Maximum Likelihood Estimation

Daniel B. Rowe, Ph.D.

Professor
Department of Mathematical and Statistical Sciences

Maximum Likelihood Estimation

We have been saying that $y \sim N(\mu,\sigma^2)$,

when what we actually mean is that $y = \mu + \varepsilon$ where

$\varepsilon \sim N(0,\sigma^2)$.

That is, y has some true underling value μ,

but there is additive measurement error (noise) .

We know that if $\varepsilon \sim N(0,\sigma^2)$, then from a linear

transformation of variable, we get $y \sim N(\mu,\sigma^2)$.

Maximum Likelihood Estimation - Mean

If we have a random sample of size n with $y=\mu+\varepsilon$, where $\varepsilon \sim N(0,\sigma^2)$.

Then we have $y_i=\mu+\varepsilon_i$, $\varepsilon_i \sim N(0,\sigma^2)$ for i=1,…,n.

Since these are independent observations,

the joint distribution is

$$f(y_1,...,y_n\,|\,\mu,\sigma^2)=\frac{\exp[-(y_1-\mu)^2/2\sigma^2]}{(2\pi\sigma^2)^{1/2}}\cdots\frac{\exp[-(y_n-\mu)^2/2\sigma^2]}{(2\pi\sigma^2)^{1/2}}$$

Maximum Likelihood Estimation - Mean

If we have a random sample of size n with $y = \mu + \varepsilon$, where $\varepsilon \sim N(0,\sigma^2)$.

Then we have $y_i = \mu + \varepsilon_i$, $\varepsilon_i \sim N(0,\sigma^2)$ for i=1,…,n.

Since these are independent observations,

the joint distribution is

$$f(y_1,\ldots,y_n \mid \mu,\sigma^2) = (2\pi\sigma^2)^{-n/2}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right]$$
$$= L(\mu,\sigma^2)$$

Maximum Likelihood Estimation - Mean

$L(\mu,\sigma^2)$ is called the likelihood function.

What we want to do is find the values of (μ,σ^2)

that maximize $L(\mu,\sigma^2)$. The value of μ that maximizes

$L(\mu,\sigma^2)$ is the value $\hat{\mu}$ that minimizes $\sum_{i=1}^{n}(y_i-\mu)^2$.

The value of σ² that maximizes $L(\mu,\sigma^2)$ is

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{\mu})^2 .$$

$d_i=y_i-\hat{\mu}$

minimize $\sum_{i=1}^{n}d_i^2$

note n not n-1

Maximum Likelihood Estimation - Mean

$L(\mu,\sigma^2)$ is called the likelihood function.

What we do is differentiate $L(\mu,\sigma^2)$ wrt μ and σ^2, set = 0 and solve. That is,

$$\left.\frac{\partial L(\mu,\sigma^2)}{\partial \mu}\right|_{\hat{\mu},\hat{\sigma}^2}=0 \quad \text{and} \quad \left.\frac{\partial L(\mu,\sigma^2)}{\partial \sigma^2}\right|_{\hat{\mu},\hat{\sigma}^2}=0 \ .$$

The values of μ and σ^2 that maximize $L(\mu,\sigma^2)$ are the maximum likelihood estimators (MLEs).

Maximum Likelihood Estimation - Mean

However, this is messy, but we can instead maximize

$$LL(\mu,\sigma^2)=\ln(L(\mu,\sigma^2))$$

as $\left.\frac{\partial LL(\mu,\sigma^2)}{\partial \mu}\right|_{\hat{\mu},\hat{\sigma}^2}=0$ $\left.\frac{\partial LL(\mu,\sigma^2)}{\partial \sigma^2}\right|_{\hat{\mu},\hat{\sigma}^2}=0$

to obtain MLEs μ and σ^2 because it is a monotonic function.

Maximum Likelihood Estimation - Mean

With $y_i = \mu + \varepsilon_i$ and $\varepsilon_i \sim N(0,\sigma^2)$, ε_i independent,

$$f(y_1,\ldots,y_n \mid \mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right]$$

$$LL(\mu,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2$$

$$\left.\frac{\partial LL(\mu,\sigma^2)}{\partial \mu}\right|_{\hat{\mu},\hat{\sigma}^2} = -\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n} 2(y_i - \hat{\mu})(-1) = 0$$

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

Maximum Likelihood Estimation - Mean

With $y_i = \mu + \varepsilon_i$ and $\varepsilon_i \sim N(0,\sigma^2)$, ε_i independent,

$$f(y_1,\ldots,y_n \mid \mu,\sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2 \right]$$

$$LL(\mu,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - \mu)^2$$

$$\left. \frac{\partial LL(\mu,\sigma^2)}{\partial \sigma^2} \right|_{\hat{\mu},\hat{\sigma}^2} = -\frac{n}{2}\frac{1}{\hat{\sigma}^2} - \frac{-1}{2(\hat{\sigma}^2)^2} \sum_{i=1}^{n} (y_i - \hat{\mu})^2 = 0$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{\mu})^2$$

note n not n-1

Maximum Likelihood Estimation - Mean

Solving for μ and σ^2 yields

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i \quad \text{and} \quad \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2$$

These are MLEs, most probable or modal values.

Note that the denominator is n and not n-1.

This is why we use a denominator n-1.

$\hat{\sigma}^2$ is a biased estimator of σ^2, $E(\hat{\sigma}^2) = \frac{(n-1)}{n}\sigma^2$.

$$\frac{n\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1) \longrightarrow E\left(\frac{n\hat{\sigma}^2}{\sigma^2}\right) = n-1 \longrightarrow E(\hat{\sigma}^2) = \frac{n-1}{n}\sigma^2$$

$$\frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1) \longrightarrow E\left(\frac{(n-1)s^2}{\sigma^2}\right) = n-1 \longrightarrow E(s^2) = \sigma^2$$

Maximum Likelihood Estimation - Mean

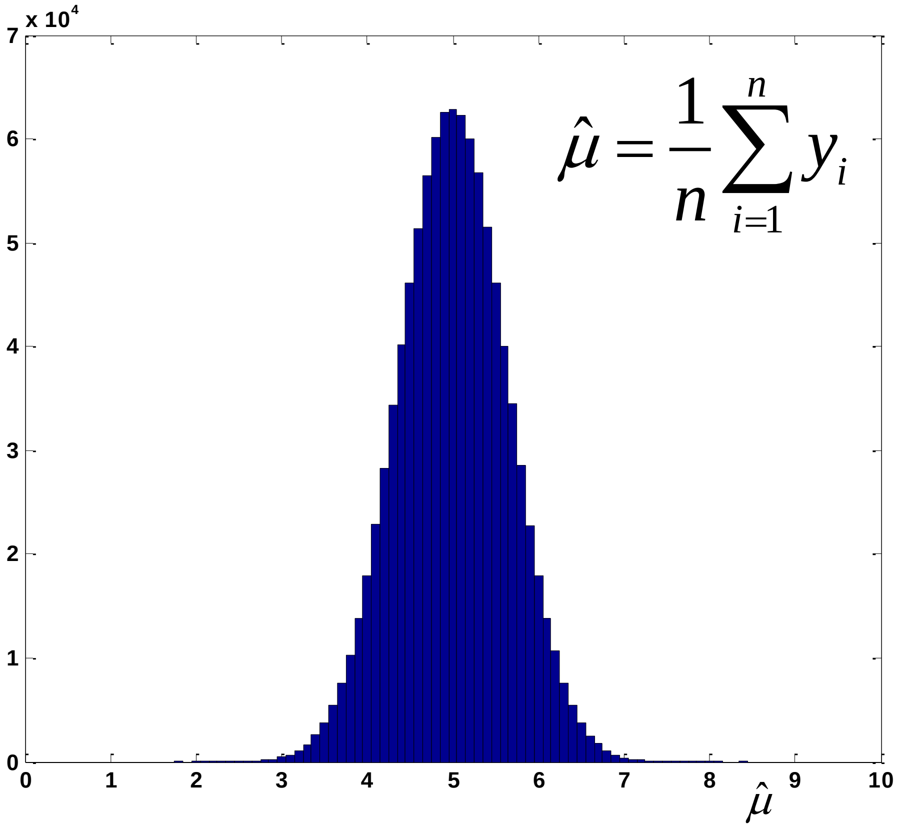
$$\hat{\mu} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$$

$\bar{y}$

```
n=10;, mu=5;, sigma=2;
y=sigma*randn(10^6,n)+mu;
ybar=mean(y,2);
figure(1)
hist(ybar,(0:.1:10)')
axis([0 10 0 70000])
mean(ybar),var(ybar)
```

