Modeling techniques for assessing outcomes from various introductory statistics curricula in a second statistics course

Abstract:

Statistics education is a new and evolving area of research, both in terms of curriculum development and assessment of the impacts of changing statistics curriculum. New simulation-based, active-learning curricula have recently been developed for introductory statistics. Recent work has suggested that these new teaching methods are improving student success rates in the introductory course, but little is known about the impacts on students in a second course. We use data from a university that has been introducing various versions of new curricula alongside traditional approaches to assess the impacts of three different curricula on student performance in a second course. We model student performance as quantitative, censored, ordinal, and binary responses, and compare the different modeling approaches. We find consistent evidence that students who take randomization-based courses have lower performances in a second statistics course on average, but these differences are small relative to those based on students’ prior grades.

**1 Introduction**

Introductory statistics courses are required for a wide variety of undergraduate degree programs including nursing, business, and most sciences. They are often viewed as “gateway” courses that need to be passed before students can progress through their degree programs. At a moderately sized, western U.S. land-grant institution a large proportion of students elect or, more typically, are required to take an introductory statistics course offered by the mathematical sciences department called STAT 216. It is the largest single course on campus, taught with about twenty sections of 40 students each semester, and had a notorious reputation with students based on a high withdrawal and failure (*W*, *D*, or *F*) rate. In order to improve the course and in response to developing instructional ideas in the field of statistics education moving away from the traditional consensus (formula-based) curriculum toward randomization-based instructional techniques, some of the sections began using randomization-based, active-learning curriculum based on the CATALST materials (<http://www.tc.umn.edu/~catalst/>). Initially a small number of sections were piloted with the provided materials and then a continually evolving curriculum inspired by the CATALST ideas was developed. Additionally, during this same time the text for more conventional lecture-based sections was changed from a traditional, consensus curriculum text to one that incorporates some randomization ideas, providing further opportunities for comparisons of outcomes.

Recent research [reference omitted] discussed results that show an increase in student success (passing) rates from a randomization curriculum over versions of the consensus curricula in STAT 216. However, less is known about the impacts of introductory statistics courses with a randomization curriculum on preparation of students for more advanced statistical course work. Fortunately, the past years of curricula modifications of STAT 216 has created a natural experiment to explore this, tracking students who took different versions of introductory statistics in their next statistics course. At this university, the second statistics course of interest is called “Intermediate Statistical Methods” (STAT 217), a course that approximately a quarter of the students who take STAT 216 continue into and that is an additional requirement of some business and science majors. It has a pre-requisite of successfully completing STAT 216 (*C-* or better) or an equivalent course.

The university administration was mostly concerned with the passing rates, the rate at which students receive and *C-* or above, but this ignores information that is available by using the entire range of grade responses. More powerful analyses are possible by analyzing students’ performance using their grades in the course as responses. The focus of this research is to compare student performance (grades) in STAT 217 based on their curriculum and grades in STAT 216, also controlling for student cumulative GPA before taking STAT 217 as a proxy for overall student performance. Focusing on student grade outcomes presents some modeling challenges that we take different approaches to address. First, a student cannot receive a score higher than an “*A*” in STAT 217. This natural upper limit to students’ success in STAT 217 prevents us from observing differences in some of the top performing student’s “true” score as higher than other students who also received an *A* – they are censored. Second, grades as reported to the registrar are categorical, ordinal variables. In this work, we start with applying ordinary least square regression, which is not appropriate for the previous reasons, to quantitatively coded grades. These results and the varying assumptions made by the approaches are then compared to censored regression models (Yee, 2015) and cumulative link models designed to model ordinal responses (Agresti, 2010). The main focus of this paper is to explore these methods with the goal of addressing whether or not the different curricula in introductory statistics courses has influenced students’ performance in a second semester statistics course.

**2 Background**

More than twenty STAT 216 sections of around 40 students are run each semester that are taught typically by a non-tenure track faculty member or graduate student. Exams and general structure were common across sections with some instructor freedom on the day-to-day aspects of the course. The course was taught with a consensus curriculum tied very closely to the textbook chosen for the course; in the era of interest here the text was Deveaux, Velleman, and Bock’s Stats: Data and Models, covering Chapters 1-22 in a 15-week semester (3rd edition, 2013). This text provided a formula-based presentation of topics that included one and two sample tests for means and proportions, conditional probability through contingency tables, and explored scatterplots and correlation. Inferences were all conducted using hand-calculation, tables, and online “apps”. In the “conventional” sections in Spring 2014, the Deveaux, Velleman, and Bock text was replaced with Lock, Lock, Lock, Lock, and Lock (2012) covering Chapters 1-6 in a 15-week semester. These classes and the text retained a relatively conventional presentation and structure while using randomization techniques to motivate inferences. All of these classes were taught in conventional, often tiered, classrooms with the website related to the text used for learning as well as data summarization and analysis. In Spring 2014, the order presented in the book was followed, while in Fall 2014 and Spring 2015 topics were reordered (Unit 1: hypothesis test and interval estimation for 1 proportion; Unit 2: test and estimate in each of three settings: two proportions, two means, one mean, and regression slope; Unit 3: z and t tests for the settings covered previously.) Statkey (<http://lock5stat.com/statkey/>) was used for simulation-based inference.

Along with these more conventional versions of the course, in Fall 2012 and Spring 2013 randomization-based sections of the course were taught using the CATALST active-learning curriculum with Tinkerplots to summarize and analyze the data sets related to each activity. Unit 1 was shortened in Fall 2013 and Unit 3 was heavily modified to better fit the follow-up class, STAT 217, dropping Tinkerplots in favor of R-Shiny web apps which were locally developed for use in STAT 216. In Fall 2014 further modifications were made that became a locally-published packet of course materials that was still using a few of the CATALST activities. Spring 2015 saw a further modification of the order of topics, so that the current course pack has only a few similarities with the original CATALST curriculum. Simulations were done using a combination of Rossman-Chance applets, StatKey, and locally developed web “apps” (reference omitted). Initially all of these active learning sections were taught in Technology Enhanced Active Learning (TEAL) rooms but in later years some of the sections were taught in regular classrooms, usually with mobile seating but without the custom-designed classroom architecture and technology. The TEAL rooms used circular desks and the ability for each group to easily project their laptop to the entire class. Class settings involved short introductions and wrap-up summaries at the end of the class along with just-in-time mini-lectures as needed.

STAT 217 is a course taken by a wide variety of students on campus, with the pre-requisite of successfully completing STAT 216 (C- or better) or an equivalent course. Typically, ten sections are run each year with around 35 students with common exams and sometimes common activities and projects across the sections in a given semester. The instructors are mostly more senior statistics Master’s and Doctoral students with an occasional non-tenure track or tenure track faculty member teaching the course. The instructors are given more freedom to develop the day-to-day materials than in STAT 216. STAT 217 covers inference with one-way and two-way ANOVAs (with interactions), multiple linear regression with interactions and multi-category categorical explanatory variables, and Chi-square testing methods for homogeneity and independence situations. The statistical software R (R Core Team, 2015) via the R-studio (RStudio Team, 2015) interface is introduced and used from the first to last day of the class. The course touches on writing R code and data management, and heavily emphasizes data visualization and interrogation of model results and model assumptions. The text adopts many ideas from the MOSAIC project and its associated R package (Pruim et al., 2015) in terms of formula notation and associated R-code for performing permutations and bootstrapping. Named distributions that should be new to the students, such as the F and Chi-square, are introduced first via permutations and then using their parametric approximations. Visualization of model results is accomplished using the effects package (Fox, 2003).

A locally developed and published textbook was written and first used in Fall 2013 (reference omitted). Two subsequent revisions were made for Spring 2014 and Fall 2015. The content and the course remained essentially the same through these revisions, with R code updates and moderate changes to the content and examples occurring in each revision. A paper copy of the text is available for purchase through the university’s bookstore for under $40 as well as a free digital download from the provided link. Because of the variety of different experiences students were having in the pre-requisite course, let alone the population of students that transferred into the university, the text begins with an attempt to unify students understanding of basic inference techniques by focusing on two-sample mean problems, both using conventional parametric inferences and using permutation and bootstrapping. In this context, basic concepts that should have been covered in STAT 216 and that are needed for more complex material are reviewed. In the first week of the class, students receive their first introduction to R via R-studio to analyze two-sample mean data. The course then proceeds into covering new topics with students required to use each technique to perform an analysis and write a summary report. The course usually contains five projects, three exams including the semi-cumulative final, and many smaller homeworks and quizzes that have a limited weight on the student grades. The same general grading framework has been used over the time of data collection and, while being inspired by previous exams and projects, all the materials change each semester. Exams are curved to maintain similar overall means and medians each semester to reduce semester to semester and section to section variability in grades.

