Advanced Quantitative Research Methodology, Lecture Notes: Theories of Inference[1]

Gary King
http://GKing.Harvard.Edu

February 10, 2013

The Problem of Inference

The Problem of Inference

1. Probability: $\mathrm{P}(y|M) = \mathrm{P}(\text{known}|\text{unknown})$

The Problem of Inference

1. Probability: $\mathrm{P}(y|M) = \mathrm{P}(\text{known}|\text{unknown})$
2. The goal of inverse probability: $\mathrm{P}(M|y) = \mathrm{P}(\text{unknown}|\text{known})$.

The Problem of Inference

1. Probability: $\mathrm{P}(y|M) = \mathrm{P}(\text{known}|\text{unknown})$
2. The goal of inverse probability: $\mathrm{P}(M|y) = \mathrm{P}(\text{unknown}|\text{known})$.
3. Define: $M = \{M^*, \theta\}$, where M^* is assumed and θ is to be estimated

The Problem of Inference

1. Probability: $\mathrm{P}(y|M) = \mathrm{P}(\text{known}|\text{unknown})$
2. The goal of inverse probability: $\mathrm{P}(M|y) = \mathrm{P}(\text{unknown}|\text{known})$.
3. Define: $M = \{M^*, \theta\}$, where M^* is assumed and θ is to be estimated
4. More limited goal: $\mathrm{P}(\theta|y, M^*) \equiv \mathrm{P}(\theta|y)$.

The Problem of Inference

5. Bayes Theorem (no additional assumptions):

5. Bayes Theorem (no additional assumptions):

$$\mathrm{P}(\theta|y) = \frac{\mathrm{P}(\theta, y)}{\mathrm{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

The Problem of Inference

5. Bayes Theorem (no additional assumptions):

$$
\begin{aligned}
\mathrm{P}(\theta|y) &= \frac{\mathrm{P}(\theta, y)}{\mathrm{P}(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y)} && [\mathrm{P}(AB) = \mathrm{P}(B)\mathrm{P}(A|B)]
\end{aligned}
$$

5. Bayes Theorem (no additional assumptions):

$$
\begin{aligned}
P(\theta|y) &= \frac{P(\theta, y)}{P(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)} && [P(AB) = P(B)P(A|B)] \\
&= \frac{P(\theta)P(y|\theta)}{\int P(\theta)P(y|\theta)d\theta} && [P(A) = \int P(AB)dB]
\end{aligned}
$$

The Problem of Inference

5. Bayes Theorem (no additional assumptions):

$$\begin{aligned} \mathrm{P}(\theta|y) &= \frac{\mathrm{P}(\theta, y)}{\mathrm{P}(y)} && \text{[Defn. of conditional probability]} \\ &= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y)} && [\mathrm{P}(AB) = \mathrm{P}(B)\mathrm{P}(A|B)] \\ &= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(y|\theta)d\theta} && [\mathrm{P}(A) = \int \mathrm{P}(AB)dB] \end{aligned}$$

6. If we knew the right side, we could compute the inverse probability.

The Problem of Inference

5. Bayes Theorem (no additional assumptions):

$$
\begin{aligned}
P(\theta|y) &= \frac{P(\theta, y)}{P(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)} && [P(AB) = P(B)P(A|B)] \\
&= \frac{P(\theta)P(y|\theta)}{\int P(\theta)P(y|\theta)d\theta} && [P(A) = \int P(AB)dB]
\end{aligned}
$$

6. If we knew the right side, we could compute the inverse probability.
7. 2 theories of inference arose to interpret this result: likelihood and Bayesian

The Problem of Inference

5. Bayes Theorem (no additional assumptions):

$$\begin{aligned} \mathrm{P}(\theta|y) &= \frac{\mathrm{P}(\theta, y)}{\mathrm{P}(y)} && \text{[Defn. of conditional probability]} \\ &= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y)} && [\mathrm{P}(AB) = \mathrm{P}(B)\mathrm{P}(A|B)] \\ &= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(y|\theta)d\theta} && [\mathrm{P}(A) = \int \mathrm{P}(AB)dB] \end{aligned}$$

6. If we knew the right side, we could compute the inverse probability.
7. 2 theories of inference arose to interpret this result: likelihood and Bayesian
8. In both, $\mathrm{P}(y|\theta)$ is a traditional probability density

The Problem of Inference

5. Bayes Theorem (no additional assumptions):

$$\mathrm{P}(\theta|y) = \frac{\mathrm{P}(\theta, y)}{\mathrm{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

$$= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y)} \qquad [\mathrm{P}(AB) = \mathrm{P}(B)\mathrm{P}(A|B)]$$

$$= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(y|\theta)d\theta} \qquad [\mathrm{P}(A) = \int \mathrm{P}(AB)dB]$$

6. If we knew the right side, we could compute the inverse probability.
7. 2 theories of inference arose to interpret this result: likelihood and Bayesian
8. In both, $\mathrm{P}(y|\theta)$ is a traditional probability density
9. The two differ on the rest

Interpretation 1: The Likelihood Theory of Inference

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathrm{P}(\theta)}{\int P(\theta)P(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(y|\theta)d\theta} = k(y)P(y|\theta)$$

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathrm{P}(\theta)}{\int P(\theta)P(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(y|\theta)d\theta} = k(y)P(y|\theta)$$

4. *Define* $K(y)$ as an *unknown* function of y with θ fixed at its true value

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathrm{P}(\theta)}{\int P(\theta)P(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(y|\theta)d\theta} = k(y)P(y|\theta)$$

4. *Define* $K(y)$ as an *unknown* function of y with θ fixed at its true value
5. $\rightsquigarrow$ the likelihood theory of inference has four axioms: the 3 probability axioms plus the likelihood axiom:

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathrm{P}(\theta)}{\int P(\theta)P(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(y|\theta)d\theta} = k(y)P(y|\theta)$$

4. *Define* $K(y)$ as an *unknown* function of y with θ fixed at its true value
5. $\rightsquigarrow$ the likelihood theory of inference has four axioms: the 3 probability axioms plus the likelihood axiom:

$$L(\theta|y) \equiv k(y)P(y|\theta)$$

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathrm{P}(\theta)}{\int P(\theta)P(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(y|\theta)d\theta} = k(y)P(y|\theta)$$

4. *Define* $K(y)$ as an *unknown* function of y with θ fixed at its true value
5. $\rightsquigarrow$ the likelihood theory of inference has four axioms: the 3 probability axioms plus the likelihood axiom:

$$\begin{aligned} L(\theta|y) &\equiv k(y)P(y|\theta) \\ &\propto P(y|\theta) \end{aligned}$$

Interpretation 1: The Likelihood Theory of Inference

1. R.A. Fisher's idea
2. θ is fixed and y is random
3. Let:

$$k(y) \equiv \frac{\mathrm{P}(\theta)}{\int P(\theta)P(y|\theta)d\theta} \implies P(\theta|y) = \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(y|\theta)d\theta} = k(y)P(y|\theta)$$

4. *Define* $K(y)$ as an *unknown* function of y with θ fixed at its true value
5. $\rightsquigarrow$ the likelihood theory of inference has four axioms: the 3 probability axioms plus the likelihood axiom:

$$\begin{aligned} L(\theta|y) &\equiv k(y)P(y|\theta) \\ &\propto P(y|\theta) \end{aligned}$$

6. $L(\theta|y)$ is a function: for y fixed at the observed values, it gives the "likelihood" of any value of θ.

Interpretation 1: The Likelihood Theory of Inference

7. Likelihood: a relative measure of uncertainty, changing with the data set

Interpretation 1: The Likelihood Theory of Inference

7. Likelihood: a relative measure of uncertainty, changing with the data set
8. Comparing the value of $L(\theta|y)$ for different θ values in one data set y is meaningful.

7. Likelihood: a relative measure of uncertainty, changing with the data set
8. Comparing the value of $L(\theta|y)$ for different θ values in one data set y is meaningful.
9. Comparing values of $L(\theta|y)$ across data sets is meaningless. (just as you can't compare R^2 values across equations with different dependent variables.)

Interpretation 1: The Likelihood Theory of Inference

7. Likelihood: a relative measure of uncertainty, changing with the data set
8. Comparing the value of $L(\theta|y)$ for different θ values in one data set y is meaningful.
9. Comparing values of $L(\theta|y)$ across data sets is meaningless. (just as you can't compare R^2 values across equations with different dependent variables.)
10. The likelihood principle: the data only affect inferences through the likelihood function

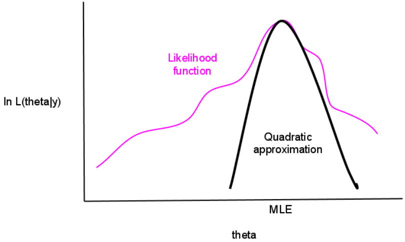
Visualizing the Likelihood

Visualizing the Likelihood

- For algebraic simplicity and numerical stability, we use the log-likelihood (the shape changes, but the max is in the same place)

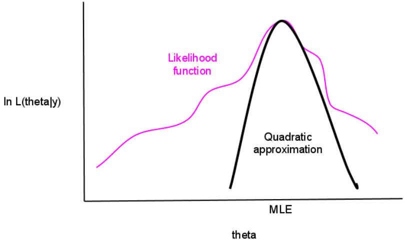
Visualizing the Likelihood

- For algebraic simplicity and numerical stability, we use the log-likelihood (the shape changes, but the max is in the same place)
- If θ has one element, we can plot:



Visualizing the Likelihood

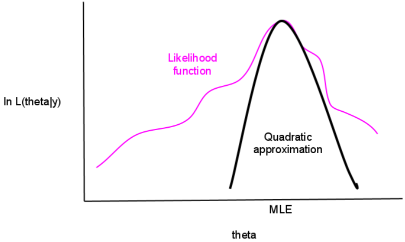
- For algebraic simplicity and numerical stability, we use the log-likelihood (the shape changes, but the max is in the same place)
- If θ has one element, we can plot:



- The full likelihood curve is a Summary Estimator. The likelihood principle means that once this is plotted, we can discard the data (if the model is correct!).

Visualizing the Likelihood

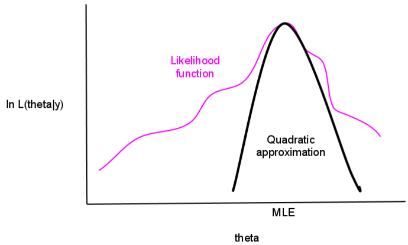
- For algebraic simplicity and numerical stability, we use the log-likelihood (the shape changes, but the max is in the same place)
- If θ has one element, we can plot:



- The full likelihood curve is a Summary Estimator. The likelihood principle means that once this is plotted, we can discard the data (if the model is correct!).
- A one-point summary at the maximum is the MLE

Visualizing the Likelihood

- For algebraic simplicity and numerical stability, we use the log-likelihood (the shape changes, but the max is in the same place)
- If θ has one element, we can plot:



- The full likelihood curve is a Summary Estimator. The likelihood principle means that once this is plotted, we can discard the data (if the model is correct!).
- A one-point summary at the maximum is the MLE
- Uncertainty of point estimate: curvature at the maximum

Interpretation 2: The Bayesian Theory of Inference

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' unpublished idea, and later rediscovered.

