Linear Regression

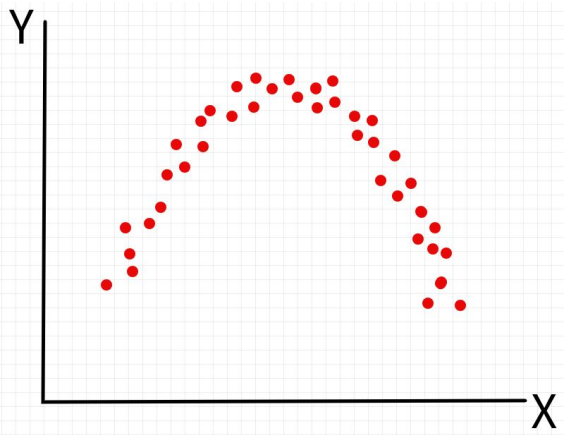
Boston University CS 506 - Lance Galletti

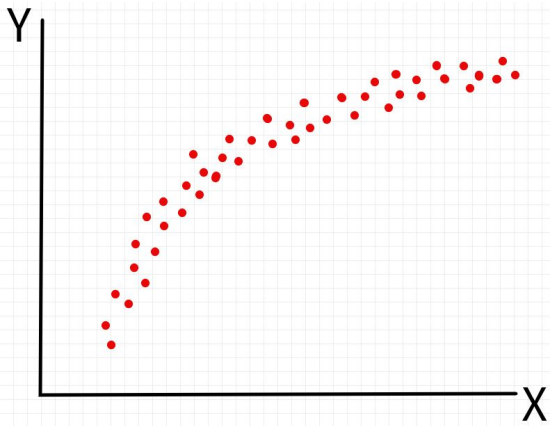
Challenge for those who have LR experience

- Find the data.csv file in the regression folder of our course repo
- Challenge:
  - Every day my alarm goes off at seemingly random times...
  - I've recorded the times for the past year of so (1 - 355 days)
  - Today is day 356
  - Can you predict when my alarm will ring?

Motivation

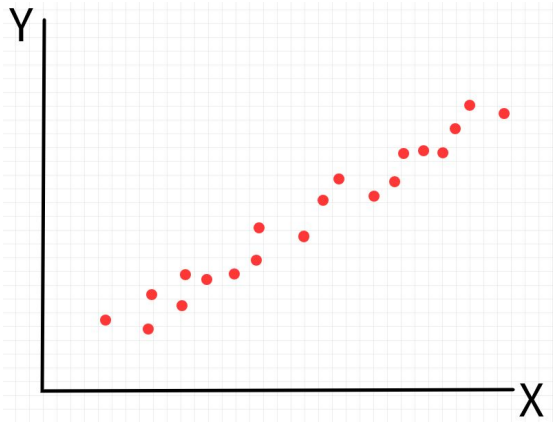
Given **n** samples / data points ($\mathbf{y_i}$, $\mathbf{x_i}$). Y is a continuous variable (as opposed to classification).

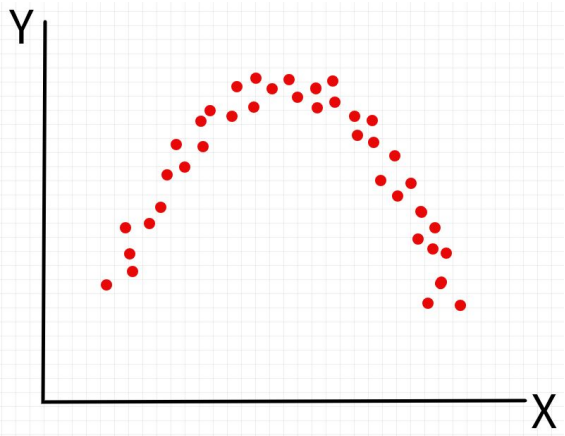


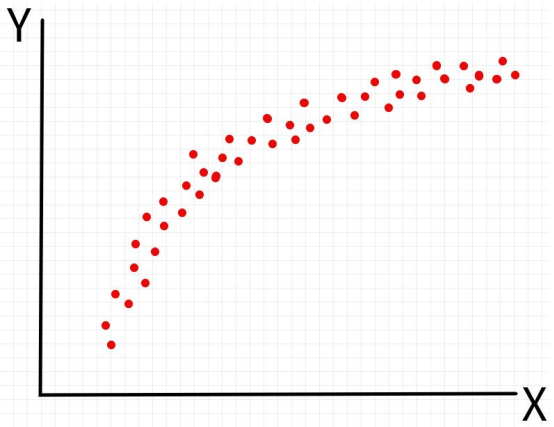


Motivation

Understand/explain how **y** varies as a function of **x** (i.e. find a function **y = h(x)** that best fits our data)







Motivation

Suppose we are given a curve **y = h(x)**, how can we evaluate whether it is a good fit to our data?

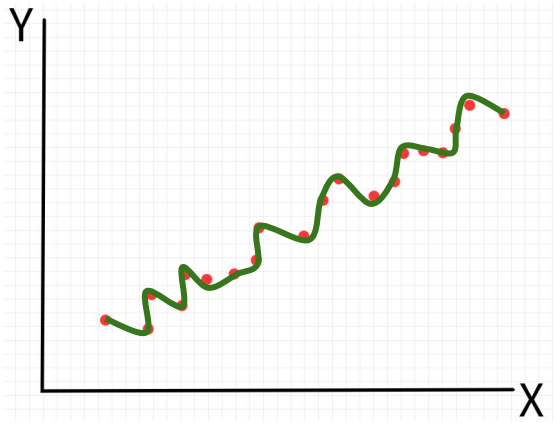
Compare $\mathbf{h(x_i)}$ to $\mathbf{y_i}$ for all **i**.

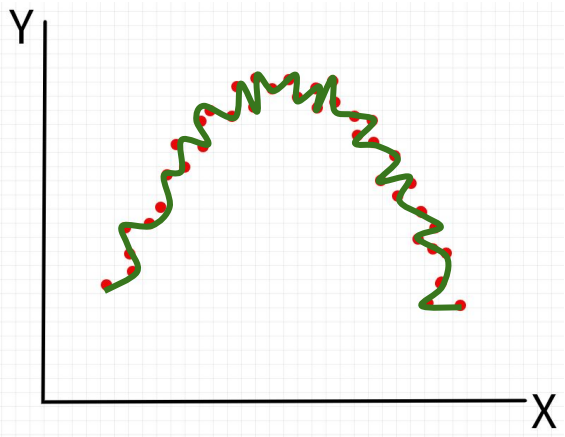
Goal: For a given distance function **d**, find **h** where **L** is smallest.

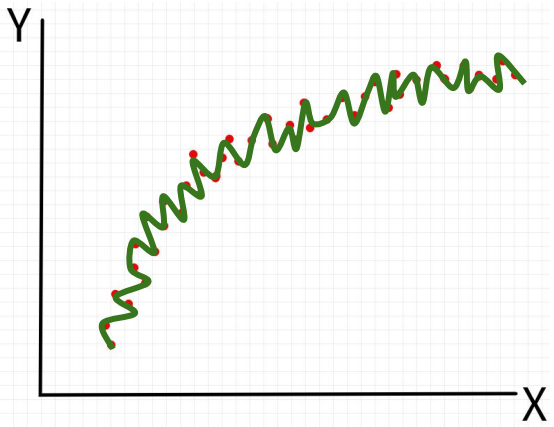
$$L(h) = \sum_i d(h(x_i), y_i)$$

Motivation

Should **h** be the curve that goes through the most samples? I.e. do we want $\mathbf{h(x_i) = y_i}$ for the maximum number of **i**?



