LECTURE #10: PROBABILISTIC MODELS FOR INFERENCE, PART II

CS 109A, STAT 121A, AC 209A: Data Science

Weiwei Pan, Pavlos Protopapas, Kevin Rader

Fall 2016

Harvard University

ANNOUNCEMENTS

- Special office hours today with Pavlos at 4-5pm, WP at 5-6pm (open for Zoom)
- An anonymous midterm survey is now available on Canvas!
- New Project Milestone due dates are on course calendar.
- **HW feedback:** If you're doing a fine job with no obvious errors, you may not get comments beyond "good job".

  If you are making errors in code or analysis, you should have specific comments indicating the type of error we see and some directions on how to improve.

  If you do not see these comments, or don't know how to interpret them, contact the helpline! **Providing useful feedback to you is our top priority**!
- **Midterm policies:** open course resources and open Google (and resources linked by Google). Midterm is **not** open to collaboration or discussion with human beings.

  Multiple choice part on Canvas (like Quiz), code part off-line, upload ipynb to Canvas.

  During the exam, email all questions to the helpline!

QUIZ TIME

Password: thirdoneistheworst

LECTURE OUTLINE

Last Time in CS109A...

Estimating Model Parameters from the Posterior

Summary

Loose Ends and Lingering Questions

Being a Responsible Data Scientist

Hint for Midterm

TALK OUTLINE

Last Time in CS109A...

Estimating Model Parameters from the Posterior

Summary

Loose Ends and Lingering Questions

Being a Responsible Data Scientist

Hint for Midterm

OUR ORIGIN STORY FOR THE DATA IN LINEAR REGRESSION

Our belief: The relationship between *price* (y) and *square footage* (x) is linear, and that observed prices differ from our pricing rule by some random amount, ϵ, which we call residual or ***noise***.

$$y = \underbrace{\beta_1 \cdot x + \beta_0}_{\text{theoretical price}} + \underbrace{\epsilon}_{\text{noise}}, \qquad \underbrace{\epsilon \sim \mathcal{N}(0, \sigma^2)}_{\text{noise is normally distributed}}$$

For a set of observations $\mathbf{X} = \{x_1, \ldots, x_N\}$, $\mathbf{Y} = \{y_1, \ldots, y_N\}$, and a set of noise $\epsilon = \{\epsilon_1, \ldots, \epsilon_N\}$, we **believe** that

1. the noise is identically distributed, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, so then...

$$\underbrace{y_i | \beta_1, \beta_0, x_i \sim \mathcal{N}(\beta_1 \cdot x_i + \beta_0, \sigma^2)}_{\text{each } y_i \text{ is normally distributed, fixing } \beta_1, \beta_0, x_i}$$

2. the noise is independent, $\epsilon_i \perp \epsilon_j$, so then...

$$\underbrace{L(\beta_1, \beta_0) = p(\mathbf{Y} | \beta_1, \beta_0, \mathbf{X})}_{\text{likelihood of the data given the model}} = \underbrace{\prod_{i=1}^{N} \mathcal{N}(\beta_1 \cdot x_i + \beta_0, \sigma^2)}_{\text{product of the pdf of each } y_i}$$

HOW TO INTERPRET LIKELIHOOD

The likelihood function,

$$L(\textbf{model}) = p(\textbf{response} \mid \textbf{model}, \ \textbf{predictors}),$$

tells us how likely or probable it is to observe a set of data while assuming a particular model.

Thus,

$$L(\beta_1, \beta_0) = p(\mathbf{Y}|\beta_1, \beta_0, \mathbf{X})$$

tells us how likely are we to observe $\mathbf{Y} = \{y_1, \ldots, y_N\}$ if $y = \beta_1 x + \beta_0 + \epsilon$ is the model.

Your mandate to me: We **must** choose the model which renders the observed data the most probable. I.e. if we believe in our story for the data, we **are obligated** to choose parameters to **maximize likelihood**.

MLE FOR LINEAR REGRESSION

For linear regression with one predictor, we saw that

$$\max \underbrace{L(\beta_1, \beta_0)}_{\text{likelihood}} \Leftrightarrow \max \underbrace{\ln L(\beta_1, \beta_0)}_{\text{log likelihood}} \Leftrightarrow \min \underbrace{\sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2}_{\text{loss function for OLS}}$$

Observation: Maximizing likelihood is equivalent to minimizing Residual Sum of Squares!

Model parameters that maximize the likelihood of the data are called ***maximum likelihood estimates***, or MLE, and are denoted $\beta_0^{MLE}, \beta_1^{MLE}$.

THE NECESSITY OF PRIORS

Often, we **learn** "best fitting models" by optimizing one set of metrics (e.g. MLE), but we **evaluate** the models we find by an entirely different set of standards.

We often reject models b/c they do not conform to our implicit, **prior beliefs** about what the model should look like.

Your mandate to me: If we have strong prior beliefs about the model parameters, we **must** explicitly incorporate them into our model!

Simple Linear Regression

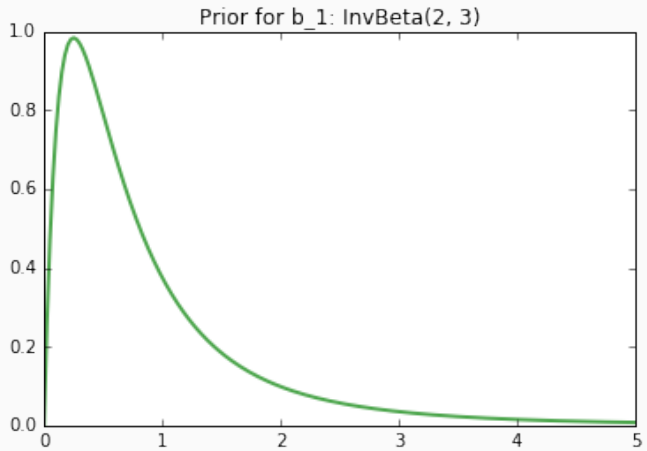
When we're modeling the housing prices data, with $y = \beta_1 x + \beta_0$,

- We believe that β_1 can't be negative
- We believe that β_0 is probably positive, and can't be too large

Simple Linear Regression

When we're modeling the housing prices data, with $y = \beta_1 x + \beta_0$,

- We believe that β_1 can't be negative



$$\beta_1 \sim \text{InvBeta}(2,3), \quad p(\beta_1) = \frac{\beta_1(1+\beta_1)^{-5}}{B(2,3)}, \quad B(2,3) = \int_0^\infty t/(1+t)^5 dt$$

