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 12, 2020

Announcements

- grade estimates should now be correct
- forecasting homework due 2:30pm ET tomorrow (Wednesday)
- project milestone II due tonight (Tuesday) 11:59pm
- peer review for project milestone II due next Sunday 11:59pm
  - complete peer review to avoid losing up to 30% of your project grade
  - we'll use a different system (Google forms, not Canvas)

(Optional:) Survey of Engineering students about COVID impact

Survey of engineering students about COVID impact:
`https://www.research.net/r/rapidcovid19`

Subject: Help us understand how universities can better support engineering students through the COVID-19 crisis
Body: Are you an undergraduate engineering student and 18 or older? If so, we need your help!
We are writing to ask for your help in determining how engineering students are coping with the COVID-19 pandemic and how COVID-19 is affecting student stress levels. This information will be used to help inform universities and engineering programs nationwide on how best to support students during this crisis and in future emergencies. If you are willing to participate, please use the link below to take the included survey. We estimate participation will take about 15 minutes. As an added bonus, the first 1000 participants will receive a $5 amazon.com gift card!

Outline

Factors

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 ith level
- The responses are

$$Y_{ij}, \;\; i = 1, \ldots, I \;\; \text{and} \;\; j = 1, \ldots, n_i$$

example:

- there are 4 different temperature settings you try for baking bread: 300°, 325°, 350°, 375°, 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.

Single factor

- The statistical model is

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

- We have decomposed μ_i into an overall mean μ and the deviation $\alpha_i = \mu_i - \mu$ of the ith mean from the overall mean (so $\mu_i = \mu + \alpha_i$).
- We assume the ϵ_{ij} are independent $N(0, \sigma_\epsilon^2)$
- $\alpha_1, \ldots, \alpha_I$ are called the "effects" of the factor

Two possible goals:

- **fixed effect model.** Estimate α_i for each i.
- **random effect model.** Estimate distribution of α_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: $\alpha_1, \dots, \alpha_I$ are assumed independent $N(0, \sigma_\alpha^2)$

- "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 σ_α^2 is the parameter of interest
- we assume $\mathrm{E}\alpha_i = 0$ for each $i = 1, \dots, I$

Random Effects, with One Factor

Example

Example: Y_{ij} is the quality of the jth item from the ith 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.
- σ_α is the standard deviation of the batch means
- σ_ϵ 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 σ_α^2 and σ_ϵ^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 σ_ϵ is due to the equipment that mixes and packages the raw ingredients
- Variation across the batch σ_α 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 σ_α^2 and σ_ϵ^2 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

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

Define

$$\overline{Y}_i := n^{-1} \sum_{j=1}^{n} Y_{ij} = \mu + \alpha_i + \underbrace{\left(n^{-1} \sum_{j=1}^{n} \epsilon_{ij} \right)}_{\text{expected value is 0}} \approx \mu + \alpha_i$$

$$\begin{aligned}
\hat{\mu} &= \overline{Y} \\
\hat{\alpha}_i &= \overline{Y}_i - \hat{\mu} = \overline{Y}_i - \overline{Y} \\
s_i^2 &= (n-1)^{-1} \sum_{j=1}^{n} (Y_{ij} - \overline{Y}_i)^2 \quad \text{(within-sample variance)} \\
\hat{\sigma}_\epsilon^2 &= I^{-1} \sum_{i=1}^{I} s_i^2 \qquad \text{(average within-sample variance)}
\end{aligned}$$

Estimating the variance components: the ANOVA method

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

$$\mathrm{Var}[\overline{Y}_i - \mu] = \mathrm{Var}[\overline{Y}_i] = \sigma_\alpha^2 + \frac{\sigma_\epsilon^2}{n}$$

Since $\hat{\alpha}_i = \overline{Y}_i - \overline{Y}$ and $\overline{Y}$ is the sample average of the $\overline{Y}_i$,

$$(I-1)^{-1} \sum_{i=1}^{I} \hat{\alpha}_i^2 = (I-1)^{-1} \sum_{i=1}^{I} (\overline{Y}_i - \overline{Y})^2$$

is an unbiased estimator of $\mathrm{Var}[\overline{Y}_i] = \sigma_\alpha^2 + \sigma_\epsilon^2 / n$.

