Curve Fitting: Alumni Employment vs. Faculty Quality

Kevin Chen, Zachary Couvillion, Max Telemaque, Calvin Atkeson, Allison Anderson

Our Data Set

- 2019-2020 CWUR data set
  - Comprehensive ranking of the top 2000 world universities based on CWUR scaling
  - Quantitative approach used to determine rankings rather than subjective
- What we analyzed
  - Isolated the top 150 universities from this ranking
  - Furthermore identified and analyzed two variables from these rankings
  - Alumni Employment to CEO Positions in Forbes Global 2000 companies is compared to quality of education

Our Data Set cont.

$$r_F = C \cdot \exp\left[-k \cdot ((Y-1)-x)^2\right]$$

Y: current year
X: year award is given to faculty member
C: always set to 1 unless a faculty member holds a full time position at more than one location

The positive constant k is chosen so that $r_F = 0.01$ when $(Y - 1) - x = 99$ and $C = 1$. This gives $k = 99^{-2} \ln(100)$

$$p_E = \frac{q^2}{\max\ (n,\ 2000)}$$

q: yearly weighted average of CEO alumni
n: current number of students enrolled

Theory

The Best Approximation Theorem

Let W be a subspace of $\mathbb{R}^n$, let $\mathbf{y}$ be any vector in $\mathbb{R}^n$, and let $\hat{\mathbf{y}}$ be the orthogonal projection of $\mathbf{y}$ onto W. Then $\hat{\mathbf{y}}$ is the closest point in W to $\mathbf{y}$, in the sense that

$$\|\mathbf{y} - \hat{\mathbf{y}}\| < \|\mathbf{y} - \mathbf{v}\| \tag{3}$$

for all $\mathbf{v}$ in W distinct from $\hat{\mathbf{y}}$.

If A is $m \times n$ and $\mathbf{b}$ is in $\mathbb{R}^m$, a **least-squares solution** of $A\mathbf{x} = \mathbf{b}$ is an $\hat{\mathbf{x}}$ in $\mathbb{R}^n$ such that

$$\|\mathbf{b} - A\hat{\mathbf{x}}\| \leq \|\mathbf{b} - A\mathbf{x}\|$$

for all $\mathbf{x}$ in $\mathbb{R}^n$.

Problem

A = [1 x1; 1 x2; 1 x3; …; 1 xn]

B = [y1;y2;y3;...;yn]

The equation Ax=B is inconsistent, as the data spread is not a true linear relationship. By the Best Approximation Theorem, we can interpret the closest linear fit as the solution to Ax = B*, where B* is the orthogonal projection of B onto the column space of A.

Problem

A = [1 1; 1 2; 1 3; 1 5; …]
B = [1 10 3 19 ….]
Ax = B is inconsistent.
If we project B onto the column space of A, we must first express the column space as an orthogonal basis in order to apply the following theorem.

The Orthogonal Decomposition Theorem

Let W be a subspace of $\mathbb{R}^n$. Then each $\mathbf{y}$ in $\mathbb{R}^n$ can be written uniquely in the form

$$\mathbf{y} = \hat{\mathbf{y}} + \mathbf{z} \quad (1)$$

where $\hat{\mathbf{y}}$ is in W and $\mathbf{z}$ is in $W^{\perp}$. In fact, if $\{\mathbf{u}_1, \ldots, \mathbf{u}_p\}$ is any orthogonal basis of W, then

$$\hat{\mathbf{y}} = \frac{\mathbf{y}\cdot\mathbf{u}_1}{\mathbf{u}_1\cdot\mathbf{u}_1}\mathbf{u}_1 + \cdots + \frac{\mathbf{y}\cdot\mathbf{u}_p}{\mathbf{u}_p\cdot\mathbf{u}_p}\mathbf{u}_p \quad (2)$$

and $\mathbf{z} = \mathbf{y} - \hat{\mathbf{y}}$.

Problem

Col([1 1; 1 2; 1 3; 1 5; …]) = span{(1 1 1 1 …), (-103.4600 -102.4600 -101.4600 -99.4600 ...)} (calculated with Gram-Schmidt Process)

The projection of B onto our orthogonal column space can be calculated by adding the projections of B onto each basis vector of the orthogonal column space.

