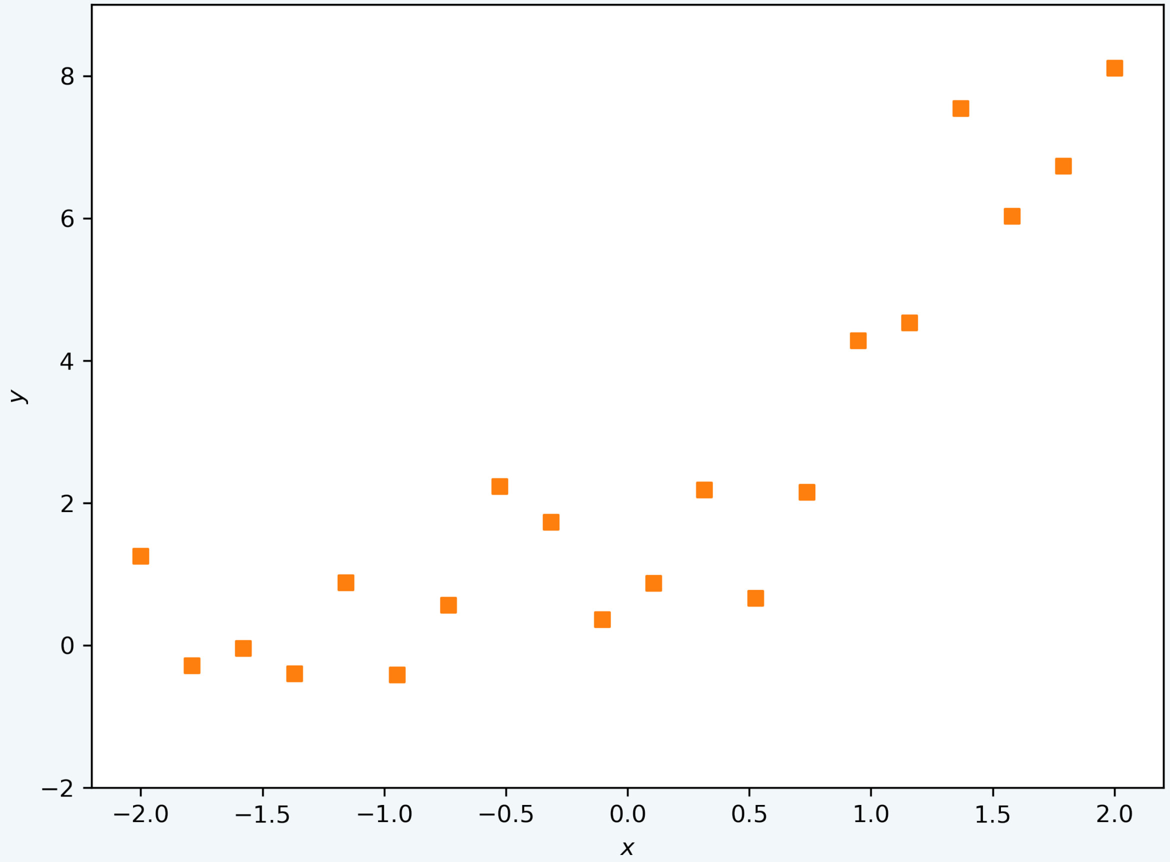
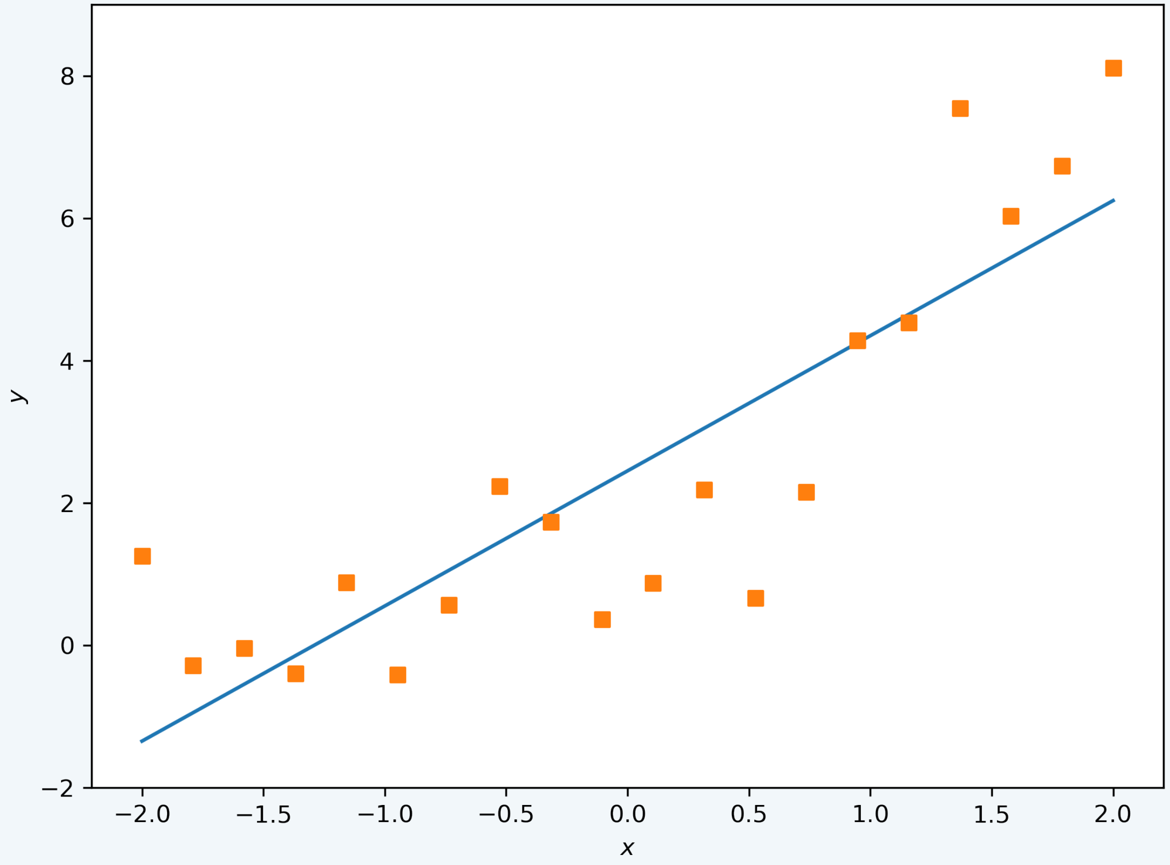
GENERALIZED LINEAR MODELS

Motivation: Normal Linear Models



Let's start with some data…

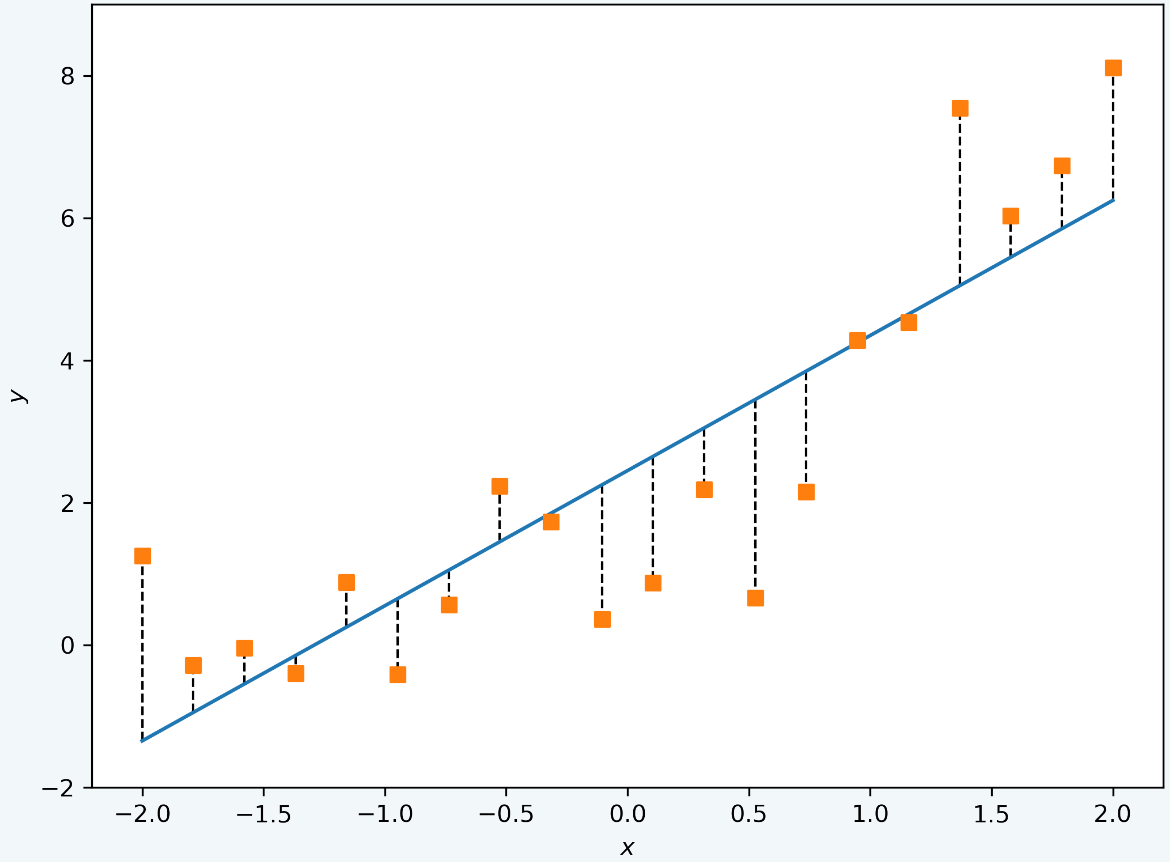
Motivation: Normal Linear Models



Let's start with some data…

Here's an attempt to fit the data. How do we judge it?

