Math 5071 Final Project Write-up: What Factors Influence the Enrollment of Math courses?

Adrian Cao

2023-12-06

Confidentiality Agreement

"I understood and agreed to that the data used in this report must be kept confidential and must not be shared or distributed outside this class. The statistical analysis, the result, and the report were agreed to be only used in this class and will not be posted or published without the authorization of Prof. Jimin Ding."

Enrollment Trends in Mathematics Courses: A Statistical Analysis

In the academic environment, understanding enrollment trends is crucial for curriculum planning, resource allocation, and policy making. This analysis delves into the enrollment patterns in the Department of Mathematics and Statistics at Washington University over a span of six years, from 2017 to 2022. The study aims to unearth the factors influencing these trends, with a particular focus on lower-level mathematics courses.

The data encompasses over 35,000 records from approximately 10,000 students, detailing their enrollment in various mathematics courses. These courses range across different levels, with the study specifically honing in on 100- and 200-level courses. The dataset is a comprehensive collection that includes information on student demographics, course specifics, and enrollment periods. The key variables included in the analysis are:

Student ID (Encrypted): A unique identifier for each student, ensuring confidentiality.

Primary Department: The department to which the student primarily belongs, indicative of their major field of study.

Course Number (Crs): A numerical identifier for each course, with focus on 100- and 200-level courses.

Enrollment Semester (Sem): The semester and year of course enrollment, with specific codes indicating Fall or Spring semesters.

COVID Period Analysis: An additional variable introduced to assess the impact of the COVID-19 pandemic on enrollment trends.

More details could be seen from the full data set provided by Professor Ding.

The primary objective is to statistically analyze the factors affecting enrollment trends in lower-level mathematics courses. Special attention is given to understanding the impact of variables starting semester, to course level, to primary department, and the influence of the COVID-19 pandemic. The analysis seeks to provide insights into how these factors collectively influence student enrollment decisions.

Using statistical models and techniques, the study aims to quantify the relationships between enrollment numbers and the aforementioned variables. The approach involves an exploratory data analysis to identify patterns and trends, followed by more complex statistical modeling to determine the significance and strength of these relationships.

Data Filter and Preprocess

Before performing the data analysis and modeling, the first thing we need to do is to meticulously preprocessing the dataset, which entailed selecting relevant courses and refining the data for analysis. The dataset initially consisted of individual student records for various mathematics courses. Our primary focus was on 100- and 200-level courses, given their foundational role in the curriculum. To enhance the dataset's relevance to our study, we performed the following preprocessing steps: 1. Filtered the data to include only 100 and 200-level courses. 2. Omitted discussion sections due to their distinct enrollment patterns. 3. Handling Missing Values: Evaluated and deleted missing values, given the low missing rate.

The transformation from individual student data to aggregated group data was central to our analysis. This aggregation was performed with the intent to discern the characteristics of different groups within the enrollment data. The aggregation methods varied based on the specific aspect of the analysis:

1. **Temporal Analysis**: Utilized average enrollments across all courses to examine temporal trends.
2. **Course-Level Analysis**: Employed total enrollments per semester for each course, offering insights into course-specific trends.
3. **Primary Department Analysis**: Aggregated data to reflect the number of students from each department enrolled in a particular course each semester, revealing departmental preferences and tendencies.

In our case, the aggregated data could reveal, for instance, which departments have higher enrollments in certain math courses. However, it doesn't provide insight into individual student choices or the specific factors influencing each student's decision to enroll in a course. Also, we hold the assumption that enrollment patterns of students from a particular department or year level do not vary significantly across different semesters or course levels, which may not be a realistic representation.

Exploratory Data Analysis

For the data reading and summary statistic, as it is also in out supplement file, so I would not include this here, but I will show how the variables I care hold different enrollment distribution.

We first see how total enrollment in all math courses would be differed by year and semester. This would be shown in Figure 1. It could be seen that there is more total enrollments in all math courses in fall semeste compared to that in spring semester. Also, there is a slightly increasing trend of enrollments over years. This could be due to the reason that modern studies now needs more and more mathematical foundations for prepare disciplinary understandings and for the entry level course, students tends to learn it as soon as possible, mainly their first semester, which leads to the popularity of more enrollments in fall. However, it could also may because there are more courses opening in fall which leads us to a more specific look at how courses enrollment in each courses would be differed by time.

Here in figure 2, we can tell that there are approximate the same number of courses opening in spring and fall. And it is indeed in general more students enrolled in fall semester compared to the spring semester, notably popular courses like 131, 132, 233. These calculus related courses, which are required for many majors and departments, hold more enrollments during the fall semester while the growing by years is not that obivous. Also, one interesting finding is that, 217, the ordinary differential equations class, holds very similiar performance for fall and spring semester.

