How to Learn and Teach on the Internet

Lessons from the Open University (UK)

Ravi Charan
Gabriel Seemann

// Flatiron (NYC Data Science)
December 2, 2019

Despite Potential, Students don't Engage with Online Courses

Representative Situation

The first 4 years of EdX (MIT/Harvard):

- 8.3M Signups
- 4.4M Participants (access some material)
- 0.7M Explorers (access >50% of material)
- 0.2M Certificates (finished class)

Source: MIT Office of Digital Learning

Potential Research Questions

Prediction and Interventions for:

- Drop-outs
- **Learning outcomes**
- Career outcomes

We used Student Level Data to Explore Learning Outcome Correlates

Dataset

- The **Open University** is a large, public research university in the UK
  - 174K students enrolled per year
  - Mostly distance learning
- Anonymized data from 6 classes over 4 semesters from 2013–2014
  - 19 class instances
  - **10,151 students** with grades

Metrics

- **Grade:** on assignments throughout the class (0–100)
- **Student Information**: age band, residence SES, prior course history, credit load, disability status, prior education
- **Activity**
  - Numerical distribution: number of days, mean day, variability, skew, number of unique materials accessed, start and end dates
- **Attention**
  - Clicks by type of material
  - Distribution of materials accessed compared to class average (f-Divergence e.g. Hellinger distance)

Task: Regression (inference)
Identification Strategy: None

- Results should be interpreted as suggesting interventions with randomized controls

Attention and Activity Correlate with Outcomes

Attention Matters

- Overall number of clicks is **not significantly correlated** with outcome
- Significance of engagement **varies widely** with type of materials

Consistency matters

- **Front-loading** of activity correlates with negative outcomes
- High **variation in activity** correlates with negative outcomes

Findings can be:

- **Communicated to students**,
- Serve as **early warning signs** and provide the basis for **randomized control trials** to see if changing patterns leads to improved outcomes

	Coefficient	Log_{10} p
Residence SES	0.0441	-14
Previous Attempts	-2.4123	-11
Activity Skew	-0.9065	-3
Activity Variability	-0.0589	-11
Clicks: External Quiz	0.0857	-7
Clicks: Folder	-0.7095	-4
Clicks: Forum	0.0022	-18
Clicks: Glossary	0.0081	-3
Clicks: Collaboration Material	0.0355	-3
Clicks: University Content	0.0029	-19
Clicks: University Wiki	0.0159	-15
Clicks: Resource	0.0045	-3
Clicks: URL	0.0112	-3

Our model selection process focused on identifying the best correlates

Step 1: ANOVA for categorical variables

- No effect for gender
- No other significant interactions
- Class Instance significant over semester + class

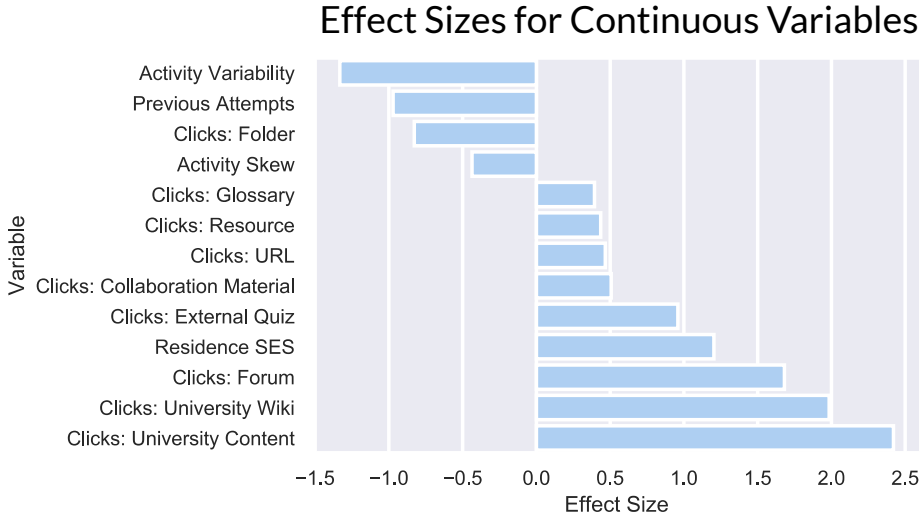
ANOVA Type I	DF	Mean Squares	F	log p
Disability	1	6,521	46.3	-11
Prior Education	2	22,253	158.1	-68
Age	1	3,451	24.5	-6
Class Instance	18	10,479	74.5	-259
Class:Prior Education	10	774	5.5	-7
Class:Age	5	684	4.9	-4
Residual	12462	141		