$\mu = 5$ $\qquad \sigma^2 / n = 0.4$

$\bar{y}_{\hat{\mu}} = 5.0003$ $\qquad s^2_{\hat{\mu}} = 0.3987$



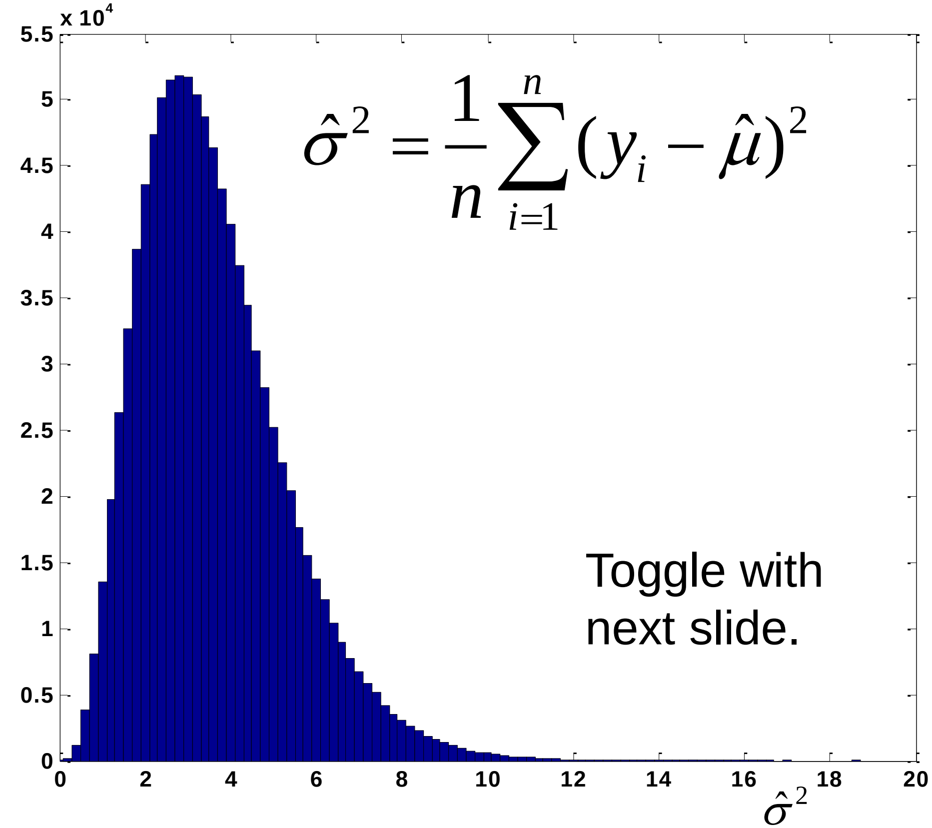
Maximum Likelihood Estimation - Mean

$$\hat{\sigma}^2 \sim \frac{\sigma^2}{n}\chi^2(n-1)$$

$$(n-1)\frac{\sigma^2}{n} = 3.6 \qquad 2(n-1)\frac{\sigma^4}{n^2} = 2.88$$

$$\bar{y}_{\hat{\sigma}^2} = 3.600 \qquad s^2_{\hat{\sigma}^2} = 2.8805$$

```
sigma2hat=var(y',1)';
figure(2)
hist(sigma2hat,(0:.2:20)')
axis([0 20 0 55000])
mean(sigma2hat)
var(sigma2hat)
```



$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2$$

Toggle with next slide.

horizontal- axis scale →

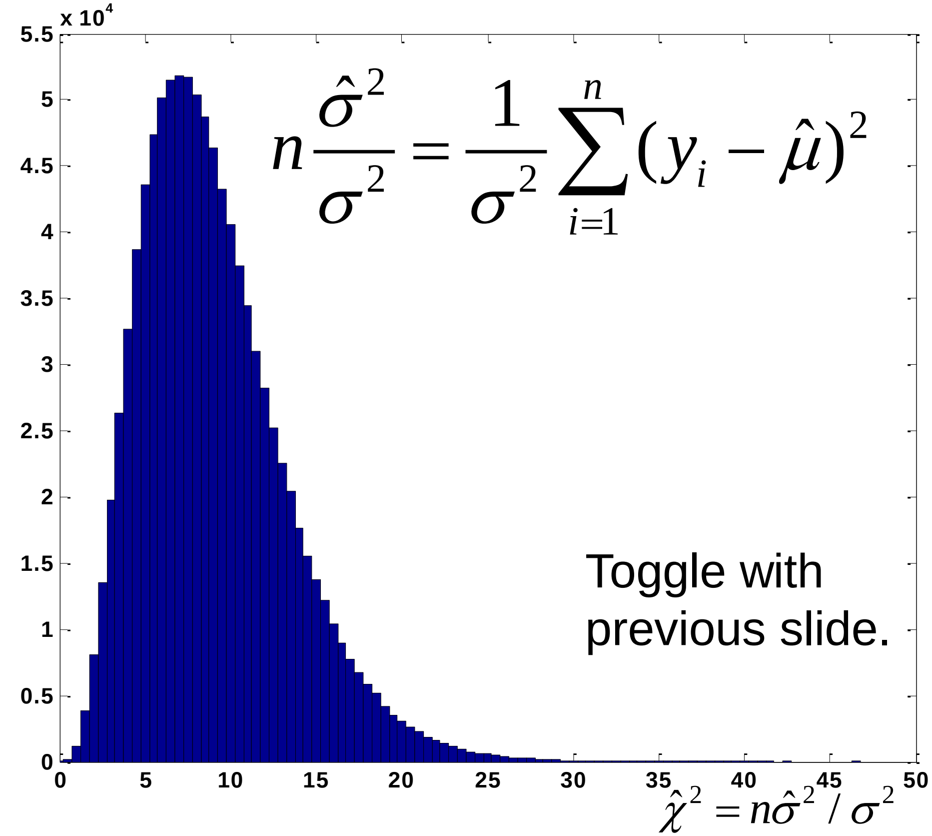
Maximum Likelihood Estimation - Mean

$$\hat{\chi}^2 = n\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-1)$$

$(n-1)=9$ $2(n-1)=18$

$\bar{y}_{\hat{\chi}^2} = 9.000$ $s^2_{\hat{\chi}^2} = 18.0031$

```
chi2=n*sigma2hat/sigma^2;
figure(3)
hist(chi2,(0:.5:50)')
axis([0 50 0 55000])
mean(chi2)
var(chi2)
```



horizontal- axis scale →

Maximum Likelihood Estimation - Linear

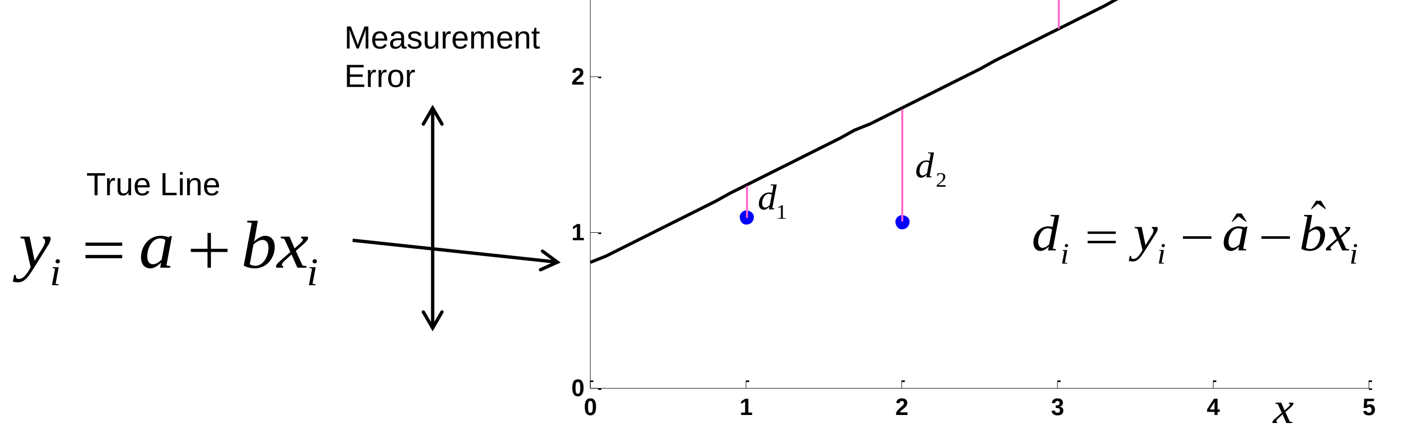
This technique, can be generalized to linear regression.

Let $y_i = a + bx_i + \varepsilon_i$,

where $\varepsilon_i \sim N(0,\sigma^2)$

are independent.

