Understanding regression models

ST552 Lecture 10

Charlotte Wickham

2019-01-30

Today

- Lecture: mathematical strategies for understanding models
- Lab: understanding models through visualization

Inference so far

We've talked about the machinery to perform:

- t-tests, t-based confidence intervals for individual βs and linear combinations of βs
- F-tests for hypotheses about many βs

But what can we do with this machinery?

Two stages in understanding regression models

1. Understand a model in the context of a problem
2. Define a set of models to answer questions of interest

Focus on #1 to gain intuition in how to approach #2.

Example

LA Dodgers (baseball team) sometimes give out “bobbleheads” at home games. They are curious if this increases attendance at games.

81 games in 2012 season, 11 of which bobbleheads were given out.

Have measurements on:

- attendance at game (number of people)
- day of the week the game was played
- some other variables, that are probably important, but we will ignore for now

(*Inspired by Chapter 2 in Modeling Techniques in Predictive Analytics: Business Problems and Solutions with R. Get data from http://www.informit.com/promotions/modeling-techniques-in-predictive-analytics-141183 if you interested.*)

A model

$$\begin{aligned}\text{attendance}_i = \beta_0 + \beta_1 1\{\text{bobblehead YES}\}_i + \\ \beta_2 1\{\text{Tue}\}_i + \beta_3 1\{\text{Wed}\}_i + \beta_4 1\{\text{Thu}\}_i + \\ \beta_5 1\{\text{Fri}\}_i + \beta_6 1\{\text{Sat}\}_i + \beta_7 1\{\text{Sun}\}_i + \epsilon_i\end{aligned}$$

What does this model say about the relationship between attendance and whether bobbleheads are given out and day of the week?

Aside some terminology

variable: a measurement made on the observational units.

E.g. bobblehead (yes/no) and day of week (mon/tue/wed/thu/fri/sat/sun).

term: a column of the design matrix

E.g. bobblehead, $1\{\text{Fri}\}$

One useful strategy

Ask about the **effect** of a *variable*?

What does the model say about the mean response when a variable is varied, holding all other variables constant.

Categorical variable: Find the mean response for each level and compare.

Continuous variable: Find the change in mean response if the variable increases by 1 unit.

The effect of bobbleheads

The variable has two levels: yes, no

We'll find $\mathrm{E}\left(\text{attendence}\middle|\,\text{bobblehead} = \text{Yes}\,\right)$ and $\mathrm{E}\left(\text{attendence}|\text{bobblehead} = \text{No}\right)$ then compare them.

$$\mathrm{E}\left(\text{attendence}\middle|\,\text{bobblehead} = \text{Yes}\,\right)$$

$$\mathrm{E}\left(\text{attendence}\middle|\,\text{bobblehead} = \text{No}\,\right)$$

The effect of bobbleheads

$$\mathrm{E}\left(\text{attendence}\middle|\ \text{bobblehead} = \text{Yes}\right) - \mathrm{E}\left(\text{attendence}\middle|\ \text{bobblehead} = \text{No}\right)$$

- For a fixed day of the week, the model predicts the mean attendance increases by β_1 when bobbleheads are given out.
- The model predicts the mean attendance when bobbleheads are given is β_1, more than when bobbleheads aren't given, after accounting for day of the week.

Relationship to inference

If β_1 is zero, then bobbleheads don't have an effect on attendance.

We could answer the questions:

- Is the mean attendance higher when bobbleheads are given out? t-test on $\beta_1 = 0$.
- How much higher is the mean attendance higher when bobbleheads are given out? Confidence interval on β_1

But this is an observational study, so we need to be careful with our language!
OK "It is estimated distributing bobbleheads is **associated** with an increased mean attendance of XX".
Not OK "It is estimated distributing bobbleheads increases the mean attendance by XX".

Your turn: Day of the week

What does model say about the expected attendance on Monday?

What does model say about the expected attendance on Tuesday?

What does model say about the expected attendance on Wednesday?

Your turn: Day of the week

Day of the week: inference

If $\beta_2 = \beta_3 = \beta_4 = \beta_5 = \beta_6 = \beta_7$ is zero, then day of the week doesn't have an effect on attendance.

What tools could we use to answer the questions:

- Does day of the week have an effect on mean attendance (after accounting for the bobblehead promotion)?
- How much does the mean attendance differ between Friday and Saturday?

An example with a continuous variable

In lab today:

$$\text{weight}_i = \beta_0 + \beta_1 1\{\text{male}\}_i + \beta_2 \text{height}_i + \epsilon_i$$

What is the effect of height?

What is the effect of height?

Compare $\mathrm{E}\left(\text{weight}\middle|\, \text{height} = h\,\right)$ to $\mathrm{E}\left(\text{weight}\middle|\, \text{height} = h + 1\,\right)$, holding other variables constant.

Interactions

Interactions: describe situations where the effect of one variable depends on the level of another explanatory variable.

E.g.

$$\text{weight}_i = \beta_0 + \beta_1 1\{\text{male}\}_i + \beta_2 \text{height}_i + \beta_3 \left(1\{\text{male}\} \times \text{height}\right)_i + \epsilon_i$$

The same strategy will work.