B* = [209.1569 210.4138 211.6706 214.1844 …]

Problem

$Ax = B^*$ → This is solvable since B^* is in the column space of A by definition of orthogonal projection.

We can let Matlab row reduce the augmented matrix $[A\ B^*]$ to arrive at

$X = [207.9000, 1.2569]$

$Y = 1.2569x + 207.9000$

Theory

Theorem 14

Let A be an $m \times n$ matrix. The following statements are logically equivalent:

a. The equation $A\mathbf{x} = \mathbf{b}$ has a unique least-squares solution for each $\mathbf{b}$ in $\mathbb{R}^m$.

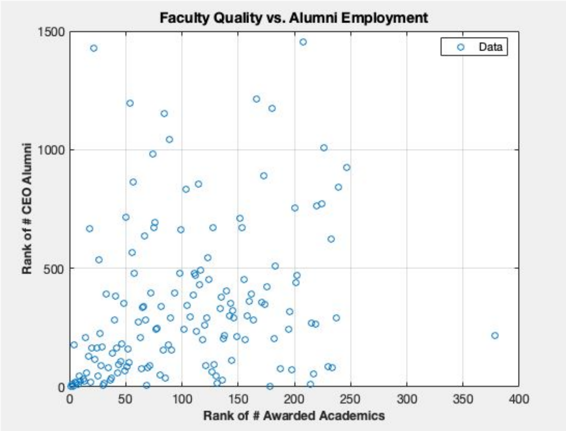
b. The columns of A are linearly independent.

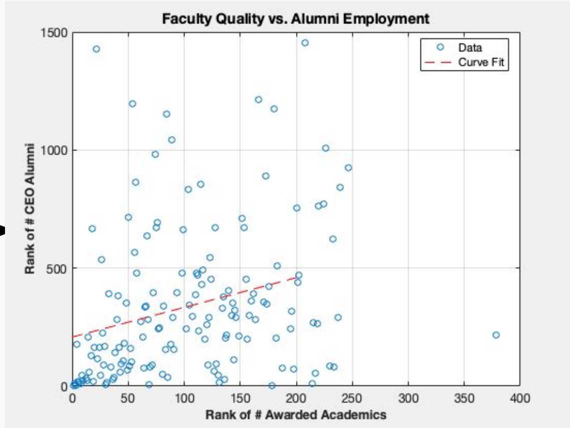
c. The matrix A^TA is invertible.

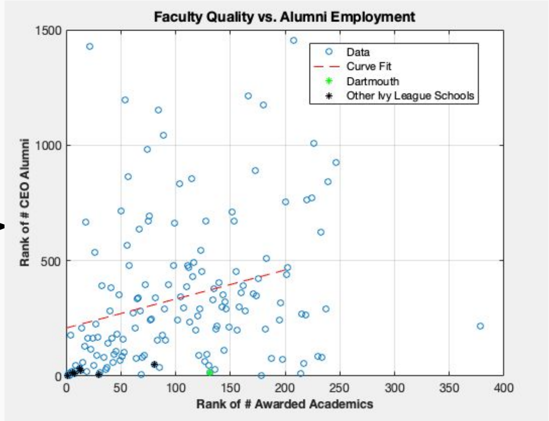
When these statements are true, the least-squares solution $\hat{\mathbf{x}}$ is given by

