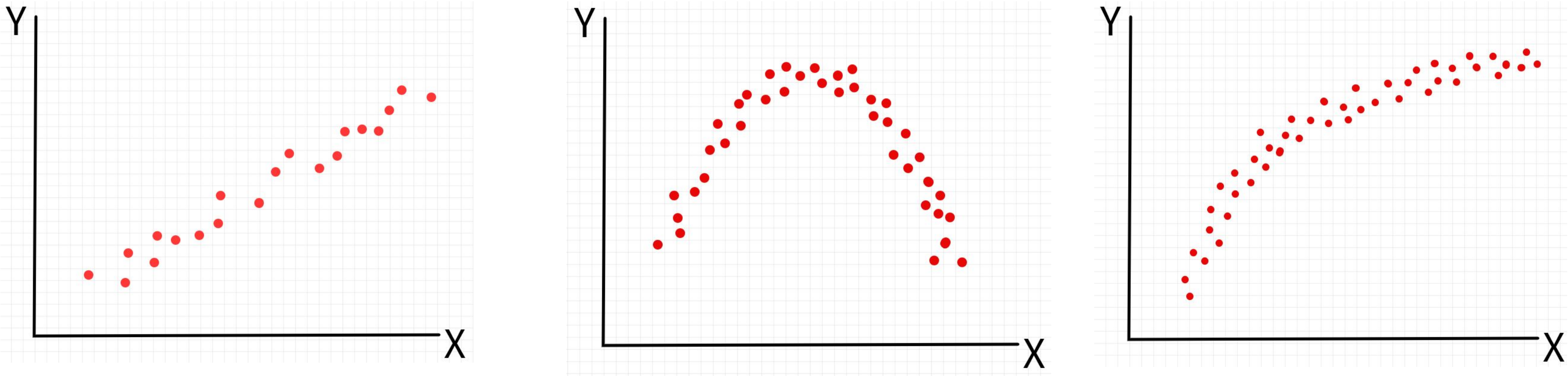
Linear Regression

Boston University CS 506 - Lance Galletti

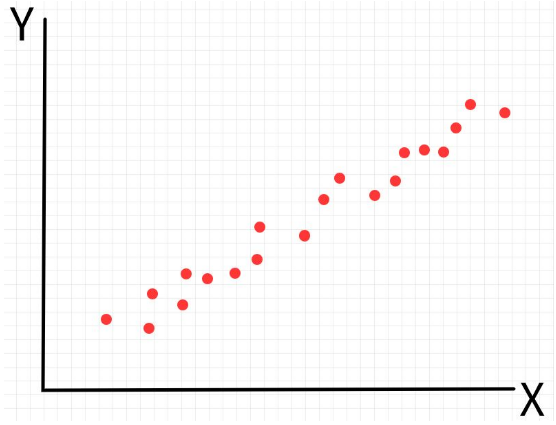
Motivation

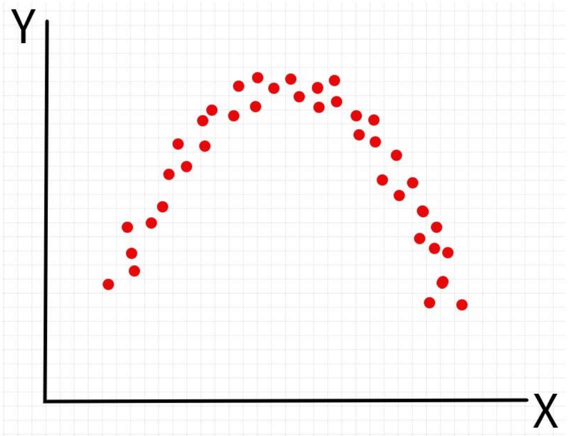
Given **n** samples / data points ($\mathbf{y}_i$, $\mathbf{x}_i$)

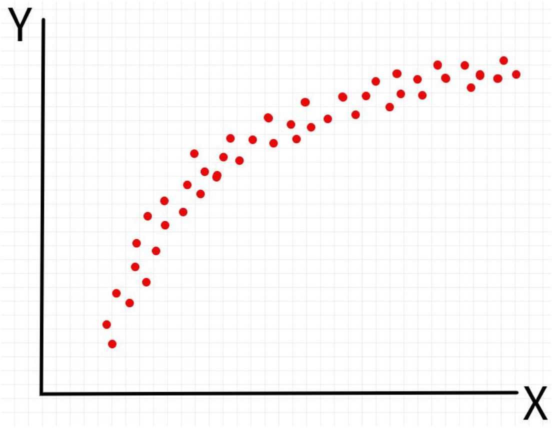


Motivation

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?

