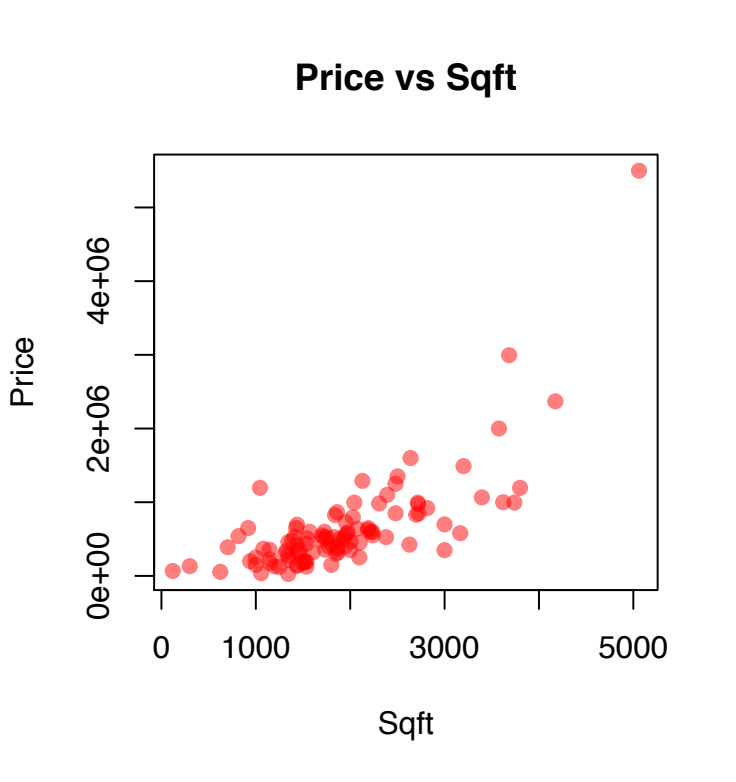
PROBABILISTIC MODELS FOR INFERENCE

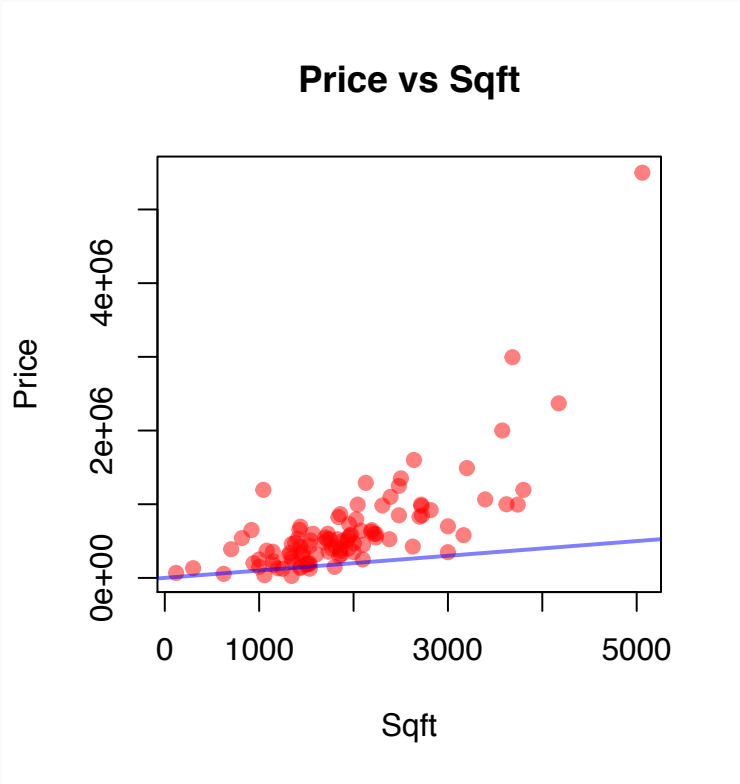
La Serena School For Data Science
Pavlos Protopapas

LINEAR REGRESSION REVISITED



This is a scatter plot of home prices vs square footage of some homes in southern California.

