What is a “good” estimator?

Criteria that measure the goodness of estimators are

1. Computational cost: low because of technologies.

2. **Least squares:** For any set of values of parameters characterizing a relationship, estimated values for dependent variable can be calculated using the values of the independent variables in the data set; these estimated values are called $\hat{y}$. These estimates of dependent variables can be subtracted from the actual values of dependent variable y to produce residuals or errors, (y - $\hat{y}$). There are various schools of thoughts on how to minimize and generate small residuals, the most popular school of thought to create “small” residual is the minimization of the sum of squared residuals/errors. The estimator that generates parameters with estimates that have smallest sum of squared error is ordinary least squares (OLS) estimators often referred as β^{OLS}. Greatest popularity of OLS estimator is because of 1) OLS estimators scores well on some (not all!) criteria listed below which is a valuable property & 2) Computational ease. OLS always minimize the sum of squared residuals but it does not always meet other criteria that are more important.

3. **Highest $\boldsymbol{R^2}$:** Coefficient of determination, R^2, explains proportion of variation in the dependent variable as explained by variation in the independent variables. In OLS, the sum of squared deviations of the dependent variable from mean (total variation in the dependent variable) can be broken into “explained” variation (the sum of squared deviations of the estimated values of the dependent variable from mean) and “unexplained” variation. (sum of squared residuals). R^2 = explained variations/total variations = (Model SE/TSE) = 1 - (unexplained variations/total variation) = 1- (SS Error/TSE). OLS, as minimize the sum of squared of error (SSE), it automatically maximizes the R^2 and no additional effort is needed. Also, R^2 can also be used to identify a good estimator when the function form and included independent variables are known. However, it can be misused.

4. **Unbiasedness:** An estimator $\hat{\beta}$ is said to be an unbiased estimator of β if the mean of

sampling distribution [expected value] of $\hat{\beta}$ is equal to β. The sample distribution is a probability density function of $\hat{\beta}s$ when $\hat{\beta}$ obtained from repeated sampling is plotted as histogram. [Fig: Kennedy, P. (2008). A Guide to Econometrics, 6th Edition, p: 14.] The property of unbiasedness does not mean that $\hat{\beta}$ = β; it only means that if we could take repeated sampling an infinite number of times, we would get the correct estimate "on an average". We hope that particular estimate from available sample will be close to the mean of the estimator sampling distribution; being centered over the parameter to be estimated is only one good property that the sampling distribution of an estimator can have. Another important property is variance of sampling distribution.

5. **Efficiency:** It is impossible to find an unbiased estimator. But whenever it is possible to find one, it is possible to find many unbiased estimators. In this case, sampling distribution with smallest variance is the most desirable unbiased estimators among several unbiased estimators. This is called *best unbiased estimator*. *The estimator with smallest variance is the efficient estimator among unbiased estimators*. Estimators with wider sampling distribution is less desirable as it has larger variance (standard deviation) and thus less certainty about if our estimates of $\hat{\beta}$ is close to real β though both estimators would give same estimates with same average β i.e E[β] in repeated sampling. [Fig: Kennedy, P. (2008). A Guide to Econometrics, 6th Edition, p: 16].

   Whenever the minimum variance (efficiency) criterion is mentioned, there must exist at least implicitly, some additional constraints, such as unbiasedness, accompanying the criterion. When the additional constraints accompanying the minimum variance, criterion is that the estimator under consideration is unbiased, the estimator is referred to as the best unbiased estimator. Unfortunately, it is, in many cases, impossible to determine which unbiased estimator (among all) has the smallest variance. Due to this problem, econometricians add further restriction that estimator to be a *linear* function ($\hat{\beta}_s$ adds or subtract but does not square or multiply) of the observations in the dependent variables. *An estimator that is linear, unbiased and has* *minimum variance among all unbiased estimator is called the best linear unbiased estimator (BLUE). BLUE is very popular among econometricians.*

