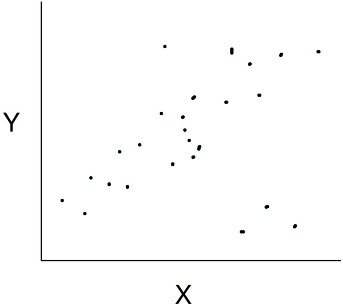
Quantitative Social Science Methods, I, Lecture Notes: Inference

Gary King[1]
Institute for Quantitative Social Science
Harvard University

August 17, 2020

[1] GaryKing.org

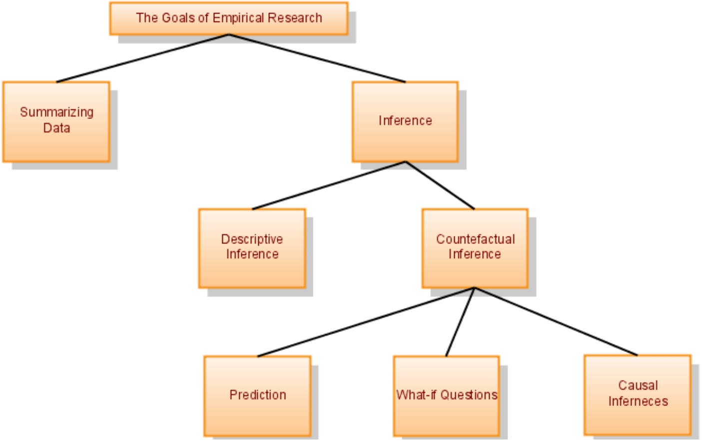
How to fit a line to a scatterplot?



- some "rule": Least squares? Least absolute deviations?
- visually, by hand (tends to be principal components)
- a statistical criterion: (unbiasedness, efficiency, consistency, etc.)
- a full theory of inference, and for a specific purpose (like causal estimation, prediction, etc.)
- (It's a pretty dumb question, don't you think?)

Quantities of Interest

- Summarizing data: functions of facts you have
- Inference: using facts you know to learn facts you don't know



The Problem of Inference

- Probability:

$$P(y \mid M) = P(\text{known} \mid \text{unknown})$$

- The goal of inverse probability:

$$P(M \mid y) = P(\text{unknown} \mid \text{known})$$

- A more reasonable, limited goal. Let $M = \{M^*, \theta\}$, where M^* is assumed & θ is to be estimated:

$$P(\theta \mid y, M^*) \equiv P(\theta|y)$$

Building Theories of Inference

Everything on this page is true; no assumptions

- Bayes Theorem (as distinct from Bayesian inference):

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

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

Interpretation 1: The Likelihood Theory of Inference

- R.A. Fisher's idea
- θ is fixed and y is random
- Let:

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

- Define $k(y)$ as an unknown function of y with θ fixed at its true value
- The likelihood theory of inference has four axioms: the 3 probability axioms plus the likelihood axiom (neither true nor false):

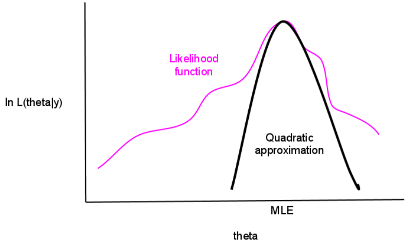
$$\begin{aligned} L(\theta|y) &\equiv k(y)P(y|\theta) \\ &\propto P(y|\theta) \end{aligned}$$

Interpretation 1: The Likelihood Theory of Inference

- $L(\theta|y)$ is a function: for y fixed at the observed values, it gives the "likelihood" of any value of θ you might want to try
- Likelihood: a relative measure of uncertainty, changing with the data
- Comparing value of $L(\theta|y)$ for different θ values:
    - within a data set: meaningful
    - across data sets: meaningless
    - You also can't compare R^2 values across equations with different dependent variables
- The likelihood principle: the data y only affect inferences through the likelihood function, $L(\theta|y) = k(y)P(y|\theta)$

Visualizing the Likelihood

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



- Summary Estimator: The likelihood curve. (Likelihood principle: we can now discard the data—if the model is correct!)
- One-point summary: at the maximum is the "MLE"
- Uncertainty of the MLE: curvature at the maximum

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

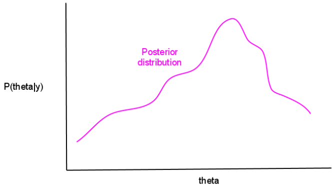
What is the prior density, $P(\theta)$?