Now, let's examine the distribution of enrollments by department across these various courses. To enhance the clarity and interpretability of our analysis, we have focused on undergraduate students, specifically excluding graduate student enrollments in these introductory-level mathematics courses. Our attention is centered on the following schools: the College of Arts and Sciences (LA), the McKelvey School of Engineering (EN), and the Olin Business School (BU), while grouping all other enrollments into a separate category.

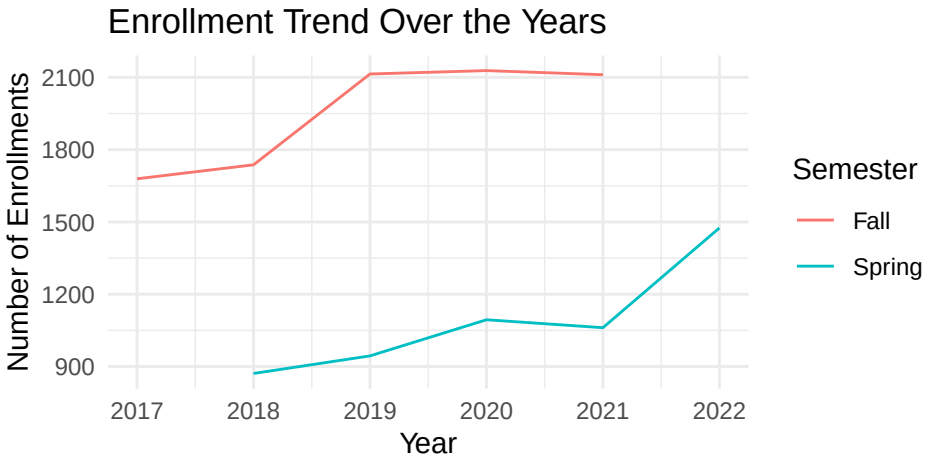

Figure 1: Enrollment Trend Over the Years

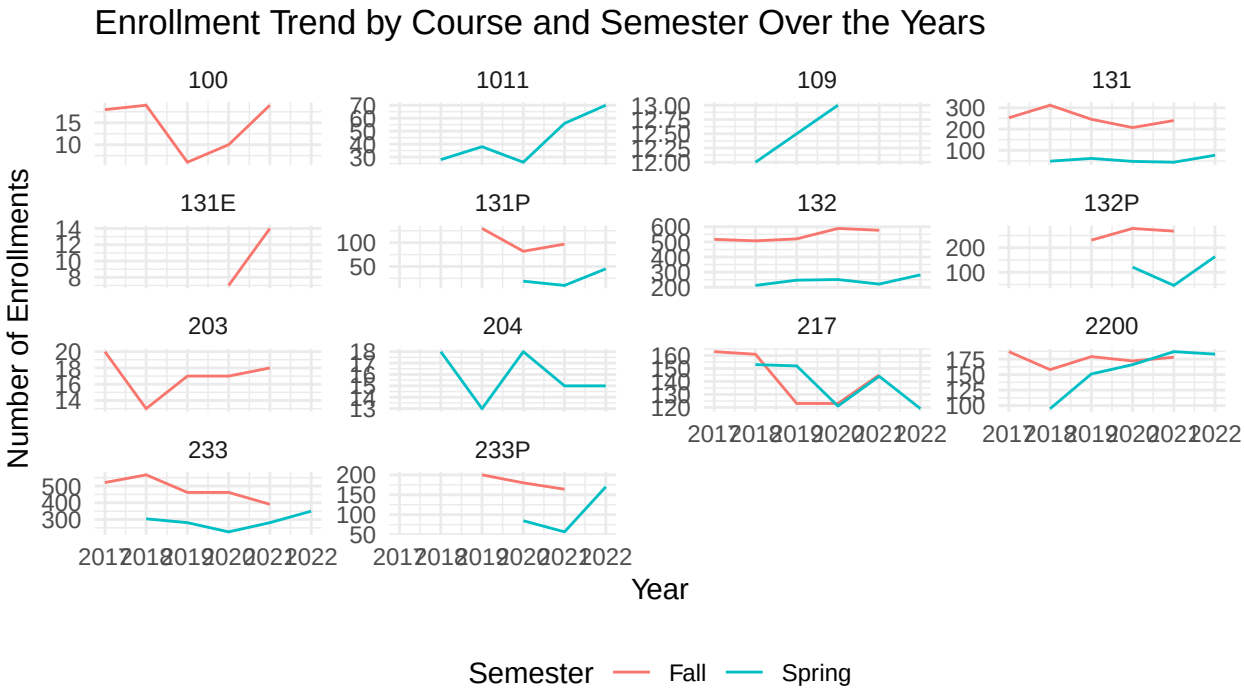

Figure 2: Course Enrollment Trend Over the Years

As depicted in Figure 3, it is evident that for the majority of the courses, students from LA represent the highest enrollment numbers. This trend could be attributed to the fact that mathematics is housed within the Arts and Sciences, coupled with the larger student body within this college. Engineering students follow as the second largest group enrolling in mathematics courses, a reflection of the intrinsic connection between engineering disciplines and mathematical foundations.

