Multivariate Modelling and Simulation - MLE

COSC/DATA 405/505

Maximum Likelihood Estimation (MLE)

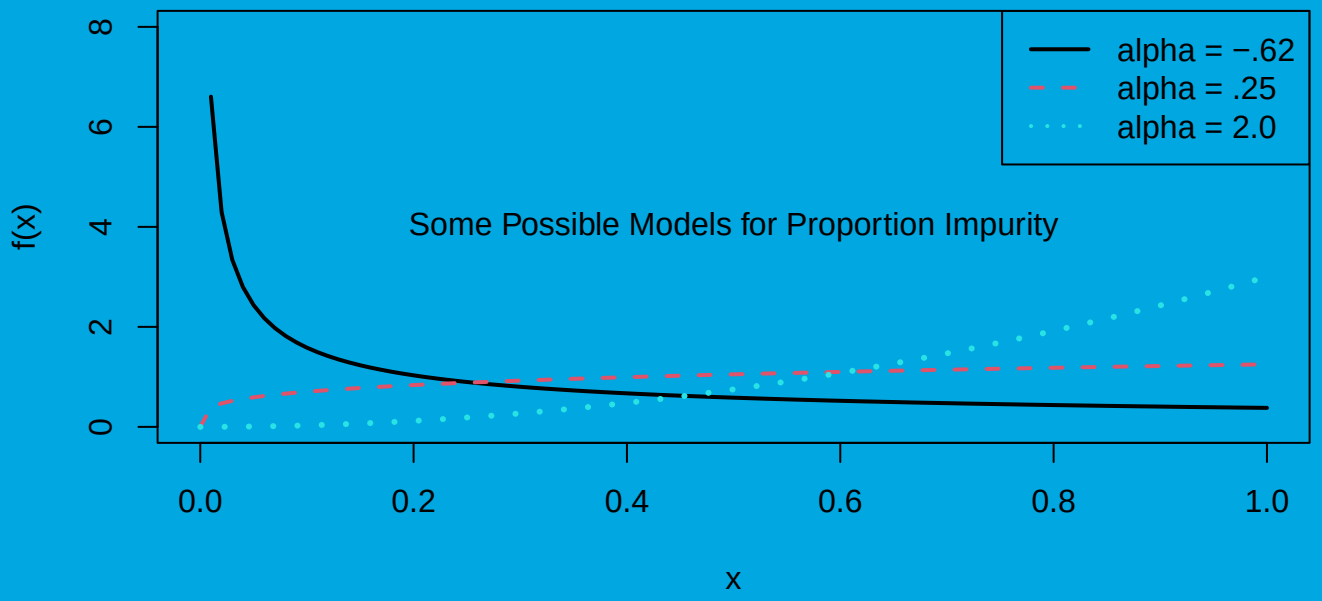
- a motivating example
- maximizing the likelihood when there are a small number of possible parameter values
- maximizing the likelihood when there is a continuous infinity of possible parameter values, using calculus
- Generalized Linear Models
  - Linear Models and the need for more flexibility
  - Poisson regression
  - Binary logistic regression

Motivating Example

Suppose the proportion of impurity X in an iron ore specimen is modelled with the pdf

$$f_X(x) = (\alpha + 1)x^{\alpha}, \quad 0 \leq x \leq 1.$$

α is an unknown parameter.



Because α is unknown, it is impossible to know which probability density function is the correct one.

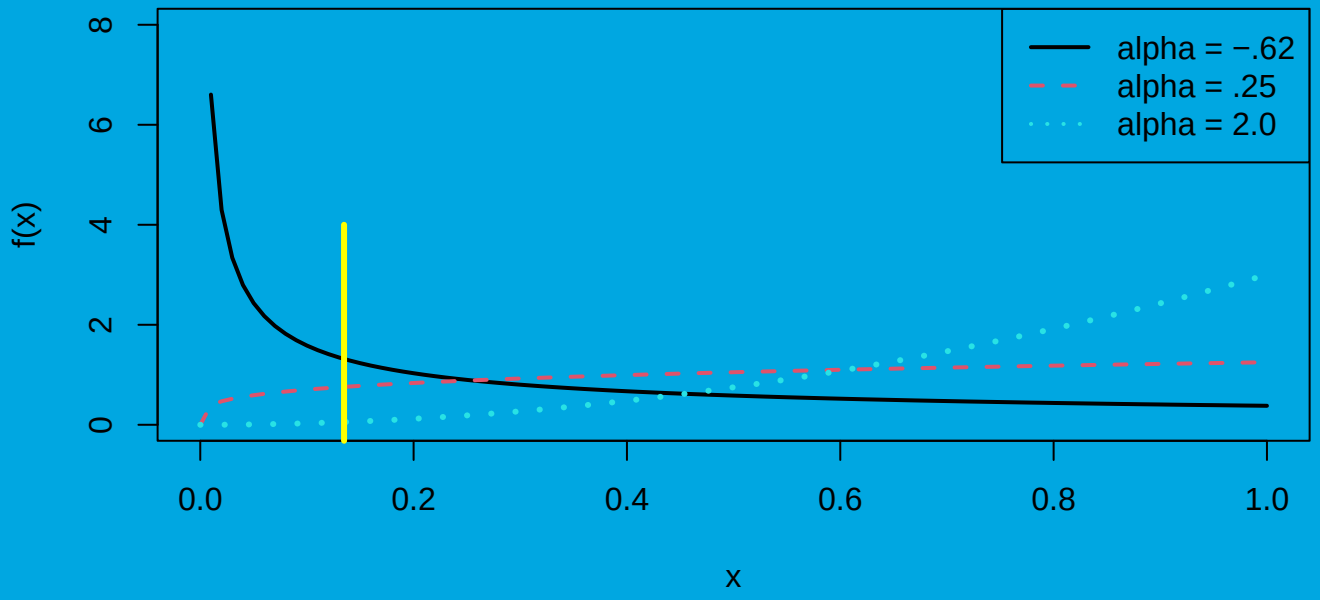
Motivating Example

By taking measurements of the impurity in a sample of one or more specimens, we can estimate α using maximum likelihood estimation to determine which is the best model for the impurity measurements.

The density function is highest at values of x that are most probable.

Taking one measurement

Suppose an impurity measurement is taken: $x = .1348$.



Unless we have taken an unusual measurement, the probability density function at our measurement should be high.

Which of the above curves appears to be the most likely?

Maximizing the Likelihood

If the parameter space consists only of the values $\{-.62, .25, 2.0\}$, then we would choose $\hat{\alpha} = -.62$, since it seems to be the most likely value.

This is the maximum likelihood estimate.

The maximum likelihood estimate is the value of the parameter which gives maximum probability density at the given data.

It is important to keep the possible set of parameter values in mind. Here there were 3 possibilities only.

In many cases there are an infinite number of possibilities.

Maximizing the Likelihood

Suppose the parameter space is the set where $\alpha > -1$.

The likelihood principle tells us to choose the value that gives highest density at the data point.