- A probability density that represents all prior evidence about θ
- An opportunity: a way of getting other information outside the data set into the model and estimator
- An annoyance: the "other information" is required
- A philosophical assumption that nonsample information:
  - should matter — as it always does
  - should be formalized and included in all inferences — which is more debatable
- Quiz: Example of nonsample information you want included
- Quiz 2: Example of nonsample information you're skeptical of including

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:
$$P(A \mid B) = \frac{P(A, B)}{P(B)}$$
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, $\mathrm{P}(\theta|y)$



- 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 relative, just like likelihood.
- Bayesian inference obeys the likelihood principle: the data set only affects inferences through the likelihood function
- If $\mathrm{P}(\theta) = 1$, i.e., is uniform in the relevant region, then $L(\theta|y) = \mathrm{P}(\theta|y)$.

How to think about Bayes v. Likelihood

- Summary:
  - Likelihood is simpler; start there
  - Bayes opens up more possibilities; use if needed
- Philosophical differences from likelihood: Huge
- Practical differences: Minor, unless the prior matters
- Example where prior matters: demographic forecasting model
- Bayesians are happier people: If $P(\theta)$ is *diffuse*, differences from likelihood are minor, but numerical stability (and "identification") improves $\rightsquigarrow$ your programs will run better!
- Advantages of Bayes: more information $\rightsquigarrow$ more efficiency; MCMC algorithms are easier
- Few fights now between Bayesians and likelihoodists

A 3rd Theory: Neyman-Pearson Hypothesis Testing

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

For example, is $\beta = 0$ in $E(Y) = \beta_0 + \beta X$?

- H_0: $\beta = 0$ vs. H_1: $\beta > 0$
- Choose Type I error, probability of deciding H_1 is right when H_0 is really true: say $\alpha = 0.05$
- (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.)
- Assume n is large enough for the CLT to kick in
- Then $b|(\beta = 0) \sim N(0, \sigma_b^2)$
- 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

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

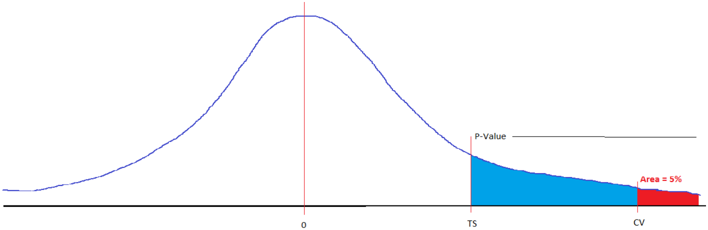
$$\int_{(CV)}^{\infty} N(b|0, \sigma_b^2)\,db = \alpha$$

- In educational psychology and some other fields: write your prospectus, plan your experiment, report the CV, and write the concluding chapter:

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

- Then collect your data. You may not revise your hypothesis or chapter
- Once discredited; making a comeback through the preregistration movement

Neyman-Pearson Hypothesis Testing



- In this example, $(TS) < (CV) \rightsquigarrow$ conclude $\beta = 0$.
- Decision will be wrong 5% of the time
- Quiz: What is the probability it's right this time?
- Quiz 2: What happens when n is large (or under your control)?
- Relaxed approach, use p-values: The probability under the null of getting a value as or more extreme than the value we got — the area to the right of the realized value of (TS).
- Star-gazing is often silly; where's the QOI?
- $\rightsquigarrow$ Can use likelihood: to compute tests and p-values.

What's the best 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
4. Methods for applied researchers: either useful or irrelevant

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
    - 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

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* y, $\Pr(\text{data}|\text{model})$

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

reparameterizing with $\mu_i = \beta$:

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

Quiz: What can you do with this probability density?