- We believe that β_0 is probably positive, and can't be too large

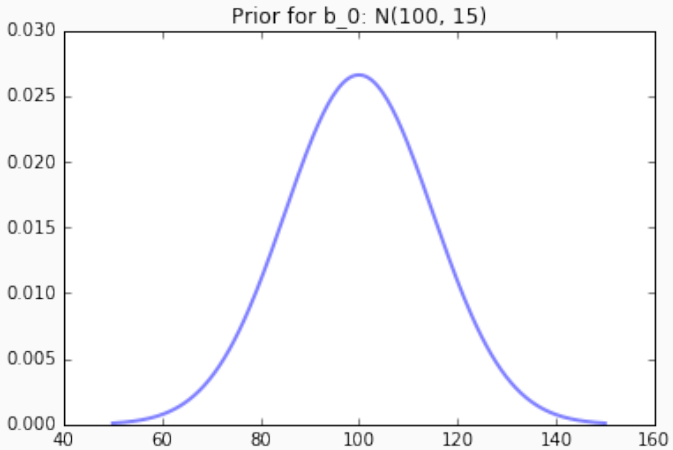
Simple Linear Regression

When we're modeling the housing prices data, with $y = \beta_1 x + \beta_0$,

- We believe that β_1 can't be negative

  $\beta_1 \sim \text{InvBeta}(2,3), \quad p(\beta_1) = \frac{\beta_1(1+\beta_1)^{-5}}{B(2,3)}, \quad B(2,3) = \int_0^\infty t/(1+t)^5 dt$

- We believe that β_0 is probably positive, and can't be too large



$$\beta_0 \sim \mathcal{N}(100, 15), \quad p(\beta_0) = C * \exp\{-\frac{(\beta_0 - 100)^2}{K}\}$$

THE POSTERIOR DISTRIBUTION

If we want to consider the likelihood and priors in conjunction we should multiply their pdf's:

$$\underbrace{p(\mathbf{Y}|\beta_1, \beta_0, \mathbf{X})}_{\text{old: likelihood}} * \underbrace{p(\beta_1) * p(\beta_0)}_{\text{new: priors}} \quad \textit{or} \quad \underbrace{p(\mathbf{Y}|\beta_1, \beta_0, \mathbf{X})}_{\text{old: likelihood}} * \underbrace{p(\beta_1, \beta_0)}_{\text{new: priors}}$$

If we want to consider the likelihood and priors in conjunction we should multiply their pdf's:

$$\underbrace{p(\mathbf{Y}|\beta_1, \beta_0, \mathbf{X})}_{\text{old: likelihood}} * \underbrace{p(\beta_1) * p(\beta_0)}_{\text{new: priors}} \quad \textit{or} \quad \underbrace{p(\mathbf{Y}|\beta_1, \beta_0, \mathbf{X})}_{\text{old: likelihood}} * \underbrace{p(\beta_1, \beta_0)}_{\text{new: priors}}$$

Using **Bayes Rule**, we can express the above product succinctly:

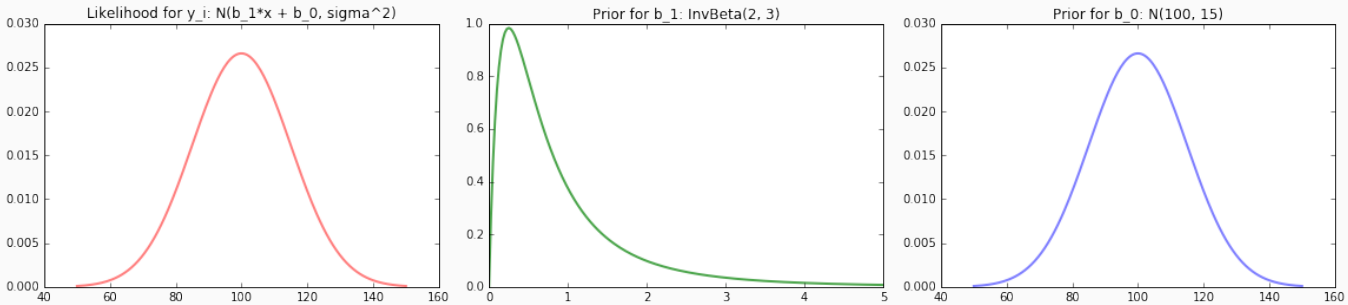
$$\underbrace{p(\mathbf{Y}|\beta_1, \beta_0, \mathbf{X})}_{\text{old: likelihood}} * \underbrace{p(\beta_1) * p(\beta_0)}_{\text{new: priors}} \propto \underbrace{p(\beta_1, \beta_0|\mathbf{Y}, \mathbf{X})}_{\text{posterior}}$$

The distribution of the model parameters *given* the data is called the ***posterior distribution***.

Simple Linear Regression

For data $\mathbf{X} = \{x_1, \ldots, x_N\}$, $\mathbf{Y} = \{y_1, \ldots, y_N\}$ with i.i.d noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Using the priors we selected for the housing prices dataset, our posterior looks like

$$\underbrace{p(\beta_1, \beta_0 | \mathbf{Y}, \mathbf{X})}_{\text{Posterior}} \propto \underbrace{\prod_{i=1}^{N} \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2)}_{\text{Likelihood}} * \underbrace{\text{InvBeta}(\beta_1; 2, 3) * \mathcal{N}(\beta_0; 100, 15)}_{\text{Priors}}$$



TALK OUTLINE

Last Time in CS109A...

Estimating Model Parameters from the Posterior

Summary

Loose Ends and Lingering Questions

Being a Responsible Data Scientist

Hint for Midterm

Question: What does the posterior distribution mean? And what is it good for?

Example

Say we're considering two linear models for the data $\{(1, 2)\}$:

- $M_1 : y = x + 2$

  $$\underbrace{p(\beta_1 = 1, \beta_0 = 2 | x = 1, y = 2)}_{posterior} = 1.5$$

- $M_2 : y = 2 - 2x$

  $$\underbrace{p(\beta_1 = -2, \beta_0 = 2 | x = 1, y = 2)}_{posterior} = 0.0001$$

Which model is the most appropriate for the data?

Question: What does the posterior distribution mean? And what is it good for?

Example

Say we're considering two linear models for the data $\{(1, 2)\}$:

- $M_1 : y = x + 2$

  $$\underbrace{p(\beta_1 = 1, \beta_0 = 2 | x = 1, y = 2)}_{posterior} = 1.5$$

- $M_2 : y = 2 - 2x$

  $$\underbrace{p(\beta_1 = -2, \beta_0 = 2 | x = 1, y = 2)}_{posterior} = 0.0001$$

Which model is the most appropriate for the data?