6. **Mean square error (MSE):** Sometime, unbiased estimator (mean of distribution of estimator $\hat{\beta}$ centered around β) might have larger variance and only slightly biased estimator might have very small variance. Using best unbiased criteria, So, we might select unbiased estimator but with larger variance and become more uncertain about our prediction of value of β due to larger variance. So, better choice is selecting only slightly biased estimator with more certainty (viz. small variances). [Fig: Kennedy, P. (2008). A Guide to Econometrics, 6th Edition, p: 17] This trade-off between bias and variance is decided by using a criterion that minimize weighted average of the bias and variance. One way is to use absolute value of bias; more popular way is to use it square. When the estimator is chosen to minimize the weighted average of the variance and the square of the bias, the estimator is said to be chosen on the *weighted square error* criterion. When weights are equal, the criterion is the popular MSE criterion. Also, it happens that *the expected value of a loss function consisting of the square of the difference between β and its estimate (i.e. the squares of the estimation error) is the sum of the variance and the squared bias*. Minimization of the expected value of this loss function makes good intuitive sense as a criterion for choosing an estimator. In practice MSE is not usually used unless the best criterion is unable to produce estimates with small variances. One such example is multicollinearity.

7. **Asymptotic Properties:** The properties of estimators' unbiasedness, efficiency and MSE are related to the nature of sampling distribution. An unbiased estimator has its value centered around the true value of parameter estimated. This is true for both small and large sample. However sometime, it is impossible to find estimators having these desirable sample distribution properties in small sample; we have to use asymptotic properties—the nature of the estimator's sampling distribution in extremely large samples. The sampling distribution of estimator changes with sample size; the sample mean statistics (estimates of estimators?) have sampling distribution centered around the population mean but variance decrease with increase in sample

size and biased estimator become less and less biased as sample size increases. *As the sample size become larger its sampling distribution changes, such that the mean of its sampling distribution shifts closer to the true value of the population parameter being estimated. This is "the concept of an asymptotic distribution" and defining desirable asymptotic or "large sample properties" of an estimator.*

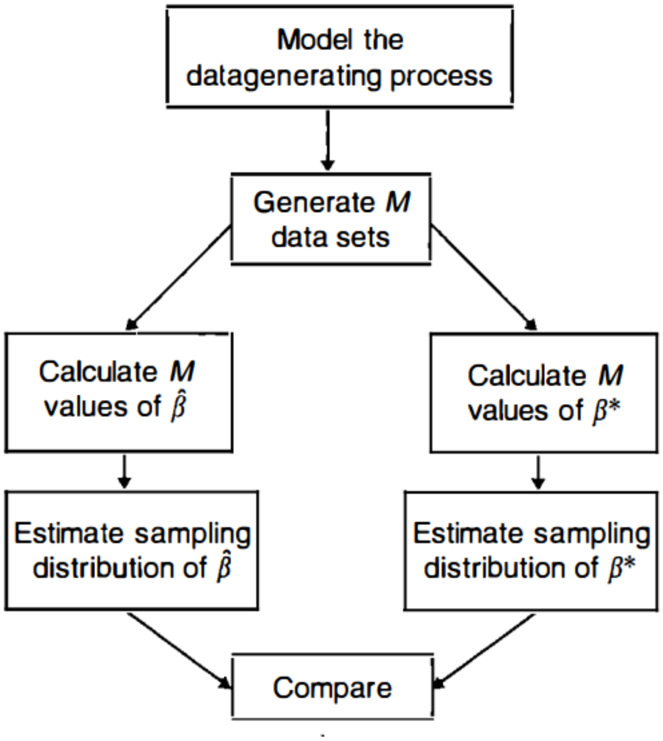
Consider a sequence of sampling distributions of an estimator $\hat{\beta}$ [Consider taking sample size of 5, 40,70, 100, 200 and repeating each sampling 2000 times from same population and estimate $\hat{\beta}$.], formed by calculating the sampling distribution of $\hat{\beta}$ for successively large sample sizes. [Fig: Kennedy, P. (2008). A Guide to Econometrics, 6th Edition, p: 20] If the distribution in this sequence follows a (pre-known) specific distribution (such as normal distribution) as the sample size become extremely large, this specific distribution is called the *asymptotic distribution of $\hat{\beta}$*. Two basic estimators are defined in terms of asymptotic distribution: **1)** if the asymptotic distribution of $\hat{\beta}$ (viz. $\hat{\beta}$ following a specific distribution) becomes concentrated on a particular value *k* as the sample size approaches infinity, *k* is said to be probability limit of $\hat{\beta}$ (written as plim $\hat{\beta}$ = *k*). If plim $\widehat{\beta}$ = β, then $\hat{\beta}$ is said to be *consistent*. **2)** The variance of the asymptotic distribution of $\hat{\beta}$ is called *asymptotic variance* of $\hat{\beta}$ if $\hat{\beta}$ is consistent and if its asymptotic variance is smaller than the asymptotic variance of all other consistent estimator, $\hat{\beta}$ is said to be *asymptotically efficient variance*. plim $\widehat{\beta}$ = β is large sample equivalent of unbiasedness.

