Deconstructing Data Science

David Bamman, UC Berkeley

Info 290
Lecture 13: Linear regression

Mar 7, 2016

Regression

A mapping from input data x (drawn from instance space $\mathcal{X}$) to a point y in $\mathbb{R}$

($\mathbb{R}$ = the set of real numbers)

x = the empire state building
y = 17444.5625"

Regression problems

task	x	y
predicting box office revenue	movie	$\mathbb{R}$

TWEETS 508
FOLLOWING 400
FOLLOWERS 799
LIKES 133
LISTS 2
Tweets
Tweets & replies
Photos & videos
David Bamman
@dbamman
Assistant Professor, School of Information, UC Berkeley. Natural language processing, machine learning, computational social science, digital humanities.
Berkeley, CA
people.ischool.berkeley.edu/~dbamman/
Joined October 2009
10 Photos and videos
David Bamman Retweeted
Ted Underwood @Ted_Underwood · 6h
How have the differences between descriptions of men and women in fiction changed over the last 200 yrs? (ICYMI) tedunderwood.com/2016/01/09/the...
8
13
View summary
David Bamman @dbamman · Jan 6
"Figure Eights" (Max Roach/Buddy Rich, 1959) is just dazzling. Probably no video of them anywhere? open.spotify.com/track/23EssvWY...
View summary
David Bamman Retweeted
Anders Søgaard @soegaarducph · Jan 6
@stanfordnlp @brendan642 @jacobeisenstein Here goes: twitter-research.ccs.neu.edu/language/
Enter a term to display: mountain

x = feature vector

Feature	Value
follow clinton	0
follow trump	0
"benghazi"	0
negative sentiment + "benghazi"	0
"illegal immigrants"	0
"republican" in profile	0
"democrat" in profile	0
self-reported location = Berkeley	1

β = coefficients

Feature	β
follow clinton	-3.1
follow trump	6.8
"benghazi"	1.4
negative sentiment + "benghazi"	3.2
"illegal immigrants"	8.7
"republican" in profile	7.9
"democrat" in profile	-3.0
self-reported location = Berkeley	-1.7