Derive the Log-Likelihood

The likelihood of β having generated the data we observe:

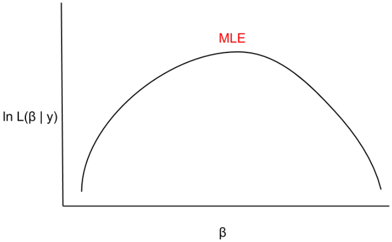
$$L(\beta|y) = k(y)\prod_{i=1}^{n} f_{\text{stn}}(y_i|\beta)$$
$$= k(y)\prod_{i=1}^{n}(2\pi)^{-1/2}\exp\left(\frac{-(y_i-\beta)^2}{2}\right)$$

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

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

Quiz: What subs for β to make $\ln L$ the largest? What's that called?

Log-likelihood interpretation



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

Summarizing k-dimensional space

- Graphs
- The problem of Flatland
- The curse of dimensionality
- We'll often use:
  - $\hat{\beta}$, a vector of point estimates, the MLE
  - Curvature at the maximum (standard errors, about which more shortly)

How to find the maximum?

Goal: Find the value of $\theta \equiv \{\theta_1, \ldots, \theta_k\}$ that maximizes $L(\theta|y)$

1. Analytically — sometimes possible
   - Take derivative of $\ln L(\theta|y)$ w.r.t. θ
   - Set to 0, substituting $\hat{\theta}$ for θ

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

   - If possible, solve for θ, and label it $\hat{\theta}$
   - Check second order condition: make sure 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
   - We'll show you how
   - (Sound good?)

Finite Sample Properties of the MLE

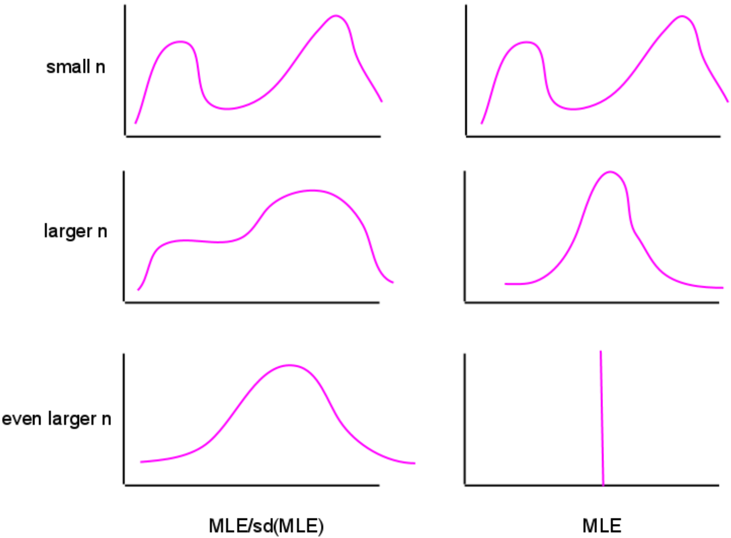
1. Minimum variance unbiased estimator (MVUE)
   - Unbiasedness:
     - Definition: $E(\hat{\theta}) = \theta$
     - Example: $E\left(\bar{Y}\right) = E\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(Y_i) = \frac{1}{n} n\mu = \mu$
   - Minimum variance ("efficiency")
     - Variance to be minimized: $V(\hat{\theta})$
     - Example: $V\left(\bar{Y}\right) = V\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right) = \frac{1}{n^2}\sum_{i=1}^{n} V(Y_i) = \frac{1}{n^2} n\sigma^2 = \sigma^2/n$
     - Efficiency: Define $\hat{\theta}$ to minimize $V(\hat{\theta})$, s.t. $E(\hat{\theta}) = \theta$
   - If there is a MVUE, ML will find it
   - If there isn't one, ML will still usually find a good estimator
2. Invariance to Reparameterization
   - Both are MLEs: Estimate σ^2 with $\hat{\sigma}^2$
     or estimate σ with $\hat{\sigma}$ and calculate $\hat{\sigma}^2$
   - Not true for other methods of inference: e.g. $E(\bar{y}) = \mu$. What is an unbiased estimate of $1/\mu$? Is it $1/\bar{y}$? Nope: $E(1/\bar{y}) \neq 1/E(\bar{y})$
3. Invariance to sampling plans
   - OK to look at results while deciding how much data to collect
   - In fact, it's a great idea! (e.g., King, Schneer, White 2017)

