If we're not and the long run / population means really would differ, **can we say the difference is due to the difference in treatments?**
- Yes! **Because people were randomly assigned** to groups, **if there really is a systematic difference** between the groups, it must in some way be **caused** by the differences in treatments

Outline

Example: Did Public Opinion Change?

The financial firm Lehman Bros. declared bankruptcy in mid-September 2008, during the height of the presidential campaign between then Sen. Barack Obama and Sen. John McCain.

- Was **public opinion** about the election different **before vs. after** the bankruptcy?

Cases/Obs Units	Individual election polls
Response (Y_i)	% Supporting McCain
Predictor (X_i)	Before or after Lehman Bankruptcy?

A Process of Statistical Modeling

1. **CHOOSE** — Pick a form (or forms) for the model (or models)
2. **FIT** — Estimate parameters (if any)
3. **ASSESS** — Is the model adequate? Does it strike a good balance between simplicity and fidelity? Are conditions met? (If no go back to 1 and consider more candidates)
4. **USE** — Answer the question of interest (e.g., draw conclusions, make predictions)

CHOOSE: Define possible models

Population Model 0: Single Mean (No Difference)

$$Y_i = \mu + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2)$$

$$f(X_i) = \mu$$

Population Model 1: Group Means (Change in Opinion)

$$Y_i = \mu_{X_i} + \varepsilon_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \sigma^2_{X_i})$$

$$f(X_i) = \begin{cases} \mu_{\text{before}} & \text{if } X_i = \text{before} \\ \mu_{\text{after}} & \text{if } X_i = \text{after} \end{cases}$$

FIT: Parameter Estimation

Model 0:

- Just one parameter in this model: the constant μ (this value is our prediction, $\hat{Y}_i$ for every i. Could choose
  - Sample mean $\hat{\mu} = \bar{Y}$
  - Sample median $\hat{\mu} = Q_2$

Model 1:

- Two parameters: μ_{before} and μ_{after}. Could choose
  - Sample means by group
  - Sample medians by group

Prediction Error: the Residual

The **model** (population level):

$$Y_i = \mu + \varepsilon_i$$

Prediction Error: the Residual

The **model** (population level):

$$Y_i = \mu + \varepsilon_i$$

The **prediction** (based on sample data):

$$\hat{Y}_i = \hat{\mu} = \bar{Y}$$

Prediction Error: the Residual

The **model** (population level):

$$Y_i = \mu + \varepsilon_i$$

The **prediction** (based on sample data):

$$\hat{Y}_i = \hat{\mu} = \bar{Y}$$

The **prediction error**: Actual Minus Predicted

$$\hat{\varepsilon}_i = Y_i - \hat{Y}_i$$

FIT: Parameter Estimation

Estimating parameters from the sample (FIT step):

Model 0:

$$Y_i = \bar{Y} + \hat{\varepsilon}_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \hat{\sigma}^2)$$

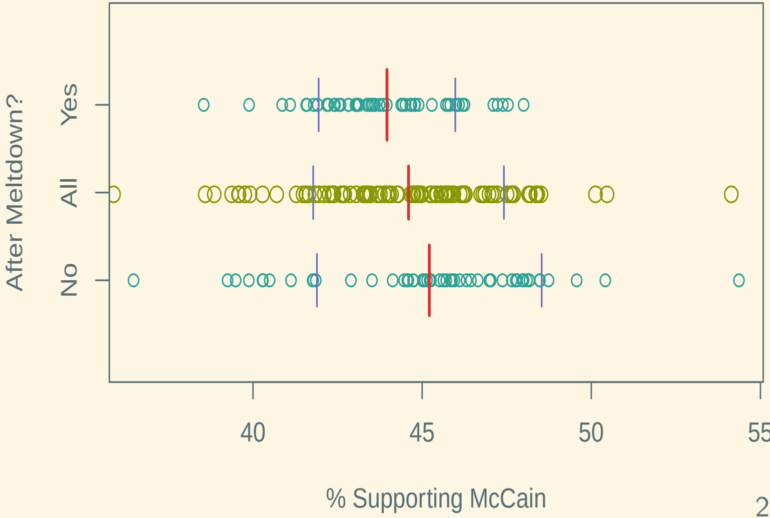
Model 1:

$$Y = \bar{Y}_{X_i} + \hat{\varepsilon}_i, \qquad \varepsilon_i \sim \mathcal{N}(0, \hat{\sigma}^2_{X_i})$$

