Appendix: Simulation

Hongyu Mou

January 23, 2024

1 Extrapolation Model

- Extrapolation 1 based on GPA: want to have Y_{GPA} for all scores[1].

$$\begin{aligned} \boldsymbol{Y_{GPA,b}} &= \underbrace{\alpha_1 + \beta_1 \times GPA(\%) + \gamma \times b}_{\hat{\boldsymbol{Y}}_{\boldsymbol{GPA,b}}} + \epsilon_1 \\ &= 10.01 + 0.09 \times GPA(\%) + 0.2 \times b + \epsilon_1 \\ &\text{where } \boldsymbol{\epsilon_1 \sim N(0, 1.002)} \end{aligned} \tag{1}$$

  use the estimated relationship (estimated for $GPA >$ cutoff) to extrapolate Y_{GPA} for $GPA <$ cutoff.

- Extrapolation 2 based on Quota 2 Score: want to have $Y_{s,GPA,Q2}$ for all scores. Assuming that s is continuous even below the cutoff, we can again assume a linear relationship in quota 2 score and GPA so that:

$$\begin{aligned} \boldsymbol{Y_{GPA,s,b,Q2>0}} &= \underbrace{\alpha_2 + \beta_2 \times GPA(\%) + \gamma_2 \times s(\%) + \gamma_3 \times b}_{\hat{\boldsymbol{Y}}_{\boldsymbol{GPA,s,b,Q2>0}}} + \epsilon_2 \\ &= 9.12 - 0.04 \times GPA(\%) + 2.33 \times s(\%) + 0.3 \times b + \epsilon_2 \\ &\text{where } \boldsymbol{\epsilon_2 \sim N(0, 0.703)} \end{aligned} \tag{2}$$

  use the estimated relationship (estimated for $s >$ cutoff) to extrapolate $Y_{s,GPA,Q2}$ for $s <$ cutoff.

- **Test if ϵ_1 and ϵ_2 are normally distributed:** Figure 1 show distributions and Kolmogorov-Smirnov tests of residuals ϵ_1 and ϵ_2:

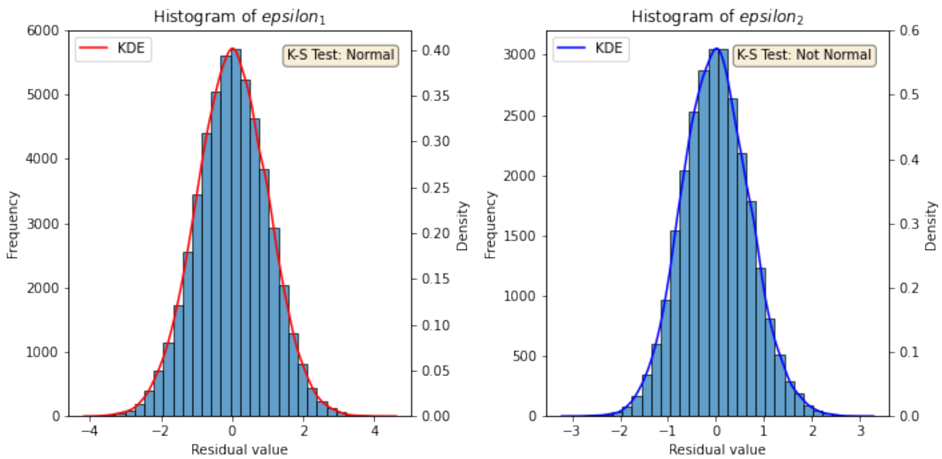


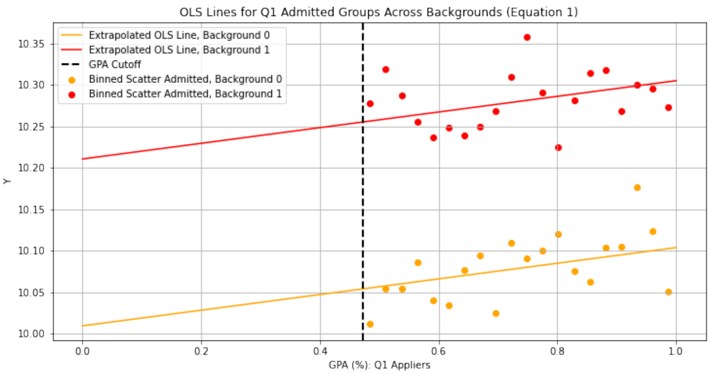
Figure 1: Distribution of ϵ_1 and ϵ_2 from Equations (1) and (2)

Note: ϵ_1 comes from the OLS fitting as Equations (1): $Y_{GPA,b} = \alpha_1 + \beta_1 \times GPA(\%) + \gamma \times b + \epsilon_1$. ϵ_2 comes from the OLS fitting as Equations (2): $Y_{GPA,s,b,Q2>0} = \alpha_2 + \beta_2 \times GPA(\%) + \gamma_2 \times s(\%) + \gamma_3 \times b + \epsilon_2$.

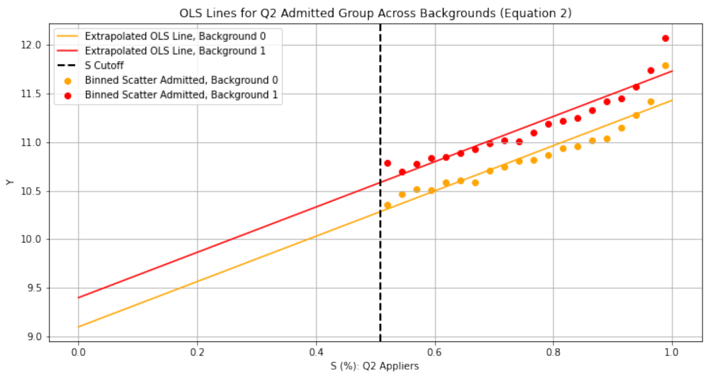
[1] Since in the empirical data we observe that the higher the GPA, the larger the earning potential.

1.1 Extrapolation Plots

Note that in Figure 2a and 2b, with real data, we would not be able to fit the below-admission cutoff lines as we cannot observe the outcome of the not-admitted students.



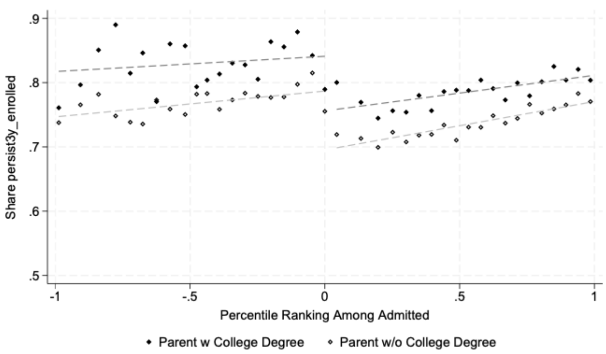
(a) OLS Regression Lines of Equation (1)



(b) OLS Regression Lines of Equation (2) using $Q2 > 0$ Sub-sample

Figure 2: Extrapolation Plots for Each Background from Equation (1) and (2)

Note: For both figures, we first fit the OLS line using the observed above-cutoff sample that is admitted by Q1 or Q2. Next, we extrapolate the fitted line to the below-cutoff sample that is not observed in real data.



(b) Persistence, Year 3

Figure 3: Pooled Fitted lines (Figure A9b) for Each Background in Real Data

Note: This figure differs the above two in the following sense: (1) GPA (%) and Score (%) are calculated using the **admitted** sample, instead of the **applied** sample. (2) The below-zero part plots the observed relationship for those who were admitted to Q2, and the above-zero part plots the observed relationship for those who were admitted to Q1.