M_2 is 10,000 more likely than M_1, given the observed data.

THE MAXIMUM A POSTERIORI ESTIMATE

Observation: The posterior distribution tells us how likely is a set of model parameters given the data.

Goal: We want to find the model parameters that maximizes the posterior distribution.

Model parameters that maximize the posterior are called ***maximum a posteriori estimates***, or MAP, and are denoted $\beta_0^{MAP}, \beta_1^{MAP}$.

Goal: We want to find the model parameters that maximizes the posterior distribution.

Simple Linear Regression

For data $\mathbf{X} = \{x_1, \ldots, x_N\}$, $\mathbf{Y} = \{y_1, \ldots, y_N\}$ with i.i.d. noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Using the priors we selected for the housing prices dataset, our posterior looks like

$$\underbrace{p(\beta_1, \beta_0 | \mathbf{Y}, \mathbf{X})}_{\text{Posterior}} \propto \underbrace{\prod_{i=1}^{N} \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2)}_{\text{Likelihood}} * \underbrace{\mathrm{InvBeta}(\beta_1; 2, 3) * \mathcal{N}(\beta_0; 100, 15)}_{\text{Priors}}$$

$$= \left(\prod_{i=1}^{N} C_y * \exp\left\{ -\frac{(y_i - \beta_1 x_i - \beta_0)^2}{K_y} \right\} \right) * \left(\frac{\beta_1 (1+\beta_1)^{-5}}{B(2,3)} \right) * \left(C_{\beta_0} * \exp\left\{ -\frac{(\beta_0 - 100)^2}{K_{\beta_0}} \right\} \right)$$

Goal: We want to find the model parameters that maximizes the posterior distribution.

Simple Linear Regression

For data $\mathbf{X} = \{x_1, \ldots, x_N\}$, $\mathbf{Y} = \{y_1, \ldots, y_N\}$ with i.i.d. noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$. Using the priors we selected for the housing prices dataset, our posterior looks like

$$\underbrace{p(\beta_1, \beta_0 | \mathbf{Y}, \mathbf{X})}_{\text{Posterior}} \propto \left(\prod_{i=1}^{N} C_y * \exp\left\{ -\frac{(y_i - \beta_1 x_i - \beta_0)^2}{K_y} \right\} \right) * \left(\frac{\beta_1 (1+\beta_1)^{-5}}{B(2,3)} \right) * \left(C_{\beta_0} * \exp\left\{ -\frac{(\beta_0 - 100)^2}{K_{\beta_0}} \right\} \right)$$

Maximizing $p(\beta_1, \beta_0 | y, x)$ will involve taking the (partial) derivative(s) of the above and solving a system of nonlinear equations. **That sounds hard!**

Goal: We want to find the model parameters that maximizes the posterior distribution.

Let's choose some easier priors of β_1 and β_0. Say, $\beta_0, \beta_1 \sim \mathcal{N}(0, 1/\lambda)$ (assuming $\sigma^2 = 1$).

Then, our posterior looks like:

$$\underbrace{p(\beta_1, \beta_0 | \mathbf{Y}, \mathbf{X})}_{\text{Posterior}} \propto \underbrace{\prod_{i=1}^{N} \mathcal{N}(y_i; \beta_1 x_i + \beta_0, 1)}_{\text{Likelihood}} * \underbrace{\mathcal{N}(\beta_1; 0, 1/\lambda) * \mathcal{N}(\beta_0; 0, 1/\lambda)}_{\text{Priors}}$$

Let's make the posterior friendlier by taking the log

$$\underbrace{\ln p(\beta_1, \beta_0 | \mathbf{Y}, \mathbf{X})}_{\text{log posterior}} \propto \ln \prod_{i=1}^{N} \mathcal{N}(y_i; \beta_1 x_i + \beta_0, 1) * \mathcal{N}(\beta_1; 0, 1/\lambda) * \mathcal{N}(\beta_0; 0, 1/\lambda)$$

$$= \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, 1) + \ln \mathcal{N}(\beta_1; 0, 1/\lambda) + \ln \mathcal{N}(\beta_1; 0, 1/\lambda)$$

$$= \underbrace{\sum_{i=1}^{N} \ln C_y - \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2}_{\text{log likelihood from before}} + \underbrace{\ln C_{\beta_0} - \frac{1}{2} \lambda \beta_0^2}_{\text{log of } p(\beta_0)} + \underbrace{\ln C_{\beta_1} - \frac{1}{2} \lambda \beta_1^2}_{\text{log of } p(\beta_1)}$$

To maximize the posterior, we can ignore the constants (highlighted), and minimize the quantity:

$$\min \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2 + \lambda(\beta_0^2 + \beta_1^2)$$

Let's make the posterior friendlier by taking the log

$$\underbrace{\ln p(\beta_1, \beta_0 | \mathbf{Y}, \mathbf{X})}_{\text{log posterior}} \propto \ln \prod_{i=1}^{N} \mathcal{N}(y_i; \beta_1 x_i + \beta_0, 1) * \mathcal{N}(\beta_1; 0, 1/\lambda) * \mathcal{N}(\beta_0; 0, 1/\lambda)$$

$$= \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, 1) + \ln \mathcal{N}(\beta_1; 0, 1/\lambda) + \ln \mathcal{N}(\beta_1; 0, 1/\lambda)$$

$$= \underbrace{\sum_{i=1}^{N} \ln C_y - \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2}_{\text{log likelihood from before}} + \underbrace{\ln C_{\beta_0} - \frac{1}{2} \lambda \beta_0^2}_{\text{log of } p(\beta_0)} + \underbrace{\ln C_{\beta_1} - \frac{1}{2} \lambda \beta_1^2}_{\text{log of } p(\beta_1)}$$

To maximize the posterior, we can ignore the constants (highlighted), and minimize the quantity:

$$\min \underbrace{\sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2 + \lambda(\beta_0^2 + \beta_1^2)}_{\text{loss function of ridge regression}}$$

Goal: We want to find the model parameters that maximizes the posterior distribution.