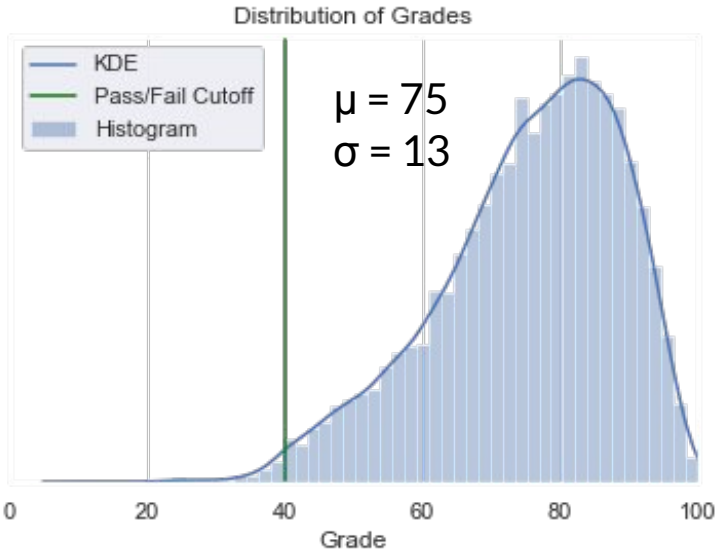
Step 2: Regularization for Continuous Variables

- 36 Continuous independent variables including alternate measures
  - 20: Clicks by type of material
  - 9: Activity Patterns
  - 3: Attention distribution statistics
  - 3: Student registration data
  - 1: Student demographic data
- Backwards Stepwise Selection ($p < 0.01$)
- Remove Multi-collinearity (VIF > 10)
- Repeat backwards stepwise selection
- Alternate model: total number of clicks instead of breakdown

Combined R^2: 21%

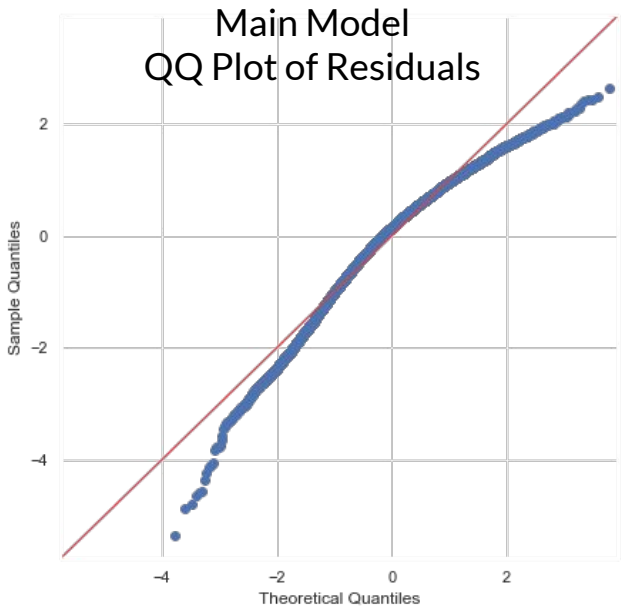
University Content, Forums, SES, and Consistency are the most substantial correlates of performance