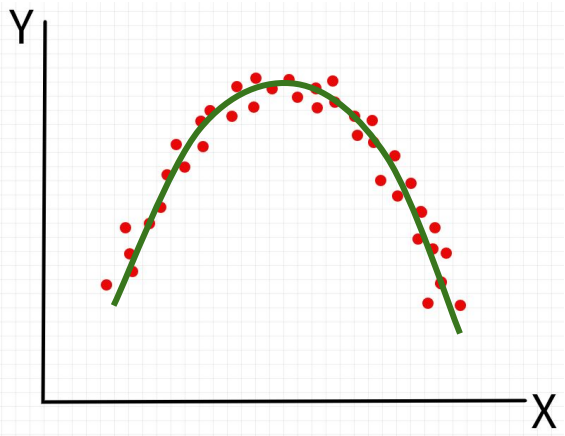




h may be too complex
overfitting - may not perform well on unseen data

Motivation

The following curves seem the most intuitive “best fit” to our samples. How can we define this best fit mathematically? Is it just about finding the right distance function?



Motivation

Another way to define this problem is in terms of probability.

Define **P(Y | h)** as the probability of observing **Y** given that it was sampled from **h**.

Goal: Find **h** that maximizes the probability of having observed our data.

Motivation

To sum up we can either:

1. Minimize

$$L(h) = \sum_i d(h(x_i), y_i)$$

1. Maximize

L(h) = P(Y | h)

Getting Started

Do we have enough to get started?

Seems like there are too many possible **h** and our problem statements are still too vague to effectively find solutions.

What can we do to constrain the problem?

Let’s make some assumptions!

Assumptions

Let's start by assuming our data was generated by a **linear function** plus some **noise**:

$$\vec{y} = h_\beta(\mathrm{X}) + \vec{\epsilon}$$

Where **h** is linear in a parameter $\boldsymbol{\beta}$.

Which functions below are linear in $\boldsymbol{\beta}$?

$\mathbf{h}(\beta) = \boldsymbol{\beta}_1 \mathbf{x}$ ✔

$\mathbf{h}(\beta) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{x}$ ✔

$\mathbf{h}(\beta) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{x} + \boldsymbol{\beta}_2 \mathbf{x}^2$ ✔

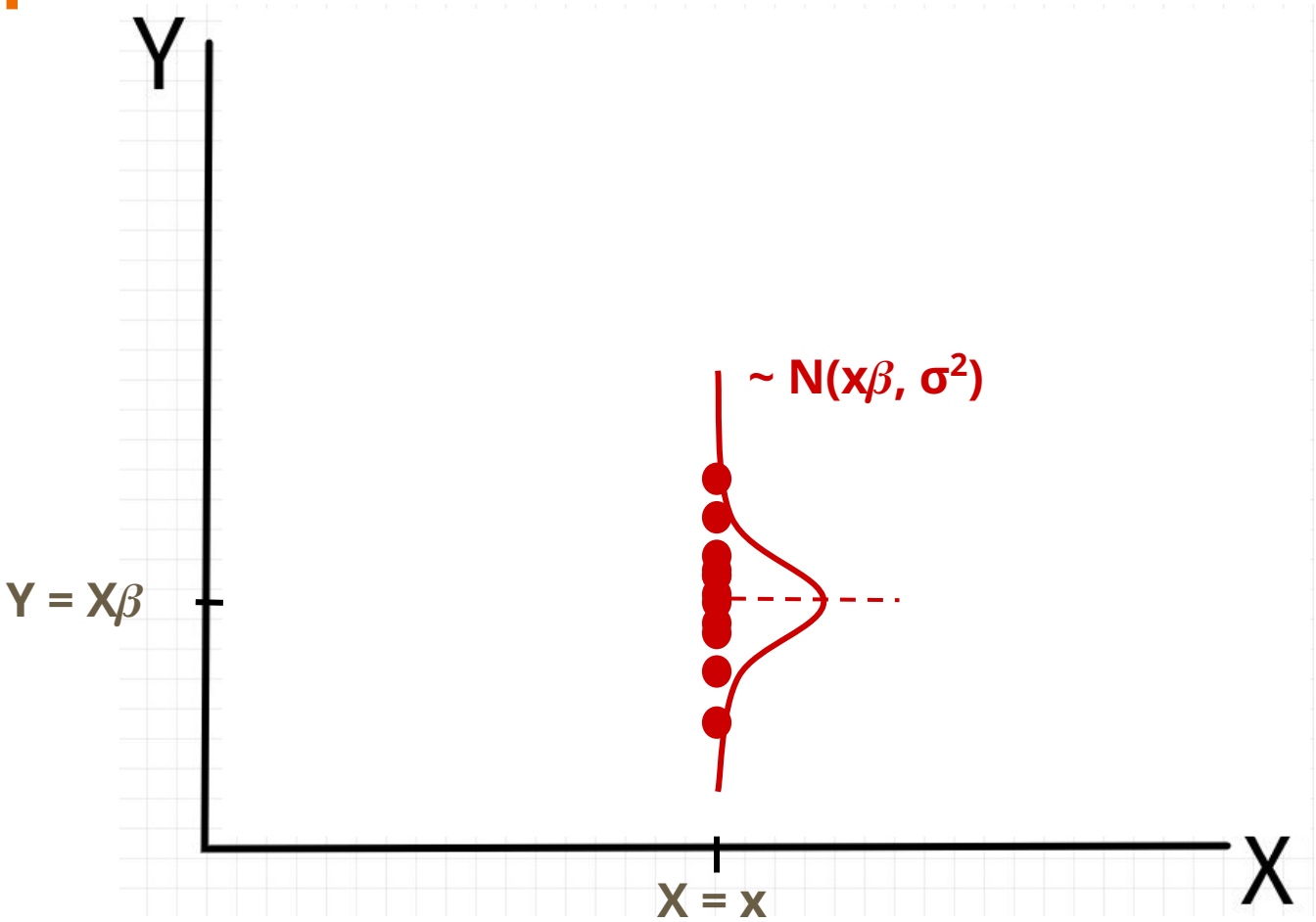
$\mathbf{h}(\beta) = \boldsymbol{\beta}_1 \mathbf{log(x)} + \boldsymbol{\beta}_2 \mathbf{x}^2$ ✔

$\mathbf{h}(\beta) = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{x} + \boldsymbol{\beta}_1^2 \mathbf{x}$ ✘

