Parameter estimation
• Introduction
• Example
• Ordinary least squares
• Ordinary least squares
• Ordinary least squares
• Ordinary least squares
• End
Introduction
What is model calibration?

• Finding the model parameter values that give the best fit to the data.
Other names for model calibration

- Parameter estimation
  - Statistics
- Inverse problem
  - Engineering
  - Instead of using model with parameter values to calculate response, we use response to calculate parameter values
- Model tuning
  - Climate science (but they also say “calibration”)
How to calibrate?

• A system model can be treated as a regression model – it relates outputs to explanatory variables

• Parameter estimation in regression is a major topic in statistics

• So treat model calibration as a statistics problem
  – But a difficult one
Difficulties of system model calibration

• Many parameters
  – There are often many (hundreds) of parameters

• Two sources of information about parameters
  – How to combine those sources?

• Complex data structure
  – Multiple measurements for each individual (e.g. multiple measurement types and/or dates from each field)
• Practical problems
  – Long execution times
• Many explanatory variables
  – Hard to examine behavior of model versus each explanatory variable to see results of calibration
Status of system model calibration

• Calibration is probably the most difficult aspect of modeling.
  – Takes the most time
• And one of the most important
  – Parameter values have a major effect on predictions
  – Calibration determines predictive quality
• And one of the least consensual
  – No agreed on approach
This lecture

• Present an approach appropriate for simple regression
  – Ordinary Least squares (OLS)
  – Show how to do OLS with R (exercise)
  – OLS will usually not be appropriate for system models
So why is this useful?

– Often good to start with OLS as first step
– There are more complex methods that build on OLS
– So OLS is an important part of the system model calibration toolkit
A simple example
• Calibrate a model for seed weight
  – seed weight versus time (measured in degree days).
1. Define model

- The model for grain filling

\[
y_{\text{mod}} = \frac{W}{1 + \exp(B - (c)(DD))}
\]

- \(y\) is grain weight (mg)
- DD is degree days from anthesis (the explanatory variable)
- \(W, B, c\) are parameters, to be estimated
2. Describe data

- A wheat field in Canada, cultivar Neepawa, standard management.
  1. In a single year, measurement at 10 dates of seed weight of a random sample of seeds. DD values from daily temperature.
  2. Measurements of DD and seed weight in each of 3 years.
3. Define a criterion of “best fit”

• A common criterion is sum of squared errors (SS)

\[
SS = \sum_{i=1}^{n} \left\{ \left[ y_i - f(X_i; \theta) \right]^2 \right\}
\]

• Calibration involves finding parameter values that minimize SS.
• This is called the ordinary least squares (OLS) criterion, because the criterion is a simple sum of squares

• The parameters are the OLS parameters

$$\theta_{OLS} = \arg \min_{\theta} \left( \frac{1}{n} \sum_{i=1}^{n} \left[ y_i - f(X_i; \theta) \right]^2 \right)$$
4. Find best fit parameters

- There are several algorithms, and many software packages
  - The R function “nls” calculates OLS parameters
  - The default algorithm is a Gauss-Newton algorithm
- Can also use trial and error
  - That is a bad idea
• Result of nls for a single year of data
• Result of nls for a 3 years of data
5. Examine results

Formula: \( y \sim \frac{W}{1 + \exp(B - c \times DD.aa)} \)

Parameters:

| Estimate | Std. Error | t value | Pr(>|t|) |
|----------|------------|---------|----------|
| W 28.171255 | 0.408284 | 69.00 | 2.17e-12 *** |
| B 4.160223 | 0.360084 | 11.55 | 2.86e-06 *** |
| c 0.016492 | 0.001415 | 11.65 | 2.68e-06 *** |

---

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8364 on 8 degrees of freedom

Number of iterations to convergence: 5
Achieved convergence tolerance: 1.961e-06
• Also, examine fit of model to data
  – We will see details soon
The estimated parameters depend on the sample

- If we redid the experiment, we would have different data and different estimated parameters

- A good model and a good estimator:
  - The average over samples are the true parameters
  - As sample size increases, less variability between samples
Variability between samples