$$\hat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b} \tag{4}$$

Curve Fitting

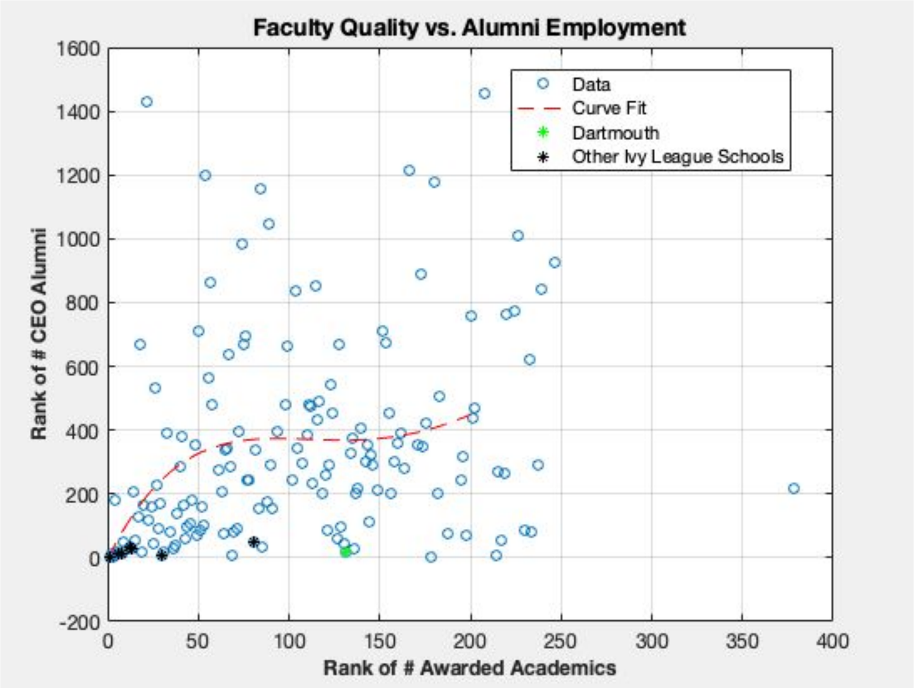






B = [1.2569, 207.9000]

Higher Degree Polynomial

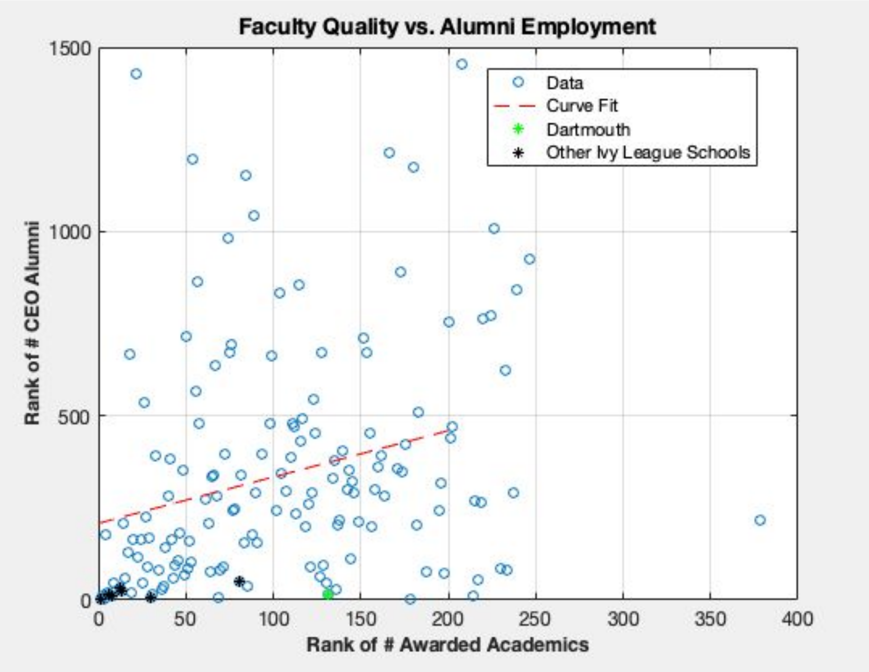


Polynomial: Degree 4

$$B = \begin{bmatrix} -0.0000 \\ 0.0006 \\ -0.1264 \\ 11.6740 \\ -5.8221 \end{bmatrix}$$

Data Significance

- Slight positive correlation between the factors
- Only 6% of data points lie on the line
- Dartmouth Residual Distance: 356
- Outliers
  - University of Helsinki (226, 1007)
  - UC Irvine (22, 1427)



Works Cited

Lay, David C, Lay, Steven R, McDonald, Judi J, et al. *Linear Algebra and Its Applications*. 5th ed., Pearson Education, Inc, 2016.

Mahassen, N. A quantitative approach to world university rankings. *Center for World University Rankings* (2019).

"MATLAB Project: Least Squares Solutions and Curve Fitting ." *Media.pearsoncmg.com*, media.pearsoncmg.com/aw/aw_lay_linearalg_updated_cw_5/student/projects/Matlab/pdf/27.lsq.pdf.