Asymptotic Properties of the MLE

1. Consistency (from the Law of Large Numbers).
   - As $n \longrightarrow \infty$, the sampling distribution of the MLE collapses to a spike over the parameter value
   - Why do we care? An approximation to: more data helps
2. Asymptotic normality (from the Central Limit Theorem):
   - As $n \longrightarrow \infty$, repeated samples of MLE/se(MLE) converge to Normal
   - Why do we care? If N is large enough, the asymptotic distribution is a good approximation
   - Quiz: Do the LLN and CLT (the 2 most important theorems in statistics) contradict each other?
3. Asymptotic efficiency
   - As $n \longrightarrow \infty$, MLE contains as much information as can be packed into a point estimator; it is the MVUE
   - Why do we care? If n is large enough, we're not wasting data

Sampling distributions of the MLE: CLT vs LLN

⇝ Why asymptotic approximations may work in small samples



Quiz: Which is Unbiased & Inconsistent

$$a_1 = \frac{1}{n}\sum_{i=1}^{n} Y_i + 15$$ biased, inconsistent

$$a_2 = \frac{1}{27}\sum_{i=1}^{27} Y_i$$ unbiased, inconsistent

$$a_3 = \frac{1}{n}\sum_{i=1}^{n} Y_i + \sum_{i=1}^{7} Y_i/n$$ biased, consistent

$$a_4 = \frac{1}{n-2}\sum_{i=1}^{n-2} Y_i$$ unbiased, consistent (inefficient)

$$a_5 = \frac{1}{n}\sum_{i=1}^{n} Y_i$$ unbiased, consistent, efficient

The Impossibility of Inference Without Assumptions

Three Theories of Inference: Overview

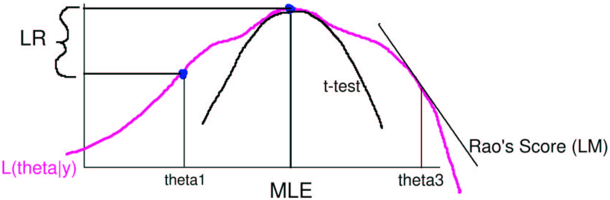
Likelihood: Example, Derivation, Properties

Uncertainty in Likelihood Inference

Simulation from Likelihood Models

Extending the Linear Model with a Variance Function

Three Measures of Uncertainty



- Relative heights at different parameter values: Likelihood Ratio
- Curvature at maximum: Standard Errors
- Slope at single parameter value: Rao's Score (LM)

Uncertainty via the Likelihood Ratio

- Compare two likelihood models
  - unrestricted model: L^*
  - restricted (nested) model: L^*_R
  - Likelihood Ratio:

$$L^* \geq L^*_R \implies \frac{L^*_R}{L^*} \leq 1$$

- Likelihood ratio: the ratio of 2 traditional probabilities

$$L^*_R \equiv L(\theta_1|y) \propto k(y)\mathrm{P}(y|\theta_1)$$
$$L^* \equiv L(\theta_2|y) \propto k(y)\mathrm{P}(y|\theta_2)$$

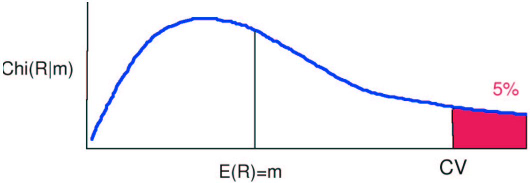
$$\frac{L(\theta_1|y)}{L(\theta_2|y)} = \frac{k(y)}{k(y)}\frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)} = \frac{\mathrm{P}(y|\theta_1)}{\mathrm{P}(y|\theta_2)}, \quad \text{a risk ratio}$$

Likelihood Ratio: Statistical Interpretation

Neyman-Pearson hypothesis testing (under the null):

