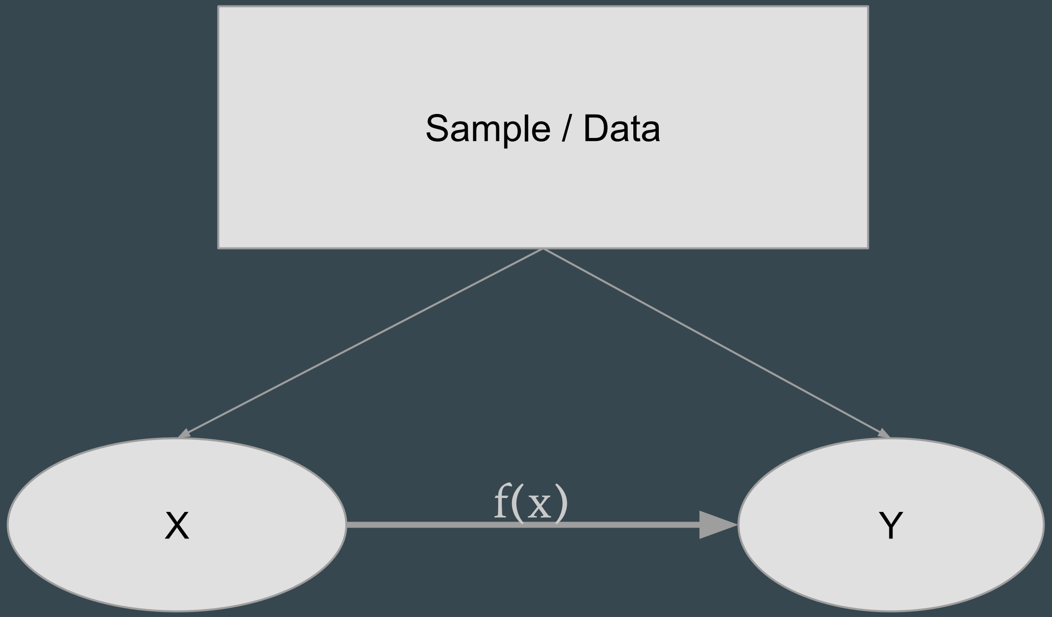
Maximum Likelihood Estimation

PLSC 309
1 June 2019

Review: OLS

- OLS is finding the best straight line fit between outcome Y and explanatory variables X
- It does this through the sum of squared errors
  - Predicted - actual outcomes
- Requires the underlying function to be additive and linear

Review: statistical modelling



Statistical modelling as conditional probability

- A statistical model gives us a function, f(x)
- This tells us how to combine our explanatory variables X to accurately guess Y
- In other words, it tells us what value for Y we should get, *conditional on X having certain values*

Conditional probability

- To get a good statistical model, we want to estimate the conditional probability

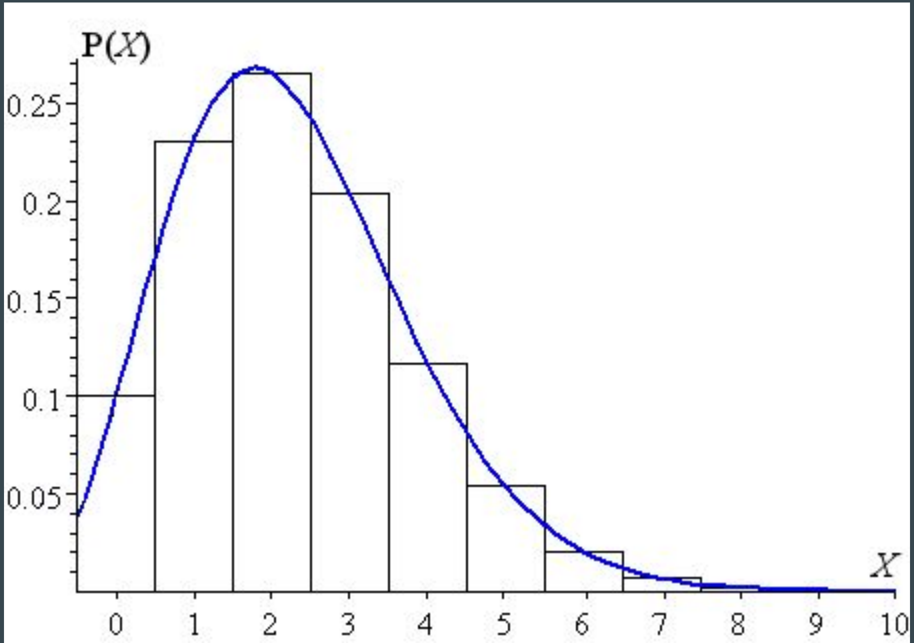
$$P(Y|X)$$

- Estimating this conditional probability directly is called *Maximum Likelihood Estimation (MLE)*
- It *maximizes* the *likelihood* of your outcome given a certain set of explanatory variables

Why do we want to do this?