2 Simulation Model[2]

Whole Sample: $N = 100000$ (Applied_Q1)

Parameters: b, GPA, and ϵ

Edu Background: $b \sim \text{Bernoulli}(\text{p} = 0.5)$

$\text{GPA} \sim \mathcal{N}(8.8, \text{sd} = 1)$ if $b = 0$, censor at 6 and 13

$\text{GPA} \sim \mathcal{N}(9.2, \text{sd} = 1)$ if $b = 1$, censor at 6 and 13

$$\epsilon_Y, \epsilon_{Q2}, \epsilon_S \sim \mathcal{N}(0, \Sigma) \text{ where } \Sigma = \begin{bmatrix} 1 & 0.5 & 0.6 \\ 0.5 & 1 & 0 \\ 0.6 & 0 & 1 \end{bmatrix}$$

$\alpha_Y = 10, \alpha_Y^1 = 0.1, \alpha_Y^2 = 0.0, \alpha_Y^3 = 0.2$

$\beta_{Q2} = 0, \beta_{Q2}^1 = 0.2, \beta_{Q2}^2 = -0.05, \beta_{Q2}^3 = 0.1$

$\gamma_S = 0, \gamma_S^1 = 0.15, \gamma_S^2 = 0.0, \gamma_S^3 = -0.2$

Simulate: $Y, Q2,$ and S

$$Y = \alpha_Y + \alpha_Y^1 \times GPA(\%) + \alpha_Y^2 \times GPA(\%)^2 + \alpha_Y^3 \times b + \epsilon_Y$$
$$Q2 = \beta_{Q2} + \beta_{Q2}^1 \times GPA(\%) + \beta_{Q2}^2 \times GPA(\%)^2 + \beta_{Q2}^3 \times b + \epsilon_{Q2}$$
$$S = \gamma_S + \gamma_S^1 \times GPA(\%) + \gamma_S^2 \times GPA(\%)^2 + \gamma_S^3 \times b + \epsilon_S$$

GPA Cutoff: GPA_cutoff = 8.93

Admission:

$$\text{Admitted\_Q1} = \begin{cases} 1 & \text{if GPA >= GPA\_cutoff} \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Applied\_Q2} = \begin{cases} 1 & \text{if } Q2 > 0 \\ 0 & \text{otherwise} \end{cases}$$

$$\text{Admitted\_Q2} = \begin{cases} 1 & \text{if } S > 0 \text{ and Applied\_Q2} = 1 \\ 0 & \text{otherwise} \end{cases}$$

Objects of Interest:

For each background $b \in \{0, 1\}$:

$$P(Q2 > 0 \mid GPA(\%), Y, B = b) = \frac{P(Y \mid Q2 > 0, GPA(\%), B = b) \times P(Q2 > 0 \mid GPA(\%), B = b)}{P(Y \mid GPA(\%), B = b)}$$

$$P(S > 0 \mid GPA(\%), Y, Q2 > 0, B = b) = \frac{P(Y \mid S > 0, GPA(\%), Q2 > 0, B = b) \times P(S > 0 \mid GPA(\%), Q2 > 0, B = b)}{P(Y \mid GPA(\%), Q2 > 0, B = b)}$$

$$P(S > 0, Q2 > 0 \mid GPA(\%), Y, B = b) = P(S > 0 \mid GPA(\%), Y, Q2 > 0, B = b) \times P(Q2 > 0 \mid GPA(\%), Y, B = b)$$

[2]See Appendix for the descriptive statistics.

3 $P(Q2 > 0 \mid GPA(\%), Y, B = b)$

The first Bayes Rule:

$$P(Q2 > 0 \mid GPA(\%), Y, B = b) = \frac{P(Y \mid Q2 > 0, GPA(\%), B = b) \times P(Q2 > 0 \mid GPA(\%), B = b)}{P(Y \mid GPA(\%), B = b)} \tag{3}$$

Since in real data we only observe Y for those who were admitted, in the feasible $GPA >= GPA_{cutoff}$ sub-sample, use local approximation instead of Bayes Rule (3) for each object of interest if $N_{GPA,Y,B} \geq \bar{N}_{GPA,Y,B}$, where $N_{GPA,Y,B}$ is the size of each $(GPA(\%), Y)$ cell for a specific background B.

1. Feasible: $\bar{N}_{GPA,Y,B} > \infty$, i.e., all from extrapolation using Bayes Rule (3)
2. Feasible: $\bar{N}_{GPA,Y,B} = 0$, i.e., local approximation without using Bayes Rule (3), but for the $GPA >= GPA_{cutoff}$ sample, since in real data we only observe Y for those who were admitted.
3. Feasible: $\bar{N}_{GPA,Y,B} = 10$, i.e., local approximation for the $GPA >= GPA_{cutoff}$ sample if $N_{GPA,Y,B} \geq 10$; Bayes Rule (3) from extrapolation if $N_{GPA,Y,B} < 10$ or $GPA < GPA_{cutoff}$.
4. Feasible: $\bar{N}_{GPA,Y,B} = 50$, i.e., local approximation for the $GPA >= GPA_{cutoff}$ sample if $N_{GPA,Y,B} \geq 50$; Bayes Rule (3) from extrapolation if $N_{GPA,Y,B} < 50$ or $GPA < GPA_{cutoff}$.

3.1 $P(Y \mid Q2 > 0, GPA(\%), B = b)$

- **From Extrapolation**: Equation (2) and S Integration. **(Figure 5)**

  Compute

$$\begin{aligned} & P[Y \in Y_{\text{bin}} \,|GPA(\%) \in GPA(\%)_{\text{bin}}\,, Q2 > 0, B = b] \\ = & \sum_s P\left[Y \in Y_{\text{bin}} \,|GPA(\%) \in GPA(\%)_{\text{bin}}\,, Q2 > 0, s(\%) \in S(\%)_{\text{bin}}\,, B = b\right] \\ & * P\left[s(\%) \in S(\%)_{\text{bin}} \,|GPA(\%) \in GPA(\%)_{\text{bin}}\,, Q2 > 0, B = b\right] \end{aligned} \tag{4}$$

  where

  - $P\left[Y \in Y_{\text{bin}} \,|GPA(\%) \in GPA(\%)_{\text{bin}}\,, Q2 > 0, s(\%) \in S(\%)_{\text{bin}}\,, B = b\right]$: for each GPA(%) and S(%) grid, assume

$$Y_{GPA,s,b,Q2>0} \sim N(\hat{Y}_{GPA,s,b,Q2>0}, std(\epsilon_2)) \tag{5}$$

    then

$$P(Y \in Y_{\text{bin}} \mid GPA(\%) \in GPA(\%)_{\text{bin}}\,, Q2 > 0, s(\%) \in S(\%)_{\text{bin}}) = \Phi\left(\frac{\overline{Y}_{\text{bin}} - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{\underline{Y}_{\text{bin}} - \mu_2}{\sigma_2}\right) \tag{6}$$

    * $\boldsymbol{\mu_2}$: local predicted mean $(\hat{Y}_{GPA,s,b,Q2>0})$ from Equation (2) in this GPA(%) and S(%) grid
    * $\boldsymbol{\sigma_2}$: global standard deviation of ϵ_2 estimated from Equation (2) using the observed sample above the Score cutoff.
  - $P\left[s(\%) \in S(\%)_{\text{bin}} \,|GPA(\%) \in GPA(\%)_{\text{bin}}\,, Q2 > 0, B = b\right]$: probability mass function (PMF) from local approximation, shown in Figure 4.

$$\begin{aligned} & P\left[s(\%) \in S(\%)_{\text{bin}} \mid GPA(\%) \in GPA(\%)_{\text{bin}}, Q2 > 0, B = b\right] \\ = & \frac{\text{Number of students with } s(\%) \text{ in } S(\%)_{\text{bin}} \text{ within the GPA(\%) bin and Background group who have } Q2 > 0}{\text{Total Number of students in the GPA(\%) bin and Background group who have } Q2 > 0} \end{aligned} \tag{7}$$

