STAT 306 Group Project:
Exploring Significant Factors Affecting GPA for High School Students

Group A4 (Anthony Leong, Deacon Sowerby, Irteza Shamsi, Sethu Kondapavuluru)

Introduction

The academic success of a high school student is affected by many factors, such as study habits, absenteeism, and parental influence. Many such factors may be beyond the control of the student, so educators should seek to understand these relationships to improve student outcomes.

This study aims to look into how the Grade Point Average (GPA) of a student is affected by their weekly study schedules and absences. Furthermore, we will investigate whether parental support and the education level of the parents affect these associations. How do the quantity of study time and the number of absences affect the GPA of a student, and to what extent do parents play a role? Our goal in investigating this topic is to find practical information that can improve education outcomes for students, and assist teachers and counselors in guiding their educational journey.

The dataset, which comprises comprehensive data on 2,392 high school students, was obtained from Kaggle and will be utilized for this investigation. The dataset includes information on extracurricular activities, academic performance, study habits, and demographics. Several of the include variables are beyond the scope of this study, but the variables we will choose to focus on are weekly study time, absenteeism, parental support, and parental education as the explanatory variables, and GPA as the response variable.

The descriptions for our chosen variables are as follows:

- **GPA**: Grade Point Average, on a scale from 0.0 to 4.0.
- **Weekly Study Time**: Weekly study time in hours.
- **Absences**: Number of absences during the school year.
- **Parental Education**: The education level of the student's parents, with levels as follows:
    - 0: None
    - 1: High School
    - 2: Some College
    - 3: Bachelor's
    - 4: Higher
- **Parental Support**: The level of parental support, with levels as follows:
    - 0: None
    - 1: Low
    - 2: Moderate
    - 3: High
    - 4: Very High

Analysis

A key fact about our response variable (GPA) is that it has a range of [0.0, 4.0]. We will therefore first need to analyze the distribution of GPA as well as how GPA varies with respect to each individual explanatory variable, as these will influence how we choose to fit a linear model for the GPA of a student.

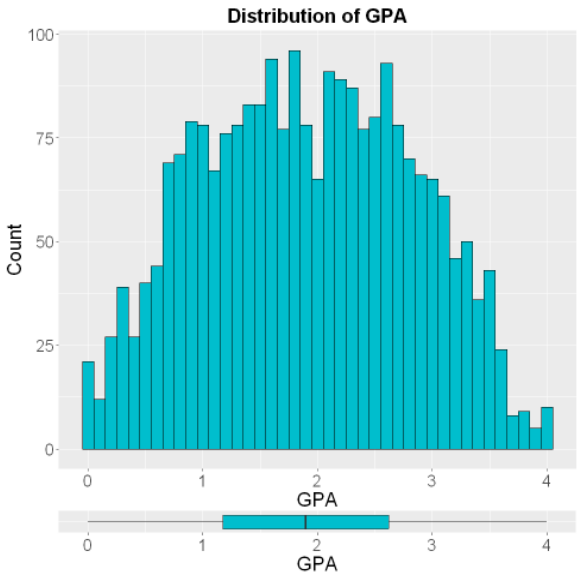

Figure 1. A histogram and a boxplot showing the distribution of GPA.

The majority of the observations have a GPA between 0.5 and 3.5. Although there are a lot less observations in the extremes of the GPA distribution, a GPA of exactly 0.0 or exactly 4.0 cannot realistically be considered an outlier.

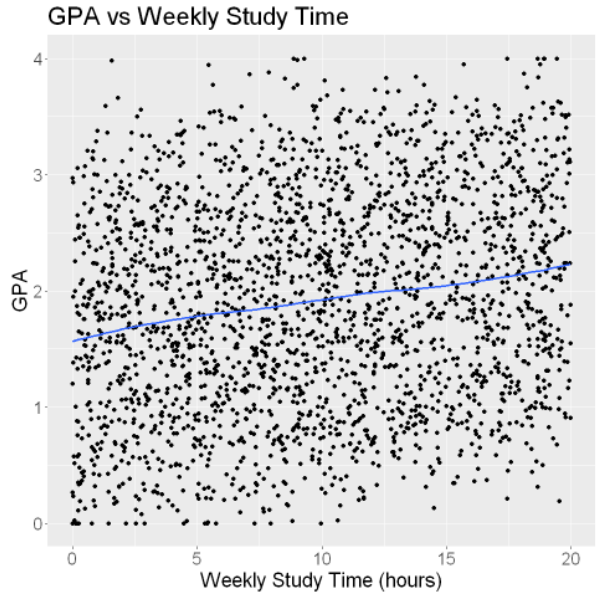

Figure 2. A scatterplot of GPA plotted against weekly study time. The blue line is the estimated conditional mean of GPA on weekly study time.

