Advanced Regression Methods for Independent Data

STAT/BIOST 570, 2020

Introduction

Mauricio Sadinle

Department of Biostatistics

UNIVERSITY *of* WASHINGTON

Introduction

This course is about *regression methods*: *modeling* how a *response* variable depends on *covariates*.

Examples

Example: home prices

There's a huge variety in the housing market: studios, condos, townhomes, houses of all sizes and styles.

Examples

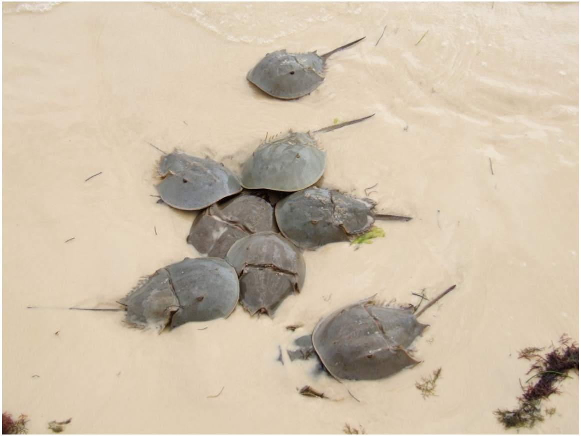
Many stakeholders (sellers, buyers, government agencies, Zillow) are interested in understanding what affects the selling price of a home

- Response variable: selling price
- Covariates: type of home, location, size, property tax, number of bedrooms, number of bathrooms, whether the house is new, ...

The cheapest house in San Francisco: one-bedroom, one-bath, 570 square-foot, listed for $675k.

Examples

Example: female horseshoe crabs and satellite males



Monandrous (front pair) and polyandrous females nesting at Seahorses Key. The polyandrous female has one attached male and four satellite males (and another male approaches the group). Taken from https://people.clas.ufl.edu/hjb/research/

Examples

A biologist might be interested in the following question: what factors explain the fact that some female horseshoe crabs attract more males than others?

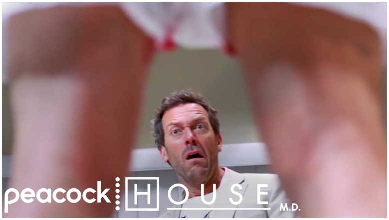
- Response variable: number of satellite males of a female crab
- Covariates: female crab's color, spine condition, weight, and carapace width, ...

Examples

Example: demand for medical care

A health services researcher might be interested in understanding what factors explain the variability in the number of doctor visits.

- Response variable: number of doctor visits
- Covariates: self-perceived health, health insurance, income, age, marital status, sex, ...



Another covariate: is Dr House your primary care provider?

Introduction

- As you can see, regression problems are ubiquitous.
- Depending on the problem, responses and covariates may be categorical (nominal or ordinal), counts, continuous, censored, or mixed (e.g. alcohol consumption).
- However, note that here we are not concerned about the distribution of the covariates: that is the focus of *multivariate modeling* – in regression we have a distinct response variable of interest.
- Depending on the context, the *response* variable also receives alternative names: outcome, dependent, output, endogenous.
- *Covariates* can also receive alternative names, including: explanatory, regressors, predictors, independent, input, exogenous.

The Big Picture

- i: index of study unit
- Y_i: response variable for unit i
- $\boldsymbol{x}_i = (1, x_{i1}, \ldots, x_{ik})$: row vector of covariates for unit i
- We think of Y_i as a random variable
- The distribution of Y_i depends on the value of $\boldsymbol{x}_i$
- Our data will consist of n independent pairs $\{(Y_i, \boldsymbol{x}_i)\}_{i=1}^{n}$