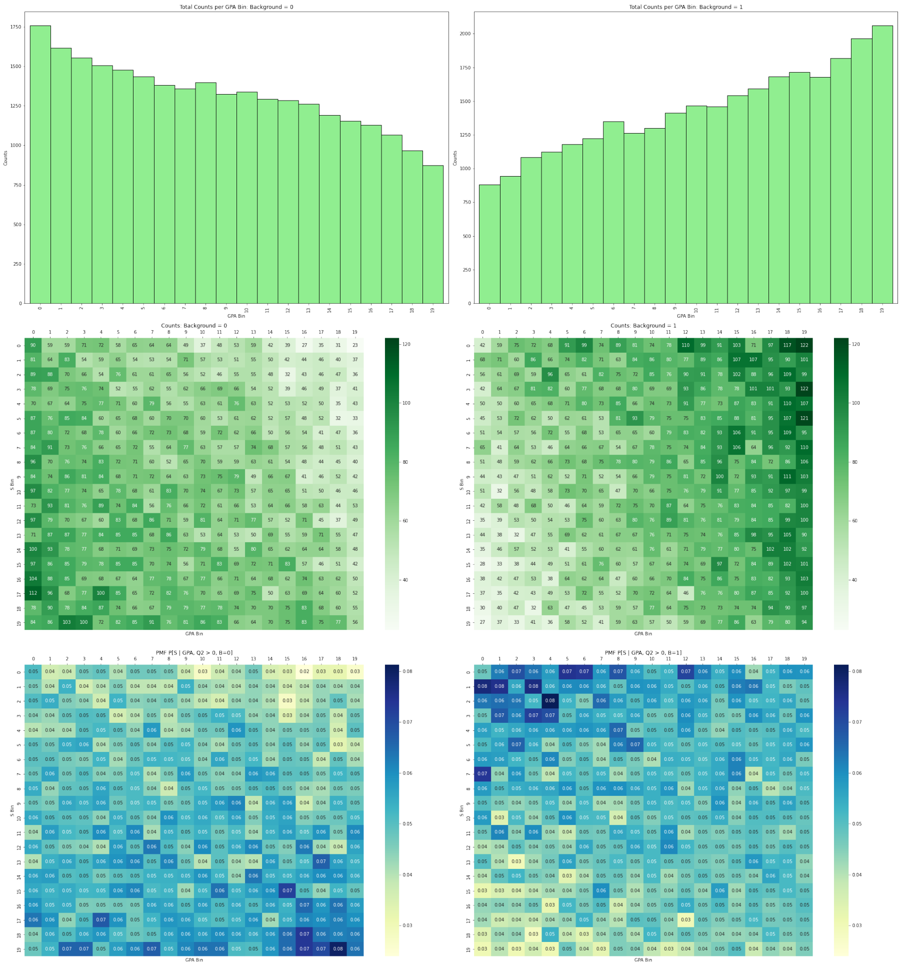


Figure 4: PMF $P[S(\%)|GPA(\%), Q2 > 0, B]$ from Local Approximation

Note: The last row of this figure shows the probability of a student's score $s(\%)$ falling within a certain score bin, given that their GPA is within a specified GPA bin, they have applied to Q2 (indicated by $Q2 > 0$), and belong to a specific background group $B = b$. The probability is calculated using $P\left[s(\%) \in S(\%)_{\text{bin}} \mid GPA(\%) \in GPA(\%)_{\text{bin}}, Q2 > 0, B = b\right] = \frac{\text{Number of students with } s(\%) \text{ in } S(\%)_{\text{bin}} \text{ within the GPA(\%) bin and Background group who have } Q2>0}{\text{Total Number of students in the GPA(\%) bin and Background group who have } Q2>0}$, where the denominator is shown in the first row of this figure, and the numerator is shown in the middle row of the figure. Note that we do not have cells where we cannot observe GPA (and thus GPA percentile) by $Q2 > 0$ and b, since the population in this practice are $Q1$ appliers who should have their own GPA available in order to apply for Quota 1. Thus, here in the last row, we are actually presenting, for each background, the probabilities of $S(\%)$ conditional on $GPA(\%)$ for those who have applied to both Quota 1 (whole population) and Quota 2 ($Q2 > 0$), which should be observed directly from the real data so we categorize it into "local approximation".

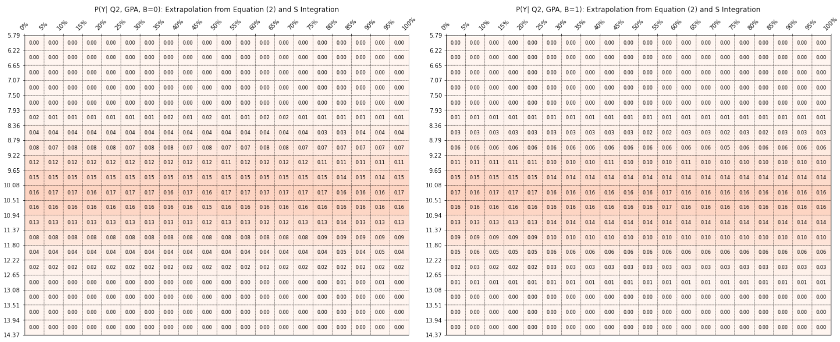


Figure 5: $P(Y|Q2>0, GPA, B)$ from Extrapolation

Note: This figure shows the probability calculated using the integration over S: $P[Y \in Y_{\text{bin}} | GPA(\%) \in GPA(\%)_{\text{bin}}, Q2 > 0, B = b] = \sum_s P[Y \in Y_{\text{bin}} | GPA(\%) \in GPA(\%)_{\text{bin}}, Q2 > 0, s(\%) \in S(\%)_{\text{bin}}, B = b] * P[s(\%) \in S(\%)_{\text{bin}} | GPA(\%) \in GPA(\%)_{\text{bin}}, Q2 > 0, B = b]$, where the first term on the right-hand side inside the summation, $P[Y \in Y_{\text{bin}} | GPA(\%) \in GPA(\%)_{\text{bin}}, Q2 > 0, s(\%) \in S(\%)_{\text{bin}}, B = b]$, is extrapolated using predicted value of Y from equation (2), standard deviation of ϵ_2, and normal distribution: for each GPA(%) and S(%) grid, assume $Y_{GPA,s,b,Q2>0} \sim N(\hat{Y}_{GPA,s,b,Q2>0}, std(\epsilon_2))$ then $P(Y \in Y_{\text{bin}} \mid GPA(\%) \in GPA(\%)_{\text{bin}}, Q2 > 0, s(\%) \in S(\%)_{\text{bin}}) = \Phi\left(\frac{\overline{Y}_{\text{bin}} - \mu_2}{\sigma_2}\right) - \Phi\left(\frac{\underline{Y}_{\text{bin}} - \mu_2}{\sigma_2}\right)$, where μ_2 is the local predicted mean ($\hat{Y}_{GPA,s,b,Q2>0}$) from Equation (2) in this GPA(%) and S(%) grid; σ_2 is the global standard deviation of ϵ_2 estimated from Equation (2) using the observed sample above the Score cutoff. $P[s(\%) \in S(\%)_{\text{bin}} | GPA(\%) \in GPA(\%)_{\text{bin}}, Q2 > 0, B = b]$ comes from local approximation using PMF (more detail in the footnote of Figure 4).

3.2 $P(Y \mid GPA(\%), B = b)$

- **From Extrapolation**: Equation (1). **(Figure 6)**

  for each GPA(%) bin, assume

$$Y_{GPA,b} \sim N(\hat{Y}_{GPA,b}, std(\epsilon_2)) \tag{8}$$

  then

$$P(Y \in Y_{\text{bin}} \mid GPA(\%) \in GPA(\%)_{\text{bin}}, B = b) = \Phi\left(\frac{\overline{Y}_{\text{bin}} - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{\underline{Y}_{\text{bin}} - \mu_1}{\sigma_1}\right) \tag{9}$$

  - $\boldsymbol{\mu_1}$: local predicted mean ($\hat{Y}_{GPA,b}$) from Equation (1) in this GPA(%) bin
  - $\boldsymbol{\sigma_1}$: global standard deviation of ϵ_1 estimated from Equation (1) using the observed sample above the GPA cutoff.

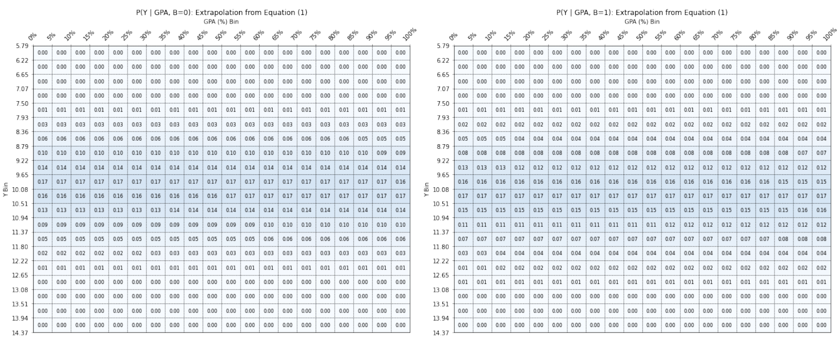


Figure 6: $P(Y|GPA, B)$ from Extrapolation

Note: This figure shows the probability extrapolated using predicted value of Y from equation (1), standard deviation of ϵ_1, and normal distribution: for each GPA(%) bin, assume $Y_{GPA,b} \sim N(\hat{Y}_{GPA,b}, std(\epsilon_2))$ then $P(Y \in Y_{\text{bin}} \mid GPA(\%) \in GPA(\%)_{\text{bin}}, B = b) = \Phi\left(\frac{\overline{Y}_{\text{bin}} - \mu_1}{\sigma_1}\right) - \Phi\left(\frac{\underline{Y}_{\text{bin}} - \mu_1}{\sigma_1}\right)$ where μ_1 is the local predicted mean ($\hat{Y}_{GPA,b}$) from Equation (1) in this GPA(%) bin; σ_1 is the global standard deviation of ϵ_1 estimated from Equation (1) using the observed sample above the GPA cutoff.