Linear regression

$$y = \sum_{i=1}^{F} x_i \beta_i + \varepsilon$$

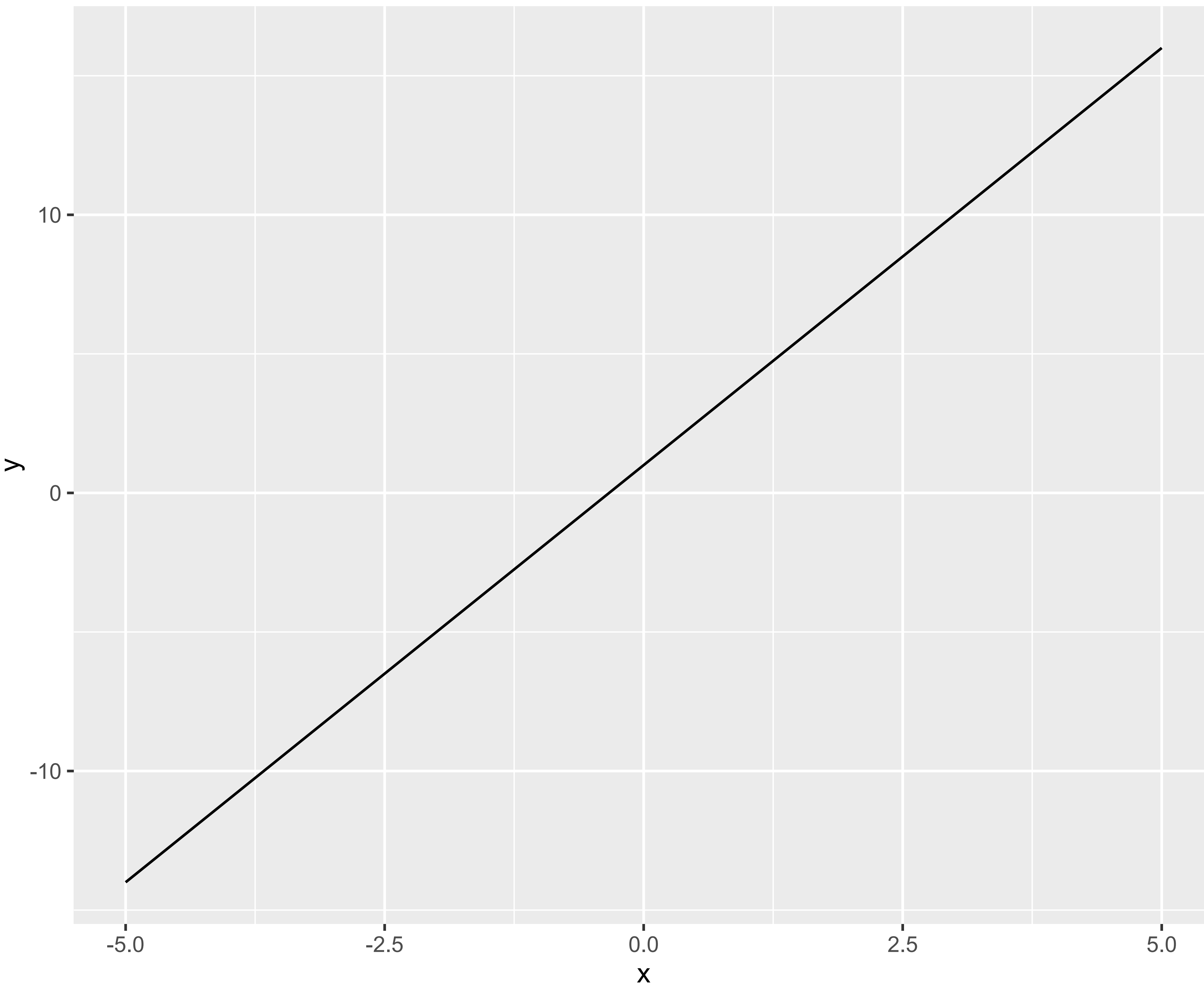
true value y

$$\hat{y} = \sum_{i=1}^{F} x_i \beta_i$$

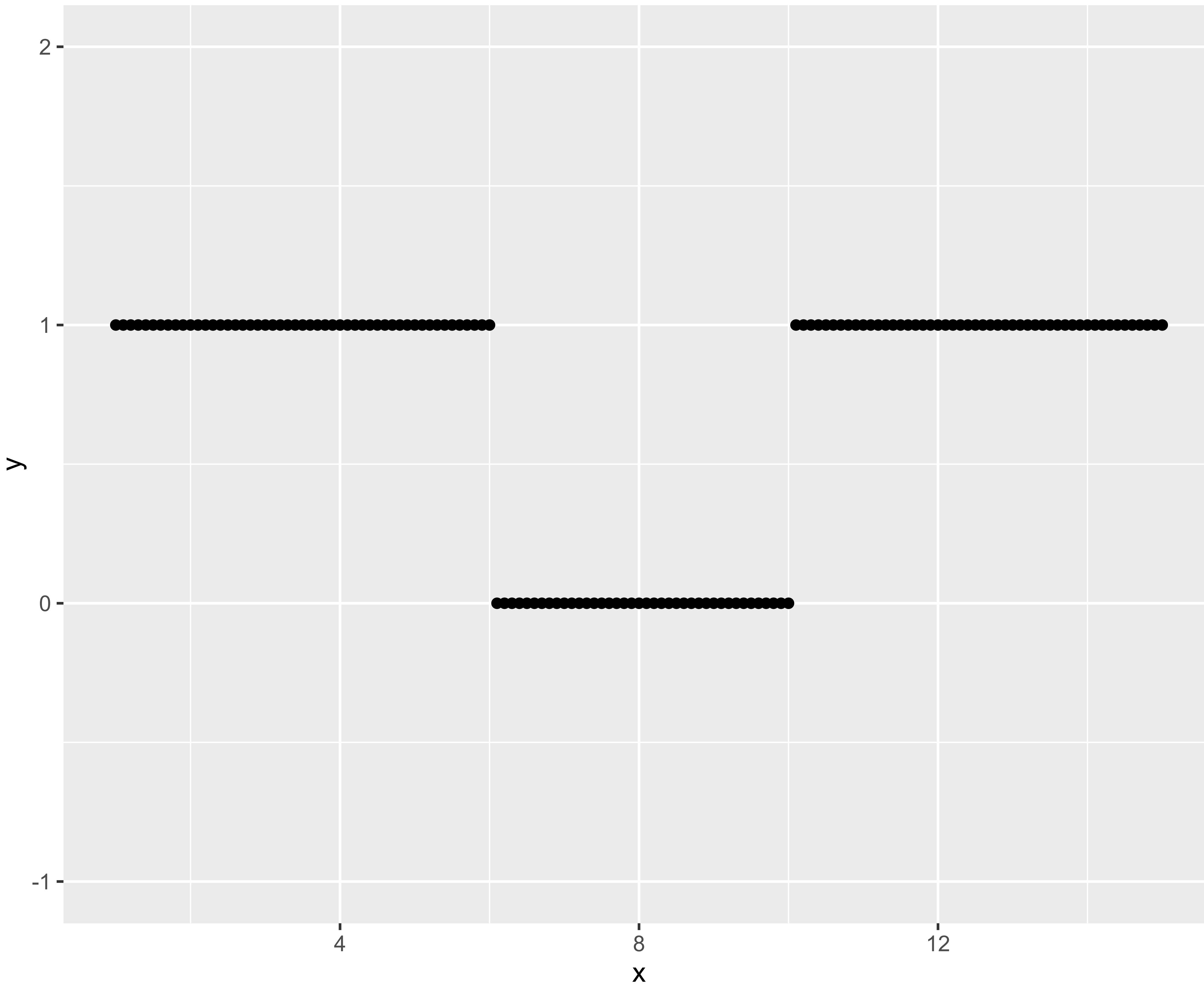
prediction ŷ

$$\varepsilon = y - \hat{y}$$

ε is the difference between the prediction and true value



10
0
-10
y
-5.0
-2.5
0.0
2.5
5.0
x



2
1
0
-1
y
4
8
12
x

How do we get good values for β?

β = coefficients

Feature	β
follow clinton	-3.1
follow trump	6.8
“benghazi”	1.4
negative sentiment + “benghazi”	3.2
“illegal immigrants”	8.7
“republican” in profile	7.9
“democrat” in profile	-3.0
self-reported location = Berkeley	-1.7

Least squares

$$\beta = \min_{\beta} \sum_{i=1}^{N} \varepsilon^2$$

we want to minimize the errors we make

$$\beta = \min_{\beta} \sum_{i=1}^{N} (y - \hat{y})^2$$

$$\beta = \min_{\beta} \sum_{i=1}^{N} \left(y - \sum_{j=1}^{F} x_j \beta_j \right)^2$$

Least squares

$$\beta = \min_{\beta} \sum_{i=1}^{N} \left(y - \sum_{j=1}^{F} x_j \beta_j \right)^2$$

- We can solve this in two ways:
    - Closed form (normal equations)
    - Iteratively (gradient descent)

Algorithm 3 Linear regression stochastic gradient descent

Data: training data $x \in \mathbb{R}^F, y \in \mathbb{R}$
$\beta = 0^F$
while not converged **do**
    for $i = 1$ to N **do**
        $\beta_{t+1} = \beta_t + \alpha \left(y_i - \hat{y} \right) x_i$
    end for
end while

Algorithm 3 Linear regression stochastic gradient descent

Data: training data $x \in \mathbb{R}^F, y \in \mathbb{R}$
$\beta = 0^F$
while not converged **do**
  for $i = 1$ to N **do**
    $\beta_{t+1} = \beta_t + \alpha\,(y_i - \hat{y})\,x_i$
  end for
end while

Algorithm 2 Logistic regression stochastic gradient descent

Data: training data $x \in \mathbb{R}^F, y \in \{0,1\}$
$\beta = 0^F$
while not converged **do**
  for $i = 1$ to N **do**
    $\beta_{t+1} = \beta_t + \alpha\,(y_i - \hat{p}(x_i))\,x_i$
  end for
end while

Code

Many features that show up rarely may likely only appear (by chance) with one label

More generally, may appear so few times that the noise of randomness dominates

β = coefficients

Feature	β
follow clinton	-3.1
follow trump + follow NFL + follow bieber	7299302
“benghazi”	1.4
negative sentiment + “benghazi”	3.2
“illegal immigrants”	8.7
“republican” in profile	7.9
“democrat” in profile	-3.0
self-reported location = Berkeley	-1.7

Ridge regression

$$\beta = \min_{\beta} \underbrace{\sum_{i=1}^{N} (y - \hat{y})^2}_{\text{error}} \underbrace{+\eta \sum_{i=1}^{F} \beta_i^2}_{\text{coefficient size}}$$

We want both of these to be small!