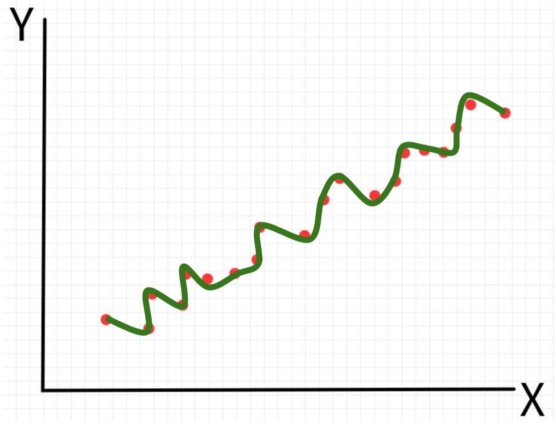
Compare $\mathbf{h(x_i)}$ to $\mathbf{y_i}$ for all **i**.

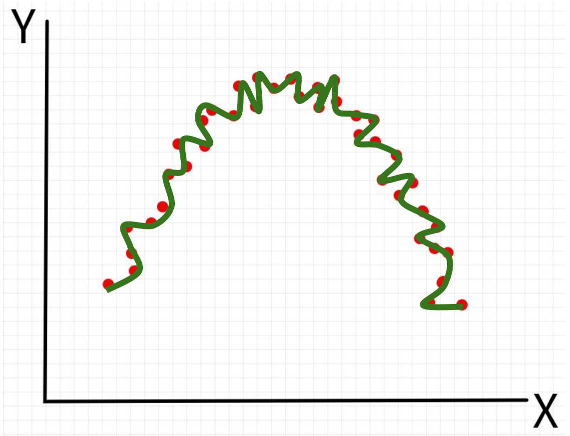
Goal: For a given distance function **d**, find **h** where **L** is smallest.

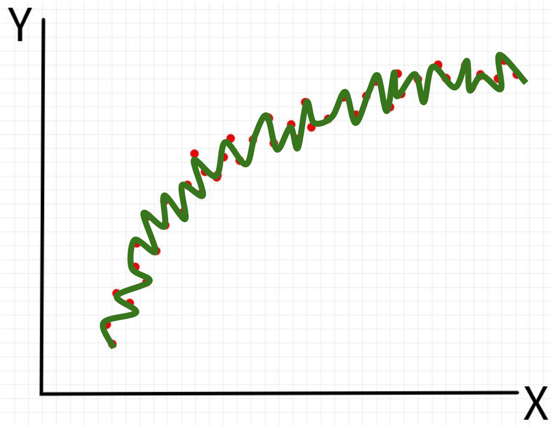
$$L(h) = \sum_i d(h(x_i), y_i)$$

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**?