3.3 $P(Q2 > 0 \mid GPA(\%), B = b)$

The probability that a student applies to Q2 given their GPA and Background. See Figure 7.

$$P(Q2 > 0 \mid GPA(\%), B = b) = \frac{\text{Number of students who applied to Q2 in the GPA (\%) bin and Background group}}{\text{Total Number of students in the GPA (\%) bin and Background group}} \tag{10}$$

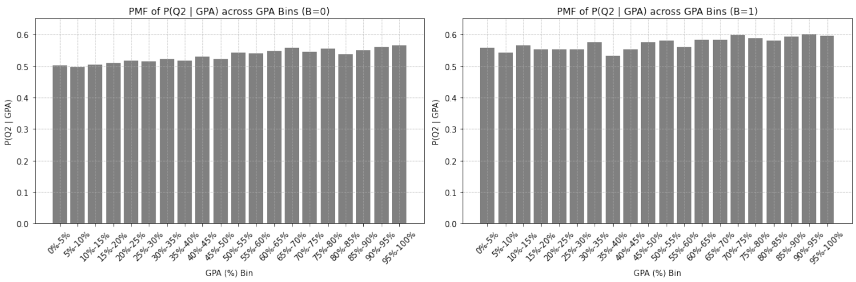
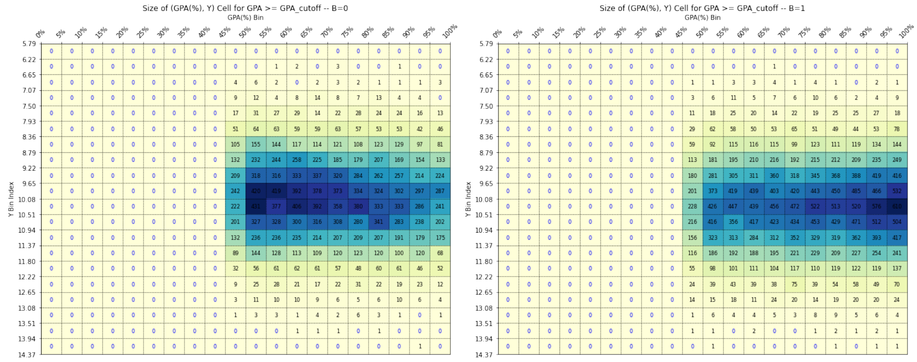


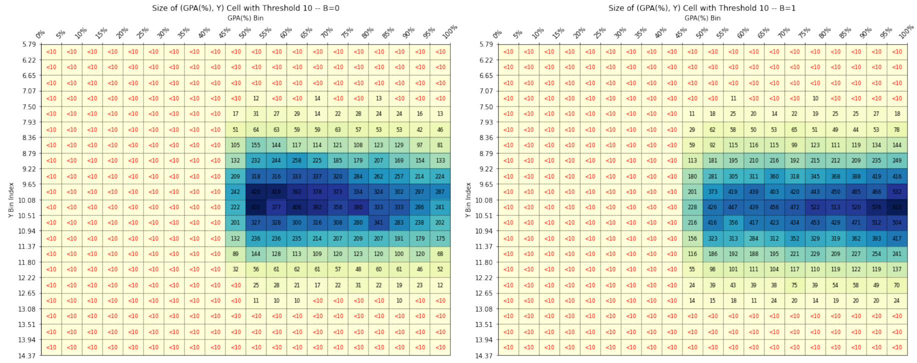
Figure 7: PMF $P(Q2 > 0 \mid GPA(\%), B)$ from Local Approximation

Note: The probability is calculated from the local approximation of PMF: $P(Q2 > 0 \mid GPA(\%), B = b) = \frac{\text{Number of students who applied to Q2 in the GPA (\%) bin and Background group}}{\text{Total Number of students in the GPA (\%) bin and Background group}}$.

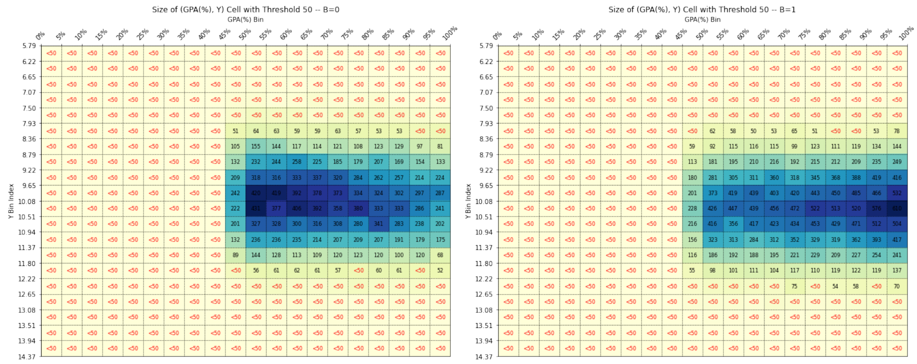
3.4 Cell Size for GPA vs. Y



(a) $GPA >= GPA_{cutoff}$: applied to Figure 9a



(b) $\bar{N}_{GPA,Y,B} = 10$ and $GPA >= GPA_{cutoff}$: applied to Figure 9c

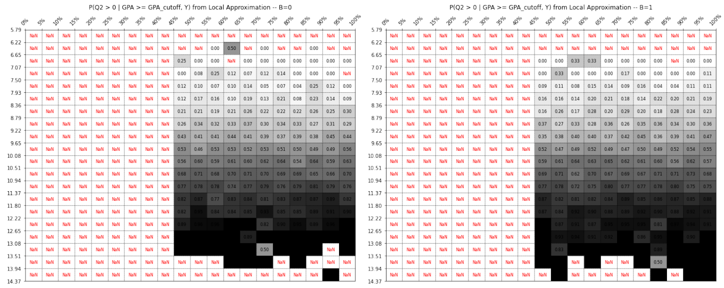


(c) $\bar{N}_{GPA,Y,B} = 50$ and $GPA >= GPA_{cutoff}$: applied to Figure 9d

Figure 8: Size of $(GPA(\%), Y)$ Cell

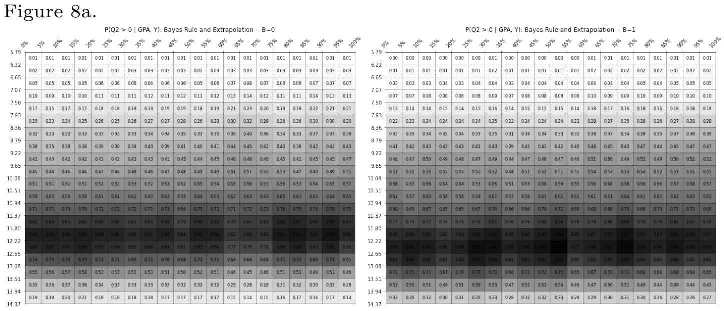
Note: These figures report the number of observations in each $GPA(\%)$ and Y cell, where we also indicate whether the cell size is less than 10 or 50. Figure 8a is applied to Figure 9a where we only consider the local approximation using the observed data. Figure 8b (Figure 8c) is applied to Figure 9c (Figure 9d) as threshold rule to decide whether using local approximation or extrapolation: cells with red letters here are using extrapolation model in Figure 9c (Figure 9d), and cells with black letters are using local approximation in Figure 9c (Figure 9d).

3.5 $P(Q2 > 0 \mid GPA(\%), Y, B = b)$



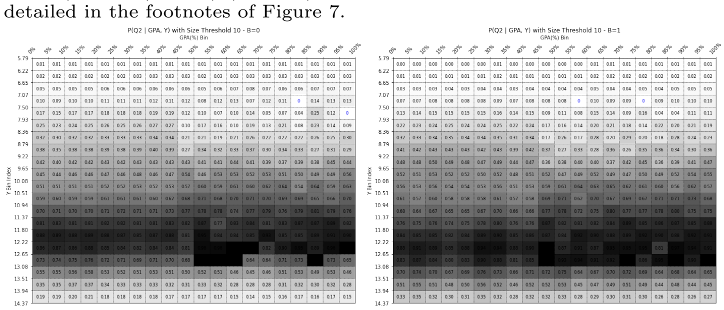
(a) Local Approximation: $\bar{N}_{GPA,Y,B} = 0$ (Cell Size in Figure 8a)

Note: This figure shows the observed $P(Q2 > 0 \mid GPA(\%), Y, B = b)$ from local approximation in each $GPA(\%)$ and Y cell, without using any information from the extrapolation models or Bayes Rules. Note that we can only plot for the $GPA >= GPA_{cutoff}$ sample, since in real data we only observe Y for those who were admitted. The corresponding cell size counts are showed in Figure 8a.