**Sample size 22**

- **Histogram of BHist**
- **Histogram of cHist**

**Sample size 110**

- **Histogram of BHist**
- **Histogram of cHist**

ENDURE Volterra Nov. 10-14, 2016
Is OLS a good estimator? (and what do we mean by “good” model?)
• OLS is a good estimator if certain assumptions are satisfied
  – Important to test those assumptions
• Write \( y = f(X; \theta) + \varepsilon \)
  – \( y \) is true response (e.g. true seed weight)
  – \( f(X; \theta) \) is the model.
  – \( \varepsilon \) is model error (the difference between the model and the true response)
  – We can always write that. No assumptions so far.
  – Assumptions concern \( \varepsilon \) for the whole population
    • e.g. all samples, all dates in 1988 for seed weight
The assumptions

There is some $\theta^*$ such that, for $\theta = \theta^*$

1. “Correct model” assumption.
   \[ E(\varepsilon) = 0 \text{ for all } X \]

2. “Homoscedasticity” assumption.
   \[ \text{var}(\varepsilon) = \sigma^2 \text{ same for all } X \]

3. “No correlation” assumption
   All $\varepsilon$ are uncorrelated
If the assumptions are satisfied

• Then as the amount of data tends toward infinity
  – The expectation of the OLS parameters tends toward $\theta^*$
  – The variance of the OLS parameters tends toward 0
  – The model tends toward the best possible predictor

• If the assumptions aren’t satisfied, we can’t ensure these properties
1. “Correct model” assumption

- For some parameter vector $\theta^*$, $E(\varepsilon)=0$ for all $X$.
  - The model $f(X;\theta^*)$ goes through the middle of the points
  - This implies that the model $f(X;\theta^*)$ takes $X$ into account correctly.
  - If it didn’t, then $Y-f(X;\theta^*)$ would depend on $X$
\[
\frac{W^*}{1 + e^{B^* - c^*X}}
\]
• Assumption 1 defines “correct” model
• Assumption 1 also defines “true” parameter values.
  – Parameters such that $E(\varepsilon)=0$ for all $X$
• The correct or best model is not necessarily a good predictor
• If \( \text{var}(\varepsilon) \) large, the model may go through the middle of the points, but be far from individual points.
• A “correct model” correctly describes the effect of explanatory variables
• But those explanatory variables may not describe all or even most of the variability in the output
• So a correct model may have small or large errors compared to observations.
  – Depends on choice of explanatory variables.
To test “correct model” assumption

• Do OLS.
• Examine residuals $y - f(X; \theta_{\text{OLS}})$
  – Vocabulary: Residual is difference between an observed value and a simulated value, using parameters estimated from data.
  – Model error is difference, when parameter values aren’t estimated from data.
• Residuals should show no structure as a function of $X$
• That’s easy for a simple model, not for model with many explanatory variables

ENDURE Volterra Nov. 10-14, 2016
Two different models for seed weight

\[ y = \frac{W}{1 + \exp(B - c \times DD)} \]

\[ y = a \times (1 - \exp(-b \times DD)) \]

Graphs showing the relationship between seed weight and DD (°C days) for both models.
• When is assumption 1 likely to be violated?
  – For complex models with many explanatory variables, where the response to each explanatory variable cannot be thoroughly tested.
    • That is often the case for system models
  – For models with many parameters, where some parameters are fixed (not estimated by calibration)
    • That is often the case for crop models
    • This gives incorrect model, even if form of model is correct
• Consequences of violation of assumption 1
  – The bad news:
    • Parameters are just empirical adjustment factors. Not the true values.
  – The good news:
    • The model tends (with lots of data) toward best model with those equations
Prediction

- $f(X; \theta_{\text{OLS}})$ tends toward best predictor of that form
- for the population that is sampled
Assumption 1 and system models

• System models are likely to be incorrect
  – Because can’t test response to all explanatory variables
  – Because many parameters fixed at approximate values

• The OLS parameters then don’t estimate the true parameter values
  – The OLS parameters are just adjustment factors

• Calibration, on average improves prediction
  – For the sampled population. Beware extrapolation.
What to do?

