# STAT 213
# Statistical Modeling

Colin Reimer Dawson

Oberlin College

February 7, 2018

# Outline

# Course Structure/Logistics

- R expectations / opportunities for practice
- Feedback/keeping track of progress
- "Good faith effort"
- Nature of the two projects

# Models are...

- simplifications
- approximations
- not perfectly correct
- useful for a particular purpose

# Data: Numbers With a "Story"

$$\text{DATA} = \text{PATTERN} + \text{IDIOSYNCRACIES}$$

How do we decide what "the pattern" is? This, in a nutshell, is the project of modeling

# Purposes of Statistical Models

1. Making predictions
2. Understanding relationships
3. Assessing differences

## The Project, More Formally

Find a relationship between a **response variable** $(Y)$ and one or more **predictor/explanatory variables**, $X_1, \ldots, X_k$.

$$Y = f(X) + \varepsilon$$

$$\text{DATA} = \text{PATTERN} + \text{IDIOSYNCRACIES}$$

# Examples

- $Y =$ Home Price
  $X =$ Home size
- $Y =$ Exam score
  $X =$ Hours spent studying
- $Y =$ State % in poverty
  $X =$ State % with no health insurance
- $Y =$ SAT score
  $X =$ Family income

# Your Questions

I am not sure what "pattern" means. Since $Y = f(X) + \varepsilon$, the "pattern", I'm guessing, is not a slope, but something else.

## The 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? Could it be simpler? Are conditions met?
4. **Use** — Answer the question of interest (e.g., make predictions)


## 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 caffeine pill. The response variable of interest is the number of words participants are able to recall following the break. We are testing to see if there is a difference in the average number of words a person can recall depending on whether the person slept or ingested caffeine.

# Prediction and Testing: Sleep vs. Caffeine

How can we predict how many words someone will remember?

Data: Results of a recall experiment (Person $i$ has group and number of words recalled: $(X_i, Y_i)$)

Model 1: No predictors (CHOOSE step)

$$Y_i = c + \varepsilon_i$$

Words = "Typical" Number + Individual/Situational Influence

$$f(X) = c$$

Each individual $i$ is different, but not based on whether they slept or took caffeine.

# FIT/ASSESS/USE

- Later, we will discuss how to estimate $c$ (**FIT**ting the model to data), and how to **ASSESS** the model
- What about **USE**ing the model?
- Predict based on the inputs (in this case, none): estimate an individual's outcome using the "typical" number, $c$

# Model 2: Now With A Predictor!

CHOOSE step:

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

$$c_{X_i} = c_{\text{sleep}} \text{ if } X_i = \text{sleep}$$
$$c_{X_i} = c_{\text{caffeine}} \text{ if } X_i = \text{caffeine}$$

$$f(X) = \begin{cases} c_{\text{sleep}} & \text{if } X = \text{sleep} \\ c_{\text{caffeine}} & \text{if } X = \text{caffeine} \end{cases}$$

# How can we decide between two models?

Model 1: No predictors

$$Y_i = c + \varepsilon_i$$

Model 2: Predictor based on group

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

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

# Simplicity vs. Fit

- The more complex model is guaranteed fit the data better (or at least no worse). (Why?)
- Need to balance fit by simplicity.
- "All else equal", prefer the simpler model.
- But what counts as "equal"? Exactly equal only?

# Hypothesis Testing as Model Selection

Can adopt the simpler model by default, and see if there's enough evidence to reject.

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

$$H_0 \Leftrightarrow \text{Model 1}$$
$$H_1 \Leftrightarrow \text{Model 2}$$

# USE and Interpretation

- Suppose we reject $H_0$ and favor the more complex model. Now we can make predictions. What can we conclude?
- In using the model to draw conclusions, we need to be sensitive to how the data was collected. (Really, should keep this in mind at every step)

## 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? |

# CHOOSE: Define possible models

## Population Model 1: Single Mean (No Difference)

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

$$f(X_i) = \mu$$

## Population Model 2: 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 1:

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

Model 2:

- Two parameters: $\mu_{\text{before}}$ and $\mu_{\text{after}}$. Could choose
  - Sample means by group
  - Sample medians by group

# 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

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

# FIT: Parameter Estimation

Estimating parameters from the sample (FIT step):