In other words, we need to see where

$$f_X(.1348) = (\alpha + 1).1348^{\alpha}$$

is large. This is a *function of* α:

$$L(\alpha) = (\alpha + 1).1348^{\alpha}$$

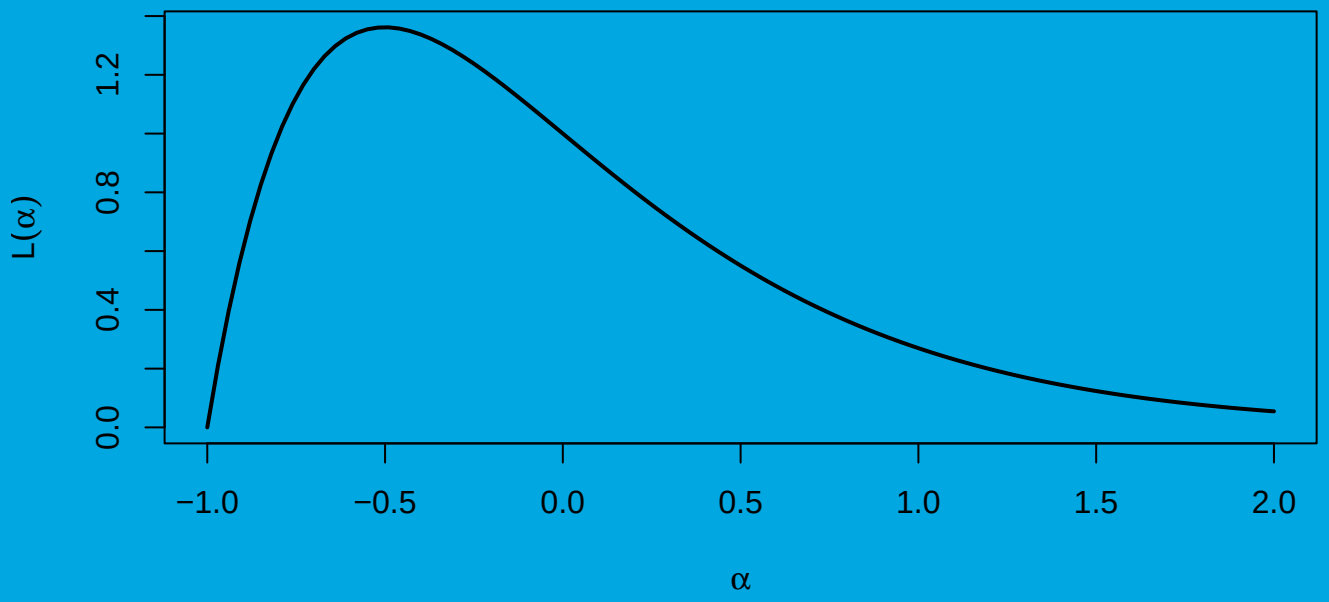
$L(\alpha)$ is called the likelihood function.

Our goal is to find the value of α for which this function is maximized.

Choosing the Most Likely Density Function

When there is only one parameter, it is easy to plot the graph of the likelihood function.

It is always a good idea to plot the likelihood function.

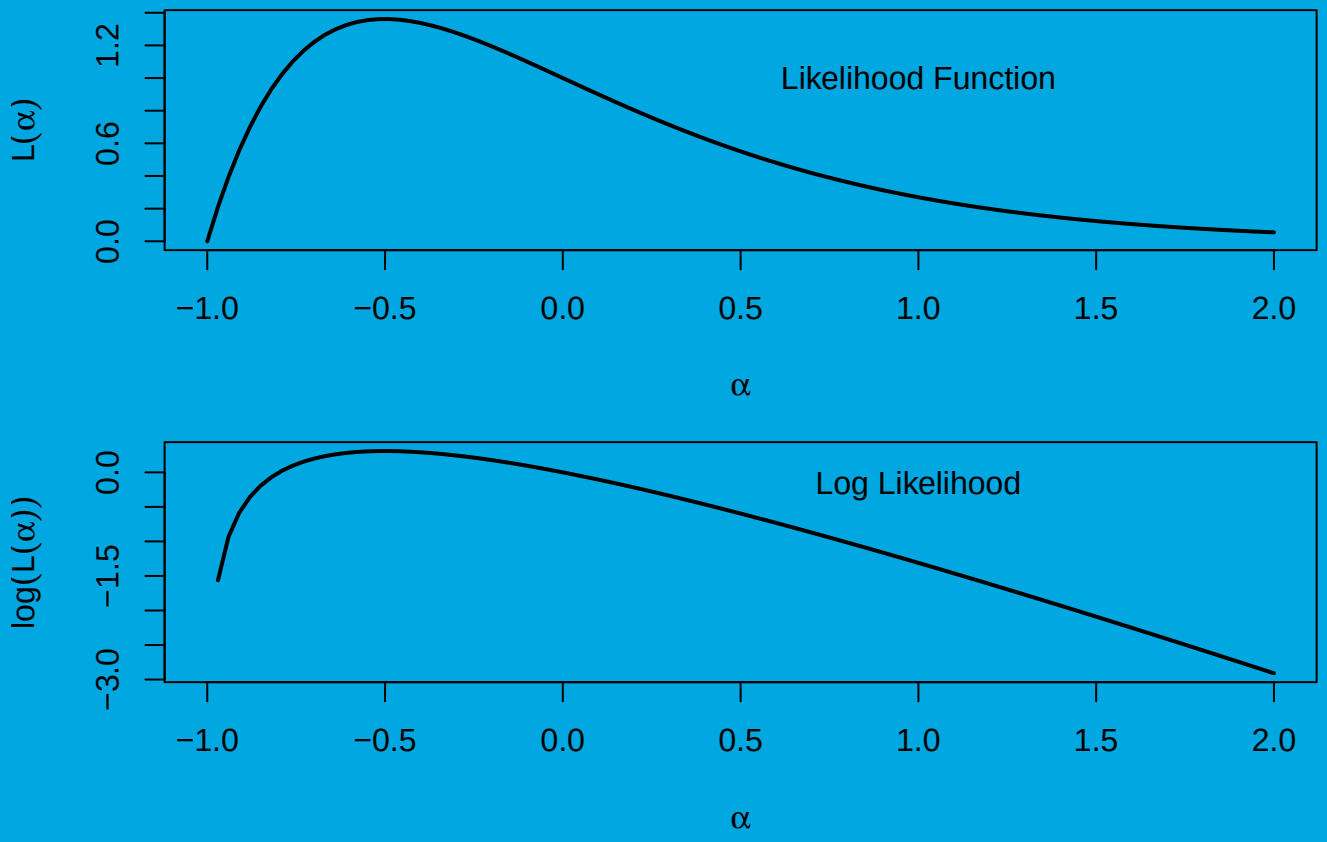


The function above measures the likelihood of our observed measurement as a function of α:

We can estimate α by maximizing this likelihood. The maximum value occurs near $\alpha = -.5$.

Choosing the Most Likely Density Function

Maximizing the log likelihood gives the equivalent result, but is often computationally more convenient:



The maximum value occurs near $\alpha = -.5$ in both cases.

Choosing the Most Likely Density Function

The exact value that maximizes the likelihood can *sometimes* be determined using calculus:

1. **Differentiate $L(\alpha)$ (or $\ell(\alpha) = \log L(\alpha)$) with respect to α**

2. **Solve for α in $L'(\alpha) = 0$.**

Example:

$$\ell(\alpha) = \log(f(x)) = \log(\alpha+1) + \alpha \log(x) = \log(\alpha+1) + \alpha \log(.1348)$$

Differentiate with respect to α:

$$\ell'(\alpha) = \frac{1}{\alpha+1} + \log(.1348)$$

Solve $\ell'(\alpha) = 0$ for α:

$$\widehat{\alpha} = -1 - \frac{1}{\log(.1348)} = -0.501.$$

Estimating α with 2 Measurements

We should hope to get a better estimate of α if we have more than one measurement.

The joint density function for 2 independent impurity measurements, x_1 and x_2, is