We observe distinct course preferences among the departments. Business students predominantly enroll in calculus courses, whereas students from the LA exhibit a more diverse course selection. A noteworthy trend is that engineering students occupy a significant proportion of seats in differential equations courses. This pattern may be linked to specific requirements within engineering curricula, which is also supported by the consistent enrollment figures for these courses across different semesters, as shown in Figure 2. Build upon this, the enrollment habits of engineering students appear to differ from their peers in liberal arts and business, who tend to enroll in introductory courses at the beginning of their college education.

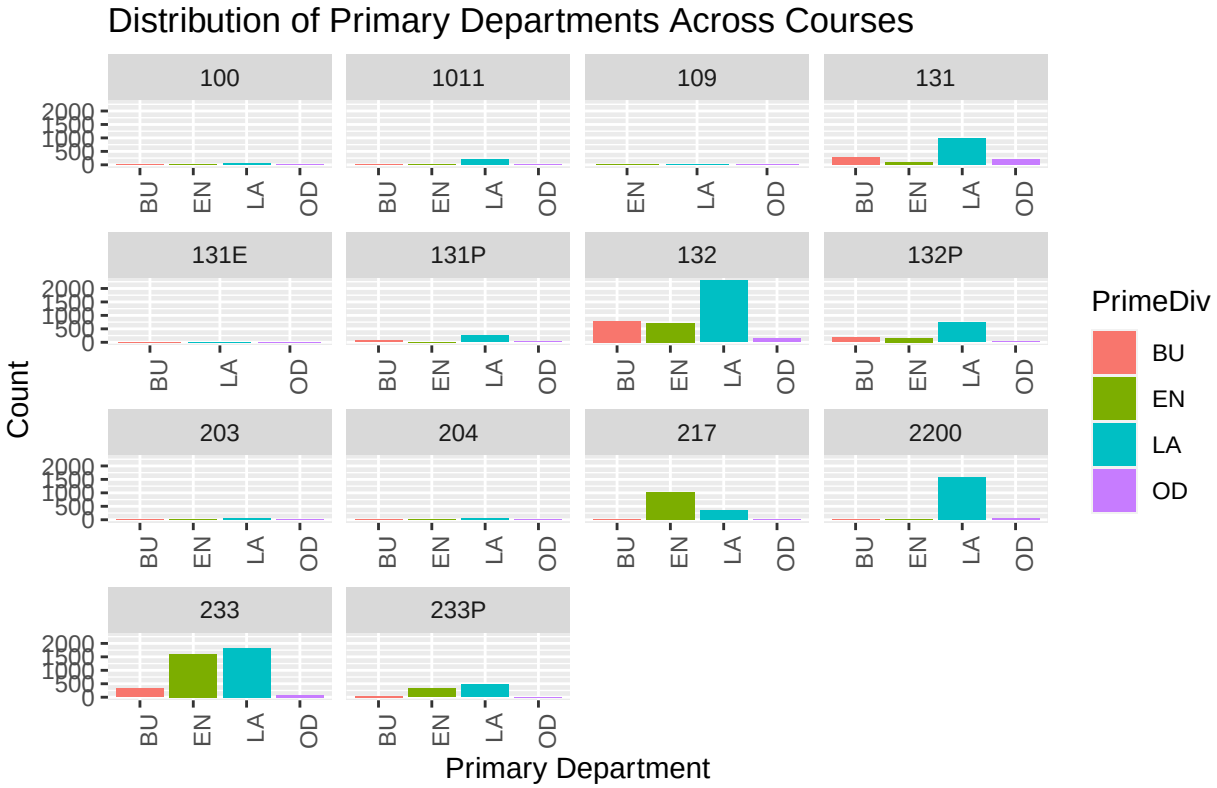

Figure 3: Distribution of Primary Departments Across Courses

In analyzing enrollment figures, we aimed to compare the numbers during the COVID period against non-COVID times to gauge the pandemic's impact on course registrations. This analysis was stratified by semester to discern if the pandemic's effects differed, particularly at the beginning of 2020, which marked a Spring semester. The period analyzed, extending from Spring 2020 to Fall 2021, is significant as it represents the timeframe when a fully online learning option was available to students. Even as the university shifted to a hybrid teaching model in Fall 2021—bringing most students back to campus—that semester was still considered part of the pandemic era to maintain the consistency of our analysis.