8. **Maximum Likelihood:** The maximum likelihood estimate (MLE) of a vector of parameter values β is simply the particular vector β^{MLE} that gives the greatest probability of obtaining the observed data. MLE has several desirable asymptotic properties. It is asymptotically unbiased, consistent, asymptotically efficient and distributed asymptotically normally and its asymptotic variance can be found via a standard formula. However, we must assume normal distribution of error terms.

9. **Robustness:** (Insensitivity to violations to the assumptions under which the estimator has desirable properties as measured by the criteria above).

10. Monte Carlo Studies:

The general idea behind a Monte Carlo study is to (1) model the data-generating process, (2) generate several sets of artificial data, (3) employ these data and an estimator to create several estimates, and (4) use these estimates to gauge the sampling distribution properties of that estimator for the particular data-generating process under study.


Model the datagenerating process
Generate M data sets
Calculate M values of $\hat{\beta}$
Calculate M values of β^*
Estimate sampling distribution of $\hat{\beta}$
Estimate sampling distribution of β^*
Compare

Structure of a Monte Carlo study.

****

Summary: Bias Vs. Consistency from https://eranraviv.com/

Consistency: Convergence in Probability.

$\lim_{n\to\infty} pr[(|\hat{\beta}-\beta|)<\delta]=1$; usually δ = ϵ (error term).

i.e. As the sample increases to become sufficiently large, *the probability of* **the difference between the parameter estimated and the real population parameter** becoming larger i.e. **larger than error term (ϵ)** is zero. OR the probability that the error becoming bigger than the difference is one. Then is it consistent.

As the sample size increases, the probability of an unusual outcome become less, less than error term.

Unbiased Estimator: An estimator ($\hat{\beta}$) is said to be an unbiased estimator of β if the mean of sampling distribution [expected value] of ($\hat{\beta}$) is equal to β. The sample distribution is a probability density function of $\hat{\beta}s$ when $\hat{\beta}$ obtained from repeated sampling is plotted as histogram. See p. 14,

A Biased Estimator:

A biased estimator means that the estimate we see comes from a distribution which is not centered around the real parameter. In other words, the LS estimator we obtained E(β-hatt) from the sample is not accurate estimator of parameter β if E(β-hatt) is biased.

For an estimator to be unbiased E($\hat{\beta}$) = β also written as μ. where, E($\hat{\beta}$) is expectation (mean) of parameter estimate (obtained from sample) of population parameter (mean) written as β or μ.

Note: Unbiases talks about the distribution and E($\hat{\beta}$) but consistence talks about $\hat{\beta}$.

Note: Theoretically, unbiased does not meant necessarily consistency and consistency does not necessarily meant unbiases. However, estimator which are unbiased tends to be consistent. If data is unbiased but not consistent (not a general case), increase sample size.

To obtain (XX\`)$^{-1}$(Xy), Find ee\` as matrix multiplication as ee\` = (y – Xb)( (y – Xb)\`. Multiply, differentiate with respect to b to obtain least squared estimate (ee\`), equate to 0 and solve for b.