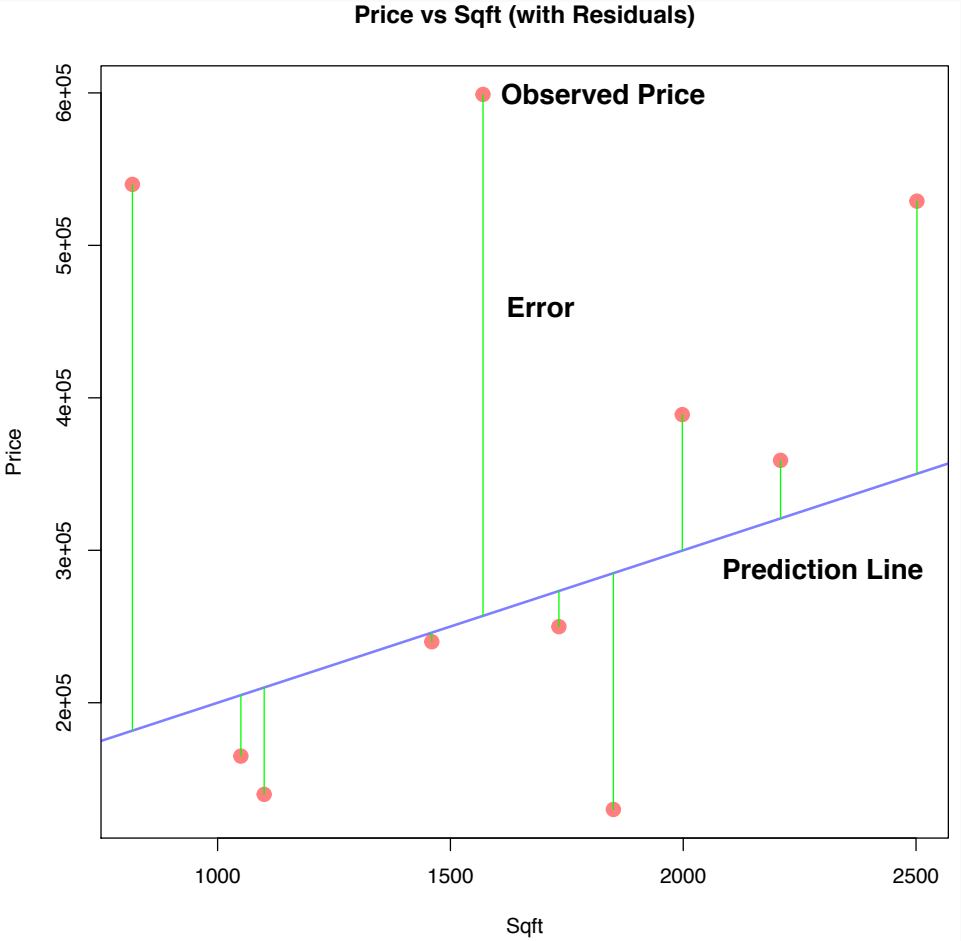
LINEAR REGRESSION REVISITED



This is a scatter plot of home prices vs square footage of some homes in southern California.

Say we want to model the data with a linear function. How do we measure how good is our model?

NOTIONS OF ERROR



An ***absolute residual*** is the absolute difference between the actual price of a home and the price predicted by the line for a given square footage.

$$\text{Res}_i = |\text{Observed}_i - \text{Predicted}_i|$$

The i-th absolute residual measures the magnitude of the "error" made by the i-th prediction.

Recall: Fitting a (linear) model means finding parameters that minimizes a choice of loss function.

1. (**Max absolute deviation**) Count only the biggest "error"

$$\text{Loss Function} = \max_i |\text{Observed}_i - \text{Predicted}_i|$$

2. (**Sum of absolute deviations**) Add up all the "errors"

$$\text{Loss Function} = \sum_i |\text{Observed}_i - \text{Predicted}_i|$$

   We can also average them.

3. (**Sum of squared errors**) Add up the squares of the "errors"

$$\text{Loss Function} = \sum_i |\text{Observed}_i - \text{Predicted}_i|^2$$

Question: Which loss function should we choose?

NOTIONS OF FITNESS

Which loss function we choose to minimize depends on ***how***, we believe, the “residual” (difference between observed and predicted values) arise.

LECTURE OUTLINE

Probabilistic Model for Linear Regression

Introduction to Bayesian Inference

Summary

Loose Ends and Lingering Questions

TALK OUTLINE

Probabilistic Model for Linear Regression

Introduction to Bayesian Inference

Summary

Loose Ends and Lingering Questions

THE STOCHASTIC MODEL FOR 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***.

$$y = \underbrace{\beta_1 \cdot x + \beta_0}_{\text{theoretical price}} + \underbrace{\epsilon}_{\text{noise}}$$

In class, we have believed that the ϵ is a *random variable* which is normally distributed

$$\epsilon \sim \mathcal{N}(0, \sigma^2)$$

Question: What can we deduce about y given our model for price and our assumption about ϵ?

Based on our model

$$y = \beta_1 \cdot x + \beta_0 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2)$$

We see that:

1. y is a random variable.
2. If we are **given** fixed values for β_1, β_0, x, then the **corresponding** price y is a *random variable*, $y|\beta_1, \beta_0, x$, whose is determined by the distribution of ϵ,

$$y|\beta_1, \beta_0, x \quad \sim \quad ?$$

THE RANDOM VARIABLE $y|\beta_1, \beta_0, x$

Let's fix $x = 500$ sqft, $\beta_1 = 2$, $\beta_0 = 0.5$ mil, $\sigma^2 = 1$.

Let's sample ϵ from $\mathcal{N}(0, \sigma^2)$, and guess at the distribution of $y|\beta_1, \beta_0, x$.

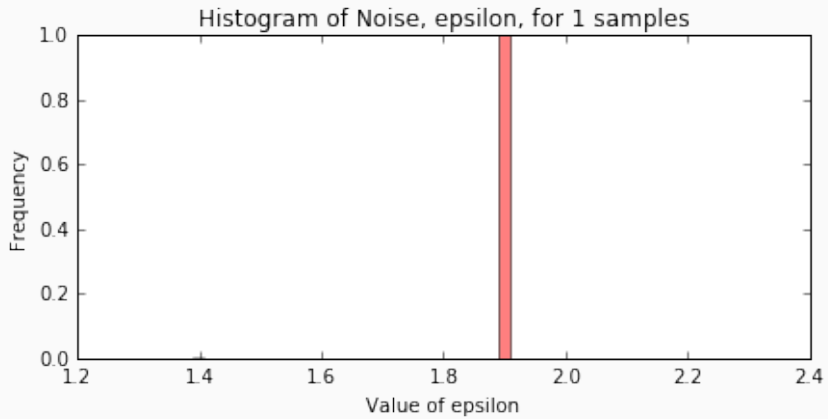
That is, we make histograms of the following table:

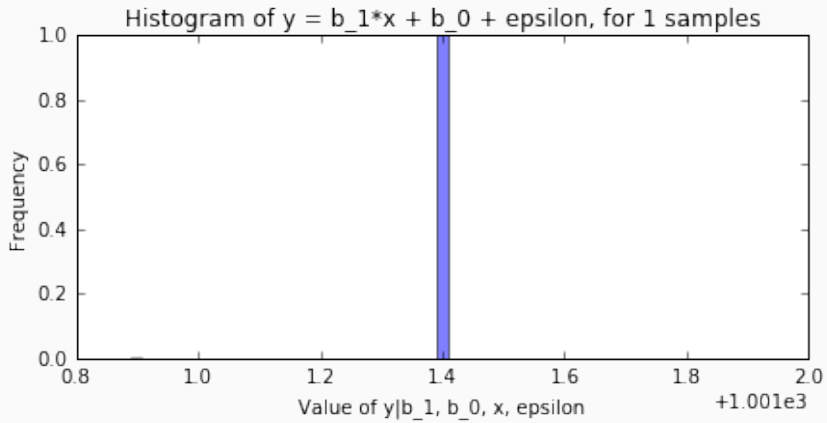
ϵ	$y = \beta_1 x + \beta_0 + \epsilon$
0	$\beta_1 x + \beta_0 = 2 * 500 + 0.5$
0.1	$\beta_1 x + \beta_0 + 0.1 = 2 * 500 + 0.5 + 0.1$
-0.024	$\beta_1 x + \beta_0 - 0.024 = 2 * 500 + 0.5 - 0.024$
...	...

THE RANDOM VARIABLE $y|\beta_1, \beta_0, x$

Let's fix $x = 500$ sqft, $\beta_1 = 2$, $\beta_0 = 0.5$ mil, $\sigma^2 = 1$.

Let's sample ϵ from $\mathcal{N}(0, \sigma^2)$, and guess at the distribution of $y|\beta_1, \beta_0, x$.

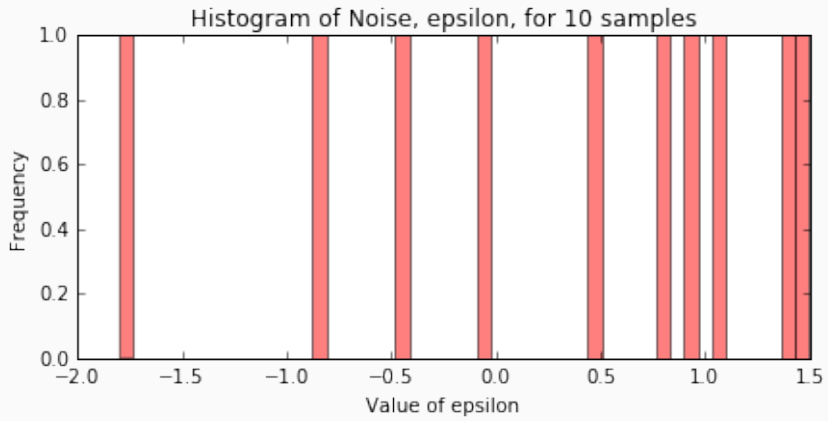


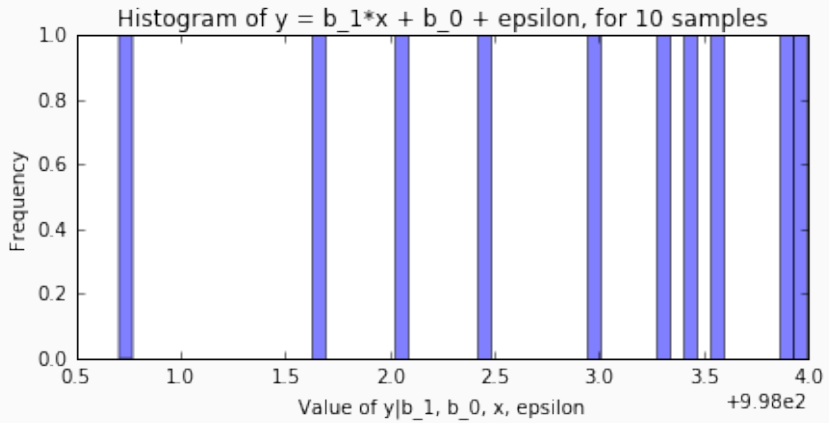


THE RANDOM VARIABLE $y|\beta_1, \beta_0, x$

Let's fix $x = 500$ sqft, $\beta_1 = 2$, $\beta_0 = 0.5$ mil, $\sigma^2 = 1$.

Let's sample ϵ from $\mathcal{N}(0, \sigma^2)$, and guess at the distribution of $y|\beta_1, \beta_0, x$.