Referencing Figure 4, there is a discernible trend suggesting that the COVID-19 pandemic influenced student enrollment patterns, with a downward shift in the average number of course enrollments during the period in question. Despite an observed increase in average enrollments during the Fall, which may indicate students' adjustment to the pandemic's challenges, the Spring semester data reflects a notable enrollment decrease, especially at the pandemic's outset. This initial decline could be attributed to a range of factors, from the

sudden pivot to online learning modalities to students opting for gap years, amid the pervasive uncertainties of the pandemic's early phases.

It should be noted that this effect may not appear as pronounced in the graph, partly because the university underwent an expansion in 2020 following a change in administration, with a larger cohort of students being admitted. This increase in the student body could mask the true extent of the enrollment decline, as the growing numbers overall could offset the visibility of any decreases due to the pandemic.

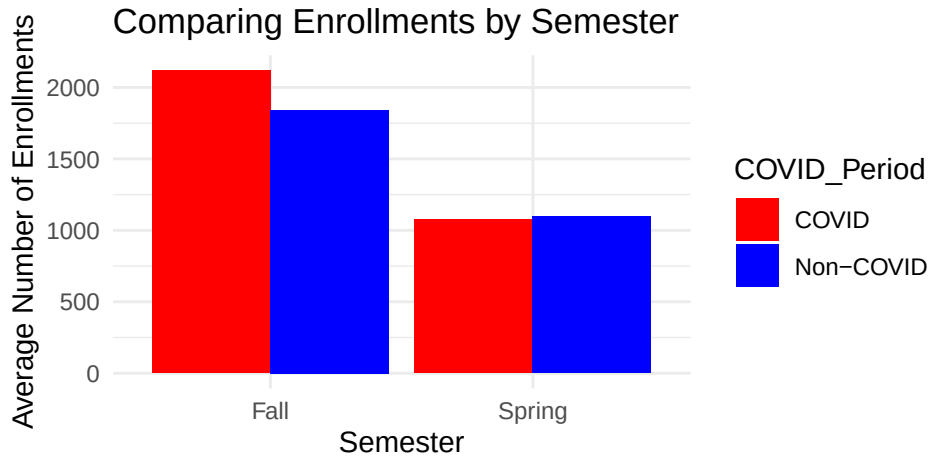

Figure 4: Comparing Enrollments by Semester

Model Building and Diagnostics

The crux of our statistical modeling begins with the construction of an appropriate response variable. Given that our dataset encapsulates individual student enrollments per course, we have a couple of avenues to construct our response variable:

1. Binary Response: This would involve a logistic regression setup, creating a binary variable indicating enrollment status. However, this seems redundant as our dataset exclusively includes enrolled students.
2. Count Response: Aggregating the data at a course level for each semester and using the total enrollments as a response variable seems more fitting. This approach naturally lends itself to count models such as Poisson or negative binomial regression.

Given this context, we aggregated the data by semester to assess dependency on various covariates. The initial model we employed is a simple linear regression aimed at understanding how average enrollment in introductory-level math courses differs by semester.

Temporal Analysis

This model is simply build to see how enrollment change by year and semester:

$$Enrollment_t = \beta_0 + \beta_1 * Year + \beta_2 * \mathbb{1}_{Semester=Spring} + \epsilon_t$$

Here, year is a continuous variable but for simpler its interpreability, I have change it to the raw year divide by 2017 to see how much, on average, enrollment would be changing since 2017. And semester is a categorical variable where I choose fall as the reference group for the reason that there are larger than the spring semester.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	1695.8	71.22311	23.809687	0.0000001
Year_Since_2017	129.0	25.18117	5.122875	0.0013644
SemesterSpring	-993.8	75.54352	-13.155332	0.0000034

Again, in our regression model, we've treated 'Year' as a continuous variable, with 2017 set as the baseline, making the intercept a direct reflection of the average enrollments for that year in fall semester, which is also the first point of our data. Each subsequent year has seen a significant increase in average enrollments by about 129 students, indicating a growing trend. Conversely, the Spring semester experiences a notable decrease, with enrollments dropping by nearly 994 compared to the Fall, suggesting a strong seasonal pattern. Overall, the model shows a consistent yearly rise in enrollment numbers, pointing towards an increased interest in or availability of introductory math courses, although the pronounced seasonal fluctuation warrants consideration of additional factors that may affect these trends. Before getting to talk about more factors, let's first check about the model diagnostics to see if it holds some poential problems with linear regression assumptions.