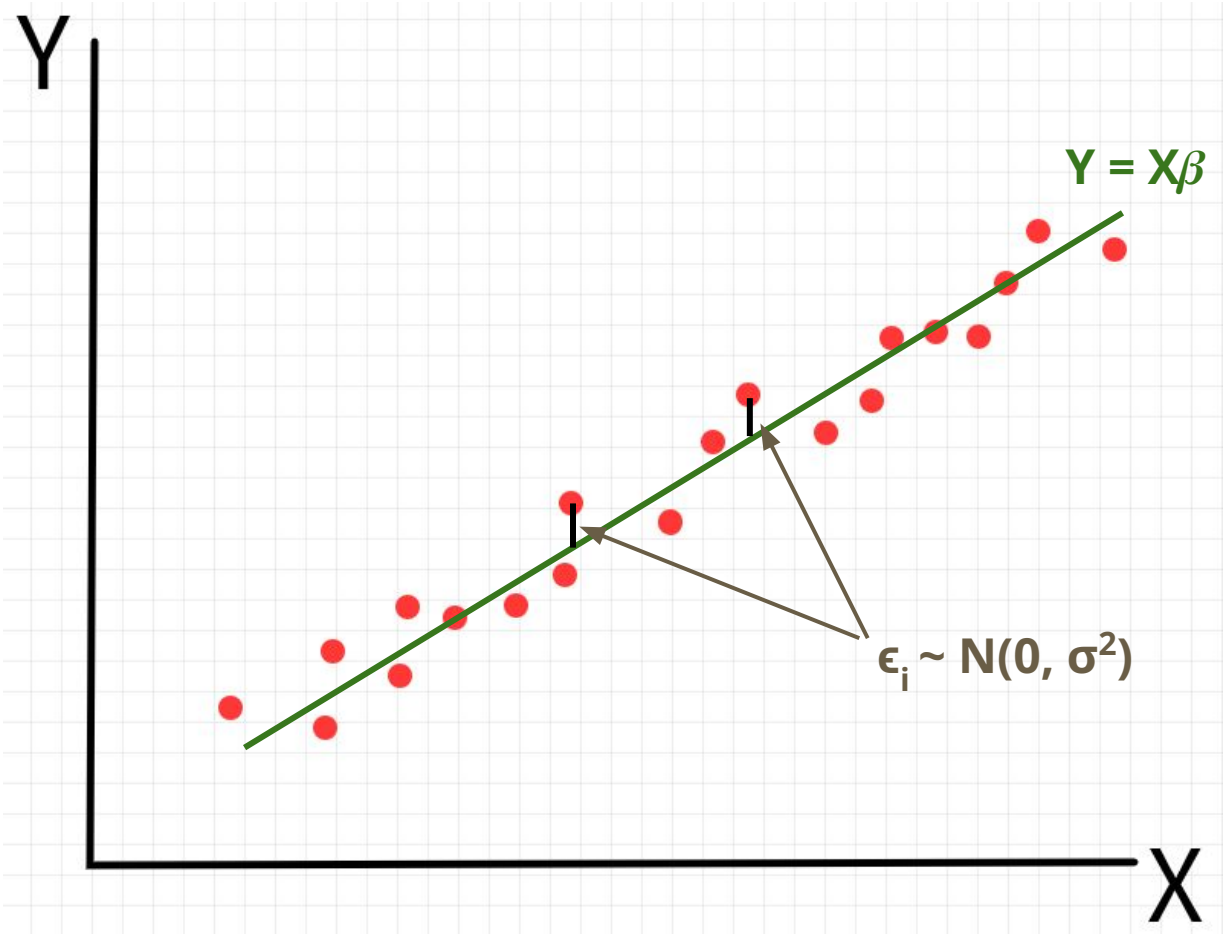
Assumptions

1. The relation between **x** (independent variable) and **y** (dependent variable) is linear in a parameter $\boldsymbol{\beta}$.
2. $\boldsymbol{\epsilon_i}$ are independent, identically distributed random variables following a $\mathbf{N(0, \sigma^2)}$ distribution. (Note: $\boldsymbol{\sigma}$ is constant)

Assumptions



Assumptions



Goal

Given these assumptions, let's try to solve the max and min problems we defined earlier!

Q: What does solving these mean?

A: Finding $\boldsymbol{\beta}$ is equivalent to finding **h**

Least Squares

$$\beta_{LS} = \arg\min_{\beta} \sum_{i} d(h_\beta(x_i), y_i)$$

$$= \arg\min_{\beta} \|\vec{y} - h_\beta(\mathrm{X})\|_2^2$$

$$= \arg\min_{\beta} \|y - X\beta\|_2^2$$

Least Squares

$$\frac{\partial}{\partial \beta} = 0$$

$$\frac{\partial}{\partial \beta}(y - X\beta)^T(y - X\beta) = 0$$

$$\frac{\partial}{\partial \beta}(y^T y - y^T X\beta - \beta^T X^T y - \beta^T X^T X\beta) = 0$$

$$\frac{\partial}{\partial \beta}(y^T y - 2\beta^T X^T y - \beta^T X^T X\beta) = 0$$

$$-2X^T y - X^T X\beta = 0$$

$$X^T X\beta = X^T y$$

$$\boxed{\beta_{LS} = (X^T X)^{-1} X^T y}$$

Maximum Likelihood

Since **$\epsilon \sim N(0, \sigma^2)$** and **$Y = X\beta + \epsilon$** then **$Y \sim N(X\beta, \sigma^2)$**.

$$
\begin{aligned}
\beta_{MLE} &= \arg\max_{\beta} \frac{1}{\sqrt{(2\pi)^n \sigma^n}} \exp\left(-\frac{\|y - X\beta\|_2^2}{2\sigma^2}\right) \\
&= \arg\max_{\beta} \exp\left(-\|y - X\beta\|_2^2\right) \\
&= \arg\max_{\beta} -\|y - X\beta\|_2^2 \\
&= \arg\min_{\beta} \|y - X\beta\|_2^2 \\
&= \beta_{LS} = (X^T X)^{-1} X^T y
\end{aligned}
$$

An Unbiased Estimator

$\boldsymbol{\beta}_{\mathbf{LS}}$ is an unbiased estimator of the true $\boldsymbol{\beta}$. That is $\mathbf{E}[\boldsymbol{\beta}_{\mathbf{LS}}]=\boldsymbol{\beta}$.

$$
\begin{aligned}
E[\beta_{LS}] &= E[(X^T X)^{-1} X^T y] \\
&= (X^T X)^{-1} X^T E[y] \\
&= (X^T X)^{-1} X^T E[X\beta + \epsilon] \\
&= (X^T X)^{-1} X^T X\beta + E[\epsilon] \\
&= \beta
\end{aligned}
$$

Demo

Logistic Regression

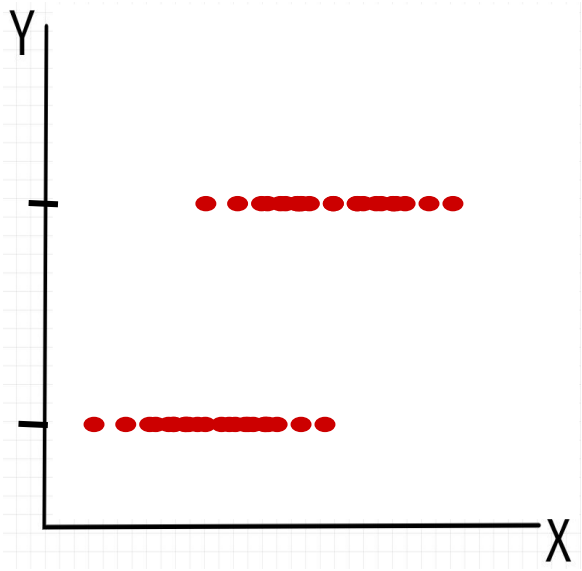
So far y_i was a continuous variable. What if y_i is categorical? Can we use a linear function to predict y_i?

