ORIE 3120: Practical Tools for OR, DS, and ML

Design of Experiments, Factor Models, and Quality Improvement

Professor Udell
Operations Research and Information Engineering
Cornell

May 5, 2020
Announcements

- submit recitation by 4:30pm ET Friday (last recitation!)
- logistic regression homework due 2:30pm ET Wednesday (tomorrow)
- have a question about your grade? (eg, “should I go S/U or GRV?”)
  → ask in TA office hours!
  (we use breakout rooms so this discussion will be private)
- ask questions about project after class, or in office hours
Project milestone II rubric

- Is the project driven by asking and answering interesting questions?
- How well does the report answer the questions posed?
- Are the visualizations easy to understand? Do they add value?
- Is the report well-written and interesting?
- Does the project use at least 3 tools from class?
  - Linear regression
  - Logistic regression
  - Checking assumptions of linear regression to ensure validity of pvalues
  - Cross-validation or out-of-sample validation
  - Model selection
  - Assessing collinearity
  - Forecasting (with trend / with seasonality)
- How creative are the analyses? Did this project surprise you? Did you learn something?
- Are the techniques they use well-explained and easy to understand?
- Does the project comply with the technical requirements (eg, page limit)? Is it well-formatted and pretty?
Outline

DOX

One-factor models

- Fixed and random effects
- ANOVA estimates of variance components
- REML estimates of variance components
Outline

DOX

One-factor models

Fixed and random effects
ANOVA estimates of variance components
REML estimates of variance components
A **factor** is a “categorical” predictor

- the values of the factor are called “levels”
  - the factor “machine” might have three levels, machine 1, machine 2, machine 3
  - the variable “temperature” might have four levels, 300°, 325°, 350°, 375° (treated as categories)
  - the individual batches from a production process are levels of the factor “batch”
- typically, the number of levels is small, even 2
Single factor

We begin by considering data with a single factor

- There are \( I \) levels of the factor
- There are \( n_i \) observations at the \( i \)th level
- The responses are \( Y_{ij}, \ i = 1, \ldots, I \) and \( j = 1, \ldots, n_i \)

**example:**

- there are 4 different temperature settings you try for baking bread: 300\(^\circ\), 325\(^\circ\), 350\(^\circ\), 375\(^\circ\), which we refer to as temperatures 1,2,3, and 4.
- you bake \( n_i \) loaves of bread at each temperature \( i \).
- you observe the response \( Y_{ij} \): how high loaf \( j \), baked at temperature \( i \), rises.

**Goal:** find the temperature best for baking fluffy bread.
The statistical model is

\[ Y_{ij} = \mu_i + \epsilon_{ij} = \mu + \alpha_i + \epsilon_{ij} \]

We have decomposed \( \mu_i \) into an overall mean \( \mu \) and the deviation \( \alpha_i = \mu_i - \mu \) of the \( i \)th mean from the overall mean (so \( \mu_i = \mu + \alpha_i \)).

We assume the \( \epsilon_{ij} \) are independent \( N(0, \sigma^2_\epsilon) \).

\( \alpha_1, \ldots, \alpha_I \) are called the “effects” of the factor.

Two possible goals:

- **fixed effect model.** Estimate \( \alpha_i \) for each \( i \).
- **random effect model.** Estimate distribution of \( \alpha_i \).
A single factor with fixed effects

From previous page:

\[ Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \]

**Fixed effects:** the effects \( \alpha_1, \ldots, \alpha_I \) of the levels are viewed as fixed parameters

- **Example:** the levels are the only three suppliers of silicon wafers used by the company
- **Example:** the levels are the only four operators employed by the company
- goal is to estimate \( \alpha_1, \ldots, \alpha_I \)
- use fixed effect if you **can control** the level
- we assume \( \alpha_1 + \cdots + \alpha_I = 0 \)
A single factor with random effects

From previous page

\[ Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \]

Random effects: \( \alpha_1, \ldots, \alpha_I \) are assumed independent \( N(0, \sigma^2_\alpha) \)

- “levels” are a sample from a larger population
- Example: levels are a sample of silicon wafers from supplier
- Example: levels are a sample of operators from large pool of workers
- use random effect if you cannot control the level
- instead, goal is to generalize conclusions to the larger population
- the particular levels in the sample are not of much interest
- the population variance \( \sigma^2_\alpha \) is the parameter of interest
- we assume \( E\alpha_i = 0 \) for each \( i = 1, \ldots, I \)
Random Effects, with One Factor
Example: \( Y_{ij} \) is the quality of the \( j \)th item from the \( i \)th batch

- Use a random effects model with a single factor:
  \[
  Y_{ij} = \mu + \alpha_i + \epsilon_{ij}
  \]

- The levels of this factor are the batches of the product.