While the diagnostic plots in figure 5 suggest that some assumptions may not be fully met, they do not show strong violations that would necessarily undermine the validity of the model. However, considering transformations or alternative modeling approaches could potentially improve the model fit and the reliability of its inferences. However, with the checking about box-cox transformation, we noticed the indicator function, meaning no change, would be good enough to give us a solid result. Hence, we can lead to the conclusion that the currect model we have are good enough to tell us some relationships between the enrollment and year and semester one is enrolled in.

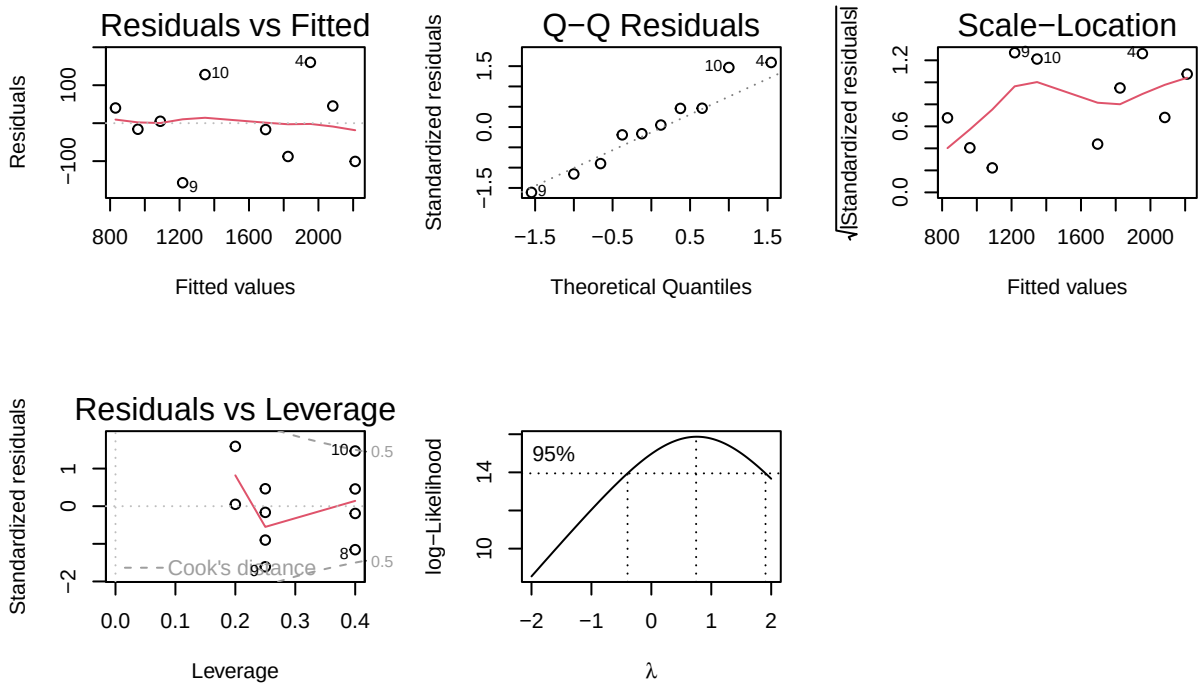

Figure 5: Regression by time diagnostics

The model provides a clear indication of temporal trends in enrollment, but it's essential to consider external factors that could influence these trends. Future models could explore additional predictors, like specific course attributes or broader educational trends, to gain more comprehensive insights.

Course-Level Analysis

Next, we add the course factor to see how different course would holds different performance for enrollments

$$Enrollment_{t,crs_i} = \beta_0 + \beta_1 * Year + \beta_2 * \mathbb{1}_{Semester=Spring} + \beta_i \times \mathbb{1}_{Crs=Crs_i} + ... + \epsilon_{t,crs}$$