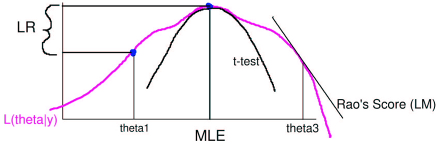
$$R = -2\ln\left(\frac{L_R^*}{L^*}\right) = 2(\ln L^* - \ln L_R^*) \sim f_{\chi^2}(r|m)$$

r is realized value of R; m is number of restricted parameters



- If restrictions have no effect: $E(R) = m$.
- Parameters are different from zero if: $r \gg m$
- Works well, but: Lots of likelihood ratio tests may be required to test all points of interest

Uncertainty via Standard Errors



- Instead of (a) plotting the entire likelihood hyper-surface or (b) computing numerous likelihood ratio tests, we summarize the likelihood curvature near the maximum with one number
- We use the normal likelihood to approximate all likelihoods
- (one justification: as $n \longrightarrow \infty$, likelihoods become normal)
- Reformulate the normal (not stylized) likelihood with $E(Y) = \mu_i = \beta$:

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

(Continued) Standard Errors, Linear Normal Model

$$
\begin{aligned}
\ln L(\beta|y) &= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \beta)^2 \\
&= -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i^2 - 2y_i\beta + \beta^2) \\
&= \left(-\frac{n}{2}\ln(2\pi\sigma^2) - \frac{\sum_{i=1}^{n} y_i^2}{2\sigma^2}\right) + \left(\frac{\sum_{i=1}^{n} y_i}{\sigma^2}\right)\beta + \left(\frac{-n}{2\sigma^2}\right)\beta^2 \\
&= a + b\beta + c\beta^2, \qquad \text{A quadratic equation}
\end{aligned}
$$

- $c = \left(\frac{-n}{2\sigma^2}\right)$ is the degree of curvature. Curvature is larger when:
  - n is large
  - σ^2 is small
- For normal likelihood, $\left(\frac{-n}{2\sigma^2}\right)$ is a summary. The bigger the (negative) number...
  - the better
  - the more information exists in the MLE
  - the larger the likelihood ratio would be in comparing the MLE with *any* other parameter value.

Standard Errors: Any Likelihood Model

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

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

  Second derivative: coefficient c on squared term for any model
- We invert the curvature to provide a statistical interpretation:

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

  - The variance (aka covar or var-covar) across repeated samples
  - Quiz: How do we interpret $\hat{\sigma}_1$? What about $\hat{\sigma}_{21}$?
  - Works in general for a k-dimensional θ vector
  - Can be computed numerically
- An estimate of a quadratic approximation to the log-likelihood
- Asymptotically, it is an estimate of the exact log-likelihood

Parameter Simulation for *any* ML Model

- Assume model is correct (we'll come back to this!)
- Write down likelihood, calculate the MLE: $\hat{\theta}$
- Properties of $\hat{\theta}$ as n gets large:
  - Distribution of $\hat{\theta}$ collapses to spike over θ (LLN $\rightsquigarrow$ consistency)
  - The standardized sampling distribution of $\hat{\theta}$ becomes normal (CLT $\rightsquigarrow$ asymptotic normality)
  - Quadratic approximation to the log-likelihood (from the second derivative) improves
- True variance of sampling distribution of $\hat{\theta}$: $V(\hat{\theta})$
- Estimate of $V(\hat{\theta})$: $\hat{V}(\hat{\theta})$, the inverse of the negative of the matrix of second derivatives of $\ln L(\theta|y)$, evaluated at $\hat{\theta}$.
- To simulate θ,
  - Draw θ from the multivariate normal: $\tilde{\theta} \sim N\left(\hat{\theta}, \hat{V}(\hat{\theta})\right)$
  - This is an asymptotic approximation and can be wrong sometimes (we'll ignore now, improve later)
- Quiz: What's the QOI? Is it θ?