Anecdotally, STAT 217 instructors have not detected differences in student performance based on the curriculum used in their prior course but often the instructors are not aware which version of the course students had taken. With all the changes to STAT 216 curricula, the students also have trouble identifying their version of STAT 216; and while this is also a somewhat interesting issue, it has made it difficult to understand whether there are differences in outcomes in the second course. This research quantitatively explores whether there are differences in performance in STAT 217 based on student background, starting with interests in whether different grades in STAT 216 were associated with different performances in STAT 217, after controlling for overall student performance using prior semester cumulative GPA.

**3 Data and Visual Assessments**

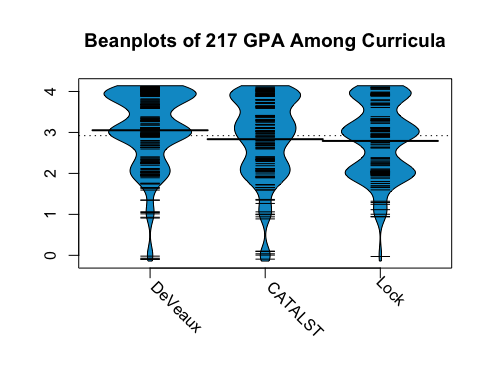
**3.1 Data**

The data were collected from student records provided by the university’s Registrar from Fall of 2013 to Spring of 2016, with an IRB approved data management protocol for this project. The data set contains the students’ grades in STAT 217 and STAT 216, cumulative GPA up to the semester before they took STAT 217, and enough information to identify the curriculum of STAT 216 that the student took and section and semester of STAT 217. GPA is recorded on a 0 to 4.0 scale with individual grades measured using a +/- system. For students who took STAT 216 before 2012, we assumed that their curriculum was “DeVeaux,” as that was the text used for many previous years. There were 55 students who took STAT 217 multiple times out of 925 unique grades over this timeframe; only students’ first attempts in STAT 217 were considered to limit this source of confounding. The data set also included students who did not meet the STAT 217 prerequisite requirement by taking STAT 216; these students were excluded from this analysis as we were not interested in comparing unknown statistics backgrounds to those of interest.

In one case a transfer student who received an A in STAT 216 and had a 4.0 cumulative GPA before taking STAT 217 received an F in STAT 217. The student’s cumulative GPA from our university was based on one term of summer courses and did not include any grades earned from their previous institution. Other information available suggests that this cumulative GPA based only on these summer courses was a poor estimate of the students' true overall performance. This observation was also influential on the models and was removed because of its unusual pattern of courses and results, restricting our scope of inference slightly by not retaining this observation. It does suggest that there might be some rare and extreme patterns of responses that could be pursued qualitatively for better understanding of why students fail at an institution, but a second semester statistics course is not the best course to pursue this research and this was not the purpose of this project. This left us with a data set of 483 observations who completed STAT 217 after taking a version of STAT 216 of interest with 215, 147, and 121 students from the DeVeaux, CATALST, and Lock curricula, respectively.

Our response variable, student grade in STAT 217, was treated three different ways. This variable is recorded as ordinal with 11 levels: *F, D, D+, C-, C, C+, B-, B, B+, A-, A* We converted this into a binary response variable to measure passing rates for students who finish STAT 217 with a final grade of *C-* or higher considered passing. Finally, we converted the ordinal grades into earned GPA points based on the 4.0 scale (*GPA217*). On this scale an *A* earns 4 points, *A-* 3.7 points, *B+* 3.3 points, etc. This was treated as a continuous response variable with potential censoring at 4.0 incorporated into the censored regression model.

The main explanatory variable of interest is the STAT 216 Curriculum (*Curric*) type and whether the type modifies the relationship between grade in STAT 216 and results in STAT 217. Regardless of the modifications made during the time period of interest, all randomization curricula were based on or inspired by techniques developed by the team that created the CATALST materials and were treated as one level of *Curric*. The other two levels are based on the DeVeaux and Lock books. Initially we also considered some of the sub-types of the CATALST curricula as separate levels, but some had small numbers of students that had taken STAT 217 and so were combined to reflect the general types of curricula and courses that students took. Results were generally similar to those presented here when we considered these further subdivisions of the active learning, randomization-based sections based on the differences in materials used. We are also interested in controlling for the students’ general academic performance using their cumulative GPA (*PREVGPA*) before taking STAT 217. We are using STAT 216 GPA as an indication of student learning in STAT 216 although it is also surely related to student general academic performance. Both of these are treated as continuous variables with STAT 216 grade converted to earned GPA points on a 4.0 scale and used in the same way as the grade in STAT 217.

**3.2** **Visual Assessments**

Preliminary visual assessment of the relationships between the quantitatively coded STAT 217 GPA responses and the other variables were made to guide model building in early stages of data exploration. Beanplots (Kampstra, 2008) were used to visually compare the distributions of STAT 217 GPAs among the different 216 curricula (Figure 1). We can see left-skewed distributions for each caused by the censoring of students’ performance based on a maximum of 4.0=*A* caused by inability of the grade scale to measure differences in top ability students and the discreteness of the plus/minus grading system. We can see that the DeVeaux group has the highest mean STAT 217 GPA followed by CATALST and Lock.

Figure 1: Beanplot of STAT 217 GPAs by STAT 216 curriculum type. Small lines are jittered responses, wide lines are means and the shaded areas are mirrored, nonparametric density estimates.

The ggplot2 package (Wickham, 2009) was used along with LOWESS smoothers to visually assess the relationship between STAT 217 GPA and cumulative GPA and the relationship between STAT 217 GPA and STAT 216 GPA, all by curriculum type. We see a strong, positive linear relationship between cumulative GPA and STAT 217 GPA for all curricula (Figure 2). There is a little visual evidence of an interaction between cumulative GPA (*PREVGPA*) and curriculum, as the relationship appears to be relatively consistent for all STAT 216 curricula groups. There is also a strong, positive linear relationship between STAT 216 GPA and STAT 217 grade for all curricula (Figure 3). It appears that the DeVeaux grades start a little higher but that the LOWESS smoother levels off a little a higher levels of the predictor variable. Figures 2 and 3 also highlight the censoring of the responses at 4.0 that become more pronounced for higher levels of both predictor variables.

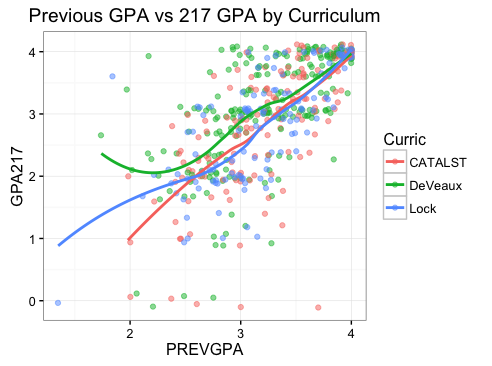
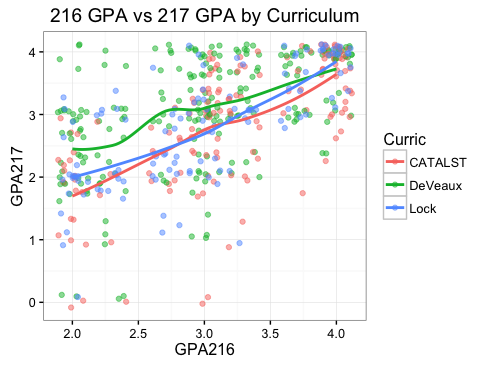


Figure 2: Cumulative GPA vs 217 GPA with LOWESS smoothers for each curriculum.

Figure 3: 216 GPA vs 217 GPA with LOWESS smoothers for each curriculum.

**4 Methods**

We considered four different approaches to modeling these responses, naïve ordinary least squares that assumes continuous and normally distributed responses, a censored regression model that retains the normality assumptions of OLS and incorporates the information on censoring of responses that could or should have been over 4.0, an ordinal response model that treat STAT 217 grades as an 11-level ordinal variable, and a binary response model based on passing STAT 217. The associated assumptions are considered and compared to the characteristics of the data set being analyzed for each model in Section 4. For each, the evidence related to our research questions are addressed and models are refined. Pairwise comparisons between the STAT 217 curricula were explored to estimate differences in mean STAT 217 performance where evidence of an overall difference was found. A Bonferroni adjustment for multiple comparisons multiplies the p-values of the contrasts by the number of pairs (3) compared. Final results are discussed in detail in Section 5.

