Not test relevant

Estimation Principles:

- Besides ***ordinary least squares*** estimation principle, there are several alternative estimation principles such as the ***method of moments***, ***maximum likelihood*** or ***Bayesian*** used in statistics

  Each of these estimation methods employs some form of optimization techniques to estimate the unknown parameters from the observed (or hypothetical) sample data.

- Historically, the ***maximum likelihood*** and ***OLS*** estimators were (and still are) the most prominent estimation methods.

  stronger assumption

  However, other principles are gaining momentum under specific conditions.

- The maximum likelihood estimator explicitly requires

  1. that the functional form $f(x_i|\theta)$of the ***underlying distribution*** of each random variable is explicitly ***known*** and that the sample data match this distribution (the parameters of the underlying distribution do not need to be fully specified),

2. that the observations are ***statistically independent***, i.e., $f(x_1, \dots, x_n|\theta) = f(x_1|\theta) \cdots f(x_n|\theta)$ (or can be ***transformed*** to be independence as we will see when we discuss *General Least Squares*). allow transfer auto-correlated observation to independent

- Recall: ***OLS is more relaxed***: it satisfies the Gauss-Markov theorem (unbiasedness and efficiency if some assumptions are satisfied) without relying on any distributional assumptions.

The Maximum Likelihood Principle (see FOX Appendix)

- Change of perspective from joint-distribution to likelihood function:
  - Under the classical statistical perspective the underlying distributional parameters of the populating are fixed and the observations are random outcomes of the sampling process: $f(x_1, \dots, x_n \mid \theta)$
  - For the ***likelihood perspective*** the data $\{x_1, \dots, x_n\}$ of a particular sample are considered fixed and the unknown population parameter θ becomes a random variable.

    This reversal of perspective gives the likelihood function: $L(\theta \mid x_1, \dots, x_n)$

- Maximum likelihood asks the question:

  Given the observed data $\{x_1, \ldots, x_n\}$***, which is the <u>most likely</u> population parameter*** θ ***that has generated the observations?***

- This becomes an optimization task of finding the unknown parameter θ which maximizes the likelihood function of the observed data: $\max_{\theta} L(\theta \mid x_1, \ldots, x_n)$.

- This optimization is often carried out by transforming the likelihood function $L()$ into its logarithmic form $l(\theta \mid x_1, \ldots, x_n) = \log L(\theta \mid x_1, \ldots, x_n)$ because:

  - The logarithm is monotonically increasing transformation and therefore preserves the order of the observations.

    Consequently, if the likelihood function is at its maximum then ***also*** the log-likelihood function will be at its maximum.

  - The reason for this transformation is that the ***product*** of the probabilities under the independence assumptions becomes now a ***summation***:

$$\Rightarrow \boxed{\begin{aligned} L(\theta \mid x_1,\ldots,x_n) &= f(\theta \mid x_1)\cdot\cdots\cdot f(\theta \mid x_n) \\ l(\theta \mid x_1,\ldots,x_n) &= \log f(\theta \mid x_1)+\cdots+\log f(\theta \mid x_n) \end{aligned}}$$

using log transformation to transfer product to sum

Derivatives of summations are easier to evaluate because one can skip the product rule.

- Properties of the maximum likelihood estimator:

  1. ML estimators are ***asymptotically unbiased***
  2. ML estimators are asymptotically ***normally distributed*** and, for testing purposes, the ***covariance matrix*** of the estimated parameters can be calculated, e.g., their standard errors their standard errors are available.

     FYI: The covariance matrix is based on the expected value of inverse of what is called the information matrix (i.e., second derivative of the log-likelihood function)
  3. ***Comparable*** test statistics to the *t*-test, *F*-test and partial *F*-test can be calculated.

     unbiased
  4. The ML estimator is asymptotically ***consistent*** (unbiased with smaller or equal variance than any alternative estimator for large sample sizes).

- **<u>Example of the maximum likelihood principal:</u>**
  - Flipping a coin *n=10* times with the given outcome HHTHHHTTHH. Each throw is conducted independently of the other throws.