- \( \sigma_\alpha \) is the standard deviation of the batch means

- \( \sigma_\epsilon \) is the within-batch standard deviation of the product

- \( \sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2} \) is the overall standard deviation of the product
Example

- Keeping $\sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2}$ small is the objective.
- Decomposing the product’s standard deviation into $\sigma_\alpha^2$ and $\sigma_\epsilon^2$ helps locate the source of variability.
Example: Lipton Dry Soup Mix

Problem: Too much variation in one component, the “intermix”, in dry soup mix

- too little intermix $\Rightarrow$ flavor too weak
- too much intermix $\Rightarrow$ flavor too strong
Example: Lipton Dry Soup Mix

- every 15 minutes, a batch of 5 samples is produced
- “batch” is a random factor
- Variation within the batch $\sigma_\epsilon$ is due to the equipment that mixes and packages the raw ingredients
- Variation across the batch $\sigma_\alpha$ is due to variation in the raw ingredients
- If $\sigma_\epsilon > \sigma_\alpha$, then we can try to improve how we mix and package the raw ingredients
- If $\sigma_\alpha > \sigma_\epsilon$, then we can try to reduce variation in the raw ingredients
Estimating the variance components

We call $\sigma^2_\alpha$ and $\sigma^2_\epsilon$ the “variance components”. We will describe two different ways to estimate them:

- Analysis of Variance (ANOVA)
- Restricted Maximum Likelihood (REML)

ANOVA is simpler but less accurate.
Estimating the variance components: the ANOVA method

For simplicity assume $n_i = n$ for all $i$

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$  (From a previous page)

Define

$$\overline{Y}_i := n^{-1} \sum_{j=1}^{n} Y_{ij} = \mu + \alpha_i + \left( n^{-1} \sum_{j=1}^{n} \epsilon_{ij} \right) \approx \mu + \alpha_i$$

$$\hat{\mu} = \overline{Y}$$
$$\hat{\alpha}_i = \overline{Y}_i - \hat{\mu} = \overline{Y}_i - \overline{Y}$$

$$s_i^2 = (n - 1)^{-1} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_i)^2$$  (within-sample variance)

$$\hat{\sigma}_\epsilon^2 = l^{-1} \sum_{i=1}^{l} s_i^2$$  (average within-sample variance)
Estimating the variance components: the ANOVA method

\[
\bar{Y}_i - \mu = n^{-1} \sum_{j=1}^{n} (\alpha_i + \epsilon_{ij}) = \alpha_i + n^{-1} \sum_{j=1}^{n} \epsilon_{ij}
\]

\[
\text{Var}[\bar{Y}_i - \mu] = \text{Var}[\bar{Y}_i] = \sigma^2 + \frac{\sigma^2_\epsilon}{n}
\]

Since \( \hat{\alpha}_i = \bar{Y}_i - \bar{Y} \) and \( \bar{Y} \) is the sample average of the \( \bar{Y}_i \),

\[
(l - 1)^{-1} \sum_{i=1}^{l} \hat{\alpha}_i^2 = (l - 1)^{-1} \sum_{i=1}^{l} (\bar{Y}_i - \bar{Y})^2
\]

is an unbiased estimator of \( \text{Var}[\bar{Y}_i] = \sigma^2 + \sigma^2_\epsilon / n \).

Therefore

\[
\hat{\sigma}^2_\alpha = (l - 1)^{-1} \sum_{i=1}^{l} \hat{\alpha}_i^2 - \frac{\hat{\sigma}^2_\epsilon}{n}
\]
Estimating the variance components: the REML method

- REML = REstricted Maximum Likelihood
- these are maximum likelihood estimators with a correction for bias
- (a better name would be bias-corrected maximum likelihood)
## Wafer example: data

<table>
<thead>
<tr>
<th>wafer</th>
<th>location</th>
<th>thickness</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>89.98323716</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>90.08198204</td>
</tr>
<tr>
<td>1</td>
<td>3</td>
<td>89.95311004</td>
</tr>
<tr>
<td>1</td>
<td>4</td>
<td>89.97625452</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>89.94101068</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>89.94956078</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>89.92025096</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>89.86716804</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>90.07129618</td>
</tr>
<tr>
<td>3</td>
<td>2</td>
<td>89.94329401</td>
</tr>
<tr>
<td>3</td>
<td>3</td>
<td>89.96035653</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>90.02122963</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>89.8763338</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>90.0948249</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>90.03716224</td>
</tr>
<tr>
<td>8</td>
<td>2</td>
<td>89.9546313</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>90.01625186</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>90.04526308</td>
</tr>
</tbody>
</table>
Wafer example: plot

![Graph showing wafer thickness comparison](image-url)
Demo: DOE

Wafer example: random effects model

```python
wdata = pd.read_csv('waferdata.csv')
smf.mixedlm("thickness ~ 1", wdata, groups=wdata["wafer"])
```

- thickness is the response
- "thickness ~ 1" specifies a fixed intercept
- "groups=wdata["wafer"]" specifies a random effect for each wafer
Wafer Example: random effects model