Along with assessing evidence related to model components, it is important to be able to interpret the model coefficients. To aid in this interpretation, the effects package (Fox, 2003) was used for the OLS model to generate term-plots which are plots of each model component holding the other variables at their means or most common level. We also developed a similar set of plots for the censored regression model results and extended some previous results to our models based on term-plot ideas in Fox and Hong (2009) for cumulative probit regression models.

**4.1 Ordinary Least Square Regression**

Ordinary least square (OLS) regression is one of the most widely used statistical tools. It is remarkably effective for answering questions involving many variables (Ramsey and Schafer, 2002). In this instance we are using it to model the mean STAT 217 grade as a function of students’ cumulative GPA, STAT 216 GPA, their STAT 216 curriculum, and an interaction between curriculum type and STAT 216 GPA. The model for the *i*th response, ***Y****i,* can be specified as ***Y****i*~ N(µ*i*,) starting with

,

where when a student’s STAT 216 course was taught using the Lock, Lock, Lock, Lock, and Lock (2012) text and 0 otherwise; when a student’s STAT 216 course was taught using a version of the CATALST materials and 0 otherwise. The DeVeaux curriculum is the model’s baseline meaning both and are 0 when a student’s STAT 216 course was taught using the DeVeaux text. The coefficients are estimated using ordinary least squares (equivalent to Maximum Likelihood when assuming independent and normally distributed responses) with inferences conducted using t-tests and extra sums of squares F-tests.

Homogeneity of STAT 217 class grade distributions across sections and over time has been attempted through the use of common projects and exams in each semester, including common grading rubrics, but it is possible that there is systematic variation in the grades assigned. We explored using linear mixed models to assess possible section-to-section variability in grades, but the differences associated with the random section effect were found to be minimal and the fixed effect results were relatively unchanged. Because the average grade assigned didn’t seem to vary much across sections and we could not identify extensions of the censored regression model to incorporate random effects, we are not going to present results that account for the variation across sections. It would also be possible to consider instructor effects from either 216 or 217 but many instructors only taught 217 once and the sample size per 216 instructor taking STAT 217 was too small to make considering either in the modeling process viable. This also means that we are going to assume that individual student grades are independent observations - an assumption that these unreported result suggests is reasonable.

However, diagnostic plots of the observed residuals from the OLS model show that there are issues with the assumptions of homogeneity of variance and normality of residuals in this model. There is a clear funneling pattern in the residuals for larger fitted values indicating heteroscedasticity and a slight departure from what is expected from under normality in the residuals in the lower tail of the distribution (Figure 4 (a) and (b), respectively). Both of these issues seem to be caused by the censoring of our response variable at a maximum value of 4,

(b) Normal Q-Q plot of residuals

1. Residuals vs Fitted plot

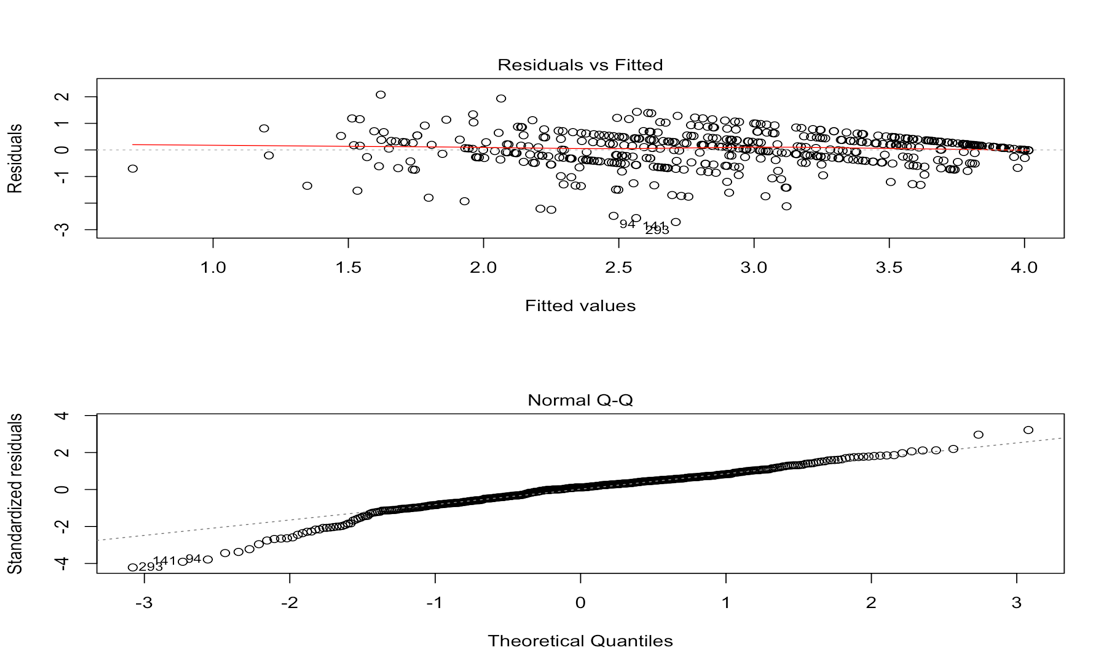


Figure 4: Diagnostic plots for OLS model.

something that OLS is not designed to handle. Inferences can be robust to moderate departures from normality but censored responses can lead to biased estimators as the slope coefficients here are attenuated to 0 because they cannot attain the higher values that the responses should have demonstrated. Generally, linear models are not robust to heteroscedasticity and failing to account for this may provide inaccurate inferences.

Despite these issues, we continue to report inferences from this model, mainly for comparison to later results. An analysis of variance F-test using Type II sums of squares shows moderate evidence of an interaction between Curriculum and STAT 216 GPA (p-value = 0.0335, F(2,476) = 3.4222) suggesting that the influences of the STAT 216 curriculum on the true mean STAT 217 GPA points earned by students depends on the students’ STAT 216 grades, controlling for the students’ previous GPA. Further inferences based on this model are explored in Section 5.1. These inferences are drawn from a model with clearly violated assumptions so should be treated with suspicion. The next two approaches attempt to better deal with the measurement characteristics of the grade response variable.

**4.2 Censored Regression**

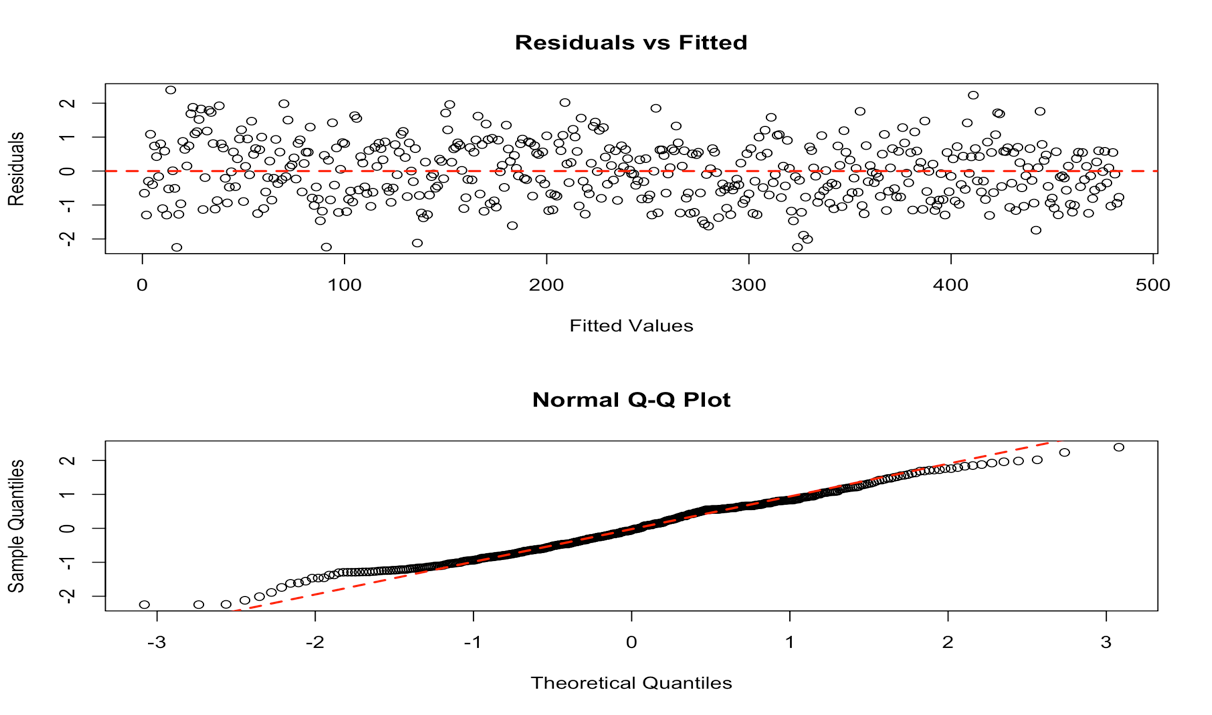
Tobit regression was developed for modeling censored observations as an extension of the regular normal response model framework (Tobin, 1958). This method was originally created to handle left censored responses, particularly a response that could not be observed below 0 and was recorded as 0 instead. The technique was later generalized for any censored response, either left, right, or both left and right censored responses (for more details, see Yee, 2015) and is available in the R package VGAM (Yee, 2016). We most clearly encountered right censoring in the data set with many students that likely should have obtained over 4.0 only able to obtain *A*s in the course, especially students with high cumulative GPAs. In order to account for this censoring, the regular linear model normal likelihood is modified based on:

where ***Y*\*** is a true value that is either observed or unobserved (censored) and ***Y*** is the observed result. The model can be specified as ***Y*\*i** ~ N(µ*i*,) starting with