QOI Simulation from *any* ML Model

Overview here; Application to Linear Models Next; Then any QOI

- Recall Generalized ML Model:

$$\begin{aligned} Y_i &\sim f(\theta_i, \alpha) && \text{stochastic} \\ \theta_i &= g(x_i, \beta) && \text{systematic} \end{aligned}$$

- Choose values of X: X_c
- Estimate: MLE $\hat{\gamma} = \{\hat{\beta}, \hat{\alpha}\}$ and its variance $\hat{V}(\hat{\gamma})$
- Simulate estimation uncertainty: $\tilde{\gamma} \sim N\left[\hat{\gamma}, \hat{V}(\hat{\gamma})\right]$
- Calculate (often expected value of y): $\tilde{\theta}_c = g(X_c, \tilde{\beta})$
- Simulate fundamental uncertainty: $\tilde{y}_c \sim f(\tilde{\theta}_c, \tilde{\alpha})$
- Calculate QOI: Calculate histogram, mean, variance, etc. of $\tilde{y}_c$

Example: Forecasting Presidential Elections

The Data

i	U.S. state, for $i = 1, \ldots, 50$
t	election year, for $t = 1948, 1952, \ldots, 2016$
y_{it}	Democratic proportion of the two-party vote
X_{it}	Constant, economics, polls, home state, ideology, etc.
$X_{i,2020}$	the same covariates as X_{it} but measured in 2020
C_i	The number of electoral College delegates in i in 2020

The Model

1. $Y_{it} \sim N(\mu_{it}, \sigma^2)$.
2. $\mu_{it} = x_{it}\beta$, where x_{it} includes a constant
3. Y_{it} and $Y_{i't'}$ are independent $\forall\ i \neq i'$ and $t \neq t'$ (given X)

Quiz: What are this model's weaknesses?

The Likelihood Model

- Likelihood for observation *it*

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

- Likelihood for all n observations

$$\begin{aligned} L(\beta, \sigma^2 | y) &= \prod_{i=1}^{n} \prod_{t=1}^{T} L(\mu_{it}, \sigma^2 | y_{it}) \\ &= \prod_{i=1}^{n} \prod_{t=1}^{T} (2\pi\sigma^2)^{-1/2} e^{\frac{-(y_{it}-\mu_{it})^2}{2\sigma^2}} \end{aligned}$$

Log-Likelihood

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

Estimation

- Reparameterize to unbounded scale
  - numerical optimizers work better this way
  - the CLT kicks in faster
  - β is already unbounded
  - $\sigma > 0 \rightsquigarrow$ transform with $\sigma = e^{\eta}$, and estimate η
- Stack: $\gamma = \{\beta, \eta\}$, a $k + 2 \times 1$ vector (k: number of covariates)
- Turn log-likelihood into code; maximize so we can get:
  - Point estimates: save the MLE, $\hat{\gamma} = \{\hat{\beta}, \hat{\eta}\}$
  - Uncertainty estimates: $\hat{V}(\hat{\gamma})$, which is $k + 2 \times k + 2$

R Code for the Log-Likelihood

- (Recall) mathematical Form:

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

- An R function:

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

- Calling it:

```
loglik(c(2,1,2,1,33,4,2),x,y)
loglik(c(2,1,2,1,33,4,7),x,y)
loglik(c(2,1,2,1,33,4,5),x,y)
```

Quantities of Interest in this election data set

- Quiz: What are the QOIs?
- There's no right answer; here's mine:
  - (Reasons we care about the regression coefficients: None)
  - Predictive distribution of Dem electoral college delegates
  - Expected number of Dem electoral college delegates
  - Probability that Dem candidate is elected: gets more than $\sum_{i=1}^{n} C_i/n > 0.5$ proportion of electoral college delegates

Predicting Allocations of Electoral College Delegates

- Quiz: how to simulate predictions of C_i in state i?
- Options:
  1. if $\hat{y}_{i,2020} > 0.5$, Dems get all C_i; otherwise, Reps get all C_i
     - Quiz: What's your prediction if $\hat{y}_{i,2020} = 0.51\ \forall\ i$?
     - Problem: ignores fundamental uncertainty
  2. Allocate $C_i\hat{y}_{i,2020}$ to Dems; $C_i(1 - \hat{y}_{i,2020})$ to Reps
     - Quiz: What happens if $\hat{y}_{i,2020}$ is uncertain?
     - Problem: Ignores estimation uncertainty
     - Quiz: How might we also include estimation uncertainty?