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	455.086789	21.067800	21.6010588	0.0000000
Year_Since_2017	1.459547	4.520315	0.3228861	0.7476682
SemesterSpring	-132.871312	14.357416	-9.2545419	0.0000000
Crs100	-443.605882	31.519939	-14.0738179	0.0000000
Crs1011	-282.994118	31.519939	-8.9782572	0.0000000
Crs109	-312.634571	44.283415	-7.0598569	0.0000000
Crs131	-238.500000	25.127422	-9.4916223	0.0000000
Crs131E	-449.695203	44.570869	-10.0894421	0.0000000
Crs131P	-329.926214	29.364656	-11.2354870	0.0000000
Crs132P	-208.926214	29.364656	-7.1148871	0.0000000
Crs203	-441.005882	31.519939	-13.9913305	0.0000000
Crs204	-310.794118	31.519939	-9.8602386	0.0000000
Crs217	-251.900000	25.127422	-10.0249042	0.0000000
Crs2200	-226.700000	25.127422	-9.0220158	0.0000000
Crs233	-7.800000	25.127422	-0.3104178	0.7570929
Crs233P	-251.092880	29.364656	-8.5508537	0.0000000

Interestingly, we found the coefficient for *Year* is not statistically significant ($p > 0.05$), indicating that there is no strong evidence of a year-over-year change in enrollments when controlling for semester and course. On the other hand, the coefficient for semester is significantly negative ($p < 0.001$), indicating that enrollments in the Spring semester are, on average, 132.871 less than in the Fall, suggesting a strong seasonal effect. The coefficients for other courses compared to course 132 show significant differences. For instance, course 100 has 435.81 fewer enrollments on average compared to course 132 ($p < 0.001$). This pattern holds for other courses as well, where negative coefficients indicate fewer enrollments compared to course 132. However, interestingly, the course 233, which is the Calc 3 higher level course to 132 Calc 2, is the only one that is not significant related. This could be due to the reason that this two courses holds very similiar sizes, making them not that distinguishable in enrollments compared to each other.

For the model diagnostics in figure 6, it is also very similar to the previous model. While the model seems reasonable in terms of linearity, there may be concerns about the normality of residuals, potential heteroscedasticity, and a few influential observations. This problem would be addressed further with check for non-linearity, interaction terms, and cross-validations.

In conclusion, The model suggests semester timing (Spring vs. Fall) is a strong predictor of enrollment numbers, with fewer students enrolling in the Spring. Specific courses have a significant impact on enrollment numbers, with most courses having fewer enrollments than course 132. The year since 2017 does not appear to be a significant predictor of enrollment numbers when controlling for semester and course. The lack of significance for the year could be due to multicollinearity, especially if certain courses are only offered in specific years or if there are other time-related variables that are not included in the model.

Now, since we have identified that the year variable is not contributing much to the explanatory power of the model, we consider it may be better to delete it. Or since it is used to see how enrollments changes over year, we may also interested to see the impact of semester for this year-by-year change. For this reason, check their interaction term would help us better understand the yearly change effect of the model. Here, these two models are build and the result could be seem in the supplement R file.

Across all the models assessed, R-squared values are closely aligned, indicating that each model accounts for a comparable portion of variance in enrollment figures. This similarity extends to Adjusted R-squared values,