,

where and are parametrized the same way as the OLS version of the model. The coefficients are estimated using Maximum Likelihood with inferences conducted using z-tests and drop in deviance tests. Tobit regression coefficients are interpreted the same way as regular regression coefficients; however, the linear effect is on the mean of the uncensored latent variable when above the censoring value, not the observed outcome (McDonald and Moffitt, 1980). The (latent) response variable in this case is STAT 217 performance and the observed outcome is the earned grade. For this model, STAT 217 performance is only a latent trait when it is censored.

We are going to assume that individual student grades are independent observations for the same reasons stated above even though we could not explore the impacts of clustering of observations by sections in this modeling framework. The residuals for the initial model now do not suggest heteroscedasticity which suggests that this was an artifact of the previous naïve model for the censored response (Figure 5a). The residuals also follow what is expected under normality reasonably well (Figure 5b). Given these results, the inferences from the censored regression model should be more trustworthy than those from standard linear model discussed previously.



1. Residuals vs Fitted plot

Figure 5: Censored regression model diagnostic plots.

(b) Normal Q-Q plot of residuals

A Likelihood Ratio test for an interaction between Curriculum Type and STAT 216 grade showed little evidence of an interaction (p-value = 0.1321, = 4.06) so it was removed from the model. The additive model,

is used for inferences and is fully explored in Section 5.2.

**4.3 Cumulative Probit Model with an Ordinal Response**

Ordinary regression models for continuous responses, like the OLS method explored in Section 4.1, are special cases of generalized linear models (GLMs). The regression models assume a normal distribution for ***Y*** and model its mean directly. A GLM generalizes regular regression models in two ways: First, it allows ***Y*** to follow any exponential family distribution. For ordered or unordered categorical responses the distribution used is a multinomial or binomial if only two categories are possible. Second, it allows modeling some function of the mean through *link functions* (Agresti, 2007) that relate the systematic component (**)** based on the predictor variables to the mean. One link function, called the *probit link function*, defines the relationship between a probability of observing a response in a single category, , and the predictors in what is called a cumulative link model as

where. This link was chosen for this project because of its easy interpretability, as it provides inferences for a normally distributed underlying latent response variable. The function applied to gives the standard normal z-score at which the left-tail probability equals . Instead of predicting the average GPA points earned in STAT 217, these models are used to estimate the probability of receiving a particular grade or one lower.

Generalizations of the probit GLM can be used to model categorical responses with more than just two levels. When response categories have a natural ordering, the probit model can use the ordering to generate a model for the probabilities of being at or below a category. This compares to multinomial models for unordered categorical responses where J-1 sub-models are needed to predict probabilities for the J categories. The cumulative link models are simpler models with simpler interpretations and potentially greater power than the baseline-category models (Agresti, 2007) if the ordinal response model assumption and the slopes for all the related categories are reasonable. These models estimate the cumulative probability for ***Y***, the probability of ***Y*** occurs at or below a certain level of the ordinal response. For the outcome category j, the cumulative probability is

,

where the probability ***Y*** falls in level *j*. The cumulative probabilities are a result of the ordered response, where The *cumulative probits* are

One motivation for the cumulative probability structure relates to a model for an assumed underlying continuous variable (Agresti, 2007). Here this variable is STAT 217 performance. We believe that students’ observed STAT 217 grade, treated as an ordinal response, is a categorical realization of the underlying latent trait, STAT 217 performance. Let ***Y\**** denote this latent trait, and denote the *cut points* or *thresholds* for the continuous scale of ***Y\**** such that

where ***Y*** is the observed response and *j* is defined as above. If we assume that ***Y\**** is distributed normally with constant variance, the probit model holds for cumulative probabilities (Agresti, 2007). This is a very useful result as we can estimate relationships with the true mean on the latent trait (STAT 217 performance) scale. The coefficients are estimated using Maximum Likelihood with inferences conducted using z-tests and Likelihood Ratio tests.

Our starting model is

,

where *j = F, D, D+, C-, C, C+, B-, B, B+, A-, A* and and are parametrized the same way as above. A Likelihood Ratio test gave moderately weak evidence for an interaction between Curriculum and STAT 216 GPA (p-value = 0.0876, = 4.87), so it was removed. The reduced model,

is used for inferences and is fully explored in Section 5.4.1.

If our response variable had only two levels, these models reduce to commonly used models for a binomially distributed response such as logistic regression (if a logit link function is used) or, for consistency with our other models, the probit-link binomial GLM. We use this to explore the impacts of the different curricula on passing rates for students who completed STAT 217, where a passing grade is a *C-* or higher. We use the same set of explanatory variables in our fullest model, starting with

where all explanatory variables are parametrized the same as in previous models. A Likelihood Ratio test showed no evidence for the interaction between Curriculum and STAT 216 GPA (p-value = 0.5938, = 1.04) or evidence of differences in mean STAT 217 performance among the STAT 216 curricula (p-value = 0.2479, = 2.79) controlling for the two grade-based predictors so these terms were removed. The reduced model,

,

is used for inferences and is fully explored in Section 5.4.2.

**4.4** **Additive OLS Regression**

Since the censored regression and ordinal response regression both suggest an additive model we decided to explore the same additive model fit using OLS with

and ***Y****i*~ N(µ*i*,) so that we can directly compare it with the results from the censored response model. These results are discussed further in Section 5.3.

**5 Results**

**5.1 OLS Regression Results**

As stated in Section 4.1, there is evidence of an interaction between STAT 216 Curriculum and STAT 216 GPA. There is also strong evidence of a linear relationship between students’ cumulative GPA and the true mean STAT 217 GPA points earned by students (p-value < 0.0001, F(1,146) = 115.8), controlling for the 216 GPA and curriculum interaction. The model’s individual coefficients are fully summarized in Table 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficient | Estimate | Standard Error | P-value |
| Intercept | **-0.513** | **0.25203** | **0.0423** |
| GPA 216 | **0.244** | **0.07829** | **0.0020** |
| PREVGPA | **0.885** | **0.08224** | **< 0.0001** |
| CATALST | **-0.996** | **0.33048** | **0.0027** |
| LOCK | **-0.945** | **0.33599** | **0.0051** |
| GPA216 : CATALST | **0.229** | **0.10458** | **0.0288** |
| GPA216 : LOCK | **0.239** | **0.10993** | **0.0299** |

Table 1: Summary output from OLS model.

The slope for the STAT 216 GPA is steeper for both the CATALST and Lock curricula than DeVeaux (Table 1). This implies that the students’ STAT 216 result is more impactful on their average 217 grade in the curricula with randomization methods than the original consensus curriculum. But also note that the DeVeaux group started closest to the upper measurable limit – the difference in slopes might be partially attributed to the censoring, especially when compared to the results from the censored regression model. The effects package (Fox, 2003) was used to aid in visualization of these relationships (Figure 6), displaying the estimated effects while holding other model components constant. The difference in the slope for the Deveaux group (left sub-panel of the second panel) is most apparent. The importance of previous GPA is also very clear.