Motivation: Normal Linear Models

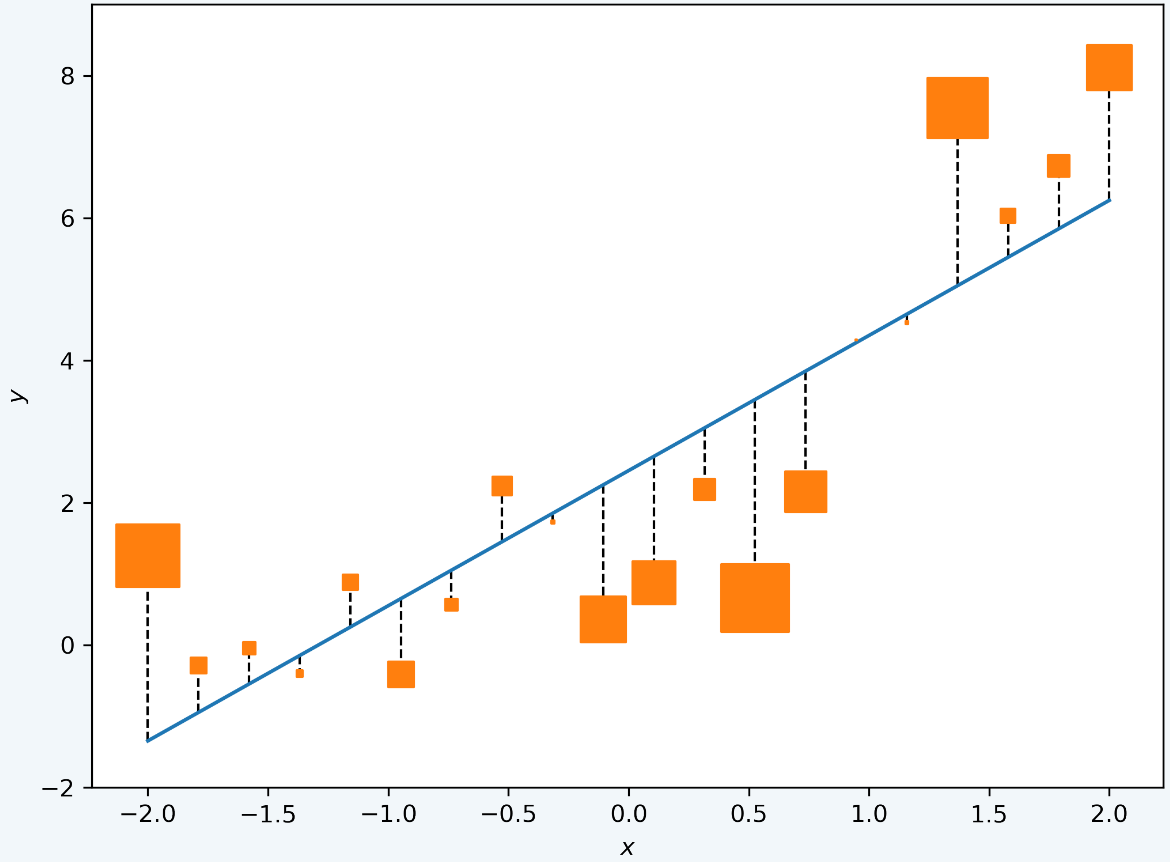


Let's start with some data…

Here's an attempt to fit the data. How do we judge it?

Find the residuals, $r = y - \hat{y}$

Motivation: Normal Linear Models



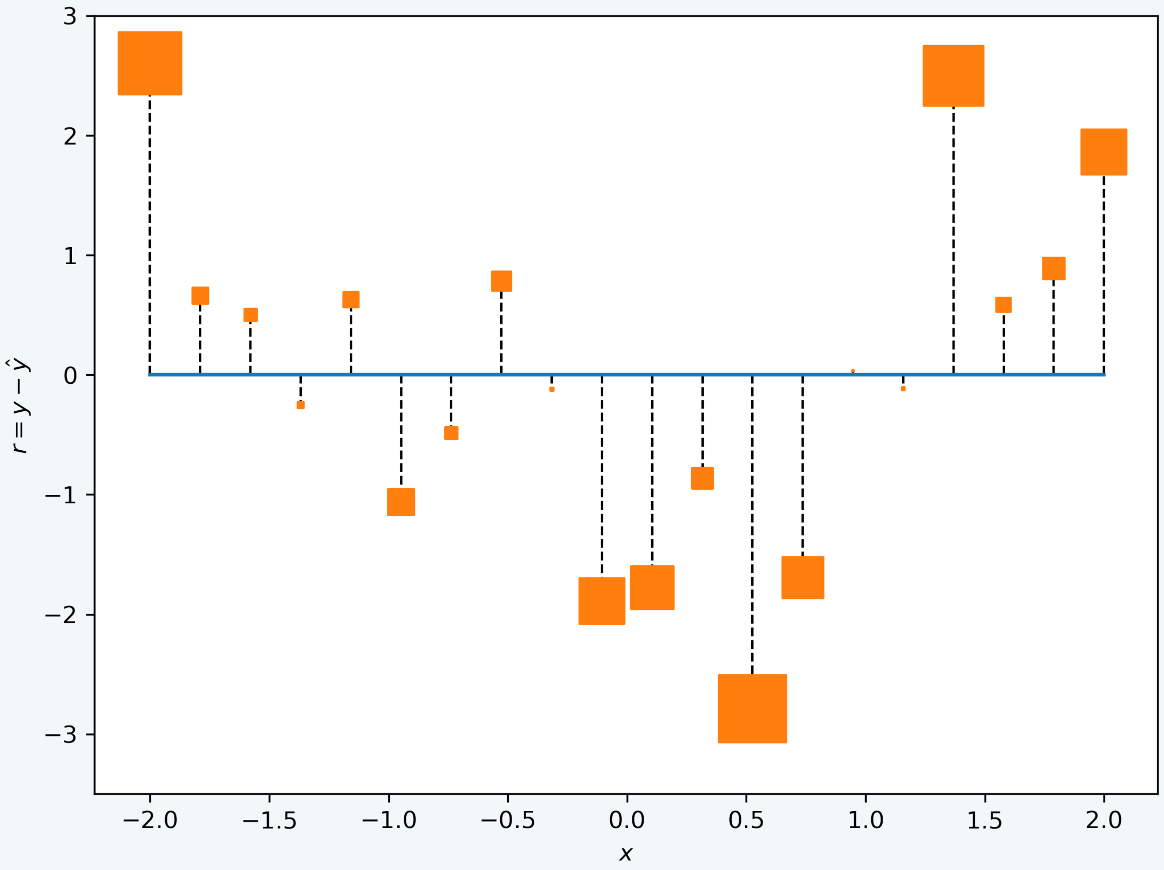
Let's start with some data…

Here's an attempt to fit the data. How do we judge it?

Find the residuals, $r = y - \hat{y}$

Each residual gives a contribution (its square) to the cost

Motivation: Normal Linear Models



Let's start with some data…

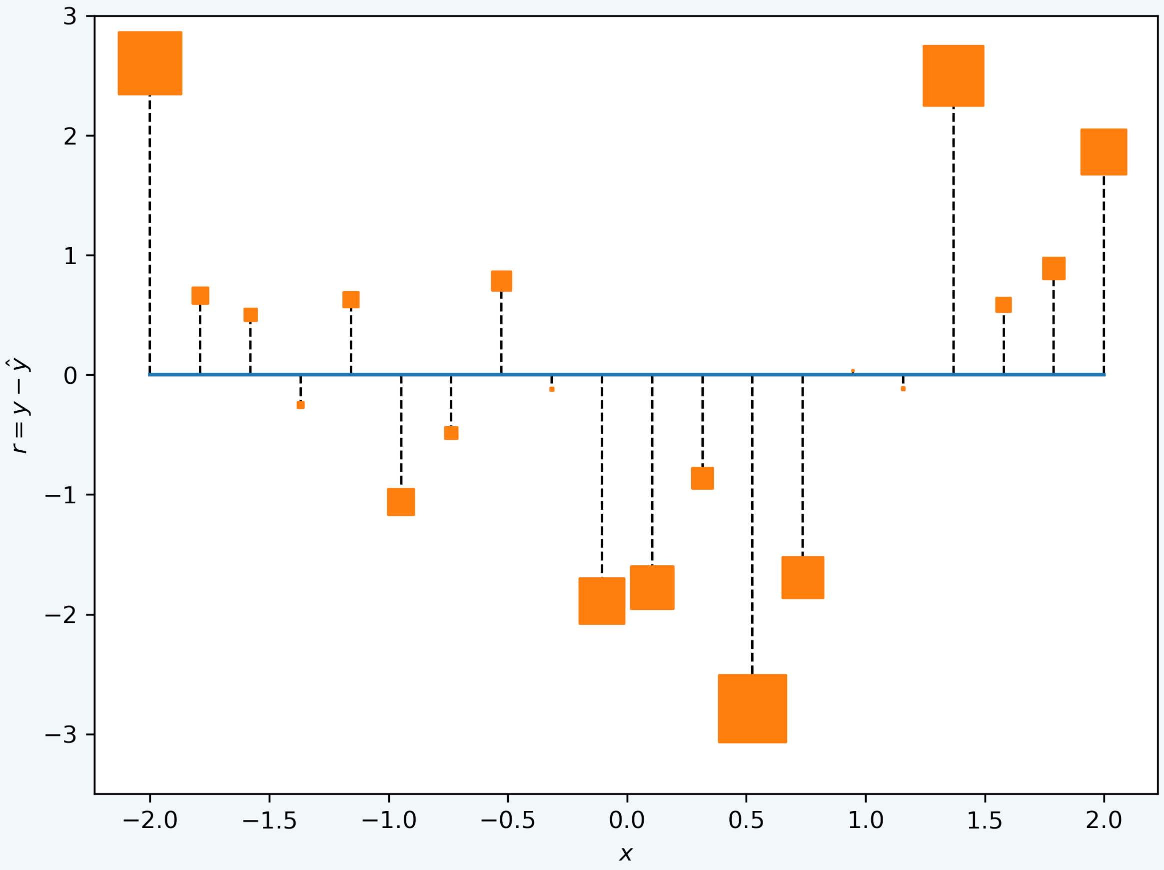
Here's an attempt to fit the data. How do we judge it?

Find the residuals, $r = y - \hat{y}$

Each residual gives a contribution (its square) to the cost

We often plot residuals on the *y*-axis

Motivation: Normal Linear Models



Let’s start with some data…

Here’s an attempt to fit the data. How do we judge it?

Find the residuals, $r = y - \hat{y}$

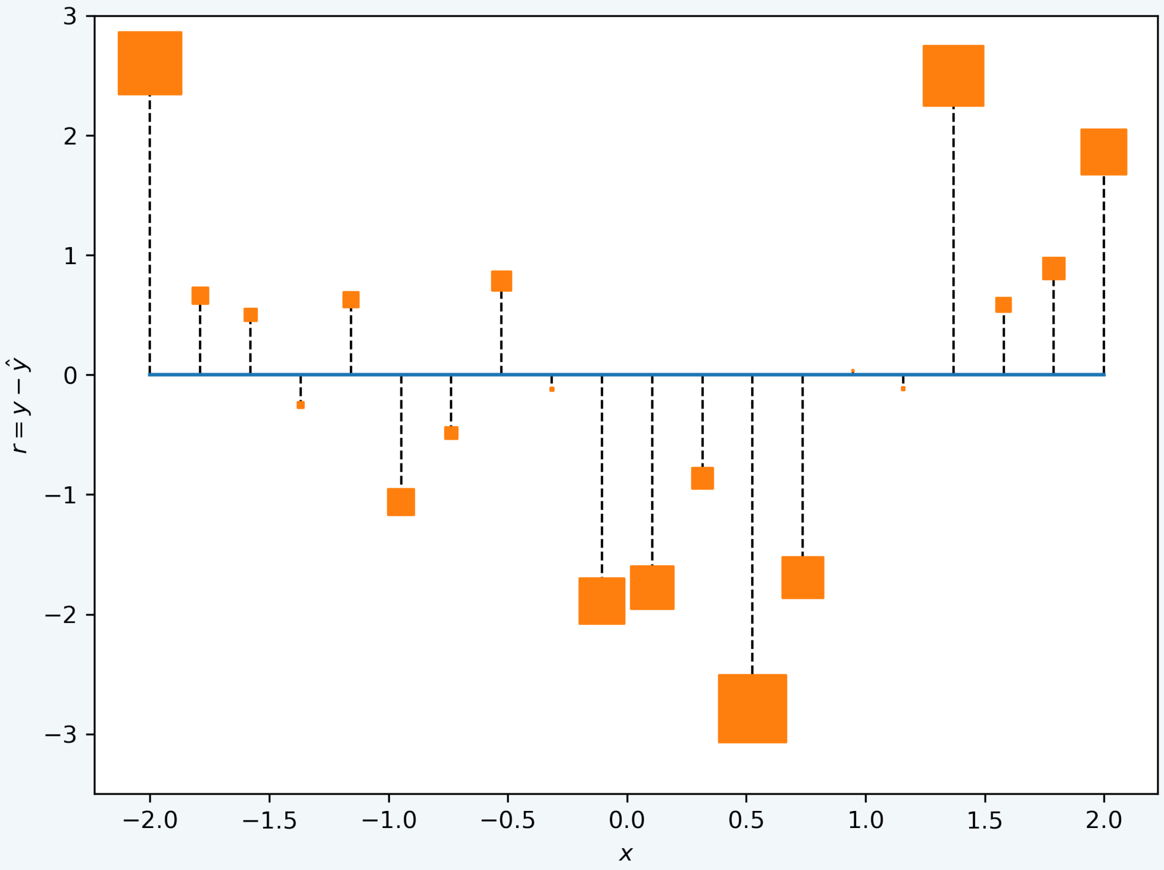
Each residual gives a contribution (its square) to the cost