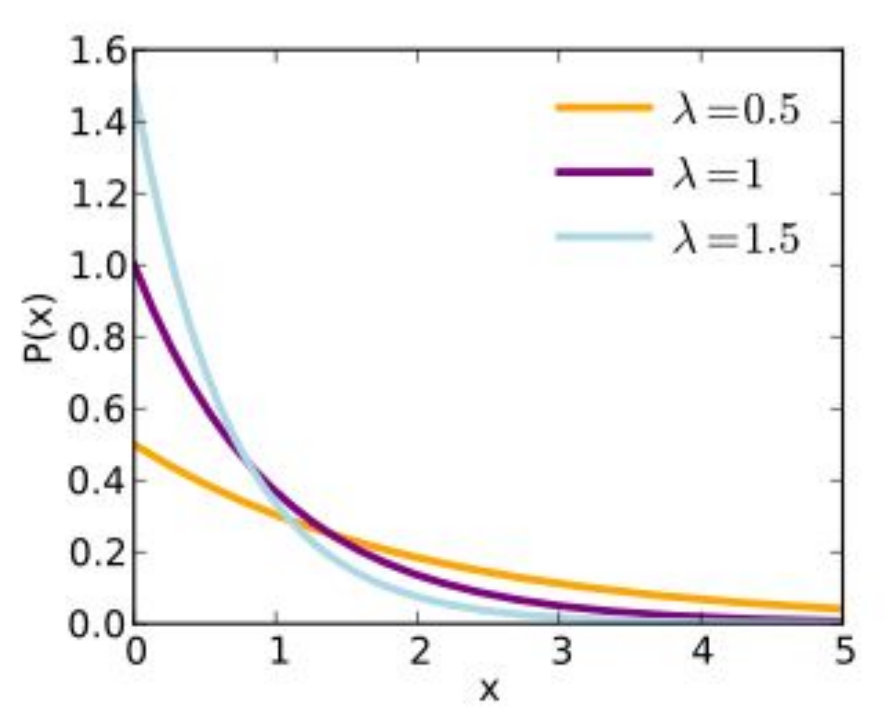
- OLS finds P(Y|X), *if and only if* X combines in a strictly linear way to produce Y
- MLE lets us find *any type of relationship* between P(Y|X)
- With MLE, we can find P(Y|X) that takes any probability distribution

What if P(Y|X) looked like this?



$$e^{-\lambda}\frac{\lambda^x}{x!}$$

What if P(Y|X) looked like this?



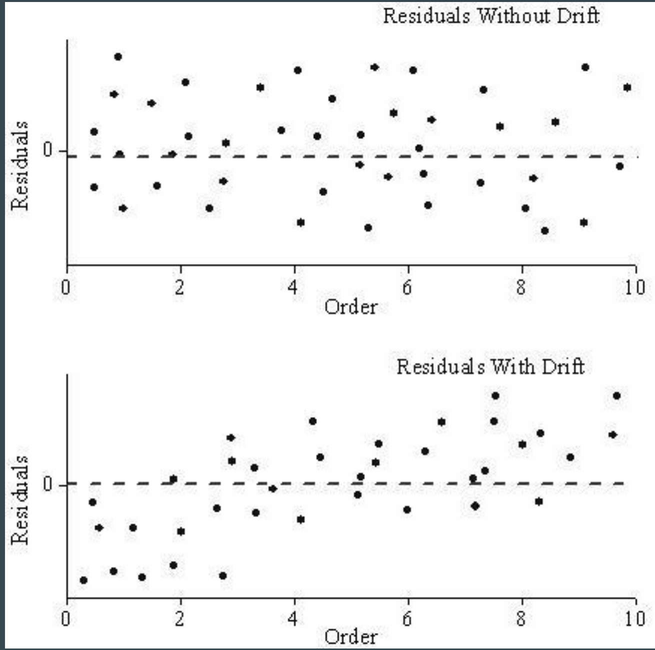
$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