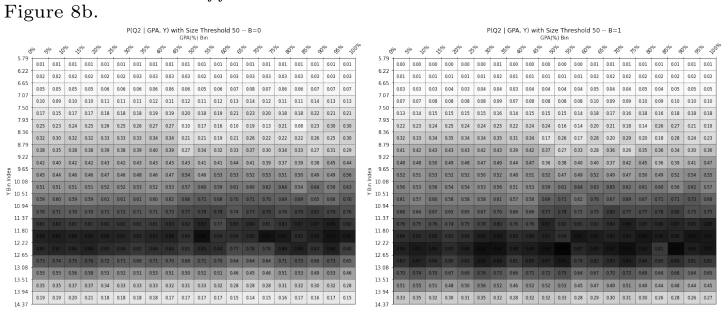
(b) Extrapolation and Bayes Rule (3): $\bar{N}_{GPA,Y,B} > \infty$

Note: The figure shows the extrapolated $P(Q2 > 0 \mid GPA(\%), Y, B = b)$ that comes from Bayes Rule (3): $P(Q2 > 0 \mid GPA(\%), Y, B = b) = \frac{P(Y \mid Q2>0, GPA(\%), B=b) \times P(Q2>0 \mid GPA(\%), B=b)}{P(Y \mid GPA(\%), B=b)}$, where the blue terms $P(Y \mid Q2 > 0, GPA(\%), B = b)$ and $P(Y \mid GPA(\%), B = b)$ are constructed using extrapolation models as detailed in the footnotes of Figure 5 and Figure 6, and $P(Q2 > 0 \mid GPA(\%), B = b)$ comes from local approximation of PMF as detailed in the footnotes of Figure 7.



(c) $\bar{N}_{GPA,Y,B} = 10$ (Cell Size in Figure 8b)

Note: $\bar{N}_{GPA,Y,B} = 10$, i.e., local approximation for the $GPA >= GPA_{cutoff}$ sample if $N_{GPA,Y,B} \geq 10$; extrapolation and Bayes Rule (3) if $N_{GPA,Y,B} < 10$ or $GPA < GPA_{cutoff}$. The corresponding cell size counts are showed in Figure 8b.



(d) $\bar{N}_{GPA,Y,B} = 50$ (Cell Size in Figure 8c)

Note: $\bar{N}_{GPA,Y,B} = 50$, i.e., use local approximation for the $GPA >= GPA_{cutoff}$ sample if $N_{GPA,Y,B} \geq 50$; use extrapolation and Bayes Rule (3) if $N_{GPA,Y,B} < 50$ or $GPA < GPA_{cutoff}$. The corresponding cell size counts are showed in Figure 8c.

Figure 9: $P(Q2 > 0 \mid GPA(\%), Y, B)$

Note: Since in real data we can observe Y for those who were admitted, in the feasible $GPA >= GPA_{cutoff}$ sub-sample, we use local approximation instead of extrapolation and Bayes Rule (3) for each object of interest if $N_{GPA,Y,B} \geq \bar{N}_{GPA,Y,B}$, where $N_{GPA,Y,B}$ is the size of each $(GPA(\%), Y)$ cell for a specific background B.

4 $P(S > 0 \mid GPA(\%), Y, Q2 > 0, B = b)$

The second Bayes Rule:

$$P(S > 0 \mid GPA(\%), Y, Q2 > 0, B = b) = \frac{P(Y \mid S > 0, GPA(\%), Q2 > 0, B = b) \times P(S > 0 \mid GPA(\%), Q2 > 0, B = b)}{P(Y \mid GPA(\%), Q2 > 0, B = b)} \tag{11}$$

So now, in the feasible $GPA >= GPA_{cutoff}$ & $Q2 > 0$ sub-sample, we change to use local approximation for each object of interest if $N_{GPA,Y,Q2>0} \geq \bar{N}_{GPA,Y,B,Q2>0}$, where $N_{GPA,Y,B,Q2>0}$ is the size of each $(GPA(\%), Y)$ cell for a specific background B.

1. Feasible: $\bar{N}_{GPA,Y,B,Q2>0} > \infty$, i.e., all from extrapolation using Bayes Rule (11)
2. Feasible: $\bar{N}_{GPA,Y,B,Q2>0} = 0$, i.e., local approximation without using Bayes Rule (11), but for the $GPA >= GPA_{cutoff}$ sample, since in real data we only observe Y for those who were admitted
3. Feasible: $\bar{N}_{GPA,Y,B,Q2>0} = 100$, i.e., local approximation for the feasible $GPA >= GPA_{cutoff}$ & $Q2 > 0$ sub-sample if $N_{GPA,Y,B,Q2>0} \geq 100$, Bayes Rule (11) from extrapolation if $N_{GPA,Y,B,Q2>0} < 100$ or $GPA < GPA_{cutoff}$ & $Q2 > 0$

4.1 Feasible PMF from Local Approximation (Figure 12a)

The empirical PMF approach counts the number of occurrences where $S > 0$ and divides it by the total number of occurrences in each (GPA, Y) cell for the $Q2 > 0$ and $B = b$ sub-sample. See Figure 12a.

4.2 Feasible from Extrapolation by Bayes Rule (Figure 12b)

Remember Bayes Rule (11):

1. $P(Y \mid S > 0, GPA(\%), Q2 > 0, B = b)$ from Extrapolation: Figure 10a that comes from Equation (2) using the $S > 0$ and $Q2 > 0$ sample:

$$\begin{aligned}
\boldsymbol{Y_{GPA,s,b,Q2>0}} &= \underbrace{\alpha_2 + \beta_2 \times GPA(\%) + \gamma_2 \times s(\%) + \gamma_3 \times b}_{\boldsymbol{\hat{Y}_{GPA,s,b,Q2>0}}} + \epsilon_2 \\
&= 9.12 - 0.04 \times GPA(\%) + 2.33 \times s(\%) + 0.3 \times b + \epsilon_2 \\
&\text{where } \boldsymbol{\epsilon_2 \sim N(0, 0.703)}
\end{aligned}$$

In the $S > 0$ and $Q2 > 0$ subsample, for each GPA (%) bin, we assume

$$Y_{GPA,S>0,b,Q2>0} \sim N\left(\hat{Y}_{GPA,S>0,b,Q2>0}, \text{std}(\epsilon_2)\right) \tag{12}$$

then

$$P\left(Y \in Y_{\text{bin}} \mid GPA(\%) \in GPA(\%)_{\text{bin}}, S > 0, Q2 > 0\right) = \Phi\left(\frac{\bar{Y}_{\text{bin}} - \mu_2'}{\sigma_2}\right) - \Phi\left(\frac{\underline{Y}_{\text{bin}} - \mu_2'}{\sigma_2}\right) \tag{13}$$

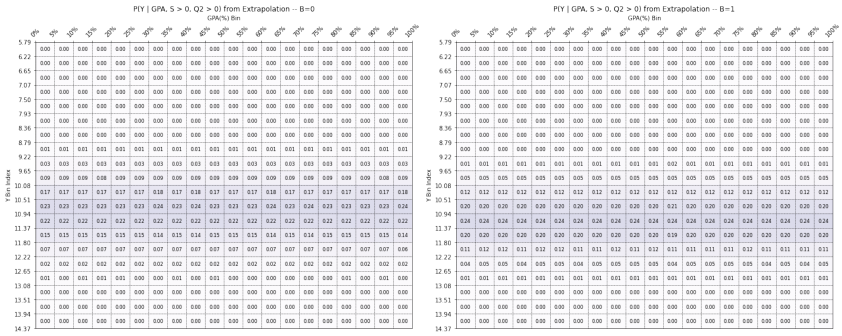
where

- $\boldsymbol{\mu_2'}$: averaged local predicted mean: $\hat{Y}_{GPA,S>0,b,Q2>0}$, which is the average of $\hat{Y}_{GPA,S,b,Q2>0}$ from Equation (2) for each $GPA(\%)$ bin.
- $\boldsymbol{\sigma_2}$: global standard deviation of ϵ_2 estimated from Equation (2) using the observed sample above the Score cutoff.

2. $P(S > 0 \mid GPA(\%), Q2 > 0, B = b)$ from LPM: linear probability model (LPM) applied to the $Q2 > 0$ sub-sample for each background $B = b$. See Figure 10b.

$$1(S > 0, B = 0) = \underbrace{\alpha_{3,B=0} + \beta_{3,B=0} \times GPA(\%)}_{P(S>0|GPA,Q2>0,B=0) \textbf{ from LPM}} + \epsilon_{3,B=0} = 0.51 + 0.05 \times GPA(\%) + \epsilon_{3,B=0} \tag{14}$$

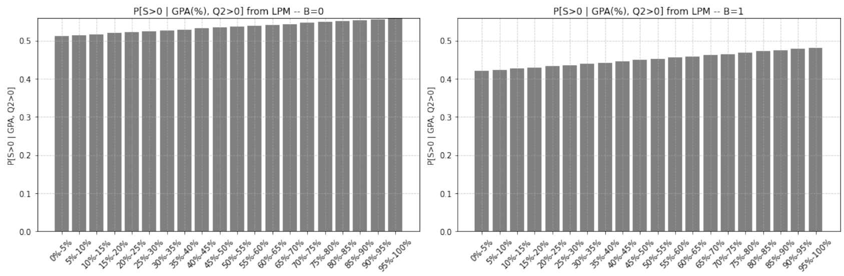
$$1(S > 0, B = 1) = \underbrace{\alpha_{3,B=1} + \beta_{3,B=1} \times GPA(\%)}_{P(S>0|GPA,Q2>0,B=1) \textbf{ from LPM}} + \epsilon_{3,B=1} = 0.42 + 0.06 \times GPA(\%) + \epsilon_{3,B=1} \tag{15}$$



(a) $P(Y|S > 0, GPA(\%), Q2 > 0, B = b)$ from Extrapolation

Note: This figure shows the probability extrapolated using predicted value of Y from equation (2) that is averaged for the $S > 0$ sample, standard deviation of ϵ_2, and normal distribution: from Equation (2): $Y_{GPA,s,b,Q2>0} = \underbrace{\alpha_2 + \beta_2 \times GPA(\%) + \gamma_2 \times s(\%) + \gamma_3 \times b}_{\hat{Y}_{GPA,s,b,Q2>0}} + \epsilon_2$.