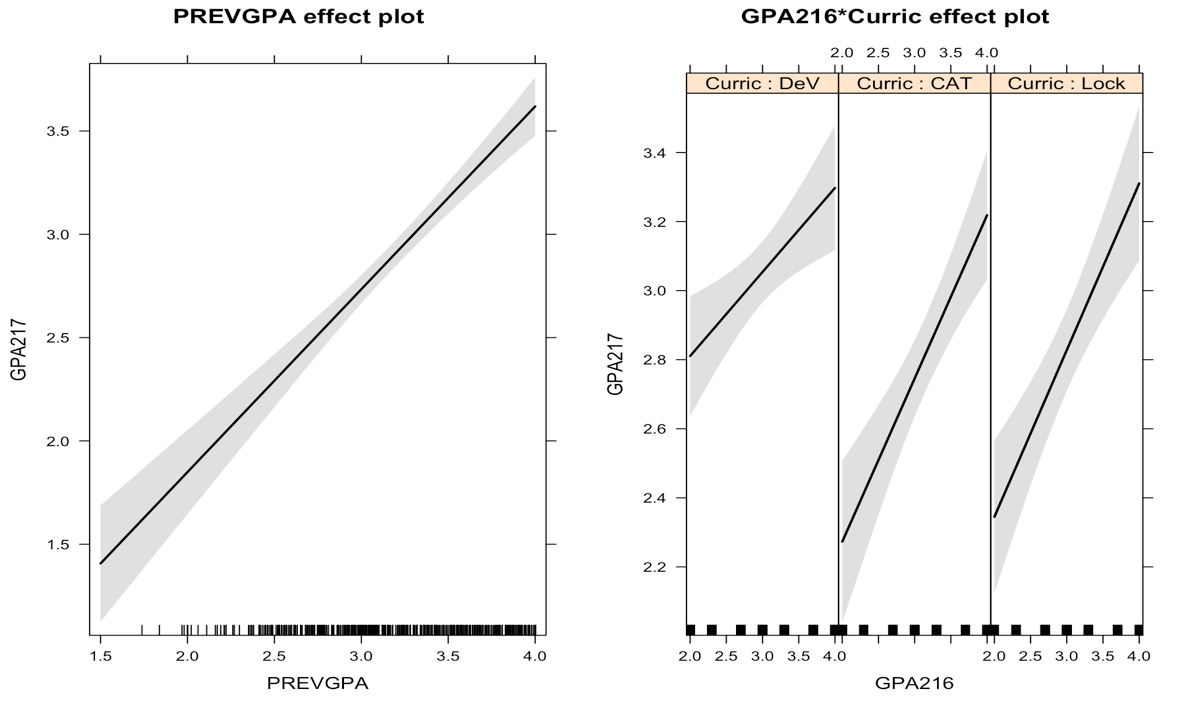


Figure 6: Term-plot for PREVGPA (left) and STAT 216 GPA by Curriculum (right).

Even though this model suggests that the impacts of 216 grade vary by curriculum, the other, more trustworthy models do not. For comparison to those results, the additive OLS model,

provides the estimated coefficients provided in Table 2. There is strong evidence of at least one difference in mean STAT 217 GPA points earned among the STAT 216 curricula (p-value < 0.0001, F(2,478) = 9.82), controlling for STAT 216 GPA and cumulative GPA. There is also strong evidence of a linear relationship between students’ STAT 216 GPA and the true mean STAT 217 GPA points earned by students (p-value < 0.0001, F(1,478) = 36.93), controlling for cumulative GPA and Curriculum. And there is strong evidence of of a linear relationship between students’ cumulative GPA and the true mean STAT 217 GPA points earned by students (p-value < 0.0001, F(1,478) = 117.35), controlling for STAT 216 GPA and Curriculum.

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficient | Estimate | Standard Error | P-value |
| Intercept | **-0.924** | **0.19774** | **< 0.0001** |
| GPA 216 | **0.372** | **0.06125** | **< 0.0001** |
| PREVGPA | **0.893** | **0.08251** | **< 0.0001** |
| CATALST | **-0.296** | **0.07163** | **< 0.0001** |
| LOCK | **-0.231** | **0.07529** | **0.0023** |

Table 2: Summary output from additive OLS model.

**5.2 Censored Regression Results**

When moving to treating the responses as right censored at 4.0, we did not find evidence of an interaction between 216 grade and curriculum, but there is strong evidence of at least one difference in mean STAT 217 performance among the STAT 216 curricula (p-value < 0.0001, = 19.39), controlling for STAT 216 GPA and cumulative GPA. There is also strong evidence of a linear relationship between students’ STAT 216 GPA and the true mean STAT 217 GPA points earned by students (p-value < 0.0001, = 38.85), controlling for cumulative GPA and Curriculum. And there is strong evidence of a linear relationship between students’ cumulative GPA and the true mean STAT 217 GPA points earned by students (p-value < 0.0001, = 121.38), controlling for STAT 216 GPA and Curriculum. The model’s individual coefficients are fully summarized in Table 3.

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficient | Estimate | Standard Error | P-value |
| Intercept | **-1.866** | **0.25301** | **< 0.0001** |
| GPA 216 | **0.477** | **0.07663** | **< 0.0001** |
| PREVGPA | **1.144** | **0.10366** | **< 0.0001** |
| CATALST | **-0.380** | **0.09075** | **< 0.0001** |
| LOCK  Table 3: Summary output from censored regression model. | **-0.275** | **0.09525** | **0.0042** |

The mean STAT 217 performance is estimated to increase by 0.477 GPA points for a one-point increase in STAT 216 GPA and 1.144 GPA points for a one-point increase in cumulative GPA, controlling for STAT 216 curriculum. The results also imply that students who took STAT 216 from the consensus curriculum perform better in STAT 217 on average than the other two versions of the course. The term-plots also show that for DeVeaux students (the most common STAT 216 curriculum) and for the average 216 grade of 3.032, the predicted mean 217 grade exceeds 4.0 (Figure 7). This illustrates the importance (and oddity) of the censored regression model as the predictions exceed the maximum observable value on the original scale. The results from the other categories that had lower intercepts are less impacted by censoring. These results also suggest why the slope in the DeVeaux group in the OLS model with an interaction was smaller than in this model where the data did not provide much evidence for different slopes for the pre-requisite class grade based on curriculum type.

The mean STAT 217 GPA for students from the DeVeaux curriculum is estimated to be 0.38 and 0.275 points higher than students from the CATALST and Lock curricula, respectively, with Bonferroni adjusted p-values of < 0.0001 and 0.0126, respectively, after controlling for STAT 216 GPA and cumulative GPA. The mean STAT 217 GPA for students from the CATALST curriculum is estimated to be 0.107 points higher than students from the Lock curriculum, with a Bonferroni adjusted p-value of 0.897, after controlling for STAT 216 GPA and cumulative GPA.

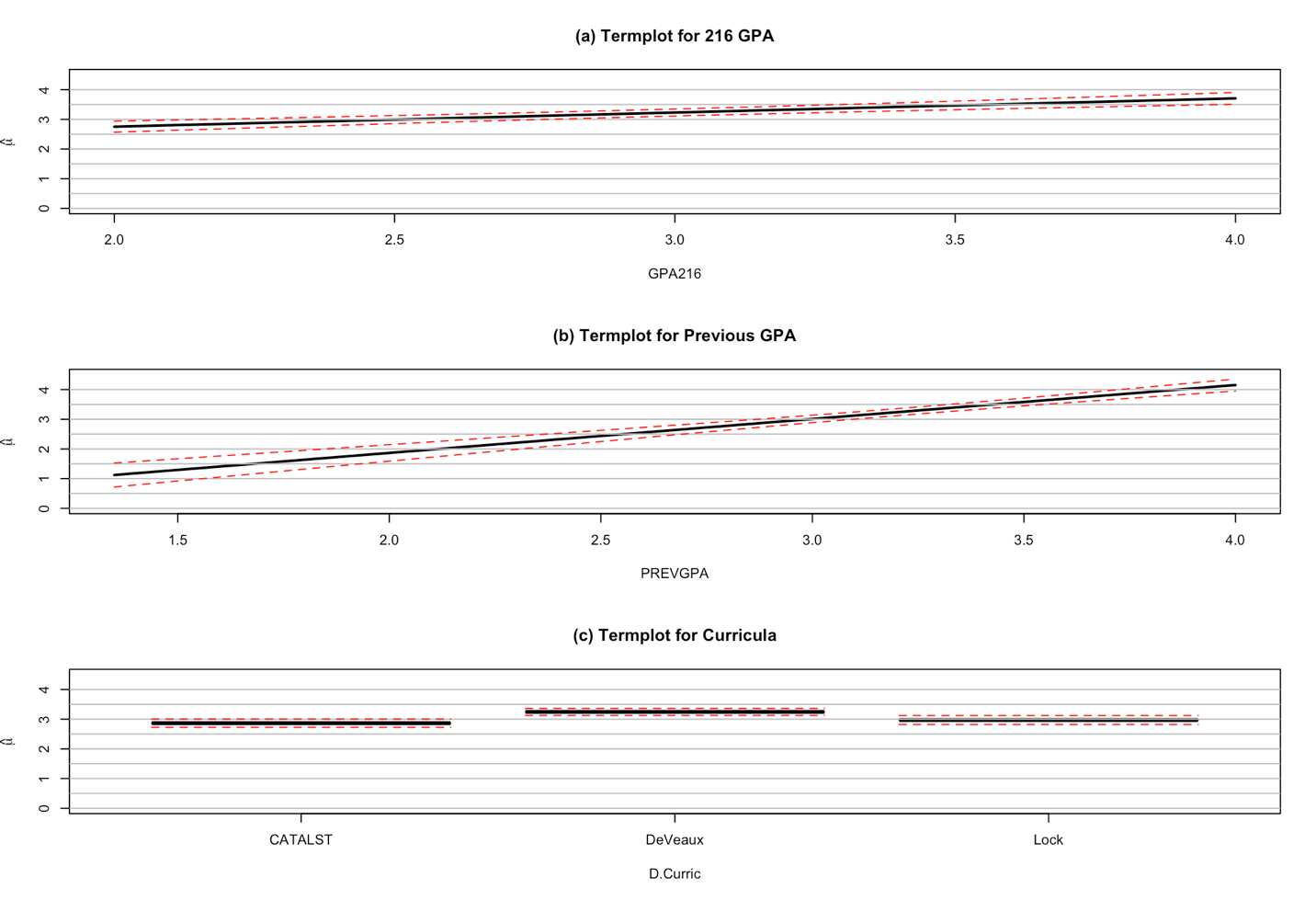


Figure 7: Term-plots for censored regression model with 95% confidence intervals.