Therefore

$$\hat{\sigma}_\alpha^2 = (I-1)^{-1} \sum_{i=1}^{I} \hat{\alpha}_i^2 - \frac{\hat{\sigma}_\epsilon^2}{n}$$

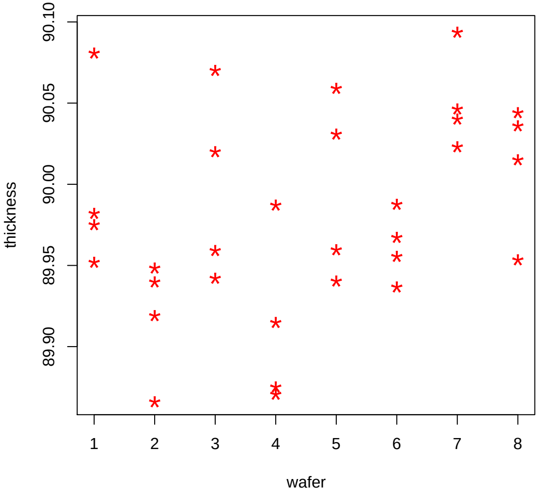
Estimating the variance components: the REML method

- 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

```
wafer location thickness
1    1    89.98323716
1    2    90.08198204
1    3    89.95311004
1    4    89.97625452
2    1    89.94101068
2    2    89.94956078
2    3    89.92025096
2    4    89.86716804
3    1    90.07129618
3    2    89.94329401
3    3    89.96035653
3    4    90.02122963
4    1    89.8763338
   .
   .
   .
7    4    90.0948249
8    1    90.03716224
8    2    89.9546313
8    3    90.01625186
8    4    90.04526308
```

Wafer example: plot



Demo: DOE

https://github.com/madeleineudell/orie3120-sp2020/blob/master/demos/doe.ipynb

Wafer example: random effects model

```
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:

$$\underbrace{Y_{ij}}_{\text{thickness}} = \underbrace{\mu}_{\text{fixed effect}} + \underbrace{\alpha_i}_{\text{random effect}} + \underbrace{\epsilon_{ij}}_{\text{noise}}$$

Y_{ij} is the thickness at the ith location within the jth wafer

Wafer Example: model summary

```
  Mixed Linear Model Regression Results
=======================================================
Model:             MixedLM Dependent Variable: thickness
No. Observations: 32      Method:              REML
No. Groups:        8       Scale:               0.0022
Min. group size:   4       Likelihood:          44.4661
Max. group size:   4       Converged:           Yes
Mean group size:   4.0
-------------------------------------------------------
    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: μ is deterministic, $\alpha_i \sim N(0, \sigma_\alpha^2)$, $\epsilon_{ij} \sim N(0, \sigma_\epsilon^2)$

What is the variance of Y_{ij}?

- (up) σ_α^2
- (down) σ_ϵ^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: μ 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) σ_α^2
- (down) σ_ϵ^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: μ 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 $\mathrm{Cov}(A)/\sqrt{\mathrm{Var}(A)\mathrm{Var}(B)}$.

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: μ 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 different batches?

- 0
- $\sigma_\alpha^2/(\sigma_\alpha^2 + \sigma_\epsilon^2)$
- $\sigma_\epsilon^2/(\sigma_\alpha^2 + \sigma_\epsilon^2)$
- $\sigma_\alpha/\sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2}$
- $\sigma_\epsilon/\sqrt{\sigma_\alpha^2 + \sigma_\epsilon^2}$

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

Outline

Two Factor Models, No Interactions

Two factors

We now look at experiments with two factors

- we will call them “A” and “B”
- we will model
  - the effect of A alone (the “main effect of A”)
  - the effect of B alone (the “main effect of B”)
  - in some cases, we model an effect due to A and B together beyond what is modeled by their main effects (the “interaction” of A and B)

Two-factor model, no interactions

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

- α_i = effect of factor A at level i. We call this the "main effect of A".
- β_j = effect of factor B at level j. We call this the "main effect of B".
- The model above does not include the "interaction" between A and B. We will consider models with interactions soon.