$i = 1,\ldots,n$



Maximum Likelihood Estimation - Linear

This technique, can be generalized to linear regression.

Let $y_i = a + bx_i + \varepsilon_i$,

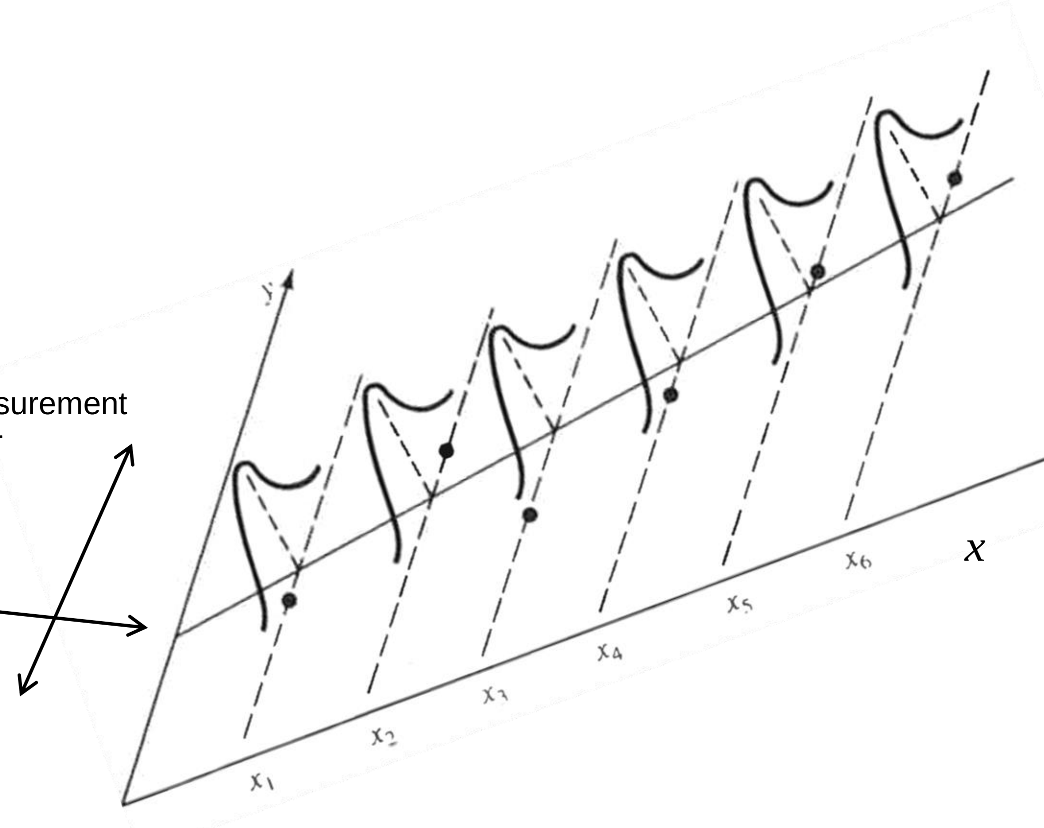
where $\varepsilon_i \sim N(0, \sigma^2)$

are independent.

$i = 1, \ldots, n$

Measurement
Error

True Line
$y_i = a + bx_i$



Maximum Likelihood Estimation - Linear

This technique, can be generalized to linear regression.

Let $y_i = a + bx_i + \varepsilon_i$, where $\varepsilon_i \sim N(0,\sigma^2)$ are independent.

Then, the likelihood is

$$f(y_1,\ldots,y_n \mid a,b,\sigma^2) = \frac{\exp[-(y_1 - a - bx_1)^2 / 2\sigma^2]}{(2\pi\sigma^2)^{1/2}} \cdots \frac{\exp[-(y_n - a - bx_n)^2 / 2\sigma^2]}{(2\pi\sigma^2)^{1/2}}$$

Maximum Likelihood Estimation - Linear

This technique, can be generalized to linear regression.

Let $y_i = a + bx_i + \varepsilon_i$, where $\varepsilon_i \sim N(0,\sigma^2)$ are independent.

Then, the likelihood is

$$f(y_1,\ldots,y_n \mid a,b,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2} \sum_{i=1}^{n} (y_i - a - bx_i)^2 \right]$$

and the log likelihood is

$$LL(a,b,\sigma^2) = \underbrace{-\frac{n}{2}\log(2\pi)}_{\text{no } a \text{ or } b} \underbrace{-\frac{n}{2}\log(\sigma^2)}_{\text{no } a \text{ or } b} \underbrace{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2}_{a \text{ or } b} .$$

Maximum Likelihood Estimation - Linear

$L(a,b,\sigma^2)$ is again called the likelihood function.

What we want to do is find the values of (a,b,σ^2)

that maximize $L(a,b,\sigma^2)$. The values (a,b) that maximize

$L(a,b,\sigma^2)$ are the values $(\hat{a},\hat{b})$ that minimize $\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2$.

The value of σ^2 that maximizes $L(a,b,\sigma^2)$ is

$$\hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2 .$$

$d_i=y_i-\hat{a}-\hat{b}x_i$

minimize $\sum_{i=1}^{n}d_i^2$

wrt a, b

Maximum Likelihood Estimation - Linear

Differentiate $LL(a,b,\sigma^2)$ wrt a, b, and σ^2, then set = 0

$$LL(a,b,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2$$

$$\left.\frac{\partial LL(a,b,\sigma^2)}{\partial a}\right|_{\hat{a},\hat{b},\hat{\sigma}^2} = -\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}2(y_i - \hat{a} - \hat{b}x_i)(-1) = 0$$

$$\left.\frac{\partial LL(a,b,\sigma^2)}{\partial b}\right|_{\hat{a},\hat{b},\hat{\sigma}^2} = -\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}2(y_i - \hat{a} - \hat{b}x_i)(-x_i) = 0$$

$$\left.\frac{\partial LL(a,b,\sigma^2)}{\partial \sigma^2}\right|_{\hat{a},\hat{b},\hat{\sigma}^2} = -\frac{n}{2}\frac{1}{\hat{\sigma}^2} - \frac{-1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2 = 0$$

Maximum Likelihood Estimation - Linear

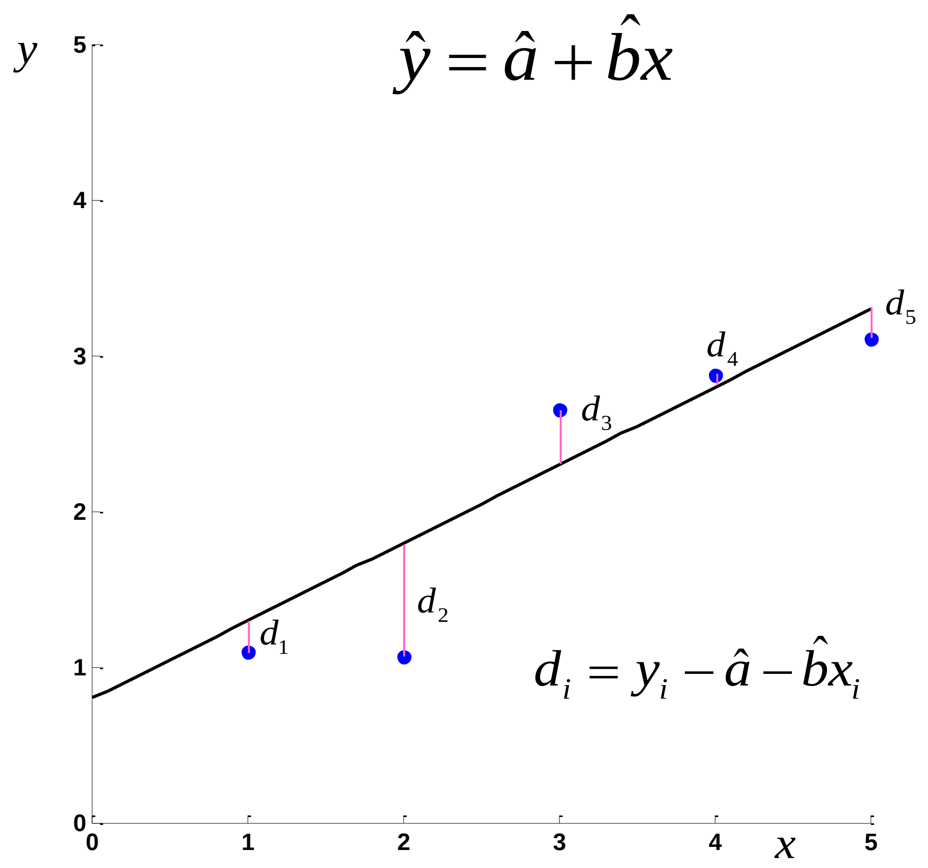
Solving for the estimated parameters yields

$$\hat{b} = \frac{n(\sum_{i=1}^{n} x_i y_i) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$\hat{a} = \frac{(\sum_{i=1}^{n} y_i)(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)(\sum_{i=1}^{n} x_i y_i)}{n(\sum_{i=1}^{n} x_i^2) - (\sum_{i=1}^{n} x_i)^2}$$

$$\hat{a} = \bar{y} - \hat{b}\bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$



Maximum Likelihood Estimation - Linear

The regression model $y_i = a + bx_i + \varepsilon_i$ where $\varepsilon_i \overset{iid}{\sim} N(0,\sigma^2)$

$i = 1,\ldots,n$

that we presented, can be equivalently written as

$$y = X\beta + \varepsilon \quad \text{where}$$

design matrix

regression coefficients

measurement error

measured data

$$\underset{n\times 1}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \underset{n\times 2}{X} = \begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}, \quad \underset{2\times 1}{\beta} = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \underset{n\times 1}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

and $\varepsilon \sim N(0,\sigma^2 I_n)$. I_n is an n-dimensional identity matrix.