$\hat{\beta}$ = (XX`)$^{-1}$(Xy) where, X is matrix of independent variables, X` is transformation of matrix X and y is dependent variable or predictor.

E($\hat{\beta}$) = E[(XX`)$^{-1}$(Xy)]

y = βx + ϵ

Replace y in above equation and solve for β assuming E(ϵ) = 0. You will get (XX`)$^{-1}$(X`X) which is Identity matrix (I) and equal to 1.

Assumptions of OLS:

Existence Assumptions:

(i) X and y are observed.

(ii) Rank(X) = k ⊂ T (dependent Variables are not multiplicative in nature with themselves or other for all observations.)

(iii) K < T (Less number of explanatory variables (k) than number of observations (T).)

Note: Violation of these three assumptions above means either OLS estimator does not exist or are not unique.

Classic Linear Regression Model Assumption:

1. **$\epsilon \sim N(0, \sigma^2)$** in y = Xβ + ϵ & Error terms are distributed normally with the mean/expected value = 0 and variance = σ^2 in a linear regression function.

2. E [ϵ] = 0. The expectation of error term is zero.

3. Homoscedasticity (No Heteroscedasticity) i.e.: E[ee`|X] = σ^2I. Constant variance of error term $\sigma_i^2 = \sigma_j^2$ for all i, j ie. across observations. If errors have constant variance, the error are called homoscedastic.

4. No Autocorrelation: E[e|X] = 0 ➔ Cov(e, X) = 0 & Cov(e_t, e_s) = 0.

5. Non-stochastic matrix X of dependent variables. There are three possibilities a) No problem, b) lagged endogenous and c) simultaneity. We want “No probelm”.

Note: Only 2 and 5 are necessary to prove consistency and unbiasedness.

Small Sample Properties: Unbiased, Efficient, and BLUE.
Large Sample Properties: Consistent and Asymptotic Efficiency.

1 => 4: Unbiased ➔ Consistent: No.
4 => 1: Consistent ➔ Unbiased: No.
1 => 2: Unbiased ➔ BLU: No.
2 => 1: BLU ➔ Unbiased: Yes.
3 => 5: Efficient ➔ Asymptotically efficient: Yes.
5 => 3: Asymptotically efficient ➔ Efficient: No.
5 => 4: Asymptotically efficient ➔ Consistent: Yes.
3 => 1: Efficient ➔ Unbiased: No.
3 => 2: Efficient ➔ BLU: No.
4 => 2: Consistent ➔ BLU: No.
5 => 4: Asymptotically efficient ➔ Unbiased: No.

Corrected answers after grading:

1 => 4: Unbiased (Implies) ➔ Consistent
Ans: No, may not follow the distribution of parameter and thus inconsistent.

4 => 1: Consistent ➔ Unbiased.
Ans: No. Can be consistent (follow the distribution) and still be biased.

1 => 2: Unbiased (Implies) ➔ Best Linear Unbiased
Ans: NO, may not be the best as unbiased estimators might have higher variance than another estimator.

2 => 1: BLU ➔ Efficient: Best
Ans: Yes, smallest variance (Best) and unbiased implies efficient.

3 => 5: Efficient ➔ Asymptotically Efficient
Ans: Yes. Efficient means Asymptotically efficient and vice versa. The proof is messy.

5 => 3: Asymptotically efficient ➔ Efficient.
Ans: No. Asymptotically efficient needs consistent (estimator follows the distribution), and smaller asymptotic (variance of estimator follows the distribution) variance but could be biased.

5 => 4: Asymptotically efficient ➔ Consistent
Ans: Yes. Consistency (follow its distribution) is an assumption of Asymptotically efficient estimator. Another assumption is smaller asymptotic variance (smallest variance following its distribution). So, it is consistent.

3 => 1: Efficient ➔ Unbiased.
Ans: Yes. Unbiases is one of the properties of efficiency. Another property of efficient is minimum variance.