Two-factor model, no interactions, fixed effects

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

- if A and B have fixed effects, then we view $\alpha_1, \ldots, \alpha_I$ and $\beta_1, \ldots, \beta_J$ as fixed pameters.
- $\alpha_1, \ldots, \alpha_I$ and $\beta_1, \ldots, \beta_J$ are the parameters of interest.
- We assume $\alpha_1 + \cdots + \alpha_I = 0$.
- We assume $\beta_1 + \cdots + \beta_J = 0$.

Two-factor model, no interactions, random effects

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

- if A and B have random effects, then we assume:
  - $\alpha_1, \ldots, \alpha_I$ are independent $N(0, \sigma_\alpha^2)$
  - $\beta_1, \ldots, \beta_J$ are independent $N(0, \sigma_\beta^2)$.
- σ_α^2 and σ_β^2 are the parameters of interest.

Two-factor model, no interactions, mixed effects

$$Y_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

- We can view A as a fixed effect, and B as a random effect. This is called "mixed effects".
- $\alpha_1, \ldots, \alpha_I$ are viewed as fixed parameters.
- $\beta_1, \ldots, \beta_J$ are independent $N(0, \sigma_\beta^2)$
- $\alpha_1, \ldots, \alpha_I$ and σ_β^2 are the parameters of interest.
- We assume $\alpha_1 + \cdots + \alpha_I = 0$.

Demo: DOE

`https://github.com/madeleineudell/orie3120-sp2020/blob/master/demos/doe.ipynb`

Two Factor Models, Interactions, Fixed Effects

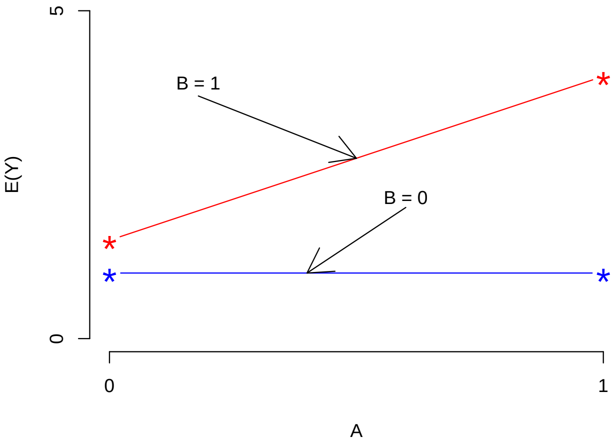
What is an interaction?

When there are two or more factors, then interactions are possible. In the plot, factors A and B have 2 levels each.

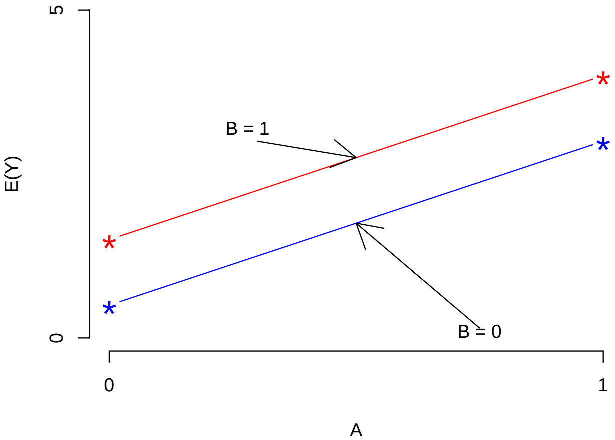
- Factors A and B interact when the effect of factor A depends on the level of factor B
- In the following plot:
  - A has no effect if B = 0
  - A has a large effect if B = 1

Plot: with an interaction

In the plot, factors A and B have 2 levels each.