We often plot residuals on the y-axis

Some things to notice:

- Equal-magnitude residuals make the same contribution to the cost
- A positive (or negative) residual means the data is greater than (or less than) we predicted

Motivation: Normal Linear Models



Let's start with some data…

Here's an attempt to fit the data. How do we judge it?

Find the residuals, $r = y - \hat{y}$

Each residual gives a contribution (its square) to the cost

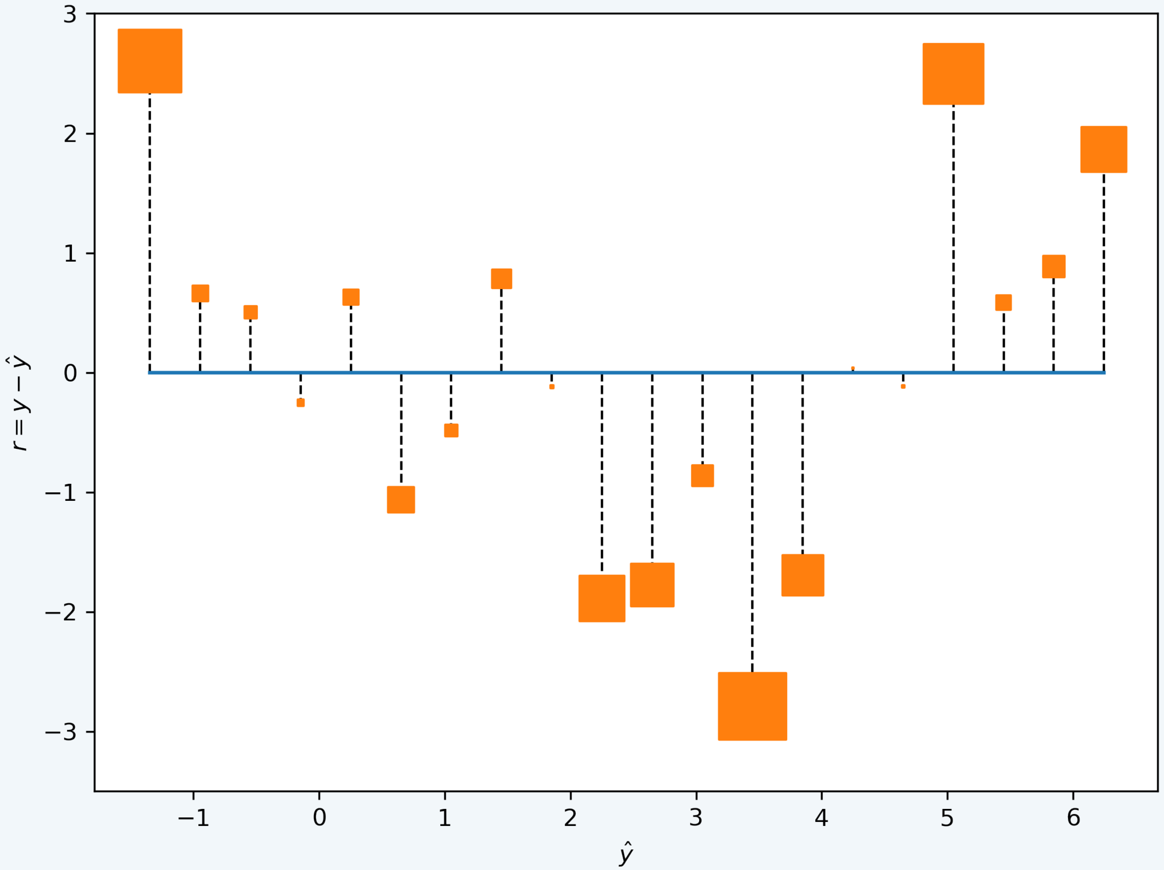
We often plot residuals on the *y*-axis

Some things to notice:

- Equal-magnitude residuals make the same contribution to the cost
- A positive (or negative) residual means the data is greater than (or less than) we predicted

Features are usually higher-dimensional, tough to plot

Motivation: Normal Linear Models



Let's start with some data…

Here's an attempt to fit the data. How do we judge it?

Find the residuals, $r = y - \hat{y}$

Each residual gives a contribution (its square) to the cost

We often plot residuals on the *y*-axis

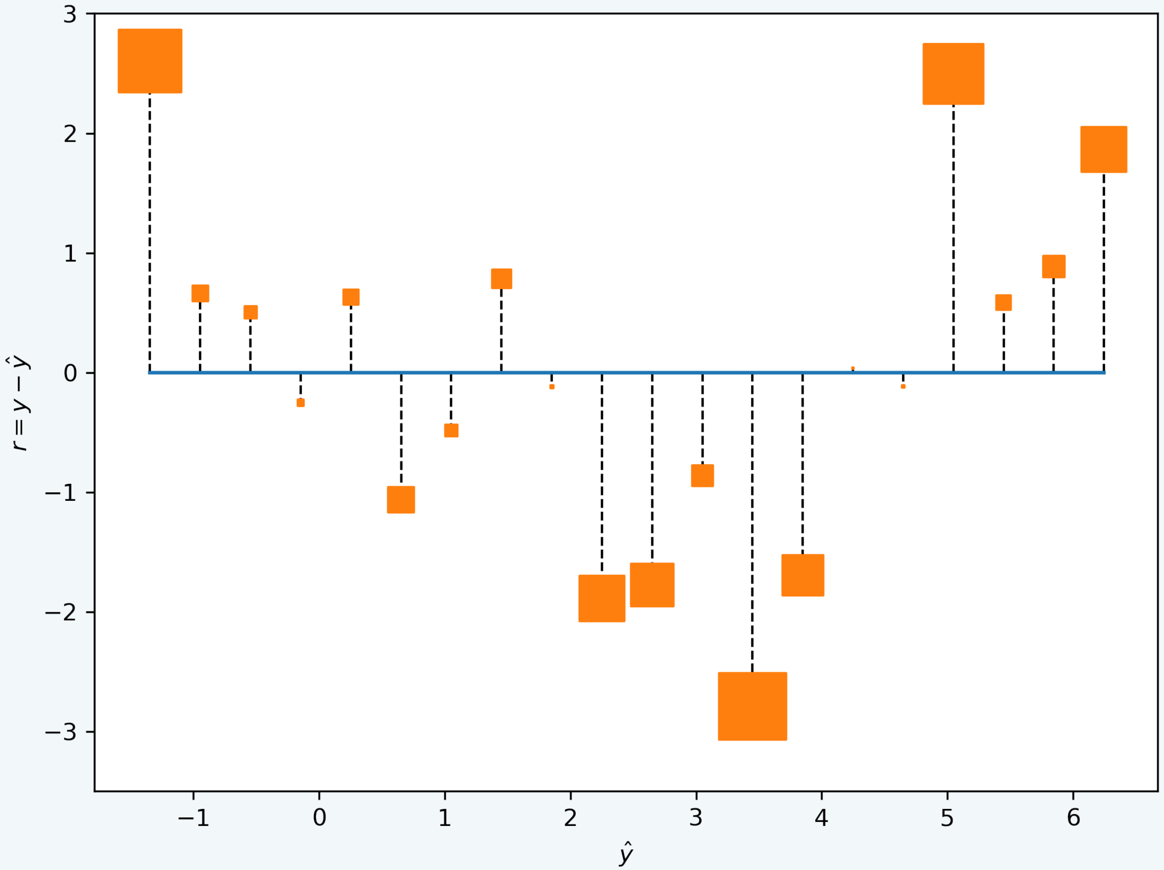
Some things to notice:

- Equal-magnitude residuals make the same contribution to the cost
- A positive (or negative) residual means the data is greater than (or less than) we predicted

Features are usually higher-dimensional, tough to plot

So you'll often see "predicted values" on the *x*-axis

Motivation: Normal Linear Models



Let's start with some data…

Here's an attempt to fit the data. How do we judge it?

Find the residuals, $r = y - \hat{y}$

Each residual gives a contribution (its square) to the cost

We often plot residuals on the *y*-axis

Some things to notice:

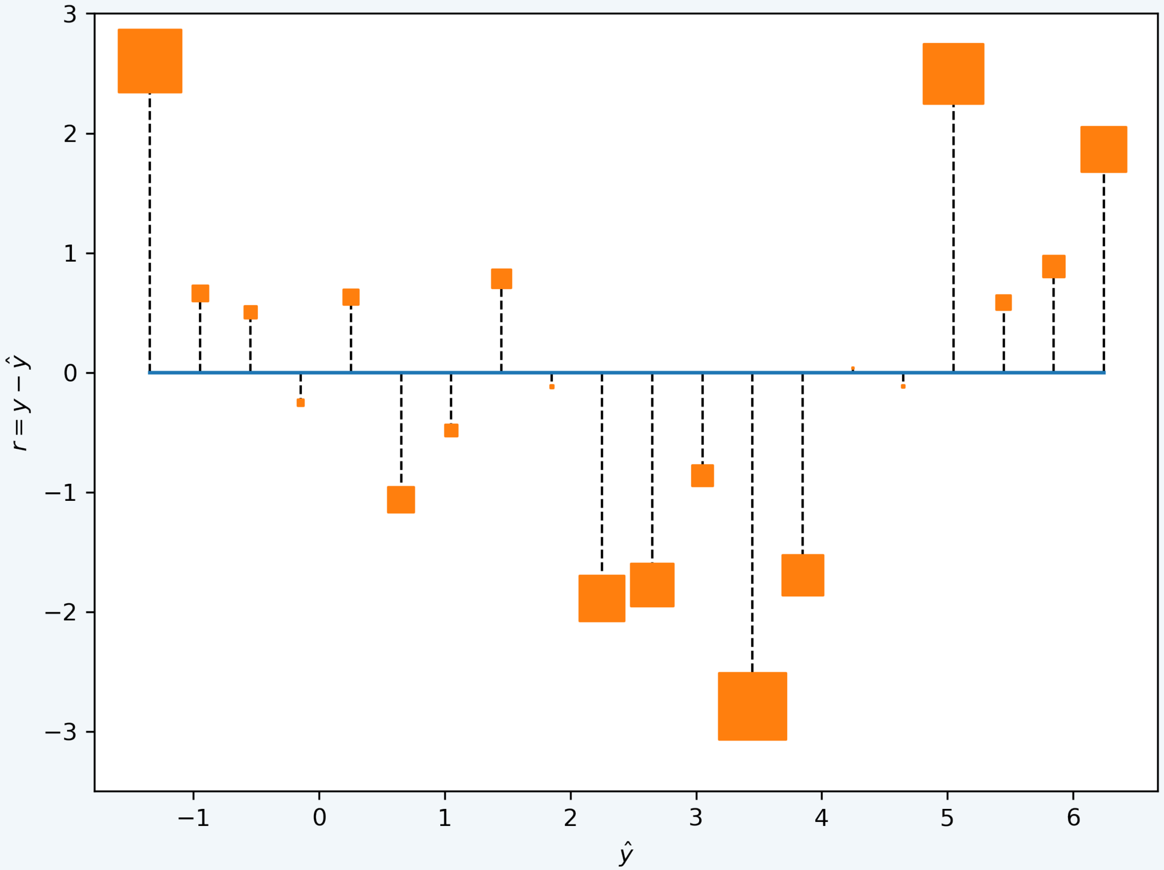
- Equal-magnitude residuals make the same contribution to the cost
- A positive (or negative) residual means the data is greater than (or less than) we predicted

Features are usually higher-dimensional, tough to plot

So you'll often see "predicted values" on the *x*-axis

We can also check for patterns this way

Motivation: Normal Linear Models



Let's start with some data…

Here's an attempt to fit the data. How do we judge it?

Find the residuals, $r = y - \hat{y}$

Each residual gives a contribution (its square) to the cost

We often plot residuals on the *y*-axis

Some things to notice:

- Equal-magnitude residuals make the same contribution to the cost
- A positive (or negative) residual means the data is greater than (or less than) we predicted

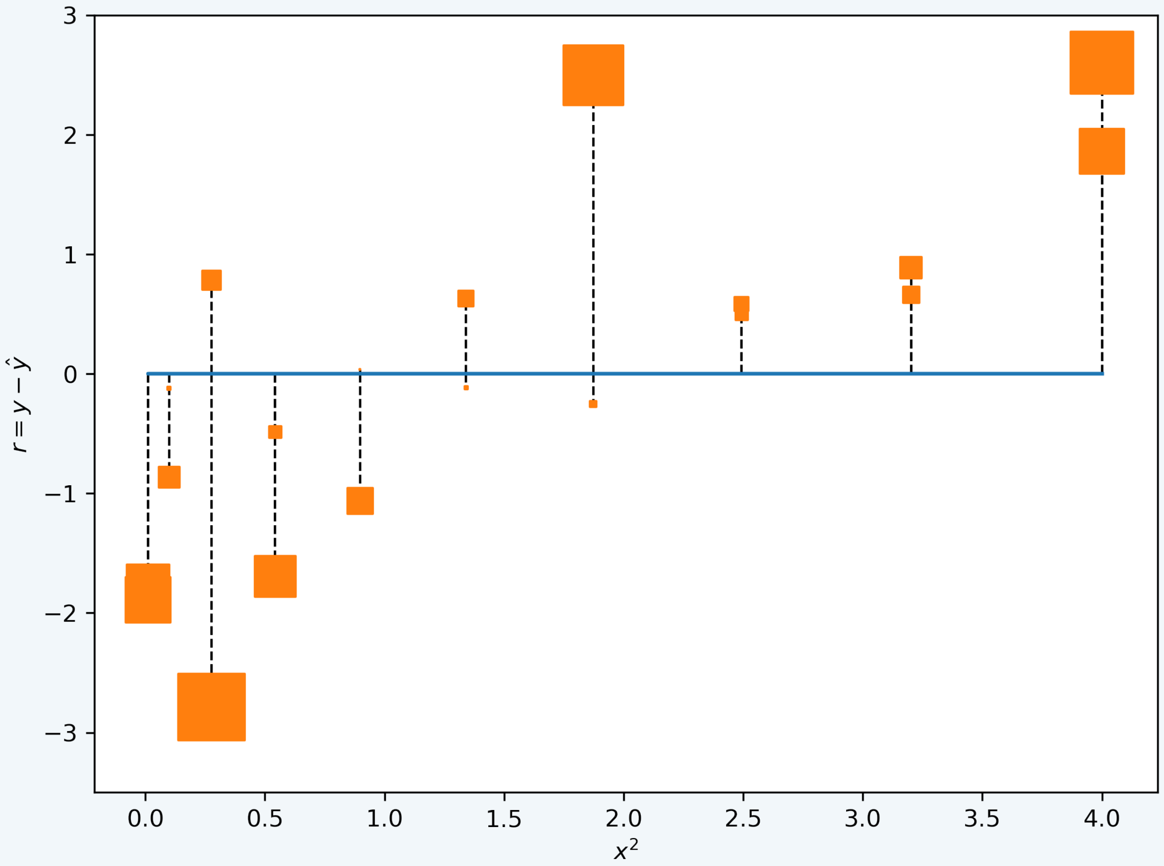
Features are usually higher-dimensional, tough to plot

So you'll often see "predicted values" on the *x*-axis

We can also check for patterns this way

Would an x^2 feature improve the fit?

Motivation: Normal Linear Models



Let's start with some data…

Here's an attempt to fit the data. How do we judge it?

Find the residuals, $r = y - \hat{y}$

Each residual gives a contribution (its square) to the cost

We often plot residuals on the *y*-axis

Some things to notice:

- Equal-magnitude residuals make the same contribution to the cost
- A positive (or negative) residual means the data is greater than (or less than) we predicted

Features are usually higher-dimensional, tough to plot

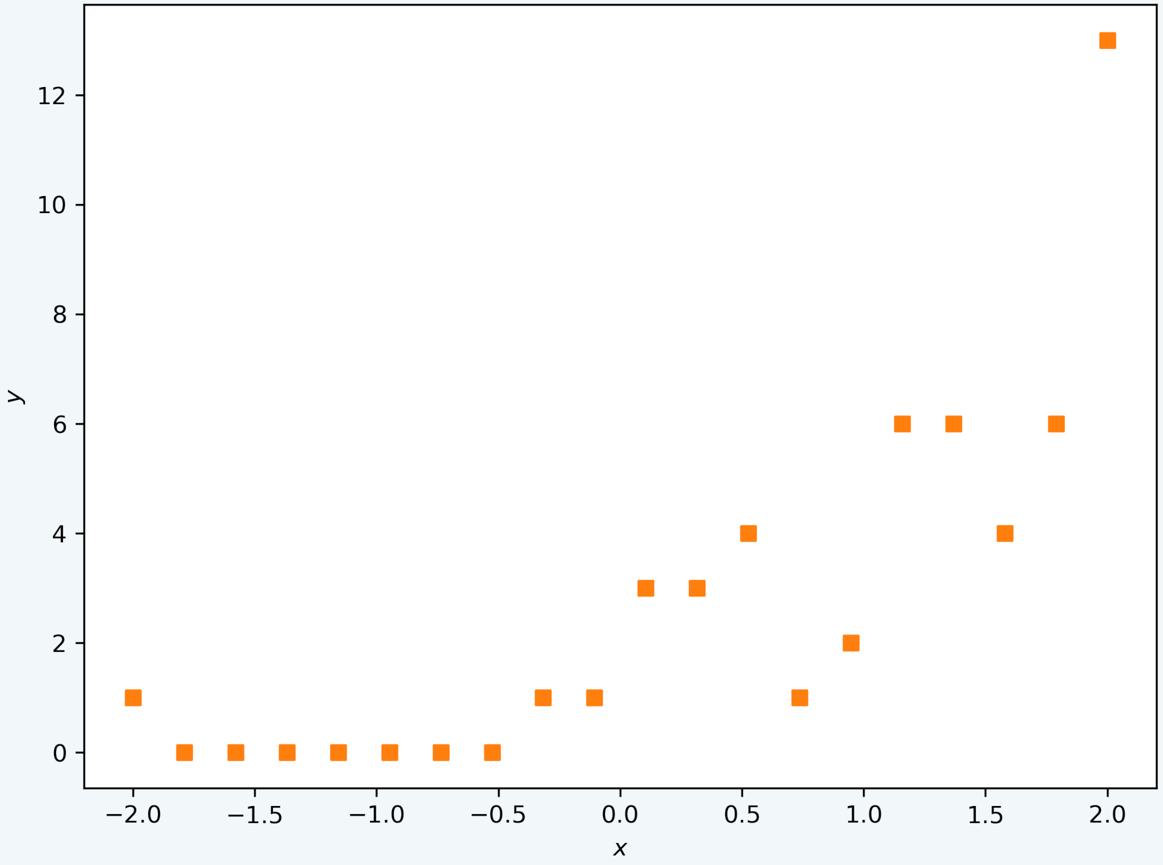
So you'll often see "predicted values" on the *x*-axis

We can also check for patterns this way

Would an x^2 feature improve the fit?