    Our intuitive guess would be $\Pr(H) = \frac{\text{\# of Heads}}{\text{\# of Trails}} = \frac{7}{10}$.
  - The probability of obtaining this sequence of observations, prior to collecting the data, is a function of the ***unknown population parameter*** π:

    $$\begin{aligned}\Pr(\text{data}|\text{parameter}) &= \Pr(HHTHHHTTHH \mid \pi) \\ &= \pi \cdot \pi \cdot (1-\pi) \cdot \pi \cdot \pi \cdot \pi (1-\pi) \cdot (1-\pi) \cdot \pi \cdot \pi \\ &= \pi^7 \cdot (1-\pi)^3 \quad \text{Probability}\end{aligned}$$

    That is, each throw is ***binary distributed*** and its outcome is independent of the other throws. Therefore, the ***product*** of individual probabilities gives the ***join-probability***.
  - Assuming the data are given and the unknown random parameter is π then the conditional probability $\Pr(data \mid \pi)$ changes into a likelihood function:

$$L(\text{unknown parameter} \mid \text{observed data}) = \Pr(\pi \mid HHTHHHTTHH)$$
$$= \pi^7 \cdot (1-\pi)^3$$

likelihood function
(reverse probability function)

Note: A likelihood function is not a proper statistical distribution function for the unknown parameter π, because it does not integrate to ***one*** over the range $\pi \in [0,1]$.

- In our imagination we can vary the parameter π within its feasible range $\pi \in [0,1]$ and find the ***maximum value*** of the likelihood function for the given sample observations at a particular value $\hat{\pi}$. See the script **`likeBinom.rmd`**.
- The probability that our given data come from a population with π is highest at $\hat{\pi} = 7/10$. This is our best guess for the unknown underlying population parameter π.
- See table in Fox's Appendix and Figure D.15

π	$L(\pi \mid \text{data}) = \pi^7(1-\pi)^3$
0.0	0.0
.1	.0000000729
.2	.00000655
.3	.0000750
.4	.000354
.5	.000977
.6	.00179
.7	.00222
.8	.00168
.9	.000478
1.0	0.0

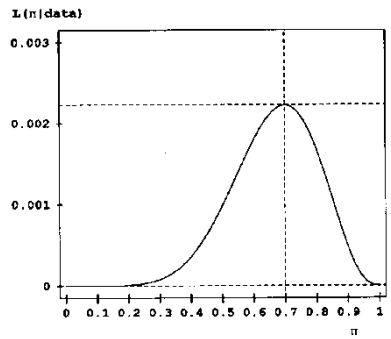

Figure D.15. The likelihood function $L(\pi|HHTHHHTTHH) = \pi^7(1-\pi)^3$.

- Discussion: If the true probabilities were $\pi = 0.0$ or $\pi = 1.0$ (the sure events) then the observed sample would have been impossible to observe.

- **Technical estimation of the unknown parameter π and its variance:**
  - For an explicit development to the estimated probability and the associated standard error see FoxRegressionMathAppendix.pdf on pages 92-98.
  - Setting the first derivative with respect to the unknown parameter π to zero leads to the rate estimator $E(\pi) = \hat{\pi} = x/n$.
  - Taking the inverse of the information matrix gives the variance:

So variance is dependent on sample(population) size,
Larger size, more accuracy.

$$Var(\pi) = \frac{\hat{\pi} \cdot (1 - \hat{\pi})}{\boxed{n}}$$

- Note that the variance of the rate estimator shrinks with an increasing denominator n, i.e., the number if trails.

- Implication for geographic data: Areas within a region vary in size, i.e., their population at risk n).

  Consequently, the ***rate estimates*** for each area have inconstant variances that depends on the areas' population at risk.

Likelihood Ratio Test, Wald Test, and Score Test

Likelihood Ratio Test (LR):

- The ***likelihood ratio test*** requires that the model is estimated (calibrated) twice:
  - first under the ***null hypothesis*** and
  - then under the ***alternative hypothesis***.

  Under the null hypothesis means that the parameters, which are tested, are ***set a priori to a given value***, which in most cases is ***zero*** (that is, not included in the estimation procedure).
- The restricted model is ***nested*** into the unrestricted model
- The test statistic becomes $\chi_H^2 = -2\left(\ln(L_{K-H}) - \ln(L_K)\right)$ where just like partial F-test
  - $\ln(L_K)$ is the log-likelihood of the ***unrestricted*** model with K parameters and
  - $\ln(L_{K-H})$ is the log-likelihood of the ***restricted*** model.
- The number of restrictions (parameters not estimated) is H.
  The relationship $\ln(L_{K-H}) \le \ln(L_K)$ holds because the restricted likelihood function is associated with a model that fit the data ***not as well*** as the full model.