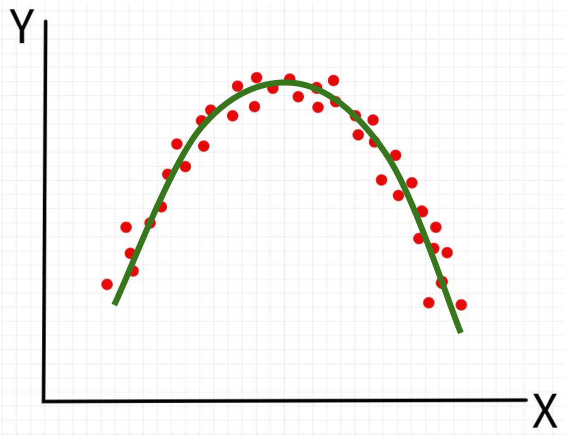


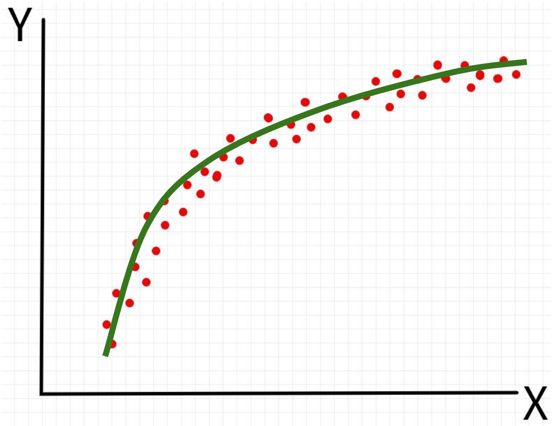


h may be too complex
overfitting - may not perform well on unseen data

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

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

2. Maximize

$$\mathbf{L(h) = P(Y \mid h)}$$

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 $\boldsymbol{\beta}$?

$\mathbf{h(x) = \beta_1 x}$ ✔

$\mathbf{h(x) = \beta_0 + \beta_1 x}$ ✔

$\mathbf{h(x) = \beta_0 + \beta_1 x + \beta_2 x^2}$ ✔

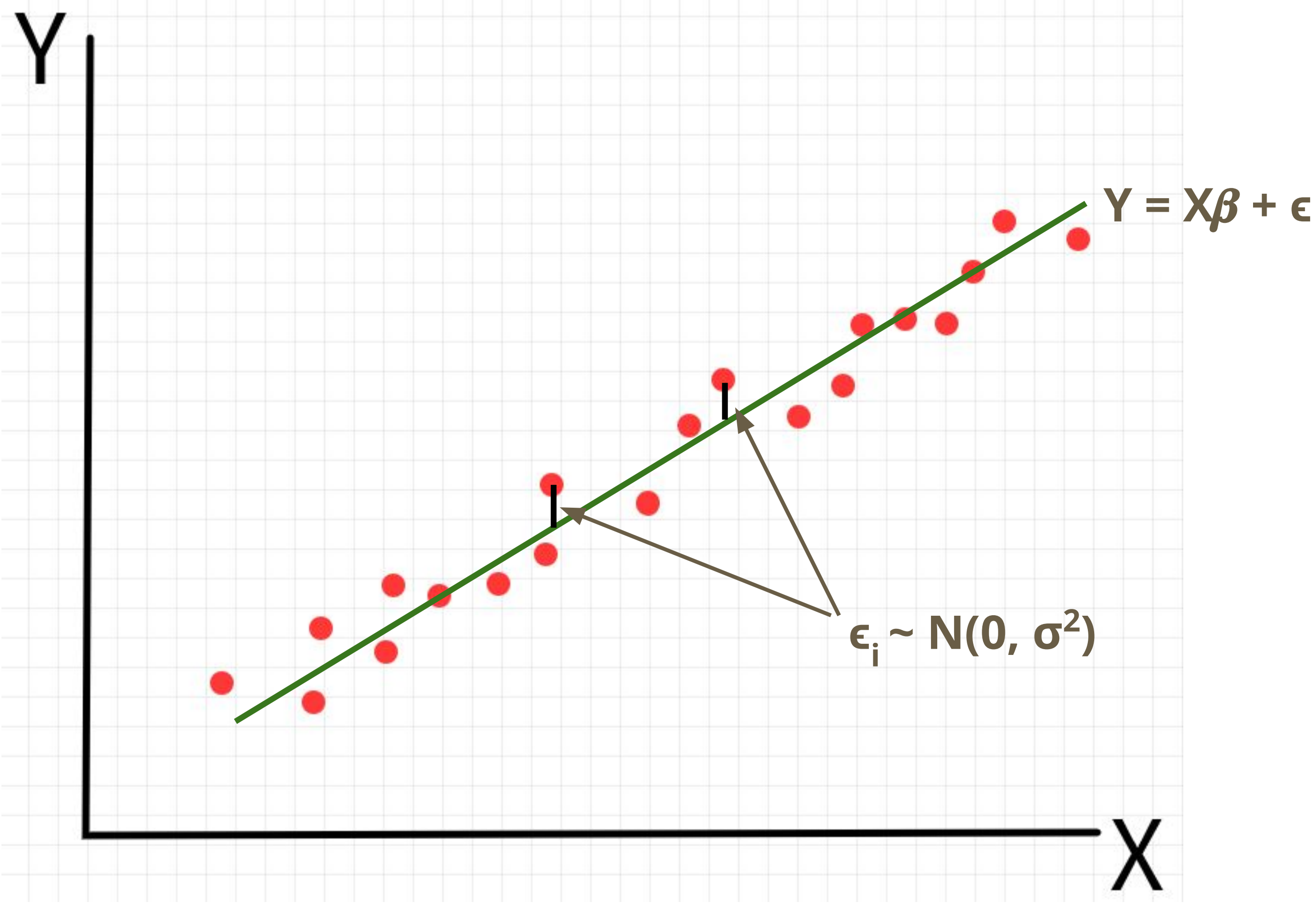
$\mathbf{h(x) = \beta_1 \log(x) + \beta_2 x^2}$ ✔