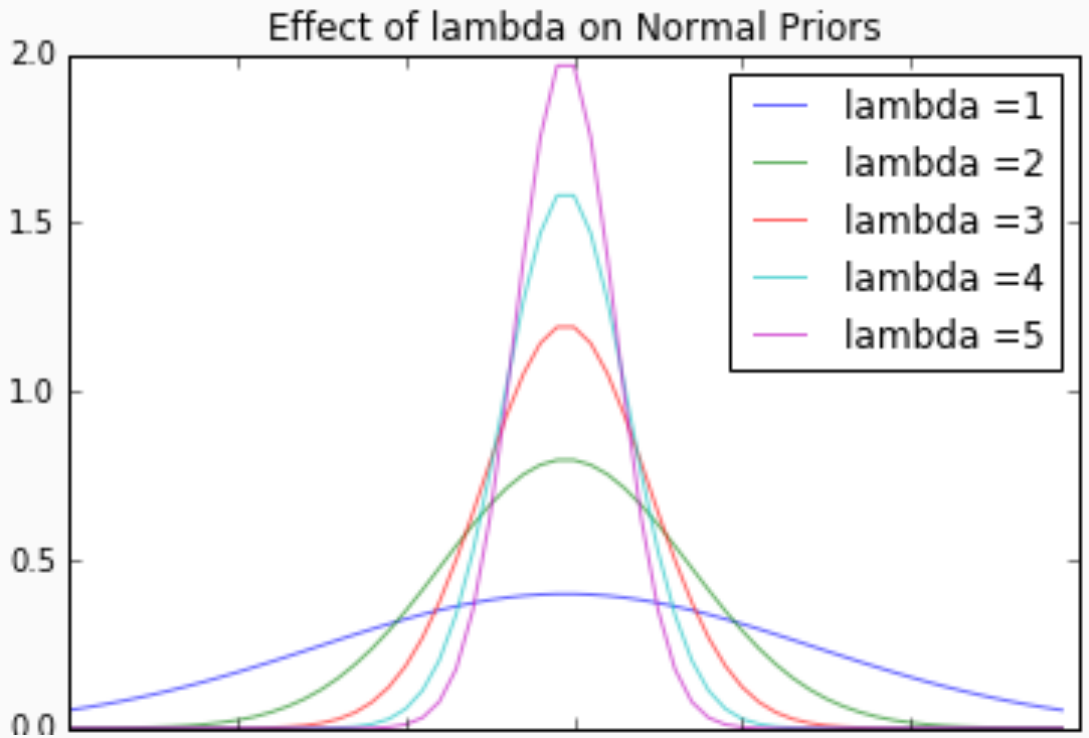
Observation: With normal priors for β_1 and β_0,

$$\underbrace{p(\beta_1, \beta_0 | \mathbf{Y}, \mathbf{X})}_{\text{Posterior}} \propto \underbrace{\prod_{i=1}^{N} \mathcal{N}(y_i; \beta_1 x_i + \beta_0, 1)}_{\text{Likelihood}} * \underbrace{\mathcal{N}(\beta_1; 0, 1/\lambda) * \mathcal{N}(\beta_0; 0, 1/\lambda)}_{\text{Priors}}$$

the **MAP estimates of β_1 and β_0 are precisely those found by ridge regression.**

Question: What kind of beliefs does a normal prior $\mathcal{N}(0, 1/\lambda)$ encode?

Question: What is the effect of λ on the normal priors?



LINEAR REGRESSION WITH LAPLACE PRIOR

Goal: We want to find the model parameters that maximizes the posterior distribution.

Let's choose some different priors for β_1 and β_0. Say, $\beta_0, \beta_1 \sim \mathcal{L}(0, 1/\lambda)$ (assuming $\sigma^2 = 1$), where $\mathcal{L}(0, 1/\lambda)$ is a Laplace distribution.

Then, our posterior looks like:

$$\underbrace{p(\beta_1, \beta_0 | \mathbf{Y}, \mathbf{X})}_{\text{Posterior}} \propto \underbrace{\prod_{i=1}^{N} \mathcal{N}(y_i; \beta_1 x_i + \beta_0, 1)}_{\text{Likelihood}} * \underbrace{\mathcal{L}(\beta_1; 0, 1/\lambda) * \mathcal{L}(\beta_0; 0, 1/\lambda)}_{\text{Priors}}$$

Let's make the posterior friendlier by taking the log

$$\underbrace{\ln p(\beta_1, \beta_0 | Y, X)}_{\text{log posterior}} \propto \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, 1) + \ln \mathcal{L}(\beta_1; 0, 1/\lambda) + \ln \mathcal{L}(\beta_0; 0, 1/\lambda)$$

$$= \sum_{i=1}^{N} \ln \left[C_y * \exp\left\{ \frac{(y - \beta_1 x_i - \beta_0)^2}{2} \right\} \right] + \ln \left[C_{\beta_1} * \exp\left\{ \frac{\lambda |\beta_1|}{2} \right\} \right] + \ln \left[C_{\beta_0} * \exp\left\{ \frac{\lambda |\beta_0|}{2} \right\} \right]$$

$$= \underbrace{\sum_{i=1}^{N} \ln C_y - \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2}_{\text{log likelihood from before}} + \underbrace{\ln C_{\beta_1} - \frac{1}{2} \lambda |\beta_1|}_{\text{log of } p(\beta_1)} + \underbrace{\ln C_{\beta_0} - \frac{1}{2} \lambda |\beta_0|}_{\text{log of } p(\beta_0)}$$

To maximize the posterior, we can ignore the constants (highlighted), and minimize the quantity:

$$\min \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2 + \lambda(|\beta_0| + |\beta_1|)$$

Let's make the posterior friendlier by taking the log

$$\underbrace{\ln p(\beta_1, \beta_0 | Y, X)}_{\text{log posterior}} \propto \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, 1) + \ln \mathcal{L}(\beta_1; 0, 1/\lambda) + \ln \mathcal{L}(\beta_0; 0, 1/\lambda)$$

$$= \sum_{i=1}^{N} \ln \left[C_y * \exp \left\{ \frac{(y - \beta_1 x_i - \beta_0)^2}{2} \right\} \right] + \ln \left[C_{\beta_1} * \exp \left\{ \frac{\lambda |\beta_1|}{2} \right\} \right] + \ln \left[C_{\beta_0} * \exp \left\{ \frac{\lambda |\beta_0|}{2} \right\} \right]$$

$$= \underbrace{\sum_{i=1}^{N} \ln C_y - \frac{1}{2} \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2}_{\text{log likelihood from before}} + \underbrace{\ln C_{\beta_1} - \frac{1}{2} \lambda |\beta_1|}_{\text{log of } p(\beta_1)} + \underbrace{\ln C_{\beta_0} - \frac{1}{2} \lambda |\beta_0|}_{\text{log of } p(\beta_0)}$$

To maximize the posterior, we can ignore the constants (highlighted), and minimize the quantity:

$$\min \underbrace{\sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2 + \lambda(|\beta_0| + |\beta_1|)}_{\text{loss function of LASSO}}$$

LINEAR REGRESSION WITH LAPLACE PRIOR

Goal: We want to find the model parameters that maximizes the posterior distribution.