GPA varies significantly with respect to weekly study time, but GPA does seem to increase at a consistent linear rate as weekly study time increases.

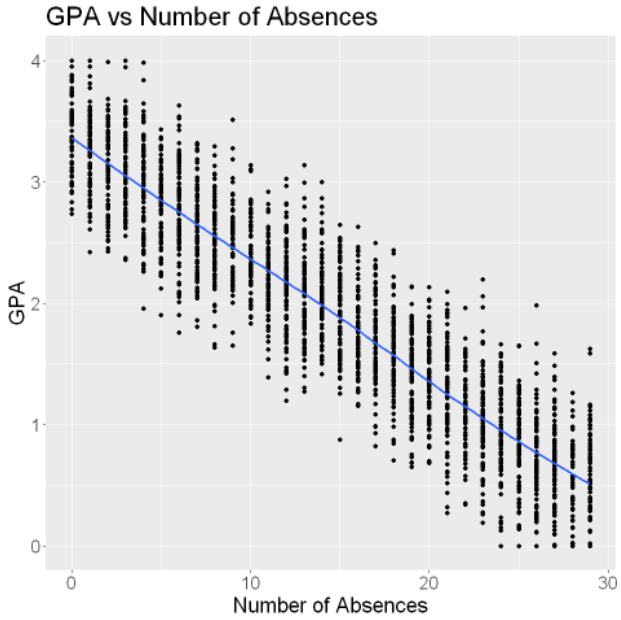

Figure 3. A scatterplot of GPA plotted against the number of absences during the school year. The blue line is the estimated conditional mean of GPA on the number of absences.

The above plot provides very strong evidence to conclude that GPA depends linearly on the number of absences. It is important to note that as the number of absences increases, the mean of GPA conditioned on the number of absences appears to linearly approach the minimum possible value of 0.0.

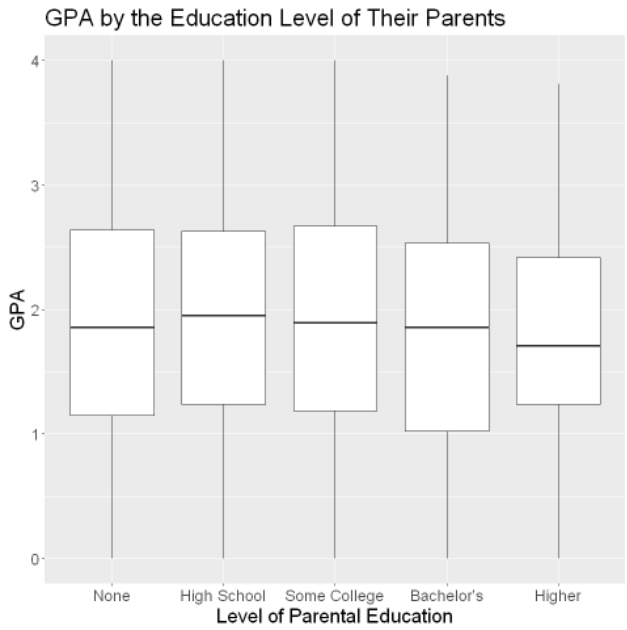

Figure 4. A boxplot of the distribution of GPA across different levels of education received by the student's parents. In the dataset, the levels are recorded as 0 to 4 from left to right.