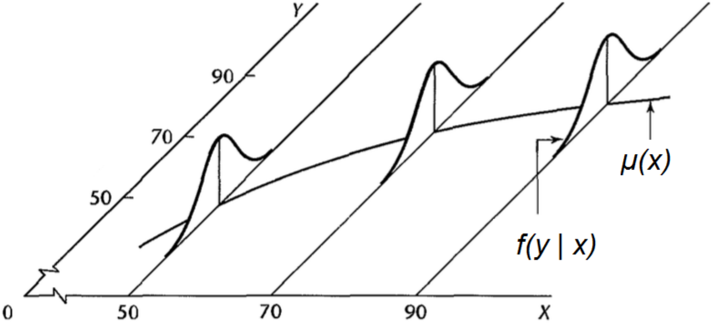
The Big Picture

- Denote the *true* conditional distribution of Y_i given $\boldsymbol{x}_i$ as $F_{\boldsymbol{x}_i}$, and we write
$$Y_i \mid \boldsymbol{x}_i \sim F_{\boldsymbol{x}_i}$$

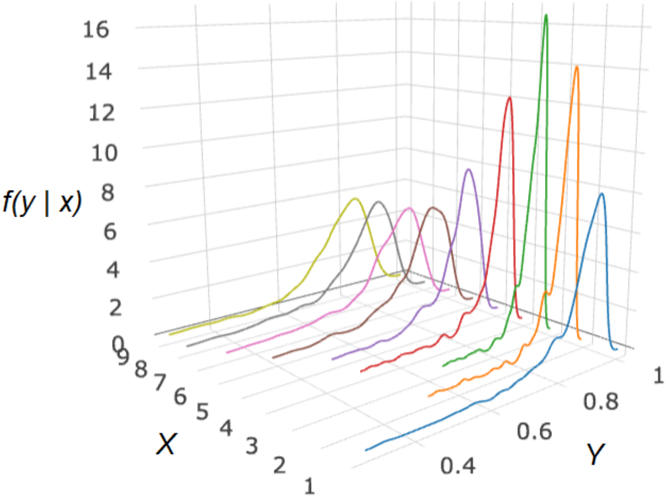
- Denote the corresponding conditional density of Y_i given $\boldsymbol{x}_i$ as $f(\cdot \mid \boldsymbol{x}_i)$

- The *regression function* $\mu(\cdot)$ is defined as the conditional mean of Y_i given $\boldsymbol{x}_i$
$$\mu(\boldsymbol{x}_i) = \mathsf{E}(Y_i \mid \boldsymbol{x}_i) = \int y \ f(y \mid \boldsymbol{x}_i) dy$$

The Big Picture



The Big Picture



The Big Picture: Models

With the above set-up we can now start talking about *modeling*. But what is a *model*?[1]

- *Model*: a class of probability distributions.

  Examples: the class of normal distributions, the class of symmetric distributions, the class of unimodal distributions, ...

- *Fully parametric model*: a class of probability distributions indexed[2] by a finite-dimensional parameter vector.

  Examples: the class of normal distributions indexed by means and variances, the Poisson distributions indexed by the mean, ...

[1] See van der Vaart (1998), p. 358

[2] By *indexed* we mean there is a one-to-one correspondence between the distributions and the parameter values.

The Big Picture: Fully Parametric Regression Models

- In a *fully parametric regression model* we specify a fully parametric model for the conditional distributions of Y_i given $\boldsymbol{x}_i$, across all possible values of $\boldsymbol{x}_i$
  - The class of *all* conditional distributions across values of the covariates is indexed by a finite-dimensional parameter vector

Examples:

- $Y_i \mid \boldsymbol{x}_i \sim Normal[\mu(\boldsymbol{x}_i), \sigma^2], \quad \mu(\boldsymbol{x}_i) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} = \boldsymbol{x}_i \boldsymbol{\beta}$
- $Y_i \mid \boldsymbol{x}_i \sim Poisson[\mu(\boldsymbol{x}_i)], \quad \mu(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i \boldsymbol{\beta})$
- $Y_i \mid \boldsymbol{x}_i \sim Normal[\mu(\boldsymbol{x}_i), \sigma^2(\boldsymbol{x}_i)], \quad \mu(\boldsymbol{x}_i) = \boldsymbol{x}_i \boldsymbol{\beta}, \ \ \sigma^2(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i \boldsymbol{\eta})$
  (this is called a *log-linear variance model*.)
- $Y_i \mid \boldsymbol{x}_i \sim Beta[\alpha(\boldsymbol{x}_i), \beta(\boldsymbol{x}_i)], \quad \alpha(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i \boldsymbol{\lambda}), \ \ \beta(\boldsymbol{x}_i) = \exp(\boldsymbol{x}_i \boldsymbol{\eta})$
  (this approach is workable but not common.)