This corresponds to a prior belief that β should be 0

Ridge regression

$$\beta = \min_{\beta} \underbrace{\sum_{i=1}^{N} (y - \hat{y})^2}_{\text{error}} \underbrace{+\eta \sum_{i=1}^{F} \beta_i^2}_{\text{coefficient size}}$$

A.K.A.

L2 regularization
Penalized least squares

low L2

Matt Gerald	$295,619,605
Peter Mensah	$294,475,429
Lewis Abernathy	$188,093,808
Sam Worthington	$186,193,754
CCH Pounder	$184,946,303
…	…
Steve Bacic	-$65,334,914
Jim Ward	-$66,096,435
Karley Scott Collins	-$66,612,154
Dee Bradley Baker	-$73,571,884
Animals	-$110,349,541

BIAS: $5,913,648

med L2

Computer Animation	$68,629,803
Hugo Weaving	$39,769,171
John Ratzenberger	$36,342,438
Tom Cruise	$36,137,757
Tom Hanks	$34,757,574
…	…
Western	-$13,223,795
World cinema	-$13,278,965
Crime Thriller	-$14,138,326
Anime	-$14,750,932
Indie	-$21,081,924

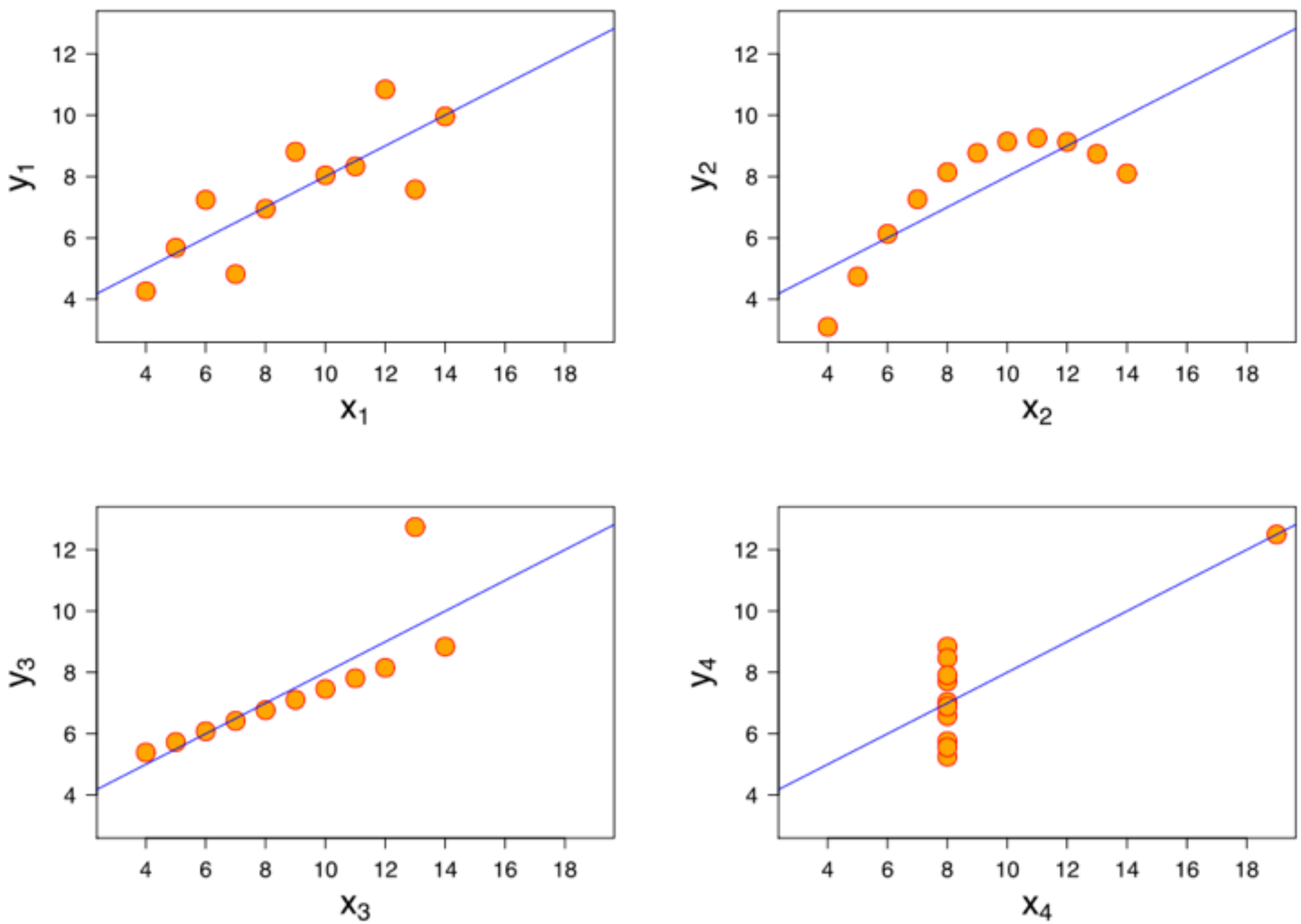
BIAS: $13,394,465

high L2

Adventure	$6,349,781
Action	$5,512,359
Fantasy	$5,079,546
Family Film	$4,024,701
Thriller	$3,479,196
…	…
Western	-$752,683
Black-and-white	-$1,389,215
World cinema	-$1,534,435
Drama	-$2,432,272
Indie	-$3,040,457