$$f(x_1, x_2) = f_X(x_1) f_X(x_2)$$

$$= (\alpha + 1)^2 (x_1 x_2)^\alpha = L(\alpha)$$

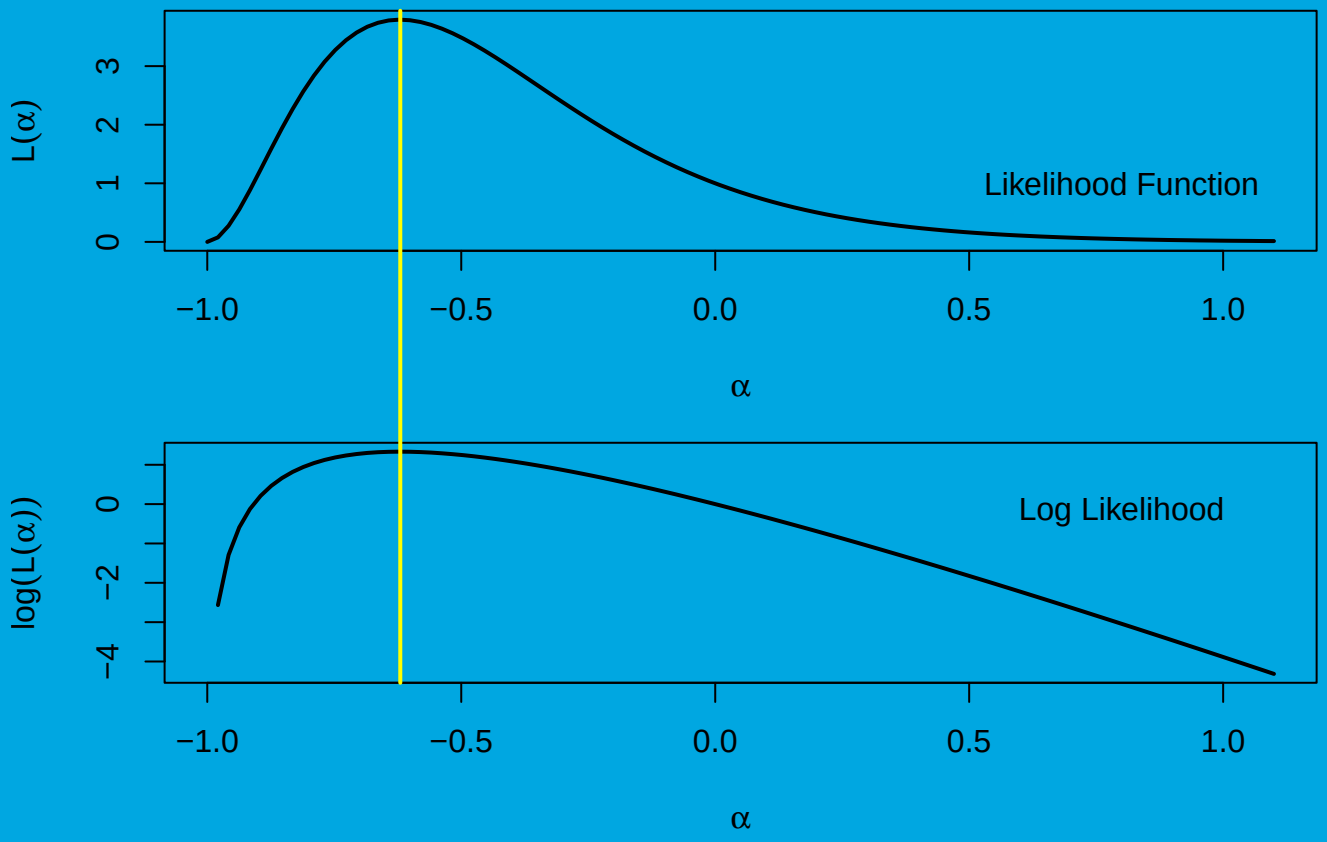
The log likelihood function is

$$\ell(\alpha) = \log L(\alpha) = 2 \log(\alpha + 1) + \alpha \log(x_1 x_2).$$

Maximizing either of these functions with respect to α gives us the maximum likelihood estimate of α.

Estimating α with 2 Measurements

A second independent impurity measurement is .0381.



The maximum value occurs near $\alpha = -.6$ in both cases.

Estimating α with 2 Measurements

Evaluating the likelihood at the 2 measurements, we have

$$L(\alpha) = (\alpha + 1)^2 (.00514)^\alpha$$

and the log likelihood, $\ell(\alpha)$ is

$$\ell(\alpha) = \log L(\alpha) = 2 \log(\alpha + 1) + \alpha \log(.00514)$$

By differentiating with respect to α, we can find the value of α that maximizes this. That is, solve

$$\ell'(\alpha) = \frac{2}{\alpha + 1} + \log(.00514) = 0.$$

The maximizer is $\widehat{\alpha} = -.62$.

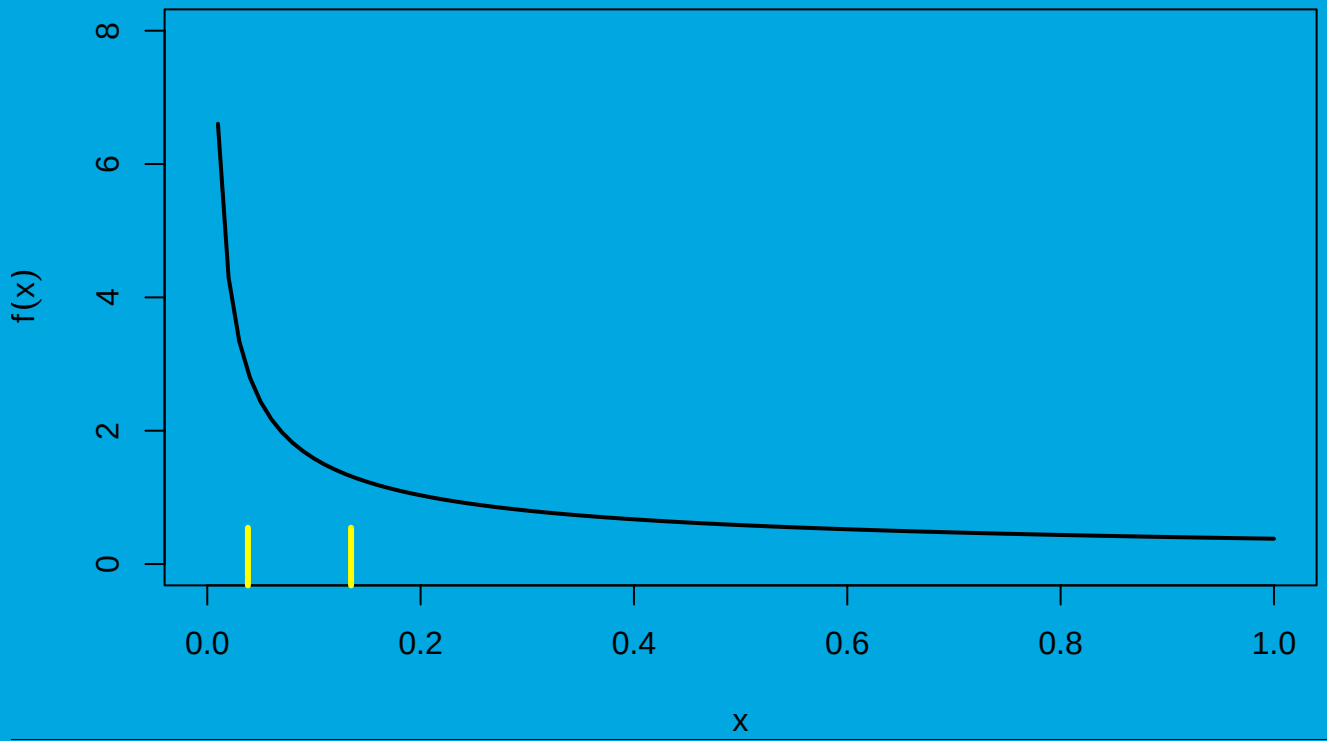
Estimating α with 2 Measurements

We can write

$$\widehat{f}_X(x) = .38x^{-.62}$$

as our estimate of the impurity pdf.

We can also plot it, together with the data, to see that it makes sense:



Estimating α with a Larger Sample

A better estimate can be obtained with a larger sample of impurity measurements.