Next, in the $S > 0$ and $Q2 > 0$ subsample, for each GPA (%) bin, we assume $Y_{GPA,S>0,b,Q2>0} \sim N\left(\hat{Y}_{GPA,S>0,b,Q2>0}, \text{std}(\epsilon_2)\right)$ then $P\left(Y \in Y_{\text{bin}} \mid GPA(\%) \in GPA(\%)_{\text{bin}}, S > 0, Q2 > 0\right) = \Phi\left(\frac{\bar{Y}_{\text{bin}} - \mu_2'}{\sigma_2}\right) - \Phi\left(\frac{\underline{Y}_{\text{bin}} - \mu_2'}{\sigma_2}\right)$ where μ_2' is the averaged local predicted mean: $\hat{Y}_{GPA,S>0,b,Q2>0}$, which is the average of $\hat{Y}_{GPA,S,b,Q2>0}$ from Equation (2) for each $GPA(\%)$ bin; σ_2 is the global standard deviation of ϵ_2 estimated from Equation (2) using the observed sample above the Score cutoff.



(b) $P(S > 0 \mid GPA(\%), Q2 > 0, B = b)$ from LPM

Note: This figure shows the probability from linear probability model (LPM) applied to the $Q2 > 0$ sub-sample for each background $B = b$: $1(S > 0, B = b) = \alpha_{3,B=b} + \beta_{3,B=b} \times GPA(\%) + \epsilon_{3,B=b}$, where $P(S > 0 \mid GPA, Q2 > 0, B = b)$ from LPM is $\alpha_{3,B=b} + \beta_{3,B=b} \times GPA(\%)$.

Figure 10: $P(Y \mid S > 0, GPA(\%), Q2 > 0, B = b)$ and $P(S > 0 \mid GPA(\%), Q2 > 0, B = b)$ from Extrapolation and LPM

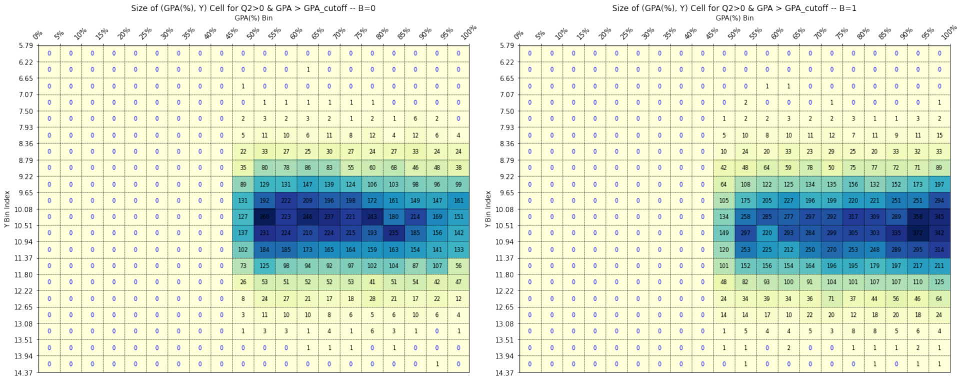
3. $P(Y \mid GPA(\%), Q2 > 0, B = b)$ from extrapolation: same as Figure 5.

4. $P(S > 0 \mid GPA, Y, Q2 > 0)$ from Bayes Rule (11) where the RHS items all come from either extrapolations or LPM. See Figure 12b.

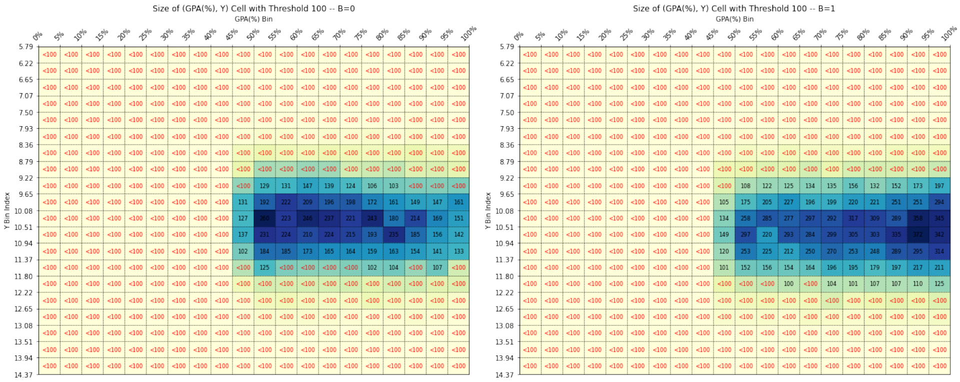
4.3 Feasible and Size Threshold 100 (Figure 12c)

Use local approximation for the feasible $GPA >= GPA_{cutoff}$ & $Q2 > 0$ sub-sample if $N_{GPA,Y,B,Q2>0} \geq 100$; use the Bayes Rule (11) from extrapolation if $N_{GPA,Y,B,Q2>0} < 100$ or $GPA < GPA_{cutoff}$ & $Q2 > 0$. See Figure 12c.

4.4 Cell Size for GPA vs. Y when $Q2 > 0$



(a) $GPA >= GPA_{cutoff}$ when $Q2 > 0$: applied to Figure 12a

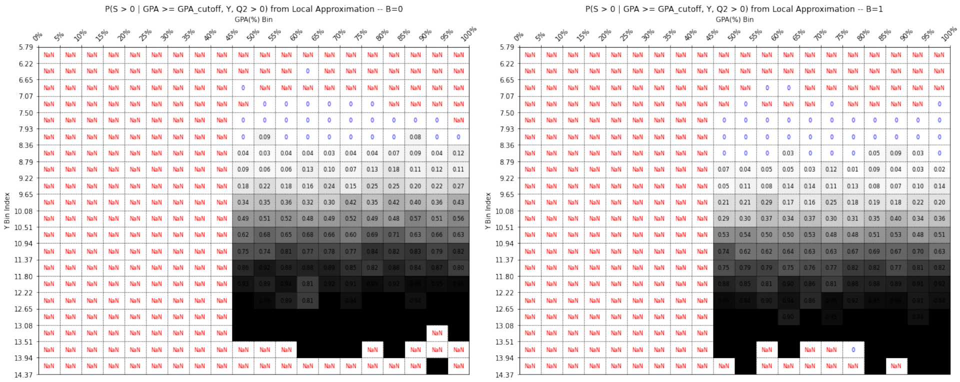


(b) $\bar{N}_{GPA,Y,B,Q2>0} = 100$ and $GPA >= GPA_{cutoff}$ when $Q2 > 0$: applied to Figure 12c

Figure 11: Size of $(GPA(\%), Y)$ Cell for the $Q2 > 0$ Sub-sample

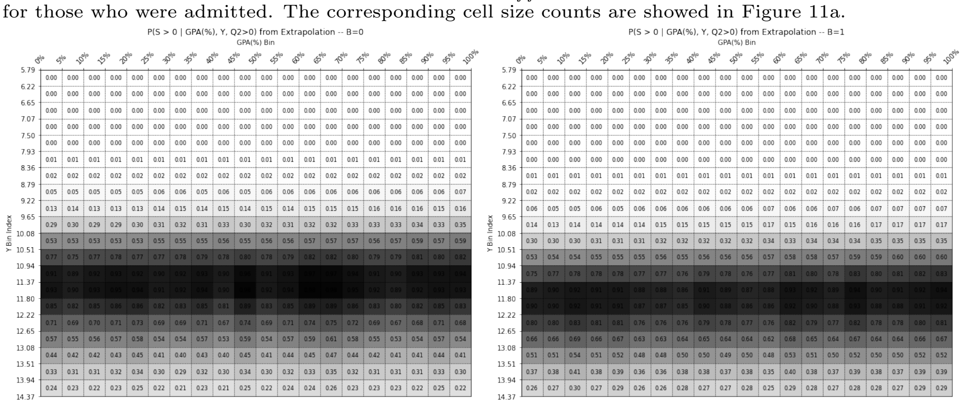
Note: These figures report the number of observations in each $GPA(\%)$ and Y cell for the $Q2 > 0$ sub-sample, where we also indicate whether the cell size is less than 100 or not. Figure 11a is applied to Figure 12a where we only consider the local approximation using the observed data. Figure 11b is applied to Figure 12c as threshold rule to decide whether using local approximation or extrapolation: cells with red letters here are using extrapolation model in Figure 12c, and cells with black letters are using local approximation in Figure 12c.