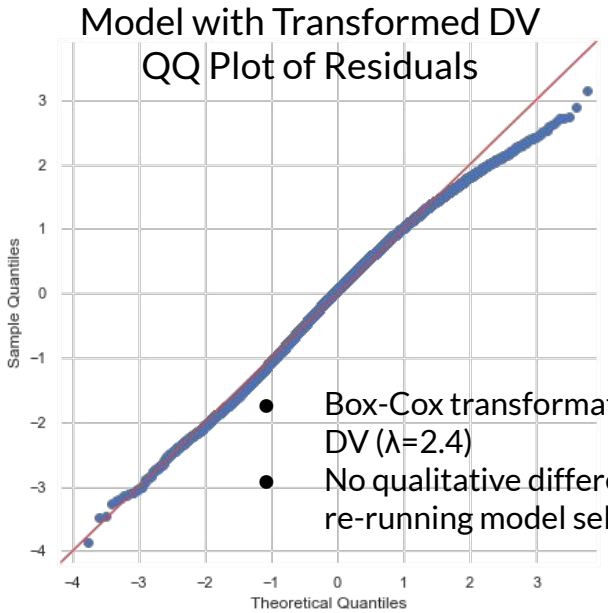


- Effect Size: Coefficient times Standard Deviation of Covariate
- Combined Marginal R^2: 10% (out of 21% with categorical variables))
- Activity Variability: Standard Deviation of Student's clicks per day on days with clicks
- Activity Skew: Skew of distribution of day on which clicks occured

Statistical Significance may be inflated due to residuals but results are robust



- D'Agostino K^2: 949 ($\sim\chi^2_2$)
- Skew = -0.97, Kurtosis = 3.5 (cf. Φ w/ 3)



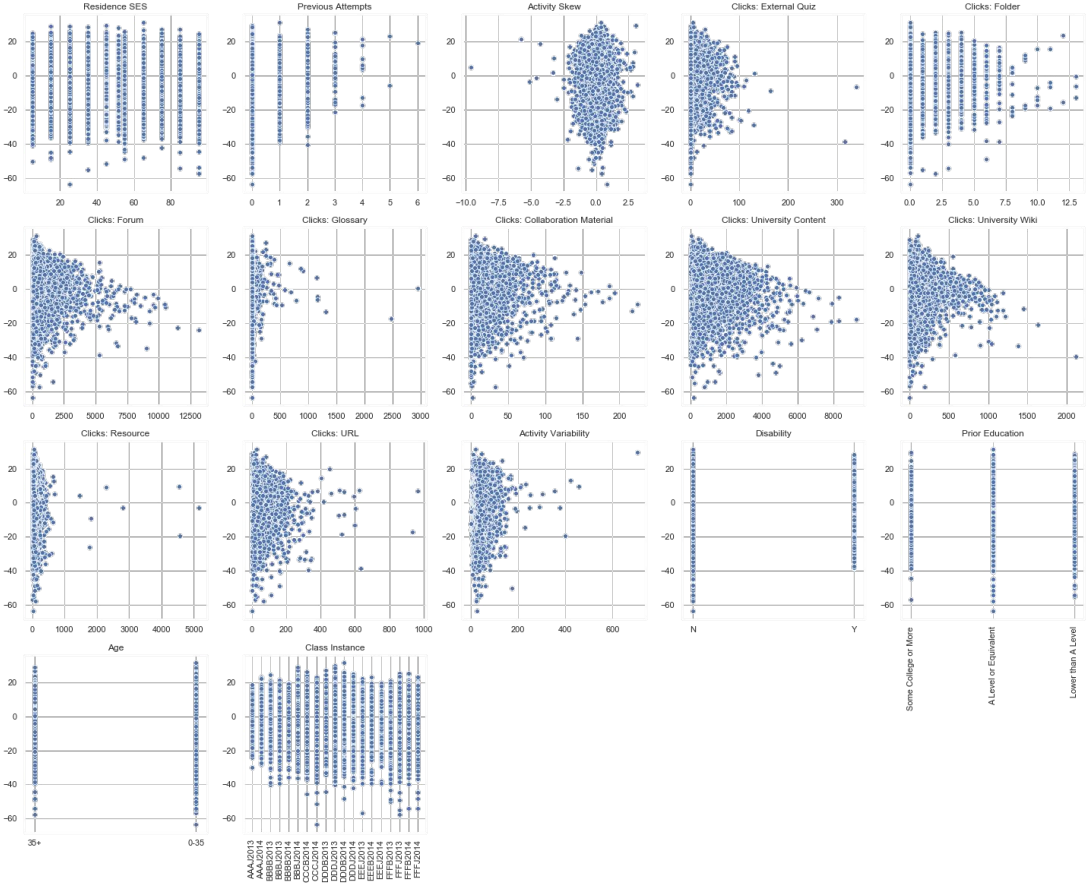
- Box-Cox transformation of the DV (λ=2.4)
- No qualitative differences when re-running model selection