Plot: no interaction model



Models without interactions are called "additive models"

Our model without without interactions is:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \alpha_i + \beta_j + \epsilon_{ijk}$$

The cell mean is additive across factors, by which we mean:

$$\mu_{ij} = \mu + \alpha_i + \beta_j$$

The effect of changing A from level i to level i^* with B=j is:

$$\mu_{ij} - \mu_{i^*j}$$

This effect does not depend on j because:

$$\mu_{ij} - \mu_{i^*j} = (\mu + \alpha_i + \beta_j) - (\mu + \alpha_{i^*} + \beta_j) = \alpha_i - \alpha_{i^*}.$$

Interactions

The interaction between A and B is:

$$(\alpha\beta)_{ij} := \underbrace{\mu_{ij}}_{\text{true cell mean}} - \underbrace{(\alpha_i + \beta_j + \mu)}_{\text{additive model cell mean}}$$

To include interactions into our model, we redefine μ_{ij} as:

$$\mu_{ij} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij}$$

Then our new model is:

$$Y_{ijk} = \mu_{ij} + \epsilon_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \epsilon_{ijk}$$

Constraints in models with interactions

Let's count parameters

$$\underbrace{\mu_{ij}}_{IJ \text{ parameters}} = \underbrace{\mu}_{1 \text{ parameter}} + \underbrace{\alpha_i}_{I \text{ parameters}} + \underbrace{\beta_j}_{J \text{ parameters}} + \underbrace{(\alpha\beta)_{ij}}_{IJ \text{ parameters}}$$

We have more parameters on the right hand side than on the left

(because $IJ < 1 + I + J + IJ$), so we need to add constraints.

Constraints on the main effects

As before

$$\alpha_1 + \cdots + \alpha_I = 0$$

$$\beta_1 + \cdots + \beta_J = 0$$

Constraints on the interactions

- $(\alpha\beta)_{i1} + \cdots + (\alpha\beta)_{iJ} = 0$ for $i = 1, \ldots, I$
- $(\alpha\beta)_{1j} + \cdots + (\alpha\beta)_{Ij} = 0$ for $j = 1, \ldots, J$.

Stated differently, $(\alpha\beta)_{ij}$ averages to 0 across either i or j

Outline

Which experiments to run?

guidelines:

- want covariates to vary as much as possible
- (within domain of reason and validity of model)

Which experiments to run?

categorical and binary variables:

- try all the values!

real-valued variables:

- grid the values (on a linear or log scale)
- try random values (from appropriate distribution) (within bounds dictated by eg safety)

Which experiments to run?

models with many variables:

- try all combinations of variables and their values + gridded values of real-valued variables
- try all combinations of categorical variables and their values + random values of real-valued variables
- subsample from the above schemes if resources are limited

Surprise! random search beats grid search

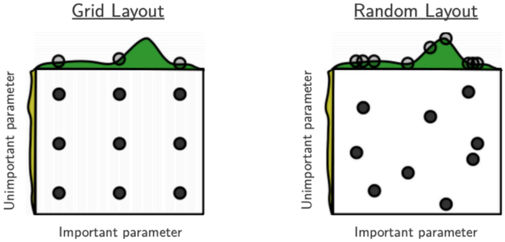


Figure 1: Grid and random search of nine trials for optimizing a function $f(x,y) = g(x) + h(y) \approx g(x)$ with low effective dimensionality. Above each square $g(x)$ is shown in green, and left of each square $h(y)$ is shown in yellow. With grid search, nine trials only test $g(x)$ in three distinct places. With random search, all nine trials explore distinct values of g. This failure of grid search is the rule rather than the exception in high dimensional hyper-parameter optimization.

from Bergstra and Bengio 2012

Outline

Review

- practical tools for
- operations research
- data science
- machine learning

Practical tools

- SQL: storing and accessing large data sets
- Excel and VBA: communicating with management
- Tableau: visualizing data
- Python: analyzing data

Practical tools

- SQL: storing and accessing large data sets
- Excel and VBA: communicating with management
- Tableau: visualizing data
- Python: analyzing data

there are more classes at Cornell on each of these!

- SQL: databases (CS)
- Excel and VBA: spreadsheets (ORIE)
- Data visualization (CS)
- Data analysis and ML: ORIE, CS, and Statistics

Operations research

continuous improvement: repeat

- measure
- analyze
- act

Operations research

continuous improvement: repeat

- measure
- analyze
- act

motivating questions:

- how to reduce supply chain costs?
- how to improve my product?
- how to predict future demand?

