Machine Learning Focus on Regression and Planting Other Seeds

Kirk Bansak

January 19, 2023

Working Example: Predicting Election Results

Goal: Use predictors to forecast Incumbent Vote Share

Potential predictors:

- GDP Growth
- Incumbent Approval
- Stock Market Trends
- Unemployment
- ...

Working Example: Predicting Election Results

Goal: Use predictors to forecast Incumbent Vote Share

Potential predictors:

- GDP Growth
- Incumbent Approval
- Stock Market Trends
- Unemployment
- ...

Conjecture: Model relationship in prior elections to predict future election

Model Variables

Input variables $\rightsquigarrow$ Approval, GDP Growth, Unemployment, etc.

- *predictors, independent variables, features, attributes, covariates*
- X; X_1 (Approval), X_2 (GDP Growth), X_3 (Unemployment), etc.

Output variable $\rightsquigarrow$ Incumbent Vote Share

- *response, dependent variable, outcome, target, label*
- Y
  - Quantitative (e.g. 15, 3.14, -82000) $\rightsquigarrow$ Regression Model
  - Categorical (e.g. Republican/Democrat, 0/1, High/Medium/Low) $\rightsquigarrow$ Classification Model

Supervised Learning

We assume some relationship between Y and $X = (X_1, X_2, ..., X_p)$, such that:

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

- f is some fixed but unknown function of $X_1, X_2, ..., X_p$
- ϵ is a random irreducible error term

Supervised Learning

We assume some relationship between Y and $X = (X_1, X_2, ..., X_p)$, such that:

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

- f is some fixed but unknown function of $X_1, X_2, ..., X_p$
- ϵ is a random irreducible error term

Supervised Learning

Using observed data (X and Y) to estimate f with $\hat{f}$

Supervised Learning

We assume some relationship between Y and $X = (X_1, X_2, ..., X_p)$, such that:

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

- f is some fixed but unknown function of $X_1, X_2, ..., X_p$
- ϵ is a random irreducible error term

Supervised Learning

Using observed data (X and Y) to estimate f with $\hat{f}$

Ultimate Goal: Build an $\hat{f}$ that is as close as possible to f

Why Estimate f?

Two main reasons that we may wish to estimate f:

Why Estimate f?

Two main reasons that we may wish to estimate f:

1. Inference
   - How is Y affected as $X_1, X_2, ..., X_p$ change?
   - The specific form/shape of $\hat{f}$ is (often) of central interest.
   - Better model = more interpretable

Why Estimate f?

Two main reasons that we may wish to estimate f:

1. Inference
   - How is Y affected as $X_1, X_2, ..., X_p$ change?
   - The specific form/shape of $\hat{f}$ is (often) of central interest.
   - Better model = more interpretable
2. Prediction
   - $\hat{Y} = \hat{f}(X)$
   - $\hat{f}$ is (often) treated as a *black box*.
   - Better model = more accurate predictions (i.e. $\hat{Y} \approx Y$)

Why Estimate f?

Two main reasons that we may wish to estimate f:

1. Inference
   - How is Y affected as $X_1, X_2, ..., X_p$ change?
   - The specific form/shape of $\hat{f}$ is (often) of central interest.
   - Better model $=$ more interpretable
2. Prediction
   - $\hat{Y} = \hat{f}(X)$
   - $\hat{f}$ is (often) treated as a *black box*.
   - Better model $=$ more accurate predictions (i.e. $\hat{Y} \approx Y$)

We'll focus mostly on prediction in this class.

Test Your Knowledge

Is the scenario a classification or regression problem? Are we most interested in inference or prediction? What are n and p?

We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Test Your Knowledge

Is the scenario a classification or regression problem? Are we most interested in inference or prediction? What are n and p?

We are considering launching a new product and wish to know whether it will be a success or a failure. We collect data on 20 similar products that were previously launched. For each product we have recorded whether it was a success or failure, price charged for the product, marketing budget, competition price, and ten other variables.

Classification, prediction, $n = 20$, $p = 13$.