The Big Picture: Semiparametric Models

- *Semiparametric model*[3]: a class of probability distributions constrained in a way that is partially characterized by a finite-dimensional parameter vector

  Example: the class of distributions where the variance equals a specific function of the mean, e.g. mean = variance

[3] See van der Vaart (1998), p. 358

The Big Picture: Semiparametric Models

- In a *semiparametric regression model* the class of conditional distributions of Y_i given $\boldsymbol{x}_i$, across all possible values of $\boldsymbol{x}_i$, is only partially characterized by a finite-dimensional parameter vector

  Examples:

  - $Y_i \mid \boldsymbol{x}_i \sim H[\mu(\boldsymbol{x}_i)], \quad \mu(\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}$, where H has mean $\mu(\boldsymbol{x}_i)$ but it is otherwise unspecified

  - $Y_i \mid \boldsymbol{x}_i \sim H[\mu(\boldsymbol{x}_i), \sigma^2(\boldsymbol{x}_i)], \quad \mu(\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta}, \quad \sigma^2(\boldsymbol{x}_i) = h[\mu(\boldsymbol{x}_i)]$, for some positive function $h(\cdot)$, where H has mean $\mu(\boldsymbol{x}_i)$ and variance $\sigma^2(\boldsymbol{x}_i)$ but it is otherwise unspecified

The Big Picture: Mean + Error

Remark: If $Z \sim H(\mu)$, for a distribution H parameterized in terms of its mean μ, we can always write

$$Z = \mu + \epsilon,$$

where

$$\epsilon = Z - \mu \sim H(\mu) - \mu$$

- Say $Z \sim \mathit{Normal}(\mu, \sigma^2)$, then $Z = \mu + \epsilon$, with $\epsilon \sim \mathit{Normal}(0, \sigma^2)$

- Indeed, this decomposition is clean if H is a member of the *location-scale family of distributions*

  Reminder: if the distribution of Z belongs to the location-scale family, then so does $a + bZ$ for any constants a and b

The Big Picture: Mean + Error

But this decomposition is not always neat. Think about the distribution of $\epsilon = Z - \mu$ in the following examples:

- $Z \sim \mathit{Bernoulli}(\pi)$
  $\Rightarrow \epsilon = Z - \pi$, with $P(\epsilon = -\pi) = 1 - \pi$ and $P(\epsilon = 1 - \pi) = \pi$
- $Z \sim \mathit{Poisson}(\lambda)$
  $\Rightarrow \epsilon = Z - \lambda$, ... ?
- $Z \sim \mathit{Multinomial}[N, (\pi_1, \pi_2, \pi_3)]$
  $\Rightarrow \epsilon = Z - (\pi_1, \pi_2, \pi_3)N$, ... ?
- $Z \sim \mathit{Beta}[\alpha, \beta]$
  $\Rightarrow \epsilon = Z - \alpha/(\alpha + \beta)$, ... ?

As you can see, the decomposition $Z = \mu + \epsilon$ leads to awkward distributions of ϵ for many distributions of Z

The Big Picture: Mean + Error

- Many regression models are represented using this decomposition, as

  $$Y_i = \mu(\boldsymbol{x}_i) + \epsilon_i,$$

  in order to separate

  - $\mu(\boldsymbol{x}_i) = \mathrm{E}(Y_i \mid \boldsymbol{x}_i)$: the regression function or *deterministic part*
  - ϵ_i: the *error term* or *stochastic part*.