- The χ^2_H test statistic is asymptotically χ^2-distributed with H degrees for freedom under the null hypothesis that the H restricted parameters are irrelevant.
  Large values of χ^2_H indicate that the ***a priori*** restricted parameters are in fact ***relevant*** for the full model (that is different from their assumed values under the null hypothesis) and, therefore, should also be estimated.
- The underlying idea of the likelihood ratio test is similar to that of the ***partial* F*-test***.

The Wald test:

- Under the ***Wald test*** the full model with all parameters is estimated and the included parameters are tested against their values under the null hypothesis.
- It requires that we estimate and evaluate the parameter's standard errors (derived from the expected value of the inverse information matrix) at the unrestricted maximum likelihood value.
- The ***Wald*** test is a substitute for the ***single* t*-test*** for the parameters in the model. It helps to decide which parameters to drop from the model.

The Score test (a.k.a. Lagrange multiplier test):

- The ***Score test*** estimates the restricted model where the parameters under scrutiny are set *a priori* to a given value (in most cases zero).

- Subsequently the test evaluates how ***binding*** the ***a priori constraints*** are:
  It requires that we evaluate the magnitude of the Lagrange multiplier relative to the gradient of the likelihood function at the constraint.
  A small value indicates that we are close to the unconstrained maximum of the log-likelihood function and therefore the constraint is not relevant $\Rightarrow$ we can ignore the parameter.
- The score test helps to decide, which parameters should be added to a restricted model and estimated freely.

General Discussion

- Critical values of the χ^2 distribution are used with the degrees of freedom equal to
  - the number of restricted parameters for the likelihood ratio test, and
  - one for the Wald-test and Score-test if just individual parameters are tested.
- These tests are only valid in large samples of statistically independent observations.
- It can be shown that $W \geq LR \geq S$ when they are applied under identical circumstances.
  Thus the Wald test will reject the null hypothesis most frequently than the likelihood ratio test and finally the scoretest..

- Visualization of the three test principles.