BIAS: $45,044,525

Assumptions

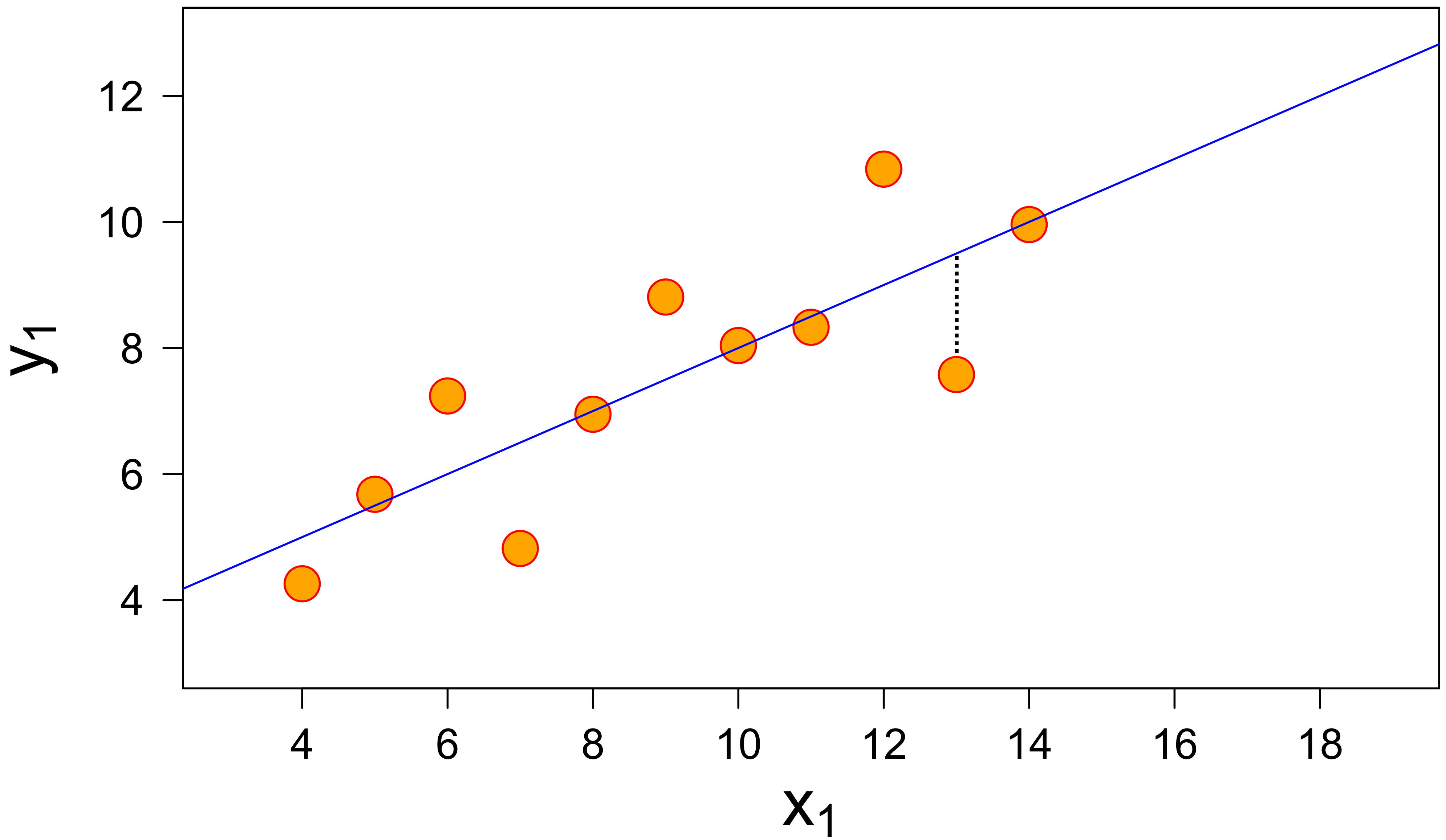


Anscombe's quartet

Probabilistic Interpretation

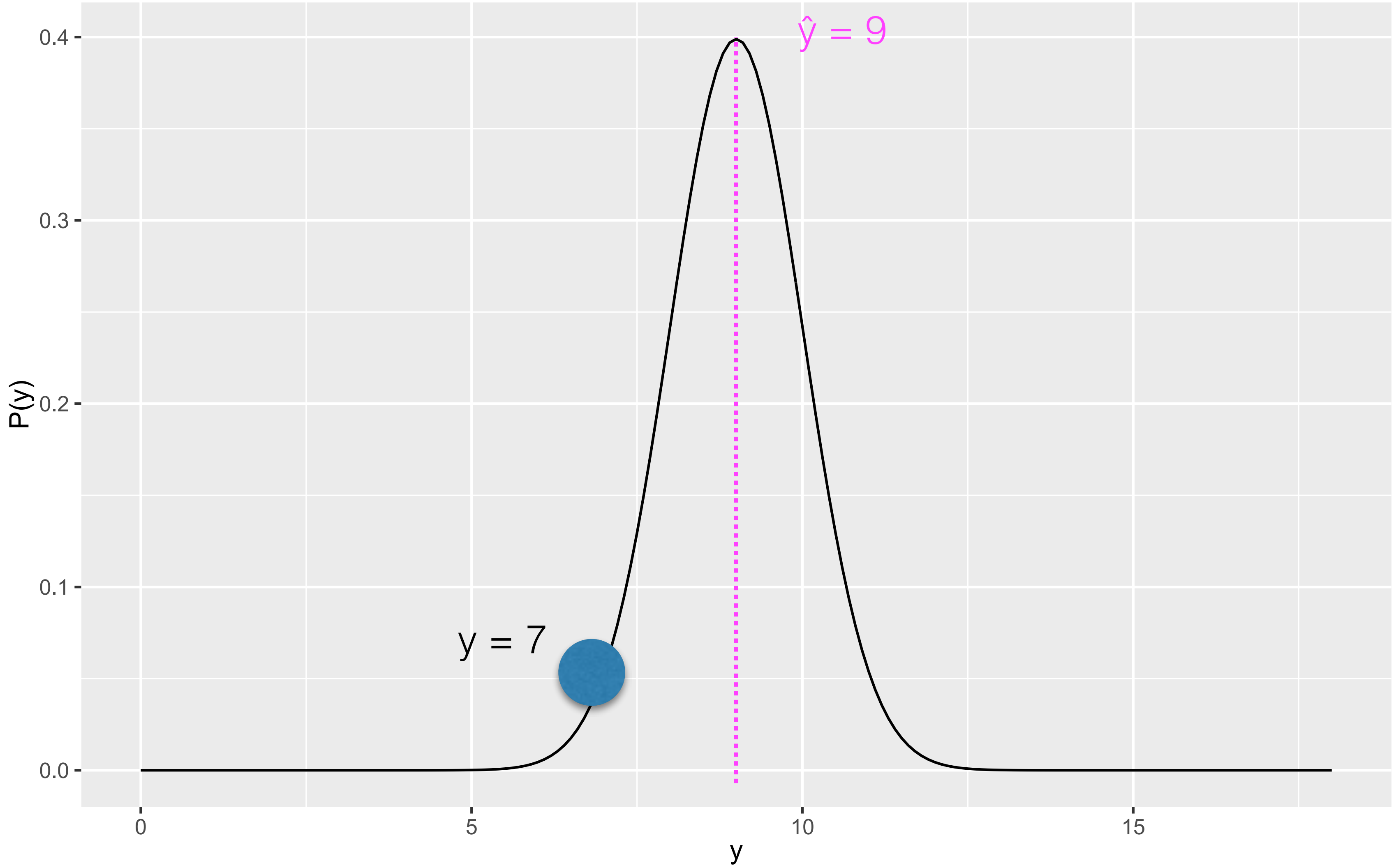
$$P(y_i \mid x, \beta) = \mathrm{Norm}(y_i \mid \hat{y}_i, \sigma^2)$$