3 => 2: Efficient ➔ BLU
Ans: No. To be efficient, estimator should have minimum variance (best), unbiased linear estimator.

4 => 2: Consistent ➔ BLU
Ans: No. Can be consistent (follow the distribution) and may not have minimum variance (best).

****

Finding Variance and Standard Error and Confidence Interval from a Function:

To find standard error, find asymptotic Variance and square root:

To find asymptotic Variance of MLE:

Asymptotic variance of MLE = $1/I(\theta_o)^{-1}$ = 1/Information matrix.

Information matrix = -E(second derivative of given function).

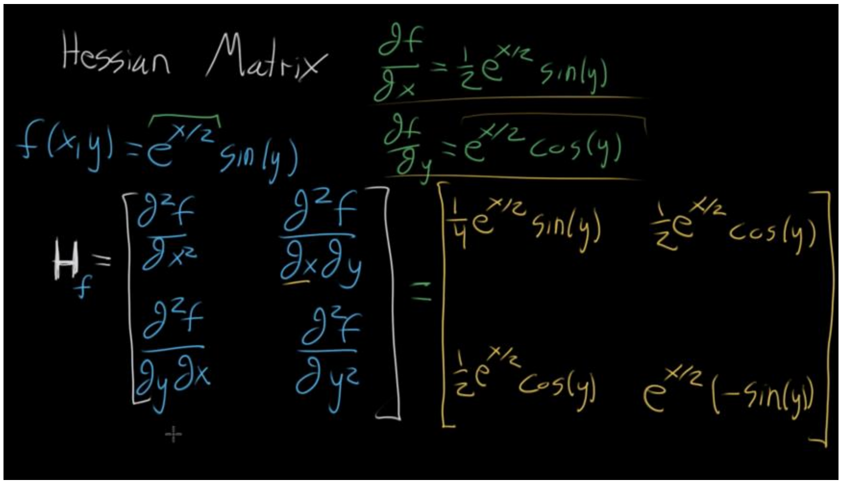
Information matrix can be computed in three different ways:

- Using Hessian matrix (Video from Khan Academy).
- Using second derivative.
- Estimated information matrix used with method of scoring.

Note: In homework, we were asked to find variance of MLE of k. So, we took likelihood function.

(check homework 3 for more details.).

Using Hessian matrix: (Two variable case):



Hessian Matrix

$f(x,y) = e^{x/2} \sin(y)$

$\frac{\partial f}{\partial x} = \frac{1}{2} e^{x/2} \sin(y)$

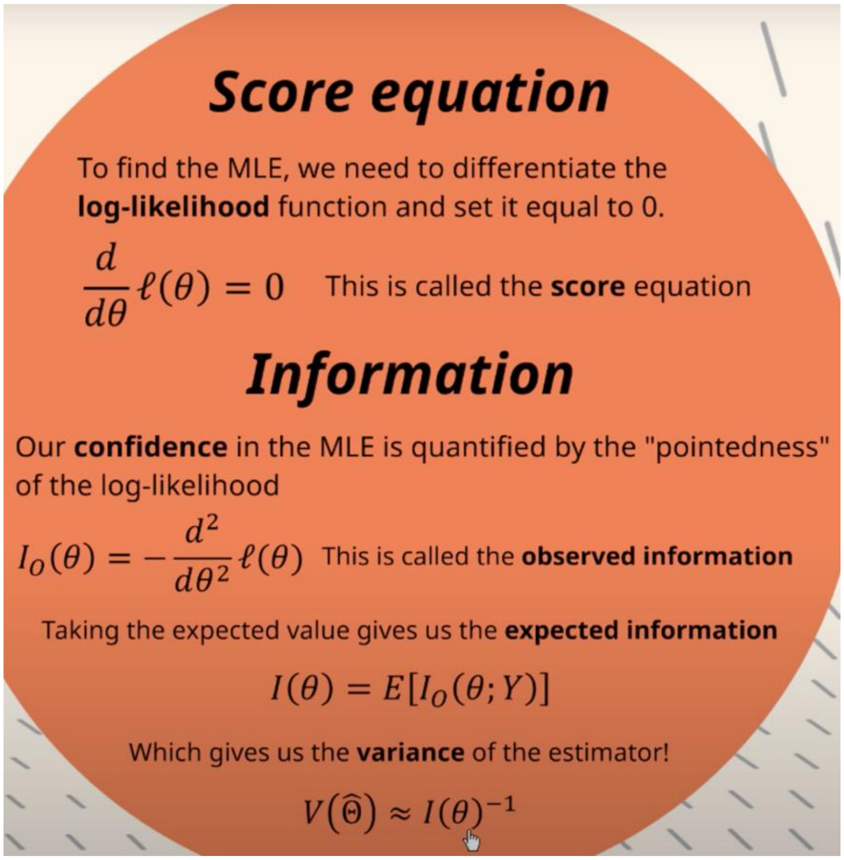
$\frac{\partial f}{\partial y} = e^{x/2} \cos(y)$