Conditional probability with model parameters

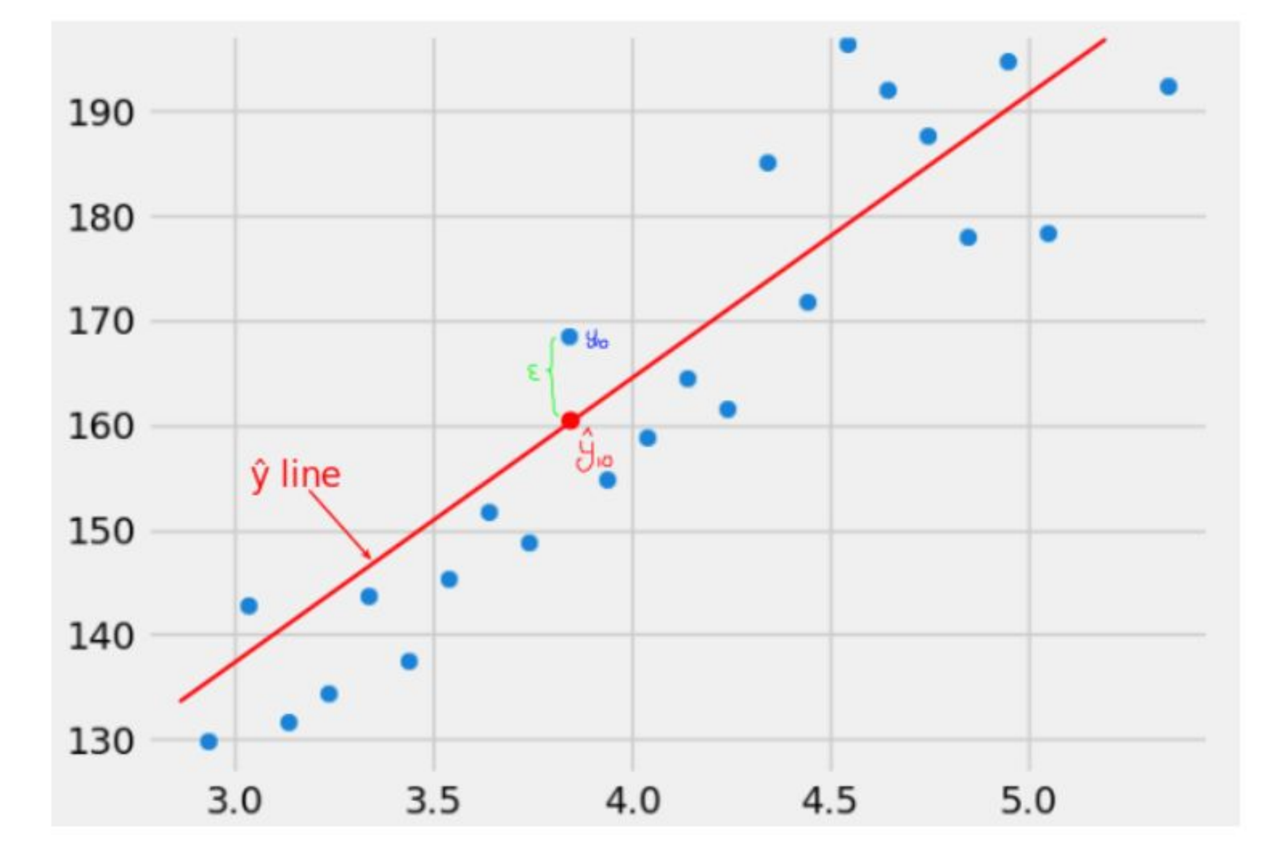
- We can also define our conditional probability in terms of model parameters
- Assume a linear relationship between X and Y
- $p(y_i \mid x_i ; \beta, \alpha)$
- This is read as the conditional probability of Y given X and a set of model parameters

Patterns in residuals

- In a linear model, a big red flag is having patterns in residuals



Normal residuals



Why normal residuals?