Maximum Likelihood Estimation - Linear

The regression model $y = X\beta + \varepsilon$ where $\varepsilon \sim N(0, \sigma^2 I_n)$.

$$\underset{n\times 1}{\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}} = \underset{n\times 2}{\begin{pmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}} \underset{2\times 1}{\begin{pmatrix} a \\ b \end{pmatrix}} + \underset{n\times 1}{\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}}$$

$$y_i = a + bx_i + \varepsilon_i$$

Maximum Likelihood Estimation - Linear

With $\underset{n\times 1}{y} = X\beta + \varepsilon$ and $\underset{n\times 1}{\varepsilon} \sim N(0, \sigma^2 I_n)$

The likelihood is

$$f(y_1,\ldots,y_n \mid a,b,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right]$$

compare to

$$f(y_1,\ldots,y_n \mid a,b,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - a - bx_i)^2\right]$$

and the log likelihood is

$$LL(a,b,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta).$$

Maximum Likelihood Estimation - Linear

$L(\beta,\sigma^2)$ is again called the likelihood function.

What we want to do is find the values of (β,σ^2)

that maximize $L(\beta,\sigma^2)$. The value of β that maximizes

$L(\beta,\sigma^2)$ is the value $\hat{\beta}$ that minimizes $(y-X\beta)'(y-X\beta)$.

The value of σ^2 that maximizes $L(\beta,\sigma^2)$ is

$$\hat{\sigma}^2 = \frac{1}{n}(y-X\hat{\beta})'(y-X\hat{\beta})$$

$d_i = y_i - \hat{a} - \hat{b}x_i$

minimize $(y-X\beta)'(y-X\beta)$
wrt β

We need to find $\hat{\beta}$.

Maximum Likelihood Estimation - Linear

We don't need to take the derivative of $L(\beta,\sigma^2)$

wrt β (although we could). We can write with algebra

$$(y-X\beta)'(y-X\beta)=(y-X\hat{\beta})'(y-X\hat{\beta})+(\beta-\hat{\beta})'(X'X)(\beta-\hat{\beta})$$

add and subtract $X\hat{\beta}$ invertible does not depend on β

where $\hat{\beta}=(X'X)^{-1}X'y$. It can be seen that $\beta=\hat{\beta}$ maximizes $LL(\beta,\sigma^2)$ because it makes $(y-X\beta)'(y-X\beta)$ smallest

$$LL(\beta,\sigma^2)=-\frac{n}{2}\log(2\pi)-\frac{n}{2}\log(\sigma^2)$$

$$-\frac{1}{2\sigma^2}\left[(y-X\hat{\beta})'(y-X\hat{\beta})+(\beta-\hat{\beta})'(X'X)(\beta-\hat{\beta})\right]$$

Maximum Likelihood Estimation - Linear

More generally, we can have a multiple regression model

$\underset{n\times 1}{y} = X\beta + \varepsilon$ where $\underset{n\times 1}{\varepsilon} \sim N(0,\sigma^2 I_n)$ and

measured data

design matrix

regression coefficients

measurement error

$$\underset{n\times 1}{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad \underset{n\times(q+1)}{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1q} \\ 1 & x_{21} & , & x_{2q} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nq} \end{pmatrix}, \quad \underset{(q+1)\times 1}{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_q \end{pmatrix}, \quad \underset{n\times 1}{\varepsilon} = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}.$$

Maximum Likelihood Estimation - Linear

The MLEs are the same,

$$\underset{(q+1)\times 1}{\hat{\beta}} = (X'X)^{-1}X'y \quad \text{and} \quad \underset{1\times 1}{\hat{\sigma}^2} = \frac{1}{n}(y - X\hat{\beta})'(y - X\hat{\beta}) \ .$$

In addition,

$$\underset{(q+1)\times 1}{\hat{\beta}} \sim N\left(\beta, \sigma^2 (X'X)^{-1}\right) \quad \text{and} \quad n\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-q-1) \ .$$

$$\underbrace{(y - X\beta)'(y - X\beta)}_{\chi^2(n)} = \underbrace{(y - X\hat{\beta})'(y - X\hat{\beta})}_{\chi^2(n-q-1)} + \underbrace{(\beta - \hat{\beta})'(X'X)(\beta - \hat{\beta})}_{\chi^2(q+1)}$$

could + & - $X\hat{\beta}$

independent

This means we should use a denominator of n-q-1 for unbiased estimator of σ^2.

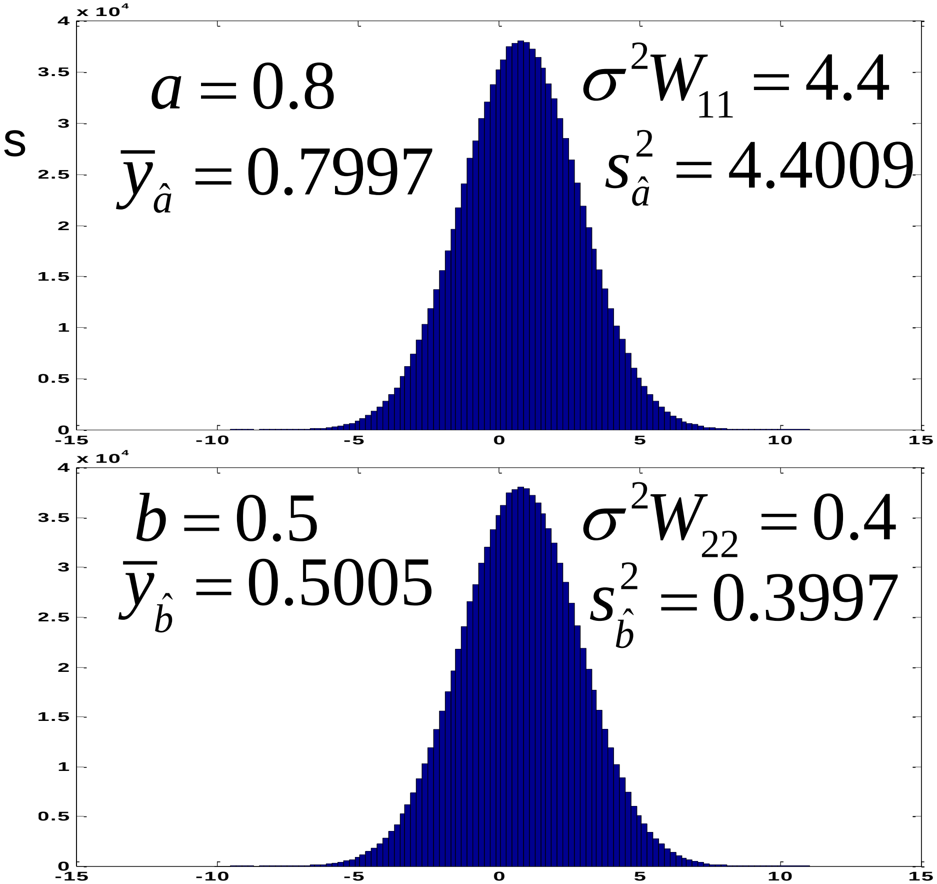
Maximum Likelihood Estimation - Linear

Let $\beta = (a,b)'$, $X = (1, x)$, then

$$W = (X'X)^{-1}$$

$$\underset{(q+1)\times 1}{\hat{\beta}} \sim N\left(\beta, \sigma^2 (X'X)^{-1}\right)$$