- This formulation is equivalent to

  $$Y_i \mid \boldsymbol{x}_i \sim H[\ \mu(\boldsymbol{x}_i)\ ],$$

  for some distribution H with mean $\mu(\boldsymbol{x}_i)$

- While the mean + error formulation is natural when $Y_i \in \mathbb{R}$, the second one is cleaner when Y_i is discrete or constrained

- We shall use one representation or another depending on what is more convenient

This Course: Fully Parametric Models

We will start by studying certain fully parametric regression models:

- *The normal linear model*:

$$Y_i \mid \boldsymbol{x}_i \sim \mathit{Normal}[\,\mu(\boldsymbol{x}_i), \sigma^2], \quad \mu(\boldsymbol{x}_i) = \boldsymbol{x}_i\boldsymbol{\beta},$$

  or equivalently,

$$Y_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \quad \epsilon_i \sim \mathit{Normal}[0, \sigma^2]$$

This Course: Fully Parametric Models

- *Generalized linear models*:

  $$Y_i \mid \boldsymbol{x}_i \sim H[\ \mu(\boldsymbol{x}_i)\ ],$$

  for H a member of the *exponential family* of distributions with mean $\mu(\boldsymbol{x}_i)$ such that

  $$g[\mu(\boldsymbol{x}_i)] = \boldsymbol{x}_i\boldsymbol{\beta}$$

  - For example, the Poisson distribution is a member of the exponential family, and so

    $$Y_i \mid \boldsymbol{x}_i \sim \textit{Poisson}[\mu(\boldsymbol{x}_i)], \quad \log[\mu(\boldsymbol{x}_i)] = \boldsymbol{x}_i\boldsymbol{\beta},$$

    is an example of a generalized linear model (GLM)

This Course: Fully Parametric Models

- Under a fully parametric model, we will be able to write down a likelihood function for the model parameters
- Given the likelihood function, inference on the model parameters can be conducted in different ways, and we will look into:
  - Maximum likelihood
  - Bayesian inference

This Course: Regression with Weaker Assumptions

Then, we will study the extent to which we can relax assumptions of the aforementioned models and parameter estimators

- *Least squares estimator in a linear model*: if our model simply says that

$$Y_i = \boldsymbol{x}_i \boldsymbol{\beta} + \epsilon_i, \quad \text{var}(\epsilon_i) = \sigma^2,$$

  what can we say about the *least squares* estimator of $\boldsymbol{\beta}$?

This Course: Regression with Weaker Assumptions

- *Quasi-likelihood*: say our model simply says that

$$Y_i \mid \boldsymbol{x}_i \sim H[\mu(\boldsymbol{x}_i), \sigma^2(\boldsymbol{x}_i)]$$

  - H has mean $\mu(\boldsymbol{x}_i)$ such that

$$g[\mu(\boldsymbol{x}_i)] = \boldsymbol{x}_i \boldsymbol{\beta}$$

  - H has variance $\sigma^2(\boldsymbol{x}_i)$

$$\sigma^2(\boldsymbol{x}_i) = h[\mu(\boldsymbol{x}_i)]$$

    for some positive function $h(\cdot)$

  - H is otherwise unspecified

  - How can we draw inferences on $\boldsymbol{\beta}$?

This Course: Regression with Weaker Assumptions

- *Estimating equations, M-estimators, and sandwich estimators*: say our model simply says that

$$Y_i \mid \boldsymbol{x}_i \sim H[\ \mu(\boldsymbol{x}_i; \boldsymbol{\beta})\ ]$$

  - H has mean $\mu(\boldsymbol{x}_i; \boldsymbol{\beta})$ but it is otherwise unspecified
  - $\mu(\cdot; \boldsymbol{\beta})$ is a function fully characterized by a finite vector of parameters $\boldsymbol{\beta}$, e.g. $\mu(\boldsymbol{x}_i; \boldsymbol{\beta}) = h(\boldsymbol{x}_i \boldsymbol{\beta})$
  - How can we draw inferences on $\boldsymbol{\beta}$?