$\mathbf{h(x) = \beta_0 + \beta_1 x + \beta_1^2 x}$ ✘

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



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 $\mathbf{h}$

Least Squares

$$
\begin{aligned}
\beta_{LS} &= \arg\min_{\boldsymbol{\beta}} \sum_i d(h_\beta(x_i), y_i) \\
&= \arg\min_{\boldsymbol{\beta}} \|\vec{y} - h_\beta(\mathrm{X})\|_2^2 \\
&= \arg\min_{\boldsymbol{\beta}} \|\vec{y} - \beta \mathrm{X}\|_2^2
\end{aligned}
$$

Least Squares

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

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

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

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

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

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

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

Maximum Likelihood

Since **ϵ ~ N(0, σ^2)** and **Y = X$\boldsymbol{\beta}$ + ϵ** then **Y ~ N(X$\boldsymbol{\beta}$, σ^2)**.

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

An Unbiased Estimator

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

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

Short Demo

Logistic Regression

So far y_i was a continuous variable. What if y_i is categorical?

Assume we have **2 classes**.

Even if we can make these classes numerical (i.e. translate labels such as “yes”/”no” into 1 / 0), these numbers don’t have a mathematical meaning in the context of linear models and what we learn will be as arbitrary as the numerical labels we assigned (i.e. using “yes” =2/”no”=7 instead of “yes”=1/”no”=0 might “fit” a better model…).

Maybe we can use the probability of belonging to a given class as a proxy for how confidently we can classify a given point? Maybe we can fit a linear model to the probability of being in a given class!

Logistic Regression

So the output of our regression model could be a probability. But how can we enforce that $X\boldsymbol{\beta}_{LS}$ from our model is always constrained to [0,1]? i.e. how can we learn a $\boldsymbol{\beta}_{LS}$ such that $0 \leq X\boldsymbol{\beta}_{LS} \leq 1$ even for unseen X?

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

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

But again how can we enforce that the $X\boldsymbol{\beta}_{LS}$ are constrained to [0, ∞)? We need (-∞, ∞) - but how?

Let's take the log! This is also convenient numerically because in the previous 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.

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

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

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

We know:

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

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

Logistic Regression

So we can define

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

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

Short Demo

Evaluating Our Regression Model

Some Notation:

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

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

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

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

Evaluating Our Regression Model

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

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

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

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

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

Exercise

Show that TSS = ESS + RSS

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

Assume for simplicity that $\hat{\mathbf{y}}_i = \boldsymbol{\beta}_0 + \boldsymbol{\beta}_1 \mathbf{x}_i$
Since $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ 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 $\boldsymbol{\beta}_0$ and $\boldsymbol{\beta}_1$ we discover that

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

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

Evaluating our Regression Model

Each parameter of an independent variable **x** has an associated confidence interval

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)

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 $\mathbf{\epsilon \sim N(0, 25)}$. Then the Least Squares estimator of $\mathbf{\mu}$ ($\mathbf{\mu_{LS}}$) is

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

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

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

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

Z-value for 95% Confidence Interval

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 number of standard deviations from the mean that contains 95% of values) you need to solve:

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

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.