- D'Agostino K^2: 203
- Skew = -0.23, Kurtosis = 2.7

Statistical significance may be inflated due to heteroskedasticity, but results are robust

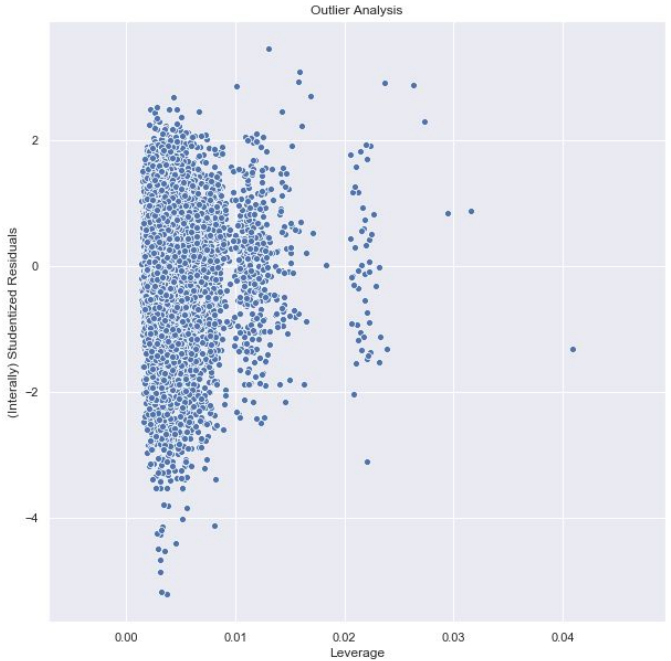
- Breusch-Pagan $\chi^2_{12} = 3181$
- Modest improvement with Box-Cox transformations of DV or IVs (results not shown)
- Heteroskedasticity Robust Standard Errors (HC0) show no qualitative differences (results not shown)
- Also: linearity appears reasonable

Residual Plots for the 17 covariates (excluding interactions)

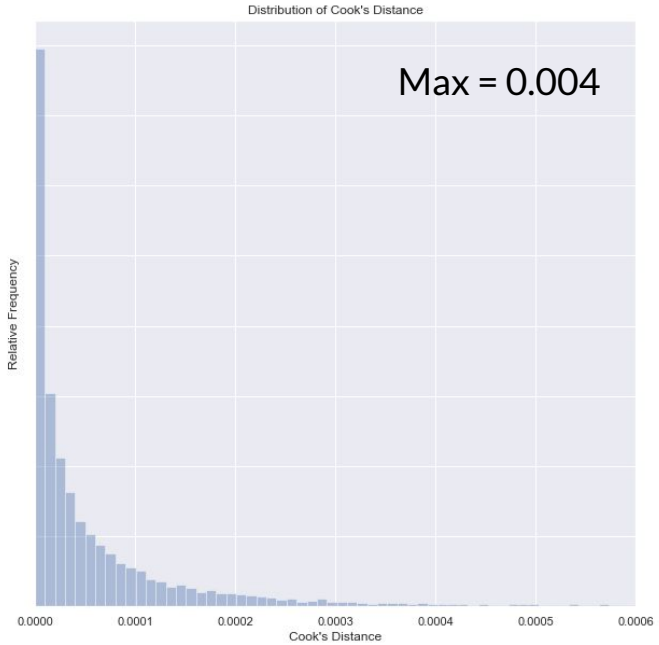


There were no outlier students

Studentized Residuals v Leverage



Cook’s Distance Histogram



Potential Future Work

- Cross-validated estimates for goodness of fit (R^2)
- Cross-validation of the model selection process
- Alternative regularization procedures (Ridge, Lasso)
- Breakdown (interactions) of coefficients by class
- Alternative specifications of the dependent variable (e.g. Final Exam score for a subset of the classes)
- Further interaction terms
- Prediction
    - Not using covariates not known in advance (e.g. class average grade)
    - Time series-analysis and cross validation
- Statistical accounting for implicit left-censorship of the data