Explanation vs Prediction

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
nanayawce@gmail.com

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2020
Columbia University

a framework to explore explanation vs prediction

Shmueli (2010)

- **theoretically...**
  - let $\mathcal{X}$ cause $\mathcal{Y}$ through the function $\mathcal{F}$

$$\mathcal{Y} = \mathcal{F}(\mathcal{X})$$

- **empirically...**
  - $\mathbf{X}$ and Y operationalize $\mathcal{X}$ and $\mathcal{Y}$
  - f is the model that operationalizes $\mathcal{F}$
- **explanatory modeling** seeks an f close to $\mathcal{F}$

$$E(Y) = f(\mathbf{X})$$

- **predictive modeling** seeks an $\hat{f}$ that best predicts Y_{new}

$$E(Y_{new}) = \hat{f}(\mathbf{X}_{new})$$

a different — but related — perspective

Expected Prediction Error (Hastie et al. 2009)

$$EPE = Var(Y) + Bias^2 + Var(\hat{f}(x)) \tag{1}$$

where:

$$EPE = \text{Expected Prediction Error}$$
$$Var(Y) = E\{Y - f(x)\}^2 : \text{random error}$$
$$Bias^2 = \{E(\hat{f}(x)) - f(x)\}^2 : \text{model mispecification}$$
$$Var(\hat{f}(x)) = E\{\hat{f}(x) - E(\hat{f}(x))\}^2 : \text{sample estimation}$$

- **explanatory modeling**

$$min\{Bias^2\}$$

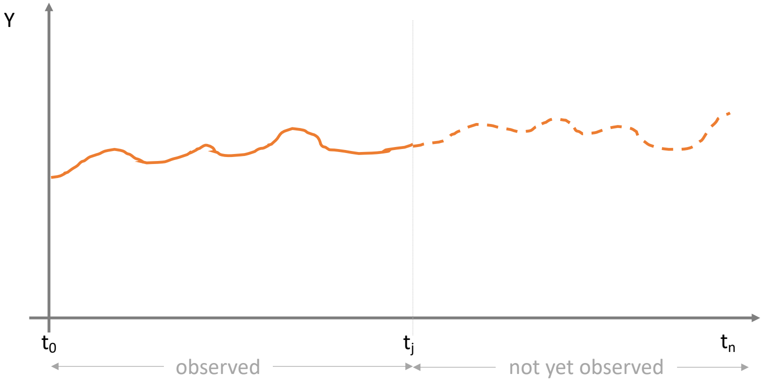
- **predictive modeling**

$$min\{Bias^2 + Var(\hat{f}(x))\}$$

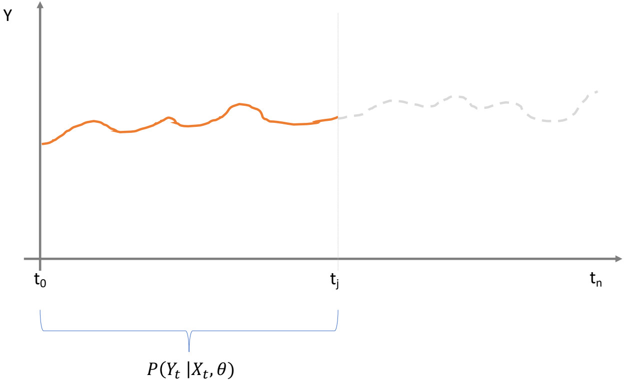
in more detail: how explanation $\neq$ prediction

Explanatory Modeling	**Predictive Modeling**
f resembles $\mathcal{F}$	$\hat{f}$ links $\mathbf{X}, Y$
theory-selected $\mathbf{X}$	**association**-selected $\mathbf{X}$
may use **alternate** $\mathbf{X}$ and Y	requires **exact** $\mathbf{X}$ and Y
backward-looking	**forward**-looking
model fit validation	**predictive error** validation
$\min(Bias^2)$ in (1)	$\min\{Bias^2 + Var(\hat{f}(x))\}$ in (1)
$\lvert E[\hat{\beta}] - \beta\rvert \to 0$	$\min(\lvert Y_{new} - \tilde{Y}_{new}\rvert)$