Column of settings




```
num=10^6; a=.8;b=.5;, sigma=2;
x=[1,2,3,4,5]'; n=length(x);
mu=a+b*x';, X=[ones(n,1),x];
y=sigma*randn(num,n)...
+ones(num,1)*mu;
betahat=inv(X'*X)*X'*y';
figure(1), hist(betahat(1,:),(-10:.2:10)')
figure(2), hist(betahat(2,:),(-5:.1:5)')
betabar=mean(betahat,2);
varbetahat=var(beta,hat1,2);
```

cov(a,b)=-1.2 corr(a,b)= -0.9045

Maximum Likelihood Estimation - Linear

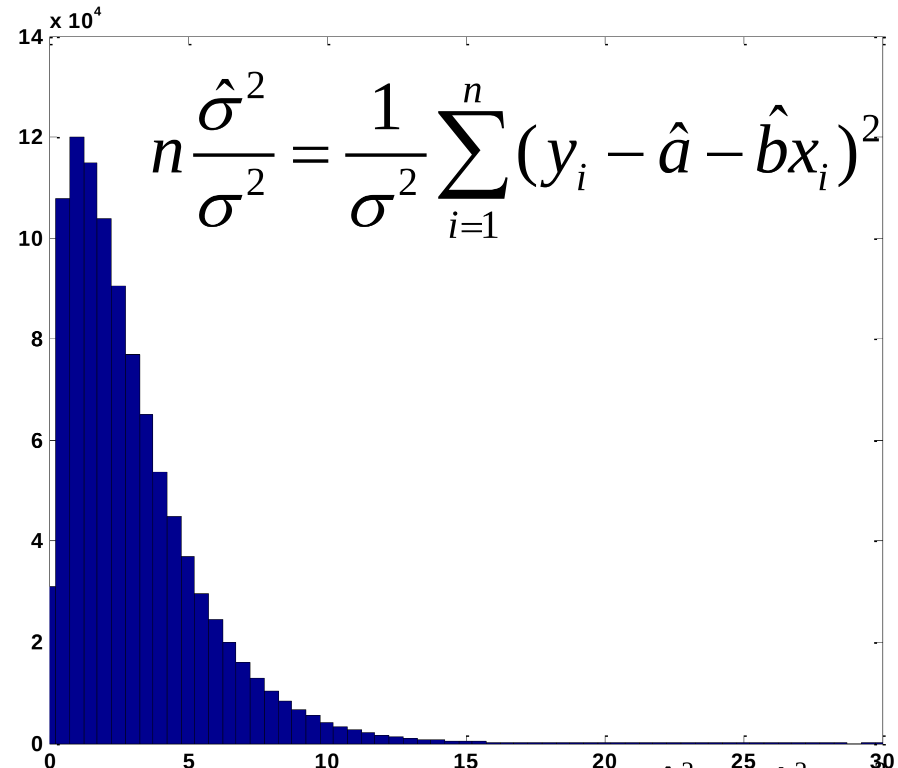
$$\hat{\chi}^2 = n\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2(n-2)$$

$$n-q-1 \qquad 2(n-q-1)$$

$$(n-2)=3 \qquad 2(n-2)=6$$

$$\bar{y}_{\hat{\sigma}^2} = 3.0006 \qquad s^2_{\hat{\sigma}^2} = 6.0038$$

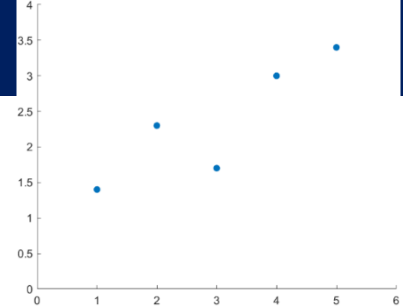
```
resid=y-(X*betahat)';
sigma2hat=var(resid',1)';
chi2=n*sigma2hat/sigma^2;
figure(3)
hist(chi2,(0:.5:30)')
xlim([0 30])
mean(chi2), var(chi2)
```



$$n\frac{\hat{\sigma}^2}{\sigma^2} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$

$$\hat{\chi}^2 = n\hat{\sigma}^2 / \sigma^2$$

Example



Given observed data (1,1.4), (2,2.3), (3,1.7), (4,3.0), (5,3.4). Estimate the slope, y-intercept, and residual variance.

Method 1

$$y = \begin{bmatrix} 1.4 \\ 2.3 \\ 1.7 \\ 3.0 \\ 3.4 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 4 \\ 1 & 5 \end{bmatrix}$$

$$X'X = \begin{bmatrix} 5 & 15 \\ 15 & 55 \end{bmatrix} \qquad (X'X)^{-1} = \begin{bmatrix} 1.1 & -0.3 \\ -0.3 & 0.1 \end{bmatrix}$$

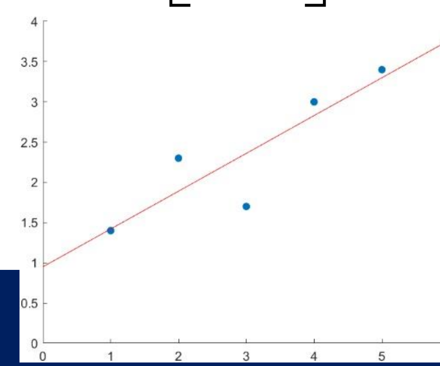
$$(X'X)^{-1}X' = \begin{bmatrix} 0.8 & 0.5 & 0.2 & -0.1 & -0.4 \\ -0.2 & -0.1 & -0.0 & 0.1 & 0.2 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'y = \begin{bmatrix} 0.95 \\ 0.47 \end{bmatrix}$$

$$\hat{\beta} = (X'X)^{-1}X'y$$

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$

$$\hat{\sigma}^2 = 0.1286$$



Example

Given observed data (1,1.4), (2,2.3), (3,1.7), (4,3.0), (5,3.4). Estimate the slope, y-intercept, and residual variance.

Because the likelihood $L(a,b,\sigma^2)=(2\pi\sigma^2)^{-\frac{n}{2}}\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-a-bx_i)^2\right]$

is maximized when we select (a,b) to minimize $\sum_{i=1}^{n}(y_i-a-bx_i)^2$,

we can set up a score function $Q=\frac{1}{n}\sum_{i=1}^{n}(y_i-a-bx_i)^2$ (i.e. σ^2)

and try (a,b) combinations to see which make Q smallest.

Example

$$\hat{y} = \hat{a} + \hat{b}x$$

Given observed data (1,1.4), (2,2.3), (3,1.7), (4,3.0), (5,3.4).
Numerically get slope, y-intercept, and residual variance.

Method 2

Select a_{min}, a_{max}, b_{min}, and b_{max} values. Use $\Delta a = \Delta b = .01$.

can make smaller

Perform an exhaustive brute force grid search.

Compute $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$ for each (a,b) combination.

Find the a and b combination that make σ^2 the smallest.

The a and b that min σ^2 are $\hat{a}$ and $\hat{b}$, and the σ^2 is $\hat{\sigma}^2$.

Example

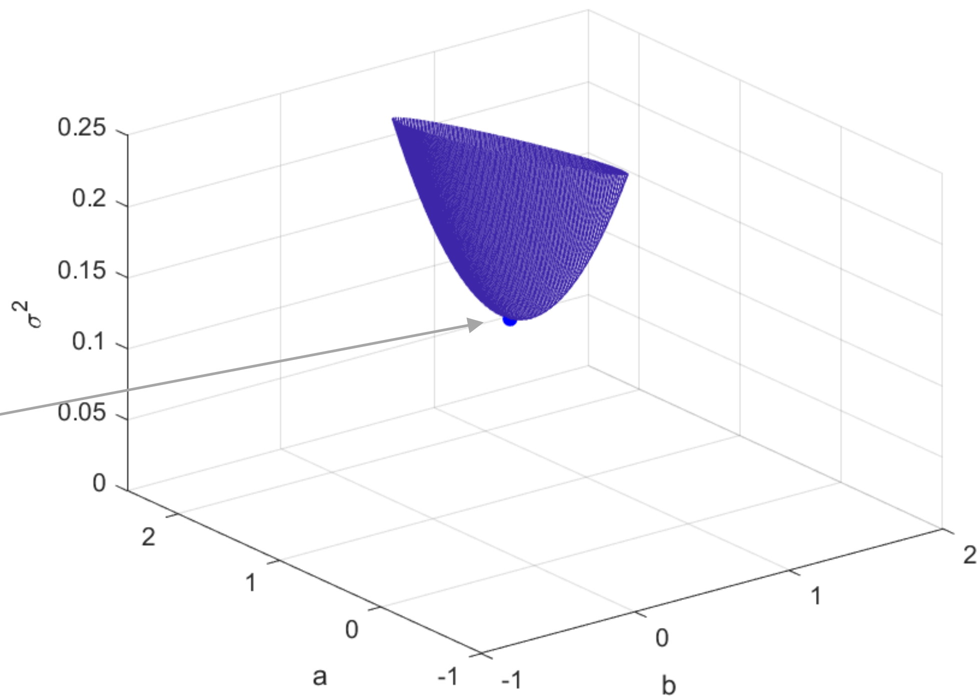
$$\hat{y} = \hat{a} + \hat{b}x$$

Given observed data (1,1.4), (2,2.3), (3,1.7), (4,3.0), (5,3.4). Numerically get slope, y-intercept, and residual variance.