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' unpublished idea, and later rediscovered.
- Recall:

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' unpublished idea, and later rediscovered.
- Recall:

$$\mathrm{P}(\theta|y) = \frac{\mathrm{P}(\theta, y)}{\mathrm{P}(y)} \qquad \text{[Defn. of conditional probability]}$$

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' unpublished idea, and later rediscovered.
- Recall:

$$\begin{aligned} \mathrm{P}(\theta|y) &= \frac{\mathrm{P}(\theta, y)}{\mathrm{P}(y)} && \text{[Defn. of conditional probability]} \\ &= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y)} && [\mathrm{P}(AB) = \mathrm{P}(B)\mathrm{P}(A|B)] \end{aligned}$$

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' unpublished idea, and later rediscovered.
- Recall:

$$\begin{aligned} P(\theta|y) &= \frac{P(\theta, y)}{P(y)} && \text{[Defn. of conditional probability]} \\ &= \frac{P(\theta)P(y|\theta)}{P(y)} && [P(AB) = P(B)P(A|B)] \\ &= \frac{P(\theta)P(y|\theta)}{\int P(\theta)P(y|\theta)d\theta} && [P(A) = \int P(AB)dB] \end{aligned}$$

- Rev. Thomas Bayes' unpublished idea, and later rediscovered.
- Recall:

$$
\begin{aligned}
\mathrm{P}(\theta|y) &= \frac{\mathrm{P}(\theta, y)}{\mathrm{P}(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\mathrm{P}(y)} && [\mathrm{P}(AB) = \mathrm{P}(B)\mathrm{P}(A|B)] \\
&= \frac{\mathrm{P}(\theta)\mathrm{P}(y|\theta)}{\int \mathrm{P}(\theta)\mathrm{P}(y|\theta)d\theta} && [\mathrm{P}(A) = \int \mathrm{P}(AB)dB] \\
&\propto \mathrm{P}(\theta)\mathrm{P}(y|\theta)
\end{aligned}
$$

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' unpublished idea, and later rediscovered.
- Recall:

$$
\begin{aligned}
P(\theta|y) &= \frac{P(\theta, y)}{P(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)} && [P(AB) = P(B)P(A|B)] \\
&= \frac{P(\theta)P(y|\theta)}{\int P(\theta)P(y|\theta)d\theta} && \left[P(A) = \int P(AB)dB\right] \\
&\propto P(\theta)P(y|\theta)
\end{aligned}
$$

- $P(\theta|y)$ the posterior density

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' unpublished idea, and later rediscovered.
- Recall:

$$
\begin{aligned}
P(\theta|y) &= \frac{P(\theta, y)}{P(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)} && [P(AB) = P(B)P(A|B)] \\
&= \frac{P(\theta)P(y|\theta)}{\int P(\theta)P(y|\theta)d\theta} && [P(A) = \int P(AB)dB] \\
&\propto P(\theta)P(y|\theta)
\end{aligned}
$$

- $P(\theta|y)$ the posterior density
- $P(y|\theta)$ the traditional probability ($\propto$ likelihood)

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' unpublished idea, and later rediscovered.
- Recall:

$$\begin{aligned}
P(\theta|y) &= \frac{P(\theta, y)}{P(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)} && [P(AB) = P(B)P(A|B)] \\
&= \frac{P(\theta)P(y|\theta)}{\int P(\theta)P(y|\theta)d\theta} && [P(A) = \int P(AB)dB] \\
&\propto P(\theta)P(y|\theta)
\end{aligned}$$

- $P(\theta|y)$ the posterior density
- $P(y|\theta)$ the traditional probability ($\propto$ likelihood)
- $P(y)$ a constant, easily computed

Interpretation 2: The Bayesian Theory of Inference

- Rev. Thomas Bayes' unpublished idea, and later rediscovered.
- Recall:

$$\begin{aligned}
P(\theta|y) &= \frac{P(\theta, y)}{P(y)} && \text{[Defn. of conditional probability]} \\
&= \frac{P(\theta)P(y|\theta)}{P(y)} && [P(AB) = P(B)P(A|B)] \\
&= \frac{P(\theta)P(y|\theta)}{\int P(\theta)P(y|\theta)d\theta} && [P(A) = \int P(AB)dB] \\
&\propto P(\theta)P(y|\theta)
\end{aligned}$$

- $P(\theta|y)$ the posterior density
- $P(y|\theta)$ the traditional probability ($\propto$ likelihood)
- $P(y)$ a constant, easily computed
- $P(\theta)$, the prior density — the way Bayes differs from likelihood

What is the prior density, $P(\theta)$?

What is the prior density, $P(\theta)$?

1. A probability density that represents all prior evidence about θ.

What is the prior density, $P(\theta)$?

1. A probability density that represents all prior evidence about θ.
2. An opportunity: a way of getting other information outside the data set into the model

What is the prior density, $P(\theta)$?

1. A probability density that represents all prior evidence about θ.
2. An opportunity: a way of getting other information outside the data set into the model
3. An annoyance: the "other information" is required

What is the prior density, $P(\theta)$?

1. A probability density that represents all prior evidence about θ.
2. An opportunity: a way of getting other information outside the data set into the model
3. An annoyance: the "other information" is required
4. A philosophical assumption that nonsample information should matter (as it always does) *and* be formalized and included in all inferences.

Principles of Bayesian analysis

Principles of Bayesian analysis

1. All unknown quantities (θ, Y) are treated as random variables and have a joint probability distribution.

Principles of Bayesian analysis

1. All unknown quantities (θ, Y) are treated as random variables and have a joint probability distribution.
2. All known quantities (y) are treated as fixed.

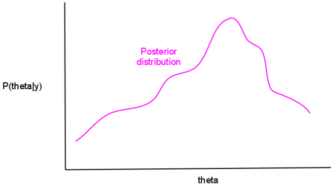
Principles of Bayesian analysis

1. All unknown quantities $(\theta,\ Y)$ are treated as random variables and have a joint probability distribution.
2. All known quantities (y) are treated as fixed.
3. If we have observed variable B and unobserved variable A, then we are usually interested in the conditional distribution of A, given B: $P(A|B) = P(A,B)/P(B)$

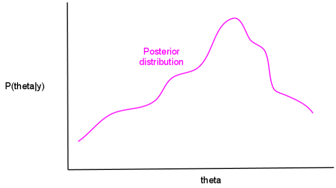
Principles of Bayesian analysis

1. All unknown quantities (θ, Y) are treated as random variables and have a joint probability distribution.
2. All known quantities (y) are treated as fixed.
3. If we have observed variable B and unobserved variable A, then we are usually interested in the conditional distribution of A, given B: $P(A|B) = P(A, B)/P(B)$
4. If variables A and B are both unknown, then the distribution of A alone is $P(A) = \int \mathrm{P}(A, B)dB = \int P(A|B)P(B)dB$.

The posterior density, $P(\theta|y)$

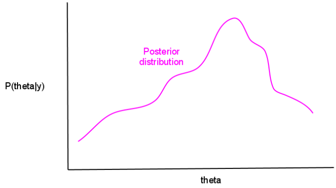


The posterior density, $P(\theta|y)$



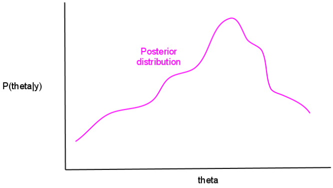
- Like L, it's a summary estimator

The posterior density, P($\theta|y$)



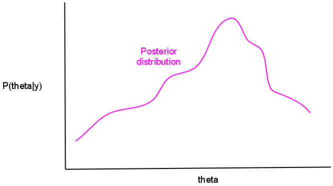
- Like L, it's a summary estimator
- Unlike L, it's a real probability density, from which we can derive probabilistic statements (via integration)

The posterior density, $P(\theta|y)$



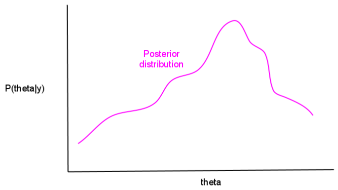
- Like L, it's a summary estimator
- Unlike L, it's a real probability density, from which we can derive probabilistic statements (via integration)
- To compare across applications or data sets, you may need different priors. So, the posterior is also relative, just like likelihood.

The posterior density, $P(\theta|y)$



- Like L, it's a summary estimator
- Unlike L, it's a real probability density, from which we can derive probabilistic statements (via integration)
- To compare across applications or data sets, you may need different priors. So, the posterior is also relative, just like likelihood.
- Bayesian inference obeys the likelihood principle: the data set only affects inferences through the likelihood function

The posterior density, $P(\theta|y)$



- Like L, it's a summary estimator
- Unlike L, it's a real probability density, from which we can derive probabilistic statements (via integration)
- To compare across applications or data sets, you may need different priors. So, the posterior is also relative, just like likelihood.
- Bayesian inference obeys the likelihood principle: the data set only affects inferences through the likelihood function
- If $P(\theta) = 1$, i.e., is uniform in the relevant region, then $L(\theta|y) = P(\theta|y)$.

Bayesians are happier people

Bayesians are happier people

- If P(θ) is *diffuse*, differences from likelihood are minor, but numerical stability (and "identification") is improved (your programs will run better!).

Bayesians are happier people

- If $P(\theta)$ is *diffuse*, differences from likelihood are minor, but numerical stability (and "identification") is improved (your programs will run better!).
- Philosophical differences from likelihood: Huge

Bayesians are happier people

- If $P(\theta)$ is *diffuse*, differences from likelihood are minor, but numerical stability (and "identification") is improved (your programs will run better!).
- Philosophical differences from likelihood: Huge
- Practical differences when we can compute both: Minor (unless the prior matters)

Bayesians are happier people

- If $P(\theta)$ is *diffuse*, differences from likelihood are minor, but numerical stability (and "identification") is improved (your programs will run better!).
- Philosophical differences from likelihood: Huge
- Practical differences when we can compute both: Minor (unless the prior matters)
- Advantages: more information produces more efficiency; MCMC algorithms are easier with Bayes.

Bayesians are happier people

- If $P(\theta)$ is *diffuse*, differences from likelihood are minor, but numerical stability (and "identification") is improved (your programs will run better!).
- Philosophical differences from likelihood: Huge
- Practical differences when we can compute both: Minor (unless the prior matters)
- Advantages: more information produces more efficiency; MCMC algorithms are easier with Bayes.
- Few fights now between Bayesians and likelihoodists

A 3rd Theory: Neyman-Pearson Hypothesis Testing

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the $\{$Bayesians, Likelihoodists$\}$

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0
4. For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0
4. For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?
   (a) H_0: $\beta = 0$ vs. H_1: $\beta > 0$

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0
4. For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?
   (a) H_0: $\beta = 0$ vs. H_1: $\beta > 0$
   (b) Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0
4. For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?
   (a) H_0: $\beta = 0$ vs. H_1: $\beta > 0$
   (b) Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$
   (c) Type II error, the power to detect H_1 if it is true, is a consequence of choosing an estimator, not an ex ante decision like choosing α.

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0
4. For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?
   (a) H_0: $\beta = 0$ vs. H_1: $\beta > 0$
   (b) Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$
   (c) Type II error, the power to detect H_1 if it is true, is a consequence of choosing an estimator, not an ex ante decision like choosing α.
   (d) Assume n is large enough for the CLT to kick in

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0
4. For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?
   (a) H_0: $\beta = 0$ vs. H_1: $\beta > 0$
   (b) Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$
   (c) Type II error, the power to detect H_1 if it is true, is a consequence of choosing an estimator, not an ex ante decision like choosing α.
   (d) Assume n is large enough for the CLT to kick in
   (e) Then $b|(\beta = 0) \sim N(0, \sigma_b^2)$

A 3rd Theory: Neyman-Pearson Hypothesis Testing

1. Huge fights between these folks and the {Bayesians, Likelihoodists}
2. Strict but arbitrary distinction: null H_0 vs alternative H_1 hypotheses
3. All tests are "under" (i.e., assuming) H_0
4. For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?
    (a) H_0: $\beta = 0$ vs. H_1: $\beta > 0$
    (b) Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$
    (c) Type II error, the power to detect H_1 if it is true, is a consequence of choosing an estimator, not an ex ante decision like choosing α.
    (d) Assume n is large enough for the CLT to kick in
    (e) Then $b|(\beta = 0) \sim N(0, \sigma_b^2)$
    (f) or

$$(TS)_\beta|(\beta = 0) \equiv \frac{b - \beta}{\hat{\sigma}_b} \equiv \frac{b}{\hat{\sigma}_b} \sim N(0, 1).$$

Neyman-Pearson Hypothesis Testing

Neyman-Pearson Hypothesis Testing

(g) Derive critical value, CV, e.g., the right tail:

$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2)db = \alpha$$

Neyman-Pearson Hypothesis Testing

(g) Derive critical value, CV, e.g., the right tail:

$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2) db = \alpha$$

(h) This means, in educational psychology and other fields, write your prospectus, plan your experiment, report the CV, and write your concluding chapter:

Neyman-Pearson Hypothesis Testing

(g) Derive critical value, CV, e.g., the right tail:

$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2) db = \alpha$$

(h) This means, in educational psychology and other fields, write your prospectus, plan your experiment, report the CV, and write your concluding chapter:

$$\text{Decision} =$$

Neyman-Pearson Hypothesis Testing

(g) Derive critical value, CV, e.g., the right tail:

$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2)db = \alpha$$

(h) This means, in educational psychology and other fields, write your prospectus, plan your experiment, report the CV, and write your concluding chapter:

$$\text{Decision} = \begin{cases} \beta > 0 \text{ (I was right)} & \text{if } (TS) > (CV) \\ \beta = 0 \text{ (I was wrong)} & \text{if } (TS) \leq (CV) \end{cases}$$

Neyman-Pearson Hypothesis Testing

(g) Derive critical value, CV, e.g., the right tail:

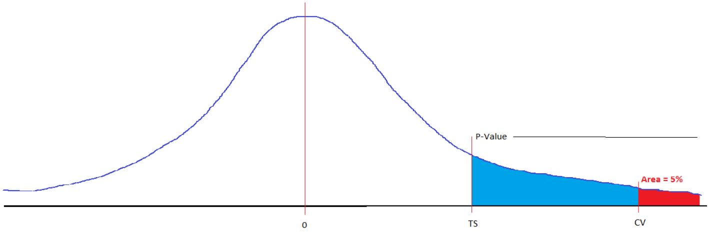
$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2)db = \alpha$$

(h) This means, in educational psychology and other fields, write your prospectus, plan your experiment, report the CV, and write your concluding chapter:

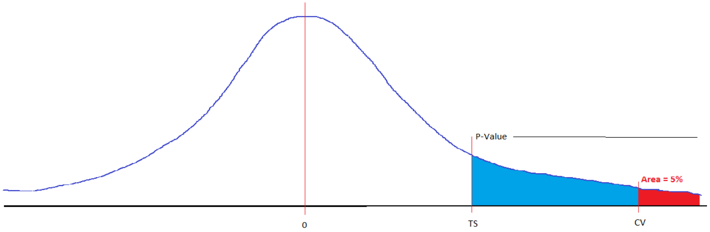
$$\text{Decision} = \begin{cases} \beta > 0 \text{ (I was right)} & \text{if } (TS) > (CV) \\ \beta = 0 \text{ (I was wrong)} & \text{if } (TS) \leq (CV) \end{cases}$$

And then first collect your data. You may not revise your hypothesis or your theory.

Neyman-Pearson Hypothesis Testing

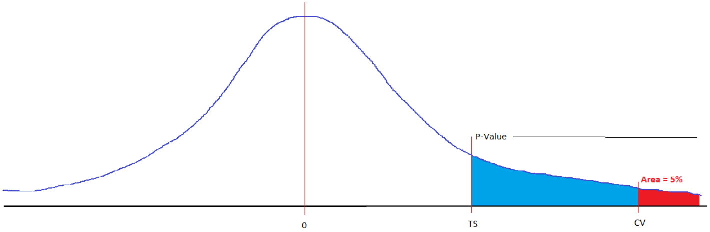


Neyman-Pearson Hypothesis Testing



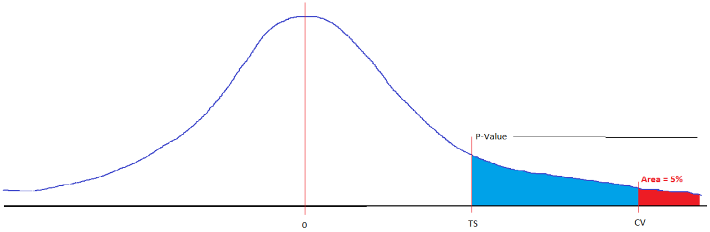
(i) In this example, $(TS) < (CV)$ and so we conclude that $\beta = 0$.

Neyman-Pearson Hypothesis Testing



(i) In this example, $(TS) < (CV)$ and so we conclude that $\beta = 0$.
(j) Decision will be wrong 5% of the time; what about this time?

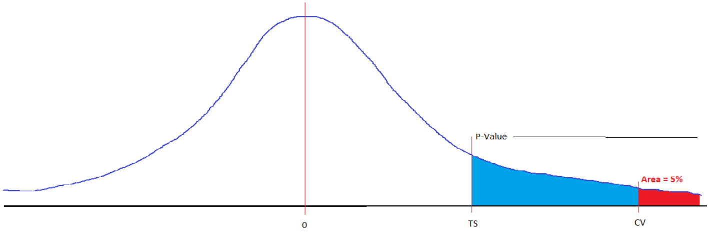
Neyman-Pearson Hypothesis Testing



(i) In this example, $(TS) < (CV)$ and so we conclude that $\beta = 0$.
(j) Decision will be wrong 5% of the time; what about this time?

5. What about when n is large or under control of the investigator?

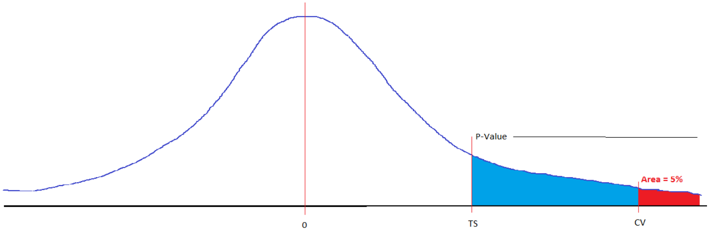
Neyman-Pearson Hypothesis Testing



- (i) In this example, $(TS) < (CV)$ and so we conclude that $\beta = 0$.
- (j) Decision will be wrong 5% of the time; what about this time?

5. What about when n is large or under control of the investigator?
6. In practice, hypothesis testing is used with p-values:

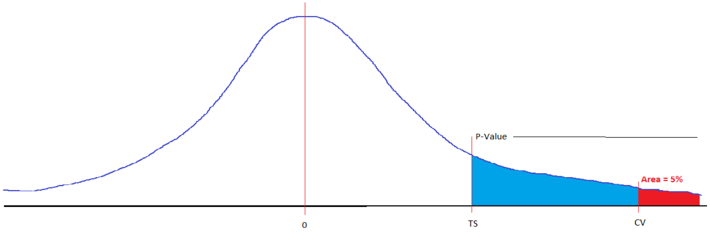
Neyman-Pearson Hypothesis Testing



(i) In this example, $(TS) < (CV)$ and so we conclude that $\beta = 0$.
(j) Decision will be wrong 5% of the time; what about this time?

5. What about when n is large or under control of the investigator?
6. In practice, hypothesis testing is used with p-values: The probability under the null of getting a value as weird or weirder than the value we got — the area to the right of the realized value of (TS).

Neyman-Pearson Hypothesis Testing

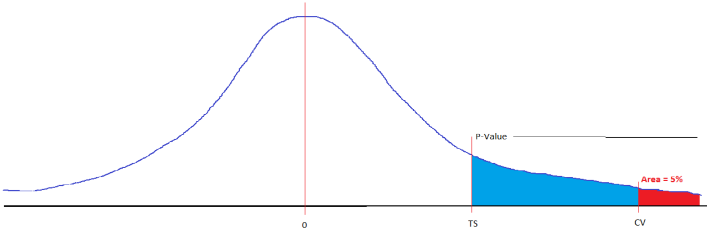


(i) In this example, $(TS) < (CV)$ and so we conclude that $\beta = 0$.

(j) Decision will be wrong 5% of the time; what about this time?

5. What about when n is large or under control of the investigator?
6. In practice, hypothesis testing is used with p-values: The probability under the null of getting a value as weird or weirder than the value we got — the area to the right of the realized value of (TS).
7. Star-gazing is silly; where is the quantity of interest?

Neyman-Pearson Hypothesis Testing



(i) In this example, $(TS) < (CV)$ and so we conclude that $\beta = 0$.
(j) Decision will be wrong 5% of the time; what about this time?

5. What about when n is large or under control of the investigator?
6. In practice, hypothesis testing is used with p-values: The probability under the null of getting a value as weird or weirder than the value we got — the area to the right of the realized value of (TS).
7. Star-gazing is silly; where is the quantity of interest?
8. We can use likelihood to compute hypothesis tests and p-values.

What is the right theory of inference?

What is the right theory of inference?

1. Likelihood?

What is the right theory of inference?

1. Likelihood? Bayes?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference?

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. No

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None of

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None of these.

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None of these.
3. The right theory of inference:

What is the right theory of inference?

1. Likelihood? Bayes? Neyman-Pearson? Criteria estimators? Finite or asymptotic based theory? Decision theory? Nonparametrics? Semiparametrics? Conditional inference? Superpopulation-based inference? etc.
2. None of these.
3. The right theory of inference: utilitarianism

Unification of Theories of Inference

Unification of Theories of Inference

- Can't bank on agreement on normative issues!

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occuring: different methods giving the same result.

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occuring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occuring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occuring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods
  - Matching for use as preprocessing for parametric analysis

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occuring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods
  - Matching for use as preprocessing for parametric analysis
  - Bayesian model averaging, with a large enough class of models to average over

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occuring: different methods giving the same result.
    - Likelihood or Bayes with careful goodness of fit checks
    - Various types of robust or semi-parametric methods
    - Matching for use as preprocessing for parametric analysis
    - Bayesian model averaging, with a large enough class of models to average over
    - Committee methods, mixture of experts models

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occuring: different methods giving the same result.
    - Likelihood or Bayes with careful goodness of fit checks
    - Various types of robust or semi-parametric methods
    - Matching for use as preprocessing for parametric analysis
    - Bayesian model averaging, with a large enough class of models to average over
    - Committee methods, mixture of experts models
    - Some models with highly flexible functional forms

Unification of Theories of Inference

- Can't bank on agreement on normative issues!
- Even if there is agreement, it won't hold or shouldn't
- Alternative convergence is occuring: different methods giving the same result.
  - Likelihood or Bayes with careful goodness of fit checks
  - Various types of robust or semi-parametric methods
  - Matching for use as preprocessing for parametric analysis
  - Bayesian model averaging, with a large enough class of models to average over
  - Committee methods, mixture of experts models
  - Some models with highly flexible functional forms
- The key: No assumptions can be trusted; all theories of inference condition on assumptions and so data analysts always struggle trying to understand and get around them

A Simple Likelihood Model: Stylized Normal, no X

A Simple Likelihood Model: Stylized Normal, no X

The model:

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of *all* observations, Pr(data|model)
(Recall: if A and B are independent, $\text{P}(AB) = \text{P}(A)\text{P}(B)$):

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of *all* observations, Pr(data|model)
(Recall: if A and B are independent, $\mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B)$):

$$\mathrm{P}(y|\mu) \equiv \mathrm{P}(y_1, \ldots, y_n|\mu_1, \ldots, \mu_n) = \prod_{i=1}^{n} f_{\text{stn}}(y_i|\mu_i)$$

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of *all* observations, Pr(data|model) (Recall: if A and B are independent, $\text{P}(AB) = \text{P}(A)\text{P}(B)$):

$$\begin{aligned}
\text{P}(y|\mu) \equiv \text{P}(y_1, \ldots, y_n|\mu_1, \ldots, \mu_n) &= \prod_{i=1}^{n} f_{\text{stn}}(y_i|\mu_i) \\
&= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2}\right)
\end{aligned}$$

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall \ i \neq j$.