"the errors are normally distributed"


12
10
8
6
4
4
6
8
10
12
14
16
18
y_1
x_1

Probabilistic Interpretation

$$P(y_i \mid x, \beta) = \mathrm{Norm}(y_i \mid \hat{y}_i, \sigma^2)$$



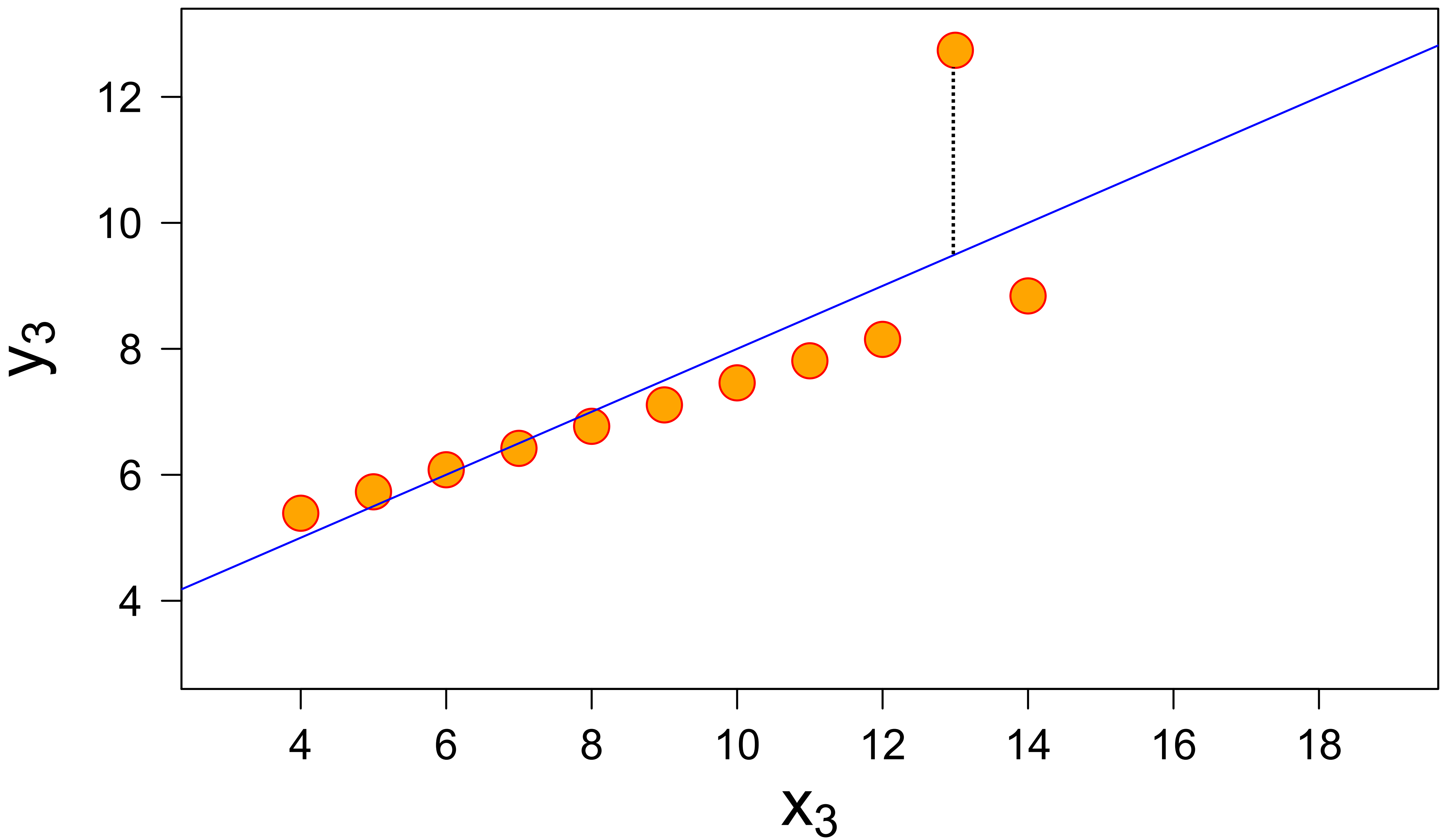
Conditional likelihood

$$\prod_{i}^{N} P(y_i \mid x_i, \beta)$$

For all training data, we want probability of the true value y for each data point x to high