Example: think of a simple time-series

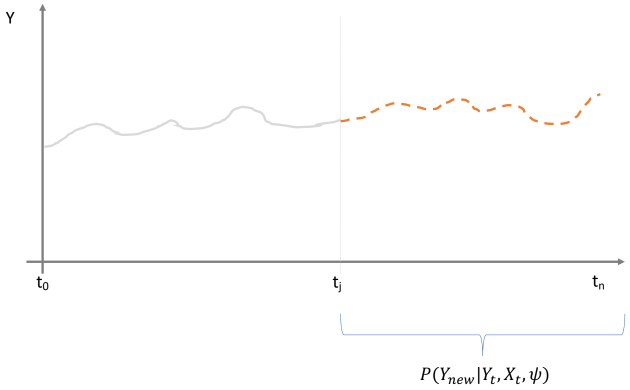


Explanation: problem to solve is observed Y



- **explanatory models:** characterize Y exactly as observed — and not otherwise
- must estimate θ w/o bias to make valid **inferences**

Prediction: problem to solve is Y_{new}



- **predictive models:** project Y_{new} based on X_{new}
- most likely, $\psi \neq \theta$ where ψ exists but may be useless for inference

to put it in perspective...

- any model will contain a combination of degrees of:
  - **explanatory power**
  - **predictive accuracy**
- two different dimensions or one with tradeoffs?
- a "good" model is **sophisticatedly simple** (Zellner 2001)

Explanation

what do we mean by explanation?

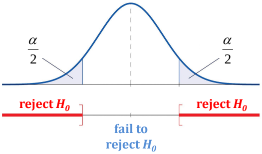
- **explanation:** provide a rationale for **why something happened the way it happened** and not in a different way
  - necessary for meaningful **inference**
  - more formally, find a **model** that describes

$$E(Y) = f(X, \theta)$$

- **problems to solve:**
  - approximate the "true" **data-generating mechanism**
    - find the f that is sufficiently close to $\mathcal{F}$
  - recover the "true" **parameter values** that govern the observed data
    - find the appropriate θ in $P(Y|X, \theta)$

focus for inference: successful recovery of θ

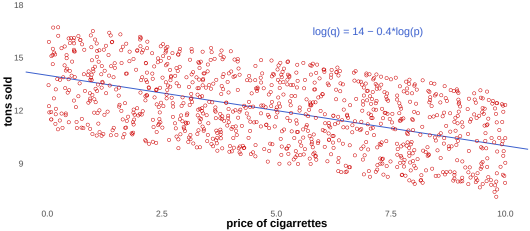
- inference requires having the **"correct" estimated parameters** $\hat{\theta}$, so we spend time
  i) carefully **describing null hypotheses** for estimated parameters: $H_0 : \theta = 0$
  ii) **characterizing distributions under the null**: $\theta \overset{H_0}{\sim} t_v(0, 1)$
  iii) evaluating **how unlikely** is the estimated value under the null: $Pr(\hat{\theta} > k)$



  iv) using theory to ensure that estimated parameters are **unbiased**
  v) using statistics to ensure that **appropriate variances** were estimated

an example: price elasticity of demand

- based on economic theory



- we hypothesize a relation between price (p) and quantity sold (q) of a good — **price elasticity**:

$$\eta_D = \frac{\%\Delta q^D}{\%\Delta p}$$

- we fit a model and estimate $\hat{\beta}$ to recover η_D

$$\log(q) = \alpha + \beta * \log(p) + \epsilon$$

- from the estimated parameter $\hat{\beta}$, we infer that a 1% increase in p decreases q by 0.4%

do not forget: experiments and causal inference

- recovering parameters for inference is hard — too many things could go wrong
  - **observational data** is messy; must rely on theory as guidance
- **causal inference** is even harder: must **compare potential outcomes** under treatment and without treatment
  - **experiments** are golden standard for **causal inference**: treatment assignment is **unconfounded**
  - **observational data** for causal inference requires balance in pre-treatment covariates to compare appropriate groups
    - observational data does not always have “untreated” observations

Prediction

what do we mean by prediction?

- ethimologically:
  - ***predict***: `prae-` before + `dicere` to say
  - ***forecast***: `fore-` before + `casten` to prepare
  - ***prognosticate***: `pro-` before + `gnoscere` to know
- generically, the use of a **model** that leverages **observed information** to project **new information**

$$\tilde{Y}_{new} = \hat{f}(Y_{obs}, X_{obs})$$

- **problem to solve:** find an "appropriate" model $\hat{f}$ that can produce $\tilde{Y}_{new}$ with small errors ($\min\{|Y_{new} - \tilde{Y}_{new}|\}$)

some empirical considerations for prediction

- **Predictability** depends on (Hyndman et al. 2013):
  - how well we know factors that influence the predictions
  - how much data (and of what quality!)
  - recursive influence of predictions (especially forecasts)
- Key **question**: **what to predict?**
  - every item?
  - at what level of aggregation?
  - at what frequency?
- **Objective**: find a model with **consistently "small" predictive errors**
  - cope with risk of **overfitting** the model

what do we mean by “overfitting”?