Derive the full probability density of *all* observations, Pr(data|model) (Recall: if A and B are independent, $\text{P}(AB) = \text{P}(A)\text{P}(B)$):

$$\begin{aligned} \text{P}(y|\mu) \equiv \text{P}(y_1, \ldots, y_n|\mu_1, \ldots, \mu_n) &= \prod_{i=1}^{n} f_{\text{stn}}(y_i|\mu_i) \\ &= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2}\right) \end{aligned}$$

reparameterizing with $\mu_i = \beta$:

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of *all* observations, Pr(data|model)
(Recall: if A and B are independent, $\mathrm{P}(AB) = \mathrm{P}(A)\mathrm{P}(B)$):

$$\mathrm{P}(y|\mu) \equiv \mathrm{P}(y_1, \dots, y_n|\mu_1, \dots, \mu_n) = \prod_{i=1}^{n} f_{\text{stn}}(y_i|\mu_i)$$

$$= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2}\right)$$

reparameterizing with $\mu_i = \beta$:

$$\mathrm{P}(y|\beta) \equiv \mathrm{P}(y_1, \dots, y_n|\beta) = \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \beta)^2}{2}\right)$$

A Simple Likelihood Model: Stylized Normal, no X

The model:

1. $Y_i \sim f_{\text{stn}}(y_i|\mu_i)$, normal stochastic component
2. $\mu_i = \beta$, a constant systematic component (no covariates)
3. Y_i and Y_j are independent $\forall\ i \neq j$.

Derive the full probability density of *all* observations, Pr(data|model) (Recall: if A and B are independent, $\text{P}(AB) = \text{P}(A)\text{P}(B)$):

$$\begin{aligned}\text{P}(y|\mu) \equiv \text{P}(y_1, \ldots, y_n|\mu_1, \ldots, \mu_n) &= \prod_{i=1}^{n} f_{\text{stn}}(y_i|\mu_i) \\ &= \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2}\right)\end{aligned}$$

reparameterizing with $\mu_i = \beta$:

$$\text{P}(y|\beta) \equiv \text{P}(y_1, \ldots, y_n|\beta) = \prod_{i=1}^{n} (2\pi)^{-1/2} \exp\left(\frac{-(y_i - \beta)^2}{2}\right)$$

- What can you do with this density?

Stylized Normal Likelihood Function

$$L(\beta|y) = k(y) \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)$$

Stylized Normal Likelihood Function

$$
\begin{aligned}
L(\beta|y) &= k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\
&\propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)
\end{aligned}
$$

Stylized Normal Likelihood Function

$$
\begin{aligned}
L(\beta|y) &= k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\
&\propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\
&= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)
\end{aligned}
$$

Stylized Normal Likelihood Function

$$\begin{aligned} L(\beta|y) &= k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\ &\propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\ &= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right) \end{aligned}$$

the relative likelihood of β (conditional on the model) having generated the data we observe.

Stylized Normal Likelihood Function

$$\begin{aligned} L(\beta|y) &= k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\ &\propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\ &= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right) \end{aligned}$$

the relative likelihood of β (conditional on the model) having generated the data we observe.

The log-likelihood (Recall: $\ln(ab) = \ln(a) + \ln(b)$):

Stylized Normal Likelihood Function

$$
\begin{aligned}
L(\beta|y) &= k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\
&\propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\
&= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)
\end{aligned}
$$

the relative likelihood of β (conditional on the model) having generated the data we observe.

The log-likelihood (Recall: $\ln(ab) = \ln(a) + \ln(b)$):

$$
\ln L(\beta|y) = \ln[k(y)] + \sum_{i=1}^{n} \ln f_{\text{stn}}(y_i|\beta)
$$

Stylized Normal Likelihood Function

$$\begin{aligned}
L(\beta|y) &= k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\
&\propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\
&= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)
\end{aligned}$$

the relative likelihood of β (conditional on the model) having generated the data we observe.

The log-likelihood (Recall: $\ln(ab) = \ln(a) + \ln(b)$):

$$\begin{aligned}
\ln L(\beta|y) &= \ln[k(y)] + \sum_{i=1}^{n}\ln f_{\text{stn}}(y_i|\beta) \\
&= \ln[k(y)] + \sum_{i=1}^{n}\ln[(2\pi)^{-1/2}] - \sum_{i=1}^{n}\frac{1}{2}(y_i-\beta)^2
\end{aligned}$$

Stylized Normal Likelihood Function

$$\begin{aligned} L(\beta|y) &= k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\ &\propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta) \\ &= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right) \end{aligned}$$

the relative likelihood of β (conditional on the model) having generated the data we observe.
The log-likelihood (Recall: $\ln(ab) = \ln(a) + \ln(b)$):

$$\begin{aligned} \ln L(\beta|y) &= \ln[k(y)] + \sum_{i=1}^{n} \ln f_{\text{stn}}(y_i|\beta) \\ &= \ln[k(y)] + \sum_{i=1}^{n} \ln[(2\pi)^{-1/2}] - \sum_{i=1}^{n} \frac{1}{2}(y_i-\beta)^2 \\ &\doteq \sum_{i=1}^{n} -\frac{1}{2}(y_i-\beta)^2 \end{aligned}$$

Stylized Normal Likelihood Function

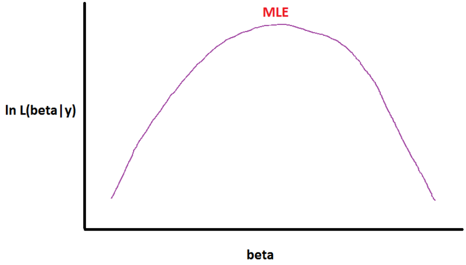
$$
\begin{aligned}
L(\beta|y) &= k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)\\
&\propto \prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)\\
&= \prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)
\end{aligned}
$$

the relative likelihood of β (conditional on the model) having generated the data we observe.

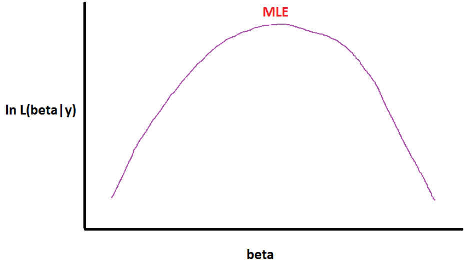
The log-likelihood (Recall: $\ln(ab)=\ln(a)+\ln(b)$):

$$
\begin{aligned}
\ln L(\beta|y) &= \ln[k(y)] + \sum_{i=1}^{n}\ln f_{\text{stn}}(y_i|\beta)\\
&= \ln[k(y)] + \sum_{i=1}^{n}\ln[(2\pi)^{-1/2}] - \sum_{i=1}^{n}\frac{1}{2}(y_i-\beta)^2\\
&\doteq \sum_{i=1}^{n} -\frac{1}{2}(y_i-\beta)^2\\
&= -\frac{1}{2}\sum_{i=1}^{n}(y_i-\beta)^2
\end{aligned}
$$

Log-likelihood interpretation

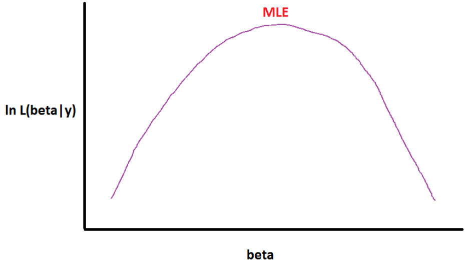


Log-likelihood interpretation



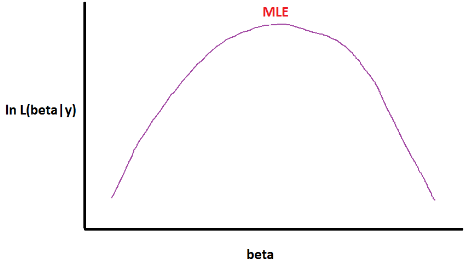
1. The log-likelihood is quadratic

Log-likelihood interpretation



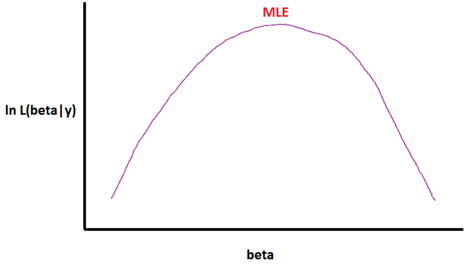
1. The log-likelihood is quadratic
2. The MLE is at the same point as the MVLUE

Log-likelihood interpretation



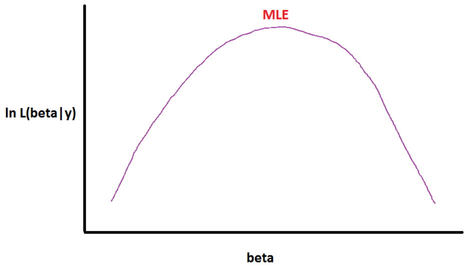
1. The log-likelihood is quadratic
2. The MLE is at the same point as the MVLUE
3. The maximum is at the same point as the least squares point

Log-likelihood interpretation



1. The log-likelihood is quadratic
2. The MLE is at the same point as the MVLUE
3. The maximum is at the same point as the least squares point
4. By the theory, this curve summarizes all information the data gives about β, assuming the model.

Log-likelihood interpretation



1. The log-likelihood is quadratic
2. The MLE is at the same point as the MVLUE
3. The maximum is at the same point as the least squares point
4. By the theory, this curve summarizes all information the data gives about β, assuming the model.
5. No reason to summarize this curve with only the MLE

Summarizing k-dimensional space

Summarizing k-dimensional space

1. The problem of Flatland

Summarizing k-dimensional space

1. The problem of Flatland
2. Graphs

Summarizing k-dimensional space

1. The problem of Flatland
2. Graphs
3. Maximum

Summarizing k-dimensional space

1. The problem of Flatland
2. Graphs
3. Maximum
4. The curvature at the maximum (standard errors, about which more shortly)

Summarizing k-dimensional space

1. The problem of Flatland
2. Graphs
3. Maximum
4. The curvature at the maximum (standard errors, about which more shortly)
5. The curse of dimensionality

How to find the maximum?

How to find the maximum?

1. Analytically — often impossible or too hard

How to find the maximum?

1. Analytically — often impossible or too hard
   (a) Take the derivative of $\ln L(\theta|y)$ w.r.t. θ

How to find the maximum?

1. Analytically — often impossible or too hard
   (a) Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   (b) Set to 0, substituting $\hat{\theta}$ for θ

How to find the maximum?

1. Analytically — often impossible or too hard
   (a) Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   (b) Set to 0, substituting $\hat{\theta}$ for θ

$$\left|\frac{\partial \ln L(\theta|y)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

How to find the maximum?

1. Analytically — often impossible or too hard
   (a) Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   (b) Set to 0, substituting $\hat{\theta}$ for θ

$$\left|\frac{\partial \ln L(\theta|y)}{\partial\theta}\right|_{\theta=\hat{\theta}} = 0$$

   (c) If possible, solve for θ, and label it $\hat{\theta}$

How to find the maximum?

1. Analytically — often impossible or too hard
   (a) Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   (b) Set to 0, substituting $\hat{\theta}$ for θ

$$\left|\frac{\partial \ln L(\theta|y)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

   (c) If possible, solve for θ, and label it $\hat{\theta}$
   (d) Check the second order condition: see if the second derivative w.r.t. θ is negative (so its a maximum rather than a minimum)

How to find the maximum?

1. Analytically — often impossible or too hard
   (a) Take the derivative of $\ln L(\theta|y)$ w.r.t. θ
   (b) Set to 0, substituting $\hat{\theta}$ for θ

$$\left|\frac{\partial \ln L(\theta|y)}{\partial \theta}\right|_{\theta=\hat{\theta}} = 0$$

   (c) If possible, solve for θ, and label it $\hat{\theta}$
   (d) Check the second order condition: see if the second derivative w.r.t. θ is negative (so its a maximum rather than a minimum)
2. Numerically — let the computer do the work for you

Properties of the maximum

Properties of the maximum

1. Finite Sample Properties

Properties of the maximum

1. Finite Sample Properties
   (a) Invariance to Reparameterization

Properties of the maximum

1. Finite Sample Properties
   (a) Invariance to Reparameterization
      i. Estimate σ with $\hat{\sigma}$ and calculate $\hat{\sigma}^2$ or estimate $\hat{\sigma}^2$: both are MLEs

Properties of the maximum

1. Finite Sample Properties
   (a) Invariance to Reparameterization
      i. Estimate σ with $\hat{\sigma}$ and calculate $\hat{\sigma}^2$ or estimate $\hat{\sigma}^2$: both are MLEs
      ii. Not true for other methods of inference: e.g. $\bar{y}$ is an unbiased estimate of μ. What is an unbiased estimate of $1/\mu$? $E(1/\bar{y}) \neq 1/E(\bar{y})$.