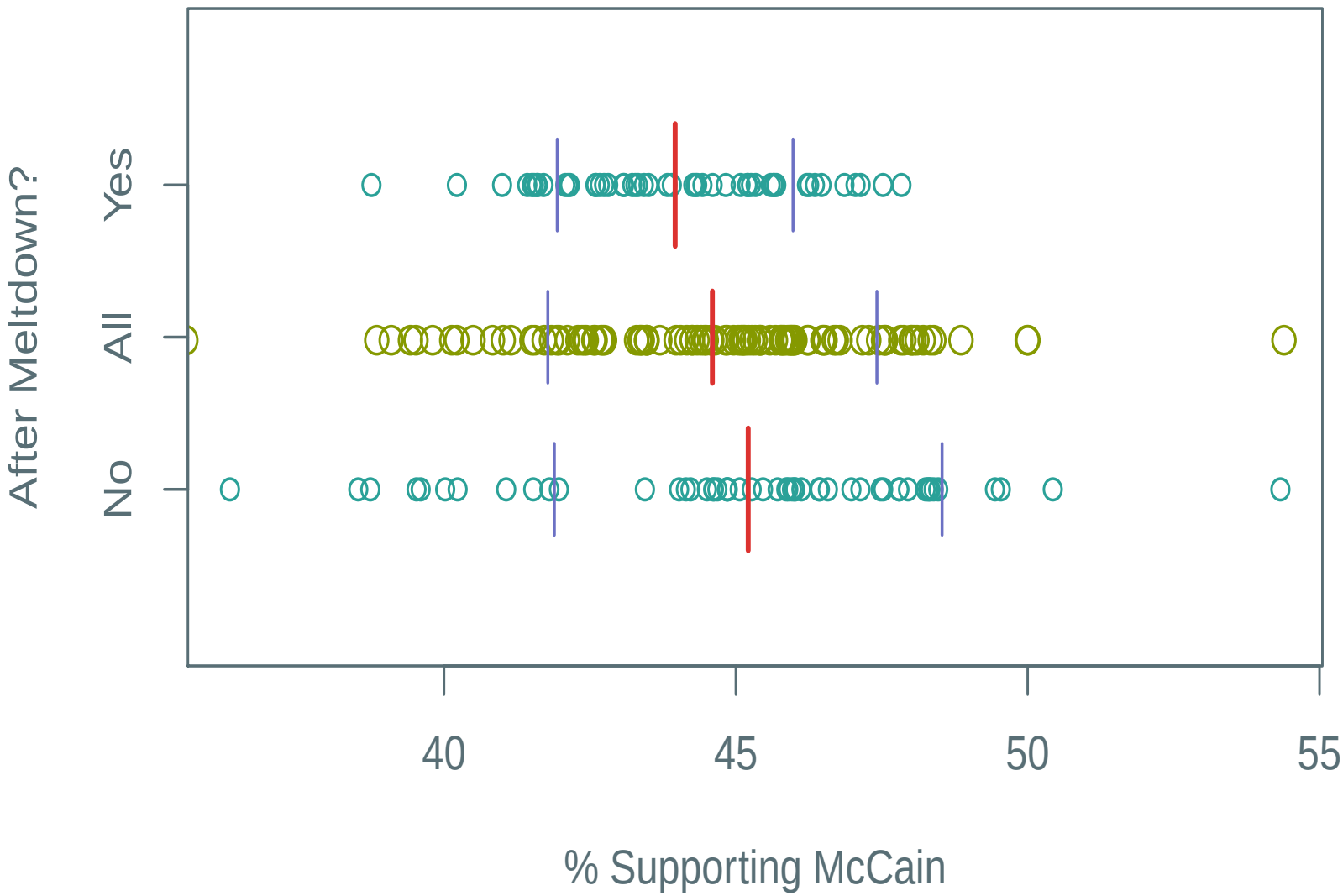
Model 1:

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

Model 2:

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

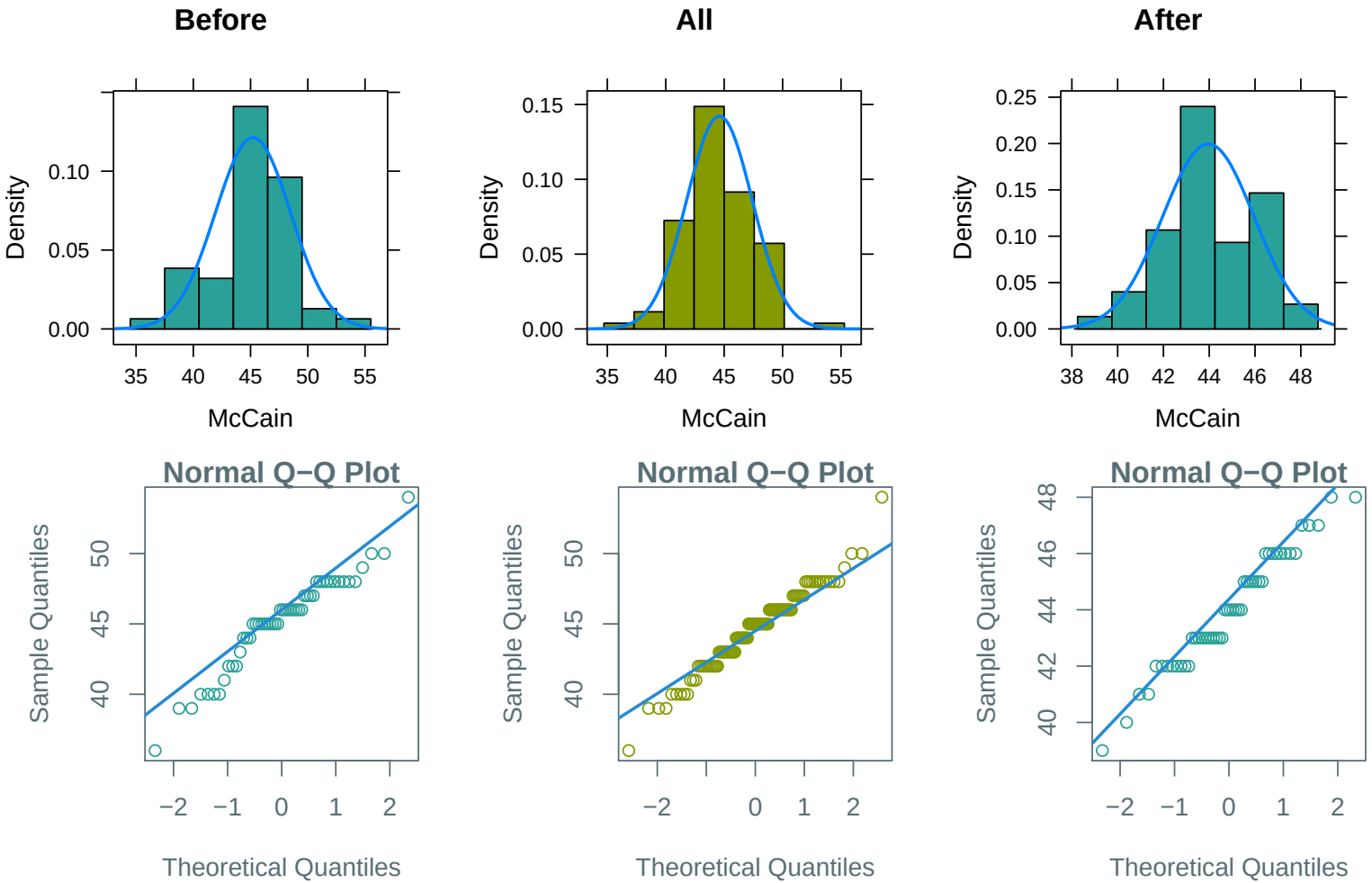
# FIT: Parameter Estimation



# ASSESS/TEST: Checking Conditions

We assumed Normal residuals. Is that justified?

Plot the residuals!



# Simplicity vs. Fit

- The more complex model is guaranteed fit the data better (or at least no worse). (Why?)
- Need to balance fit by simplicity.
- "All else equal", prefer the simpler model.
- But what counts as "equal"? Exactly equal only?

# Hypothesis Testing as Model Selection

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

$$H_0 \Leftrightarrow \text{Model 1}$$
$$H_1 \Leftrightarrow \text{Model 2}$$

# ASSESS/TEST: Select Among Competing Models

- Need to balance fit by simplicity.
- Can adopt the simpler model by default, and see if there's enough evidence to reject.
  1. Randomization test
  2. Two-sample $t$-test

# ASSESS/TEST: Hypothesis Testing as Model Selection

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

$$H_0 \Leftrightarrow \text{Model 1}$$
$$H_1 \Leftrightarrow \text{Model 2}$$

# ASSESS/TEST: Select Among Competing Models

Hypothesis testing logic:

- $\bar{Y}_{\text{before}}$ and $\bar{Y}_{\text{after}}$ differ even when $\mu_{\text{before}}$ and $\mu_{\text{after}}$ do not.
- $P$-value: What is the likelihood (over possible samples from a pop. with no diff. in means) that $\bar{Y}s$ would differ as much as in our data, *if* $\mu$s are the same?
- If $P$ small, can reject $H_0$; conclude that we need the more complex model.

# USE and Interpretation

- Suppose we reject $H_0$ and favor the more complex model. Now we can make predictions. What can we conclude?
- In using the model to draw conclusions, we need to be sensitive to how the data was collected. (Really, should keep this in mind at every step)