$$H_f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} \frac{1}{4} e^{x/2} \sin(y) & \frac{1}{2} e^{x/2} \cos(y) \\ \frac{1}{2} e^{x/2} \cos(y) & e^{x/2}(-\sin(y)) \end{bmatrix}$$

Source: Khan academy YouTube video: https://youtu.be/LbBcuZukCAw

This video has demonstrated how to expand this method for 3 variables as well.

Using Estimated Information Matrix using Scoring Method:



Score equation

To find the MLE, we need to differentiate the **log-likelihood** function and set it equal to 0.

$\frac{d}{d\theta} \ell(\theta) = 0$ This is called the **score** equation

Information

Our **confidence** in the MLE is quantified by the "pointedness" of the log-likelihood

$I_O(\theta) = -\frac{d^2}{d\theta^2} \ell(\theta)$ This is called the **observed information**

Taking the expected value gives us the **expected information**

$$I(\theta) = E[I_O(\theta; Y)]$$

Which gives us the **variance** of the estimator!

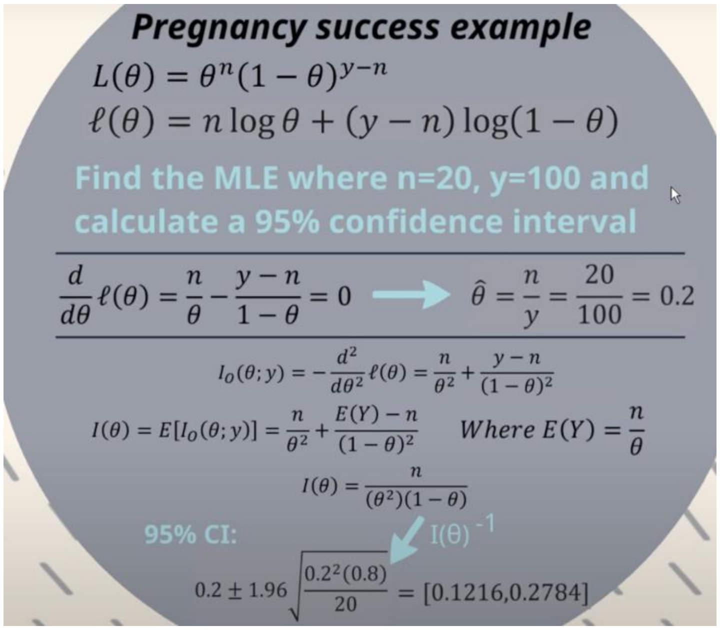
$$V(\widehat{\Theta}) \approx I(\theta)^{-1}$$

Example:

Notes: L(θ) or l(θ) is log likelihood function.

Do first derivative of l(θ) wrt θ and equate to zero, solve for θhatt gives MLE. MLE is regression coefficient.