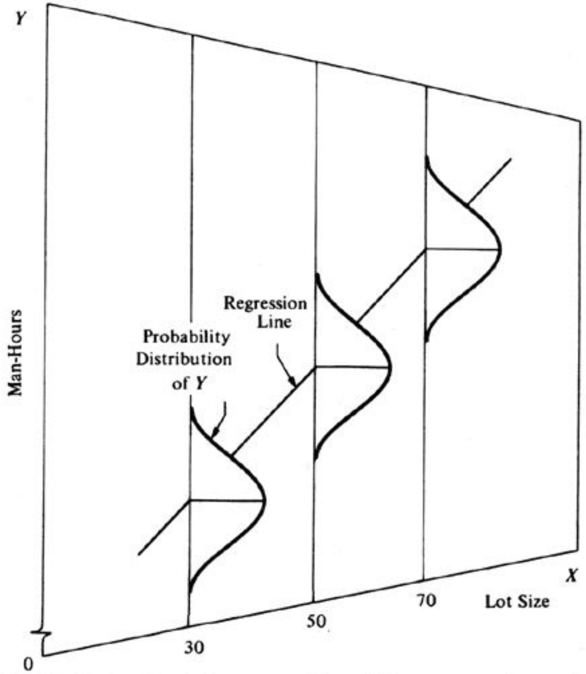


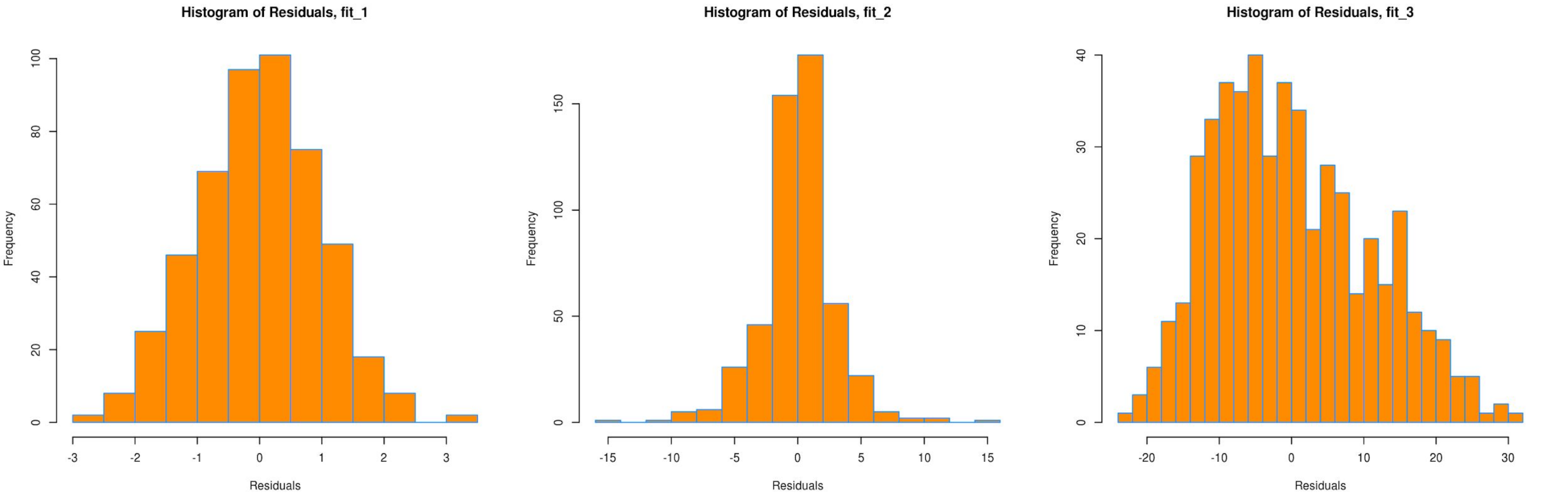
FIGURE 2.4 Pictorial representation of linear regression model

source: Neter, Wasserman & Kutner (1983) *Applied Linear Regression Models*

Why normal residuals?

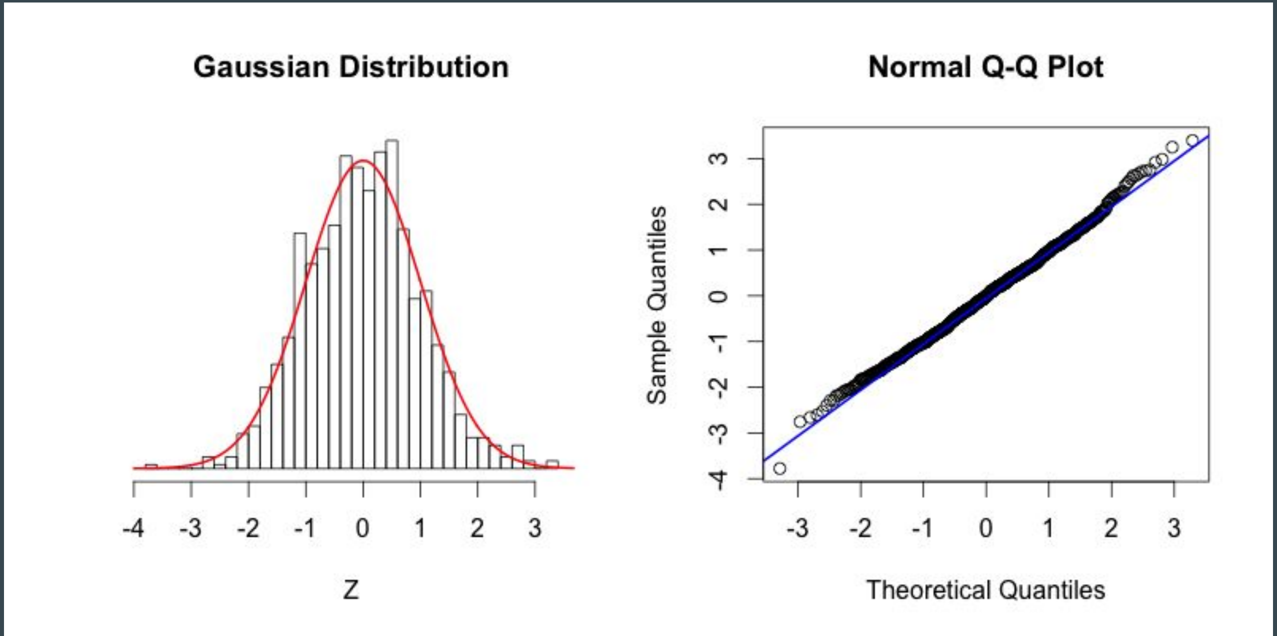
- $\hat{Y}$, is correctly interpreted as a mean of a probability distribution
- Each prediction has a probability distribution, centered around $\hat{Y}$, but with some random noise
- A linear model assumes $\hat{Y}$ is the *mean of $P(Y|X)$, assuming that $P(Y|X)$ is normal*