**5.3 OLS and Censored Regression Comparison**

It is clear from all of the models made that the different STAT 216 curricula are impacting student performance in STAT 217 and different intercepts are necessary. However, the OLS model is the only method considered that suggests that an interaction between STAT 216 curriculum and GPA needs to be retained. This is likely a result of the slope coefficients being attenuated to 0 because they cannot attain the higher values that the responses should have demonstrated.

The estimated differences between curricula from the additive versions of the censored and OLS models are similar. Their main differences come from their estimation of the impacts of the quantitative explanatory variables since the censored model is able to predict above the response’s observed upper bound. The estimated slope for STAT 216 GPA increases from 0.372 in the OLS model to 0.477 in the censored model, suggesting estimated change in mean 217 score of 0.744 or 0.954 over the 2 point range in observed 216 grades. The estimate slope for cumulative GPA increases from 0.894 in the OLS model and 1.144 in the censored model. This is a noticeable change in the size of the slope and corresponds to predicting almost 0.6 GPA points more change in the mean 217 grade over the 2.3 range of results in cumulative GPA when using the censored response model (2.05 vs 2.63). In either model the results for both 216 grade and cumulative grade show much larger impacts on the 217 grades than the curriculum types which had a maximum estimated difference of 0.38 points.

To visualize the change in results, we provide component plus residual plots (crPlots) in Figure 8. These plots are used to assess and understand estimates of individual coefficients in multiple linear regression (Fox and Weisburg, 2011). These crPlots display residuals after correcting for other model components with the estimated marginal effect from least squares. The estimated marginal effects from the censored regression model were added in Figure 8 for comparison. The censoring in the responses shows up with an accumulation of points in the upper right quadrant of the second panel.

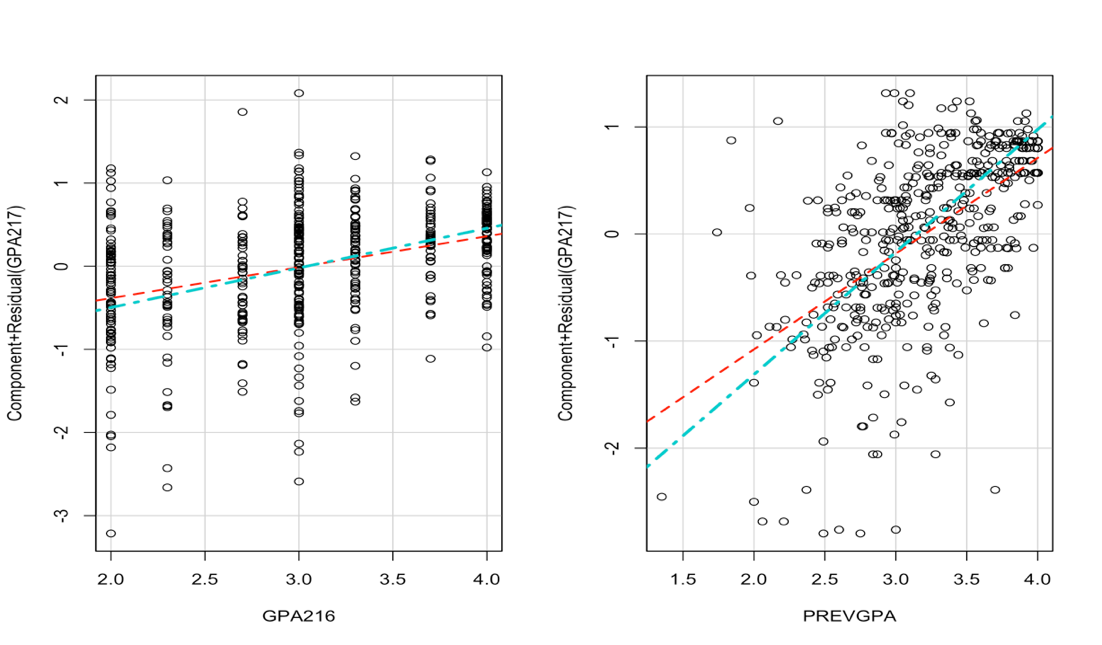
****

Figure 8: Component plus residuals plots for quantitative effects from OLS and censored regression models.

**5.4 Cumulative Probit Models**

**5.4.1 Ordinal Response**

When the response is treated as an ordinal categorical variable there still was no evidence of an interaction between STAT 216 GPA and curriculum, but there is strong evidence of at least one difference in mean STAT 217 performance among the STAT 216 curricula (p-value < 0.0001, = 20.21), controlling for STAT 216 GPA and cumulative GPA. There is also strong evidence of a linear relationship between students’ STAT 216 GPA and the true mean STAT 217 performance (p-value < 0.0001, = 42.36), controlling for cumulative GPA and Curriculum. And there is strong evidence of of a linear relationship between students’ cumulative GPA and the true mean STAT 217 performance (p-value < 0.0001, = 113.88), controlling for STAT 216 GPA and Curriculum. The model’s individual coefficients are fully summarized in Table 4 and term-plots containing predicted probabilities in each category are displayed in Figure 9.

The slope coefficients for this model can be interpreted as changes in the mean of the continuous latent trait that is defined to have a standard deviation of 1 because we are using the probit version of the cumulative link model (Agresti, 2010). The mean STAT 217 performance is estimated to increase by 0.642 standard deviations for every 1 point increase in STAT 216 GPA points, after controlling for STAT 216 curriculum and cumulative GPA. The mean STAT 217 performance is estimated to increase by 1.472 standard deviations for every 1 point increase in cumulative GPA points, after controlling for STAT 216 GPA and curriculum. Over the range of the observed predictors (2 points for 216 grade and 2.5 for cumulative GPA), the 216 GPA suggests a change of 1.28 standard deviations and the cumulative GPA a change of 3.7 standard deviations which are much larger than the differences estimated based on curriculum. Table 4 also contains the estimated thresholds between the categories that would provide the estimated probabilities for a DeVeaux response when both quantitative predictors were 0s.

The mean STAT 217 performance for students from the DeVeaux curriculum is estimated to be 0.48 and 0.39 standard deviations higher than students from the CATALST and Lock curricula, respectively, with Bonferroni adjusted p-values of < 0.0001 and 0.0041, respectively, after controlling for STAT 216 GPA and cumulative GPA. The mean STAT 217 performance for students from the CATALST curriculum is estimated to be 0.092 standard deviations higher than those from the Lock curriculum, with a Bonferroni adjusted p-value of 1.

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficient | Estimate | Standard Error | P-value |
| GPA 216 | **0.642** | **0.0989** | **< 0.0001** |
| PREVGPA | **1.472** | **0.1392** | **< 0.0001** |
| CATALST | **-0.481** | **0.1157** | **< 0.0001** |
| LOCK | **-0.390** | **0.1215** | **0.0014** |
| Thresholds | Estimate | Standard Error | z-score |
| F|D | **3.52** | **0.3651** | **9.643** |
| D|D+ | **4.03** | **0.3552** | **11.343** |
| D+|C- | **4.21** | **0.3539** | **11.889** |
| C-|C | **4.53** | **0.3535** | **12.817** |
| C|C+ | **5.28** | **0.3614** | **14.617** |
| C+|B- | **5.67** | **0.3677** | **15.331** |
| B-|B | **5.93** | **0.3738** | **15.853** |
| B|B+ | **6.74** | **0.3933** | **17.137** |
| B+|A- | **7.12** | **0.4011** | **17.747** |
| A-|A | **7.59** | **0.4112** | **18.449** |

Table 4: Summary output from ordinal response model.