Assume we have **2 classes**.

Logistic Regression

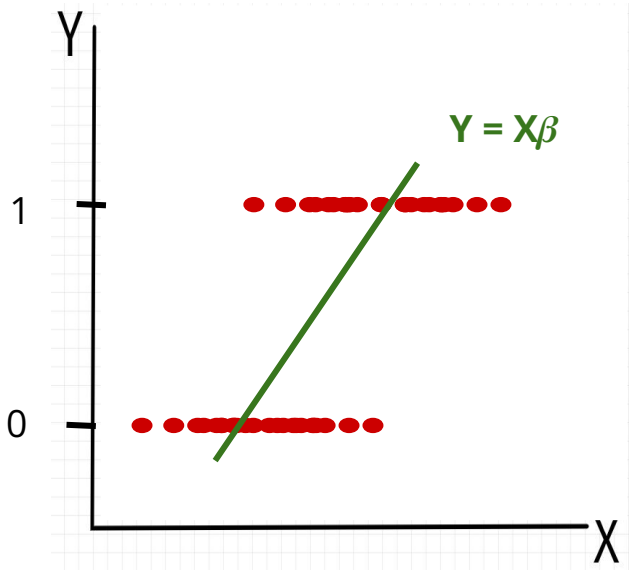
So far y_i was a continuous variable. What if y_i is categorical? Can we use a linear function to predict y_i?

Assume we have **2 classes**.



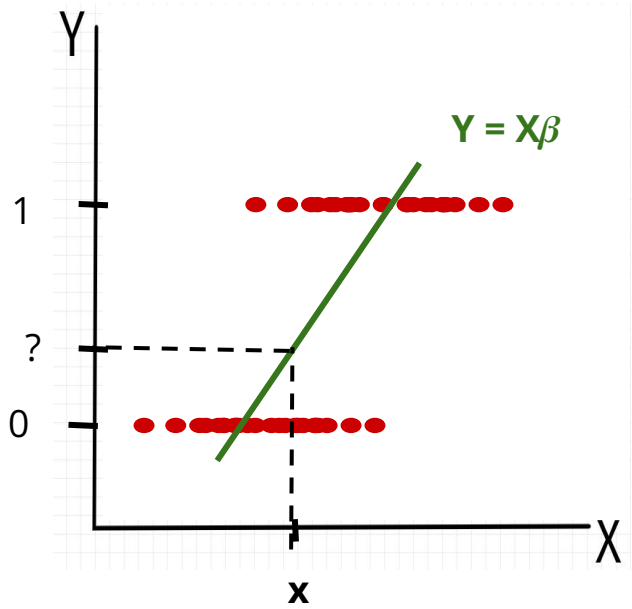
Logistic Regression

What will a linear model look like?



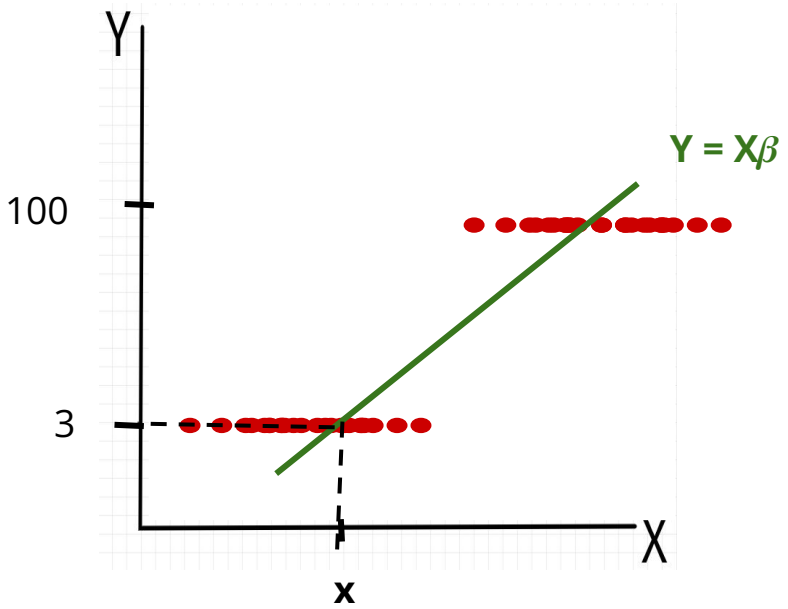
Logistic Regression

What will a linear model look like?



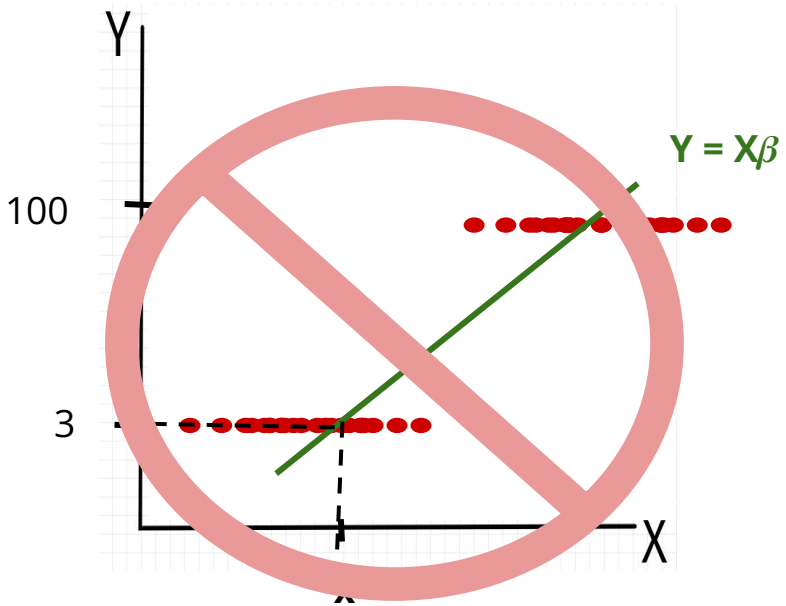
Logistic Regression

What if the numerical values of the classes change?



Logistic Regression

What if the numerical values of the classes change?



Logistic Regression

The numerical values associated with the class are arbitrary. A model based on these numbers would be meaningless...