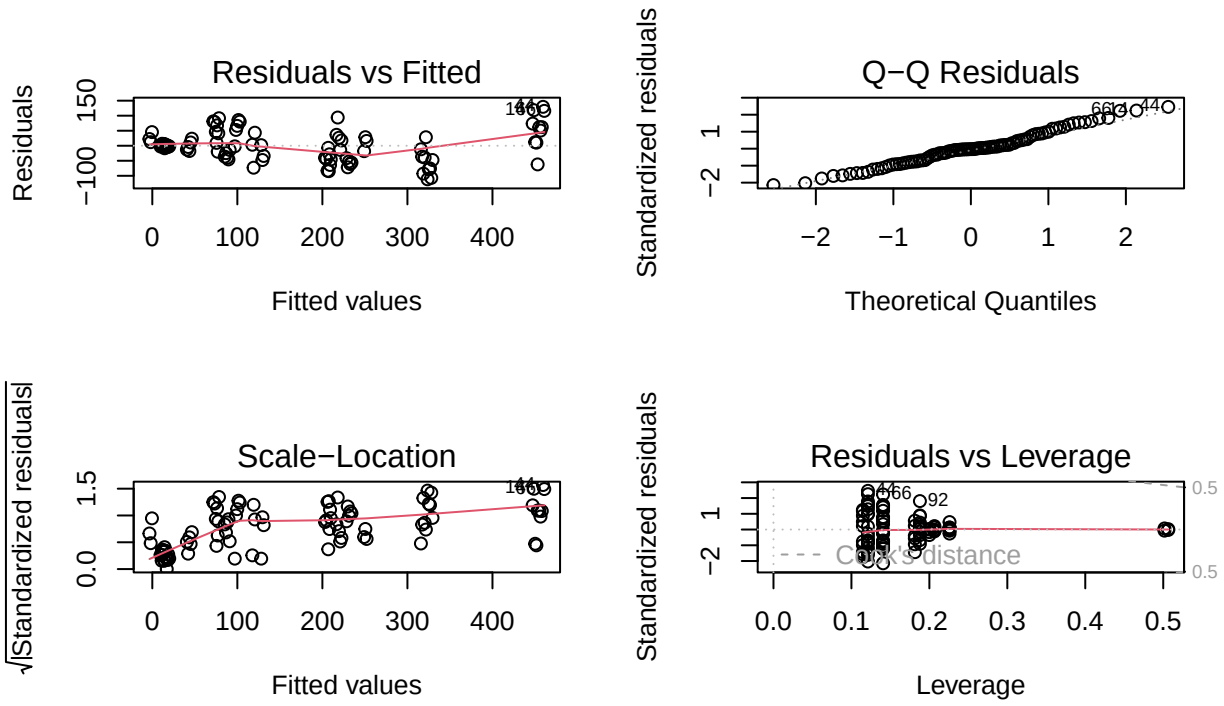

Figure 6: Regression by course diagnostics

suggesting the quantity of predictors is well-suited for the volume of data. Interestingly, introducing the variable 'Year' or the interaction term between 'Year' and 'Semester' does not significantly enhance explanatory power, implying that the impact of semesters on enrollments remains consistent over time. Conversely, 'Semester' emerges as a significant factor in all models, underscoring a seasonal enrollment trend. Moreover, course numbers ('Crs') stand out as substantial predictors, corroborating the notion that enrollment varies markedly across different courses. Considering the parallelism in R-squared and Adjusted R-squared values, the simplest model may be the most judicious choice for its parsimony. Nevertheless, the final selection of a model should be informed not only by simplicity but also by the intended application, whether it be for predictive accuracy or for elucidating the influence of specific variables. Hence, the preferable model for course level analysis would be:

$$Enrollment_{t,crs_i} = \beta_0 + \beta_2 * \mathbb{1}_{Semester=Spring} + \beta_i \times \mathbb{1}_{Crs=Crs_i} + ... + \epsilon_{t,crs}$$

There is one more caveat to consider before diving into the next model is overfitting problem as we capture the random noise in the data instead of the underlying relationship. Since we are using all the number of courses included to build the model, we may want to make sure this model is stable by checking the cross-validation, which is a technique used to assess how the results of a statistical analysis will generalize to an independent dataset. We checked the cross-validation results and notice the previous model is pretty stable for that its similar performance on unseen data.

In the supplemental R file, I also tried some nonlinear model to see if it fits better than our original one. It turns out the nonlinear model also holds similar performance to our model, showing out model may be good enough for both robustness and simplicity

Primary Department Analysis

Lastly, we go a little further to see how student from different schools would holds different preference for their entry level math courses choice. This is achieved by aggregating the data through different departments in certian courses each semester and run the linear model for that. Notice, as I mentioned earlier, as we are investigating towards the 100- and 200- level math courses, we may just care about undergraduate profolios in these courses. Hence, all the graduate students are removed in this data.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1180.1492275	3703.432710	-0.3186636	0.7501983
Year	0.6140182	1.834343	0.3347347	0.7380538
SemesterSpring	-34.7136159	5.627289	-6.1687988	0.0000000
Crs1011	40.6477003	16.831532	2.4149733	0.0163206
Crs109	20.4535122	22.973739	0.8902997	0.3740022
Crs131	53.7785746	13.287265	4.0473773	0.0000656
Crs131E	2.6971986	20.952916	0.1287266	0.8976582
Crs131P	27.7338054	14.845847	1.8681188	0.0626980
Crs132	113.4035746	13.287265	8.5347567	0.0000000
Crs132P	60.9228897	14.528379	4.1933714	0.0000360
Crs203	-15.8059612	16.742415	-0.9440670	0.3458778
Crs204	17.7936365	17.571098	1.0126650	0.3120176
Crs217	51.4195578	13.601725	3.7803704	0.0001881
Crs2200	57.2727149	13.380344	4.2803619	0.0000250
Crs233	111.4535746	13.287265	8.3879996	0.0000000
Crs233P	50.3812231	14.528379	3.4677800	0.0005997
PrimeDivBU	-82.7900653	6.728374	-12.3046174	0.0000000
PrimeDivEN	-52.3140054	6.507087	-8.0395434	0.0000000
PrimeDivOD	-96.2253529	6.765438	-14.2230788	0.0000000