4.5 $P(S > 0 \mid GPA(\%), Y, Q2 > 0, B = b)$



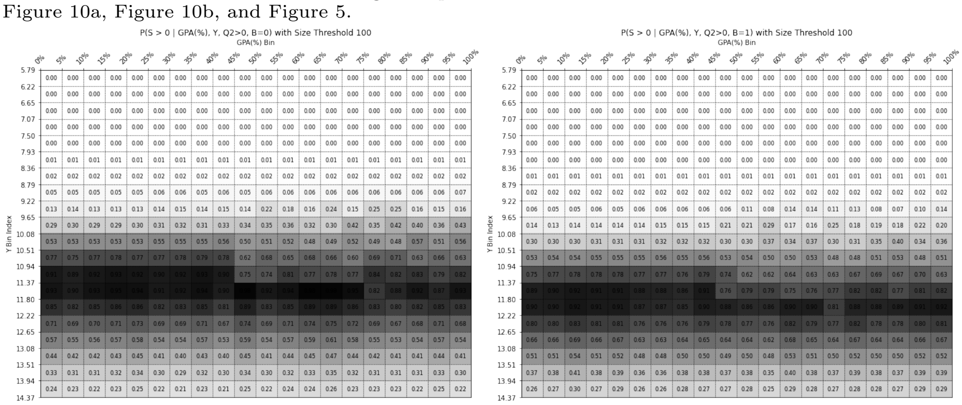
(a) Local Approximation: $\bar{N}_{GPA,Y,B,Q2>0} = 0$ (Cell Size in Figure 11a)

Note: This figure shows the observed $P(S > 0 \mid GPA(\%), Y, Q2 > 0, B = b)$ from local approximation in each $GPA(\%)$ and Y cell, without using any information from the extrapolation models or Bayes Rules. Note that we can only plot for the $GPA >= GPA_{cutoff}$ sample, since in real data we only observe Y for those who were admitted. The corresponding cell size counts are showed in Figure 11a.



(b) Extrapolation and Bayes Rule (11): $\bar{N}_{GPA,Y,B,Q2>0} > \infty$

Note: The figure shows the extrapolated $P(S > 0 \mid GPA(\%), Y, Q2 > 0, B = b)$ using Bayes Rule (11): $P(S > 0 \mid GPA(\%), Y, Q2 > 0, B = b) = \frac{P(Y \mid S>0, GPA(\%), Q2>0, B=b) \times P(S>0 \mid GPA(\%), Q2>0, B=b)}{P(Y \mid GPA(\%), Q2>0, B=b)}$, where all the blue terms are constructed using extrapolation or LPM models as detailed in the footnotes of Figure 10a, Figure 10b, and Figure 5.



(c) $\bar{N}_{GPA,Y,B,Q2>0} = 100$ (Cell Size in Figure 11b)

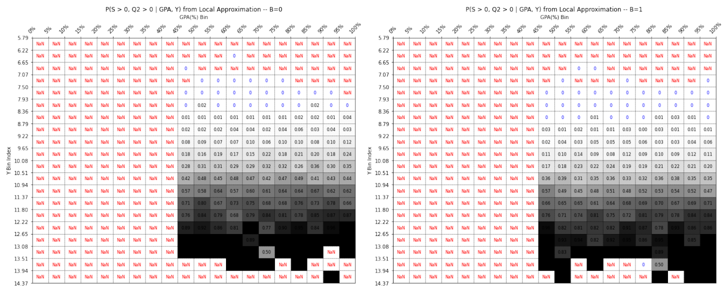
Note: $\bar{N}_{GPA,Y,B,Q2>0} = 100$, i.e., use local approximation for the feasible $GPA >= GPA_{cutoff}$ in the $Q2 > 0$ sub-sample if $N_{GPA,Y,B,Q2>0} \geq 100$; use extrapolation and Bayes Rule (11) if $N_{GPA,Y,B,Q2>0} < 100$ or $GPA < GPA_{cutoff}$ in the $Q2 > 0$ sub-sample. The corresponding cell size counts are showed in Figure 11b.

Figure 12: $P(S > 0 | GPA(\%), Y, Q2 > 0, B)$

Note: For the observed $GPA >= GPA_{cutoff}$ & $Q2 > 0$ sub-sample, we use local approximation if $N_{GPA,Y,Q2>0} \geq \bar{N}_{GPA,Y,B,Q2>0}$, where $N_{GPA,Y,B,Q2>0}$ is the size of each $(GPA(\%), Y)$ cell for a specific background B in the $Q2 > 0$ sub-sample.

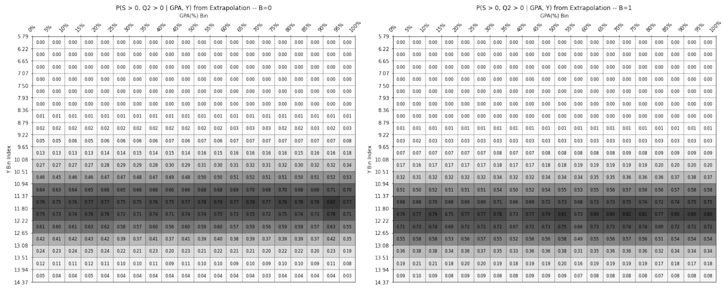
5 $P(S > 0, Q2 > 0 \mid GPA(\%), Y, B = b)$

$$P(S > 0, Q2 > 0 \mid GPA(\%), Y, B = b) = P(S > 0 \mid GPA(\%), Y, Q2 > 0, B = b) \times P(Q2 > 0 \mid GPA(\%), Y, B = b) \tag{16}$$



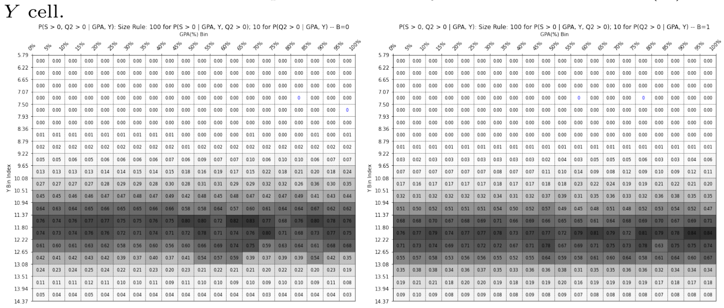
(a) Local Approximations: (Figure 9a*Figure 12a)

Note: This figure shows the observed $P(S > 0, Q2 > 0 \mid GPA(\%), Y, B = b)$ calculated from the multiplication of values in Figure 9a and Figure 12a, which both come from local approximations in each $GPA(\%)$ and Y cell.



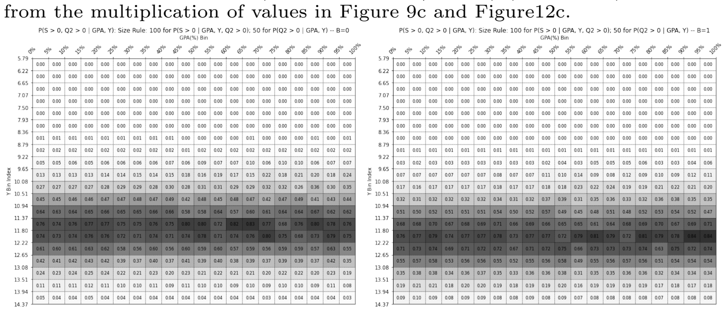
(b) Extrapolations and Bayes Rules: (Figure 9b*Figure 12b)

Note: This figure shows the extrapolated $P(S > 0, Q2 > 0 \mid GPA(\%), Y, B = b)$ calculated from the multiplication of values in Figure 9b and Figure 12b, which both come from extrapolations and Bayes Rules in each $GPA(\%)$ and Y cell.



(c) $\bar{N}_{GPA,Y,B} = 10$ & $\bar{N}_{GPA,Y,B,Q2>0} = 100$: (Figure 9c*Figure12c)

Note: This figure shows $P(S > 0, Q2 > 0 \mid GPA(\%), Y, B = b)$ calculated from the multiplication of values in Figure 9c and Figure12c.