Method 2

a_{min} =-1.0 a_{max} = 2.0
b_{min} =-1.0 b_{max} = 2.5

Compute σ^2 for each (a,b) combination. Make surface.



$$\hat{\beta} = \begin{bmatrix} 0.95 \\ 0.47 \end{bmatrix} \qquad \hat{\sigma}^2 = 0.1286$$

Example

Given observed data (1,1.4), (2,2.3), (3,1.7), (4,3.0), (5,3.4). Use Gradient Descent to iteratively find the $(\hat{a}, \hat{b})$ that

Method 3

that minimize $Q = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$.

$$\frac{dQ}{da} = \frac{2}{n}\left[-\sum_{i=1}^{n} y_i + an + b\sum_{i=1}^{n} x_i\right]$$

$$\frac{dQ}{db} = \frac{2}{n}\left[-\sum_{i=1}^{n} x_i y_i + a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2\right]$$

$$S_x = \sum_{i=1}^{n} x_i \qquad S_y = \sum_{i=1}^{n} y_i \qquad S_{xx} = \sum_{i=1}^{n} x_i^2 \qquad S_{xy} = \sum_{i=1}^{n} x_i y_i$$

Example

Given observed data (1,1.4), (2,2.3), (3,1.7), (4,3.0), (5,3.4).
Use Gradient Descent to iteratively find the $(\hat{a}, \hat{b})$ that

Method 3

that minimize $Q = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$.

$$\frac{dQ}{da} = \frac{2}{n}\left[-S_y + an + bS_x\right]$$

$$\frac{dQ}{db} = \frac{2}{n}\left[-S_{xy} + aSx + bS_{xx}\right]$$

$$\nabla Q = \begin{bmatrix} \frac{dQ}{da} \\ \frac{dQ}{db} \end{bmatrix} \qquad \nabla Q = \frac{2}{n}\begin{bmatrix} -S_y & n & S_x \\ -S_{xy} & S_x & S_{xx} \end{bmatrix}\begin{bmatrix} 1 \\ a \\ b \end{bmatrix}$$

$$S_x = \sum_{i=1}^{n} x_i$$

$$S_y = \sum_{i=1}^{n} y_i$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i$$

Example

Given observed data (1,1.4), (2,2.3), (3,1.7), (4,3.0), (5,3.4).
Use Gradient Descent to iteratively find the $(\hat{a}, \hat{b})$ that

Method 3

that minimize $Q = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$.

$$S_x = \sum_{i=1}^{n} x_i$$

$$S_y = \sum_{i=1}^{n} y_i$$

$$S_{xx} = \sum_{i=1}^{n} x_i^2$$

$$S_{xy} = \sum_{i=1}^{n} x_i y_i$$

$$\nabla Q(\hat{a},\hat{b}) = \frac{2}{n}\begin{bmatrix} -S_y & n & S_x \\ -S_{xy} & S_x & S_{xx} \end{bmatrix}\begin{bmatrix} 1 \\ \hat{a} \\ \hat{b} \end{bmatrix}$$

Start with initial $(\hat{a}^{(0)}, \hat{b}^{(0)})$ or $\hat{\beta}^{(0)}$.

$$\hat{\beta}^{(0)} = \begin{bmatrix} 1 \\ \hat{a}^{(0)} \\ \hat{b}^{(0)} \end{bmatrix}$$

step size

$$\gamma = .0001$$

Calculate new $\hat{\beta}^{(1)} = \hat{\beta}^{(0)} - \gamma \nabla Q(\hat{\beta}^{(0)})$

Calculate new $\hat{\beta}^{(2)} = \hat{\beta}^{(1)} - \gamma \nabla Q(\hat{\beta}^{(1)})$

Continue until convergence $\hat{\beta}^{(k+1)} = \hat{\beta}^{(k)} - \gamma \nabla Q(\hat{\beta}^{(k)})$ at $k=L$.

Example

Given observed data (1,1.4), (2,2.3), (3,1.7), (4,3.0), (5,3.4).
Use Gradient Descent to iteratively find the $(\hat{a}\,,\hat{b})$ that

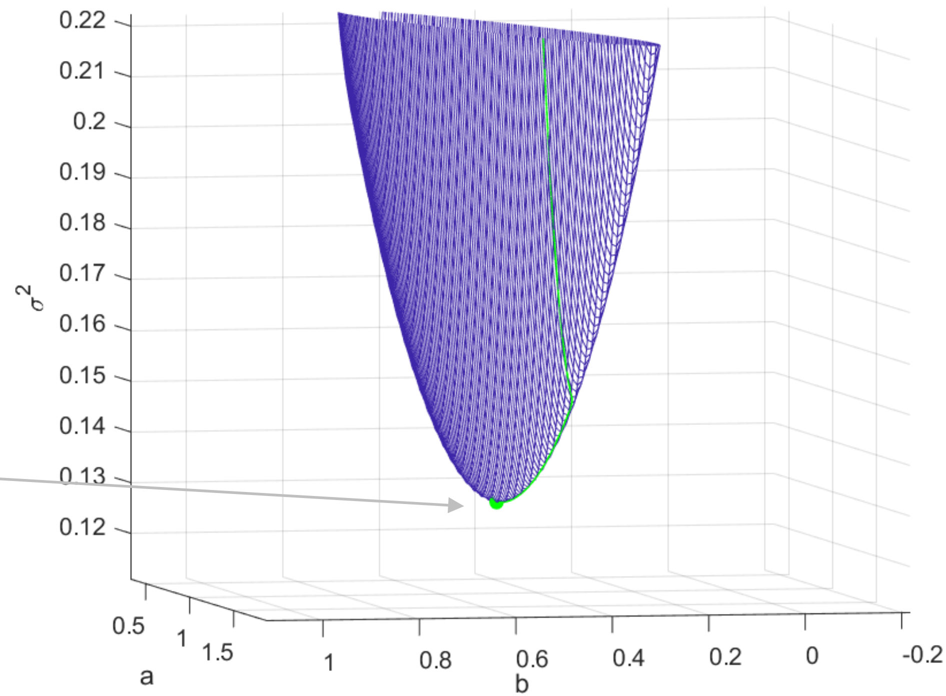
Method 3

that minimize $Q=\frac{1}{n}\sum_{i=1}^{n}(y_i-a-bx_i)^2$.

MLE is last value $\hat{\beta}^{(L)}$
or $(\hat{a}^{(L)},\hat{b}^{(L)})$.

Just set L=5×10[5].

$$\hat{\beta}=\begin{bmatrix}0.95\\0.47\end{bmatrix} \qquad \hat{\sigma}^2=0.1286$$



Maximum Likelihood Estimation - Exponential

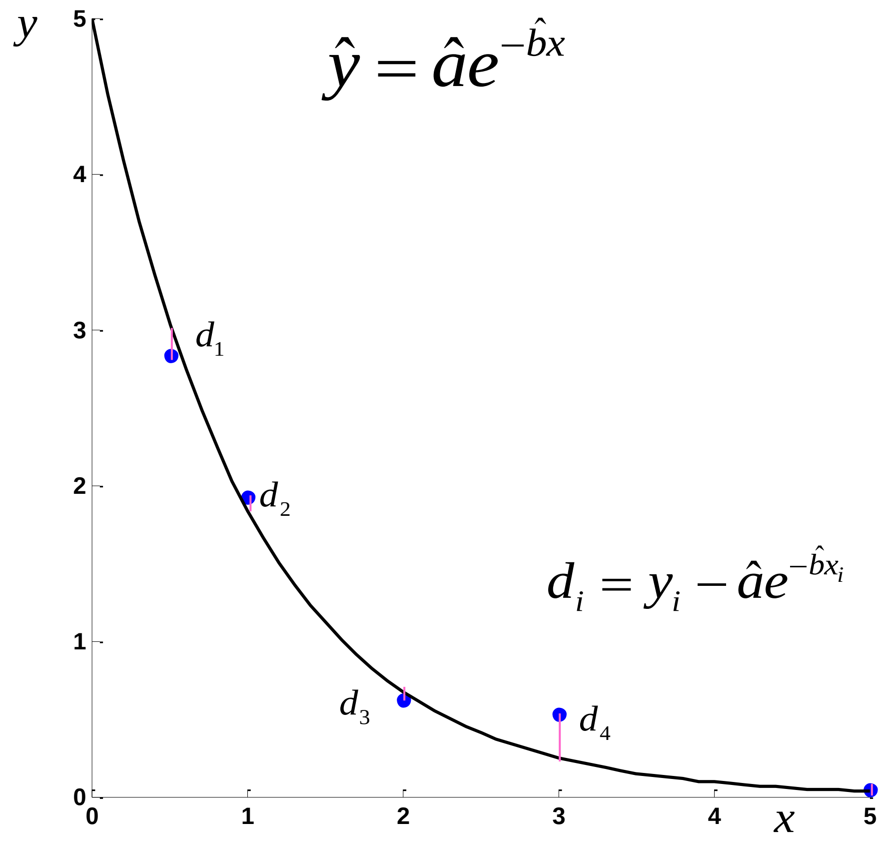
This is a more general method than just for linear functions

Let $y_i = ae^{-bx_i} + \varepsilon_i$,

where $\varepsilon_i \sim N(0,\sigma^2)$

are independent.