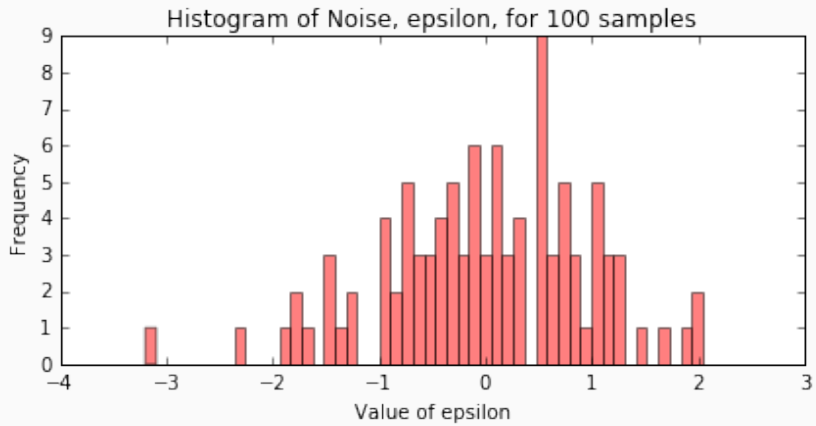


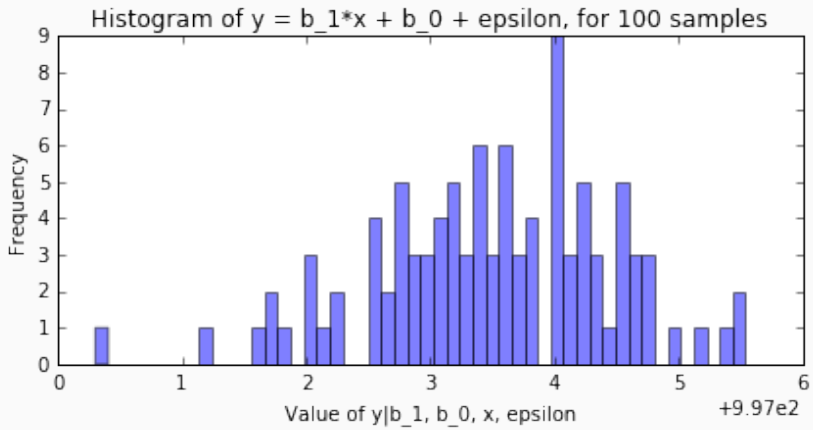


THE RANDOM VARIABLE $y|\beta_1, \beta_0, x$

Let's fix $x = 500$ sqft, $\beta_1 = 2$, $\beta_0 = 0.5$ mil, $\sigma^2 = 1$.

Let's sample ϵ from $\mathcal{N}(0, \sigma^2)$, and guess at the distribution of $y|\beta_1, \beta_0, x$.

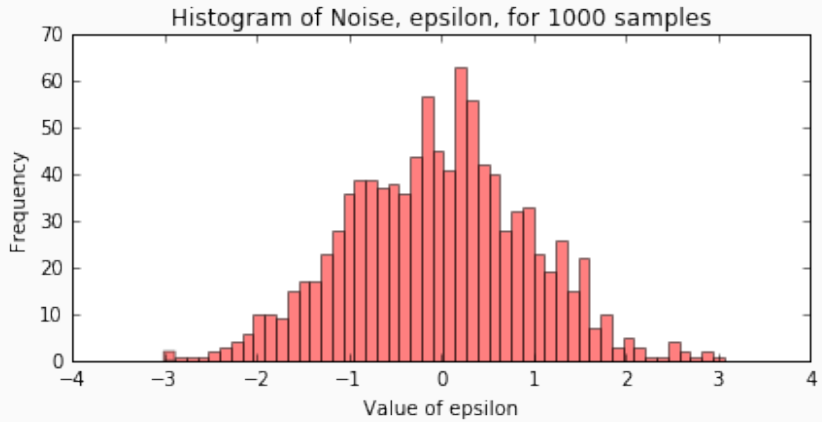


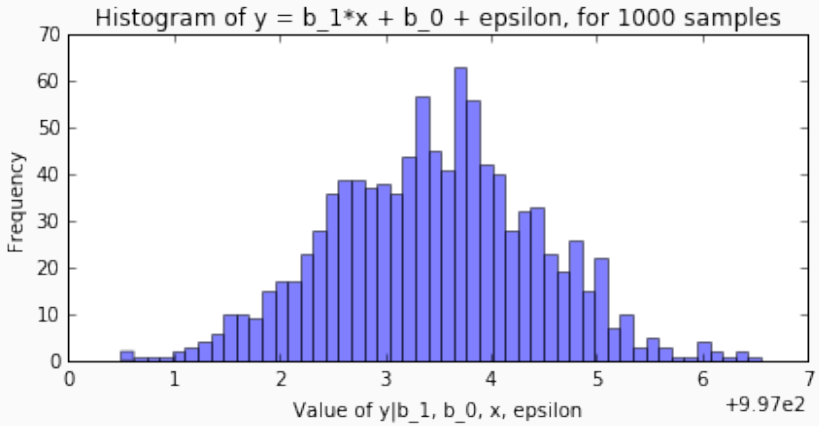


THE RANDOM VARIABLE $y|\beta_1, \beta_0, x$

Let's fix $x = 500$ sqft, $\beta_1 = 2$, $\beta_0 = 0.5$ mil, $\sigma^2 = 1$.

Let's sample ϵ from $\mathcal{N}(0, \sigma^2)$, and guess at the distribution of $y|\beta_1, \beta_0, x$.