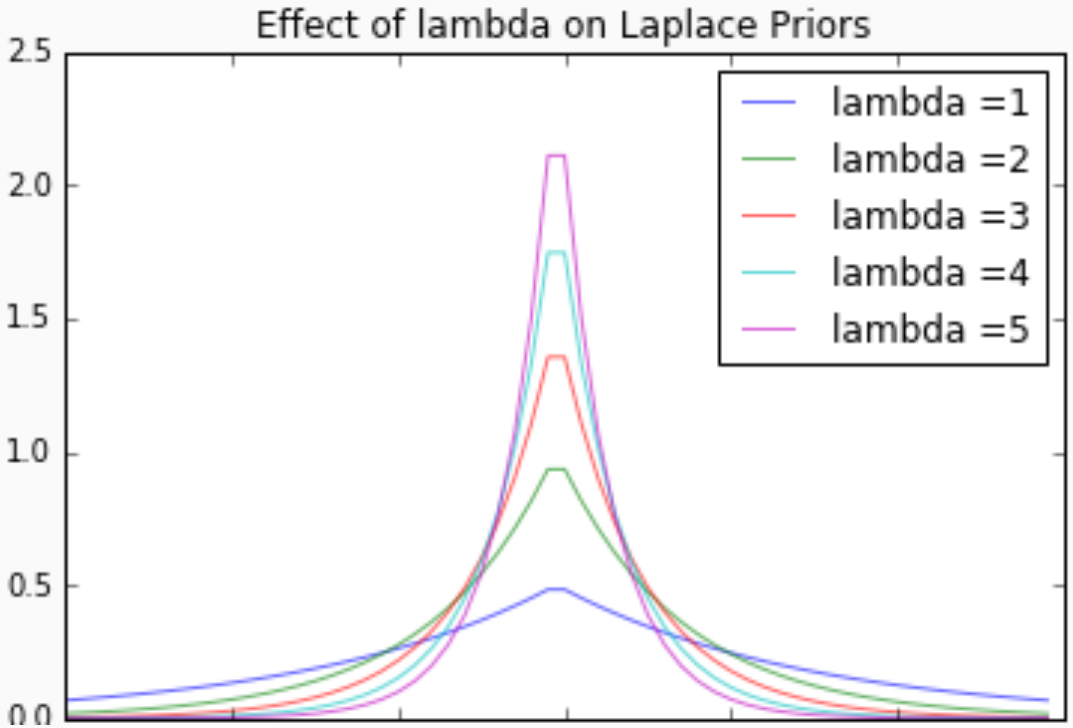
Observation: With Laplace priors for β_1 and β_0,

$$\underbrace{p(\beta_1, \beta_0 | \mathbf{Y}, \mathbf{X})}_{\text{Posterior}} \propto \underbrace{\prod_{i=1}^{N} \mathcal{N}(y_i; \beta_1 x_i + \beta_0, 1)}_{\text{Likelihood}} * \underbrace{\mathcal{L}(\beta_1; 0, 1/\lambda) * \mathcal{L}(\beta_0; 0, 1/\lambda)}_{\text{Priors}}$$

the **MAP estimates of β_1 and β_0 are precisely those found by LASSO.**

Question: What kind of beliefs does a Laplace prior $\mathcal{L}(0, 1/\lambda)$ encode?

Question: What is the effect of λ on the Laplace priors?



TALK OUTLINE

Last Time in CS109A...

Estimating Model Parameters from the Posterior

Summary

Loose Ends and Lingering Questions

Being a Responsible Data Scientist

Hint for Midterm

WAIT...WHAT WAS ALL THAT AGAIN?

1. (**Non-Probabilistic Regression**) Learn parameters, β_0 and β_1, to minimize a loss function, e.g. in OLS we solve

$$\min RSS = \min \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2$$

2. (**Probabilistic Regression**) Learn parameters, β_0 and β_1, to maximize the likelihood, i.e. the probability of *data given the parameters*

$$\max \underbrace{p(\mathbf{Y}|\beta_1, \beta_0, \mathbf{X})}_{\text{Likelihood}}$$

   The **maximum likelihood estimators** (MLE) parameters are the ones from OLS.

3. (**Bayesian Regression**) Learn parameters, β_0 and β_1, to maximize the posterior, i.e. the probability of *parameters given the data*

$$\max \underbrace{p(\beta_1, \beta_0, |\mathbf{Y}, \mathbf{X})}_{\text{Posterior}} \propto \max \underbrace{p(\mathbf{Y}|\beta_1, \beta_0, \mathbf{X})}_{\textit{Likelihood}} \underbrace{p(\beta_1)p(\beta_0)}_{\textit{Priors}} \tag{1}$$

   The **maximum a posteriori estimators** (MAP) parameters are the ones from regularized least squares (ridge or LASSO).

TALK OUTLINE

Last Time in CS109A...

Estimating Model Parameters from the Posterior

Summary

Loose Ends and Lingering Questions

Being a Responsible Data Scientist

Hint for Midterm

AFFECTS OF PRIORS

We've see that different choices of prior lead to different MAP estimates of model parameters!

- Choosing normal priors in linear regression leads to ridge regression
- Choosing Laplacian priors in linear regression leads to LASSO

Question: So how do we choose a "good prior"? Is there a universally "good" set of priors we should always be choosing?

USING NON-CONJUGATE PRIORS

Question: Isn't it too arbitrary to choose priors simply because they are mathematically convenient?

Question: If we choose complicated priors, how do we find the MAP?

MLE VS MAP

Question: Is there even a point to finding MAP? That is, are MLE and MAP estimates different?

Question: If they are different, which one is “better”?

ANALOGUE OF CONFIDENCE INTERVALS

Question: How “certain” are we in our MAP estimate?

Question: What’s the difference between a point estimate (MLE, MAP) and “confidence intervals” or “intervals of certainty”?

TALK OUTLINE

Last Time in CS109A...