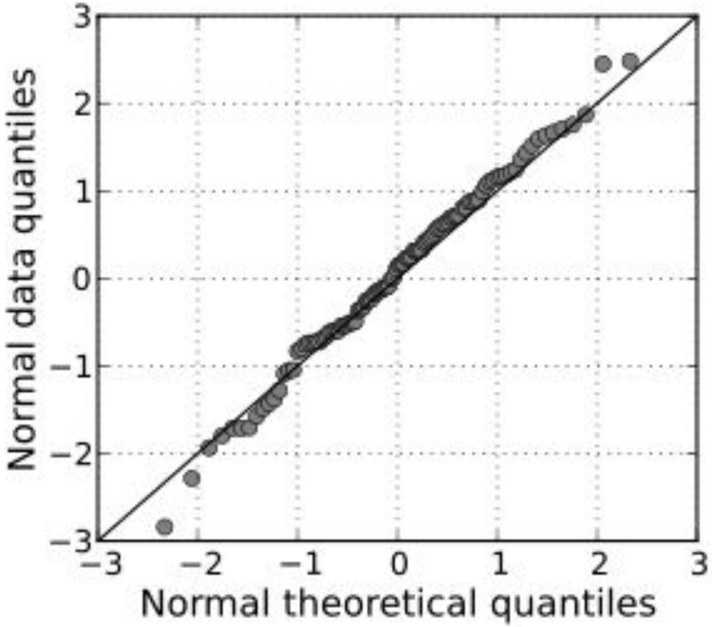
Plot the quantile of your target distribution against the quantiles of your data/ variable! If they match then your data probably comes from that distribution.

QQ plot

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

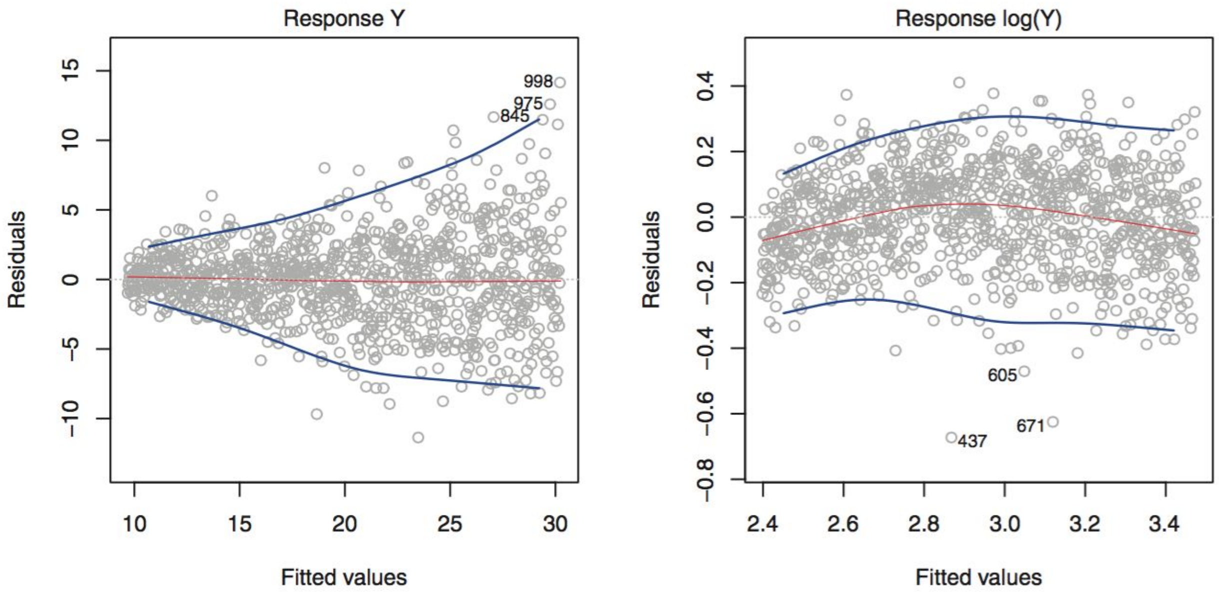


Normal data quantiles
Normal theoretical quantiles
3
2
1
0
−1
−2
−3
−3
−2
−1
0
1
2
3

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)