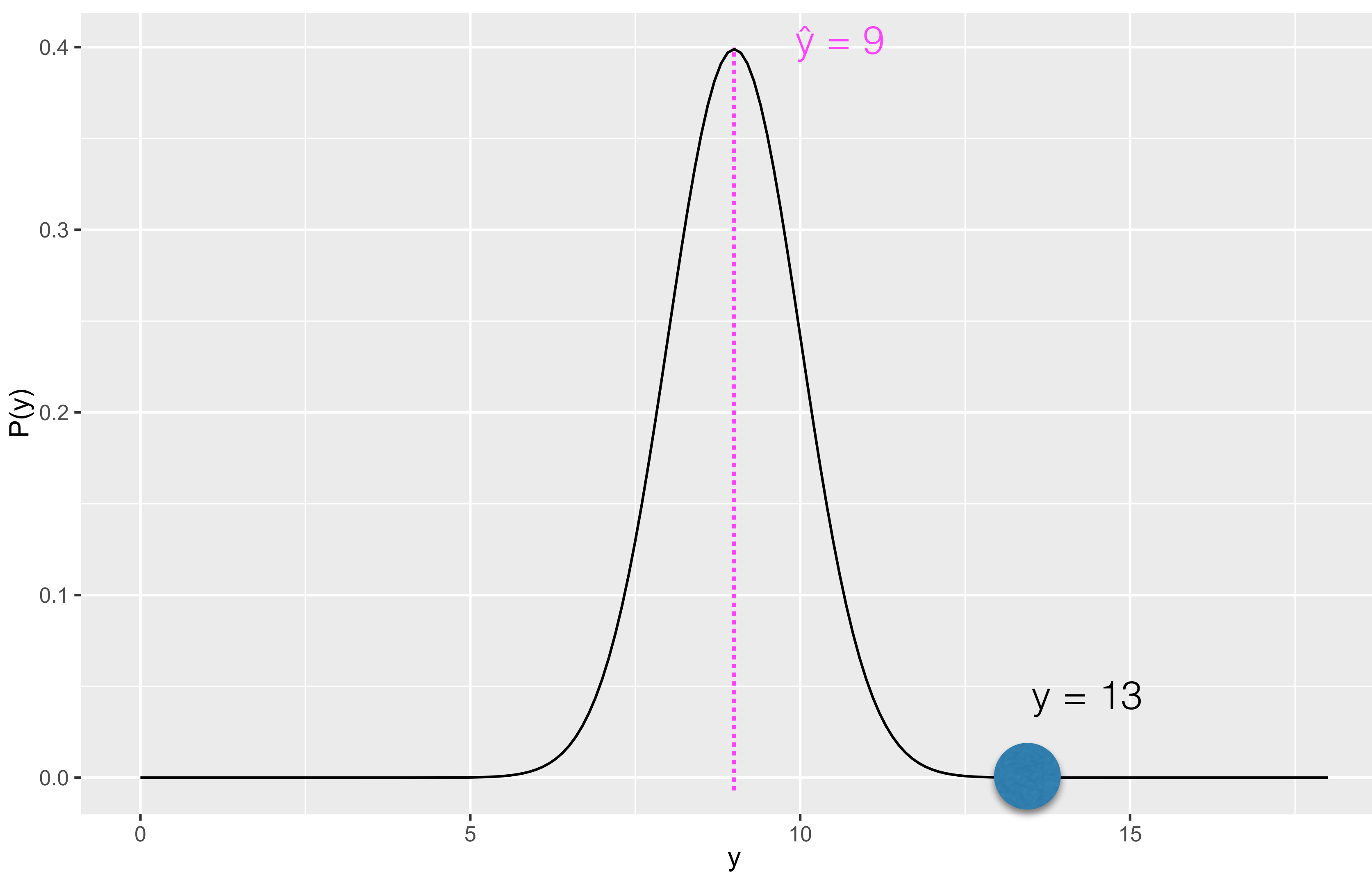
This principle gives us a way to pick the values of the parameters β that maximize the probability of the training data <x, y>

Outliers



Probabilistic Interpretation

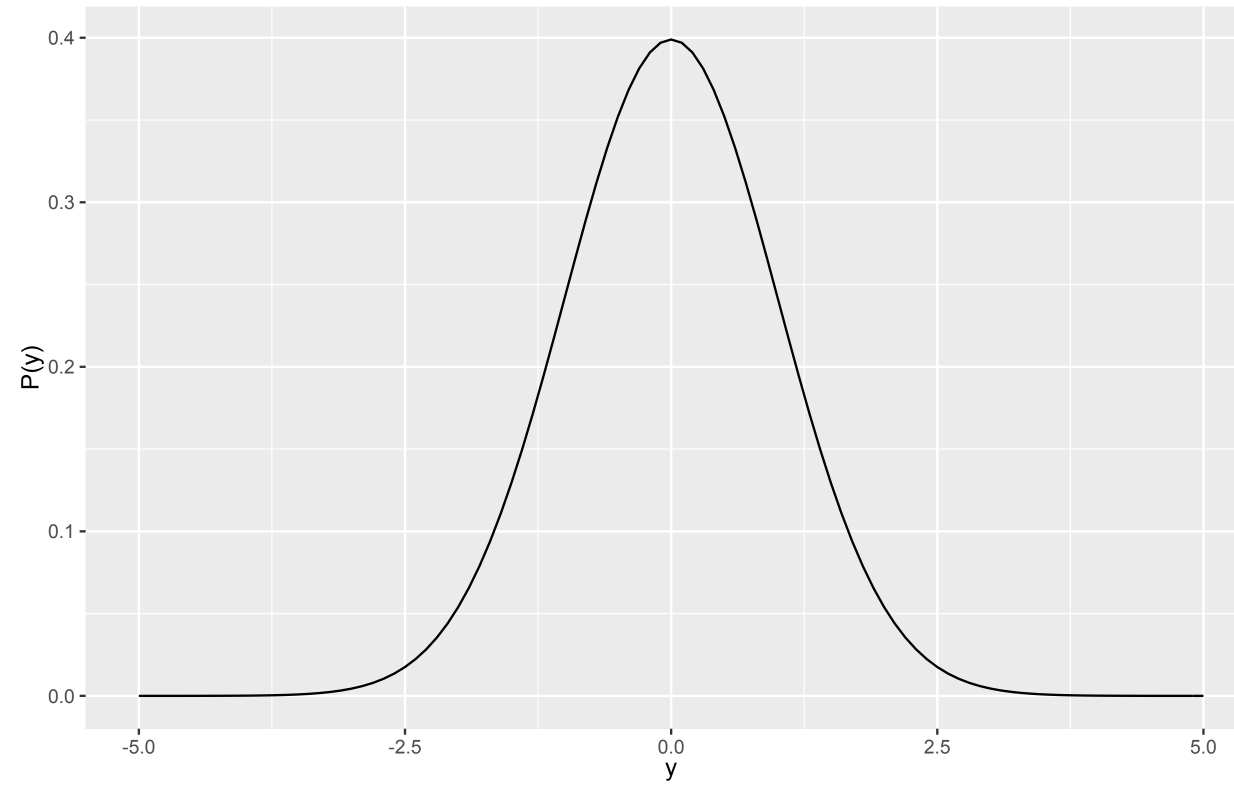
$$P(y_i \mid x, \beta) = \mathrm{Norm}(y_i \mid \hat{y}_i, \sigma^2)$$