The shape of noise

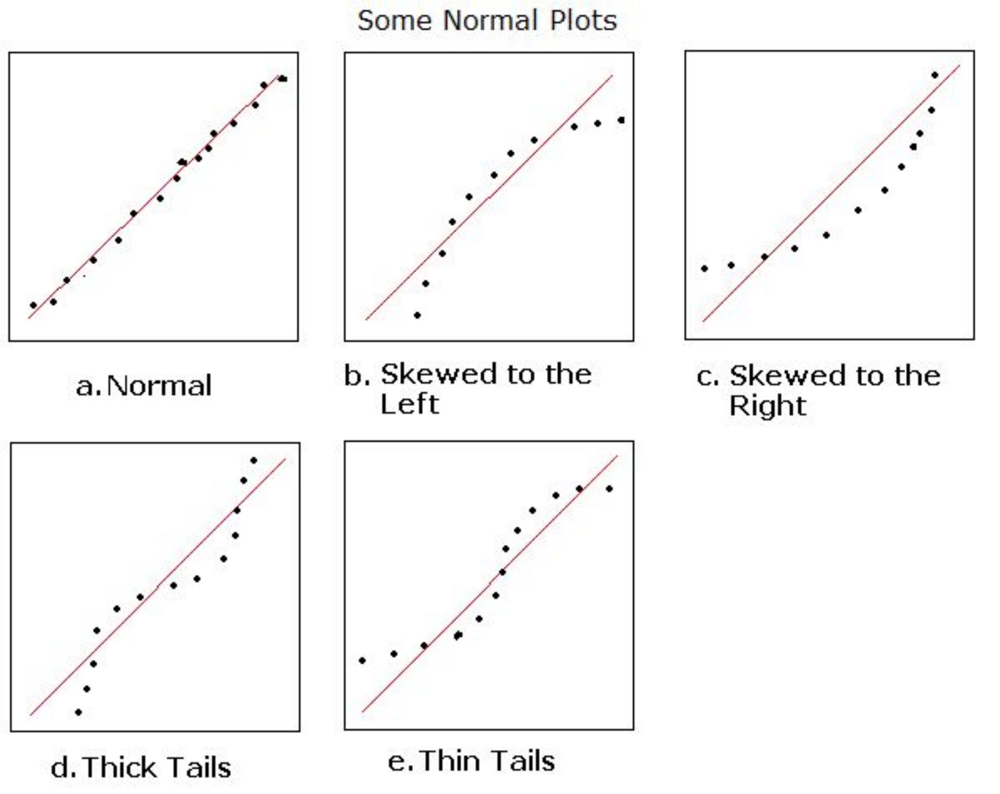


Visualizing normality

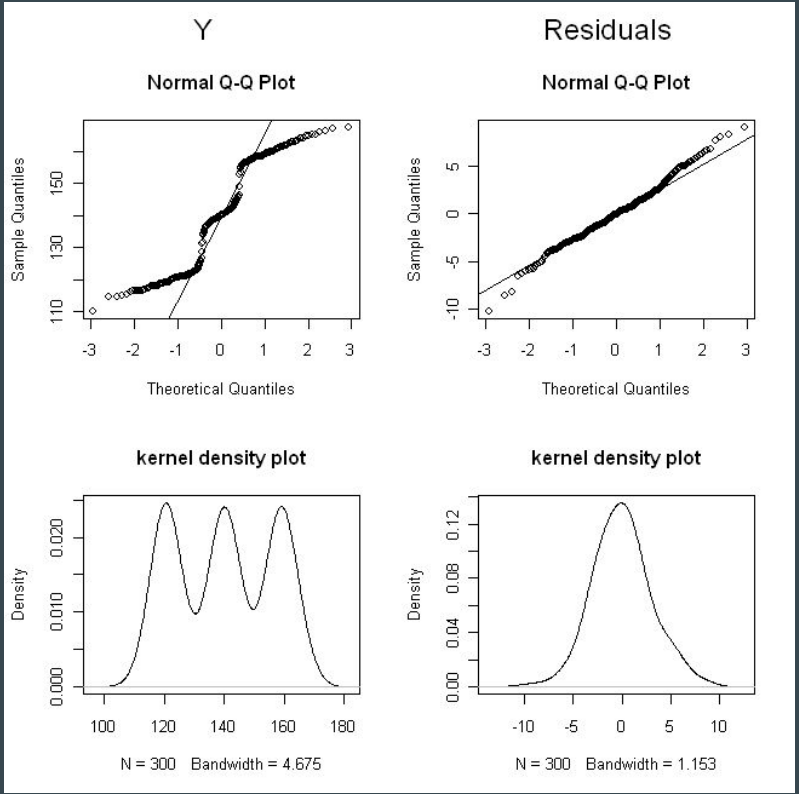
- A QQ plot provides a different way to visualize normality than a histogram
- It compares all points (dots) to their place on a perfectly normal line



Visualizing normality



Non-normal functions produce non-normal residuals



OLS: Linear model with specific error term

- Linear models argue that the real relationship, f(x), between X and Y
- Can be approximated by a straight line g(x)
- If this is correct, then we should have normally distributed residuals
    - Equal chance of over and under guessing
    - More extreme values are less likely

Statistical modelling as learning parameters

- OLS gave us the model parameters we needed to do our prediction
  - β
  - α