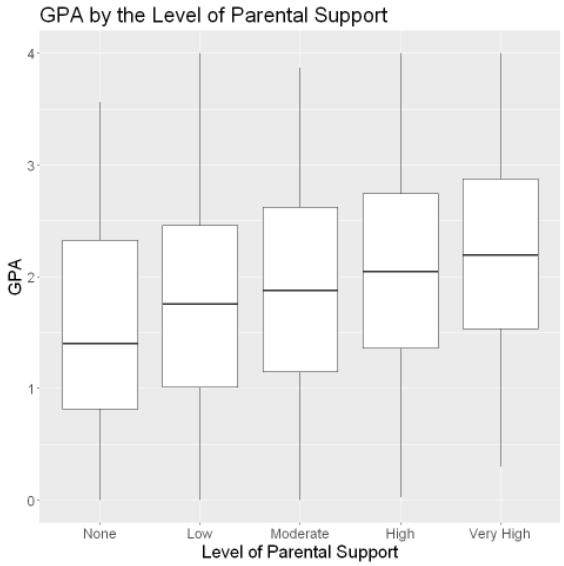

Figure 5. A boxplot of the distribution of GPA across different levels of parental support. In the dataset, the levels are recorded as 0 to 4 from left to right.

Figure 5 provides reasonable evidence that GPA depends linearly on parental support, whereas Figure 4 does not provide much evidence of GPA depending linearly on the education of the student's parents. One last bit of preliminary analysis we will perform is to assess potential collinearity via the following correlation matrix:

	GPA	StudyTimeWeekly	Absences	ParentalEducation	ParentalSupport
GPA	1.00000000	0.179275127	-0.919313576	-0.03585364	0.190773728
StudyTimeWeekly	0.17927513	1.000000000	0.009325535	-0.01105118	0.035799964
Absences	-0.91931358	0.009325535	1.000000000	0.03651750	0.002107808
ParentalEducation	-0.03585364	-0.011051182	0.036517503	1.00000000	-0.017463038
ParentalSupport	0.19077373	0.035799964	0.002107808	-0.01746304	1.000000000

Table 1. A correlation matrix of the chosen variables.

The correlation between all pairs of explanatory variables is very small, so collinearity will not be an issue for our fitted model.

For the purposes of our analysis, we will treat parental education and parental support as ordinal variables. Parental education and parental support both have five categories, which means we may potentially have to work with a lot of dummy variables in our model. Although the values of parental education and parental support are defined as categories (e.g. ParentalSupport = 1 is defined as low parental support), the values are numeric and ordered such that a larger value of ParentalSupport is necessarily a larger amount of parental support and a larger value of ParentalEducation is necessarily a higher level of parental education. As a result, it is possible to treat parental education and parental support as ordinal variables instead of

categorical variables. From Figures 4 and 5, it seems that increasing the value of parental education or parental support has a fairly consistent effect (or lack thereof) on GPA, regardless of the initial value of parental education or parental support. As a result, we have elected to treat parental education and parental support as ordinal variables in order to eliminate the need for a large number of dummy variables in our model.

As mentioned previously, we need to carefully consider our model fitting methodology to accommodate the fact that GPA has a finite range of [0.0, 4.0]. We have also previously noted that GPA seems to have a fairly consistent linear dependence on weekly study time (Figure 2) and the number of absences (Figure 3), at least when each explanatory variable is considered individually and within their respective observed ranges. As a result, a linear model with the identity link function is likely to be the best fit for the data we have, so we will use the identity link function in our model. However, the downside of using the identity link function is that GPA will not be bounded to [0.0, 4.0]. Fortunately, there are no outliers in terms of the two continuous variables, so we should not expect the identity link function to return fitted values that are significantly outside the range of [0.0, 4.0], although we will still need to inspect how our model fits the data with GPA close to or exactly 0.0 or 4.0 and adjust our model accordingly.

We will consider the "full" model as the model that includes all four explanatory variables, does not include any interactions, and does not include polynomial terms beyond degree one. The summary of the full model is as follows:

```
Call:
lm(formula = GPA ~ StudyTimeWeekly + Absences + ParentalSupport +
    ParentalEducation, data = student)

Residuals:
     Min       1Q   Median       3Q      Max
-0.99300 -0.19535 -0.00691  0.18392  1.00149

Coefficients:
                   Estimate Std. Error  t value Pr(>|t|)
(Intercept)        2.7406648  0.0198537  138.043   <2e-16 ***
StudyTimeWeekly    0.0293408  0.0009664   30.361   <2e-16 ***
Absences          -0.0995963  0.0006451 -154.380   <2e-16 ***
ParentalSupport    0.1518297  0.0048655   31.205   <2e-16 ***
ParentalEducation  0.0027932  0.0054613    0.511    0.609
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2669 on 2387 degrees of freedom
Multiple R-squared:  0.9151,    Adjusted R-squared:  0.9149
F-statistic:  6430 on 4 and 2387 DF,  p-value: < 2.2e-16
```