Test Your Knowledge

Is the scenario a classification or regression problem? Are we most interested in inference or prediction? What are n and p?

We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry, and the CEO salary. We are interested in understanding which factors affect CEO salary.

Test Your Knowledge

Is the scenario a classification or regression problem? Are we most interested in inference or prediction? What are n and p?

We collect a set of data on the top 500 firms in the US. For each firm we record profit, number of employees, industry, and the CEO salary. We are interested in understanding which factors affect CEO salary.

Regression, inference, $n = 500$, $p = 3$.

How Do We Estimate f?

1. Collect a set of n data points, which include both the *output* and *input* variables, called training data.

Year	**Incumbent Vote Share**	Incumbent Net Approval	GDP Growth
2020	51.1	1.5	3.2
2018	49.1	-2.5	2.3
2016	56.8	22	3.0
⋮	⋮	⋮	⋮

How Do We Estimate f?

1. Collect a set of n data points, which include both the *output* and *input* variables, called training data.

Year	**Incumbent Vote Share**	Incumbent Net Approval	GDP Growth
2020	51.1	1.5	3.2
2018	49.1	-2.5	2.3
2016	56.8	22	3.0
$\vdots$	$\vdots$	$\vdots$	$\vdots$

2. Select a model or method for estimating f.

How Do We Estimate f?

1. Collect a set of n data points, which include both the *output* and *input* variables, called training data.

Year	**Incumbent Vote Share**	Incumbent Net Approval	GDP Growth
2020	51.1	1.5	3.2
2018	49.1	-2.5	2.3
2016	56.8	22	3.0
⋮	⋮	⋮	⋮

2. Select a model or method for estimating f.

3. Use training data to train (a.k.a. fit, estimate, build) $\hat{f}$, which will be our prediction function.

Looking Ahead (briefly)

1. Collect a set of n data points, which include both the *output* and *input* variables, called training data.
2. Select a model or method for estimating f.
3. Use training data to train (a.k.a. fit, estimate, build) $\hat{f}$, which will be our prediction function.

Looking Ahead (briefly)

1. Collect a set of n data points, which include both the *output* and *input* variables, called training data.
2. Select a model or method for estimating f.
3. Use training data to train (a.k.a. fit, estimate, build) $\hat{f}$, which will be our prediction function.
4. Use $\hat{f}$ to predict values for Y on previously unseen observations.

Year	**Incumbent Vote Share**	Incumbent Net Approval	GDP Growth
2022	?	12.2	3.4

Looking Ahead (briefly)

1. Collect a set of n data points, which include both the *output* and *input* variables, called training data.
2. Select a model or method for estimating f.
3. Use training data to train (a.k.a. fit, estimate, build) $\hat{f}$, which will be our prediction function.
4. Use $\hat{f}$ to predict values for Y on previously unseen observations.

Year	**Incumbent Vote Share**	Incumbent Net Approval	GDP Growth
2022	?	12.2	3.4

5. Compare predicted response value ($\hat{Y}$) with true response value (Y) for observations in test / validation data to evaluate performance.

Back to How We Estimate f

1. Collect a set of n data points, which include both the *output* and *input* variables, called training data.
2. Select a model or method for estimating f.
3. Use training data to train (a.k.a. fit, estimate, build) $\hat{f}$, which will be our prediction function.

Back to How We Estimate f

1. Collect a set of n data points, which include both the *output* and *input* variables, called training data.
2. Select a model or method for estimating f.
3. Use training data to train (a.k.a. fit, estimate, build) $\hat{f}$, which will be our prediction function.

There are many models and methods for estimating f!!!

Back to How We Estimate f

1. Collect a set of n data points, which include both the *output* and *input* variables, called training data.
2. Select a model or method for estimating f.
3. Use training data to train (a.k.a. fit, estimate, build) $\hat{f}$, which will be our prediction function.

There are many models and methods for estimating f!!!

There is no free lunch in statistics: no one method dominates all others over all possible data sets. Selecting the best method is hard.