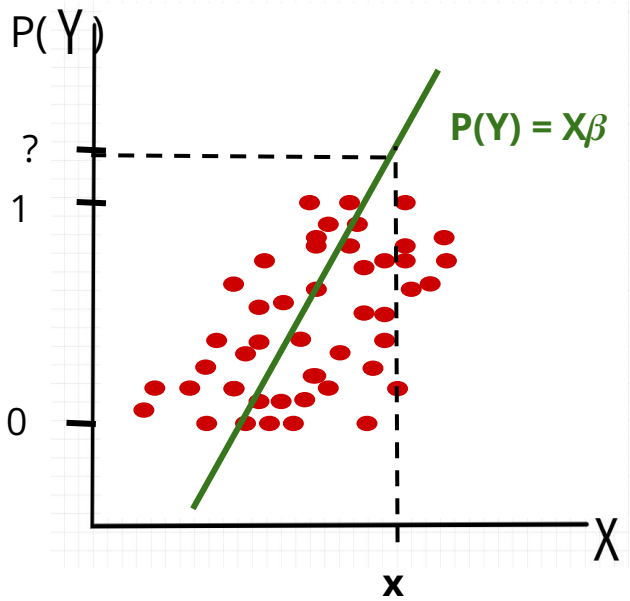
So we probably shouldn't try to predict the class itself.

Logistic Regression

Notice that a linear function will predict a **continuum** of values. So we should find an interpretation / transformation of the class that is **continuous** for us to predict.

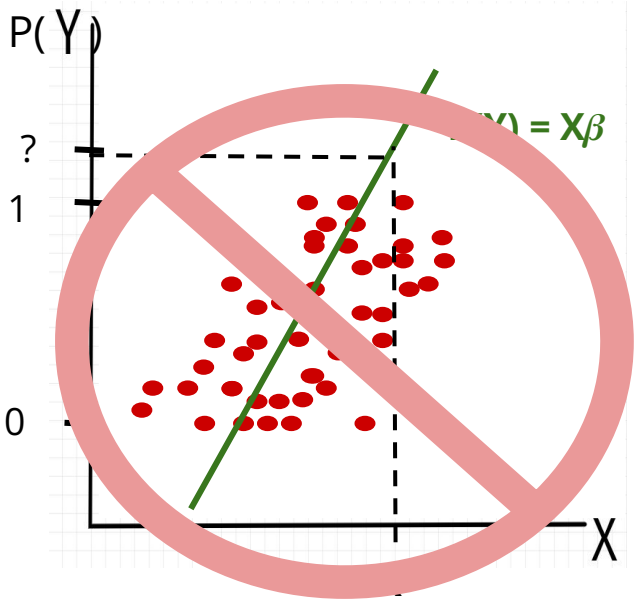
Logistic Regression

Can we use the probability of belonging to a given class as a proxy for how confidently we can classify a given point?



Logistic Regression

Can we use the probability of belonging to a given class as a proxy for how confidently we can classify a given point?



Logistic Regression

So it's not just a continuum of values - the range of values needs to be (-∞, ∞)!

Define the odds = p / 1 - p where p = P(Y = class 1 | X)

Now the range of $X\beta_{LS}$ is [0, ∞)

In order to get (-∞, ∞), let's take the log of the odds! This is also convenient numerically because in the odds format, tiny variations in p have large effects on the odds!

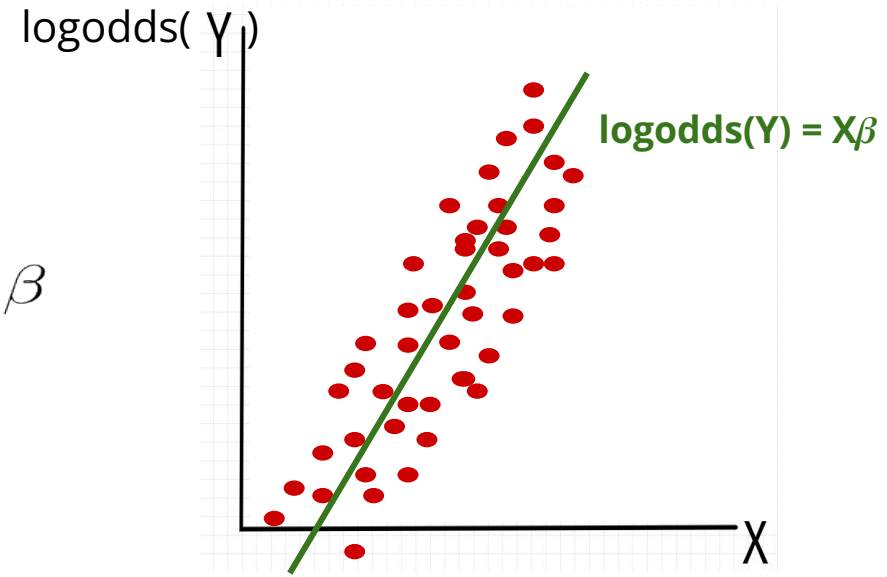
Logistic Regression

Our goal is to fit a linear model to **the log-odds of being in one of our classes** (in the 2-class case) i.e.

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = X\beta$$

Logistic Regression

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = X\beta$$



Logistic Regression

Creates a linear decision boundary between the classes. Why?

The decision boundary is where the probability is ½ (for binary classification).

Logistic Regression

Suppose we have such a model. How do we recover the P(Y=1 | X)?

$$\log\left(\frac{P(Y=1|X)}{1-P(Y=1|X)}\right) = \alpha + \beta X$$

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\alpha+\beta X}$$

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} + 1 = e^{\alpha+\beta X} + 1$$

$$\frac{P(Y=1|X)}{1-P(Y=1|X)} = e^{\alpha+\beta X} + 1$$

$$P(Y=1|X) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$

The function we apply to our probability to obtain the log odds is called the **logit** function. The function used to retrieve our probability from the log odds is called **logit^{-1}**

Logistic Regression

The probability is ½ only when $X\beta = 0$ which is the equation of a line.

Logistic Regression

How do we learn our model? I.e. the α and β parameters.

We know:

$$P(y_i|x_i) = \begin{cases} logit^{-1}(\alpha + \beta x_i) \text{ if } y_i = 1 \\ 1 - logit^{-1}(\alpha + \beta x_i) \text{ if } y_i = 0 \end{cases}$$
$$= (logit^{-1}(\alpha + \beta x_i))^{y_i}(1 - logit^{-1}(\alpha + \beta x_i))^{1-y_i}$$

Logistic Regression

So we can define the probability of having seen the data we saw:

$$L(\alpha, \beta) = \prod_i (logit^{-1}(\alpha + \beta x_i))^{y_i} (1 - logit^{-1}(\alpha + \beta x_i))^{1-y_i}$$

And try to maximize this quantity!

Unfortunately, there is no closed form solution here and we need to use numerical approximation methods to solve this optimization problem - we will talk about these soon!