Predictive Distribution of Electoral College Delegates

Including fundamental and estimation uncertainty

- Simulate 1,000 national elections ($\rightsquigarrow$ number of Dem delegates)
  - For state i (repeat for $i = 1, \ldots, 51$)
    1. Draw $y_{i,2020}$ from its distribution for state i,
    $$\tilde{y}_{i,2020} \sim \mathrm{P}(y_{i,2020} | y_{it}, t < 2020; X_{it'}, t' \leq 2020)$$
    i.e. P(unknown|data). (Details shortly.)
    2. If $\tilde{y}_{i,2020} > 0.5$ Dems "win" C_i electoral college delegates (Reps get 0); otherwise, Dems get 0 (Reps get C_i)
  - Calculate total Dem delegates nationally: add simulated winnings from all states: $\sum_{i=1}^{51} 1(\tilde{y}_{i,2020} > 0.5)C_i$
- Calculate QOIs: average, standard deviation, histogram

How to draw simulations of $y_{i,2020}$

Including fundamental and estimation uncertainty

1. Choose values of explanatory variables: $X_c = X_{i,2020}$
2. Simulate estimation uncertainty
   - Draw $\eta = \{\tilde{\beta}, \tilde{\gamma}\}$ from its sampling distribution,

   $$\tilde{\eta} \sim N(\hat{\eta}, \hat{V}(\hat{\eta}))$$

   - Pull out $\tilde{\beta}$ and save
   - Pull out $\tilde{\gamma}$, "un-reparameterize" $\tilde{\sigma} = e^{\tilde{\gamma}}$, and save
3. Compute simulated systematic component: $\tilde{\mu}_{it} = X_{i,2020}\tilde{\beta}$
4. Use stochastic component to simulate fundamental uncertainty:

   $$\tilde{y}_{i,2020} \sim N(\tilde{\mu}_{i,2020}, \tilde{\sigma}^2)$$

$\rightsquigarrow$ We can now simulate the number of Democratic delegates, in repeated elections, with fundamental and estimation uncertainty represented

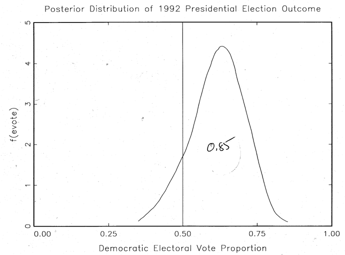
How to do it with a LS Regression Program

Useful to connect to the literature. Feel free to ignore

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, $1/\chi^2(\hat{\sigma}^2, N - k)$, labeling it $\tilde{\sigma}^2$
4. Either:
   - Draw ϵ_{it} from $N(0, \tilde{\sigma}^2)$, label it $\tilde{\epsilon}_{it}$ and compute: $\tilde{y}_{i,2020} = \tilde{X}_{i,2020}\tilde{\beta} + \tilde{\epsilon}_{it}$
   - Or, in our preferred notation, draw $\tilde{y}_{i,2020}$ from $N(X_{i,2020}\tilde{\beta}, \tilde{\sigma}^2)$

Forecasting Errors for 1992 (forecasts from early October)

- Predictive distribution of electoral vote proportion:



- Probability of Dem (Bill Clinton) victory: 0.85
- Error in Democratic 2-party electoral vote proportion: 0.01
- Error in Democratic 2-party popular vote proportion: 0.03
- Quiz: How big do you expect these errors will be if the model is correct and the election were run again?

A Gaussian Variance Function Model

The Model

1. $Y_i \sim N(y_i|\mu_i, \sigma_i^2)$
2. $\mu_i = x_i\beta$, with covariates x_i
3. $\sigma_i^2 = \exp(z_i\gamma)$, with covariates z_i possibly overlapping x_i
4. Y_i and $Y_{i'}$ are independent $\forall\ i \neq i'$, given X and Z.

The Log-Likelihood Derivation

$$\begin{aligned}
\ln L(\beta, \sigma^2|y) &= -\frac{1}{2}\sum_{i=1}^{n}\left[\ln \sigma^2 + \frac{(y_i - \mu_i)^2}{\sigma^2}\right] \\
&= -\frac{1}{2}\sum_{i=1}^{n}\left[z_i\gamma + \frac{(y_i - X_i\beta)^2}{\exp(z_i\gamma)}\right]
\end{aligned}$$

Any questions?