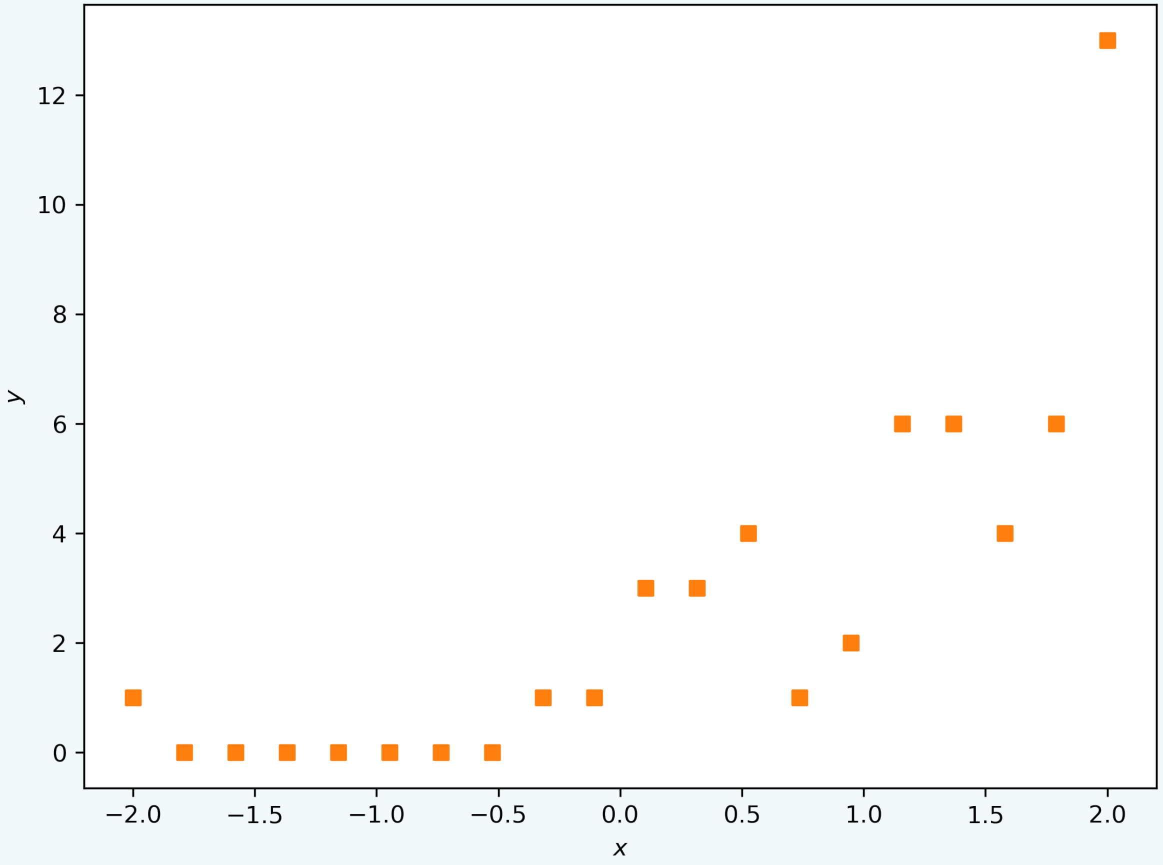
Looks like a yes!

Can we Generalize?



Say we have some count data like this

Can we Generalize?

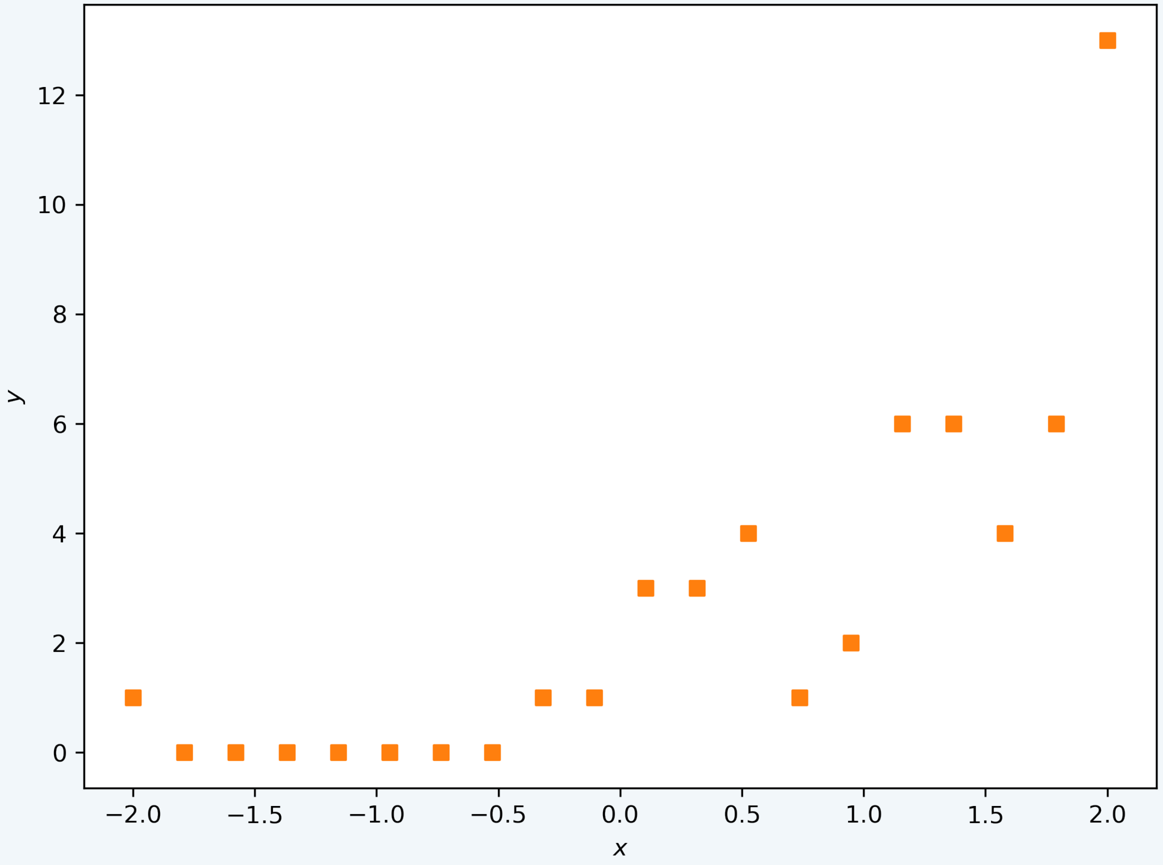


Say we have some count data like this

To this point we've seen conditional expectations that are…

- Normal (linear regression)
- Bernoulli or binomial (logistic regression)

Can we Generalize?



Say we have some count data like this

To this point we've seen conditional expectations that are…

- Normal (linear regression)
- Bernoulli or binomial (logistic regression)

It would be great to…

- Find something similar for count data
- Find the analog of residuals and squared residuals

Poisson Regression: Just Another GLM

Here are three generalized linear models:

Linear regression	identity $\mathbb{E}(y \mid X) = X\beta$	$y \sim$ Normal$(X\beta, \sigma)$
Logistic regression	logit $\mathbb{E}(y \mid X) = X\beta$	$y \sim$ Bernoulli $\left(\text{logit}^{-1}(X\beta)\right)$
Poisson regression	log $\mathbb{E}(y \mid X) = X\beta$	$y \sim$ Poisson $\left(\exp(X\beta)\right)$

Poisson Regression: Just Another GLM

Here are three generalized linear models:

Linear regression	identity $\mathbb{E}(y \mid X) = X\beta$	$y \sim$ Normal$(X\beta, \sigma)$
Logistic regression	logit $\mathbb{E}(y \mid X) = X\beta$	$y \sim$ Bernoulli $\left(\text{logit}^{-1}(X\beta)\right)$
Poisson regression	log $\mathbb{E}(y \mid X) = X\beta$	$y \sim$ Poisson $\left(\exp(X\beta)\right)$

$X\beta$ is the *systematic component*

Poisson Regression: Just Another GLM

Here are three generalized linear models:

Linear regression	identity $\mathbb{E}(y\,|\,X) = X\beta$	$y \sim$ Normal$(X\beta, \sigma)$
Logistic regression	logit $\mathbb{E}(y\,|\,X) = X\beta$	$y \sim$ Bernoulli $\left(\text{logit}^{-1}(X\beta)\right)$
Poisson regression	log $\mathbb{E}(y\,|\,X) = X\beta$	$y \sim$ Poisson $\left(\exp(X\beta)\right)$

The distribution of *y* is the *stochastic component*

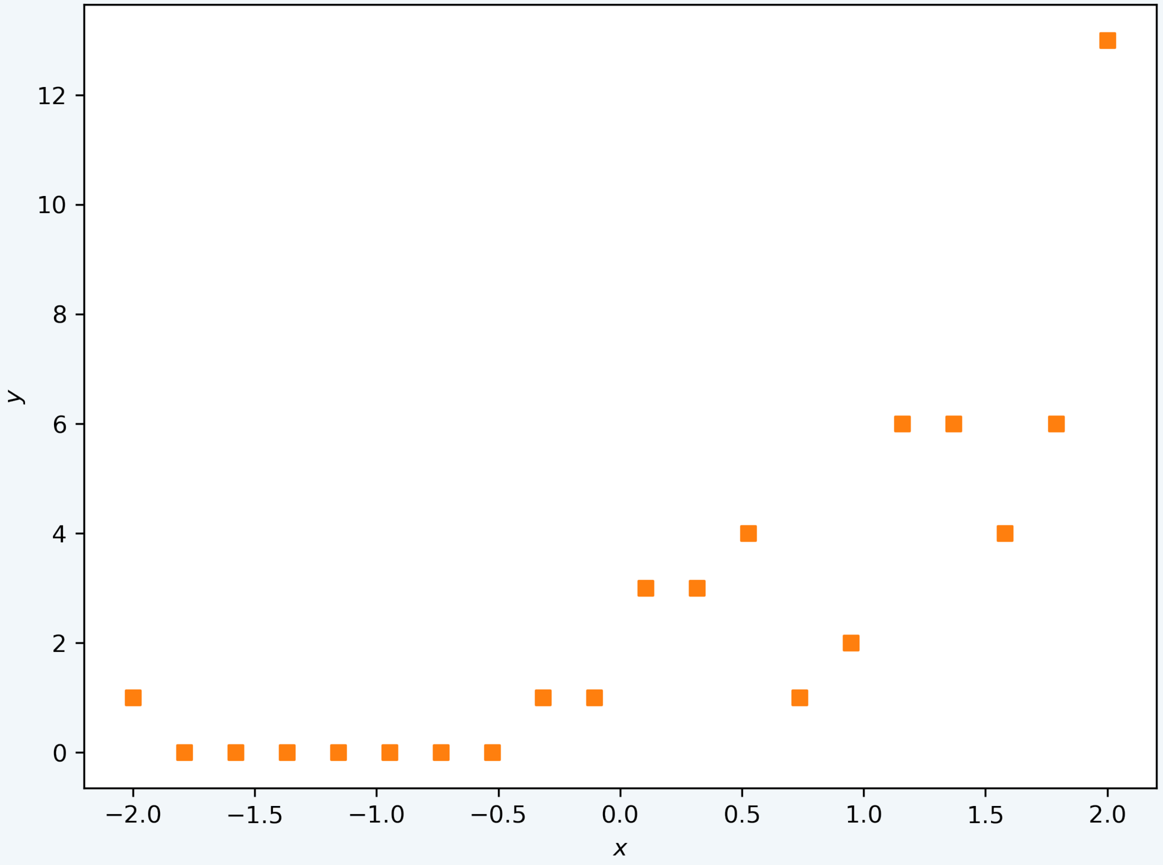
Poisson Regression: Just Another GLM

Here are three generalized linear models:

Linear regression	identity $\mathbb{E}(y \mid X) = X\beta$	$y \sim$ Normal$(X\beta, \sigma)$
Logistic regression	logit $\mathbb{E}(y \mid X) = X\beta$	$y \sim$ Bernoulli $\left(\text{logit}^{-1}(X\beta)\right)$
Poisson regression	log $\mathbb{E}(y \mid X) = X\beta$	$y \sim$ Poisson $\left(\exp(X\beta)\right)$