Estimating Model Parameters from the Posterior

Summary

Loose Ends and Lingering Questions

Being a Responsible Data Scientist

Hint for Midterm

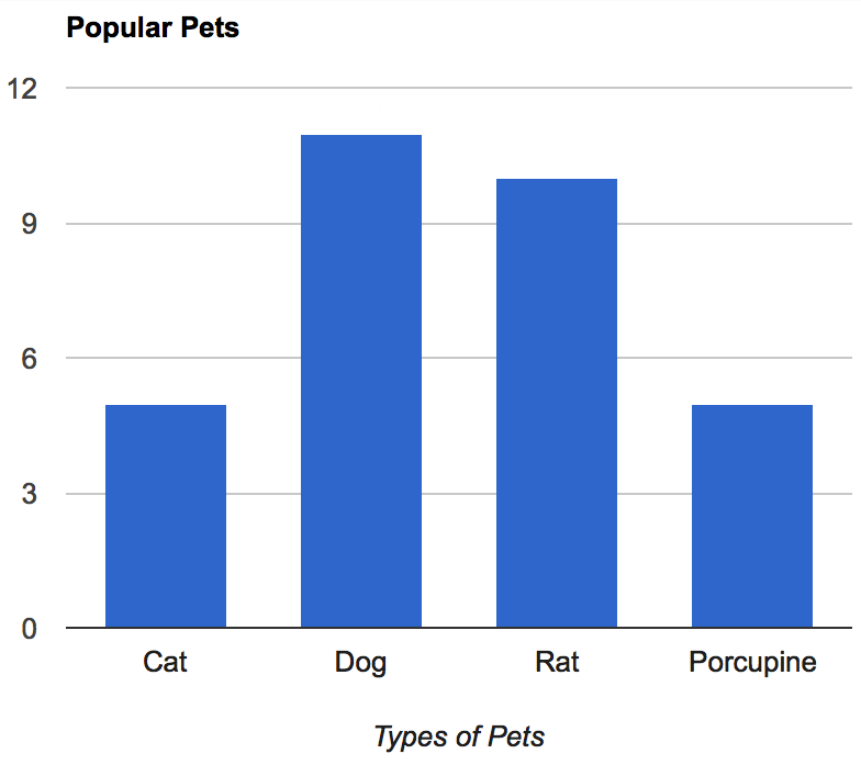
BACK TO THE BASICS

After six rigorous weeks of training in data science, we should all be able to answer the following with ease.

1. When you are asked "**what is a typical *x* value?**", you compute...
2. When you are asked "**what is the spread of *x* values?**", you compute...
3. When you are asked "**how good is your model?**", you compute...
4. When you are asked "**which predictors are most significant?**", you compute...

THE PROBLEM WITH "AVERAGE"

From Lecture #1: for samples of categorical variables, neither mean or median make sense.



The **mode** might be a better way to find the most "representative" value.

THE PROBLEM WITH "VARIANCE"

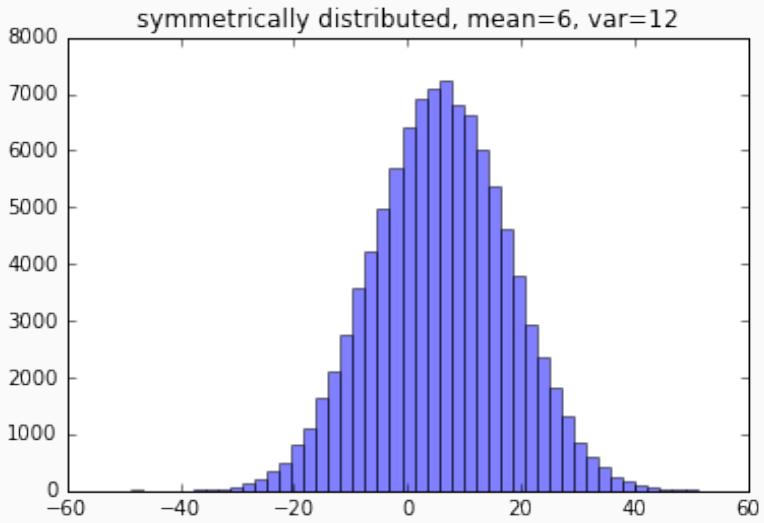
Given two sets of observations, X_1 and X_2, suppose I tell you that

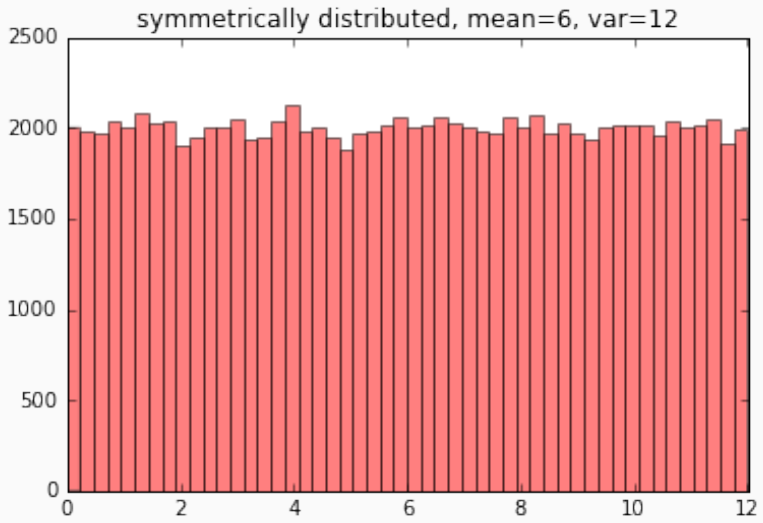
1. $\mathrm{Var}(\mathsf{X}_1) = \mathrm{Var}(\mathsf{X}_2)$
2. $\overline{\mathsf{X}_1} = \overline{\mathsf{X}_2}$
3. X_1 and X_2 are symmetrically distributed about their means

What can you conclude about the shapes of the distributions of these data sets?

THE PROBLEM WITH "VARIANCE"

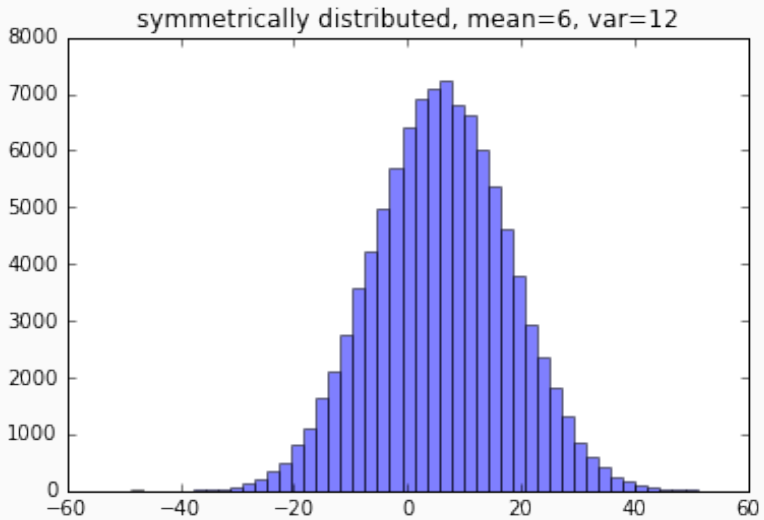
What can you conclude about the shapes of the distributions of these data sets?