Data science

grab the right data!

- what questions are you asking?
- what data would you need to answer them?
- do you need to join several datasets?

Data science

grab the right data!

- what questions are you asking?
- what data would you need to answer them?
- do you need to join several datasets?

look at the data

- map
- trends
- distribution
- correlations

Machine learning

n.b. machine learning is a buzzword; also called statistics, analytics, data science, . . .

analyze the data

- what covariates matter?
- correlations
- predictions
- forecasts

Machine learning

n.b. machine learning is a buzzword; also called statistics, analytics, data science, . . .

analyze the data

- what covariates matter?
- correlations
- predictions
- forecasts

loop back: what next to measure?

Next steps to learn more

- operations research: all ORIE courses!
- data science: ORIE, Statistics, CS, IS
- machine learning: ORIE, Statistics, CS

lots of overlap, but useful to gather many perspectives!

Next steps to learn more

- operations research: all ORIE courses!
- data science: ORIE, Statistics, CS, IS
- machine learning: ORIE, Statistics, CS

lots of overlap, but useful to gather many perspectives!

take ORIE 4741 with me next fall!

Outline

COVID and Cornell

In March I could see the future. . . now I can't

COVID and Cornell

In March I could see the future. . . now I can't
why?

- because outcomes depend on our collective actions (especially, on government action)
- and predicting politics is harder than epidemiology! (small data problem)

The landscape

my guess:

- COVID remains endemic (and scary) in the US throughout 2020, likely until mid-2021
- different regions will be more or less successful
  - Ithaca / upstate NY will do well at containment
- we need a coherent federal responce to extinguish transmission
  (as in eg Taiwan, South Korea, New Zealand)
  - because states and municipalities must balance budgets
  - because states cannot restrict interstate commerce

  only the federal gov't can print money to buy tests and PPE;
  only the federal gov't can quarantine travellers on entry to a “COVID-free” region
- we will not have a coherent federal response until 2021

Will Cornell's campus open in the fall?

my guess:

- yes, to some students
  - students will be tested on arrival
  - and periodically thereafter
  - an honor code and/or app will govern behavior
  - violations will be strictly enforced (suspension/expulsion)

Will Cornell's campus open in the fall?

my guess:

- yes, to some students
  - students will be tested on arrival
  - and periodically thereafter
  - an honor code and/or app will govern behavior
  - violations will be strictly enforced (suspension/expulsion)
- not for all students
  - some students have weakened immune systems
  - housing and classes would be too crowded if all returned
  - students not on campus will be able to make degree progress at home

Will Cornell's campus open in the fall?

my guess:

- yes, to some students
  - students will be tested on arrival
  - and periodically thereafter
  - an honor code and/or app will govern behavior
  - violations will be strictly enforced (suspension/expulsion)
- not for all students
  - some students have weakened immune systems
  - housing and classes would be too crowded if all returned
  - students not on campus will be able to make degree progress at home
- not for all classes
  - class sizes bigger than some threshold (10? 50?) will not be able to safely meet
  - we may push big classes online, or limit attendance each day
  - some professors (*e.g.*, immunosuppressed) will not be able to teach in person but will still teach
  - classes that require in-person interaction (*e.g.*, project teams or studio art or lab classes) will take priority

The big question for students

Cornell's ability to (stay) open depends on your willingness to follow distancing guidelines while you're here

how can we all cooperate enough to stay safe (and stay open)?

The big question for students

Cornell's ability to (stay) open depends on your willingness to follow distancing guidelines while you're here

how can we all cooperate enough to stay safe (and stay open)?

- half of young men 19–14 admitted to breaking UK lockdown rules: https://drive.google.com/file/d/1AOc0wCPqv2gfFSQ_DVmw12vrqQK01z0V/view

ORIE 3120 course survey

Two course surveys (part of participation grade):

- https://forms.gle/kCijZ2i6rb7fHpbG9
- Course evaluation administered by Cornell

(The course staff will delete identifying info (netID, email) from survey before looking at the contents of any of your answers.)

Thanks!

- it has been a delight to teach you
- . . . even remotely!
- stay safe, and hope to see you in the fall!

questions?