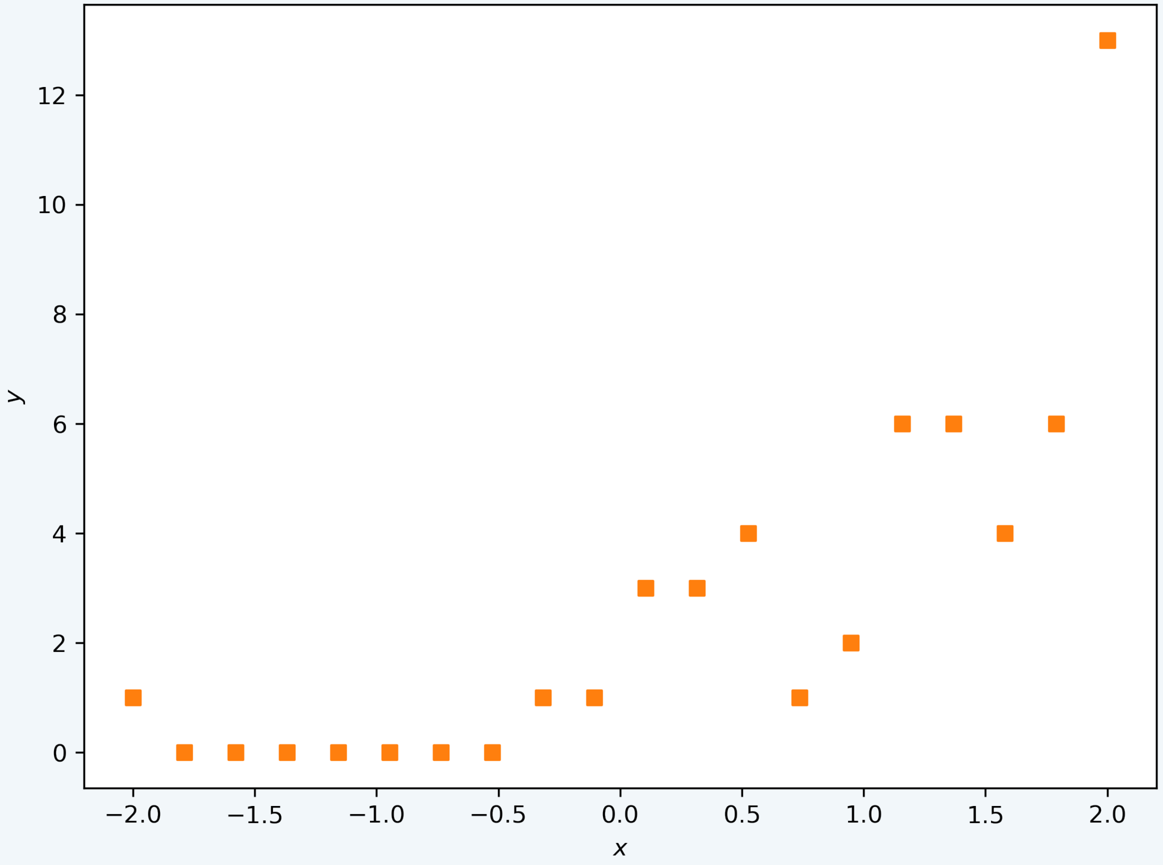
The *link function* connects the systematic and stochastic components