Symbol Table

Symbol	Definition
X_i	Value of **predictor** for case i
Y_i	Value of **response** for case i
$\hat{Y}_i$	What model would have **predicted** for case i
ε_i	**Residual** for case i (the part the model doesn't tell us)
$\hat{\varepsilon}_i$	**Estimated residual** for case i (difference between Y_i and $\hat{Y}_i$)
μ	True **population/long run mean**
$\hat{\mu}$	**Estimate** (from data) **of true population mean**
$\bar{X}$	**Sample mean** of all X_i
$\bar{Y}$	**Sample mean** of all Y_i (typically used as $\hat{\mu}$, but wouldn't have to be the case)

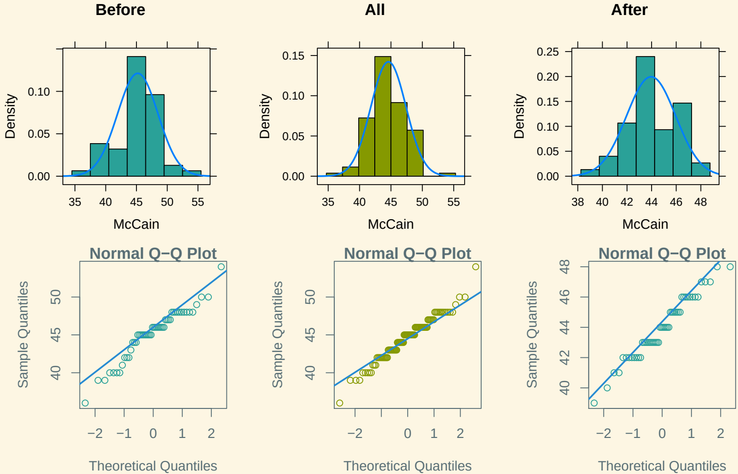
FIT: Parameter Estimation



ASSESS/TEST: Checking Conditions

We assumed Normal residuals. Is that justified?

Plot the residuals!



ASSESS/TEST: Hypothesis Testing as Model Selection

$$H_0 : \mu_{\text{Before}} = \mu_{\text{After}}$$
$$H_1 : \mu_{\text{Before}} \neq \mu_{\text{After}}$$

$$H_0 \Leftrightarrow \text{Model 0}$$
$$H_1 \Leftrightarrow \text{Model 1}$$

Presidential Polling: Hypothesis Test

```
library(Stat2Data); data(Pollster08)
t.test(McCain ~ Meltdown, data = Pollster08)
```

```

	Welch Two Sample t-test

data:  McCain by Meltdown
t = 2.3091, df = 84.729, p-value = 0.02337
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 0.1738376 2.3292393
sample estimates:
mean in group 0 mean in group 1 
       45.21154        43.96000 
```

Conclusion: The difference we see in the sample would be **suprising** if all sample variability is attributable to **random chance** / individual differences; there is **sufficient evidence** that the extra complexity of Model 2 is justified.

USE and Interpretation

- We reject H_0 and favor the more complex model. Now we can make predictions.

USE and Interpretation

- We reject H_0 and favor the more complex model. Now we can make predictions.
- If there really is a difference, **can we conclude that it is due** to the financial meltdown?

USE and Interpretation

- We reject H_0 and favor the more complex model. Now we can make predictions.
- If there really is a difference, **can we conclude that it is due** to the financial meltdown?
- We need to be sensitive to **how the data was collected.**

USE and Interpretation

- We reject H_0 and favor the more complex model. Now we can make predictions.
- If there really is a difference, **can we conclude that it is due** to the financial meltdown?
- We need to be sensitive to **how the data was collected.**
- This is an **observational** dataset, so there are many potential **confounding variables**

A Process of Statistical Modeling

1. **CHOOSE** — Pick a form (or forms) for the model (or models)
2. **FIT** — Estimate parameters (if any)
3. **ASSESS** — Is the model adequate? Does it strike a good balance between simplicity and fidelity? Are conditions met? (If no go back to 1 and consider more candidates)
4. **USE** — Answer the question of interest (e.g., draw conclusions, make predictions).
    - But be careful about conclusions not warranted by the data-collection process!