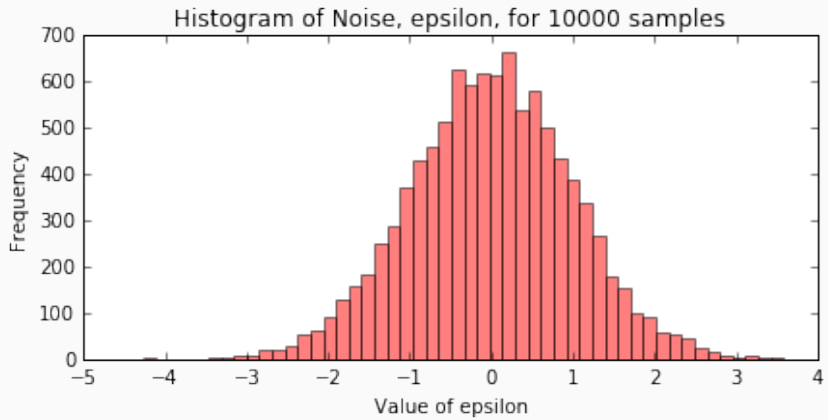


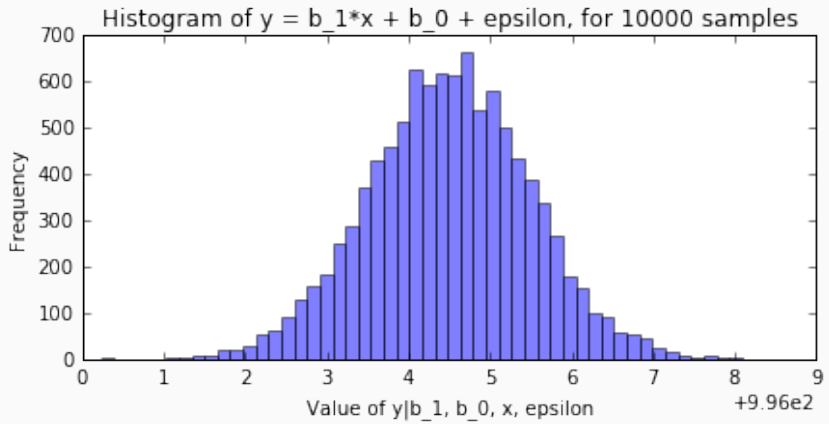


THE RANDOM VARIABLE $y|\beta_1, \beta_0, x$

Let's fix $x = 500$ sqft, $\beta_1 = 2$, $\beta_0 = 0.5$ mil, $\sigma^2 = 1$.

Let's sample ϵ from $\mathcal{N}(0, \sigma^2)$, and guess at the distribution of $y|\beta_1, \beta_0, x$.





THE RANDOM VARIABLE $y|\beta_1, \beta_0, x$

It looks like, if $\epsilon \sim \mathcal{N}(0, \sigma^2)$, then $y|\beta_1, \beta_0, x$ is normally distributed with mean $\beta_1 x + \beta_0$ and variance σ^2, i.e.

$$y|\beta_1, \beta_0, x \sim \mathcal{N}(\beta_1 x + \beta_0, \sigma^2)$$

Sanity check: Check your understanding of the normal distribution. Does this conclusion make intuitive, common sense?

THE RANDOM VARIABLE $y|\beta_1, \beta_0, x$

It was simple to guess the distribution of $y|\beta_1, \beta_0, x$ for one fixed value of x.

But in a real data set, we have multiple values of x. How do we consider the distribution of *all* the y values at once?

$$\left\{ \quad y_i|\beta_1, \beta_0, x_i \quad \right\}_{i=1}^{N}$$

Say our data set consists of $\mathbf{X} = \{x_1, \ldots, x_N\}$ and $\mathbf{Y} = \{y_1, \ldots, y_N\}$, and $\{\epsilon_1, \ldots, \epsilon_N\}$ are the corresponding noise variables.

If we **assume that the noise is identically distributed**, i.e. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, then each $y_i|\beta_1, \beta_0, x_i$ is normally distributed with the same variance,

$$y_i|\beta_1, \beta_0, x_i \sim \mathcal{N}(\beta_1 x_i + \beta_0, \sigma^2).$$

We denote the pdf by $p(y_i|\beta_1, \beta_0, x_i)$

THE LIKELIHOOD FUNCTION

Say our data set consists of $\mathbf{X} = \{x_1, \ldots, x_N\}$ and $\mathbf{Y} = \{y_1, \ldots, y_N\}$, and $\{\epsilon_1, \ldots, \epsilon_N\}$ are the corresponding noise variables.

If we **assume that the ϵ_i's are independent**, then so are the $y_i|\beta_1, \beta_0, x_i$'s.

Thus, the ***joint*** probability of all the y-values (given the x-values, noise and model parameters) is product of the pdf of each $y_i|\beta_1, \beta_0, x_i$:

$$L(\beta_1, \beta_0) = p(\mathbf{Y}|\beta_1, \beta_0, \mathbf{X}) = \prod_{i=1}^{N} p(y_i|\beta_1, \beta_0, x_i) = \prod_{i=1}^{N} \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2)$$

$L(\beta_1, \beta_0)$ is called the ***likelihood function***.

Note that since $\mathbf{X}$ is always given, L is just a function of the model parameters!

MODEL SELECTION USING THE LIKELIHOOD

Question: What exactly does the likelihood function mean? And what is it good for?

MODEL SELECTION USING THE LIKELIHOOD

Example:

Say we're considering two linear models for the data $\{(1,2),(2,3)\}$, with identically independently distributed (i.i.d.) noise, $\epsilon_i \sim \mathcal{N}(0,1)$.

- $M_1 : y = x + 2$
- $M_2 : y = 2 - 2x$

Which model is the most appropriate for the data, assuming one is correct?

Example:

Say we're considering two linear models for the data $\{(1,2),(2,3)\}$, with identically independently distributed (i.i.d.) noise, $\epsilon_i \sim \mathcal{N}(0,1)$.

- $M_1 : y = x + 2$

  $$L(\beta_1 = 1, \beta_0 = 2) = \underbrace{\mathcal{N}(y_1; x_1 + 2, 1)}_{p(y_1|x_1,\beta_1,\beta_0)} \underbrace{\mathcal{N}(y_2; x_2 + 2, 1)}_{p(y_2|x_2,\beta_1,\beta_0)}$$

- $M_2 : y = 2 - 2x$

  $$L(\beta_1 = -2, \beta_0 = 2) = \underbrace{\mathcal{N}(y_1; 2 - 2x_1, 1)}_{p(y_1|x_1,\beta_1,\beta_0)} \underbrace{\mathcal{N}(y_2; 2 - 2x_2, 1)}_{p(y_2|x_2,\beta_1,\beta_0)}$$

Which model is the most appropriate for the data, assuming one is correct?