The model specified by the demo code is:

\[ Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \]

\( Y_{ij} \) is the thickness at the \( i \)th location within the \( j \)th wafer
### Wafer Example: model summary

#### Mixed Linear Model Regression Results

<p>| | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Model:</td>
<td>MixedLM</td>
<td>Dependent Variable:</td>
<td>thickness</td>
<td></td>
</tr>
<tr>
<td>No. Observations:</td>
<td>32</td>
<td>Method:</td>
<td>REML</td>
<td></td>
</tr>
<tr>
<td>No. Groups:</td>
<td>8</td>
<td>Scale:</td>
<td>0.0022</td>
<td></td>
</tr>
<tr>
<td>Min. group size:</td>
<td>4</td>
<td>Likelihood:</td>
<td>44.4661</td>
<td></td>
</tr>
<tr>
<td>Max. group size:</td>
<td>4</td>
<td>Converged:</td>
<td>Yes</td>
<td></td>
</tr>
<tr>
<td>Mean group size:</td>
<td>4.0</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Coef. | Std.Err. | z     | P>|z| | [0.025 0.975] |
|-------|----------|-------|-------|----------------|
| Intercept | 89.982   | 0.017 | 5368.084 | 0.000 | 89.949 90.015 |
| Group Var  | 0.002    | 0.029 |                   |               |
Random effects poll

Suppose I fit a random effects model with a single factor

$$Y_{ij} = \mu + \alpha_i + \epsilon_{ij}$$

- $i$ indexes batches and $j$ indexes items within a batch
- Recall: $\mu$ is deterministic, $\alpha_i \sim N(0, \sigma^2_\alpha)$, $\epsilon_{ij} \sim N(0, \sigma^2_\epsilon)$

What is the variance of $Y_{ij}$?

- (up) $\sigma^2_\alpha$
- (down) $\sigma^2_\epsilon$
- (yes) $\sigma^2_\alpha + \sigma^2_\epsilon$
- (no) $\sqrt{\sigma^2_\alpha + \sigma^2_\epsilon}$
- (coffee) none of the above
Random effects poll

Suppose I fit a random effects model with a single factor

\[ Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \]

\( i \) indexes batches and \( j \) indexes items within a batch

Recall: \( \mu \) is deterministic, \( \alpha_i \sim N(0, \sigma_\alpha^2) \), \( \epsilon_{ij} \sim N(0, \sigma_\epsilon^2) \)

What is the covariance between two items in the same batch?

- (up) \( \sigma_\alpha^2 \)
- (down) \( \sigma_\epsilon^2 \)
- (yes) \( \sigma_\alpha^2 + \sigma_\epsilon^2 \)
- (no) \( \sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2} \)
- (coffee) none of the above
Random effects poll

Suppose I fit a random effects model with a single factor

\[ Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \]

- \( i \) indexes batches and \( j \) indexes items within a batch
- Recall: \( \mu \) is deterministic, \( \alpha_i \sim N(0, \sigma_\alpha^2) \), \( \epsilon_{ij} \sim N(0, \sigma_\epsilon^2) \)

What is the correlation between two items in the same batch?

- (up) 0
- (down) \( \sigma_\alpha^2 / (\sigma_\alpha^2 + \sigma_\epsilon^2) \)
- (yes) \( \sigma_\epsilon^2 / (\sigma_\alpha^2 + \sigma_\epsilon^2) \)
- (no) \( \sigma_\alpha / \sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2} \)
- (coffee) \( \sigma_\epsilon / \sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2} \)

Recall: the correlation between two random variables \( A \) and \( B \) is \( \text{Cov}(A) / \sqrt{\text{Var}(A)\text{Var}(B)} \).
Random effects poll

Suppose I fit a random effects model with a single factor

\[ Y_{ij} = \mu + \alpha_i + \epsilon_{ij} \]

- \( i \) indexes batches and \( j \) indexes items within a batch
- Recall: \( \mu \) is deterministic, \( \alpha_i \sim N(0, \sigma^2_{\alpha}) \), \( \epsilon_{ij} \sim N(0, \sigma^2_{\epsilon}) \)

What is the correlation between two items in different batches?

1. 0
2. \( \sigma^2_{\alpha}/(\sigma^2_{\alpha} + \sigma^2_{\epsilon}) \)
3. \( \sigma^2_{\epsilon}/(\sigma^2_{\alpha} + \sigma^2_{\epsilon}) \)
4. \( \sigma_{\alpha}/\sqrt{\sigma^2_{\alpha} + \sigma^2_{\epsilon}} \)
5. \( \sigma_{\epsilon}/\sqrt{\sigma^2_{\alpha} + \sigma^2_{\epsilon}} \)

Recall: the correlation between two random variables \( A \) and \( B \) is \( \text{Cov}(A)/\sqrt{\text{Var}(A)\text{Var}(B)} \).