(d) $\bar{N}_{GPA,Y,B} = 50$ & $\bar{N}_{GPA,Y,B,Q2>0} = 100$: (Figure 9d*Figure12c)

Note: This figure shows $P(S > 0, Q2 > 0 \mid GPA(\%), Y, B = b)$ calculated from the multiplication of values in Figure 9d abd Figure12c.

Figure 13: $P(S > 0, Q2 > 0 \mid GPA(\%), Y, B)$

Note: Based on $P(S > 0, Q2 > 0 \mid GPA(\%), Y, B = b) = P(S > 0 \mid GPA(\%), Y, Q2 > 0, B = b) \times P(Q2 > 0 \mid GPA(\%), Y, B = b)$, these figures show the different local approximated or extrapolated $P(S > 0, Q2 > 0 \mid GPA(\%), Y, B = b)$ using different values of $\bar{N}_{GPA,Y,B}$ and $\bar{N}_{GPA,Y,B,Q2>0}$.

6 Appendix

6.1 Descriptive Statistics of Random Seed 0 Sample

Table 1. Summary Statistics of Different Groups

Groups	Variables	count	mean	std	min	25%	50%	75%	max
Population (Applied to Q1)	Background	100000	0.5	0.5	0	0	1	1	1
	GPA	100000	9.01	1.01	6	8.32	9.01	9.69	13
	Applied_Q1	100000	1	0	1	1	1	1	1
	Admitted_Q1	100000	0.53	0.5	0	0	1	1	1
	GPA(%)_ApplyQ1	100000	0.5	0.29	0	0.25	0.5	0.75	1
	Y	100000	10.15	1.01	5.79	9.48	10.16	10.84	14.37
	Q2	100000	0.13	1.01	-4.71	-0.55	0.13	0.81	4.12
	S	100000	-0.02	1	-4.15	-0.7	-0.02	0.65	4.27
	Applied_Q2	100000	0.55	0.5	0	0	1	1	1
	Admitted_Q2	100000	0.27	0.44	0	0	0	1	1
	S(%)_ApplyQ2	55121	0.5	0.29	0	0.25	0.5	0.75	1
GPA>=GPA Cutoff (Addmitted to Q1)	Background	52828	0.57	0.49	0	0	1	1	1
	GPA	52828	9.77	0.63	8.93	9.28	9.65	10.14	13
	Applied_Q1	52828	1	0	1	1	1	1	1
	Admitted_Q1	52828	1	0	1	1	1	1	1
	GPA(%)_ApplyQ1	52828	0.74	0.15	0.47	0.6	0.74	0.87	1
	Y	52828	10.19	1.01	6.26	9.52	10.2	10.87	14.37
	Q2	52828	0.17	1.01	-3.89	-0.51	0.18	0.85	4.11
	S	52828	0	1	-4.01	-0.67	0	0.68	4.2
	Applied_Q2	52828	0.57	0.5	0	0	1	1	1
	Admitted_Q2	52828	0.28	0.45	0	0	0	1	1
	S(%)_ApplyQ2	30125	0.51	0.29	0	0.26	0.51	0.76	1
GPA<GPA Cutoff (Not Addmitted to Q1)	Background	47172	0.42	0.49	0	0	0	1	1
	GPA	47172	8.15	0.59	6	7.81	8.27	8.62	8.93
	Applied_Q1	47172	1	0	1	1	1	1	1
	Admitted_Q1	47172	0	0	0	0	0	0	0
	GPA(%)_ApplyQ1	47172	0.24	0.14	0	0.12	0.24	0.35	0.47
	Y	47172	10.11	1.01	5.79	9.43	10.11	10.79	14.33
	Q2	47172	0.08	1	-4.71	-0.6	0.08	0.75	4.12
	S	47172	-0.04	1	-4.15	-0.72	-0.04	0.63	4.27
	Applied_Q2	47172	0.53	0.5	0	0	1	1	1
	Admitted_Q2	47172	0.26	0.44	0	0	0	1	1
	S(%)_ApplyQ2	24996	0.49	0.29	0	0.24	0.49	0.74	1
Q2>0 (Applied to Q2)	Background	55121	0.52	0.5	0	0	1	1	1
	GPA	55121	9.05	1.01	6	8.36	9.05	9.73	13
	Applied_Q1	55121	1	0	1	1	1	1	1
	Admitted_Q1	55121	0.55	0.5	0	0	1	1	1
	GPA(%)_ApplyQ1	55121	0.51	0.29	0	0.26	0.52	0.76	1
	Y	55121	10.52	0.92	6.63	9.9	10.52	11.14	14.37
	Q2	55121	0.85	0.63	0	0.35	0.73	1.22	4.12
	S	55121	-0.02	1	-4.15	-0.7	-0.02	0.66	4.27
	Applied_Q2	55121	1	0	1	1	1	1	1
	Admitted_Q2	55121	0.49	0.5	0	0	0	1	1
	S(%)_ApplyQ2	55121	0.5	0.29	0	0.25	0.5	0.75	1
Q2<=0 (Not Applied to Q2)	Background	44879	0.47	0.5	0	0	0	1	1
	GPA	44879	8.96	1.01	6	8.27	8.95	9.64	13
	Applied_Q1	44879	1	0	1	1	1	1	1
	Admitted_Q1	44879	0.51	0.5	0	0	1	1	1
	GPA(%)_ApplyQ1	44879	0.49	0.29	0	0.23	0.48	0.73	1
	Y	44879	9.7	0.92	5.79	9.09	9.71	10.32	13.56
	Q2	44879	-0.76	0.59	-4.71	-1.09	-0.64	-0.3	0
	S	44879	-0.02	1	-3.93	-0.7	-0.02	0.65	4.2
	Applied_Q2	44879	0	0	0	0	0	0	0
	Admitted_Q2	44879	0	0	0	0	0	0	0
	S(%)_ApplyQ2	0	0	0	0	0	0	0	0
S>0 & Q2>0 (Addmitted to Q2)	Background	27115	0.48	0.5	0	0	0	1	1
	GPA	27115	9.06	1.01	6	8.38	9.06	9.75	13
	Applied_Q1	27115	1	0	1	1	1	1	1
	Admitted_Q1	27115	0.55	0.5	0	0	1	1	1
	GPA(%)_ApplyQ1	27115	0.52	0.29	0	0.27	0.52	0.77	1
	Y	27115	11	0.79	8.02	10.46	10.99	11.52	14.37
	Q2	27115	0.85	0.63	0	0.35	0.73	1.22	4.12
	S	27115	0.79	0.6	0	0.32	0.67	1.14	4.27
	Applied_Q2	27115	1	0	1	1	1	1	1
	Admitted_Q2	27115	1	0	1	1	1	1	1
	S(%)_ApplyQ2	27115	0.75	0.14	0.51	0.63	0.75	0.88	1
GPA<GPA Cutoff & S<=0 (Not Addmitted to Q1/Q2)	Background	24375	0.46	0.5	0	0	0	1	1
	GPA	24375	8.14	0.59	6	7.81	8.26	8.61	8.93
	Applied_Q1	24375	1	0	1	1	1	1	1
	Admitted_Q1	24375	0	0	0	0	0	0	0
	GPA(%)_ApplyQ1	24375	0.23	0.14	0	0.12	0.23	0.35	0.47
	Y	24375	9.65	0.89	5.79	9.05	9.66	10.26	12.74
	Q2	24375	0.07	1.01	-4.1	-0.6	0.07	0.75	3.87
	S	24375	-0.82	0.61	-4.15	-1.18	-0.7	-0.33	0
	Applied_Q2	24375	0.53	0.5	0	0	1	1	1
	Admitted_Q2	24375	0	0	0	0	0	0	0
	S(%)_ApplyQ2	12899	0.25	0.15	0	0.12	0.25	0.38	0.51

Figure 14: Descriptive Statistics

6.2 Coefficients of Extrapolation Equations

	Equation (1): Q1 Admitted	Equation (2): Q2 Admitted	Equation (3): Background 0	Equation (3): Background 1	Equation (3): Pool Background
background	0.20	0.30			-0.08
background (SE)	(0.01)	(0.01)			(0.00)
const	10.01	9.12	0.51	0.42	0.51
const (SE)	(0.02)	(0.02)	(0.01)	(0.01)	(0.00)
percentile_GPA_applyQ1	0.09	-0.04	0.05	0.06	0.06
percentile_GPA_applyQ1 (SE)	(0.03)	(0.02)	(0.01)	(0.01)	(0.01)
percentile_S_applyQ2		2.33			
percentile_S_applyQ2 (SE)		(0.03)			
Standard Deviation	1	0.7			

Figure 15: Coefficients of Equations (1)-(3)