Back to How We Estimate f

1. Collect a set of n data points, which include both the *output* and *input* variables, called training data.
2. Select a model or method for estimating f.
3. Use training data to train (a.k.a. fit, estimate, build) $\hat{f}$, which will be our prediction function.

There are many models and methods for estimating f!!!

There is no free lunch in statistics: no one method dominates all others over all possible data sets. Selecting the best method is hard.

The first thing we need to do is develop a toolkit of methods...

and linear regression will be where we start

Test Your Knowledge

What is the difference between $f(X)$ and $\hat{f}(X)$?

Test Your Knowledge

What is the difference between $f(X)$ and $\hat{f}(X)$?

Answer: $f(X)$ is the true function that maps X onto Y. $\hat{f}(X)$ is the estimated / prediction function trained on a sample of data, mapping observed X onto observed Y.

Linear Regression

What is Linear Regression?

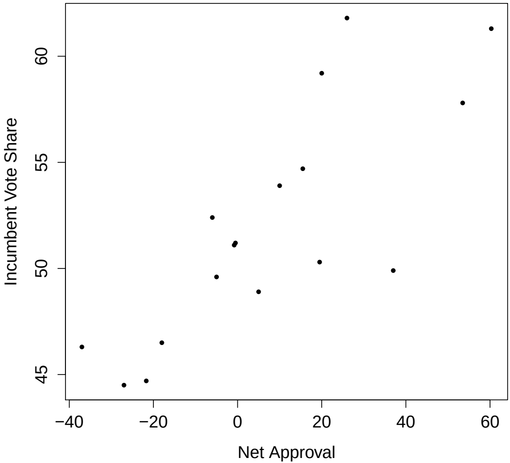
Linear regression is a simple approach for supervised learning.

- Around since 1800s.
- Still a widely used tool for predicting quantitative response.
- Building block for more sophisticated methods.
- We need to understand it before moving on!

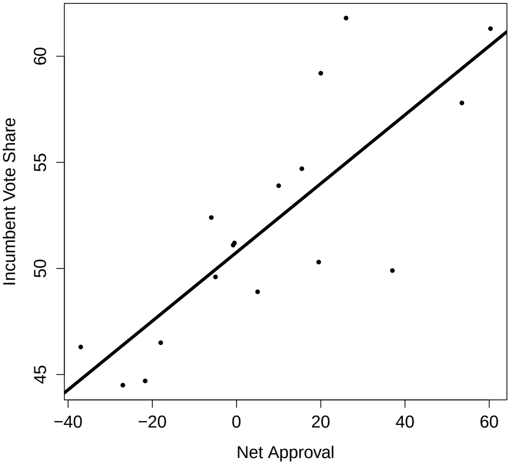
Simple linear regression:

- Assumes a linear relationship between quantitative response Y and a **single** variable X.
- Also called bivariate regression.

Bivariate Regression: Geometric Perspective

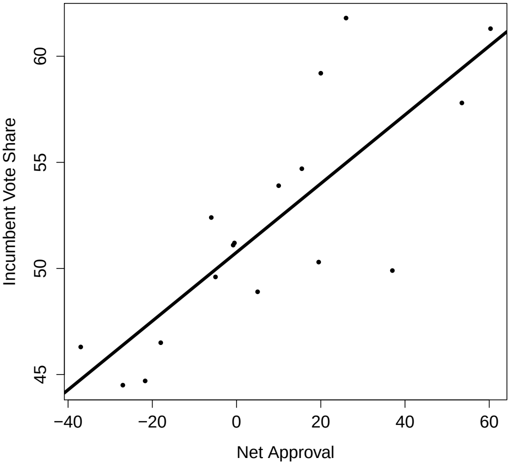


Bivariate Regression: Geometric Perspective



Bivariate Regression: Geometric Perspective

$$y = mx + b$$



Bivariate Regression: Function Perspective

For each election i (where i is used to index different observations),

Bivariate Regression: Function Perspective

For each election i (where i is used to index different observations),

Let:

$Vote_i$ = Incumbent Vote Share in election i.