Like Counting Fish in a Barrel



Back to our count data

Like Counting Fish in a Barrel

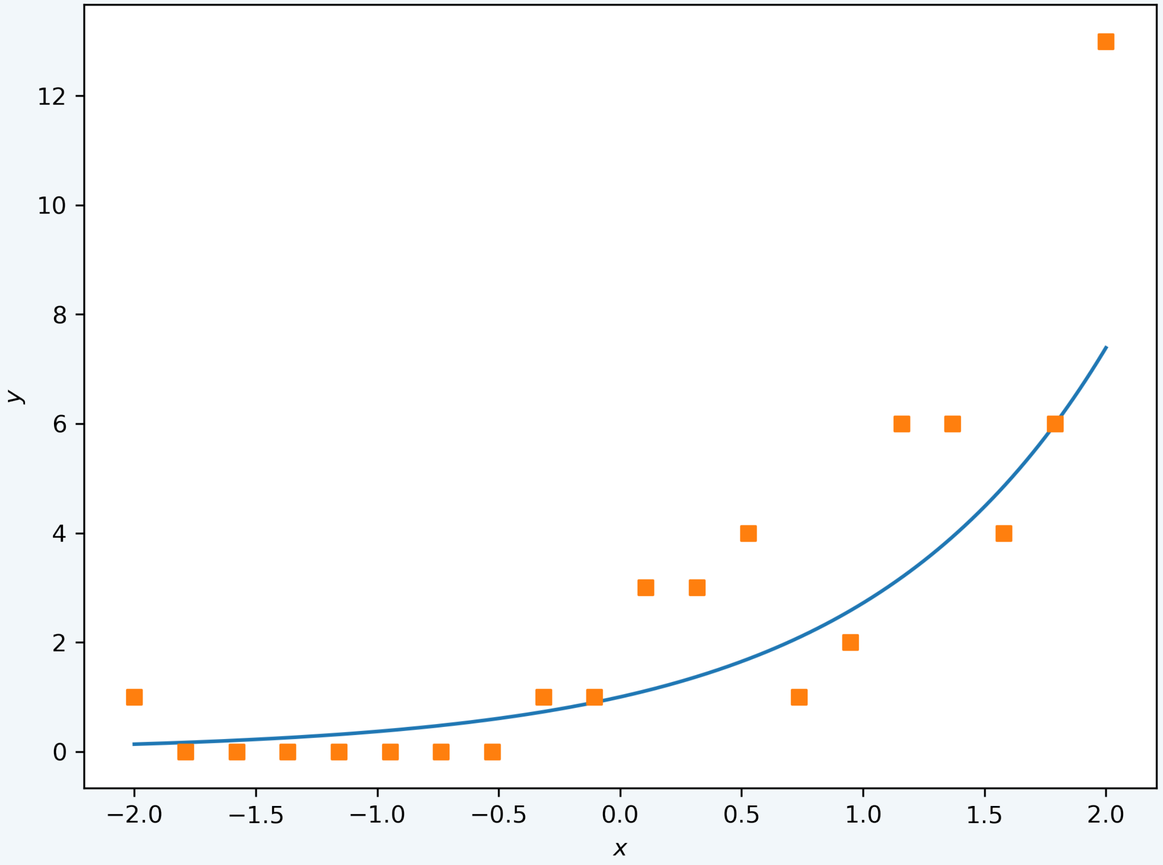


Back to our count data

The Poisson regression model is

$$y \sim \text{Poisson}\left(\exp(X\beta)\right)$$

Like Counting Fish in a Barrel



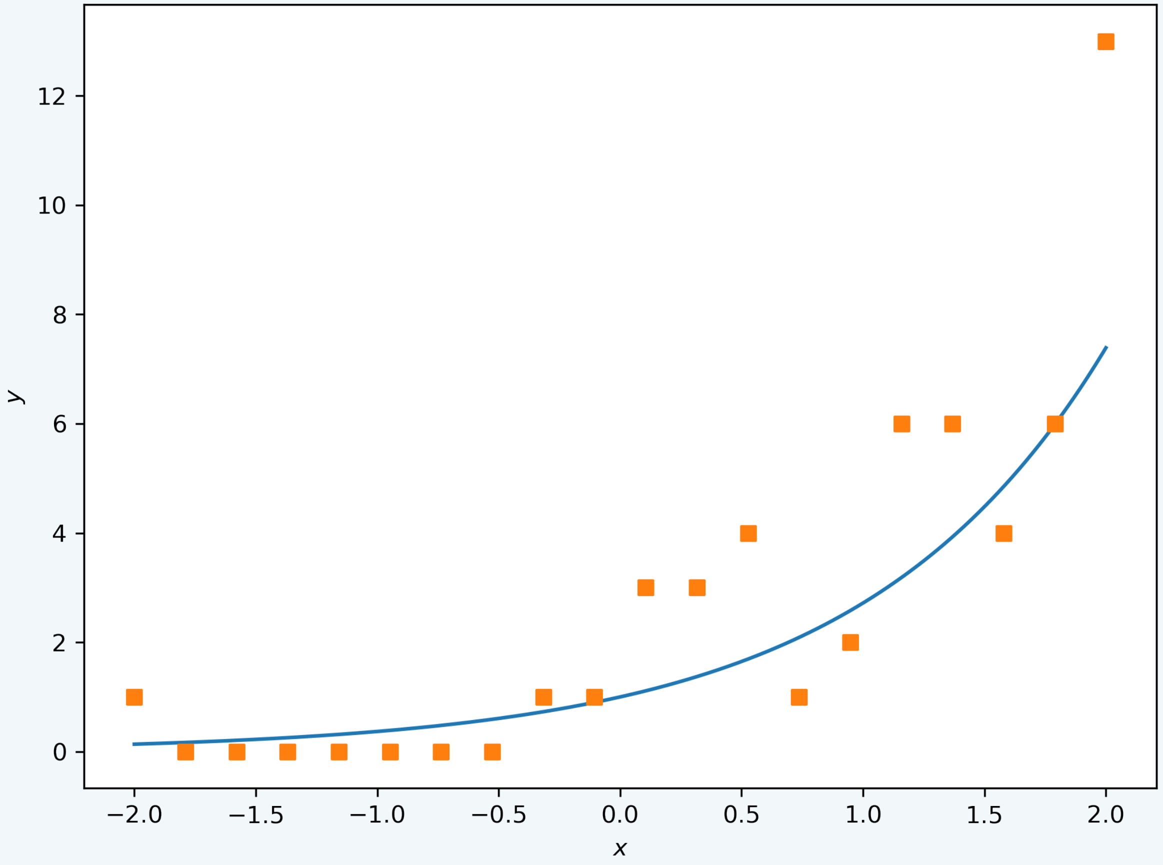
Back to our count data

The Poisson regression model is

$$y \sim \text{Poisson}\left(\exp(X\beta)\right)$$

Here's the conditional expectation

Like Counting Fish in a Barrel



Back to our count data

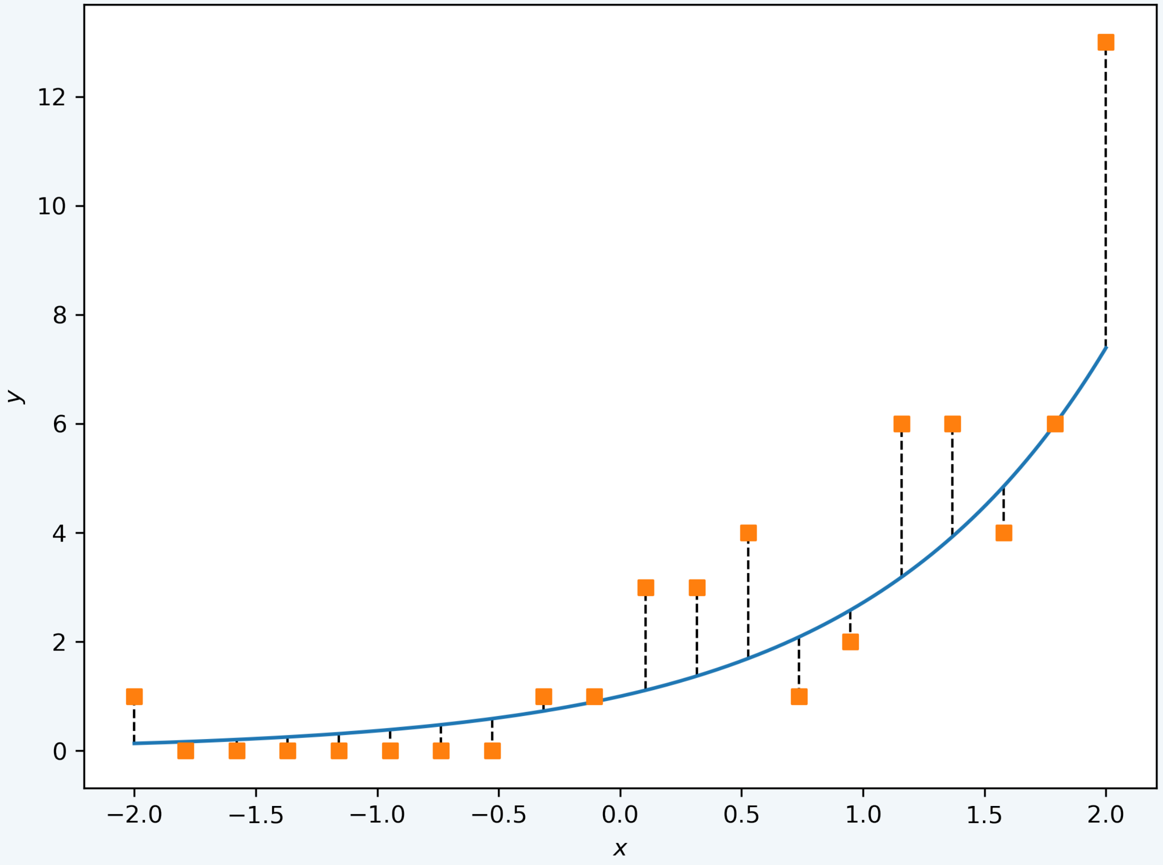
The Poisson regression model is

$$y \sim \text{Poisson}\left(\exp(X\beta)\right)$$

Here's the conditional expectation

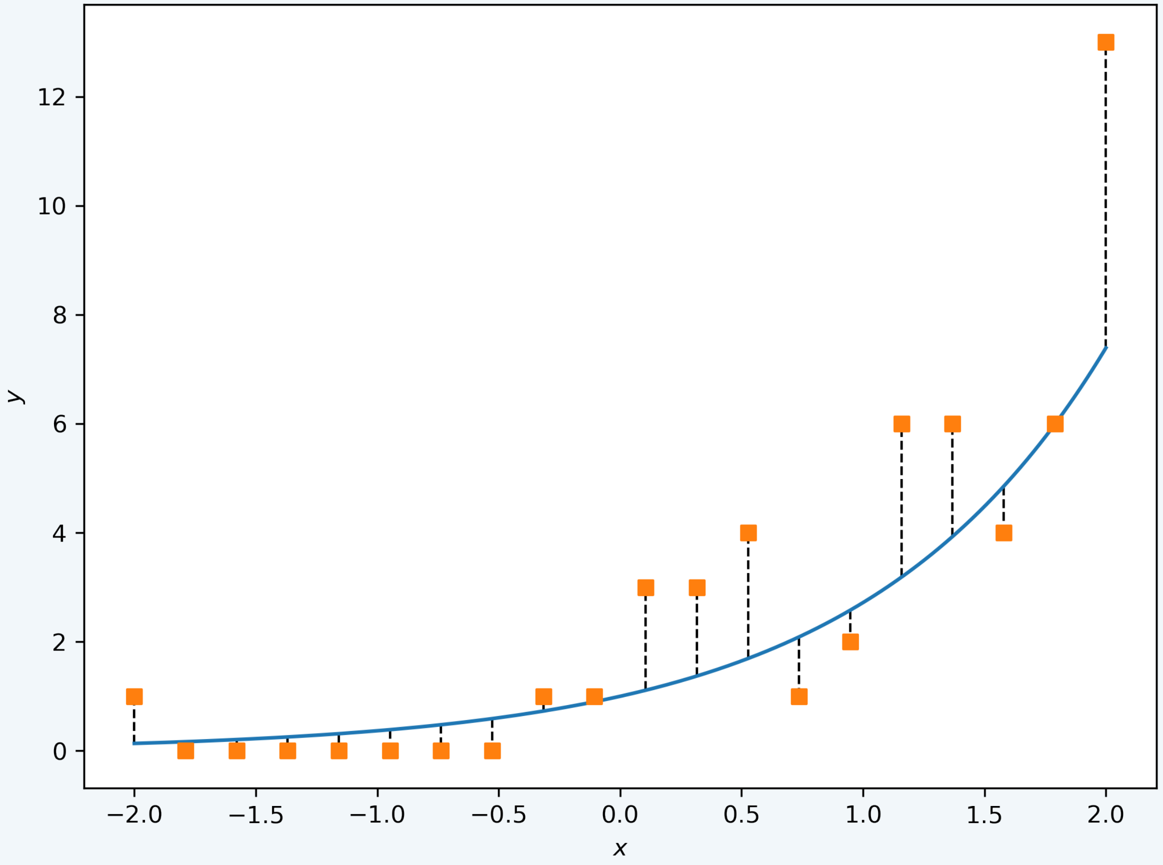
What about residuals?

These Aren't the Residuals You're Looking For



Why do we usually take "observed minus predicted"?

These Aren't the Residuals You're Looking For

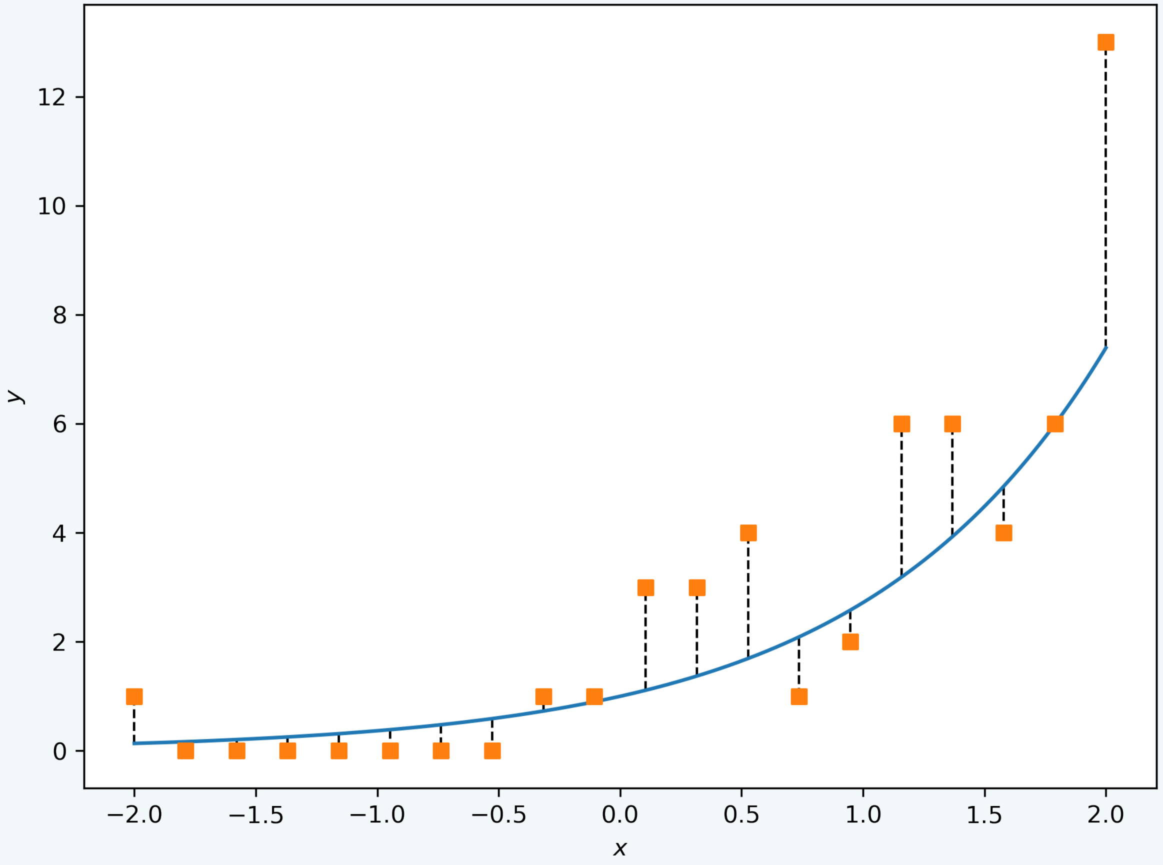


Why do we usually take "observed minus predicted"?

Likelihood for a normal looks like this:

$$L_{\text{Normal}} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

These Aren't the Residuals You're Looking For



Why do we usually take "observed minus predicted"?

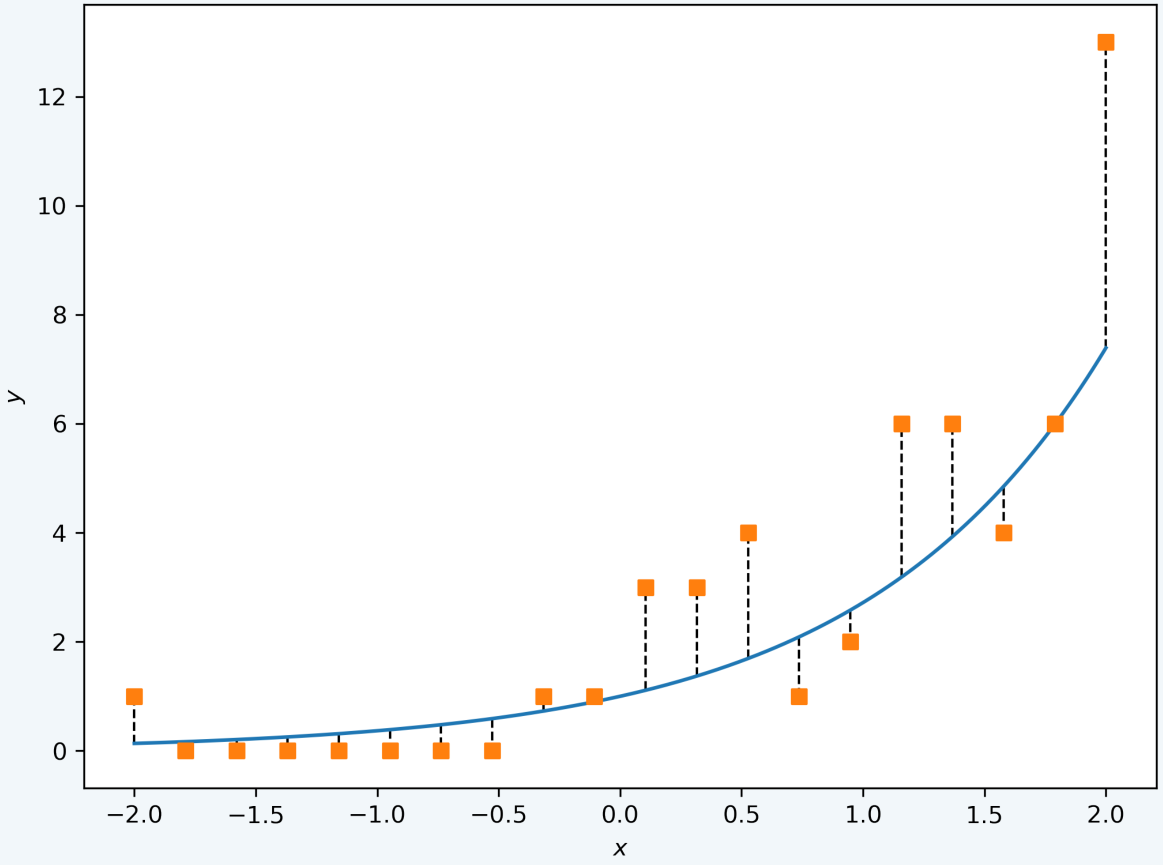
Likelihood for a normal looks like this:

$$L_{\text{Normal}} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

But here's a Poisson likelihood:

$$L_{\text{Poisson}} = \frac{\mu^y e^{-\mu}}{y!}$$

These Aren't the Residuals You're Looking For



Why do we usually take "observed minus predicted"?