$i = 1,\dots,n$



Maximum Likelihood Estimation - Exponential

This is a more general method than just for linear functions

Let $y_i = ae^{-bx_i} + \varepsilon_i$, where $\varepsilon_i \sim N(0,\sigma^2)$ are independent.

Then, the likelihood is

$$f(y_1,\ldots,y_n \mid a,b,\sigma^2) = (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - ae^{-bx_i})^2\right]$$

and the log likelihood is

$$LL(a,b,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - ae^{-bx_i})^2 .$$

Maximum Likelihood Estimation - Exponential

$L(a,b,\sigma^2)$ is again called the likelihood function.

What we want to do is find the values of (a,b,σ^2)

that maximize $L(a,b,\sigma^2)$. The values (a,b) that maximize

$L(a,b,\sigma^2)$ are the values $(\hat{a},\hat{b})$ that minimize $\sum_{i=1}^{n}(y_i - ae^{-bx_i})^2$.

The value of σ^2 that maximizes $L(a,b,\sigma^2)$ is

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}e^{-\hat{b}x_i})^2 .$$

$d_i = y_i - \hat{a}e^{-\hat{b}x_i}$

minimize $\sum_{i=1}^{n} d_i^2$
wrt a, b

Maximum Likelihood Estimation - Exponential

Differentiate $LL(a,b,\sigma^2)$ wrt a, b, and σ^2, then set = 0

$$LL(a,b,\sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - ae^{-bx_i})^2$$

$$\left.\frac{\partial LL(a,b,\sigma^2)}{\partial a}\right|_{\hat{a},\hat{b},\hat{\sigma}^2} = -\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}2(y_i - \hat{a}e^{-\hat{b}x_i})(-e^{-\hat{b}x_i}) = 0$$

$$\left.\frac{\partial LL(a,b,\sigma^2)}{\partial b}\right|_{\hat{a},\hat{b},\hat{\sigma}^2} = -\frac{1}{2\hat{\sigma}^2}\sum_{i=1}^{n}2(y_i - \hat{a}e^{-\hat{b}x_i})(-\hat{a}x_i e^{-\hat{b}x_i}) = 0$$

$$\left.\frac{\partial LL(a,b,\sigma^2)}{\partial \sigma^2}\right|_{\hat{a},\hat{b},\hat{\sigma}^2} = -\frac{n}{2}\frac{1}{\hat{\sigma}^2} - \frac{-1}{2(\hat{\sigma}^2)^2}\sum_{i=1}^{n}(y_i - \hat{a}e^{-\hat{b}x_i})^2 = 0$$

Maximum Likelihood Estimation - Exponential

Solving for the estimated parameters yields

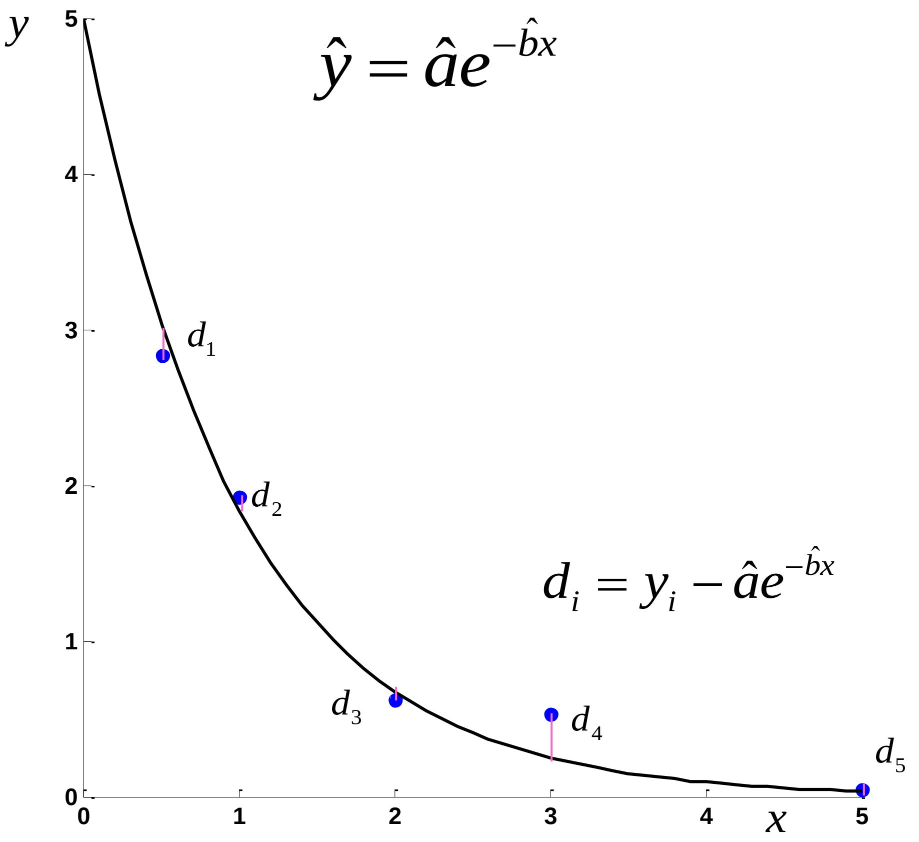
$$\hat{a} = \frac{\sum_{i=1}^{n} y_i e^{-\hat{b}x_i}}{\sum_{i=1}^{n} e^{-\hat{b}x_i}}$$

No analytic solution.

$$\hat{b} = \frac{\sum_{i=1}^{n} x_i y_i e^{-\hat{b}x_i}}{\sum_{i=1}^{n} x_i e^{-2\hat{b}x_i}}$$

Need numerical Solution.

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}e^{-\hat{b}x_i})^2$$



Maximum Likelihood Estimation - Exponential

Since we had to numerically maximize the likelihood,

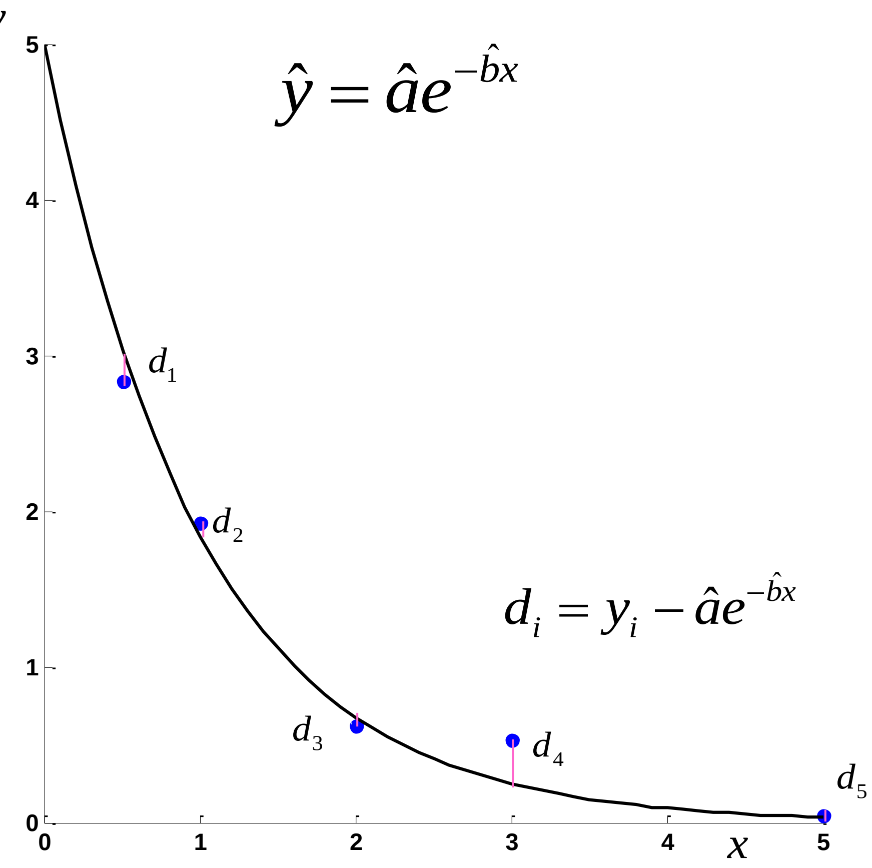
we do not have "nice" formulas

for the mean and variance of

$$(a,b,\sigma^2)$$

a and b that minimize $\sum_{i=1}^{n} d_i^2$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{a}e^{-\hat{b}x_i})^2$$