MODEL SELECTION USING THE LIKELIHOOD

Example:

In [38]:
```
#Data
x_1 = 1
x_2 = 2

y_1 = 2
y_2 = 3
```

Example:

Model 1: $L(\beta_1 = 1, \beta_0 = 2) = \mathcal{N}(y_1; x_1 + 2, 1)\mathcal{N}(y_2; x_2 + 2, 1) = 0.0585498$

In [114]:
```
# likelihood calculation for b_0 = 2, b_1 = 1
lkhd_1 = scipy.stats.norm(x_1 + 2, 1)
lkhd_2 = scipy.stats.norm(x_2 + 2, 1)

likelihood = lkhd_1.pdf(y_1) * lkhd_2.pdf(y_2)
print round(likelihood, 7)
```

```
0.0585498
```

MODEL SELECTION USING THE LIKELIHOOD

Example:

Model 1: $L(\beta_1 = 1, \beta_0 = 2) = \mathcal{N}(y_1; x_1 + 2, 1)\mathcal{N}(y_2; x_2 + 2, 1) = 0.0585498$

Model 2: $L(\beta_1 = -2, \beta_0 = 2) = \mathcal{N}(y_1; 2 - 2x_1, 1)\mathcal{N}(y_2; 2 - 2x_2, 1) = 0.0000001$

```
In [118]: # likelihood calculation for b_0 = 2, b_1 = -2
          lkhd_1 = scipy.stats.norm(2 - 2 * x_1, 1)
          lkhd_2 = scipy.stats.norm(2 - 2 * x_2, 1)

          likelihood = lkhd_1.pdf(y_1) * lkhd_2.pdf(y_2)
          print round(likelihood, 7)

          1e-07
```

Example:

Model 1: $L(\beta_1 = 1, \beta_0 = 2) = \mathcal{N}(y_1; x_1 + 2, 1)\mathcal{N}(y_2; x_2 + 2, 1) = 0.0585498$

Model 2: $L(\beta_1 = -2, \beta_0 = 2) = \mathcal{N}(y_1; 2 - 2x_1, 1)\mathcal{N}(y_2; 2 - 2x_2, 1) = 0.0000001$

Analysis: We are 10,000 times more likely to observe our data under Model 1 than under Model 2.

This means that, if Model 2 is correct, by observing the data $\{(1, 2), (2, 3)\}$, we were 10,000 times luckier than we needed to be if Model 2 is correct.

I just don't believe in luck!

MODEL SELECTION USING THE LIKELIHOOD

Example:

Model 1: $L(\beta_1 = 1, \beta_0 = 2) = \mathcal{N}(y_1; x_1 + 2, 1)\mathcal{N}(y_2; x_2 + 2, 1) = 0.0585498$

Model 2: $L(\beta_1 = -2, \beta_0 = 2) = \mathcal{N}(y_1; 2 - 2x_1, 1)\mathcal{N}(y_2; 2 - 2x_2, 1) = 0.0000001$

Conclusion: We should always select the model that makes observing our data maximally probable (least like a coincidence). I.e. we want to select parameters β_1, β_0 to **maximize the likelihood function**.

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

Let's make $L(\beta_0, \beta_1) = \prod_{i=1}^{N} \overbrace{\mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2)}^{p(y_i|\beta_0,\beta_1,x_i)}$ more friendly, by turning the multiplication into addition:

$$\log \text{likelihood} = \ln L(\beta_0, \beta_1) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2)$$

Since the ln function is monotone increasing on $\mathbb{R}$, **maximizing likelihood is equivalent to maximizing log likelihood**.

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

Let's expand the log likelihood function a bit (use C, K to rep constants that depend on σ^2):

$$\ln L(\beta_1, \beta_0) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2) \tag{1}$$

$$\tag{2}$$

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

Let's expand the log likelihood function a bit (use C, K to rep constants that depend on σ^2):

$$\ln L(\beta_1, \beta_0) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2) \tag{1}$$

$$\tag{2}$$

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

Let's expand the log likelihood function a bit (use C, K to rep constants that depend on σ^2):

$$\ln L(\beta_1, \beta_0) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2) \tag{1}$$

$$= \sum_{i=1}^{N} \ln \left[C * \exp \left\{ -\frac{(y_i - \beta_1 x_i - \beta_0)^2}{K} \right\} \right] \tag{2}$$

$$\tag{3}$$

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

Let's expand the log likelihood function a bit (use C, K to rep constants that depend on σ^2):

$$\ln L(\beta_1, \beta_0) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2) \tag{1}$$

$$= \sum_{i=1}^{N} \ln \left[C * \exp \left\{ -\frac{(y_i - \beta_1 x_i - \beta_0)^2}{K} \right\} \right] \tag{2}$$

(3)

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

Let's expand the log likelihood function a bit (use C, K to rep constants that depend on σ^2):

$$\ln L(\beta_1, \beta_0) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2) \tag{1}$$

$$= \sum_{i=1}^{N} \ln \left[C * \exp \left\{ -\frac{(y_i - \beta_1 x_i - \beta_0)^2}{K} \right\} \right] \tag{2}$$

$$= \sum_{i=1}^{N} \left[\ln C - \frac{(y_i - \beta_1 x_i - \beta_0)^2}{K} \right] \tag{3}$$

$$\tag{4}$$

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

Let's expand the log likelihood function a bit (use C, K to rep constants that depend on σ^2):

$$\ln L(\beta_1, \beta_0) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2) \tag{1}$$

$$= \sum_{i=1}^{N} \ln \left[C * \exp \left\{ -\frac{(y_i - \beta_1 x_i - \beta_0)^2}{K} \right\} \right] \tag{2}$$