Likelihood for a normal looks like this:

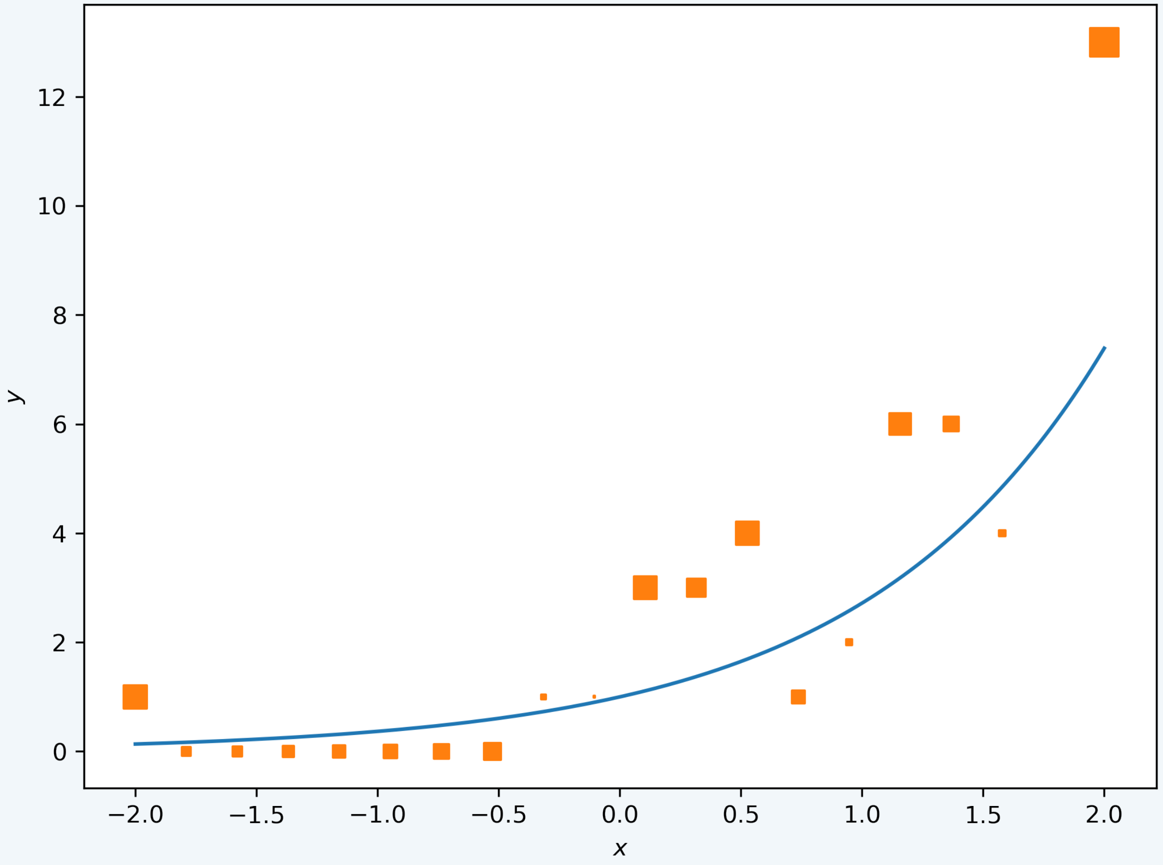
$$L_{\text{Normal}} = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\mu)^2}{2\sigma^2}\right)$$

But here's a Poisson likelihood:

$$L_{\text{Poisson}} = \frac{\mu^y e^{-\mu}}{y!}$$

"Residuals" will be different, but we can still calculate the contribution of each point to the overall cost

Unit Deviance



Here are the contributions of each point to the total cost

For a linear model, a perfectly-fit point has zero contribution (squared residual)

Here, we subtract cost for a hypothetical “saturated model” to make this work (more on this soon)

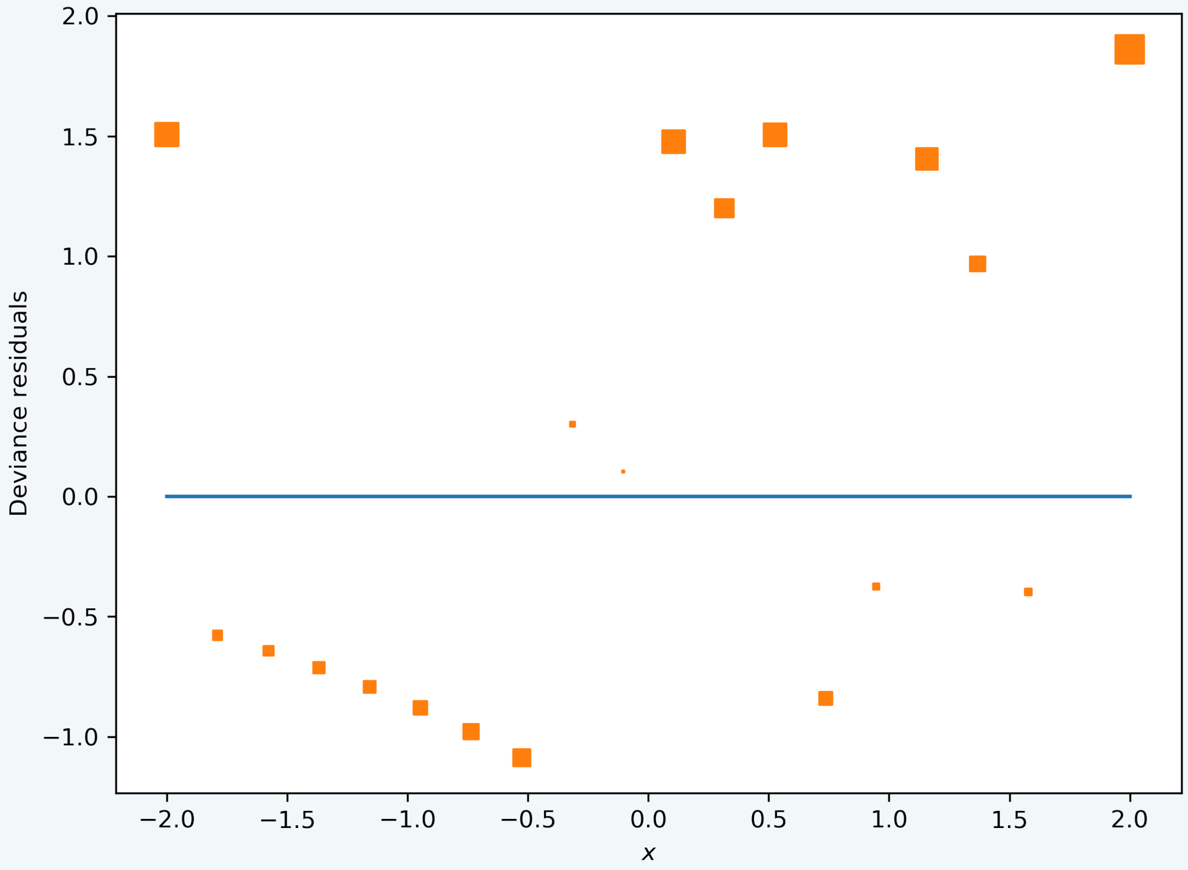
Resulting *unit deviance* values play the role of squared residuals

Remember the linear case:

- Squared residuals give contribution to cost
- Sign indicates data above or below prediction

Can we make that work out here? Spoiler: yes!

Deviance Residuals



Here's the result!

Use these *deviance residuals* just like you'd use residuals for a linear model

There can be some unavoidable patterns, like the "Hawaii" in the lower-left

Linear Models: Some more details

- For a normal linear model, the likelihood and log-likelihood for each data point are

$$L = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\hat{y})^2}{2\sigma^2}\right)$$

$$\ell = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\hat{y})^2}{2\sigma^2}$$

Linear Models: Some more details

- For a normal linear model, the likelihood and log-likelihood for each data point are

$$L = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\hat{y})^2}{2\sigma^2}\right)$$

$$\ell = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\hat{y})^2}{2\sigma^2}$$

- So the total log-likelihood is

$$\ell = -\frac{N}{2}\log(2\pi\sigma^2) \; - \; \frac{1}{2\sigma^2}\sum_{j=1}^{N}(y_j-\hat{y}_j)^2$$

Linear Models: Some more details

- For a normal linear model, the likelihood and log-likelihood for each data point are

$$L = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y-\hat{y})^2}{2\sigma^2}\right)$$

$$\ell = -\frac{1}{2}\log(2\pi\sigma^2) - \frac{(y-\hat{y})^2}{2\sigma^2}$$

- So the total log-likelihood is

$$\ell = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

- We've seen the sum of squared residuals, but what happened to the first term?

The Saturated Model, and Deviance

Imagine we could fit a model perfectly. Our log-likelihood

$$\ell = -\frac{N}{2}\log(2\pi\hat{\sigma}^2) - \frac{1}{2\hat{\sigma}^2}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2$$

would become

$$\ell_S = -\frac{N}{2}\log(2\pi\hat{\sigma}^2)$$

Deviance measures how far we are from a hypothetical perfect fit:

$$D = 2(\ell_S - \ell) = \frac{1}{\hat{\sigma}^2}\sum_{j=1}^{N}(y_j - \hat{y}_j)^2 = \sum_{j=1}^{N}\left(\frac{y_j - \hat{y}_j}{\hat{\sigma}}\right)^2$$

For a linear model, deviance is the sum of *Studentized* residuals!

GLM Hypothesis Testing

- For a linear model, deviance is

$$D = \frac{1}{\sigma^2} \sum_{j=1}^{N} (y_j - \hat{y}_j)^2$$

- If $\mathscr{M}_0$ is a “submodel” of $\mathscr{M}$ (so $\mathscr{M}_0 \subset \mathscr{M}$), which will fit training data better? $\mathscr{M}$
- What does this mean about the deviance? $D_0 > D$

- How do we measure “how much better” D fits? With a χ^2 test!

$$D_0 - D \sim \chi^2_{\Delta p}$$

“Degrees of freedom” is difference in number of parameters