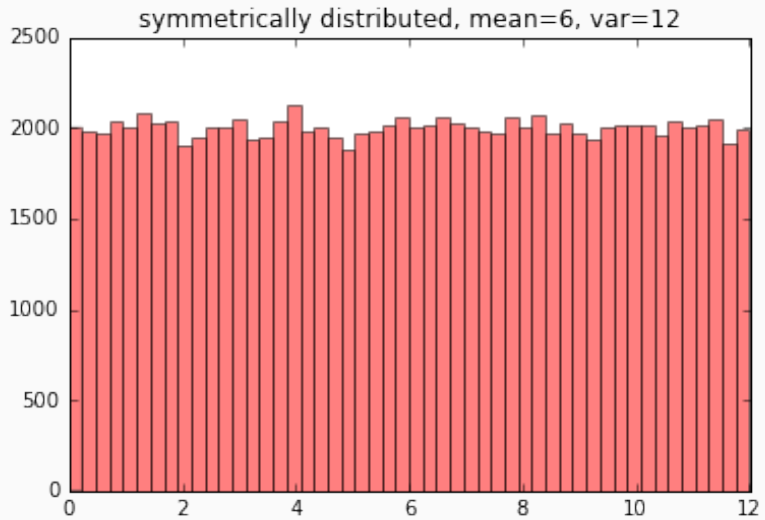




THE PROBLEM WITH "VARIANCE"

What can you conclude about the shapes of the distributions of these data sets?





Observation: Variance isn't the final word on the "spread" of the data.

THE PROBLEM WITH MODEL "GOODNESS"

We began Lecture 9 with the question of which notions of the "goodness" we should choose to evaluate a model:

1. Only look at the largest error the model makes on a data set $\mathbf{X}, \mathbf{Y}$
2. Take the sum of error the model makes on a data set $\mathbf{X}, \mathbf{Y}$
3. Take the sum of squared errors the model makes on a data set $\mathbf{X}, \mathbf{Y}$
4. Compute R^2

Question: Which notion or overall "goodness" is generally better? Is there a notion that is generally better?

THE PROBLEM WITH MODEL "GOODNESS"

When we built *probabilistic models* for the observed data (linear mode plus i.i.d. normal noise), we felt compelled to minimize sum of squared errors (MLE).

When we built *probabilistic models* for the observed data and prior beliefs about the model, we found it obvious to minimize the ridge regression loss function (MAP).

Question: Does this mean that we didn't make any choices in our modeling process?

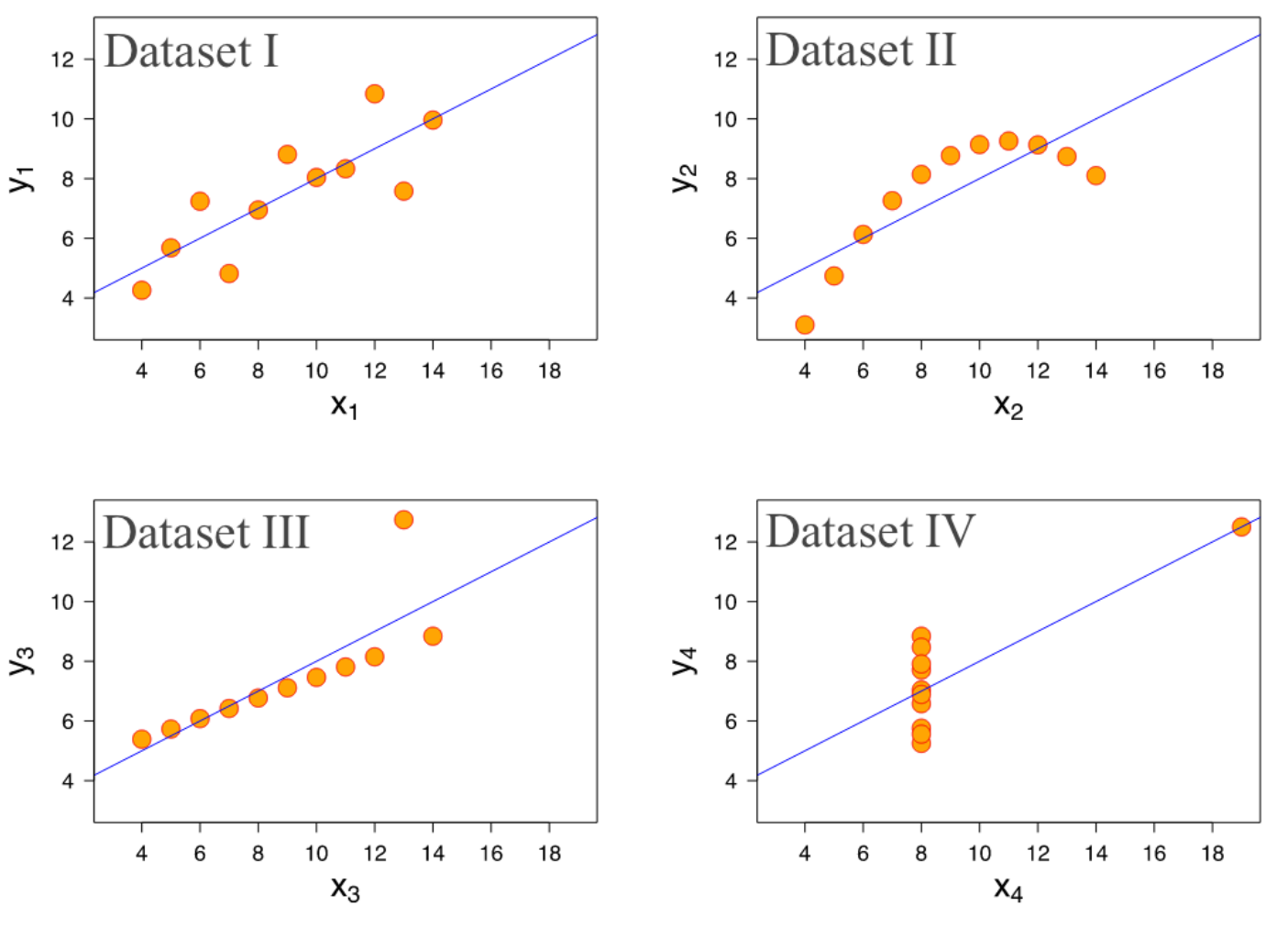
THE PROBLEM WITH R^2

The following data sets comprise the Anscombe's Quartet; which model fits the data the best?