We see that the full model has a relatively high R^2 and adjusted R^2 value, at 0.9151 and 0.9149 respectively. But the coefficient on ParentalEducation seems to be statistically insignificant, as the p-value is fairly large. Removing said term gives a four parameter model with summary as follows:

```
Call:
lm(formula = GPA ~ StudyTimeWeekly + Absences + ParentalSupport,
    data = student)

Residuals:
     Min       1Q   Median       3Q      Max
-0.98959 -0.19507 -0.00549  0.18339  1.00223

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      2.7455092  0.0174460  157.37   <2e-16 ***
StudyTimeWeekly  0.0293355  0.0009662   30.36   <2e-16 ***
Absences        -0.0995842  0.0006446 -154.49   <2e-16 ***
ParentalSupport  0.1517870  0.0048640   31.21   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.2669 on 2388 degrees of freedom
Multiple R-squared:  0.9151,    Adjusted R-squared:  0.915
F-statistic:  8576 on 3 and 2388 DF,  p-value: < 2.2e-16
```

Which gives an identical R^2 and slightly improved adjusted R^2 value. All parameters now seem to be highly statistically significant. We can conclude that the four parameter model is better than the full model, given that it uses fewer parameters to explain essentially the same amount of variance in GPA. But perhaps a model with even fewer parameters may be better. Using regsubsets() on the full model gives the following output:

```
Subset selection object
Call: regsubsets.formula(GPA ~ StudyTimeWeekly + Absences + ParentalSupport +
    ParentalEducation, data = student)
4 Variables  (and intercept)
                 Forced in Forced out
StudyTimeWeekly      FALSE      FALSE
Absences             FALSE      FALSE
ParentalSupport      FALSE      FALSE
ParentalEducation    FALSE      FALSE
1 subsets of each size up to 4
Selection Algorithm: exhaustive
         StudyTimeWeekly Absences ParentalSupport ParentalEducation
1  ( 1 ) " "             "*"      " "             " "
2  ( 1 ) " "             "*"      "*"             " "
3  ( 1 ) "*"             "*"      "*"             " "
4  ( 1 ) "*"             "*"      "*"             "*"
```

A matrix: 4 × 5 of type lgl

	(Intercept)	StudyTimeWeekly	Absences	ParentalSupport	ParentalEducation
1	TRUE	FALSE	TRUE	FALSE	FALSE
2	TRUE	FALSE	TRUE	TRUE	FALSE
3	TRUE	TRUE	TRUE	TRUE	FALSE
4	TRUE	TRUE	TRUE	TRUE	TRUE

This confirms that indeed the best model with four parameters is the one that includes a parameter for the intercept, weekly study time, number of absences, and parental support. Comparing Mallow's C_p values gives the following graph:

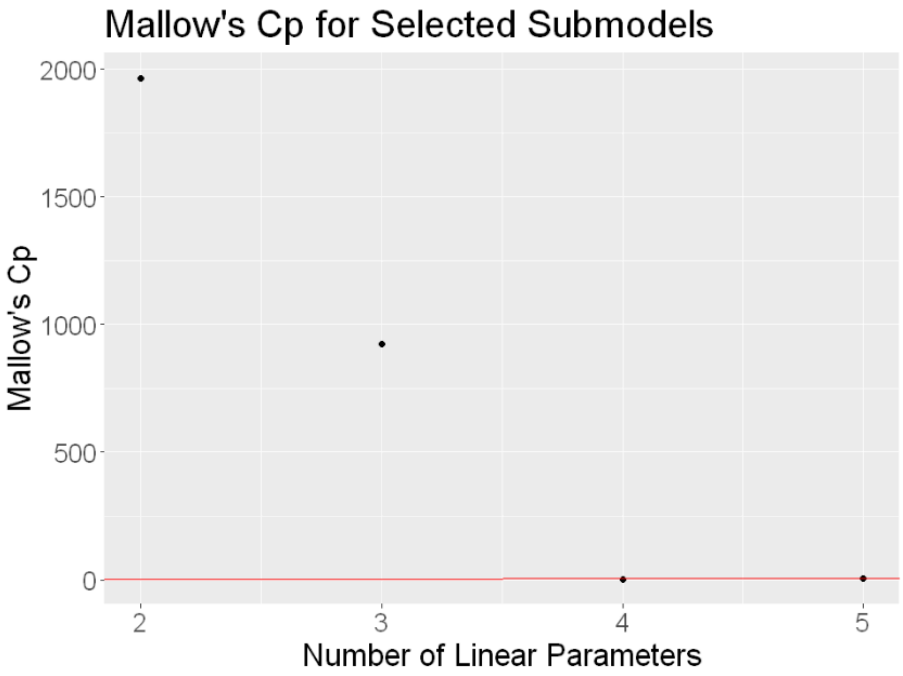

Figure 6. Mallow's Cp vs number of parameters for selected submodels

A graph with a cropped y-axis is shown below for clarity's sake:

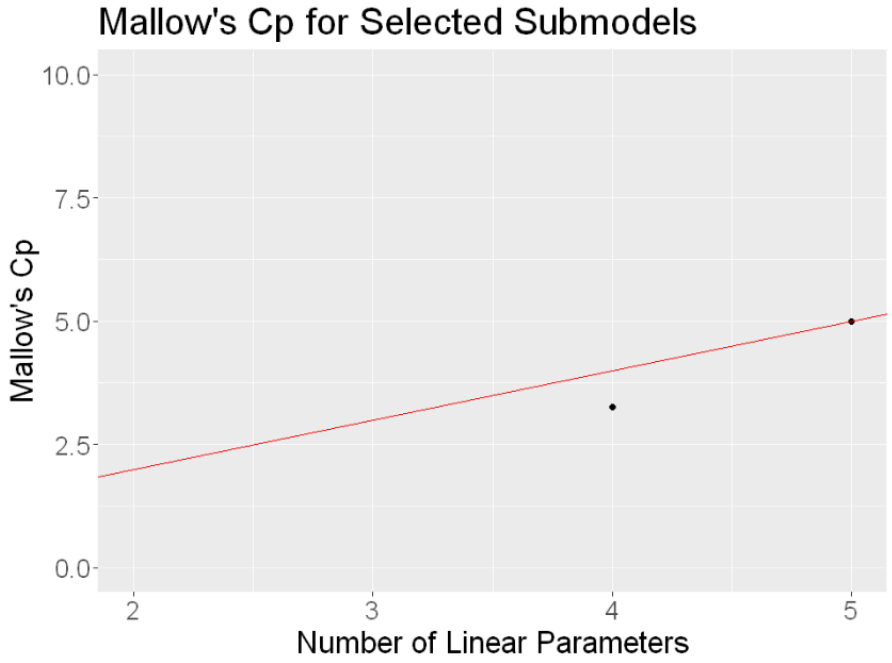

Figure 7. Mallow's C_p vs number of parameters for selected submodels, with a limited y-axis range

Outside of the full model, the model with four parameters seems to have the best Mallow's C_p value. The graph of the standardized residuals for the four parameter model is below:

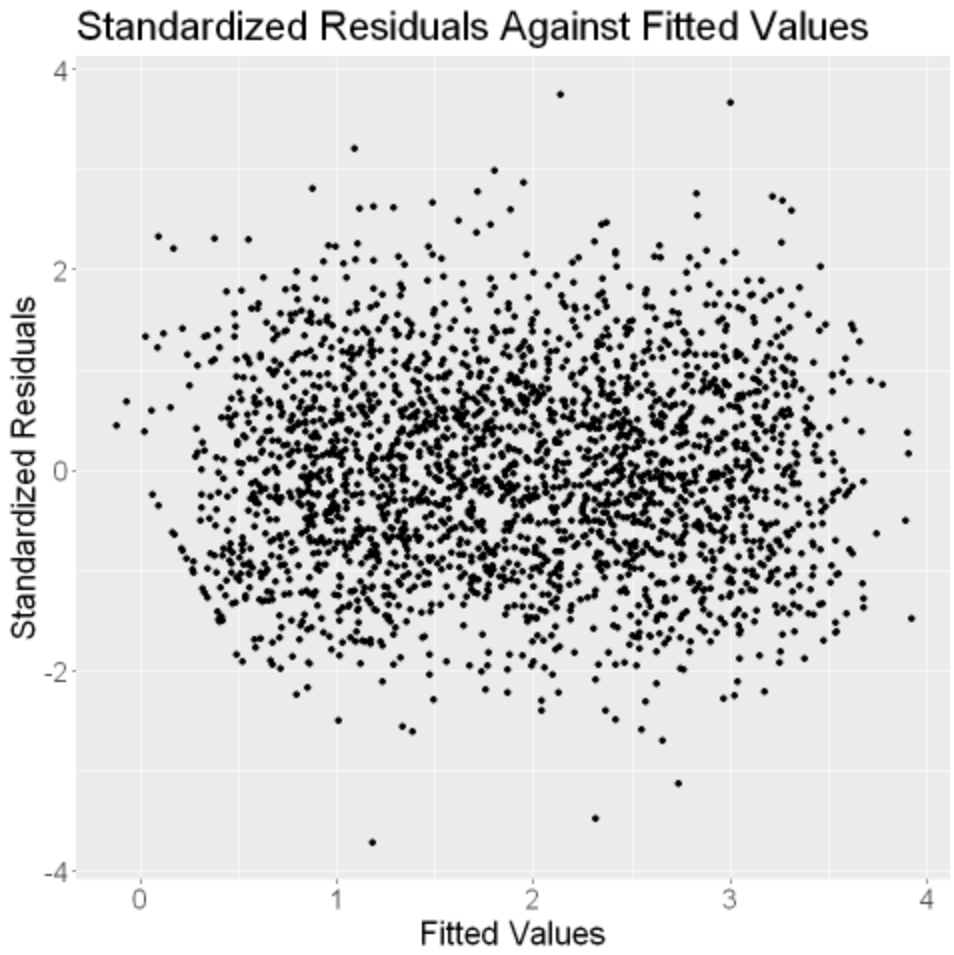

Figure 8. Residuals vs Fitted Values of the four-parameter model

A uniform, patternless cluster can be found in the 0.5 to 3.5 fitted GPA range, but this makes sense when considering the majority of the data in the dataset is also dense in said region. Variance also seems to be constant throughout the 0.5 to 3.5 fitted GPA range. The amount of variance in the residuals seems to decrease as the fitted GPA approaches 0.0 and 4.0, but this is almost certainly a consequence of it being possible for the fitted values to be less than 0 or greater than 4, whereas the actual observed values must fall within the range [0.0, 4.0].

Further Exploration

For further exploration of the dataset, considering a model using covariates not considered in this analysis would be an interesting place to start. For example, it is possible that time spent on extracurricular activities or even the type of extracurricular activities could significantly affect overall GPA. Many potentially significant factors were also included in the dataset, but for the purposes of this study, we decided to focus on a select few variables that we felt could plausibly have an effect on the GPA of a high school student. Additionally, it is

possible that interaction between certain variables can improve the model we have fitted. Interactions between our chosen explanatory variables were briefly explored but we were unable to come up with plausible reasons why GPA would be affected by interaction between our chosen explanatory variables. Therefore, we have largely chosen to ignore interactions in this study. If more explanatory variables were to be considered, the idea of including interactions in our model would certainly need to be revisited.

As we have discussed previously, we only considered linear models using the identity link function in our model selection process. This gave good results with high R^2 scores and highly significant parameters, but led to interesting issues regarding the fitted values.

```
#returning a list of fitted values <0 or >4
fourModel$fitted.values[fourModel$fitted.values < 0 | fourModel$fitted.values > 4]
```

485: -0.119924957928209 **986:** -0.0681218390718283

Our best linear model returns two fitted values that are less than 0.0, which we anticipated as a potential possibility. Obviously, this is not possible, as GPA is an aggregate of grades that must be between 0.0 and 4.0. There are a number of ways to approach this problem.

The first is to simply clamp fitted values that are outside of the acceptable range to be within the GPA range, that is to say, set any fitted values less than 0 to 0.0 and greater then 4 to 4.0. This method is attractive in its simplicity, but employing this method may potentially cause problems in determining the best model, as this method will fundamentally change how residuals are calculated. In our case, since the number of observations outside the acceptable range is so small relative to the size of the dataset (2 observations out of a dataset of 2392 observations), the negative effects of this approach are probably negligible. In practice, we would likely use the clamping method if we wanted to use our model to predict the GPA of a student at given values of weekly study time, number of absences, and parental support. This method also has the benefit of making intuitive sense in that if we already predicted a GPA of 0.0 at 30 absences, increasing the number of absences cannot further decrease GPA.