$Approval_i$ = Incumbent Net Approval in election i.

Bivariate Regression: Function Perspective

For each election i (where i is used to index different observations),

Let:

$Vote_i$ = Incumbent Vote Share in election i.
$Approval_i$ = Incumbent Net Approval in election i.

We want to predict incumbent vote share using approval as our input. Thus, we need a function f that relates $Approval_i$ to $Vote_i$:

Bivariate Regression: Function Perspective

For each election i (where i is used to index different observations),

Let:

$Vote_i$ = Incumbent Vote Share in election i.
$Approval_i$ = Incumbent Net Approval in election i.

We want to predict incumbent vote share using approval as our input. Thus, we need a function f that relates $Approval_i$ to $Vote_i$:

$$\text{Vote}_i \quad = \quad f(\text{Approval}_i) + \epsilon_i$$

Bivariate Regression: Function Perspective

For each election i (where i is used to index different observations),

Let:

$Vote_i$ = Incumbent Vote Share in election i.
$Approval_i$ = Incumbent Net Approval in election i.

We want to predict incumbent vote share using approval as our input. Thus, we need a function f that relates $Approval_i$ to $Vote_i$:

$$\begin{aligned} \text{Vote}_i &= f(\text{Approval}_i) + \epsilon_i \\ \text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \end{aligned}$$

Bivariate Regression: Function Perspective

For each election i (where i is used to index different observations),

Let:

$Vote_i$ = Incumbent Vote Share in election i.
$Approval_i$ = Incumbent Net Approval in election i.

We want to predict incumbent vote share using approval as our input. Thus, we need a function f that relates $Approval_i$ to $Vote_i$:

$$\begin{aligned} \text{Vote}_i &= f(\text{Approval}_i) + \epsilon_i \\ \text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \end{aligned}$$

β_0 and β_1 are two unknown quantities known as the model coefficients or parameters.

Bivariate Regression: Machine Learning Perspective

Employing linear function for f as one way to relate $Approval_i$ to $Vote_i$:

$$\begin{aligned} \text{Vote}_i &= f(\text{Approval}_i) + \epsilon_i \\ \text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \end{aligned}$$

β_0 and β_1 are two unknown quantities known as the model coefficients or parameters, which we (the humans) do not choose.

Bivariate Regression: Machine Learning Perspective

Employing linear function for f as one way to relate $Approval_i$ to $Vote_i$:

$$\begin{aligned} \text{Vote}_i &= f(\text{Approval}_i) + \epsilon_i \\ \text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \end{aligned}$$

β_0 and β_1 are two unknown quantities known as the model coefficients or parameters, which we (the humans) do not choose.

Instead, our training data will be used to estimate β_0 and β_1.

Bivariate Regression: Machine Learning Perspective

Employing linear function for f as one way to relate $Approval_i$ to $Vote_i$:

$$\begin{aligned} \text{Vote}_i &= f(\text{Approval}_i) + \epsilon_i \\ \text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \end{aligned}$$

β_0 and β_1 are two unknown quantities known as the model coefficients or parameters, which we (the humans) do not choose.

Instead, our training data will be used to estimate β_0 and β_1. By estimating β_0 and β_1, we can then make predictions for $Vote_i$.

Bivariate Regression: Machine Learning Perspective

Employing linear function for f as one way to relate $Approval_i$ to $Vote_i$:

$$\begin{aligned} \text{Vote}_i &= f(\text{Approval}_i) + \epsilon_i \\ \text{Vote}_i &= \beta_0 + \beta_1 \text{Approval}_i + \epsilon_i \end{aligned}$$

β_0 and β_1 are two unknown quantities known as the model coefficients or parameters, which we (the humans) do not choose.

Instead, our training data will be used to estimate β_0 and β_1. By estimating β_0 and β_1, we can then make predictions for $Vote_i$.