Differentiate first derivative of l(θ) wrt θ again and take expectation and solve to get information matrix, I(θ). The inverse of information matrix is Variance. Square root of variance is standard error and CI is given by “MLE ± 1.96√Variance”; Variance = 1/ I(θ).



Pregnancy success example

$$L(\theta) = \theta^{n}(1-\theta)^{y-n}$$

$$\ell(\theta) = n\log\theta + (y-n)\log(1-\theta)$$

Find the MLE where n=20, y=100 and calculate a 95% confidence interval

$$\frac{d}{d\theta}\ell(\theta) = \frac{n}{\theta} - \frac{y-n}{1-\theta} = 0 \longrightarrow \hat{\theta} = \frac{n}{y} = \frac{20}{100} = 0.2$$

$$I_O(\theta; y) = -\frac{d^2}{d\theta^2}\ell(\theta) = \frac{n}{\theta^2} + \frac{y-n}{(1-\theta)^2}$$

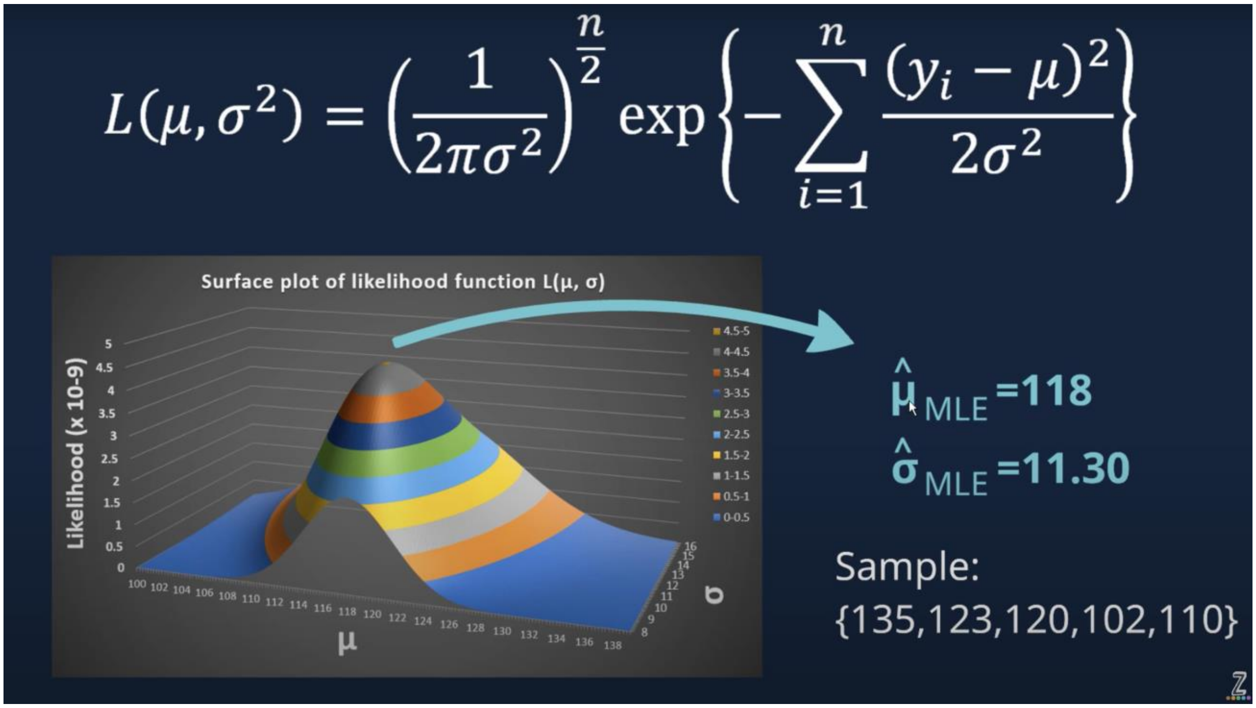
$$I(\theta) = E[I_O(\theta; y)] = \frac{n}{\theta^2} + \frac{E(Y)-n}{(1-\theta)^2} \qquad Where\ E(Y) = \frac{n}{\theta}$$