Properties of the maximum

1. Finite Sample Properties
   (a) Invariance to Reparameterization
      i. Estimate σ with $\hat{\sigma}$ and calculate $\hat{\sigma}^2$ or estimate $\hat{\sigma}^2$: both are MLEs
      ii. Not true for other methods of inference: e.g. $\bar{y}$ is an unbiased estimate of μ. What is an unbiased estimate of $1/\mu$? $E(1/\bar{y}) \neq 1/E(\bar{y})$.
   (b) Invariance to sampling plans: ok to look at results while deciding how much data to collect

Properties of the maximum

1. Finite Sample Properties
   (a) Invariance to Reparameterization
      i. Estimate σ with $\hat{\sigma}$ and calculate $\hat{\sigma}^2$ or estimate $\hat{\sigma}^2$: both are MLEs
      ii. Not true for other methods of inference: e.g. $\bar{y}$ is an unbiased estimate of μ. What is an unbiased estimate of $1/\mu$? $E(1/\bar{y}) \neq 1/E(\bar{y})$.
   (b) Invariance to sampling plans: ok to look at results while deciding how much data to collect
   (c) Minimum variance unbiased estimator (MVUE)

Properties of the maximum

1. Finite Sample Properties
   (a) Invariance to Reparameterization
      i. Estimate σ with $\hat{\sigma}$ and calculate $\hat{\sigma}^2$ or estimate $\hat{\sigma}^2$: both are MLEs
      ii. Not true for other methods of inference: e.g. $\bar{y}$ is an unbiased estimate of μ. What is an unbiased estimate of $1/\mu$? $E(1/\bar{y}) \neq 1/E(\bar{y})$.
   (b) Invariance to sampling plans: ok to look at results while deciding how much data to collect
   (c) Minimum variance unbiased estimator (MVUE)
      i. If there is one, ML will find it

Properties of the maximum

1. Finite Sample Properties
   (a) Invariance to Reparameterization
      i. Estimate σ with $\hat{\sigma}$ and calculate $\hat{\sigma}^2$ or estimate $\hat{\sigma}^2$: both are MLEs
      ii. Not true for other methods of inference: e.g. $\bar{y}$ is an unbiased estimate of μ. What is an unbiased estimate of $1/\mu$? $E(1/\bar{y}) \neq 1/E(\bar{y})$.
   (b) Invariance to sampling plans: ok to look at results while deciding how much data to collect
   (c) Minimum variance unbiased estimator (MVUE)
      i. If there is one, ML will find it
      ii. If there isn't one, ML will still find a good estimator, even when other methods of inference won't.

Properties of the maximum

2. Asymptotic Properties

Properties of the maximum

2. Asymptotic Properties
   (a) Consistency (from the Law of Large Numbers). As $n \rightarrow \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value

Properties of the maximum

2. Asymptotic Properties
   (a) Consistency (from the Law of Large Numbers). As $n \to \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value
   (b) Asymptotic normality (from the central limit theorem). As $n \to \infty$, the distribution of MLE/se(MLE) converges to a Normal.

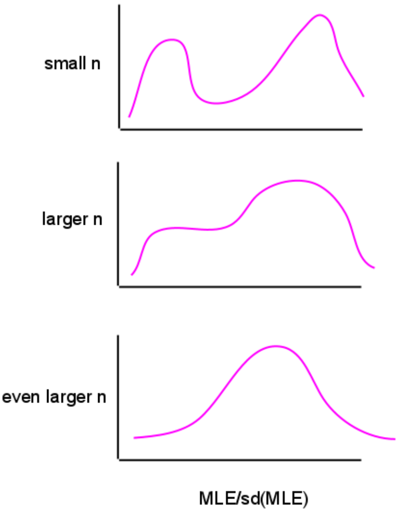
Properties of the maximum

2. Asymptotic Properties
   (a) Consistency (from the Law of Large Numbers). As $n \to \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value
   (b) Asymptotic normality (from the central limit theorem). As $n \to \infty$, the distribution of MLE/se(MLE) converges to a Normal.
   • Do the LLN and CLT (the 2 most important theorems in statistics) contradict each other?

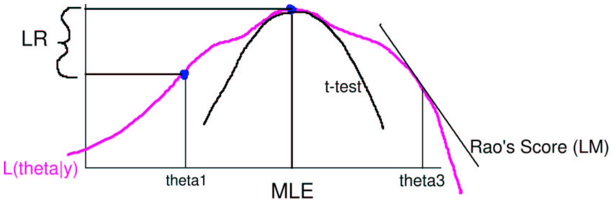
Properties of the maximum

2. Asymptotic Properties
   (a) Consistency (from the Law of Large Numbers). As $n \rightarrow \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value
   (b) Asymptotic normality (from the central limit theorem). As $n \rightarrow \infty$, the distribution of MLE/se(MLE) converges to a Normal.
   • Do the LLN and CLT (the 2 most important theorems in statistics) contradict each other?
   (c) Asymptotic efficiency. The MLE contains as much information as can be packed into a point estimator.

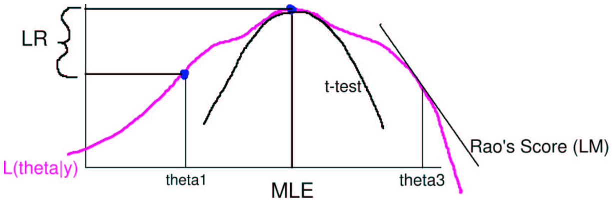
Sampling distributions of the MLE: CLT vs LLN



Uncertainty: Likelihood Ratios for nested models

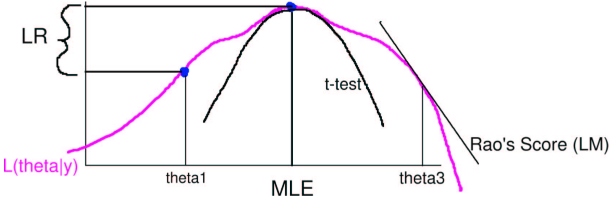


Uncertainty: Likelihood Ratios for nested models



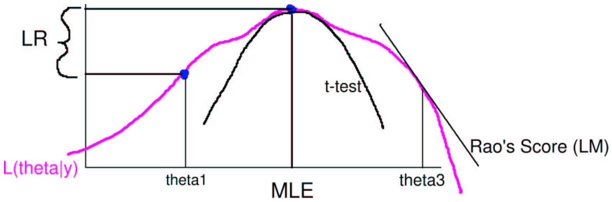
- L^* is the likelihood value for the unrestricted model

Uncertainty: Likelihood Ratios for nested models



- L^* is the likelihood value for the unrestricted model
- L^*_R is the likelihood value for the (nested) restricted model

Uncertainty: Likelihood Ratios for nested models



- L^* is the likelihood value for the unrestricted model
- L^*_R is the likelihood value for the (nested) restricted model
- $\implies L^* \geq L^*_R \implies \frac{L^*_R}{L^*} \leq 1$

Meaning of the likelihood ratio

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

$$L(\theta_1|y) \propto k(y)\mathrm{P}(y|\theta_1)$$

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

$$L(\theta_1|y) \propto k(y)\mathrm{P}(y|\theta_1)$$
$$L(\theta_2|y) \propto k(y)\mathrm{P}(y|\theta_2)$$

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

$$\begin{aligned}
L(\theta_1|y) &\propto k(y)\mathrm{P}(y|\theta_1) \\
L(\theta_2|y) &\propto k(y)\mathrm{P}(y|\theta_2) \\
\frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)}
\end{aligned}$$

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

$$
\begin{aligned}
L(\theta_1|y) &\propto k(y)\mathrm{P}(y|\theta_1)\\
L(\theta_2|y) &\propto k(y)\mathrm{P}(y|\theta_2)\\
\frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)}\\
&= \frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)}
\end{aligned}
$$

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

$$
\begin{aligned}
L(\theta_1|y) &\propto k(y)\mathrm{P}(y|\theta_1) \\
L(\theta_2|y) &\propto k(y)\mathrm{P}(y|\theta_2) \\
\frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)} \\
&= \frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)}
\end{aligned}
$$

Interpreted as a *risk ratio*.

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

$$\begin{aligned} L(\theta_1|y) &\propto k(y)\mathrm{P}(y|\theta_1) \\ L(\theta_2|y) &\propto k(y)\mathrm{P}(y|\theta_2) \\ \frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)} \\ &= \frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)} \end{aligned}$$

   Interpreted as a *risk ratio*.

2. Statistically (from the Neyman-Pearson Hypothesis Testing viewpoint), let

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

$$\begin{aligned} L(\theta_1|y) &\propto k(y)\mathrm{P}(y|\theta_1) \\ L(\theta_2|y) &\propto k(y)\mathrm{P}(y|\theta_2) \\ \frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)} \\ &= \frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)} \end{aligned}$$

   Interpreted as a *risk ratio*.

2. Statistically (from the Neyman-Pearson Hypothesis Testing viewpoint), let

$$R = -2\ln\left(\frac{L_R^*}{L^*}\right) = 2(\ln L^* - \ln L_R^*)$$

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

$$
\begin{aligned}
L(\theta_1|y) &\propto k(y)\mathrm{P}(y|\theta_1) \\
L(\theta_2|y) &\propto k(y)\mathrm{P}(y|\theta_2) \\
\frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)} \\
&= \frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)}
\end{aligned}
$$

   Interpreted as a *risk ratio*.

2. Statistically (from the Neyman-Pearson Hypothesis Testing viewpoint), let

$$
R = -2\ln\left(\frac{L_R^*}{L^*}\right) = 2(\ln L^* - \ln L_R^*)
$$

   Then, under the null of no difference between the 2 models,

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

$$
\begin{aligned}
L(\theta_1|y) &\propto k(y)\mathrm{P}(y|\theta_1) \\
L(\theta_2|y) &\propto k(y)\mathrm{P}(y|\theta_2) \\
\frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)} \frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)} \\
&= \frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)}
\end{aligned}
$$

   Interpreted as a *risk ratio*.

2. Statistically (from the Neyman-Pearson Hypothesis Testing viewpoint), let

$$R = -2\ln\left(\frac{L_R^*}{L^*}\right) = 2(\ln L^* - \ln L_R^*)$$

   Then, under the null of no difference between the 2 models,

$$R \sim f_{\chi^2}(r|m)$$

Meaning of the likelihood ratio

1. Substantively, its the ratio of 2 traditional probabilities:

$$\begin{aligned}
L(\theta_1|y) &\propto k(y)\mathrm{P}(y|\theta_1)\\
L(\theta_2|y) &\propto k(y)\mathrm{P}(y|\theta_2)\\
\frac{L(\theta_1|y)}{L(\theta_2|y)} &= \frac{k(y)}{k(y)}\frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)}\\
&= \frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)}
\end{aligned}$$

   Interpreted as a *risk ratio*.

2. Statistically (from the Neyman-Pearson Hypothesis Testing viewpoint), let

$$R = -2\ln\left(\frac{L_R^*}{L^*}\right) = 2(\ln L^* - \ln L_R^*)$$

   Then, under the null of no difference between the 2 models,

$$R \sim f_{\chi^2}(r|m)$$

   where r is the observed value of R and m is the number of restricted parameters.

Meaning of the likelihood ratio

Meaning of the likelihood ratio

(a) If restrictions have no effect, $E(R) = m$.