$$= \sum_{i=1}^{N} \left[\ln C - \frac{(y_i - \beta_1 x_i - \beta_0)^2}{K} \right] \tag{3}$$

$$\tag{4}$$

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

Let's expand the log likelihood function a bit (use C, K to rep constants that depend on σ^2):

$$\ln L(\beta_1, \beta_0) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2) \tag{1}$$

$$= \sum_{i=1}^{N} \ln \left[C * \exp \left\{ -\frac{(y_i - \beta_1 x_i - \beta_0)^2}{K} \right\} \right] \tag{2}$$

$$= \sum_{i=1}^{N} \left[\ln C - \frac{(y_i - \beta_1 x_i - \beta_0)^2}{K} \right] \tag{3}$$

$$= \sum_{i=1}^{N} \ln C - \frac{1}{K} \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2 \tag{4}$$

$$\tag{5}$$

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

Let's expand the log likelihood function a bit (use C, K to rep constants that depend on σ^2):

$$\ln L(\beta_1, \beta_0) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2) \tag{1}$$

$$= \sum_{i=1}^{N} \ln C - \frac{1}{K} \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2 \tag{2}$$

Since $\sum_{i=1}^{N} \ln C$ and $\frac{1}{K}$ are constants,

$$\max L(\beta_1, \beta_0) \Leftrightarrow \max \ln L(\beta_1, \beta_0) \Leftrightarrow \min \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2$$

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

Let's expand the log likelihood function a bit (use C, K to rep constants that depend on σ^2):

$$\ln L(\beta_1, \beta_0) = \sum_{i=1}^{N} \ln \mathcal{N}(y_i; \beta_1 x_i + \beta_0, \sigma^2) \tag{1}$$

$$= \sum_{i=1}^{N} \ln C - \frac{1}{K} \sum_{i=1}^{N} (y_i - \beta_1 x_i - \beta_0)^2 \tag{2}$$

Since $\sum_{i=1}^{N} \ln C$ and $\frac{1}{K}$ are constants,

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

THE MAXIMUM LIKELIHOOD ESTIMATE

Goal: Find values for β_0, β_1 so that the likelihood of the data $L(\beta_0, \beta_1)$ is maximized.

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 β_0^{MLE}, β_1^{MLE}.

Sanity check: Why do I care again? Why am I so excited about the equivalence of MLE and OLS?

AN EXAMPLE WITH USING THE MLE

Consider the following two sets of data and two MLE models. Do you think the models are good? Why or why not?

AN EXAMPLE WITH USING THE MLE

Consider the following two sets of data and two MLE models. Do you think the models are good? Why or why not?

In both cases, you rejected the model based on criteria (prior beliefs) that you never made explicit in the modeling process!

TALK OUTLINE

Probabilistic Model for Linear Regression

Introduction to Bayesian Inference

Summary

Loose Ends and Lingering Questions

WHAT IS A PRIOR?

A ***prior distribution*** is distribution in terms of the model parameters that encode your beliefs about the parameters (before even looking at the data).

Rather than treating parameters as underlying fixed constants that we are learning, we treat parameters like *random variables* with distributions expressing our uncertainty about them.

WHAT IS A PRIOR?

Simple Linear Regression

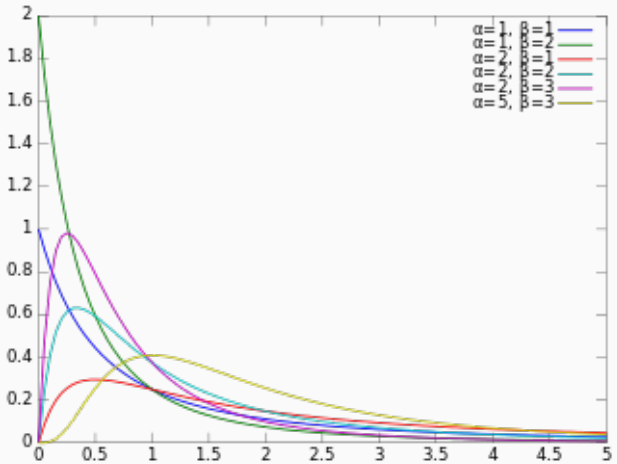
When we're modeling the housing prices data, with $y = \beta_1 x + \beta_0$,

- 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

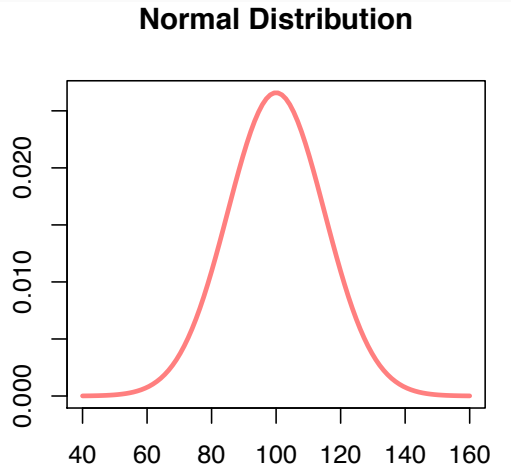
Simple Linear Regression

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



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

Now that we know how to encode our prior beliefs about the model parameters using prior distributions, how do we incorporate them into our model?

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

or **alternatively**,

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

THE POSTERIOR DISTRIBUTION

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

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

THE POSTERIOR DISTRIBUTION

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

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

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***.

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

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

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

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

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?

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

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

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

(non) conjugate priors

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

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

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.