	Dataset I		Dataset II		Dataset III		Dataset IV	
	x	y	x	y	x	y	x	y
	10	8.04	10	9.14	10	7.46	8	6.58
	8	6.95	8	8.14	8	6.77	8	5.76
	13	7.58	13	8.74	13	12.74	8	7.71
	9	8.81	9	8.77	9	7.11	8	8.84
	11	8.33	11	9.26	11	7.81	8	8.47
	14	9.96	14	8.1	14	8.84	8	7.04
	6	7.24	6	6.13	6	6.08	8	5.25
	4	4.26	4	3.1	4	5.39	19	12.5
	12	10.84	12	9.13	12	8.15	8	5.56
	7	4.82	7	7.26	7	6.42	8	7.91
	5	5.68	5	4.74	5	5.73	8	6.89
Sum:	99.00	82.51	99.00	82.51	99.00	82.51	99.00	82.51
Avg:	9.00	7.50	9.00	7.50	9.00	7.50	9.00	7.50
Std:	3.32	2.03	3.32	2.03	3.32	2.03	3.32	2.03
Reg. Line:	$y = 3 + 0.5x$		$y = 3 + 0.5x$		$y = 3 + 0.5x$		$y = 3 + 0.5x$	
R^2 :		0.816		0.816		0.816		0.816

THE PROBLEM WITH R^2

The following data sets comprise the Anscombe's Quartet; which model fits the data the best?



THE PROBLEM WITH R^2

In Challenge Problem HW4, you were asked to model a set of data on potential donors. Most of you chose to fit a linear model (OLS or ridge regression).

THE PROBLEM WITH R^2

In Challenge Problem HW4, you were asked to model a set of data on potential donors. Most of you chose to fit a linear model (OLS or ridge regression).

1. Your model had an R^2 close to zero on the testing set. What does this mean?
2. Using your model to spam only to donors predicted by the model to donate more that \$7, you made way more money than blanket mailing everyone.

Question: Is your model a good model or a bad model?

THE PROBLEM WITH R^2

In Challenge Problem HW4, you were asked to model a set of data on potential donors. Most of you chose to fit a linear model (OLS or ridge regression).

1. Your model had an R^2 close to zero on the testing set. What does this mean?
2. Using your model to spam only to donors predicted by the model to donate more that $7, you made way more money than blanket mailing everyone.

Question: Is your model a good model or a bad model?

Better Question: **What** is your model **good for**?

If your goal is to predict the exact amount each donor will donate, then your model is bad. If your goal is to maximize your net gain at the end of the day, then it's **good enough**.

THE PROBLEM WITH "SIGNIFICANCE"

Question: What's a p-value again?

The p-value for the value of particular stat $T = t$ is the probability of observing a "similarly extreme" value $T = t$, assuming some (null) model, M_0:

$$p(T \geq t | M_0) \text{ or } p(T \leq t | M_0) \text{ or combo}$$

If the p-value is small, we tend to think that observing a value "similar to" $T = t$ is unlikely assuming the model M_0, and consider this as evidence for **rejecting** the null model.

THE PROBLEM WITH "SIGNIFICANCE"

Question: Why use 0.05 or 0.01 as the cut off for p-values?

For decades, the convention have been to interpret $p < .05$ as "significant", and $p < .01$ as "highly significant.

But when p-values where introduced by Sir Ronald Fisher in 1925, he adopted .05 as a **reference point** for rejecting a null hypothesis. But 0.5 was not a sharp cutoff and should be considered in the context of other results and tests.

THE PROBLEM WITH "SIGNIFICANCE"

Question: Do significant p-values indicate the presence of an effect (a real relationship between predictor and response)?

Example

Since the p-value variable selection method **only gathers evidence against the null hypothesis**, the more tests you do, the higher the likelihood of falsely rejecting the null hypothesis.

When you have a large number of predictors, performing step-wise selection using p-values may result in predictor subsets that are not truly significant.

Question: Do significant p-values indicate the significance of an effect (does a "statistically" significant effect indicate a meaningful relationship)?

Example

Our regression model is:

$$\textbf{Price of House} \text{ (in \$)} = \underbrace{10,000}_{\beta_2} * \textbf{Area} \text{ (in SqFt)} + \underbrace{0.00001}_{\beta_1} * \textbf{Number of Blue Bathroom Tiles} + \underbrace{10,000}_{\beta_0}$$

The p-values for the coefficients are:

	β_2	β_1	β_0
p-value	0.0005	0.001	0.0001

Which predictors is are significant?

THE PROBLEM WITH "SIGNIFICANCE"

"Ultimately the problem is not with *p*-values but with null-hypothesis significance testing, that parody of falsificationism in which straw-man null hypothesis A is rejected and this is taken as evidence in favor of preferred alternative B.

Whenever this sort of reasoning is being done, the problems discussed above will arise. Confidence intervals, credible intervals, Bayes factors, cross-validation: you name the method, it can and will be twisted, even if inadvertently, to create the appearance of strong evidence where none exists."

- Andrew Gelman

(comments on the ASA statement regard the misuse of *p*-values)

THE PROBLEM WITH CHOICE

The problem problem with commanding an arsenal of sophisticated tools and theory is that it's easy to make choices based on habit, convenience, or intimidation/glamour factor. Or you're overwhelmed and are unable to make a choice.

Tips for staying grounded:

1. Do what works (for the task, for the context of the problem at hand, for time/resources available)
2. Be absolutely honest and vigilant about our assumptions and choices (justify them and identify their draw backs)
3. Be accountable to the data, to the problem (not to our favorite technique or pet theory)

TALK OUTLINE

Last Time in CS109A...

Estimating Model Parameters from the Posterior

Summary

Loose Ends and Lingering Questions

Being a Responsible Data Scientist

Hint for Midterm

HINT FOR MIDTERM

On the midterm, often when we ask you to compute certain quantities, you can either use a formula or you can use bootstrapping to simulate multiple experiments (just like we often do in lecture, lab and HW):

1. **Confidence interval:** We can use a formula or we can estimate the coefficients a bunch of times and take 95% of the most common values.
2. **Predictive interval:** We can use a formula or we can can predict the response a bunch of times and take 95% of the most common values.