STAT 213
Statistical Modeling

Colin Reimer Dawson

Oberlin College

4 February 2016

Outline

For Tuesday...

- Download R and RStudio (see URLS on the syllabus), **or** (after tomorrow) log on to `http://rstudio.oberlin.edu` and verify that you have an account.
- Write up to turn in: Ex. 0.12, 0.13, 0.19 (on Blackboard by 6pm Tuesday)
- Read: Ch. 1.1-1.3
- Be prepared to answer Ex. 1.1-1.3, 1.6-1.7 in class.

Outline

Models are...

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

All models are wrong but some models are useful.

Now it would be very remarkable if any system existing in the real world could be exactly represented by any simple model. However, cunningly chosen parsimonious models often do provide remarkably useful approximations. For example, the law $PV = RT$ relating pressure P, volume V and temperature T of an "ideal" gas via a constant R is not exactly true for any real gas, but it frequently provides a useful approximation and furthermore its structure is informative since it springs from a physical view of the behavior of gas molecules.

For such a model there is no need to ask the question "Is the model true?". If "truth" is to be the "whole truth" the answer must be "No". The only question of interest is "Is the model illuminating and useful?". — George Box, 1978

Data: Numbers With a "Story"

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

Data: Numbers With a "Story"

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

How do we decide what "the pattern" is?

Purposes of Statistical Models

1. Making predictions
2. Understanding relationships

2*. Assessing differences

The Project, More Formally

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

The Project, More Formally

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

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

The Project, More Formally

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

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

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

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)

Prediction and Testing: School Spirit

How can we predict school spirit level in college students? Do athletes differ from non-athletes?

Data: Survey responses on a 1 to 7 scale.

Prediction and Testing: School Spirit

How can we predict school spirit level in college students? Do athletes differ from non-athletes?

Data: Survey responses on a 1 to 7 scale.

Model 1: No predictors (CHOOSE step)

$$Y = c + \varepsilon$$

Individuals differ, but not based on whether they're athletes.

Estimating Parameters

Just one parameter in this model: the constant c (this value is our prediction, $\hat{Y}$, since we have no other variables). What value should we choose?

- Sample mean? $\hat{Y} = \bar{Y}$
- Sample median? $\hat{Y} = Q_2$ (or m)

Estimating Parameters

Just one parameter in this model: the constant c (this value is our prediction, $\hat{Y}$, since we have no other variables). What value should we choose?

- Sample mean? $\hat{Y} = \bar{Y}$
- Sample median? $\hat{Y} = Q_2$ (or m)

Questions to Discuss in Groups

1. How would you decide between mean and median (and possibly other things too)?
2. How would you justify your decision?
3. How would you measure how good your prediction is?

Prediction Error: the Residual

The model (population level):

$$Y = c + \varepsilon$$

Prediction Error: the Residual

The model (population level):

$$Y = c + \varepsilon$$

The prediction (based on sample data):

$$\hat{Y} = \hat{c}$$

Prediction Error: the Residual

The model (population level):

$$Y = c + \varepsilon$$

The prediction (based on sample data):

$$\hat{Y} = \hat{c}$$

The prediction error: Actual Minus Predicted

$$Y - \hat{Y}$$

What to Optimize?

Pick an overall measure of error, and make it as small as possible on the sample (FIT step):

1. Sum of residuals?

$$\sum_{i=1}^{n}(Y - \hat{Y})$$

2. Sum of absolute residuals?

$$\sum_{i=1}^{n}\left|Y - \hat{Y}\right|$$

3. Sum of squared residuals?

$$\sum_{i=1}^{n}(Y - \hat{Y})^2$$

Choice Leads To...

1. Sum of residuals?

$$\sum_{i=1}^{n}(Y - \hat{Y})$$

Not really useful, since signs cancel out.

Choice Leads To...

1. Sum of residuals?

$$\sum_{i=1}^{n}(Y - \hat{Y})$$

   Not really useful, since signs cancel out.

2. Sum of absolute residuals?

$$\sum_{i=1}^{n}\left|Y - \hat{Y}\right|$$

   Minimized by $\hat{Y} = Q_2$ (or m)

Choice Leads To...

1. Sum of residuals?

$$\sum_{i=1}^{n}(Y - \hat{Y})$$

   Not really useful, since signs cancel out.

2. Sum of absolute residuals?

$$\sum_{i=1}^{n}\left|Y - \hat{Y}\right|$$

   Minimized by $\hat{Y} = Q_2$ (or m)

3. Sum of squared residuals?

$$\sum_{i=1}^{n}(Y - \hat{Y})^2$$

   Minimized by $\hat{Y} = \bar{Y}$

What about those residuals?

Often times (though not always!), residuals are Normally distributed. We can refine Model 1 to say

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

Model 1: Parameter Estimation (FIT step

Estimating parameters from the sample:

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

Model 1: Parameter Estimation (FIT step

Estimating parameters from the sample:

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

$$f(X) = \bar{Y}$$

Model 2: Now With A Predictor!

Population model (CHOOSE step):

$$Y = \mu_i + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_i^2) \qquad i = 1, 2$$

Model 2: Now With A Predictor!

Population model (CHOOSE step):

$$Y = \mu_i + \varepsilon, \qquad \varepsilon \sim \mathcal{N}(0, \sigma_i^2) \qquad i = 1, 2$$

$$f(X) = \begin{cases} \mu_1 & \text{if } X = \text{athlete} \\ \mu_2 & \text{if } X = \text{non-athlete} \end{cases}$$

Model 2: Parameter Estimation (FIT step

Estimating parameters from the sample:

$$Y = \bar{Y}_i + \hat{\varepsilon}, \qquad \epsilon \sim \mathcal{N}(0, \hat{\sigma}_i^2) \qquad i = 1, 2$$

Model 2: Parameter Estimation (FIT step

Estimating parameters from the sample:

$$Y = \bar{Y}_i + \hat{\varepsilon}, \qquad \epsilon \sim \mathcal{N}(0, \hat{\sigma}_i^2) \qquad i = 1, 2$$

$$f(X) = \begin{cases} \bar{Y}_1 & \text{if } X = \text{athlete} \\ \bar{Y}_2 & \text{if } X = \text{non-athlete} \end{cases}$$

Checking Conditions (ASSESS step)

We assumed Normal residuals. Is that justified?

Checking Conditions (ASSESS step)

We assumed Normal residuals. Is that justified?

Plot the residuals! (More on this later)

How can we decide between two models?

Groups: 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?)

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.

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.

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.

1. Randomization test
2. Two-sample t-test

Hypothesis Testing as Model Selection

$$H_0 : \mu_{\text{Athletes}} = \mu_{\text{Non-athletes}}$$
$$H_1 : \mu_{\text{Athletes}} \neq \mu_{\text{Non-athletes}}$$

Hypothesis Testing as Model Selection

$$H_0 : \mu_{\text{Athletes}} = \mu_{\text{Non-athletes}}$$
$$H_1 : \mu_{\text{Athletes}} \neq \mu_{\text{Non-athletes}}$$

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

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)