In the refined linear regression model, the inclusion of primary department (PrimeDiv) as an additional predictor provides nuanced insights into enrollment trends. The non-significant coefficient for Year suggests that temporal changes in enrollment are not captured when controlling for semester and department, echoing findings from the previous model. Interestingly, the SemesterSpring coefficient's magnitude decreases, indicating a milder semester impact on enrollment numbers when departmental factors are considered. This implies that the primary department has a moderating effect on the seasonal variation previously observed. Course number (Crs) remains a significant predictor, with enrollment variations by course persisting in this model. The PrimeDiv coefficients reveal significant enrollment differences across departments, highlighting the importance of departmental affiliation in driving enrollment patterns. This suggests that students' primary departments are closely related to their course selection choices.

The model's goodness-of-fit, as indicated by a lower R-squared value (0.6022) compared to the initial model (0.8862), suggests that while PrimeDiv is crucial, the detailed course-level data in the earlier model accounted for more variability. The reduction in the Residual Standard Error to 42.94 points to increased predictive accuracy with the inclusion of PrimeDiv. Cross-validation results affirm the model's stability, indicating that the findings are likely to generalize across different subsets of the data. The consistency between the original and cross-validated models strengthens the confidence in the model's predictive capabilities.

The analysis underscores the significance of departmental influences on enrollment patterns and suggests that while semester timing has an effect, departmental factors may play a more nuanced role in determining student course enrollment. These insights could be particularly valuable for academic advisors and university administrators in understanding and predicting enrollment dynamics.

In our study, we explored the interaction between students' PrimeDiv and Crs to understand department-specific enrollment patterns. However, this approach led to collinearity issues in the interaction model by

Table 4: COVID Period Coefficients and Their Significance

Model	Coefficient	P_Value
Model 1	-129.750000	0.1570781
Model 2	-15.625978	0.1986381
Model 3	-3.958068	0.4220230

linear regression. To address this, we employed penalized regression techniques, such as LASSO and Ridge regression, which improved the model's predictive accuracy by managing collinearity and reducing overfitting. Despite these improvements in prediction, these methods are less suitable for explanatory analysis due to their focus on prediction over interpretation. LASSO and Ridge, while controlling for collinearity, tend to obscure the direct impact of individual predictors, especially in models with interaction terms. Consequently, while useful for enhancing prediction, these techniques do not significantly aid our goal of understanding the specific ways in which students' department affiliations influence their course choices.

COVID Impact

To find the COVID impact on enrollments, my way is to incorporate the COVID variable into our previous three model to see if it holds influence on students' enrollment by time, by course, or by department. The use of linear regression models in analyzing the impact of the COVID-19 pandemic on course enrollments is motivated by their ability to isolate and quantify the specific effects of the pandemic, while controlling for other influencing factors. These models are effective in discerning whether observed changes in enrollment patterns during the pandemic are statistically significant or if they could be attributed to random variation.

Across all three models shown in Table 4, the coefficients for the COVID period are negative, suggesting a trend of decreased enrollments during the pandemic. However, none of these decreases are statistically significant as indicated by the p-values being greater than 0.05. The lack of statistical significance implies that, based on these models, we cannot conclusively state that there was a significant difference in enrollments during the COVID period compared to the non-COVID period. This might suggest that the impact of the COVID-19 pandemic on course enrollments was not as pronounced as might be expected, or that other unexamined factors could have influenced enrollments during this period.

Conclusion and Discussion

In our analysis of enrollment trends in math courses, key findings emerged that shed light on the factors influencing student course selections. The models consistently highlighted the role of semesters, course preferences, and department affiliations in shaping enrollment numbers. A notable trend was the increase in enrollments over time, particularly in the Fall semester, suggesting both a growing interest in math courses and potential seasonal influences. The course selection emerged as a significant predictor, overshadowing the year-over-year trends, while the primary department also played a crucial role, indicating that students' majors significantly influence their course choices. Interestingly, the COVID-19 pandemic did not show a statistically significant impact on enrollment patterns, suggesting a level of resilience or adaptation in student behavior during this period.

However, these conclusions come with their own set of complexities and considerations. Despite statistical significance, the practical impact of our findings was modest, highlighting a gap between statistical and practical significance. The possibility of latent variables, not captured in our models, adds a layer of uncertainty to our interpretations. Additionally, the aggregation of data at certain levels potentially masks the nuanced effects of individual predictors, a critical aspect in data analysis that requires careful judgment. This analysis, while insightful, underscores the complexities inherent in statistical modeling, where definitive conclusions are often hedged with caveats.