Homework 10:

1) Prove

a) $$(y-X\beta)'(y-X\beta)=(y-X\hat{\beta})'(y-X\hat{\beta})+(\beta-\hat{\beta})'(X'X)(\beta-\hat{\beta})$$

b) That the MLEs for a,b and σ^2 on slide 29 are the same as those on slide 23.

$$\hat{b}=\frac{n(\sum_{i=1}^{n}x_iy_i)-(\sum_{i=1}^{n}x_i)(\sum_{i=1}^{n}y_i)}{n(\sum_{i=1}^{n}x_i^2)-(\sum_{i=1}^{n}x_i)^2}$$

$$\hat{\beta}=\begin{pmatrix}\hat{a}\\ \hat{b}\end{pmatrix}$$

$$\hat{\beta}=(X'X)^{-1}X'y$$

$$\hat{a}=\bar{y}-\hat{b}\bar{x} \qquad \hat{\sigma}^2=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{a}-\hat{b}x_i)^2$$

$$\hat{\sigma}^2=\frac{1}{n}(y-X\hat{\beta})'(y-X\hat{\beta})$$

Homework 10:

2) Given observed data points (1,1), (3,2), (2,3), (4,4).

a) Plot the points.

Method 1

b) Analytically estimate the regression slope and y-intercept.

i.e. find $\hat{y} = \hat{a} + \hat{b}x$ by estimating $\hat{a}$ and $\hat{b}$.

Use $\hat{\beta} = (\hat{a}, \hat{b})' = (X'X)^{-1}X'y$ where $X = \begin{bmatrix} 1 & 1 \\ 1 & 3 \\ 1 & 2 \\ 1 & 4 \end{bmatrix}$ and $y = \begin{bmatrix} 1 \\ 2 \\ 3 \\ 4 \end{bmatrix}$.

Estimate the residual variance σ^2,

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(y_i - \hat{a} - \hat{b}x_i)^2$$

using the estimated $\hat{a}$ and $\hat{b}$.

Homework 10:

$$\hat{y} = \hat{a} + \hat{b}x$$

2) Given observed data points (1,1), (3,2), (2,3), (4,4).

Method 2

c) Numerically fit a regression line to the points.

Select a_{min}, a_{max}, b_{min}, and b_{max} values. Use $\Delta a = \Delta b = .1$.

can make smaller

Perform an exhaustive brute force grid search.

Compute $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - a - bx_i)^2$ for each (a,b) combination.

Find the a and b combination that make σ^2 the smallest.

The a and b that min σ^2 are $\hat{a}$ and $\hat{b}$, and the σ^2 is $\hat{\sigma}^2$.

Homework 10:

2) Given observed data points (1,1), (3,2), (2,3), (4,4).

$$S_y = \sum_{i=1}^n y_i \qquad S_x = \sum_{i=1}^n x_i \qquad S_{xx} = \sum_{i=1}^n x_i^2 \qquad S_{xy} = \sum_{i=1}^n x_i y_i$$

Method 3

d) Use Gradient Descent to iteratively find the $(\hat{a}, \hat{b})$ that

that minimize $Q = \frac{1}{n}\sum_{i=1}^n (y_i - a - bx_i)^2$.

$$\frac{dQ}{da} = \frac{d}{da}\frac{1}{n}\sum_{i=1}^n (y_i - a - bx_i)^2 = \frac{2}{n}\left[-\sum_{i=1}^n y_i + an + b\sum_{i=1}^n x_i\right]$$

$$\frac{dQ}{db} = \frac{d}{db}\frac{1}{n}\sum_{i=1}^n (y_i - a - bx_i)^2 = \frac{2}{n}\left[-\sum_{i=1}^n x_i y_i + a\sum_{i=1}^n x_i + b\sum_{i=1}^n x_i^2\right]$$

$$\beta^{(k+1)} = \beta^{(k)} - \gamma \nabla Q(\beta^{(k)})$$

$$\gamma = .0001$$

$$\nabla Q = \frac{2}{n}\begin{bmatrix} -S_y & n & S_x \\ -S_{xy} & S_x & S_{xx} \end{bmatrix}\beta \qquad \beta = \begin{bmatrix} 1 \\ a \\ b \end{bmatrix}$$

Homework 10:

2) Given observed data points (1,1), (3,2), (2,3), (4,4).

$$\hat{\beta} = (\hat{a}, \hat{b})' = (X'X)^{-1} X'y \qquad \hat{y} = \hat{a} + \hat{b}x$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (y_i - a - bx_i)^2$$

e) Plot the two (three) lines on the same graph as the points.
f) Plot the surface of (a,b,σ^2) values from c).
with the estimated points from b) and c) (and d)).
g) Comment.

Homework 10:

3) Given observed data points
(1/2, 3.21), (1,1.82), (2,.86), (3,.20), (4,.06), (5,.40).
a) Plot the points.
b) Numerically fit a regression single exponential to the points.
Find $\hat{y} = \hat{a}e^{-\hat{b}x}$.
Set up an interval of possible a and b values.
Select Δa and Δb values. Compute $\sigma^2 = n^{-1}\sum_{i=1}^{n}(y_i - ae^{-bx_i})^2$
for each combination. Find a and b that make σ^2 smallest.
The a and b that min σ^2 are $\hat{a}$ and $\hat{b}$ and the σ^2 is $\hat{\sigma}^2$.
c) Plot the curve $\hat{y} = \hat{a}e^{-\hat{b}x}$ on the same graph as the points.
d) Plot the surface of (a,b,σ^2) values from b).
e) Comment.

$\Delta a = \Delta b = .1$

Homework 10:

(1/2, 3.21), (1,1.82), (2,.86), (3,.20), (4,.06), (5,.40)

4) Given same observed data points as in 3).

a) Take the natural log of each y point, y'=log(y) .

b) Plot the points (y' and old x).

b) Guess where the “best” fit line to the data is.

c) Analytically fit a linear regression line to the points.

i.e. find $\hat{y}' = \hat{c} + \hat{d}x$, where $c = \log(a)$ and $d = -b$.

d) Plot the curve $\hat{y} = e^{\hat{c}} e^{\hat{d}x}$ on the same graph as the points and the previous fitted curve from 3).

e) Compute $\hat{\sigma}^2$ from y=exp(y') and $\hat{y} = e^{\hat{c}} e^{\hat{d}x}$.

f) Comment.

Homework 10:

5) Let $x_1,\ldots,x_n$ be an independent sample from each of the following PDFs. In each case find the MLE $\hat{\theta}$ of θ.

a) $f(x|\theta)=\dfrac{\theta^x e^{-\theta}}{x!}$, $x=0,1,2,\ldots.$, $0\le\theta<\infty$, $f(0|\theta=1)\equiv 1$.

b) $f(x|\theta)=\theta x^{\theta-1}$, $0<x<1$, $0<\theta<\infty$.

c) $f(x|\theta)=\dfrac{1}{\theta}e^{-x/\theta}$, $0<x<\infty$, $0<\theta<\infty$.

$f(x|\theta)$=0 where not defined

d) $f(x|\theta)=\dfrac{1}{2}e^{-|x-\theta|}$, $-\infty<x<\infty$, $-\infty<\theta<\infty$.

e) $f(x|\theta)=e^{-(x-\theta)}$, $\theta\le x<\infty$, $-\infty<\theta<\infty$.

Homework 10:

6) Generate a random sample $x_1, \ldots, x_n$ from each of the pdfs in 5). You choose appropriate θ value for each $f(x|\theta)$. n=25

Repeat samples so you have a total of 10^6 from each $f(x|\theta)$.

Calculate the MLE from each so that you have 10^6.
Calculate the mean, variance, and make a hist of MLEs.

*How do the MLEs of θ compare to the means, modes, and medians of $f(x|\theta)$.