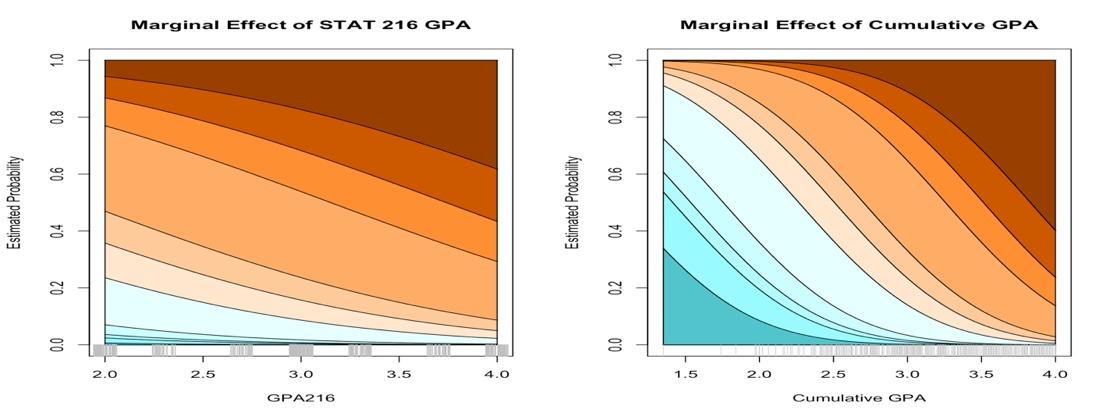
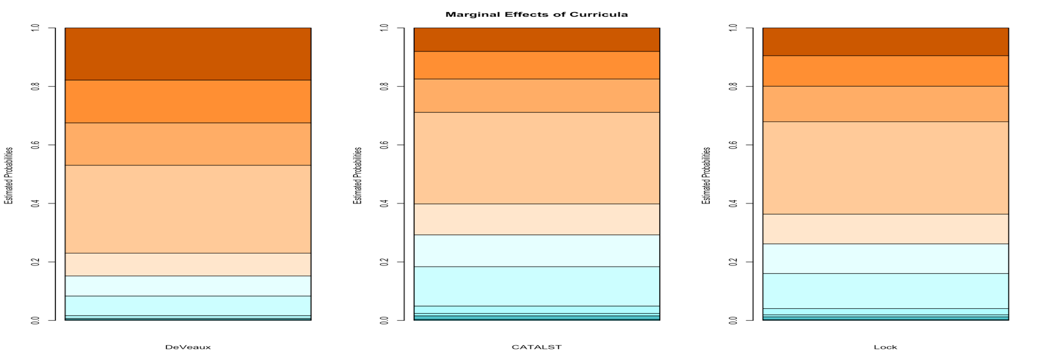


Figure 9: Term-plots of STAT 216 GPA (top left), cumulative GPA (top right), and the STAT 216 curricula (bottom) from ordinal response model.



The ordinal model also has a qualitative connection to the censored regression model. By defining a set of ordered categories, we essentially define all observations as censored to some degree or another because all observations are allocated to a category to represent their real level that is assumed to exist on a continuum in the latent trait scale. The censored regression model assigned a 4.0 to all observations that could have a 4.0 or above on the latent trait scale. In the ordinal model, the highest responses over the last estimated threshold are assigned to the highest ordered category. In comparing the two final models, the inferences are extremely similar. The z-statistics are all within 0.5 z-score units with no particular pattern to which model produced larger test statistics, suggesting that the two models are performing very similarly even though the scale of the slope coefficients are slightly different.

**5.4.2 Binary Response**

When the response is treated as a binary categorical variable there still was not any evidence of an interaction between STAT 216 GPA and curriculum or a STAT 216 curriculum impact, but there is marginal evidence of a STAT 216 GPA impact on passing STAT 217 passing (p-value = 0.0362, = 4.39) and stronger evidence for cumulative GPA explaining passing the course or not (p-value = 0.0003, = 12.96). The model’s individual coefficients are fully summarized in Table 5 and predicted probabilities of passing are displayed in Figure 10.

Because we are using a probit link, these models can continue to be interpreted in terms of a latent 217 performance trait although it is estimated based solely on being above or below the passing cutoff. The mean STAT 217 performance is estimated to increase by 0.408 standard deviations for every one point increase in earned STAT 216 GPA points. The mean STAT 217 performance is estimated to increase by 0.892 standard deviations for every one point increase in cumulative GPA points. Figure 10 shows the estimated changes in probabilities of passing as functions of STAT 216 grade and cumulative GPA, which are similar to the results that can be obtained from the cumulative probit link model above. Note that the ordinal response model used before estimates the probability of passing or not (being less than a C- versus being over it) as part of the modeling process but this result is based on estimating a slope coefficient that pertains to all steps in the 11-level ordinal variable so it may not match the results from this simpler model. The bigger difference in these models is that we lose power by focusing solely on passing/not and are not able to detect differences based on the grade in 216 or curriculum type.

|  |  |  |  |
| --- | --- | --- | --- |
| Coefficient | Estimate | Standard Error | P-value |
| Intercept | **-2.277** | **0.6556** | **0.0005** |
| GPA 216 | **0.408** | **0.1983** | **0.0399** |
| PREVGPA | **0.892** | **0.2528** | **0.0004** |

Table 5: Summary output from binary response model.

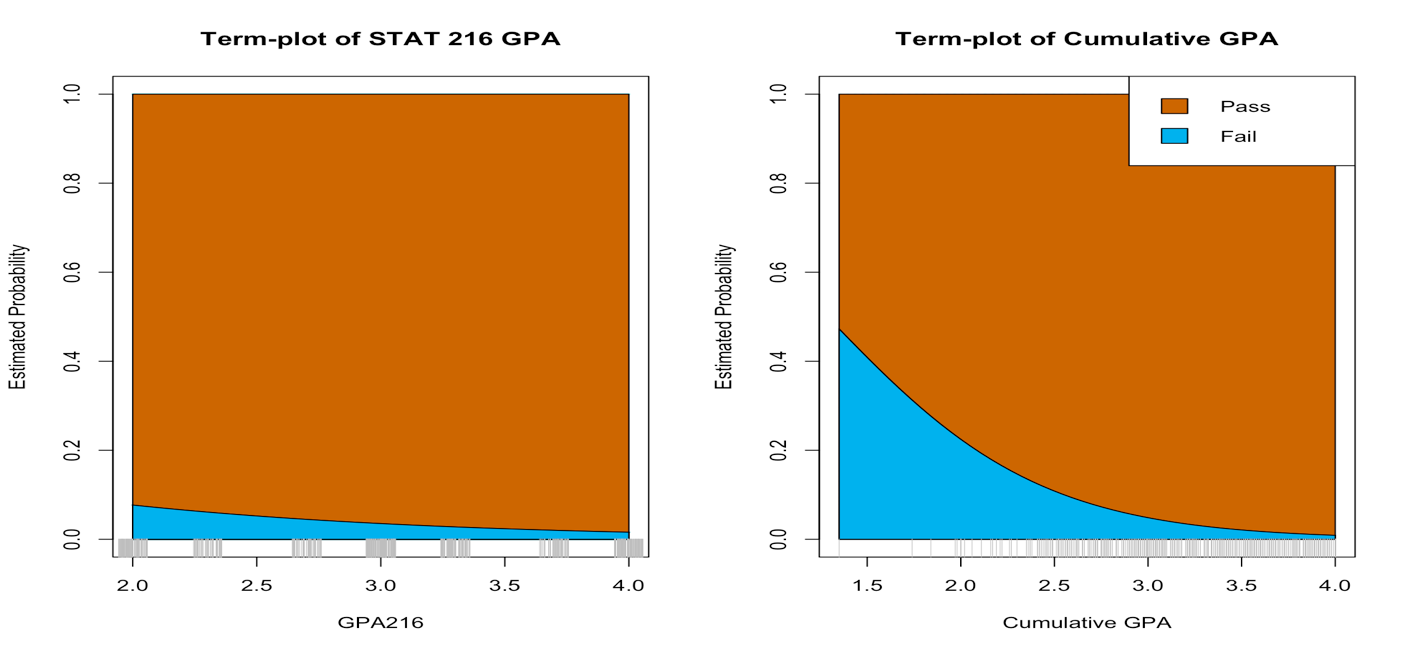


Figure 10: Term-plots of STAT 216 GPA (left) and cumulative GPA (right) for binary response model.