$$I(\theta) = \frac{n}{(\theta^2)(1-\theta)}$$

95% CI: $I(\theta)^{-1}$

$$0.2 \pm 1.96\sqrt{\frac{0.2^2(0.8)}{20}} = [0.1216, 0.2784]$$

Source: MLE, Score Equation, Information, Invariance: https://youtu.be/7kLHJ-F33GI



$$L(\mu,\sigma^2) = \left(\frac{1}{2\pi\sigma^2}\right)^{\frac{n}{2}} \exp\left\{-\sum_{i=1}^{n}\frac{(y_i-\mu)^2}{2\sigma^2}\right\}$$

$\hat{\mu}_{MLE} = 118$

$\hat{\sigma}_{MLE} = 11.30$

Sample:
{135,123,120,102,110}

For Normal distribution: μ is the line along x-axis and σ is the line along the z-axis. Y-axis is vertical line.

This is also presented in example below mathematically.

****

Summary of Estimation methods:

a) Minor methods

1) Minimize total effort or cost in choosing estimates.

Assume β = b and σ^2 = 0.

2) Minimize $\sum_j \left|\hat{e}_j\right|$ - Mean Absolute Deviation (MAD) estimator

3) Method of moments: estimate moments or distribution.

b) Least Square Method (OLS)

Minimize $\sum_{t=1}^{T} \hat{e}_t^2$

$$\min S(b) = (Y - Xb)^{/}(Y - Xb) = Y^{/}Y - 2\, b'X'Y + b'X'X\, b$$

$$\frac{\partial\, S}{\partial\, b} = -2X'Y + 2X'Xb = 0$$

$$b = (X'X)^{-1}\, X'Y$$

c) Maximum Likelihood Method (MLE)

Advantage: MLE uses distribution of y and is the most efficient estimator possible.

MLE is asymptotically unbiased and consistent in large sample.

MLE Selects parameters β and σ^2 that maximize the probability of getting the data point (X, y) i.e. Maximize likelihood function $\ell(\beta, \sigma^2 | y, X)$.

Disadvantage: Less convenient in certain circumstances and requires knowing distribution (especially in small sample).

How to find a maximum likelihood function and its variance?

This was done in homework 3 with example: $f(x) = ke^{-kx}$

Steps to find MLE: Given function (see above) ➔ Take log ➔ differentiate wrt k ➔ equate to zero and solve for k ➔ K is MLE.

Find Variance: Asymptotic variance of MLE = $1/I(\theta_o)^{-1}$

To find standard error, find asymptotic Variance and square root: To find asymptotic Variance of MLE:

Asymptotic variance of MLE = $1/I(\theta_o)^{-1}$ = 1/Information matrix.

Information matrix = -E (second derivative of given function).

Information matrix can be computed in three different ways:

a) Using Hessian matrix (Video from Khan Academy).

b) Using second derivative.

c) Estimated information matrix used with method of scoring.

Note: In homework, we were asked to find variance of MLE of k. So, we took likelihood function. (check homework 3 for more details.).

d) Generalized Method of Moments (GMM)

OLS is GMM estimator.

No distribution assumptions are made for GMM.

Estimation of θ is based on explanatory variables Xs' such that it satisfies the condition that sample moments are zero.

$$m(\theta_{GMM}) = \frac{1}{n}\sum_i x_i \, e(x_i, \theta_{GMM}) = 0 \text{ (Sample moments = 0)}$$

$$m(\theta_{GMM}) = \frac{1}{n} \, X' \, e(X, \theta_{GMM}) = 0 \text{ (Sample moments = 0)}$$

GMM estimators are available in Proc Model of SAS.

GMM estimators are weighted minimum distance estimators.

When dealing with a *simultaneous equations* system, a set of $J \geq K$ instrumental variables is likely to be used.

Then, sample moments are: $m(\theta) = \frac{1}{n}\sum_i z_i \, e(x_i, \theta) = \frac{1}{n} \, Z' e(X, \theta)$

****