• For simple empirical models, change the model
  – Choose a function that gives a better fit to the data
• For system models
  – Don’t over interpret results.
  – Don’t assume parameters estimate true values
  – Be wary of extrapolation beyond data set
2. "Homoscedasticity" assumption

- \( \text{var}(\varepsilon) = \sigma^2 \) for all \( X \)
  - The spread of \( y \) around \( f(X, \theta^*) \) is the same for all \( X \).
To test homoscedasticity

• Do OLS.

• Examine residuals $y - f(X; \theta_{OLS})$
  – variability of residuals should be about the same for all $X$
  – Can divide $X$ into zones, do statistical test
• In this example, residuals have similar spread for all values of DD

\[ y = \frac{W}{1 + \exp(B - c \cdot DD)} \]

Divide residuals into groups, DD < 420 or DD > 420. Do Bartlett’s test for equality of variances. p = 0.75.

ENDURE Volterra Nov. 10-14, 2016
• When is assumption 2 likely to be violated?
  – For variables with large range of values
    • Often, residual variance is proportional to size of response
    • System models that describe a growth cycle will often have variables like that. Examples: LAI or biomass in crop models.
    • Residual plot would look like this:

![Residual plot](image)

Divide residuals into groups, days ≤80 or days>80.
Do Bartlett’s test for equality of variances. p=0.04

ENDURE Volterra Nov. 10-14, 2016
For model with multiple types of response variable

• Different responses will have different residual variances

• System models often have multiple responses

  – Example: If data are for aphid and ladybug population densities, they probably won’t have same residual variance.
• Consequences of violation of assumption 2
  – The parameters still tend toward the true parameters (if assumption 1 is satisfied)
  – The model still tends toward the best predictor
  – But the variance for different possible data sets is not minimal
    • Convergence toward best values could be faster
  – The estimated parameter uncertainty is not realistic.
What to do

• Do weighted least squares (WLS).
  – This involves weighting outputs by $1/\text{variance}$. That makes weighted variables homoscedastic. Then can do OLS
Assumption 2 and system models

• System models are very likely to have heteroscedasticity
  – Responses that vary a lot over time
  – Multiple responses

• Use WLS to estimate parameters
3. “No correlation” assumption

• The error for one data point is unrelated to errors for other data points
• Depends a lot on sampling method
  • If every data point is drawn independently at random from population, this assumption is satisfied by construction
  • If there is hierarchical sampling, correlations may be present (need to check)
To test “no correlation”

• Consider sampling scheme
  – Does the same individual (e.g. field) contribute multiple measurements?
  – If so, assumption 3 may be violated

• Do OLS.
  – Examine residuals $y - f(X; \theta_{\text{OLS}})$, identify by individual
  – The residuals for same individual shouldn’t be similar
• Errors for the same year are similar i.e. there is nonzero correlation between data points for the same year.
Residuals from same year are related

![Graph showing observed-predicted seed weight vs. DD (°C days)]
• When is assumption 3 likely to be violated?
  – Whenever the sample is the result of other than simple random sampling.
  – Example
    • Multiple measurements of a population in the same field over time
    • If model overestimates in a field, may overestimate at all times (effect of that field)
• Consequences of violation of assumption 3
  – The parameters still tend toward the true parameters (if assumption 1 is satisfied)
  – The model still tends toward the best predictor
  – But the variance for different possible data sets is not minimal
    • Convergence toward best values could be faster
  – The estimated parameter uncertainty is not realistic.
    • It is in general underestimated
What to do

• Do generalized least squares (GLS)
  – This involves doing a transformation of the data
  – That makes transformed variables independent
  – Then can o OLS
Recap

Calibration of system models

• Use standard statistical calibration
• Start with OLS, but test assumptions
  – Good chance that model isn’t “correct model”
  – Probably have heteroscedasticity
  – Often have correlated errors
  – At least correct for heteroscedasticity and correlation
• Look at variances of estimated parameters
  – If large, there may be large uncertainty in predictions
THE END