Demo

Evaluating Our Regression Model

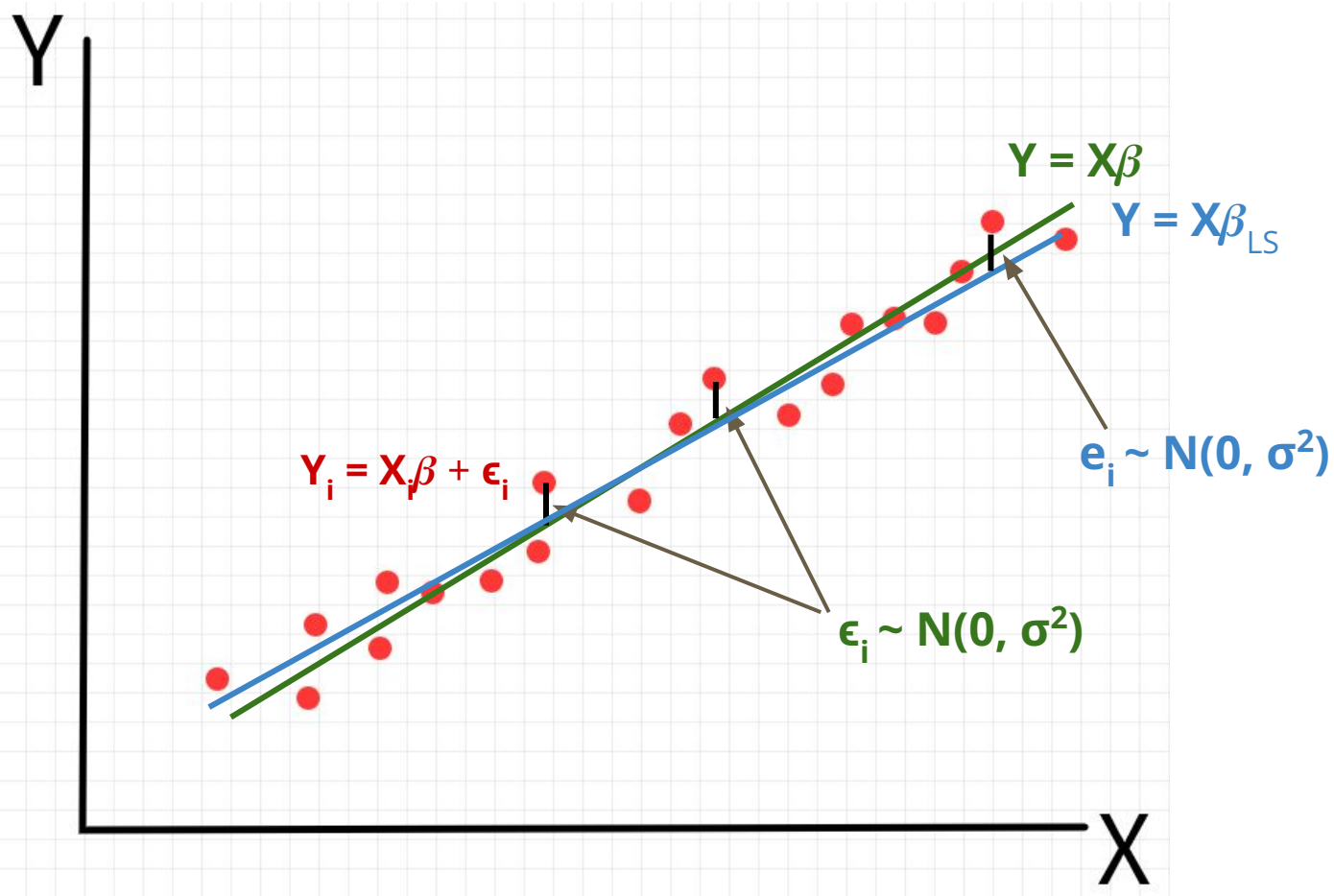
Some Notation:

$\mathbf{y_i}$ is the “true” value from our data set (i.e. $\mathbf{x_i}\boldsymbol{\beta} + \boldsymbol{\epsilon_i}$)

$\hat{\mathbf{y}}_\mathbf{i}$ is the estimate of y_i from our model (i.e. $\mathbf{x_i}\boldsymbol{\beta_{LS}}$)

$\bar{\mathbf{y}}$ is the sample mean all $\mathbf{y_i}$

$\mathbf{y_i} - \hat{\mathbf{y}}_\mathbf{i}$ are the estimates of $\boldsymbol{\epsilon_i}$ and are referred to as residuals


Y
X
$Y = X\beta$
$Y = X\beta_{LS}$
$Y_i = X_i\beta + \epsilon_i$
$e_i \sim N(0, \sigma^2)$
$\epsilon_i \sim N(0, \sigma^2)$

Evaluating Our Regression Model

$$TSS = \sum_{i}^{n} (y_i - \bar{y})^2$$

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{RSS}{TSS}$$

$$RSS = \sum_{i}^{n} (y_i - \hat{y}_i)^2$$

This is what our linear model is minimizing

$$ESS = \sum_{i}^{n} (\hat{y}_i - \bar{y})^2$$

R^2 measures the fraction of variance that is explained by ŷ

Exercise

Show that TSS = ESS + RSS

$$\begin{aligned}
\text{TSS} &= \sum_i (y_i - \bar{y})^2 \\
&= \sum_i (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2 \\
&= \sum_i (y_i - \hat{y}_i)^2 + \sum_i (\hat{y}_i - \bar{y})^2 + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) \\
&= \text{ESS} + \text{RSS} + 2\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}).
\end{aligned}$$

Assume for simplicity that $\hat{y}_i = \beta_0 + \beta_1 x_i$
Since β_0 and β_1 are least squares estimates, we know they minimize

$$\sum_i (y_i - \hat{y}_i)^2$$

By taking derivatives of the above with respect to β_0 and β_1 we discover that

$$\sum_i (y_i - \hat{y}_i) = 0 \text{ and } \sum_i (y_i - \hat{y}_i) x_i = 0$$

$$\sum_i (y_i - \hat{y}_i)(\hat{y}_i - \bar{y}) = \sum_i (y_i - \hat{y}_i)\hat{y}_i - \bar{y}\sum_i (y_i - \hat{y}_i)$$

$$= \hat{\beta}_0 \sum_i (y_i - \hat{y}_i) + \hat{\beta}_1 \sum_i (y_i - \hat{y}_i) x_i - \bar{y}\sum_i (y_i - \hat{y}_i)$$

Evaluating our Regression Model

OLS Regression Results

Dep. Variable:	y	R-squared:	0.840
Model:	OLS	Adj. R-squared:	0.836
Method:	Least Squares	F-statistic:	254.1
Date:	Sun, 20 Mar 2022	Prob (F-statistic):	2.72e-39
Time:	11:36:16	Log-Likelihood:	-482.37
No. Observations:	100	AIC:	970.7
Df Residuals:	97	BIC:	978.5
Df Model:	2		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272

Omnibus:	1.279	Durbin-Watson:	1.824
Prob(Omnibus):	0.527	Jarque-Bera (JB):	1.065
Skew:	0.253	Prob(JB):	0.587
Kurtosis:	2.999	Cond. No.	1.38

Evaluating our Regression Model

Each parameter of an independent variable **x** has an associated confidence interval and t-value + p-value.