- **overfitting:** capturing **patterns in the training data** that do not extend to new observations
- an overfit model may generate **systematically large predictive errors**
  - predictions are **not generalizable** to new data
- **challenge:** find the **set of predictors** that carry the appropriate “signal” to projections of the future
  - enough information to capture **meaningful patterns**
  - ...not so much as to also capture patterns that are **irrelevant for the future**
  - the **bias-variance tradeoff**

overfitting: the bias-variance tradeoff perspective

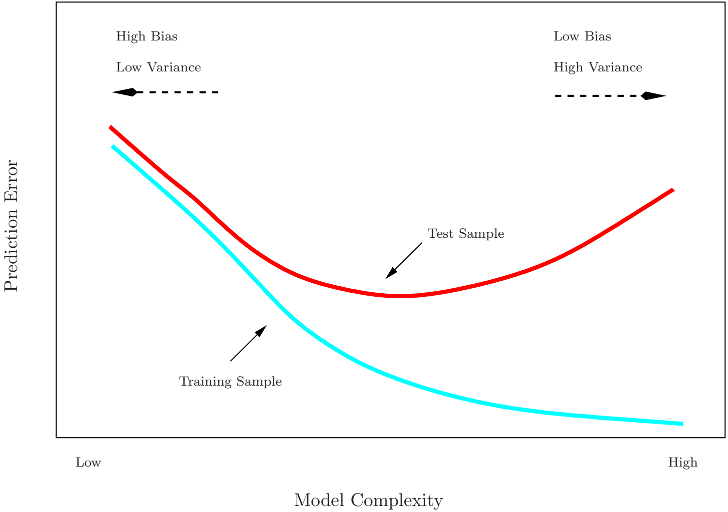
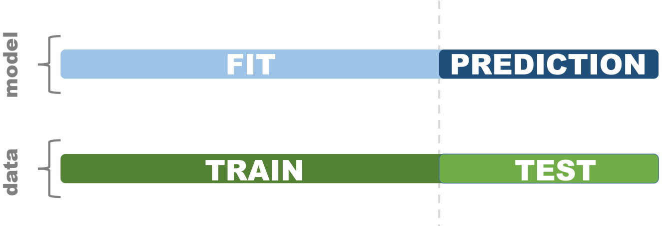


Figure: James et al (2013)

validation to minimize overfitting

i) fit model on a **training set**

ii) measure error on a **test set**

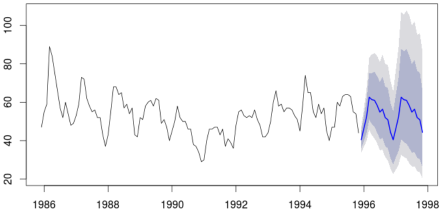


- usually, error on **test set** > error on **training set**
  - **caveat:** training and test sets should come **from the same population**
- **validation** can take many flavors (e.g. **k-fold validation**, **leave p-out cross-validation**...)

do not forget: predictions carry uncertainty!

- the future is unknown, therefore **predictions have uncertainty**
  - many predictive models only produce **point estimates** of predictions, and ignore **prediction intervals**
  - may generate **erroneous impression** that **predictions have no uncertainty**

- when possible, estimate the **range of values** where predictions may lie **with a given probability**



some (empirically validated) rules of thumb

1. **keep it simple:**
   - start parsimonious and add complexity (*iff* called for)
   - increased complexity typically reduces accuracy
2. **rely on domain expertise to select inputs**
   - statistical significance a faulty guide for inclusion
   - domain expertise should drive variables to include
3. **include more (useful) information**
   - high correlation in predictors not an issue
4. **fit $\neq$ accuracy**
   - well-fitting models may impose unwarranted "structure" and "certainty" to the forecast
5. **update models constantly**
   - update parameters as new information arrives

Explanation vs Prediction

Marco Morales
marco.morales@columbia.edu

Nana Yaw Essuman
nanayawce@gmail.com

GR5069
Topics in Applied Data Science
for Social Scientists

Spring 2020
Columbia University