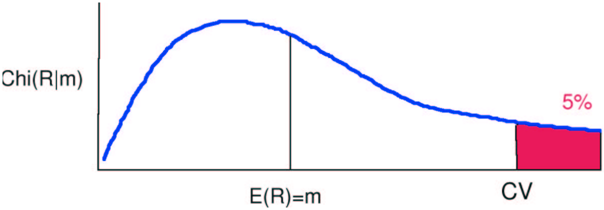
Meaning of the likelihood ratio

(a) If restrictions have no effect, $E(R) = m$.
(b) So only if $r >> m$ will the test parameters be clearly different from zero.

Meaning of the likelihood ratio

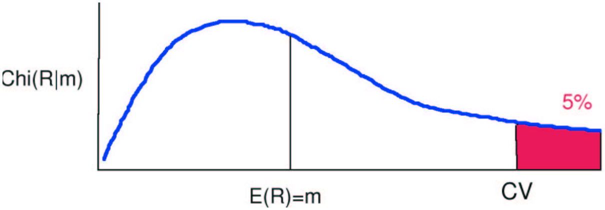
(a) If restrictions have no effect, $E(R) = m$.

(b) So only if $r >> m$ will the test parameters be clearly different from zero.



Meaning of the likelihood ratio

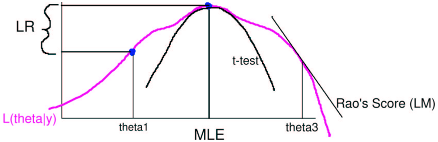
(a) If restrictions have no effect, $E(R) = m$.
(b) So only if $r >> m$ will the test parameters be clearly different from zero.



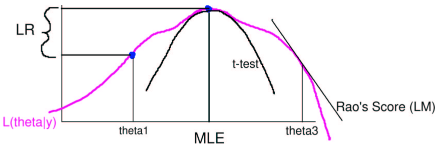
3. Disadvantage: Too many likelihood ratio tests may be required to test all points of interest

Standard Errors

Standard Errors

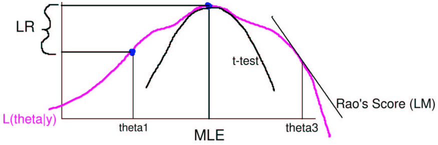


Standard Errors



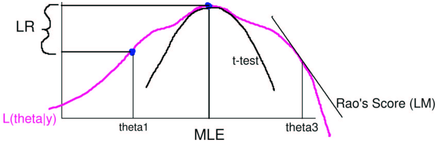
1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize all the curvature near the maximum with one number.

Standard Errors



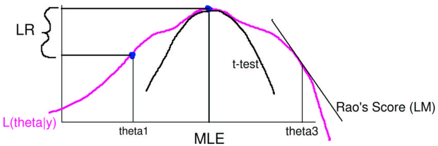
1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize all the curvature near the maximum with one number.
2. We will use the normal likelihood to approximate all likelihoods

Standard Errors



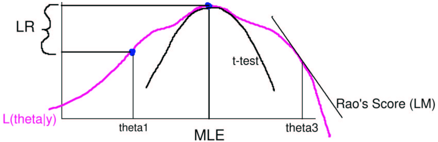
1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize all the curvature near the maximum with one number.
2. We will use the normal likelihood to approximate all likelihoods
3. (one justification) as $n \to \infty$, all likelihoods become normal.

Standard Errors



1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize all the curvature near the maximum with one number.
2. We will use the normal likelihood to approximate all likelihoods
3. (one justification) as $n \to \infty$, all likelihoods become normal.
4. Reformulate the normal (not stylized normal) likelihood.

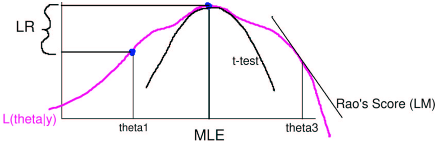
Standard Errors



1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize all the curvature near the maximum with one number.
2. We will use the normal likelihood to approximate all likelihoods
3. (one justification) as $n \to \infty$, all likelihoods become normal.
4. Reformulate the normal (not stylized normal) likelihood.

$$L(\beta|y) \propto N(y_i|\mu_i, \sigma^2)$$

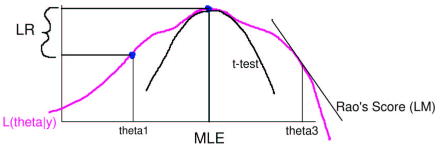
Standard Errors



1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize all the curvature near the maximum with one number.
2. We will use the normal likelihood to approximate all likelihoods
3. (one justification) as $n \rightarrow \infty$, all likelihoods become normal.
4. Reformulate the normal (not stylized normal) likelihood.

$$
\begin{aligned}
L(\beta|y) &\propto N(y_i|\mu_i, \sigma^2) \\
&= (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right)
\end{aligned}
$$

Standard Errors



1. Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we can summarize all the curvature near the maximum with one number.
2. We will use the normal likelihood to approximate all likelihoods
3. (one justification) as $n \to \infty$, all likelihoods become normal.
4. Reformulate the normal (not stylized normal) likelihood.

$$
\begin{aligned}
L(\beta|y) &\propto N(y_i|\mu_i, \sigma^2) \\
&= (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(y_i - \mu_i)^2}{2\sigma^2}\right) \\
&= (2\pi\sigma^2)^{-1/2} \exp\left(\frac{-(y_i - \beta)^2}{2\sigma^2}\right)
\end{aligned}
$$

Justifying Standard Errors

$$\ln L(\beta|y) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta)^2$$

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2)
\end{aligned}
$$

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2
\end{aligned}
$$

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2-2y_i\beta+\beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2
\end{aligned}
$$

A quadratic equation.

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2-2y_i\beta+\beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n}y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2
\end{aligned}
$$

A quadratic equation.

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:

Justifying Standard Errors

$$\ln L(\beta|y) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta)^2$$
$$= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2)$$
$$= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2$$

A quadratic equation.

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   (a) n is large

Justifying Standard Errors

$$\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2-2y_i\beta+\beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n}y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n}y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2
\end{aligned}$$

A quadratic equation.

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   (a) n is large
   (b) σ^2 is small

Justifying Standard Errors

$$\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2
\end{aligned}$$

A quadratic equation.

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   (a) n is large
   (b) σ^2 is small
6. When the log-likelihood is normal, we'll use $\left(\frac{-n}{2\sigma^2}\right)$ as the one-number summary.

Justifying Standard Errors

$$\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2
\end{aligned}$$

A quadratic equation.

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   (a) n is large
   (b) σ^2 is small
6. When the log-likelihood is normal, we'll use $\left(\frac{-n}{2\sigma^2}\right)$ as the one-number summary.
   (a) The bigger the better

Justifying Standard Errors

$$\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2
\end{aligned}$$

A quadratic equation.

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   (a) n is large
   (b) σ^2 is small
6. When the log-likelihood is normal, we'll use $\left(\frac{-n}{2\sigma^2}\right)$ as the one-number summary.
   (a) The bigger the better
   (b) The bigger, the more information exists in the MLE

Justifying Standard Errors

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2-2y_i\beta+\beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2
\end{aligned}
$$

A quadratic equation.

5. $\left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
   (a) n is large
   (b) σ^2 is small
6. When the log-likelihood is normal, we'll use $\left(\frac{-n}{2\sigma^2}\right)$ as the one-number summary.
   (a) The bigger the better
   (b) The bigger, the more information exists in the MLE
   (c) The bigger, the larger the likelihood ratio would be in comparing the MLE with *any* other parameter value.

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

More generally, this second derivative will give us a way to compute the coefficient on the squared term.

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.
8. We invert the curvature to provide a statistical interpretation:

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.
8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.

8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \dots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

   (a) Statistical interpretation: variance and covariance across repeated samples

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.

8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

   (a) Statistical interpretation: variance and covariance across repeated samples
   (b) Works in general for a k-dimensional θ vector

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

More generally, this second derivative will give us a way to compute the coefficient on the squared term.

8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \dots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \dots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

   (a) Statistical interpretation: variance and covariance across repeated samples
   (b) Works in general for a k-dimensional θ vector
   (c) Can be computed numerically

Standard Errors

7. When the log-likelihood is not normal, we'll use the best quadratic approximation to it. Under the normal,

$$\frac{\partial^2 \ln L(\beta|y)}{\partial\beta\partial\beta'} = \frac{-n}{\sigma^2}$$

   More generally, this second derivative will give us a way to compute the coefficient on the squared term.
8. We invert the curvature to provide a statistical interpretation:

$$\hat{V}(\hat{\theta}) = \left[-\frac{\partial^2 \ln L(\theta|y)}{\partial\theta\partial\theta'}\right]^{-1}_{\theta=\hat{\theta}} = \begin{pmatrix} \hat{\sigma}_1^2 & \hat{\sigma}_{12} & \cdots \\ \hat{\sigma}_{21} & \hat{\sigma}_2^2 & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}$$

   (a) Statistical interpretation: variance and covariance across repeated samples
   (b) Works in general for a k-dimensional θ vector
   (c) Can be computed numerically
   (d) Known as the variance matrix, or variance-covariance matrix, or covariance matrix

Standard errors

9. This is an estimate of a quadratic approximation to the log-likelihood.

Standard errors

9. This is an estimate of a quadratic approximation to the log-likelihood.
10. Asymptotically, it is an estimate of the exact log-likelihood.

Simulation for any ML Model

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- As n gets large, the sampling distribution of $\hat{\theta}$ becomes normal.

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- As *n* gets large, the sampling distribution of $\hat{\theta}$ becomes normal.
- As *n* gets large, the log-likelihood becomes quadratic

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- As *n* gets large, the sampling distribution of $\hat{\theta}$ becomes normal.
- As *n* gets large, the log-likelihood becomes quadratic
- As *n* gets large, the quadratic approximation assumed when using the second derivative of the log-likelihood improves

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- As n gets large, the sampling distribution of $\hat{\theta}$ becomes normal.
- As n gets large, the log-likelihood becomes quadratic
- As n gets large, the quadratic approximation assumed when using the second derivative of the log-likelihood improves
- Thus an estimate of the variance of the sampling distribution of $\hat{\theta}$ is $\hat{V}(\hat{\theta})$, which is the inverse of the negative of the matrix of second derivatives, evaluated at $\hat{\theta}$.

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- As n gets large, the sampling distribution of $\hat{\theta}$ becomes normal.
- As n gets large, the log-likelihood becomes quadratic
- As n gets large, the quadratic approximation assumed when using the second derivative of the log-likelihood improves
- Thus an estimate of the variance of the sampling distribution of $\hat{\theta}$ is $\hat{V}(\hat{\theta})$, which is the inverse of the negative of the matrix of second derivatives, evaluated at $\hat{\theta}$.
- To simulate θ, we'll draw from the multivariate normal: $N\left(\hat{\theta}, \hat{V}(\hat{\theta})\right)$

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- As n gets large, the sampling distribution of $\hat{\theta}$ becomes normal.
- As n gets large, the log-likelihood becomes quadratic
- As n gets large, the quadratic approximation assumed when using the second derivative of the log-likelihood improves
- Thus an estimate of the variance of the sampling distribution of $\hat{\theta}$ is $\hat{V}(\hat{\theta})$, which is the inverse of the negative of the matrix of second derivatives, evaluated at $\hat{\theta}$.
- To simulate θ, we'll draw from the multivariate normal: $N\left(\hat{\theta}, \hat{V}(\hat{\theta})\right)$
- This is an asymptotic approximation and can be wrong sometimes.

Simulation for any ML Model

- If the model is correct, a consistent point estimate of θ is the MLE, $\hat{\theta}$.
- As n gets large, the sampling distribution of $\hat{\theta}$ becomes normal.
- As n gets large, the log-likelihood becomes quadratic
- As n gets large, the quadratic approximation assumed when using the second derivative of the log-likelihood improves
- Thus an estimate of the variance of the sampling distribution of $\hat{\theta}$ is $\hat{V}(\hat{\theta})$, which is the inverse of the negative of the matrix of second derivatives, evaluated at $\hat{\theta}$.
- To simulate θ, we'll draw from the multivariate normal: $N\left(\hat{\theta}, \hat{V}(\hat{\theta})\right)$
- This is an asymptotic approximation and can be wrong sometimes.
- We'll discuss later how how to improve the approximation.