If the parameter / coefficient is not significantly distinguishable from 0 then we cannot assume that there is a significant linear relationship between that independent variable and the observations **y** (i.e. if the interval includes 0 or if the p-value is too large)

Hypothesis Test

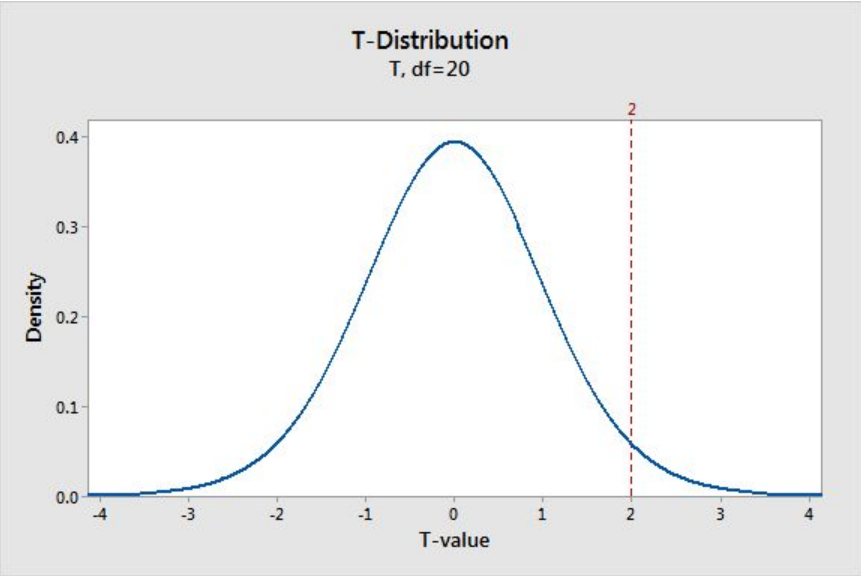
We want to know if there is evidence to reject the hypothesis $H0 : \beta = 0$ (i.e. that there is no linear relation between X and Y) using the information from β hat.

We want to know the largest probability of obtaining the data observed, under the assumption that the null hypothesis is correct.

How do we obtain that probability?

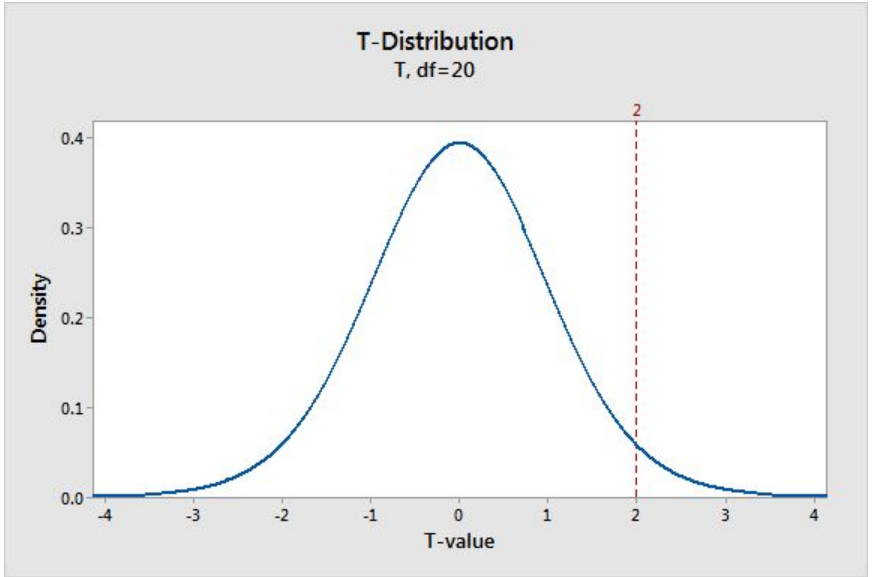
Hypothesis Test

Under the null hypothesis what should be the distribution of the estimates?
T-distribution (parametrized by the sample size)



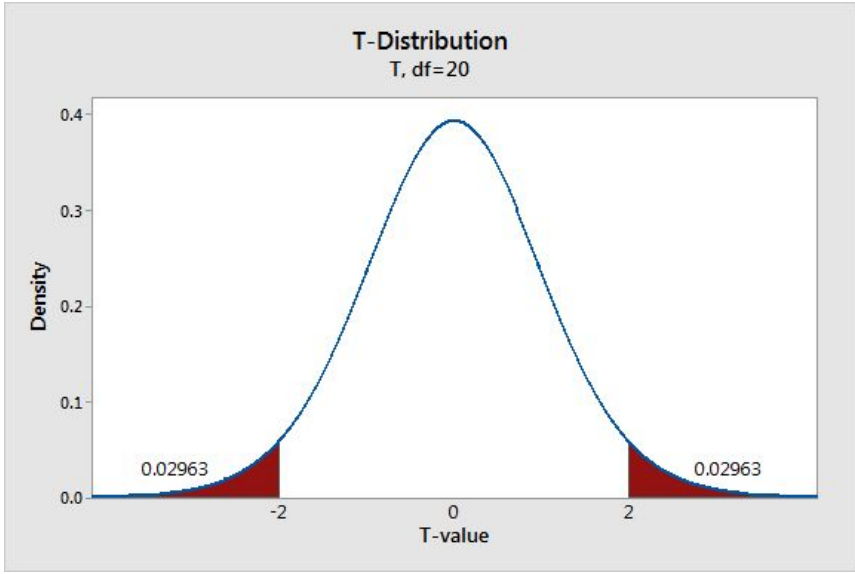
Hypothesis Test

We can then compute the t-value that corresponds to the sample we observed.



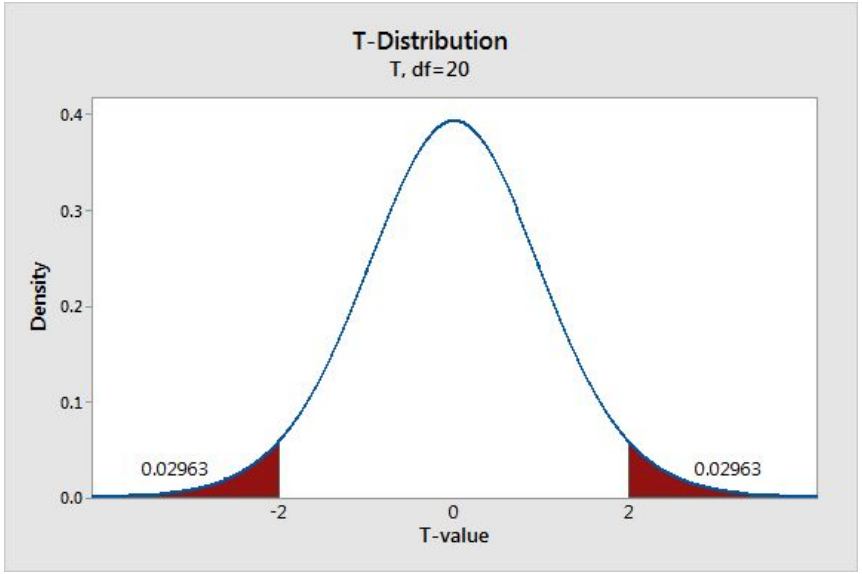
Hypothesis Test

And then compute the probability of observing estimates of β at least as extreme as the one observed. (i.e. trying to find evidence against H0)



Hypothesis Test

This probability is called a p-value.



Hypothesis Test

A p-value smaller than a given threshold would mean the data was unlikely to be observed under H0 so we can reject the hypothesis H0. If not, then we lack the evidence to reject H0.

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272

Hypothesis Test

Which parameters should we not include in our linear model?

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272

Confidence Intervals

An interval that describes the uncertainty around an estimate (here this could be β hat).