- With these model parameters we can take each x_i (each data point) and return a guess for $\hat{y}_i$
  - $\hat{y}_i = \alpha + \beta x_i$
- $\hat{Y}_i$ is the center of a normal distribution, which describes where observed values are likely to fall

MLE

In order to do MLE, we have to follow the following steps

1. Determine the probability distribution we think f(x) follows
2. Write down the probability distribution as a likelihood
3. Find the values of your parameters that maximizes the likelihood
4. Using the function you just produced, calculate the parameters with your data

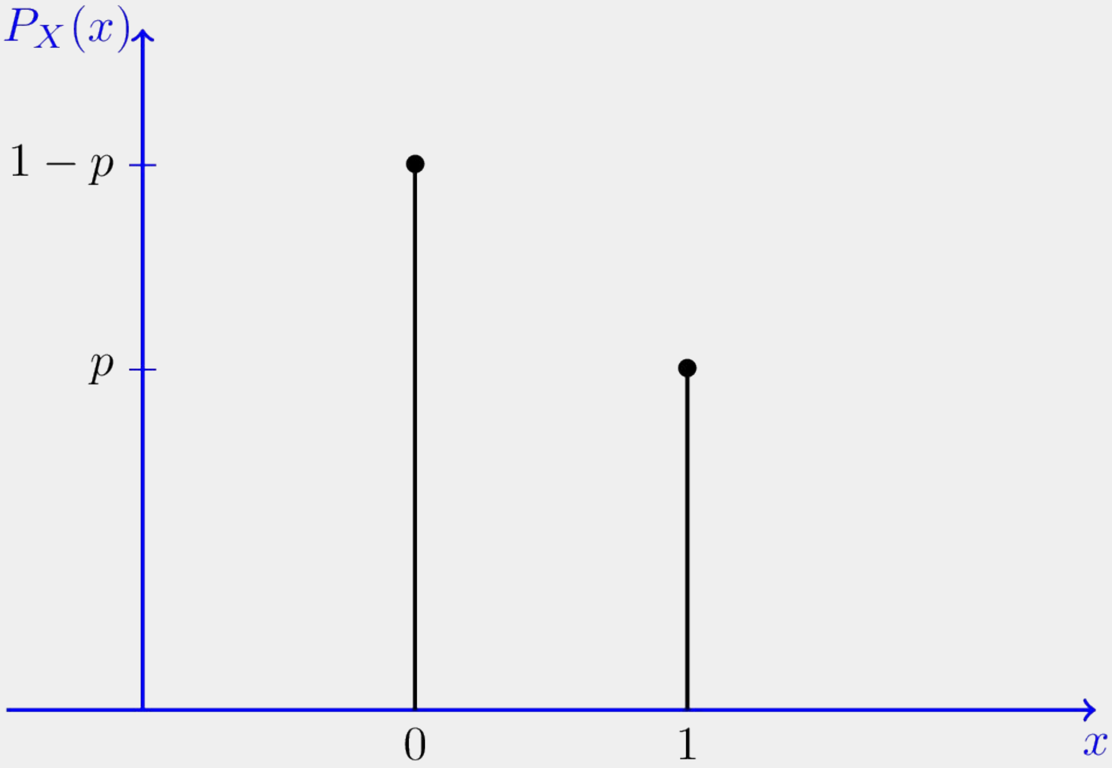
MLE example

A coin flip

- We are interested in estimating the likelihood of a coin flip
- We can use maximum likelihood estimation to do this
- Step one: what distribution does this process follow?