Here is a sample of 10 independent measurements:

```
##  [1] 0.1348 0.0420 0.0003 0.0049 0.0002
##  [6] 0.0381 0.0018 0.0264 0.0366 0.0007
```

Because of independence, the joint density evaluated at the measurements is

$$f(x_1, x_2, \ldots, x_{10}) = (\alpha + 1)^{10}(x_1 x_2 \ldots x_{10})^{\alpha}$$

Again, this is a function of the unknown parameter α. We can take the logarithm of this likelihood function:

$$\ell(\alpha) = \log L(\alpha) = 10 \log(\alpha + 1) + \alpha \sum_{j=1}^{10} \log(x_j)$$

Estimating α with 10 Measurements

The log likelihood function evaluates to

$$\ell(\alpha) = 10 \log(\alpha + 1) + \alpha(-50.9155).$$

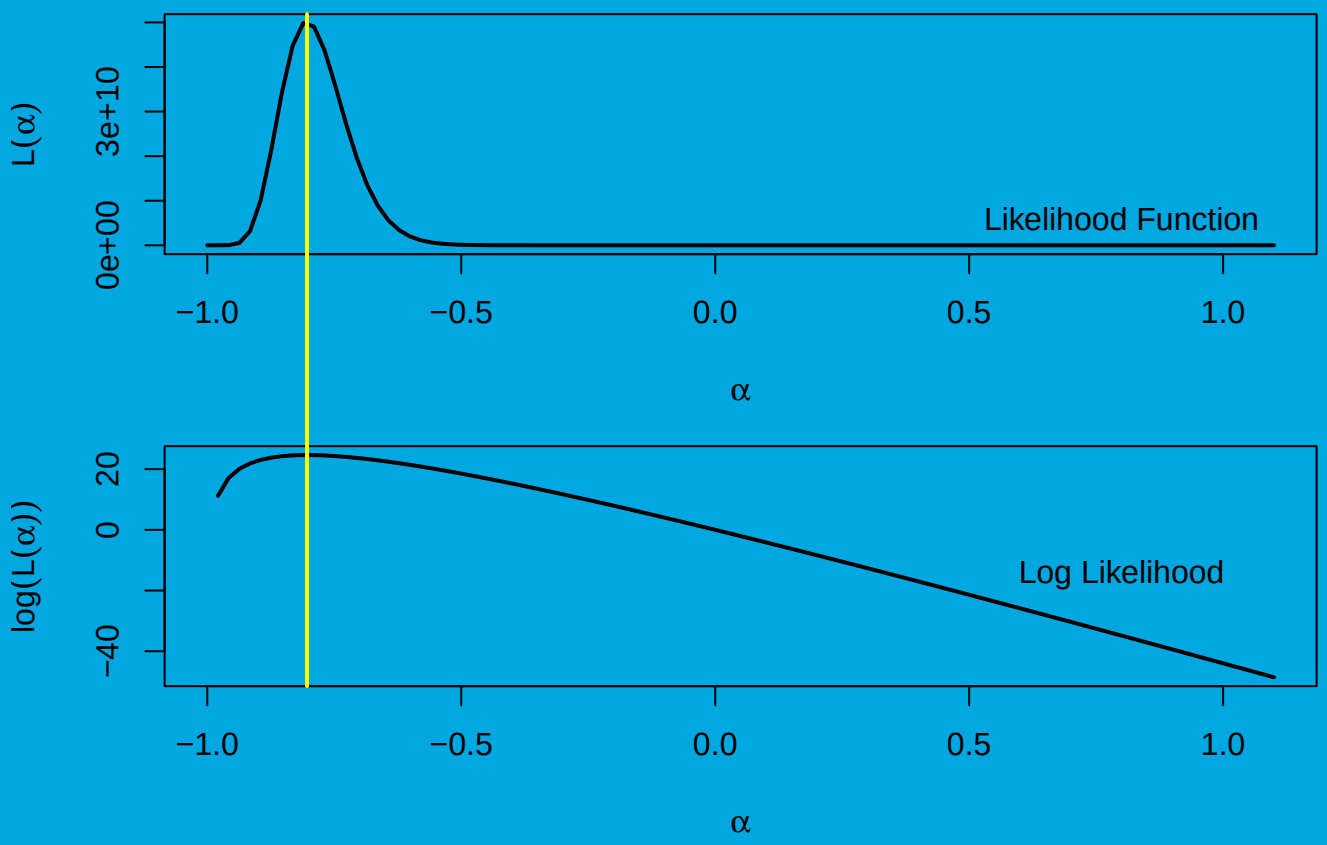
We maximize this by solving

$$\ell'(\alpha) = \frac{10}{\alpha + 1} - 50.9155 = 0.$$

$$\hat{\alpha} = -0.8036.$$

Estimating α with 10 Measurements

The likelihood and log likelihood functions can be plotted:



The maximum value occurs at $\alpha = -0.8036$ in both cases.

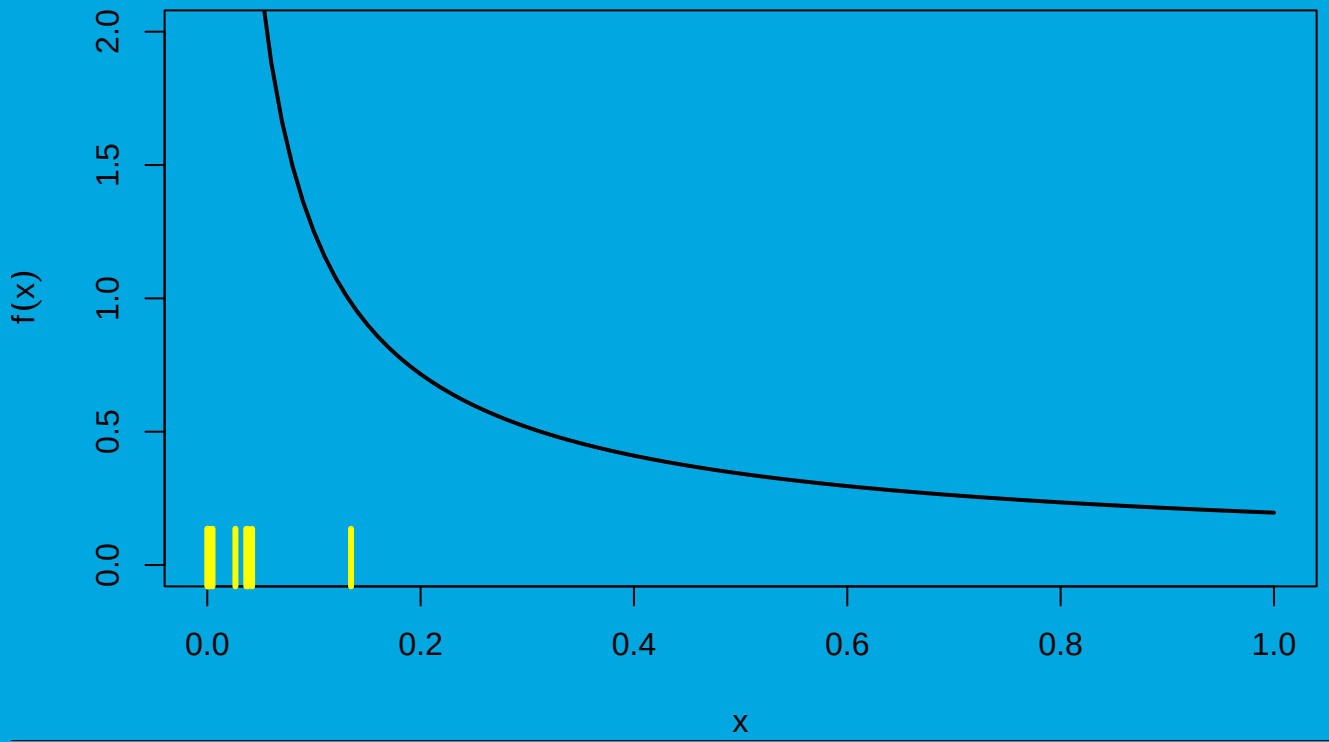
Note that the spread in the likelihood function is less than before $\rightsquigarrow$ the maximizer is more precise due to the increase in sample size.

Estimating α with 10 Measurements

The estimated pdf is now

$$\widehat{f}_X(x) = 0.1964 x^{-0.8036}.$$

Again, we plot the estimated pdf, together with the data, to see that it makes sense:



Maximum Likelihood Estimation - Summary

The likelihood function for n independent measurements $y_1, \ldots, y_n$ coming from a population modelled by a density $f(y)$ is

$$L(\theta) = f(y_1)f(y_2)\cdots f(y_n)$$

θ denotes parameter(s) to be estimated.

The maximum likelihood estimator for θ is the value which maximizes the likelihood function.

Often, the maximum likelihood estimator can be found by setting the derivative of the log likelihood to 0 and solving for θ.