Goal: for a given confidence level (let's say 90%), construct an interval around an estimate such that, if the estimation process were repeated indefinitely, the interval would contain the true value (that the estimate is estimating) 90% of the time.

	coef	std err	t	P>|t|	[0.025	0.975]
const	2.1912	3.162	0.693	0.490	-4.085	8.467
x1	29.3912	3.274	8.977	0.000	22.893	35.889
x2	78.1391	3.594	21.741	0.000	71.006	85.272

Z-values

These are the number of standard deviations from the mean of a N(0,1) distribution required in order to contain a specific % of values were you to sample a large number of times.

To find the .95 z-value (the value z such that 95% of the observations lie within z standard deviations of the mean (μ ± z * σ)) you need to solve:

$$\int_{-z}^{z} \frac{1}{2\pi} e^{-\frac{1}{2}x^2} dx = .95$$

Z-values

The .95 z-value is 1.96.

This means 95% of observations from a N(μ, σ) lie within 1.96 standard deviations of the mean ($\mu \pm 1.960 * \sigma$)

If we get a sample from a N(μ, σ) of size n, how would we create a confidence interval around the estimated mean?

Confidence Intervals

How do we build a confidence interval?

Assume $\mathbf{Y_i \sim N(5, 25)}$, for $\mathbf{1 \leq i \leq 100}$ and $\mathbf{y_i = \mu + \epsilon}$ where $\boldsymbol{\epsilon \sim N(0, 25)}$. Then the Least Squares estimator of $\boldsymbol{\mu}$ ($\boldsymbol{\mu_{LS}}$) is

the sample mean $\bar{\mathbf{y}}$

What is the 95% confidence interval for $\boldsymbol{\mu_{LS}}$?

$$\mathbf{SE(\mu_{LS}) = \sigma_\epsilon / \sqrt{n}}$$
$$\mathbf{= 5 / \sqrt{100}}$$
$$\mathbf{= .5}$$

$$\mathbf{CI_{.95} = [\bar{y} - 1.96 \times SE(\mu_{LS}), \bar{y} + 1.96 \times SE(\mu_{LS})]}$$
$$\mathbf{= [\bar{y} - 1.96 \times .5, \bar{y} + 1.96 \times .5]}$$

Z-value for 95% Confidence Interval

Checking our Assumptions

1. Normal Distribution?
2. Constant Variance?

QQ plot

Quantiles are the values for which a particular % of values are contained below it.

For example the 50% quantile of a N(0,1) distribution is 0 since 50% of samples would be contained below 0 were you to sample a large number of times.

QQ plot

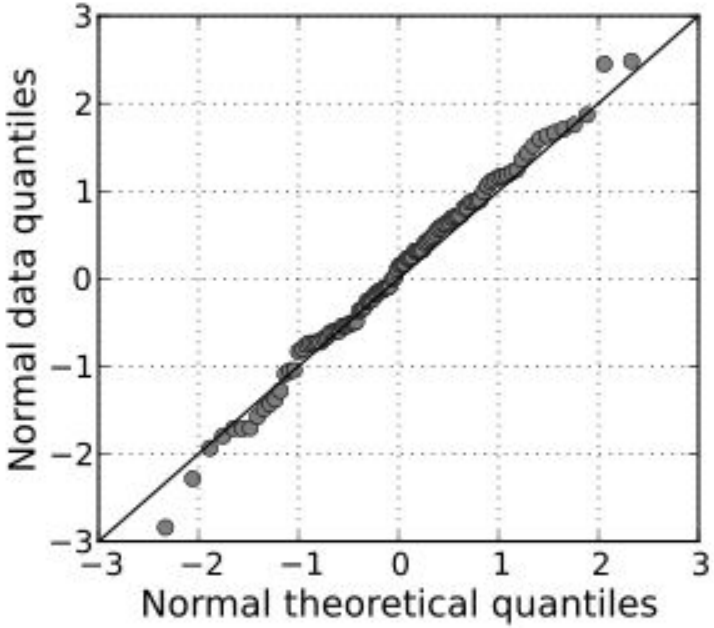
We need to check our assumption that our residuals / noise estimates are normally distributed.

How do can you check that a variable follows a specific distribution?

Need to check that our variable is **distributed** in the same way that a variable following our target distribution would be.

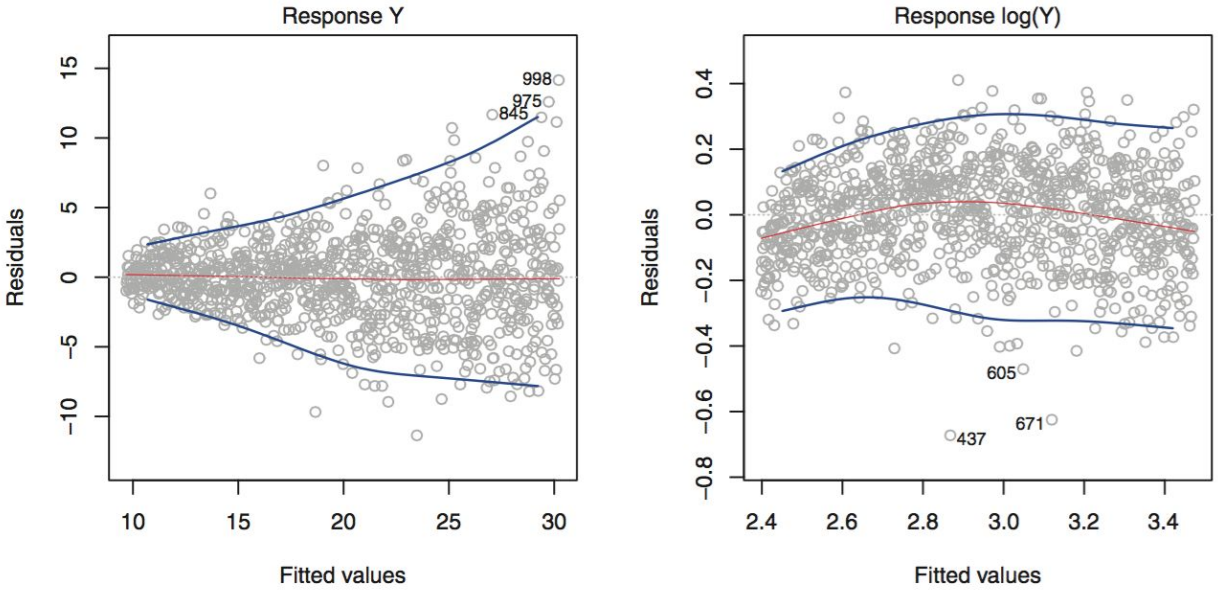
Plot the quantile of your target distribution against the quantiles of your data/ variable! If they match then your data probably comes from that distribution.

QQ plot


Normal data quantiles
Normal theoretical quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

Constant Variance

One of our assumptions was that our noise had constant variance. How can we verify this?
We can plot our fitted values against our residuals (noise estimates)



Extending our Linear Model

Changing the assumptions we made can drastically change the problem we are solving. A few ways to extend the linear model:

1. Non-constant variance - used in WLS (weighted least squares)
2. Distribution of error is not Normal - used in GLM (generalized linear models)