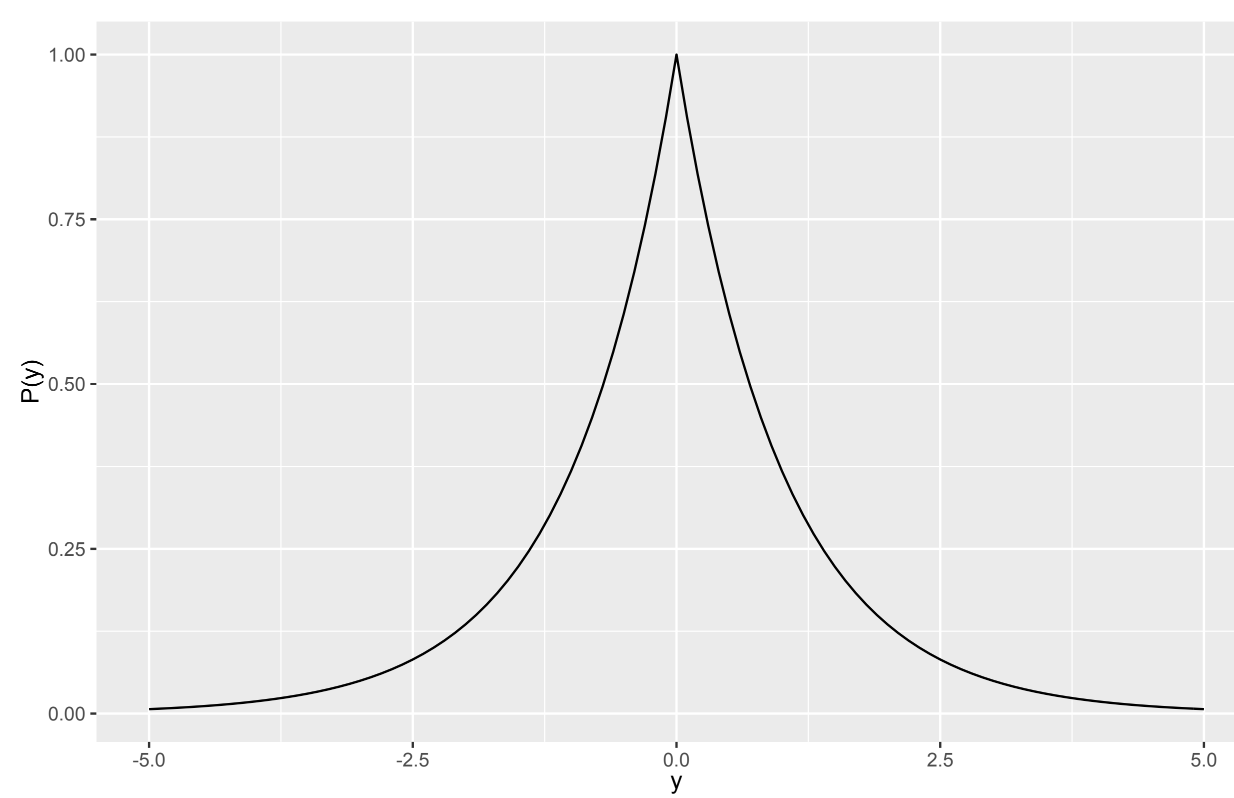
Robust regression

- Rather than modeling the errors as normally distributed, pick some heavier-tailed distribution instead
- This will assign higher likelihood to the outliers without having to move the best fit for the coefficients.