That is, we will use our training data to produce estimates, $\hat{\beta}_0$ and $\hat{\beta}_0$, which we'll use to create a prediction function:

$$\widehat{Vote}_i = \hat{\beta}_0 + \hat{\beta}_1 Approval_i$$

Another Preview of Things to Come...

Uncertainty

In this class (and in machine learning applications more broadly), we will often want to not only produce estimates/predictions, but also quantify the uncertainty surrounding those estimates/predictions.

Simple (Univariate) Example of Uncertainty

Gallup Poll ($N \approx 1000$) on Approval of Congress (Nov/Dec, 2022):

22%

Simple (Univariate) Example of Uncertainty

Gallup Poll ($N \approx 1000$) on Approval of Congress (Nov/Dec, 2022):

22%

Problem: This calculation is based upon a small sample of the U.S. voting population, so how confident can we be that it reflects the mean approval in the overall population?

Simple (Univariate) Example of Uncertainty

Gallup Poll ($N \approx 1020$) on Approval of Congress (Nov/Dec, 2022):

$$22\% \pm 3\%$$

Simple (Univariate) Example of Uncertainty

Gallup Poll ($N \approx 1020$) on Approval of Congress (Nov/Dec, 2022):

$$22\% \pm 3\%$$

Error is based on 95% confidence.

Thus, 95% Confidence Interval:

$$[19\%, 25\%]$$

Simple (Univariate) Example of Uncertainty

Gallup Poll ($N \approx 1020$) on Approval of Congress (Nov/Dec, 2022):

$$22\% \pm 3\%$$

Error is based on 95% confidence.

Thus, 95% Confidence Interval:

$$[19\%, 25\%]$$

Where does that uncertainty quantification come from?

Uncertainty

When conducting statistical analyses, we will often want to not only produce estimates/predictions, but also quantify the uncertainty surrounding those estimates/predictions.

Two possible methods:

Uncertainty

When conducting statistical analyses, we will often want to not only produce estimates/predictions, but also quantify the uncertainty surrounding those estimates/predictions.

Two possible methods:

1. Classic statistical inference

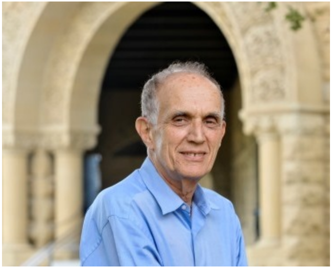
Uncertainty

When conducting statistical analyses, we will often want to not only produce estimates/predictions, but also quantify the uncertainty surrounding those estimates/predictions.

Two possible methods:

1. Classic statistical inference
2. **Bootstrap**

Uncertainty and the Bootstrap



Münchhausen O. Herrfurth pinx

Welcome to the Bootstrap World

Real World

Unknown probability distribution

Observed random sample

$$P \longrightarrow \mathrm{Z} = (\mathrm{Z}_1, \ldots, \mathrm{Z}_n)$$

$$\downarrow$$

$$\hat{\theta} = s(\mathrm{Z})$$

Statistic of interest

Welcome to the Bootstrap World

Real World

Unknown probability distribution

Observed random sample

$$P \longrightarrow \mathrm{Z} = (\mathrm{Z}_1, \ldots, \mathrm{Z}_n)$$

$$\downarrow$$

$$\hat{\theta} = s(\mathrm{Z})$$

Statistic of interest

- Given no other information, the observed sample Z contains all the available information about the underlying population distribution P

Welcome to the Bootstrap World

Real World			**Bootstrap World**	
Unknown probability distribution	Observed random sample		Empirical distribution	Bootstrap sample
$P \longrightarrow \mathrm{Z} = (\mathrm{Z}_1, \ldots, \mathrm{Z}_n)$		$\Longrightarrow$	$\hat{P} \longrightarrow \mathrm{Z}^* = (\mathrm{Z}_1^*, \ldots, \mathrm{Z}_n^*)$	
$\downarrow$			$\downarrow$	
$\hat{\theta} = s(\mathrm{Z})$			$\hat{\theta}^* = s(\mathrm{Z}^*)$	
Statistic of interest			Bootstrap replication	