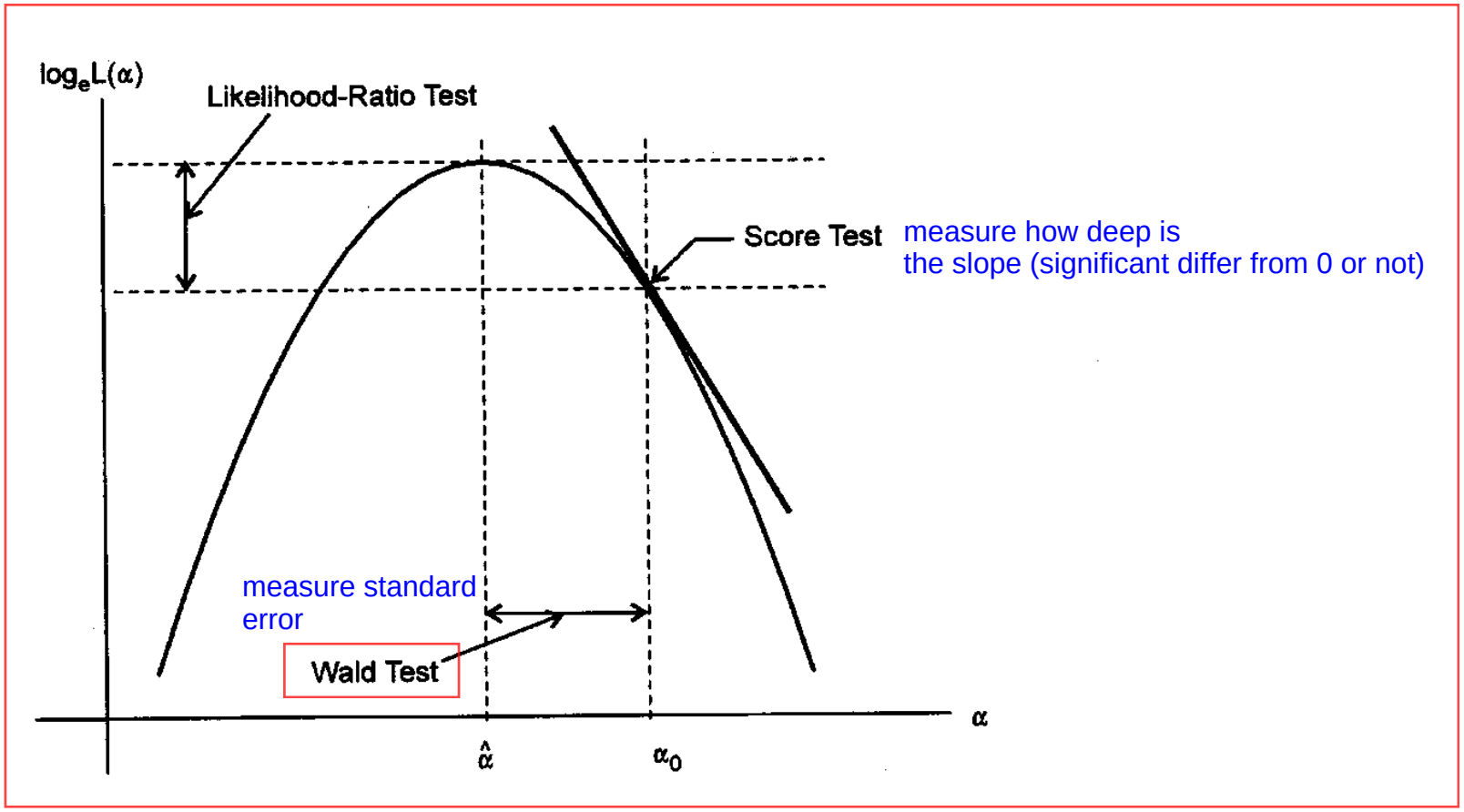

Figure D.16. Tests of the hypothesis H_0: $\alpha = \alpha_0$: The likelihood-ratio test compares $\log_e L(\hat{\alpha})$ with $\log_e L(\alpha_0)$; the Wald test compares $\hat{\alpha}$ with α_0; and the score test examines the slope of $\log_e L(\alpha)$ at $\alpha = \alpha_0$.

Goodness of Fit Statistics Associated with Likelihoods

Deviance: test the overfitting, saturate model

- The key building block of the deviance is a likelihood ratio comparison of the log-likelihood of ***current model*** against the log-likelihood of a model which ***fits the observed data perfectly***.
- A perfectly fitting model is also called a ***saturated model*** because it uses as many estimated parameters (one for each observation) as there are observations.
- A simple approach to obtain a saturated model is to substitute the observed value as predicted value into the likelihood function, i.e., $\hat{y}_i \leftarrow y_i$.
  For example: In case of a binary distributed values of dependent variable are $y_i \epsilon \{0,1\}$ with $\hat{\pi}_i = y_i$ the likelihood function becomes

$$L(saturated\ model) = \prod_{i=1}^{n} {y_i}^{y_i} \cdot (1 - y_i)^{(1-y_i)} = 1$$

  Thus $log\left(\underbrace{L(saturated\ model)}_{=1}\right) = \log(1) = 0.$

- The deviance compares the fitted model against a saturated model is simply a likelihood ratio statistic:

$$D(fitted\ model) = -2 \cdot \log\left(\frac{L(fitted\ model)}{L(saturated\ model)}\right) = -2 \cdot \log\big(L(fitted\ model)\big)$$

- Therefore, models with a small deviance fits the observed data well, whereas, a model with a large deviance fits the data poorly.

  so we like the deviance as small as possible

  Analog to linear regression one can think of the deviance as the ***residual-sum-of-squares***.

- Similar to the partial F-test in linear regression the deviance can be used to compare nested models:

  $$G = D(model\ with\ restricted\ parameters) - D(model\ with\ flexible\ parameters)$$

  This is equivalent to the likelihood ratio test of

  $$G = -2 \cdot \log\left(\frac{L(model\ restricted\ parameters}{L(model\ with\ all\ parameters)}\right)$$

Akaike Information Criterion very similar to adjusted R square

- The Akaike Information Criterion (*AIC*) is a measure of model fit. It is defined as a deviance of a model which is penalized by its number of estimated parameters:

  $$AIC = \underbrace{-2 \cdot \log\big(L(model\ with\ k\ parameters)\big)}_{=D(model\ with\ k\ parameters)} + \boxed{2 \cdot (k+1)}$$

  penalty

- In general models with a ***smaller*** AIC are preferred.