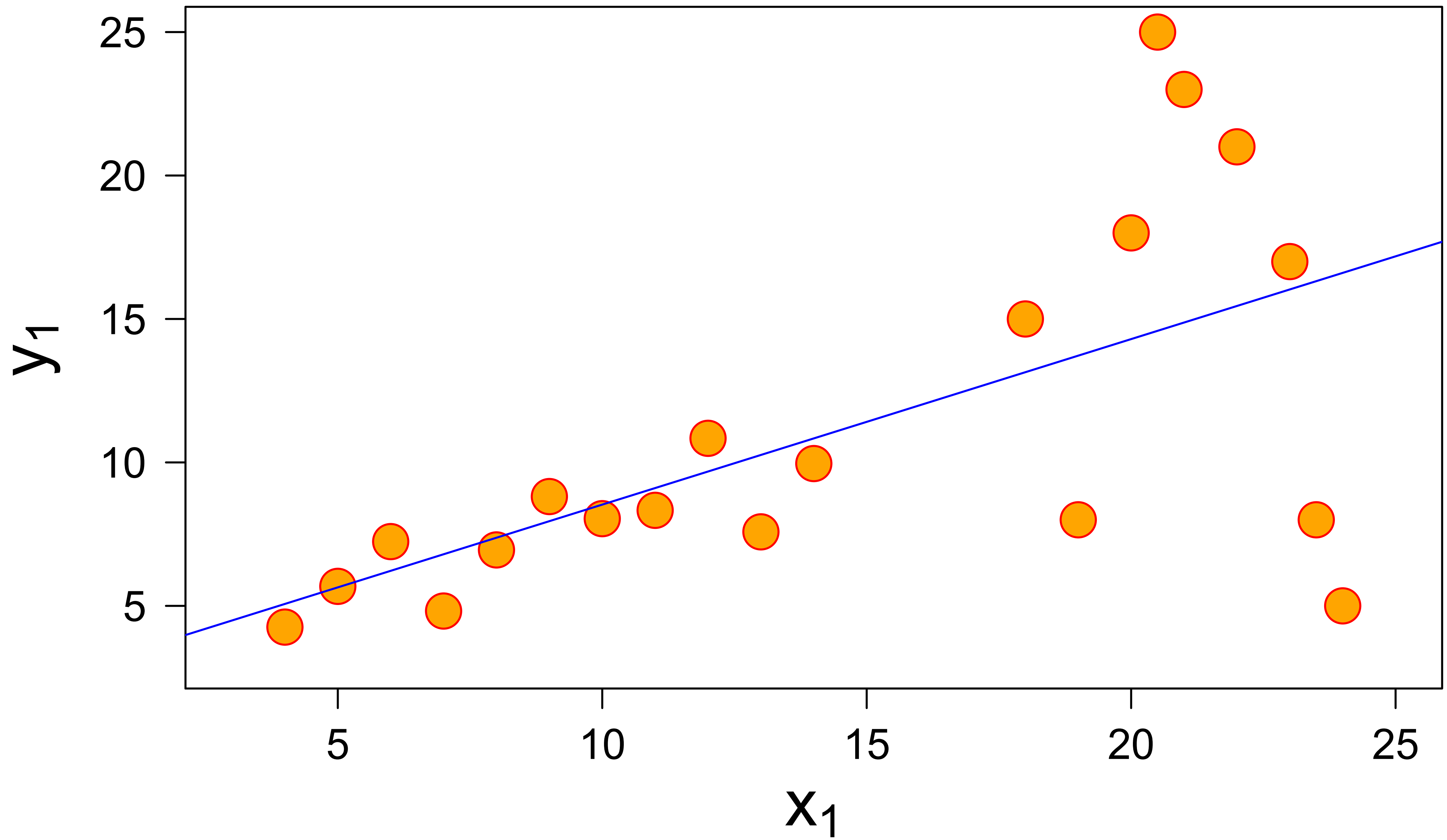
Heavy tailed distributions



Normal vs Laplace



Homoscedasticity



Assumption that the variance in y is constant for all values of x; this data is *heteroscedastic*

Evaluation

Goodness of fit (to training data)

$$R^2 = 1 - \frac{\sum_i (y_i - \hat{y}_i)^2}{\sum_i (y_i - \bar{y})^2}$$

sum of square errors

total sum of squares

For most models, R^2 ranges from 0 (no fit) to 1 (perfect fit)

Experiment design

	training	development	testing
size	80%	10%	10%
purpose	training models	model selection	evaluation; never look at it until the very end

Metrics

- Measure difference between the prediction ŷ and the true y

Mean squared error (MSE)

$$\frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2$$

Mean absolute error (MAE)

$$\frac{1}{N}\sum_{i=1}^{N}|\hat{y}_i - y_i|$$

Interpretation

$$\hat{y} = x_0\beta_0 + x_1\beta_1$$

$$x_0\beta_0 + (x_1 + 1)\beta_1$$

Let's increase the value of x_1 by 1 (e.g., from 0 → 1)

$$x_0\beta_0 + x_1\beta_1 + \beta_1$$

$$= \hat{y} + \beta_1$$

β represents the degree to which y changes with a 1-unit increase in x

Independence

benedict cumberbatch stars movie good	1
terrible acting benedict cumberbatch	0
benedict cumberbatch script excellent	1
excellent script movie good	1
benedict cumberbatch good excellent	1

- benedict
- cumberbatch
- stars
- movie
- good
- acting
- script
- excellent
- terrible

Independence

benedict_cumberbatch stars movie good	1
terrible acting benedict_cumberbatch	0
benedict_cumberbatch script excellent	1
excellent script movie good	1
benedict_cumberbatch good excellent	1

- benedict_cumberbatch
- stars
- movie
- good
- acting
- script
- excellent
- terrible

Significance

Joshi et al. (2010)

I	ngrams
II	POS ngrams
III	Dependency relations

	Features	Site	Total MAE ($M)	Total r	Per Screen MAE ($K)	Per Screen r
	Predict mean		11.672	–	6.862	–
	Predict median		10.521	–	6.642	–
meta	Best		5.983	0.722	6.540	0.272
text	I	–	8.013	0.743	6.509	0.222
		+	7.722	0.781	6.071	0.466
	see Tab. 3	B	7.627	0.793	6.060	0.411
	I ∪ II	–	8.060	0.743	6.542	0.233
		+	**7.420**	0.761	6.240	0.398
		B	7.447	0.778	6.299	0.363
	I ∪ III	–	8.005	0.744	6.505	0.223
		+	7.721	0.785	6.013	**0.473**
		B	7.595	**0.796**	†**6.010**	0.421
meta ∪ text	I	–	5.921	**0.819**	6.509	0.222
		+	5.757	0.810	6.063	0.470
		B	5.750	**0.819**	6.052	0.414
	I ∪ II	–	5.952	0.818	6.542	0.233
		+	5.752	0.800	6.230	0.400
		B	5.740	**0.819**	6.276	0.358
	I ∪ III	–	5.921	**0.819**	6.505	0.223
		+	**5.738**	0.812	6.003	**0.477**
		B	5.750	**0.819**	†**5.998**	0.423

Joshi et al. (2010)

	Feature	Weight ($M)
rating	pg	+0.085
	New York Times: adult	-0.236
	New York Times: rate_r	-0.364
sequels	this_series	+13.925
	LA Times: the_franchise	+5.112
	Variety: the_sequel	+4.224
people	*Boston Globe*: will_smith	+2.560
	Variety: brittany	+1.128
	^_producer_brian	+0.486
genre	*Variety*: testosterone	+1.945
	Ent. Weekly: comedy_for	+1.143
	Variety: a_horror	+0.595
	documentary	-0.037
	independent	-0.127
sentiment	*Boston Globe*: best_parts_of	+1.462
	Boston Globe: smart_enough	+1.449
	LA Times: a_good_thing	+1.117
	shame_$	-0.098
	bogeyman	-0.689
plot	*Variety*: torso	+9.054
	vehicle_in	+5.827
	superhero_$	+2.020