**6 Conclusions**

This research assessed an important question in statistics education: whether students with different curricula have different results in their first attempt at a second statistics course. The move to randomization techniques have been promoted to develop higher levels of statistical thinking in students and this should result in better performance in a second course. We found results that seem to suggest that consensus methods provide better outcomes in the following course than randomization-based introductions to statistics on average. We also found that the grade in the previous course and overall previous GPA were more important predictors of outcomes in the next course, having much larger impacts on results than the curriculum used.

When working with grades as response variables, there are a few important considerations in developing valid and useful statistical models. This research was focused on modeling STAT 217 performance as a measure of how well STAT 216 classes prepared students for a second statistics class; however, STAT 217 performance is a latent trait that we could not directly measure. What we can measure instead is students’ final grade in STAT 217. This response can be parametrized and modeled a few different ways: as a censored quantitative variable with an upper measurement threshold, in which case we are modeling STAT 217 grade points earned unless it is censored, as an ordinal response where STAT 217 performance is assumed to be an unobserved continuous latent trait, or collapsing the ordinal response into a binary response where students either pass or do not pass the class.

Sometimes the focus in gateway classes is on the passing rates for students using a binary version of the grade response. There are benefits to treating the responses in this fashion with the opportunity to then also include students who withdrew from the course and earned a *W* along with those that persisted in the course but did not pass it. A failing grade would then include grades *W, D+, D,* and *F*, what is sometimes called the “WDF rate”. We also explored the same models considered above with WDF or pass treated as a binary response and found similar results to those discussed in Section 5.4.2 which suggested that STAT 216 grade and cumulative GPA are related to WDF rates in STAT 217.

If we had decided to include students who withdrew from STAT 217 we would have lost our ability to treat STAT 217 grade as an ordinal variable. There is a natural and meaningful order for earned grades in a class, but there isn’t a natural place for *W*s since there is no information about why any particular student withdrew from the class. To incorporate *W*s into the categorical response variable we would need to treat earned grade as coming from an unordered multinomial response variable with 12 levels. Multinomial models can be very informative when they have at most a few response levels, but a model with 12 response levels without a natural order loses the practical interpretability that an ordinal response model offers. So if one wants to consider including *W*s in the responses, we would suggest some sort of collapsing of the response scale down to a binary response considering WDF vs passing. As we showed in Sections 5.4.1 and 5.4.2, we lost power to detect the influences of explanatory variables when we collapsed our ordinal response to a binary one. Agresti (2007) notes this can happen because of the potential loss of efficiency. By modeling all the steps in the ordinal scale, we are able to detect the more subtle differences in the influences of the different curricula and more clearly detect the impacts of the grade in STAT 216. These results suggest that the focus might be better placed on the more detailed student performance available in the recorded grade in courses rather than the less refined WDF rate.

Another common practice with ordinal variables is to convert the categories into quantitative responses as we did to generate GPA points for our OLS and censored regression models. Some caution against this approach as it is in a sense over-confident since the data are assumed to contain more information than they actually do. Observations on an ordinal scale are classified in ordered categories just like in its quantitative re-coding, but the distance between the categories is generally unknown. By using linear models, the choice of coding imposes assumptions about the distance between the response categories (Christensen, 2015). GPA points earned is how cumulative GPAs are calculated so there is no other obvious choice for the quantitative recoding of the ordinal response categories.

One criticism of our models is that we did not treat previous GPA or STAT 216 grade as censored variables but did treat STAT 217 results as censored. Some have suggested (Hosmer and Lemeshow, 2013) adding an indicator variable as an additional predictor variable for right or left censoring that is 1 for each observation that is censored of each type and 0 otherwise. Cumulative GPAs in this population do not appear to be censored because of the accumulation of information in each measurement and because we only observed 18 4.0 cumulative GPAs in the analyzed data set. STAT 216 grades more clearly appear to be censored (Figure 3) as there are 91 students at the upper limit of the x-axis of getting an *A* in STAT 216. This indicator approach essentially estimates a different mean response for all observations that are measured at the censoring limit(s). This is not nearly as elegant as the way censored responses are handled by the censored regression model and was not incorporated into these models as it would dramatically complicate the consideration of interactions with curriculum type. This censoring of STAT 216 grades may explain why it had less of an impact on responses than previous GPA. If we could have observed the differences in the top students, we might have seen a larger impact on the responses.

Students’ cumulative GPA appears to have the most influence on STAT 217 grades, which is a reasonable conclusion if we think about cumulative GPA being a measure of overall academic ability – better students tend to do better in any course. The grade in any version of STAT 216 also matters in the predicted grades for students that complete the second course, so some aspects of how much they learn in the preparatory class do matter. But doing better or worse in the different versions of the pre-requisite course did not seem to change outcomes in the next course, implying that the difference between a *B* and an *A* is the same on potential outcomes regardless of curriculum used. We did detect evidence of differences based on curriculum with the consensus curriculum producing higher average grades in STAT 217 after controlling for the other measured aspects of the students. These differences were much smaller than those based on the grade-based predictors but were still present in all but the Pass/Fail model in Section 5.4.2. This research is not intended to suggest that the randomization-based curricula are causing decreased student performance. This was not a randomized experiment and students have had some opportunity to self-select the flavor of STAT 216 that they preferred. It is possible that students could have been attracted to the active learning techniques because they were initially promoted as easier alternatives to a course that was notoriously difficult to pass. Or that there has been some sort of change in the students over this timeframe and only randomization-based courses have been offered in the last year.

There have also been suggestions that randomization methods may enhance student enjoyment of, confidence in their skills in, or attraction to statistics. It is also possible that the randomization curricula could be better preparing students for learning statistical computing but that is only a small part of the grade in STAT 217. This research has only focused on the grade outcomes in the second course and has not explored the reasons students take the course or how they feel about their skills or statistics in general after taking those courses. If the active-learning, randomization curriculum has higher pass rates, it allows more students entry into higher levels of statistics and that is an exciting outcome, even if they get slightly lower grades in our setting when taking the next course. Further research that attempts to more clearly identify any deficiencies, misconceptions, or other differences in characteristics of students based on their backgrounds is needed to more completely understand these results.

**References:**

Agresti, A. *An Introduction to Categorical Data Analysis*. Second ed. New York: Wiley, 2007.

Agresti, A. "Other Ordinal Multinomial Response Models." *Analysis of Ordinal Categorical Data*. Second ed. Hoboken, NJ: Wiley, 2010.

Christensen, R. H. B., (2015). ordinal - Regression Models for Ordinal Data. R package version 2015.6-28. <http://www.cran.r-project.org/package=ordinal/>.

Fox, J., (2003). Effect Displays in R for Generalised Linear Models. *Journal of Statistical Software*, 8(15), 1-27. URL <http://www.jstatsoft.org/v08/i15/>.

Fox, J. and Hong, J., (2009). Effect displays in R for multinomial and proportional-odds logit models: Extensions to the effects package.*Journal of Statistical Software* **32:1**, 1–24, <<http://www.jstatsoft.org/v32/i01/>>.

Fox, J. and Weisberg, S., (2011*). An R Companion to Applied Regression*, Second Edition. Thousand Oaks CA: Sage. URL: <http://socserv.socsci.mcmaster.ca/jfox/Books/Companion>

Hosmer, D. W., Jr., Lemeshow S., and Sturdivant, R. X., (2013) *Applied Logistic Regression*. Third ed. New York: Wiley. Print.

Kampstra, P., (2008). Beanplot: A Boxplot Alternative for Visual Comparison of Distributions. *Journal of Statistical Software*, Code Snippets 28(1). 1-9. URL <http://www.jstatsoft.org/v28/c01/>.

McDonald, J. F. and Moffitt, R. A., (1980). The Uses of Tobit Analysis. *The Review of Economics and Statistics* Vol 62(2): 318-321.

Pruim, R., Kaplan, D., and Horton, N., (2015). mosaic: Project MOSAIC Statistics and Mathematics Teaching Utilities. R package version 0.13.0. [https://CRAN.R-project.org/package=mosaic](https://cran.r-project.org/package=mosaic)

R Core Team, (2015). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL [https://www.R-project.org/](https://www.r-project.org/).

Ramsey, F. L., and Schafer, D. W., *The Statistical Sleuth: A Course in Methods of Data Analysis*. Second ed. Australia: Duxbury/Thomson Learning, 2002. Print.

RStudio Team, (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA

Tobin, James (1958). "Estimation of relationships for limited dependent variables". *Econometrica* **26** (1): 24–36.

Wickham, H., (2009) *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Yee, T. W., (2015). VGAM: Vector Generalized Linear and Additive Models. R package version 1.0-0. URL [http://CRAN.R-project.org/package=VGAM](http://cran.r-project.org/package=VGAM)

Yee, T. W., (2015). *Vector Generalized Linear and Additive Models: With an*

*Implementation in R*. New York, USA: Springer.