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \ldots, 50$

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \dots, 50$
t	election year, for $t = 1948, 1952, \dots, 2008$

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \dots, 50$
t	election year, for $t = 1948, 1952, \dots, 2008$
y_{it}	Democratic fraction of the two-party vote

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \ldots, 50$
t	election year, for $t = 1948, 1952, \ldots, 2008$
y_{it}	Democratic fraction of the two-party vote
X_{it}	a list of covariates (economic conditions, polls, home state, etc)

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \ldots, 50$
t	election year, for $t = 1948, 1952, \ldots, 2008$
y_{it}	Democratic fraction of the two-party vote
X_{it}	a list of covariates (economic conditions, polls, home state, etc)
$X_{i,2012}$	the same covariates as X_{it} but measured in 2012

ML Example: k Parameters, including an Ancillary Parameter, with Simulation to Interpret.

Forecasting Presidential Elections.

The Data

i	U.S. state, for $i = 1, \ldots, 50$
t	election year, for $t = 1948, 1952, \ldots, 2008$
y_{it}	Democratic fraction of the two-party vote
X_{it}	a list of covariates (economic conditions, polls, home state, etc)
$X_{i,2012}$	the same covariates as X_{it} but measured in 2012
E_i	The number of electoral college votes for each state in 2012

The Model

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X.

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X.

The Likelihood Model for the *i*th observation

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X.

The Likelihood Model for the *i*th observation

$$L(\mu_{it}, \sigma | y_{it}) \propto N(y_{it} | \mu_{it}, \sigma^2)$$

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X.

The Likelihood Model for the *i*th observation

$$\begin{aligned} L(\mu_{it}, \sigma|y_{it}) &\propto N(y_{it}|\mu_{it}, \sigma^2) \\ &= (2\pi\sigma^2)^{-1/2} e^{\frac{-(y_{it}-\mu_{it})^2}{2\sigma^2}} \end{aligned}$$

Likelihood model for all n observations

Likelihood model for all n observations

$$L(\beta, \sigma^2|y) = \prod_{i=1}^{n} \prod_{t=1}^{T} L(y_{it}|\mu_{it}, \sigma^2)$$

Likelihood model for all n observations

$$L(\beta, \sigma^2|y) = \prod_{i=1}^{n}\prod_{t=1}^{T} L(y_{it}|\mu_{it}, \sigma^2)$$

$$\ln L(\beta, \sigma^2|y) = \sum_{i=1}^{n}\sum_{t=1}^{T} \ln L(y_{it}|\mu_{it}, \sigma^2)$$

Likelihood model for all n observations

$$L(\beta, \sigma^2|y) = \prod_{i=1}^{n}\prod_{t=1}^{T} L(y_{it}|\mu_{it}, \sigma^2)$$

$$\ln L(\beta, \sigma^2|y) = \sum_{i=1}^{n}\sum_{t=1}^{T} \ln L(y_{it}|\mu_{it}, \sigma^2)$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} \left\{ -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_{it} - \mu_{it})^2}{2\sigma^2} \right\}$$

Likelihood model for all n observations

$$L(\beta, \sigma^2|y) = \prod_{i=1}^{n}\prod_{t=1}^{T} L(y_{it}|\mu_{it}, \sigma^2)$$

$$\begin{aligned}
\ln L(\beta, \sigma^2|y) &= \sum_{i=1}^{n}\sum_{t=1}^{T} \ln L(y_{it}|\mu_{it}, \sigma^2) \\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} \left\{ -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_{it}-\mu_{it})^2}{2\sigma^2} \right\} \\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\ln(2\pi) + \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln\sigma^2 + \frac{(y_{it}-\mu_{it})^2}{\sigma^2}\right]
\end{aligned}$$

Likelihood model for all n observations

$$
\begin{aligned}
L(\beta,\sigma^2|y) &= \prod_{i=1}^{n}\prod_{t=1}^{T} L(y_{it}|\mu_{it},\sigma^2)\\
\ln L(\beta,\sigma^2|y) &= \sum_{i=1}^{n}\sum_{t=1}^{T} \ln L(y_{it}|\mu_{it},\sigma^2)\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T}\left\{-\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_{it}-\mu_{it})^2}{2\sigma^2}\right\}\\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\ln(2\pi) + \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln\sigma^2 + \frac{(y_{it}-\mu_{it})^2}{\sigma^2}\right]\\
&\doteq \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln\sigma^2 + \frac{(y_{it}-\mu_{it})^2}{\sigma^2}\right]
\end{aligned}
$$

Likelihood model for all n observations

$$L(\beta,\sigma^2|y) = \prod_{i=1}^{n}\prod_{t=1}^{T} L(y_{it}|\mu_{it},\sigma^2)$$

$$\ln L(\beta,\sigma^2|y) = \sum_{i=1}^{n}\sum_{t=1}^{T} \ln L(y_{it}|\mu_{it},\sigma^2)$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} \left\{ -\frac{1}{2}\ln(2\pi\sigma^2) - \frac{(y_{it}-\mu_{it})^2}{2\sigma^2} \right\}$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\ln(2\pi) + \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln\sigma^2 + \frac{(y_{it}-\mu_{it})^2}{\sigma^2}\right]$$

$$\doteq \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln\sigma^2 + \frac{(y_{it}-\mu_{it})^2}{\sigma^2}\right]$$

$$= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln\sigma^2 + \frac{(y_{it}-X_{it}\beta)^2}{\sigma^2}\right]$$

Estimation

Estimation

1. Let k be the number of explanatory variables

Estimation

1. Let k be the number of explanatory variables
2. Reparameterize on the unbounded scale, so use $\sigma = e^{\gamma}$

Estimation

1. Let k be the number of explanatory variables
2. Reparameterize on the unbounded scale, so use $\sigma = e^{\gamma}$
3. Let $\theta = \{\beta, \gamma\}$, which is a $k + 2 \times 1$ vector.

Estimation

1. Let k be the number of explanatory variables
2. Reparameterize on the unbounded scale, so use $\sigma = e^{\gamma}$
3. Let $\theta = \{\beta, \gamma\}$, which is a $k + 2 \times 1$ vector.
4. Maximize the likelihood

Estimation

1. Let k be the number of explanatory variables
2. Reparameterize on the unbounded scale, so use $\sigma = e^{\gamma}$
3. Let $\theta = \{\beta, \gamma\}$, which is a $k + 2 \times 1$ vector.
4. Maximize the likelihood
5. Save $\hat{\theta} = \{\hat{\beta}, \hat{\gamma}\}$.

Estimation

1. Let k be the number of explanatory variables
2. Reparameterize on the unbounded scale, so use $\sigma = e^{\gamma}$
3. Let $\theta = \{\beta, \gamma\}$, which is a $k + 2 \times 1$ vector.
4. Maximize the likelihood
5. Save $\hat{\theta} = \{\hat{\beta}, \hat{\gamma}\}$.
6. Compute and save $\hat{V}(\hat{\theta})$, which is $k + 2 \times k + 2$

R Code for the Log-Likelihood

R Code for the Log-Likelihood

- Mathematical Form:

$$\ln L(\beta, \sigma^2|y) = \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2}\right]$$

R Code for the Log-Likelihood

- Mathematical Form:

$$\ln L(\beta, \sigma^2 | y) = \sum_{i=1}^{n} \sum_{t=1}^{T} -\frac{1}{2} \left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2} \right]$$

- The function:

```
ll.normal <- function(par, X, Y) {
beta <- par[1:ncol(X)]
sigma2 <- exp(par[ncol(X) + 1])
-1/2 * sum( log(sigma2) + ((Y - X %*% beta)^2)/sigma2 )
}
```

R Code for the Log-Likelihood

- Mathematical Form:

$$\ln L(\beta, \sigma^2|y) = \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2}\right]$$

- The function:

```
ll.normal <- function(par, X, Y) {
beta <- par[1:ncol(X)]
sigma2 <- exp(par[ncol(X) + 1])
-1/2 * sum( log(sigma2) + ((Y - X %*% beta)^2)/sigma2 )
}
```

- Calling it:

```
ll.normal(c(2,1,2,1,33,4,3.2),x,y)
ll.normal(c(2,1,2,1,33,4,3.7),x,y)
ll.normal(c(2,1,2,1,33,4,3.5),x,y)
```

Quantities of Interest

Quantities of Interest

1. Reasons we care about the values of the regression coefficients:

Quantities of Interest

1. Reasons we care about the values of the regression coefficients: N

Quantities of Interest

1. Reasons we care about the values of the regression coefficients: No

Quantities of Interest

1. Reasons we care about the values of the regression coefficients: Non

Quantities of Interest

1. Reasons we care about the values of the regression coefficients: None

Quantities of Interest

1. Reasons we care about the values of the regression coefficients: None
2. The posterior distribution of electoral college delegates for the Democrat.

Quantities of Interest

1. Reasons we care about the values of the regression coefficients: None
2. The posterior distribution of electoral college delegates for the Democrat.
3. Expected number of electoral college delegates for the Democrat.

Quantities of Interest

1. Reasons we care about the values of the regression coefficients: None
2. The posterior distribution of electoral college delegates for the Democrat.
3. Expected number of electoral college delegates for the Democrat.
4. Probability that the Democratic candidate gets more than $\sum_{i=1}^{n} E_i/n > 0.5$ proportion of electoral college delegates.

Computing the predictive distribution of electoral college delegates in 2012

Computing the predictive distribution of electoral college delegates in 2012

1. Draw many simulations of $y_{i,2012}$ ($\tilde{y}_{i,2012}$) from its posterior distribution for U.S. state i, $\text{P}(y_{i,2012}|y_{it}, X_{it})$, i.e. $\text{P}(\texttt{quantity of interest}|\texttt{data})$. (Details shortly.)

Computing the predictive distribution of electoral college delegates in 2012

1. Draw many simulations of $y_{i,2012}$ ($\tilde{y}_{i,2012}$) from its posterior distribution for U.S. state i, $\text{P}(y_{i,2012}|y_{it}, X_{it})$, i.e. $\text{P}(\texttt{quantity of interest}|\texttt{data})$. (Details shortly.)
2. For each simulation of state i, if $y_{i,2012} > 0.5$ the Democrat "wins" E_i electoral college delegates; if its not, he gets zero.

Computing the predictive distribution of electoral college delegates in 2012

1. Draw many simulations of $y_{i,2012}$ ($\tilde{y}_{i,2012}$) from its posterior distribution for U.S. state i, $\text{P}(y_{i,2012}|y_{it}, X_{it})$, i.e. $\text{P}(\texttt{quantity of interest}|\texttt{data})$. (Details shortly.)
2. For each simulation of state i, if $y_{i,2012} > 0.5$ the Democrat "wins" E_i electoral college delegates; if its not, he gets zero.
3. Add the number of electoral college delegates the Democrat wins in the entire country by adding simulated winnings from each state.

Computing the predictive distribution of electoral college delegates in 2012

1. Draw many simulations of $y_{i,2012}$ ($\tilde{y}_{i,2012}$) from its posterior distribution for U.S. state i, $\text{P}(y_{i,2012}|y_{it}, X_{it})$, i.e. $\text{P}(\texttt{quantity of interest}|\texttt{data})$. (Details shortly.)
2. For each simulation of state i, if $y_{i,2012} > 0.5$ the Democrat "wins" E_i electoral college delegates; if its not, he gets zero.
3. Add the number of electoral college delegates the Democrat wins in the entire country by adding simulated winnings from each state.
4. Repeat Steps 1–3 $M = 1,000$ times, and plot a histogram of the results.

Computing the predictive distribution of electoral college delegates in 2012

1. Draw many simulations of $y_{i,2012}$ ($\tilde{y}_{i,2012}$) from its posterior distribution for U.S. state i, $\text{P}(y_{i,2012}|y_{it}, X_{it})$, i.e. $\text{P}(\texttt{quantity of interest}|\texttt{data})$. (Details shortly.)
2. For each simulation of state i, if $y_{i,2012} > 0.5$ the Democrat "wins" E_i electoral college delegates; if its not, he gets zero.
3. Add the number of electoral college delegates the Democrat wins in the entire country by adding simulated winnings from each state.
4. Repeat Steps 1–3 $M = 1,000$ times, and plot a histogram of the results.
5. (How does this differ from allocating all E_i based on the point estimate?)