Another Example

The time between successive oil rig accidents can be modelled with the probability density function

$$f(y) = \lambda e^{-\lambda y}.$$

Suppose oil rig accident times have occurred at $t_1, t_2, \ldots, t_{n+1}$**.**

Let $y_i = t_{i+1} - t_i$ **for** $i = 1, 2, \ldots, n$**. Then** $y_1, y_2, \ldots, y_n$ **are the times between successive oil rig accidents.**

The likelihood function is then

$$L(\lambda) = \lambda e^{-\lambda y_1} \ldots \lambda e^{-\lambda y_n}$$

$$= \lambda^n e^{-\lambda \sum_{i=1}^{n} y_i}$$

e.g. if $n = 3$**, and** $y_1 = 8, y_2 = 2$ **and** $y_3 = 6$**, then**

$$L(\lambda) = \lambda^3 e^{-16\lambda}$$

Oil Rig Accident Times - Example

The most believable value of λ maximizes $L(\lambda)$ or $\log(L(\lambda))$.

Differentiate $\log(L(\lambda))$ with respect to λ:

$$\frac{d}{d\lambda}\log(L(\lambda)) = \frac{d}{d\lambda}\left(-\lambda\sum_{i=1}^{n} y_i + n\log(\lambda)\right) = -\sum_{i=1}^{n} y_i + n/\lambda$$

Setting this equal to 0 gives the maximum:

$$\widehat{\lambda} = n/\sum_{i=1}^{n} y_i = 1/\bar{y}$$

For the given data,

$$\widehat{\lambda} = 3/16$$

Example: Estimating more than One Parameter

An important property of air bags is permeability of the woven fabric.

This is related to their ability to absorb energy.

A possible model for permeability measurements is the normal distribution:

$$f(x) = \frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$$

where μ denotes the expected or mean permeability measurement, and σ^2 denotes the variance of such measurements.

Example: Estimating more than One Parameter

Given n independent measurements $x_1, \ldots, x_n$, the log likelihood function is

$$\log L = -\frac{\sum_{i=1}^{n}(x_i - \mu)^2}{2\sigma^2} - \frac{n}{2}\log(2\pi\sigma^2)$$

To estimate μ and σ^2, we will maximize this log likelihood with respect to μ and σ^2.

Differentiating with respect to μ and setting to 0 gives

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

Also, differentiating with respect to σ^2 and setting to 0 gives

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n}$$

Example: Estimating more than One Parameter

4 independent measurements of permeability were taken at $20°$:

```
40, 60, 50, 45
```

Substituting into the expressions above, we obtain

$$\widehat{\mu} = \bar{x} = 48.8$$

and

$$\widehat{\sigma}^2 = 54.7.$$

Example: Estimating more than One Parameter

It is important to know whether permeability is different for different temperatures.

To determine whether temperature has an effect, we consider measurements of air bag permeability at 2 different temperatures: 0°**C and** 20°**C:**

```
 0:     70, 85, 92, 80, 60
20:     40, 60, 50, 45
```

Air Bag Example

Let us *model* permeability at each temperature with normal distributions.

Denote the expected value of permeability at 0° by μ_0 and at 20° by μ_{20}.

Let the variance be σ^2 in both cases.

We are assuming that variability is the same at different temperatures.

Air Bag Example

How do we estimate μ_0, μ_{20} and σ^2?

Let $x_1, x_2, \ldots, x_m$ denote the 1st sample, and let $y_1, y_2, \ldots, y_n$ denote the 2nd sample.

$m = 5$ and $n = 4$.

The density function for one of the X measurements is

$$f(x) = \frac{e^{-(x-\mu_0)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$$

The density function for one of the Y measurements is

$$f(y) = \frac{e^{-(y-\mu_{20})^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$$

Air Bag Example

The log likelihood function is

$$\log L = -\frac{\sum_{i=1}^{m}(x_i - \mu_0)^2}{2\sigma^2} - \frac{\sum_{j=1}^{n}(y_j - \mu_{20})^2}{2\sigma^2} - \frac{m+n}{2}\log(2\pi\sigma^2)$$

Air Bag Example

Differentiating with respect to μ_0 and setting to 0 gives

$$\widehat{\mu}_0 = \frac{1}{m}\sum_{i=1}^{m} X_i = \bar{X}$$

Similarly, differentiating with respect to μ_{20} gives

$$\widehat{\mu}_{20} = \frac{1}{n}\sum_{j=1}^{n} Y_j = \bar{Y}$$

Also, differentiating with respect to σ^2 gives

$$\widehat{\sigma}^2 = \frac{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2}{n+m}$$

$\bar{x} = 77.4$, $\bar{y} = 48.8$, $\widehat{\sigma}^2 = 94.9$

Estimate of $\mu_0 - \mu_{20}$: $\bar{x} - \bar{y} = 28.6$ **Is this difference real?**

Air Bag Example - Standard Error of the Difference

One way to check if the observed difference is not just due to chance is to calculate the standard error of the difference (SED):

$$\text{SED} = \sqrt{\text{Var}(\bar{x} - \bar{y})}.$$

Because the two samples are independent, the variance of the sum is the sum of the variances:

$$\text{Var}(\bar{x} - \bar{y}) = \text{Var}(\bar{x}) + \text{Var}(\bar{y}) = \sigma^2/m + \sigma^2/n.$$

Therefore, we can estimate the SED by plugging in the estimated value of σ:

$$\text{SED} = \sqrt{18.98 + 23.725} = 6.5349.$$

This standard error is small relative to the observed difference (28.6), so the observed difference is not likely due to chance alone.

Linear Models and the Need for a More General Approach

A linear model is a predictive model where the expected value of the response or outcome variable can be expressed as a linear combination of predictor variables.

Noise or error is modelled by adding it to the expected value.

Normally, if the noise is modelled at all, it is modelled as a normal random variable with mean 0.

Usually, linear models can be fit using least-squares methods, where a linear system of equations must be solved in order to find the parameter estimates.

Examples: simple and multiple linear regression; autoregressive time series models

For count data and many other kinds of data, normality is not realistic.

How Generalized Linear Models Differ from Linear Models

A generalized linear model is a predictive model where the expected value of the response variable is a *function* of a linear combination of predictor variables.

Noise or error is modelled with specific distributions.

Count data is better modelled with binomial, Poisson, or negative binomial, and so on.

If continuous measurements are always positive, such as time until failure, other models, such as Weibull or lognormal are appropriate.

Generalized linear models are usually fit using maximum likelihood estimation or a related method.

Poisson Regression

The `cigbutts` data set (in the *MPV* package) gives counts of cigarette butts at locations along a sidewalk as a function of distance from a smoking gazebo.

We can use Poisson regression to model this count data as a function of distance.

Poisson Distribution

Recall:

the Poisson distribution with mean λ is given by

$$P(X = x) = \frac{e^{-\lambda}\lambda^x}{x!}.$$

A common approach to fitting the Poisson distribution regression is to include a predictor x through a *log-linear* relationship:

$$\log(\lambda) = \beta_0 + \beta_1 x. \tag{1}$$

The log function used in this way is referred to as the *link* function since it links the distribution parameter λ with the predictor x.

This link function is preferred here because it recognizes that λ must be positive, while the right hand side of the above equation can take any real value.

Likelihood for the cigarette butts data

If X_i is the number of butts observed at the ith location, its expected value would be λ_i and the corresponding distance might be denoted as d_i.