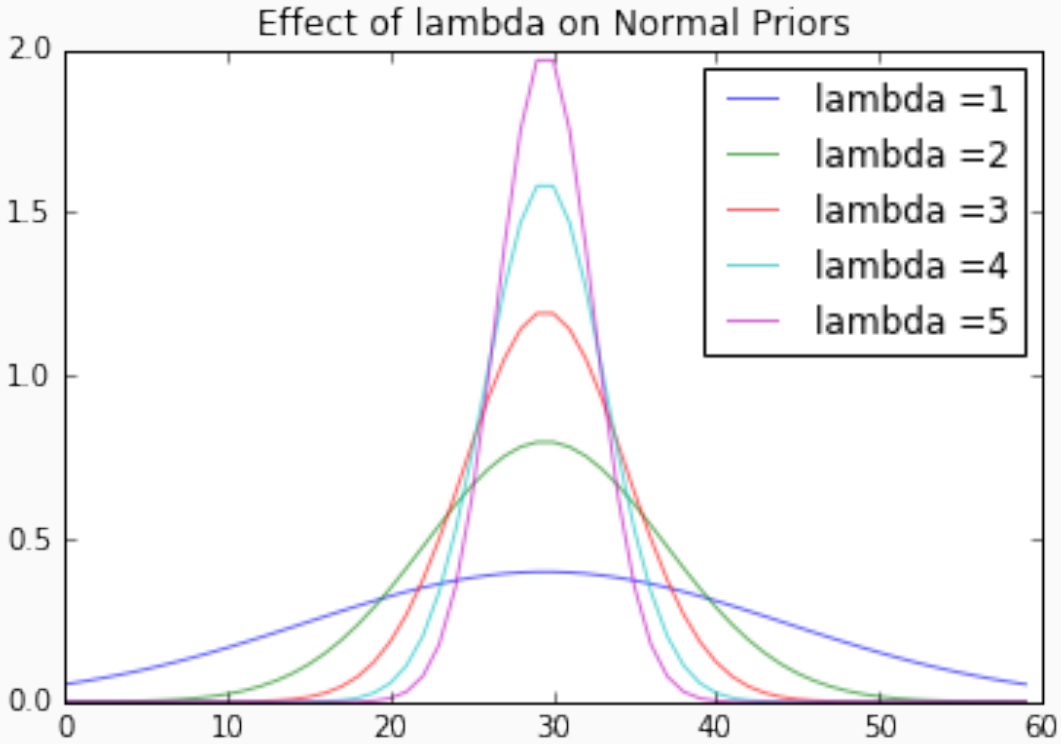
Observation: With normal priors for β_1 and β_0,

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

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?



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:

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

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

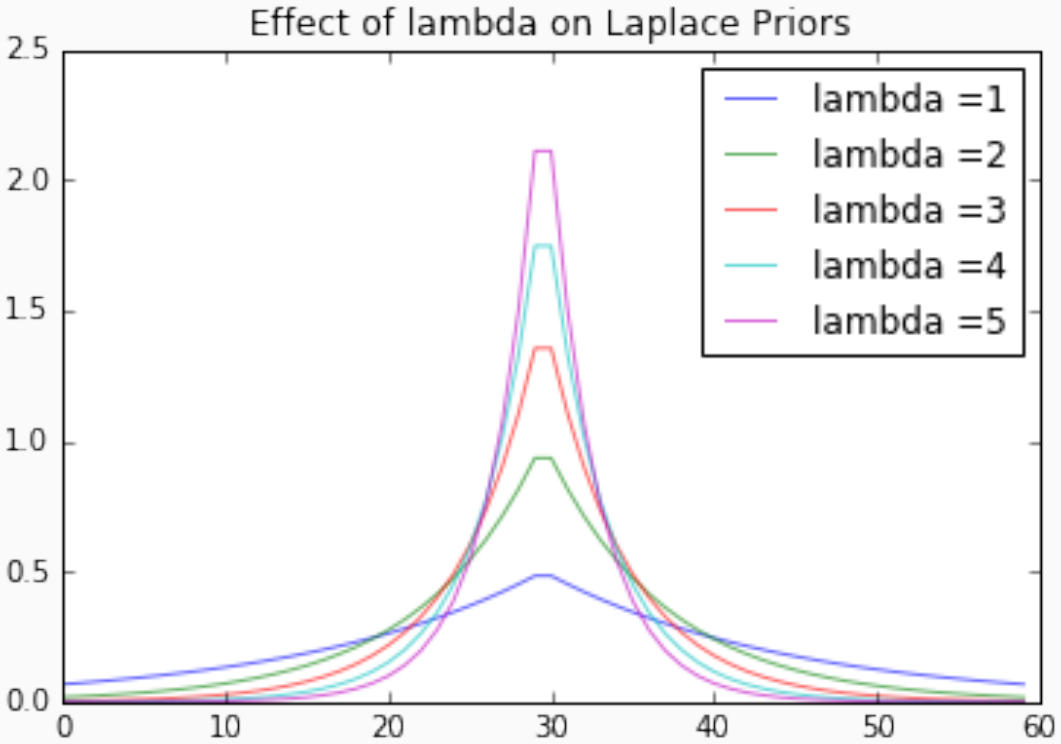
Observation: With Laplace priors for β_1 and β_0,

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

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

Probabilistic Model for Linear Regression

Introduction to Bayesian Inference

Summary

Loose Ends and Lingering Questions

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

Probabilistic Model for Linear Regression

Introduction to Bayesian Inference

Summary

Loose Ends and Lingering Questions

WHAT IS THE POINT OF PROBABILISTIC MODELS?

In the last 5 weeks, we've already covered regression, regularization, model selections in so many different ways.

Question: Why do we need the formalism of probabilistic models?

Question: What do we gain by *reformulating* loss function minimization, regularization, etc, in terms of finding point estimates from probability distributions?

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"?

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”?

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?