Bernoulli Distribution

$$X \sim Bernoulli(p)$$



$$\begin{cases} q = 1 - p & \text{if } k = 0 \\ p & \text{if } k = 1 \end{cases}$$

A coin flip

- We have figured out that the process we're interested follows a Bernoulli distribution, which can be expressed as

$$f(y \mid x_i; p) = p^x(1-p)^{(1-x)}$$

- Where x = 0 or x = 1

From probability to likelihood

- Our bernoulli distribution has a single parameter: p
- The likelihood takes the probability, and extends it infinitely
- It does this by multiplying the probability of each individual data point
- In other words, we go from $f(y|\ x_i\ ;\ p)$ to $f(y\ |\ x_1, x_2, \dots, x_3\ ;\ p)$

Likelihood is each prediction multiplied

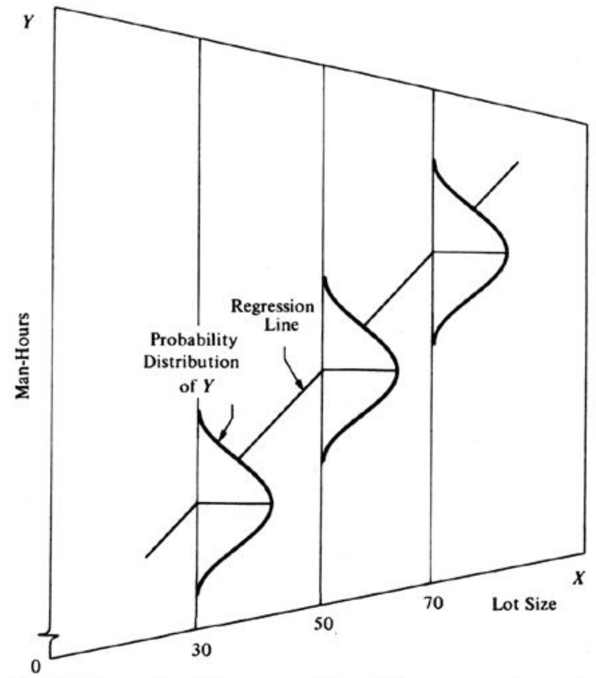


FIGURE 2.4 Pictorial representation of linear regression model

source: Neter, Wasserman & Kutner (1983) *Applied Linear Regression Models*

Likelihood for Bernoulli

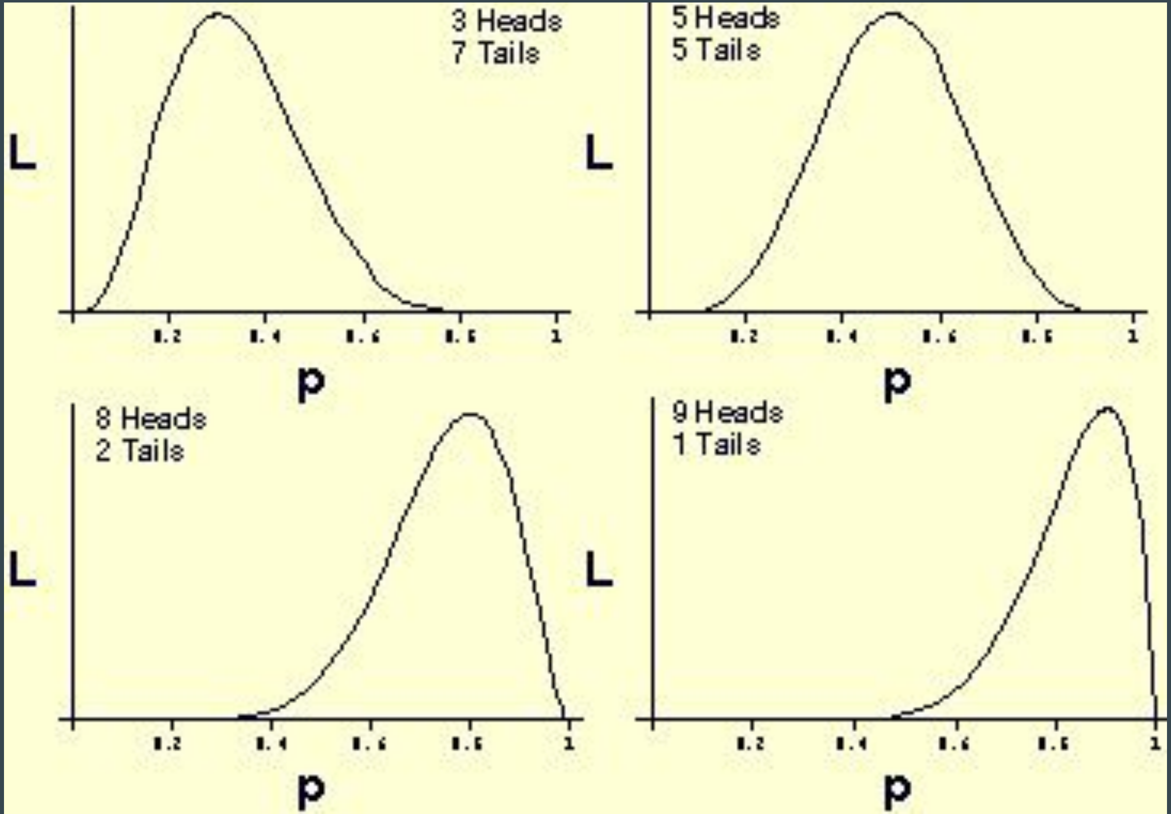
- PDF for each point: $f(x_i \mid p) = p^x(1\text{-}p)^{(1\text{-}x)}$
- We want to find $f(x_1, x_2, \ldots , x_n)$
- Likelihood:

$$L(p) = \prod_{i=1} p^{x_i}(1-p)^{(1-x_i)}$$

- Log-likelihood:

$$\ell(p) = \log p \sum_{i=1}^{n} x_i + \log(1-p) \sum_{i=1}^{n} (1-x_i)$$

What the likelihood looks like



How we can use the likelihood

- Now we have this function, the log likelihood, which expresses the combined probability distributions for all our possible data points
- Note that because we have each individual x_i in the likelihood, x is no longer a variable
- So the likelihood is just a function that expresses the probability distribution for a predicted outcome *in terms of its parameters*
- We can maximize this likelihood to find the best estimator for our parameters

Maximizing a function

$$\ell(p) = \log p \sum_{i=1}^{n} x_i + \log(1-p) \sum_{i=1}^{n} (1 - x_i)$$

- All probability distributions have a single peak
- The mathematic definition of a peak is where the slope is zero
- We find the maximum likelihood by taking the *derivative* (slope) of the log-likelihood and setting that function = 0

Maximizing the Bernoulli

- The derivative of the log-likelihood is:

$$\sum x_i(1-p) - (n - \sum x_i)p$$

- Set this equal to 0 to maximize

$$\sum x_i(1-p) - (n - \sum x_i)p = 0$$

- Multiply through

$$\sum x_i - np = 0$$

Maximizing the Bernoulli

- Now solve for p:

$$\sum x_i = np$$

$$p = \sum x_i / n$$

- In other words, the estimator for our parameter p that maximizes the likelihood of p is the total number of successes divided by the total number of outcomes
- So if I get 10 heads on 20 coin flips, the best estimator for p is 50%

Benefit of the MLE

- MLE allows us to estimate *any* type of relationship between X and Y
- As long as we can specify a probability distribution that these follow
- The MLE for any likelihood is guaranteed to be MVUE
  - Minimum Variance
  - Unbiased
- We now don't have to make the assumption that f(x) is a straight line

The costs of MLE: conditional independence

- To go from probability distributions for each point to a likelihood function across our entire data, we multiplied the probabilities
- We could do this because we made an assumption of conditional independence
- *Conditional independence:* each observation of Y is independent from the others, except for their relationship to X
    - Regular independence: a random sample of 100 students
    - Conditional independence: a random sample of 100 people over 45 years old
- Remember, if probabilities are independent, they can multiplied to find their joint probability

The costs of MLE: functional form

- We still have to assume the shape of the relationship between X and Y
- In OLS, we were restricted to only straight line shapes
- For MLE, we have many more options:
    - Exponential regression
    - Poisson regression
    - Logistic regression
    - T-distribution
- But this is still an assumption!

Review: MLE

- We want a general model that can estimate different kinds of relationships between X and Y
- To do this we define X and Y as a conditional probability
- We combine these different probabilities into a likelihood
- The likelihood tells us the relationship between our data and parameters
- We maximize this likelihood to find an estimator for our model parameters