The model for the ith observation is really

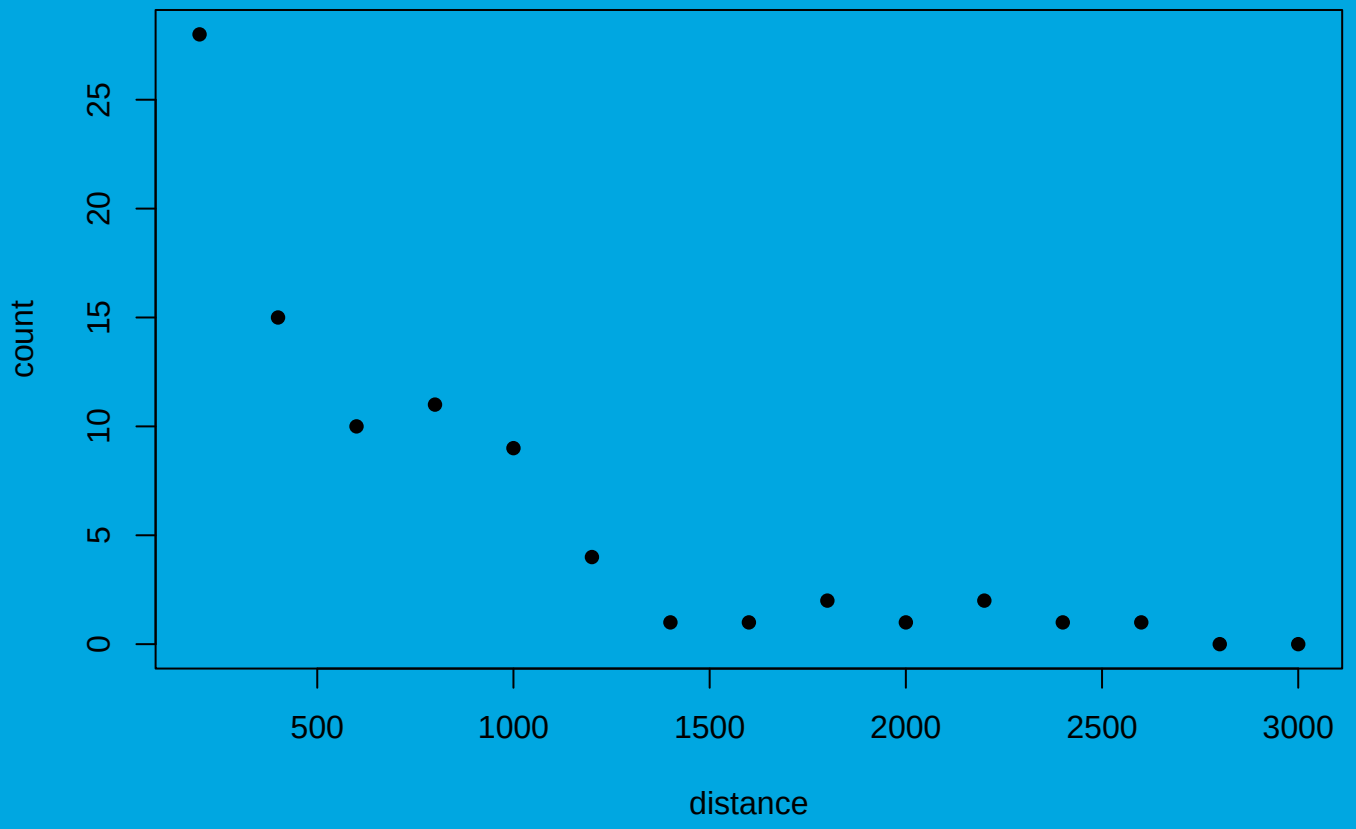
$$P(X_i = x_i) = \frac{e^{-\lambda_i}\lambda_i^{x_i}}{x_i!}.$$

The likelihood function is the product of such probabilities

The log likelihood is then the sum of the logs of the probabilities:

$$\log L = -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} x_i \log \lambda_i - \sum_{i=1}^{n} \log x_i!.$$

Visualizing the Cigarette Butt Counts



Scatterplot of log of cigarette butt counts versus log of distance.

Predicting Counts Using Distance

We suspect a linear relation between the log of λ_i and d_i:

$$\log(\lambda_i) = \beta_0 + \beta_1 d_i$$

Let's suppose we know that $\beta_0 = 3.55$. Then we could say that

$$\log(\lambda_i) = 3.55 + \beta_1 d_i$$

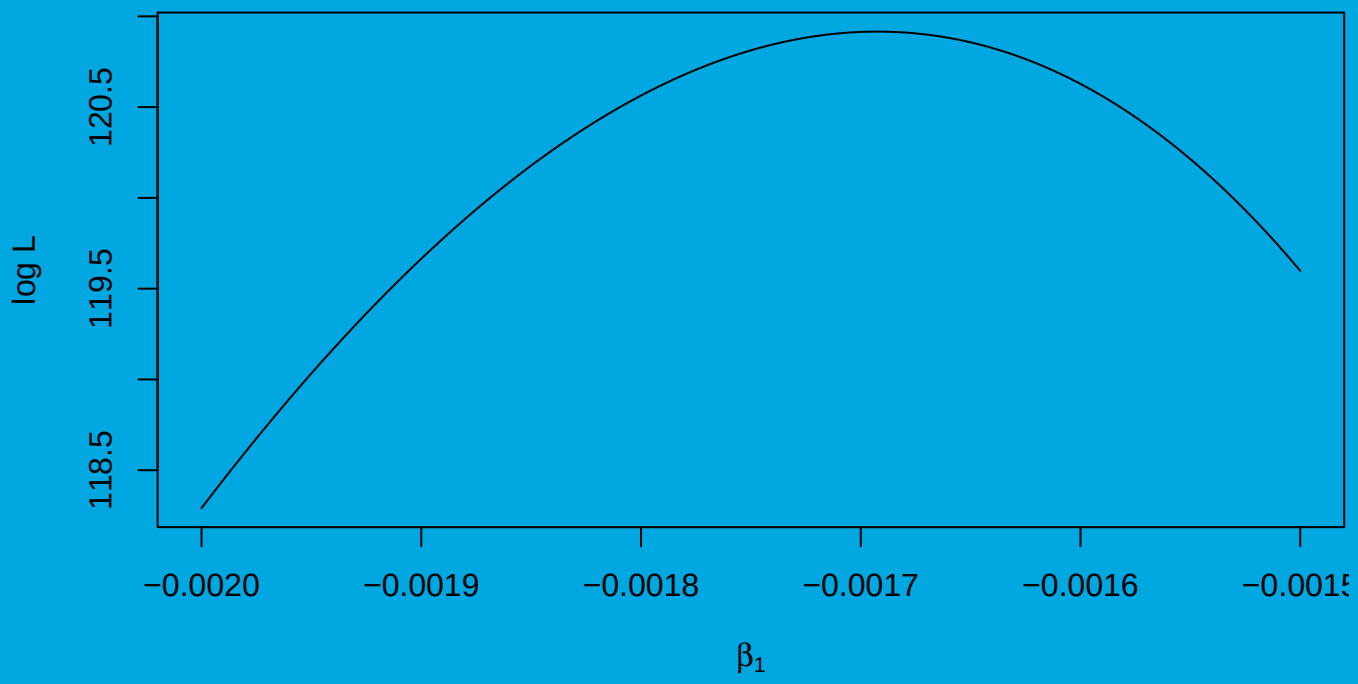
Plugging this into the log likelihood expression gives

$$\log L = -\sum_{i=1}^{n} e^{3.55+\beta_1 d_i} + \sum_{i=1}^{n} x_i(3.55 + \beta_1 d_i) - \sum_{i=1}^{n} \log x_i!.$$

The x_i's are the observed counts and the d_i's are the distances, so we can plot this as a function of β.

Predicting Counts Using Distance

Log likelihood function:



Log likelihood curve for the cigarette butts example. The maximizer is near -.0017.