- Given no other information, the observed sample Z contains all the available information about the underlying population distribution P
- Thus, resampling from Z is the best guide to what can be expected from resampling from P

Welcome to the Bootstrap World

Real World			**Bootstrap World**	
Unknown probability distribution	Observed random sample		Empirical distribution	Bootstrap sample
$P \longrightarrow \mathrm{Z} = (\mathrm{Z}_1, \ldots, \mathrm{Z}_n)$		$\Longrightarrow$	$\hat{P} \longrightarrow \mathrm{Z}^* = (\mathrm{Z}_1^*, \ldots, \mathrm{Z}_n^*)$	
$\downarrow$			$\downarrow$	
$\hat{\theta} = s(\mathrm{Z})$			$\hat{\theta}^* = s(\mathrm{Z}^*)$	
Statistic of interest			Bootstrap replication	

- Given no other information, the observed sample Z contains all the available information about the underlying population distribution P
- Thus, resampling from Z is the best guide to what can be expected from resampling from P
- Just like we drew Z from P, let's draw a resample Z^* from Z

Welcome to the Bootstrap World

Real World			**Bootstrap World**	
Unknown probability distribution	Observed random sample		Empirical distribution	Bootstrap sample
$P \longrightarrow \mathrm{Z} = (\mathrm{Z}_1, \ldots, \mathrm{Z}_n)$		$\Longrightarrow$	$\hat{P} \longrightarrow \mathrm{Z}^* = (\mathrm{Z}_1^*, \ldots, \mathrm{Z}_n^*)$	
$\downarrow$			$\downarrow$	
$\hat{\theta} = s(\mathrm{Z})$			$\hat{\theta}^* = s(\mathrm{Z}^*)$	
Statistic of interest			Bootstrap replication	

- Given no other information, the observed sample Z contains all the available information about the underlying population distribution P
- Thus, resampling from Z is the best guide to what can be expected from resampling from P
- Just like we drew Z from P, let's draw a resample Z^* from Z
- If n is sufficiently large, the observed sample Z should be a good approximation of P (i.e. treat Z as $\hat{P}$)

(Nonparametric) Bootstrap

1. Decide on a summary statistic s of interest.
   - e.g. if Z contains only a single variable, perhaps the mean of that variable

(Nonparametric) Bootstrap

1. Decide on a summary statistic s of interest.
   - e.g. if Z contains only a single variable, perhaps the mean of that variable
2. Draw B resamples of n observations from Z **with replacement**.
   - Note: If Z contains multiple variables, resampled observations should be kept intact (i.e. full rows should be drawn).

   Call the bth resample Z_b^*.

(Nonparametric) Bootstrap

1. Decide on a summary statistic s of interest.
   - e.g. if Z contains only a single variable, perhaps the mean of that variable
2. Draw B resamples of n observations from Z **with replacement**.
   - Note: If Z contains multiple variables, resampled observations should be kept intact (i.e. full rows should be drawn).

   Call the bth resample Z_b^*.
3. For each Z_b^*, compute $s(Z_b^*)$ and store it, creating the set $\{s(Z_1^*), ..., s(Z_B^*)\}$

(Nonparametric) Bootstrap

1. Decide on a summary statistic s of interest.
   - e.g. if Z contains only a single variable, perhaps the mean of that variable
2. Draw B resamples of n observations from Z **with replacement**.
   - Note: If Z contains multiple variables, resampled observations should be kept intact (i.e. full rows should be drawn).

   Call the bth resample Z_b^*.
3. For each Z_b^*, compute $s(Z_b^*)$ and store it, creating the set $\{s(Z_1^*), ..., s(Z_B^*)\}$
4. To compute 95% CI, use 2.5/97.5 percentiles of $\{s(Z_1^*), ..., s(Z_B^*)\}$ as the lower/upper bounds (bootstrap percentile CI)

1. Section tomorrow
2. Readings for next Tuesday are listed on syllabus and posted in Files on bCourses site
3. Install/Update `R` and `RStudio` on your computer by next Thursday