How to draw simulations of $y_{i,2012}$?

How to draw simulations of $y_{i,2012}$?

1. Choose values of explanatory variables. In this case, $X_{i,2012}$

How to draw simulations of $y_{i,2012}$?

1. Choose values of explanatory variables. In this case, $X_{i,2012}$
2. Simulate estimation uncertainty:

How to draw simulations of $y_{i,2012}$?

1. Choose values of explanatory variables. In this case, $X_{i,2012}$
2. Simulate estimation uncertainty:
   (a) Draw θ randomly from its posterior (i.e., sampling) distribution, $N(\theta|\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\hat{\beta}, \hat{\gamma}\}$.

How to draw simulations of $y_{i,2012}$?

1. Choose values of explanatory variables. In this case, $X_{i,2012}$
2. Simulate estimation uncertainty:
   (a) Draw θ randomly from its posterior (i.e., sampling) distribution, $N(\theta|\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\hat{\beta}, \hat{\gamma}\}$.
   (b) Pull out $\tilde{\beta}$ and save.

How to draw simulations of $y_{i,2012}$?

1. Choose values of explanatory variables. In this case, $X_{i,2012}$
2. Simulate estimation uncertainty:
   (a) Draw θ randomly from its posterior (i.e., sampling) distribution, $N(\theta|\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\hat{\beta}, \hat{\gamma}\}$.
   (b) Pull out $\tilde{\beta}$ and save.
   (c) Pull out, un-reparameterize, and save $\tilde{\sigma} = e^{\tilde{\gamma}}$

How to draw simulations of $y_{i,2012}$?

1. Choose values of explanatory variables. In this case, $X_{i,2012}$
2. Simulate estimation uncertainty:
   (a) Draw θ randomly from its posterior (i.e., sampling) distribution, $N(\theta|\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\hat{\beta}, \hat{\gamma}\}$.
   (b) Pull out $\tilde{\beta}$ and save.
   (c) Pull out, un-reparameterize, and save $\tilde{\sigma} = e^{\tilde{\gamma}}$
3. Compute the simulated systematic component:

How to draw simulations of $y_{i,2012}$?

1. Choose values of explanatory variables. In this case, $X_{i,2012}$
2. Simulate estimation uncertainty:
   (a) Draw θ randomly from its posterior (i.e., sampling) distribution, $N(\theta|\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\hat{\beta}, \hat{\gamma}\}$.
   (b) Pull out $\tilde{\beta}$ and save.
   (c) Pull out, un-reparameterize, and save $\tilde{\sigma} = e^{\tilde{\gamma}}$
3. Compute the simulated systematic component: $\tilde{\mu}_{it} = X_{i,2012}\tilde{\beta}$

How to draw simulations of $y_{i,2012}$?

1. Choose values of explanatory variables. In this case, $X_{i,2012}$
2. Simulate estimation uncertainty:
   (a) Draw θ randomly from its posterior (i.e., sampling) distribution, $N(\theta|\hat{\theta}, \hat{V}(\hat{\theta}))$. Label the random draw $\tilde{\theta} = \{\hat{\beta}, \hat{\gamma}\}$.
   (b) Pull out $\tilde{\beta}$ and save.
   (c) Pull out, un-reparameterize, and save $\tilde{\sigma} = e^{\tilde{\gamma}}$
3. Compute the simulated systematic component: $\tilde{\mu}_{it} = X_{i,2012}\tilde{\beta}$
4. Add fundamental uncertainty: draw $y_{i,2012}$ from $N(y_{i,2012}|\tilde{\mu}_{i,2012}, \tilde{\sigma}^2)$

How to do it with a LS Regression Program

How to do it with a LS Regression Program

1. Run LS regression of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$

How to do it with a LS Regression Program

1. Run LS regression of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$
2. Draw β randomly from its posterior distribution (i.e., its sampling distribution), $N(\beta|\hat{\beta}, V(\hat{\beta}))$. Label the random draw $\tilde{\beta}$.

How to do it with a LS Regression Program

1. Run LS regression of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$
2. Draw β randomly from its posterior distribution (i.e., its sampling distribution), $N(\beta|\hat{\beta}, V(\hat{\beta}))$. Label the random draw $\tilde{\beta}$.
3. Draw σ^2 from its posterior (or sampling) distribution, $1/\chi^2(\sigma^2|\hat{\sigma}^2, N-k)$, labeling it $\tilde{\sigma}^2$

How to do it with a LS Regression Program

1. Run LS regression of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$
2. Draw β randomly from its posterior distribution (i.e., its sampling distribution), $N(\beta|\hat{\beta}, V(\hat{\beta}))$. Label the random draw $\tilde{\beta}$.
3. Draw σ^2 from its posterior (or sampling) distribution, $1/\chi^2(\sigma^2|\hat{\sigma}^2, N-k)$, labeling it $\tilde{\sigma}^2$
4. Either:

How to do it with a LS Regression Program

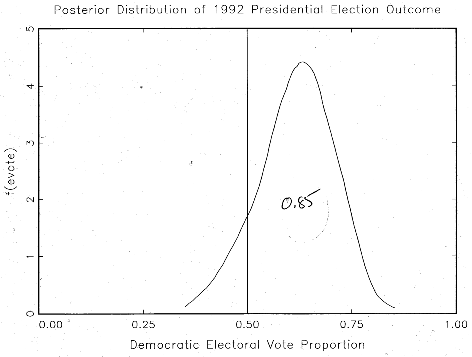
1. Run LS regression of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$
2. Draw β randomly from its posterior distribution (i.e., its sampling distribution), $N(\beta|\hat{\beta}, V(\hat{\beta}))$. Label the random draw $\tilde{\beta}$.
3. Draw σ^2 from its posterior (or sampling) distribution, $1/\chi^2(\sigma^2|\hat{\sigma}^2, N-k)$, labeling it $\tilde{\sigma}^2$
4. Either:
   - Draw ϵ_{it} from $N(\epsilon_{it}|0, \tilde{\sigma}^2)$, label it $\tilde{\epsilon}_{it}$ and compute: $\tilde{y}_{i,2012} = \tilde{X}_{i,2012}\tilde{\beta} + \tilde{\epsilon}_{it}$

How to do it with a LS Regression Program

1. Run LS regression of y_{it} on X_{it} and get $\hat{\beta}$ and $V(\hat{\beta})$
2. Draw β randomly from its posterior distribution (i.e., its sampling distribution), $N(\beta|\hat{\beta}, V(\hat{\beta}))$. Label the random draw $\tilde{\beta}$.
3. Draw σ^2 from its posterior (or sampling) distribution, $1/\chi^2(\sigma^2|\hat{\sigma}^2, N-k)$, labeling it $\tilde{\sigma}^2$
4. Either:
   - Draw ϵ_{it} from $N(\epsilon_{it}|0, \tilde{\sigma}^2)$, label it $\tilde{\epsilon}_{it}$ and compute: $\tilde{y}_{i,2012} = \tilde{X}_{i,2012}\tilde{\beta} + \tilde{\epsilon}_{it}$
   - Or, in our preferred notation, draw $\tilde{y}_{i,2012}$ from $N(X_{i,2012}\tilde{\beta}, \tilde{\sigma}^2)$

Actual Results (calculated before the election) for 1992

Actual Results (calculated before the election) for 1992



Variance Function Models

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}
4. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X and Z.

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}
4. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X and Z.

The log-likelihood:

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}
4. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X and Z.

The log-likelihood:

$$\ln L(\beta, \sigma^2|y) = \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2}\right]$$

Variance Function Models

1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}
4. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X and Z.

The log-likelihood:

$$
\begin{aligned}
\ln L(\beta, \sigma^2|y) &= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2}\right] \\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[z_{it}\gamma + \frac{(y_{it} - X_{it}\beta)^2}{\exp(z_{it}\gamma)}\right]
\end{aligned}
$$

Variance Function Models

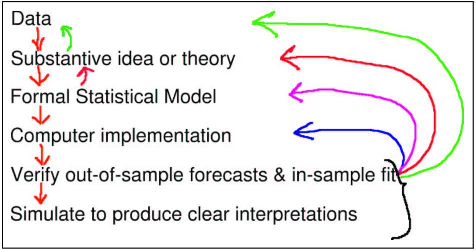
1. $Y_{it} \sim N(y_{it}|\mu_{it}, \sigma^2_{it})$
2. $\mu_{it} = x_{it}\beta$, where x_{it} is a vector of explanatory variables and a constant
3. $\sigma^2_{it} = \exp(z_{it}\gamma)$, where z_{it} is a vector of explanatory variables possibly overlapping x_{it}
4. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$, conditional on X and Z.

The log-likelihood:

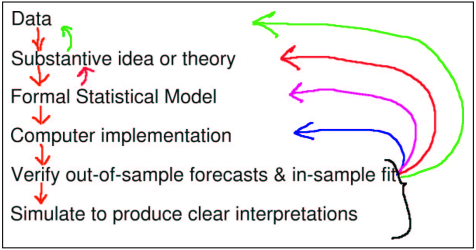
$$\begin{aligned}
\ln L(\beta, \sigma^2|y) &= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[\ln \sigma^2 + \frac{(y_{it} - X_{it}\beta)^2}{\sigma^2}\right] \\
&= \sum_{i=1}^{n}\sum_{t=1}^{T} -\frac{1}{2}\left[z_{it}\gamma + \frac{(y_{it} - X_{it}\beta)^2}{\exp(z_{it}\gamma)}\right]
\end{aligned}$$

- For what applications would this model be informative?

An Outline of the Research Process

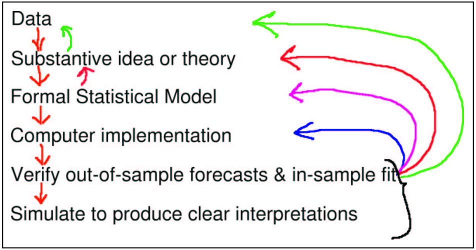


An Outline of the Research Process



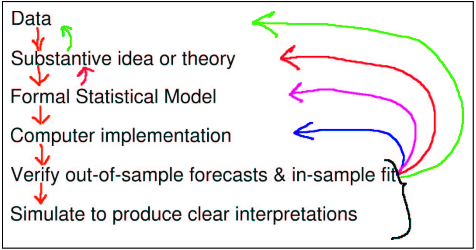
1. These figures are always wild simplifications.

An Outline of the Research Process



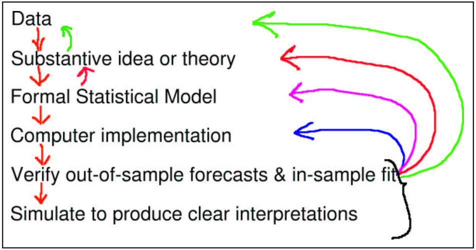
1. These figures are always wild simplifications.
2. Items are roughly in order.

An Outline of the Research Process



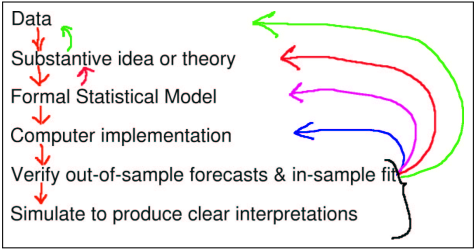
1. These figures are always wild simplifications.
2. Items are roughly in order.
3. You can start at any point.

An Outline of the Research Process



1. These figures are always wild simplifications.
2. Items are roughly in order.
3. You can start at any point.
4. A formal theory is often a useful addition, but not always necessary.

An Outline of the Research Process



1. These figures are always wild simplifications.
2. Items are roughly in order.
3. You can start at any point.
4. A formal theory is often a useful addition, but not always necessary.
5. Don't miss any parts.