Another approach is to simply remove data points that have a GPA close to 0 or 4. The philosophy here is that models removing said “outliers” would be better and more precise in capturing trends in a narrowed GPA range of say 0.1 to 3.9. However, the criterion for an observation to be considered as an outlier is arbitrary and very subjective, and given that observations with a GPA close to or exactly equal to 0 or 4 are not exactly uncommon, we as a group were not fans of this approach.

A final approach is to model GPA as asymptotically approaching 0.0 and 4.0 with respect to the explanatory variables, either by transforming the GPA variable or using a different link function. The advantage of this approach is that it would capture potential “diminishing returns” phenomena. It is reasonable to assume that if we were to indefinitely increase weekly study time, GPA would asymptotically approach 4 in some way. Likewise, it is reasonable to assume that if we were to indefinitely increase the number of absences, GPA would asymptotically approach 0 in some way. However, Figures 2 and 3 do not provide much evidence to support this assumption

of asymptotic behavior. Both figures show GPA following fairly consistent linear trends, at least within the observed ranges of weekly study time and the number of absences. This is especially notable in Figure 3, as the mean of GPA conditioned on the number of absences gets fairly close to zero. However, we have already established in our preliminary analysis that GPA seems to linearly approach 0.0 and 4.0, at least within the observed ranges of the explanatory variables. Accordingly, when a link function similar to logit was tried, the R^2 value was significantly worse compared to our best model with the identity link function (decreased by roughly 0.2 worse), and the significance of parameters dropped significantly. The link function in question is detailed below:

$$Log(\frac{GPA}{4-GPA}) = b_0 + b_1 x_1 + \ \ldots$$

This would limit the range of fitted values to [0.0, 4.0], as required. Ultimately, using any link function at all severely impacts the interpretability of the model and any potential results, especially considering the efficiency and high-fit of the simple no-transform linear model. Even the asymptotic link function that we experimented with was only ever intended to be used as a demonstration of how a model of GPA asymptotically approaching 0.0 and 4.0 would fit the data. So, while these alternative models might merit further study, we decided to stick to the simple "clamp" approach, as it is the approach that best fits the data we have. We would likely want to reconsider asymptotic link functions if we could get data that had observations with more extreme values for absences and weekly study time.

Conclusion

If we let $\widehat{y}$ be the predicted GPA, x_1 be the weekly study time in hours, x_2 be the number of absences in a school year, and x_3 the level of parental support, then our fitted model is:

$$\widehat{y} = 2.7455092 + 0.0293355x_1 - 0.0995842x_2 + 0.1517870x_3$$

The interpretations of each coefficient are as follows:

- A student that studies for zero hours per week, has zero absences, and has no parental support will have a GPA of roughly 2.7455.
- At a fixed number of absences and a fixed level of parental support, increasing the amount of weekly study time by one hour increases GPA by roughly 0.0293.
- At a fixed amount of weekly study time and a fixed level of parental support, increasing the number of absences by one decreases GPA by roughly 0.0996.
- At a fixed amount of weekly study time and a fixed number of absences, increasing the level of parental support by one level increases GPA by roughly 0.1518.

We found that parental support, weekly study time, and number of absences, form a strong model with an adjusted R^2 of approximately 0.915. Parental support and weekly study

time, as expected, have a positive relationship with GPA, while absenteeism has a negative coefficient. On the other hand, the amount of education received by a student's parents does not seem to significantly affect the GPA of the student. Consequently, this means that the educational success/failure of the parents, which is in many cases outside of their control, is not an indicator of the child's academic success. Instead, parental support is much more valuable, and should for many parents be well within their ability to improve upon.

In conclusion, parental support, absences, and weekly study time form a strong baseline in determining the academic performance of a student, as well as to reveal areas for improvement. Addressing high absenteeism, emphasizing the importance of independent study, and promoting more parental support could be effective strategies to positively affect student academic performance. Finding practical and feasible methods of addressing these three key factors could be critical for highschool teachers and counselors to help the students who are struggling academically.

Dataset

Rabie El Kharoua. (2024). 📚 Students Performance Dataset 📚 [Data set]. Kaggle. https://doi.org/10.34740/KAGGLE/DS/5195702