A predictive model for log of the expected number of cigarette butts would be

$$\widehat{\log(\lambda)} = 3.55 - .0017d$$

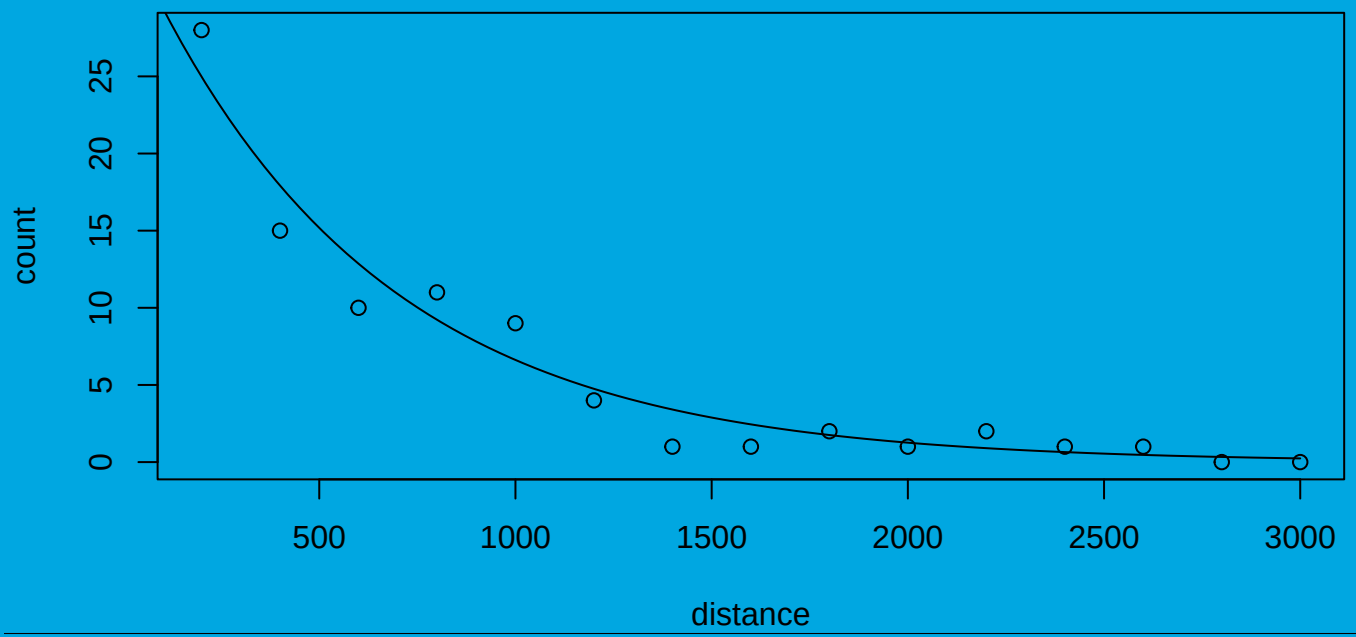
where d is distance from the gazebo.

Using Built-In Software

The `glm()` function conducts maximum likelihood estimation for Poisson regressions and logistic regressions fairly straightforwardly.

```
cig.glm <- glm(x ~ d, family = poisson)
coef(cig.glm)

## (Intercept)           d
##    3.553514   -0.001696
```



Additional output:

```
summary(cig.glm)$coefficients

##               Estimate Std. Error
## (Intercept)  3.553514  0.1735102
## d           -0.001696  0.0002009
##              z value  Pr(>|z|)
## (Intercept)  20.480 3.237e-93
## d            -8.442 3.127e-17
```

Note, in particular, the standard errors of the parameter estimates.

Simulating from the Fitted Model

Recall that we simulate n Poisson random numbers using `rpois(n, lambda)`.

If we have a vector of n predictor values in `x`, and a fitted model of the form

$$\log(\lambda) = \beta_0 + \beta_1 x$$

we can simulated `n` corresponding Poisson responses, using

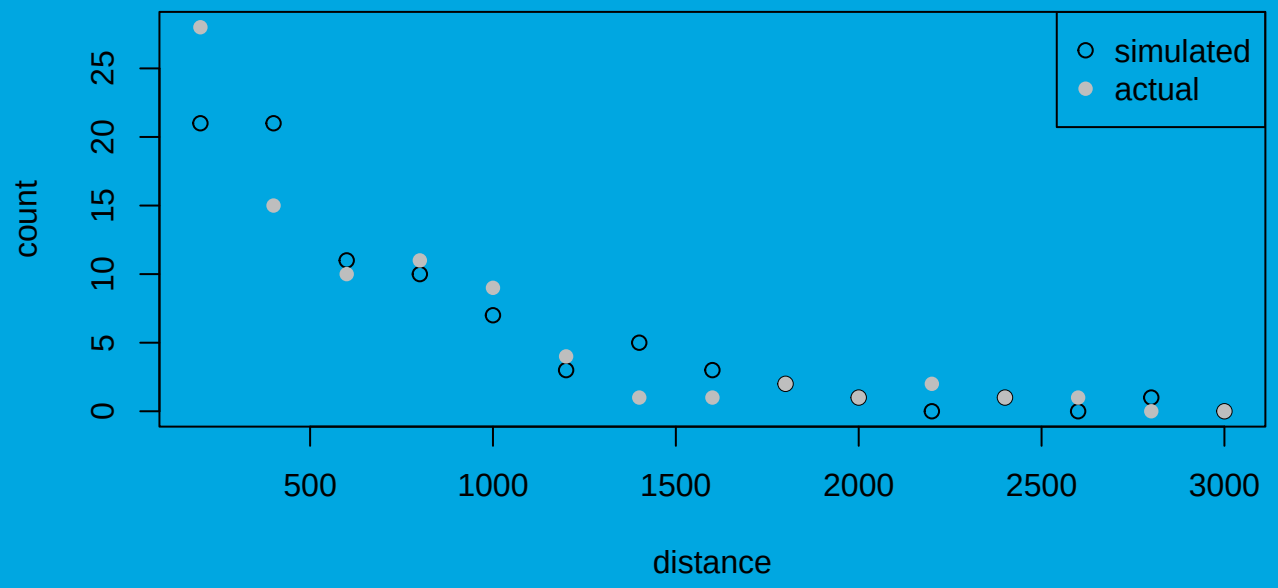
$$\lambda = \mathbf{e}^{\beta_0 + \beta_1 x}.$$

To simulate the numbers of cigarette butts as a function of distance, use

```
lambda <- exp(3.55 - 0.00166*d)
n <- nrow(cigbutts) # No. of observations in the data set
simcounts <- rpois(n, lambda = lambda)
simcigbutts <- data.frame(count = simcounts, distance = d)
```

Simulating from the Fitted Model

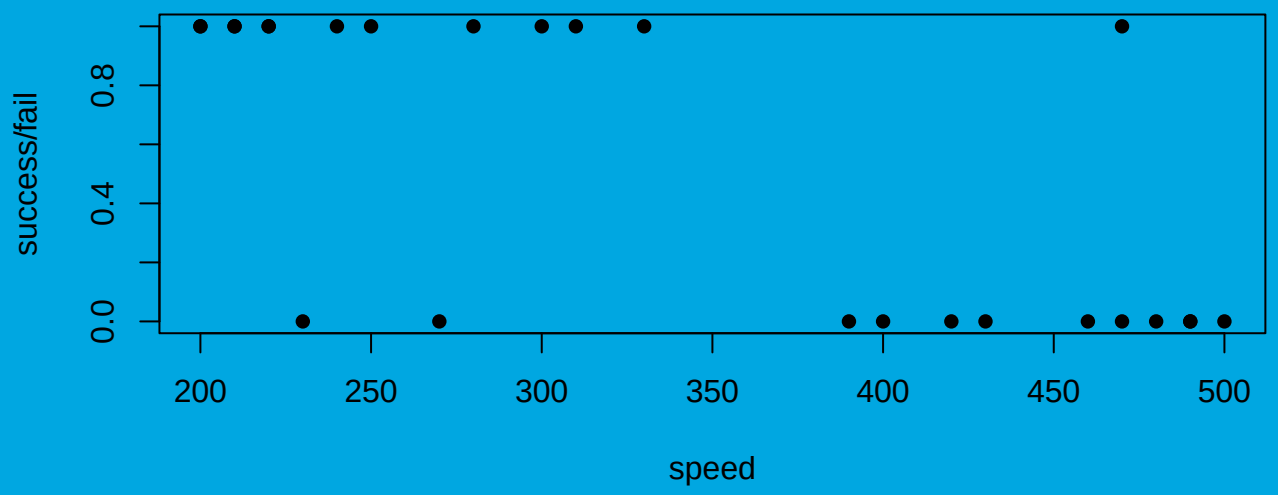
```
plot(count ~ distance, data = simcigbutts,
    ylim = range(simcigbutts$count, cigbutts$count))
points(count ~ distance, data = cigbutts, col="grey", pch=16)
legend("topright", legend = c("simulated", "actual"), pch=c(1, 16),
    col=c("black", "grey"))
```



Why might this kind of simulation exercise be useful?