This Course: Regression with Weaker Assumptions

Note:

- The regression models with weaker assumptions that we will be covering are formally *semiparametric*
- However, there are many other examples of semiparametric regression, e.g.
  - Flexible modeling of the regression function:

    $$\mu(\boldsymbol{x}_i) = \boldsymbol{x}_{i1}\boldsymbol{\beta} + \eta(\boldsymbol{x}_2), \quad \text{with } \eta \text{ in some functional space}$$

  - For more examples see Tsiatis (2006)
- Here, we work with $\mu(\boldsymbol{x}) = \mu(\boldsymbol{x}; \boldsymbol{\beta})$ being determined by a finite dimensional parameter vector $\boldsymbol{\beta}$.

This Course: Model Misspecification

What if our models are totally or partially wrong?

- *Least squares*: what are we estimating asymptotically when we incorrectly assume a linear model? Do we get anything meaningful?
- *Maximum-likelihood estimation*: what is the MLE recovering asymptotically when our model is wrong?
  - What if we wrongly assumed a parametric model for the distribution of $Y_i \mid \boldsymbol{x}_i$ but correctly specified $\mu(\boldsymbol{x}_i) = \mathrm{E}(Y_i \mid \boldsymbol{x}_i)$?
- *Estimating equations, M-estimators*: what do estimators based on estimating equations give us if our models are wrong?

The Big Picture: Study Population

This course focuses on *inference*: the task of going from sample data to *population* generalizations

- But what do we mean by *population*?

The Big Picture: Study Population

Let us consider an example:

- A biologist wants to determine whether a certain species can be classified as *sexually dimorphic* (the two sexes exhibit different characteristics beyond sexual organs)
- Males and females are indistinguishable in terms of color and markings, so the biologist wants to determine whether there are differences in their weight
- This is an endangered species, so the biologist is able to collect the weight of all of the remaining specimens
- There are 10 males and 10 females left, with average weights of 200 and 190 ounces
- The biologist argues that, since we have the information of the entire population, and the average weights are different between the sexes, we can conclude that the species is sexually dimorphic
- Do you agree with this assessment?

The Big Picture: Study Population

What are we interested in?

- Say we are interested in a finite population of size N, from where we sample $n \leq N$: in this case, if $N = n$ then no inference is required!
- If we are interested in finite population characteristics based on a sample of size $n < N$, then *survey sampling* techniques are relevant.

The Big Picture: Study Population

- However, it often makes sense to think of a *super-population*: an infinite population from which the population of size N is assumed to be drawn

$$\begin{array}{ccccc} \text{Super-population} & \rightarrow & \text{Finite Study Population} & \rightarrow & \text{Sample} \\ \infty & \rightarrow & N & \rightarrow & n \end{array}$$

- *In this course our models are for "the" super-population*
- The super-population is a (conditional) probability distribution whose characteristics (parameters) we care about
- We assume simple random sampling in the two sampling stages to go super-population → finite study population → sample

The Big Picture: Inference and Model Formulation

This course focuses on *inference*: making statements about parameters of the model for the super-population

- The methods covered here will require us to at least formulate the form of the regression function

$$\mathrm{E}(Y_i \mid \boldsymbol{x}_i) = \mu(\boldsymbol{x}_i; \boldsymbol{\beta})$$

- Given a model, the methods that we will cover will guarantee certain inferential properties of the estimators of $\boldsymbol{\beta}$

- But how are we supposed to come up with $\mu(\boldsymbol{x}_i; \boldsymbol{\beta})$, let alone a model for the conditional distribution of $Y_i \mid \boldsymbol{x}_i$?

The Big Picture: Inference and Model Formulation

To formulate a reasonable model, learn as much as possible about the problem

- The background science (make sure your model makes sense in the context)
- The data collection procedure (maybe you need to account for the sample design)
- Your target population
- The aims of the analysis

The Big Picture: Inference and Model Formulation

What is your goal?

- *Description, exploration*: summarize trends in a large dataset, model formulation, hypothesis generation.
- *Confirmation, testing* of a hypothesis. Here tweaking the model based on your data will get you in trouble. Your model should be formulated before you use your data.
- *Prediction*. Not necessarily interested in *why* or *how*.
- *Causality*. If you seek to potentially intervene units to affect outcomes, subtleties arise.

Some of these require more/less assumptions, and inference is not really that relevant for description, exploration and prediction.