The data in `p13.1` in the *MPV* package describes successes and failures of surface-to-air missiles as they relate to target speed.

```
library(MPV)
plot(p13.1, xlab = "speed", ylab = "success/fail", pch=16)
```



Surface-to-air missile successes (1) and failures (0) as they relate to target speed (in knots).

Modelling Binary Responses

The first observation to make is that fitting a straight line to such data makes no sense, since the plotted points do not at all scatter about such a line.

Furthermore, if such a line were to be fit to the data, it would necessarily take values outside the interval $[0, 1]$ on subsets of the domain; interpretation of such values would be difficult.

In fact, the preferred interpretation of output arising from the fitting of models to such data is that of probability.

That is, useful models can provide answers to questions such as, "What is the probability of success at a given target speed?"

Since probabilities must lie within the interval $[0, 1]$, we must consider models based on nonlinear functions.

There are many functions which have values in $[0, 1]$.

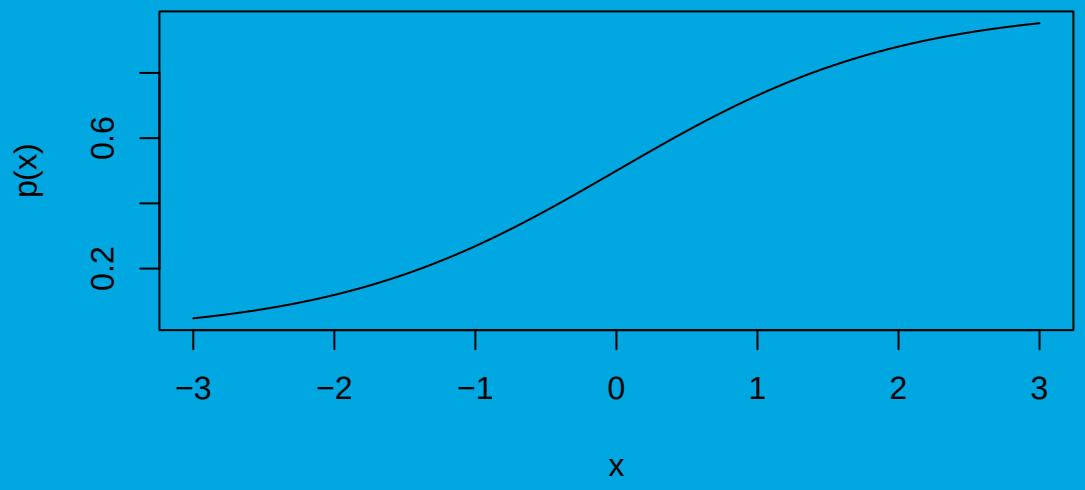
For the current example, we might reasonably believe that the probability of success decreases as target speed increases.

Perhaps the most popular function for this purpose is the *logistic* function

$$p(x) = \frac{e^x}{e^x + 1}.$$

Modelling Binary Responses

```
curve(exp(x)/(1 + exp(x)), from = -3, to = 3, ylab="p(x)")
```



The logistic function.

A bit of algebra allows us to express x in terms of p, yielding the *logit* function:

$$\textbf{logit}(p) = \log\left(\frac{p}{1-p}\right).$$

While p is restricted to take values between 0 and 1, the logit function can take any possible value, so relating the logit function to a straight line or other linear combination is a possibility. For example,

$$\textbf{logit}(p(x)) = \beta_0 + \beta_1 x$$

which means that we can express the probability of an event in terms of a covariate x, using a linear function, but the probability is related to the linear function through the logit.

The logit is an example of a *link function*, since it links the expected response, in this case the probability $p(x)$ to the linear function of the covariate(s).

To fit the logistic regression model to the missile success data, try

```
p13.glm <- glm(y ~ x, data = p13.1, family = binomial)
```

Note that we did not specify the link function; the default choice with the binomial family is the logit.

```
coef(p13.glm)

##  (Intercept)            x
##       6.0709      -0.0177
```

The Coefficient part of the output tells us that the logit of the probability of success as a linear function of target speed has intercept 6.07 and slope -.0177.

More output:

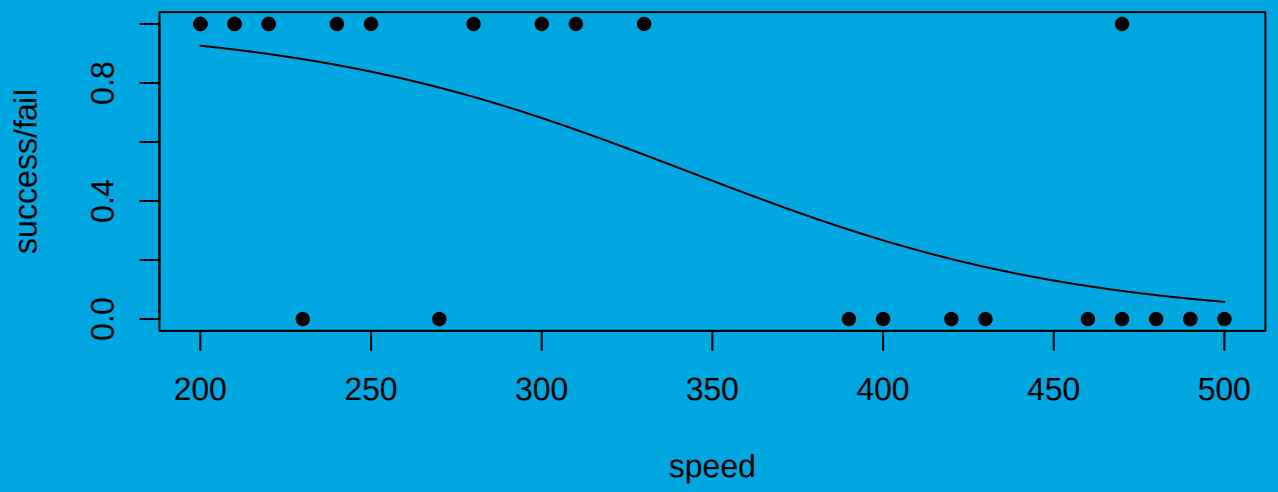
```
summary(p13.glm)$coefficients

##               Estimate Std. Error z value
## (Intercept)    6.0709   2.108996   2.879
## x             -0.0177   0.006076  -2.914
##               Pr(>|z|)
## (Intercept) 0.003995
## x           0.003567
```

Standard error estimates for these parameter estimates are supplied and indicate, in particular, that the slope is clearly negative.

Modelling Binary Responses - Visualizing the Model

```
plot(p13.1, xlab = "speed", ylab = "success/fail", pch=16)
newspeeds <- 200:500 # predict at these target speeds
lines(newspeeds, predict(p13.glm,
  newdata=data.frame(x = newspeeds),  type = "response"))
```



Surface-to-air missile successes (1) and failures (0) as they relate to target speed (in knots) with overlaid logistic curve.

Simulating from the Fitted Model

Recall that we simulate n Bernoulli random numbers using `rbinom(n, 1, p)`.

If we have a vector of n predictor values in `x`, and a fitted model of the form

$$\text{logit}(p) = \beta_0 + \beta_1 x$$

we can simulated `n` corresponding Bernoulli responses, using

$$p = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}}.$$

To simulate the numbers of missile successes as a function of speed, use

```
p <- exp(6.0709-0.0177*p13.1$x)/(1+exp(6.0709-0.0177*p13.1$x)
n <- nrow(p13.1) # No. of observations in the data set
simy <- rbinom(n, 1, prob = p)
simp13.1 <- data.frame(y = simy, x = p13.1$x)
```

Simulating from the Fitted Model

```
plot(y ~ x, data = simp13.1)
points(y ~ x, data = p13.1,  col="grey", pch=16)
legend(450, .5, legend = c("simulated", "actual"), pch=c(1, 16),
    col=c("black", "grey"))
```