Summary

Luis Pacheco & Tyler Rogers

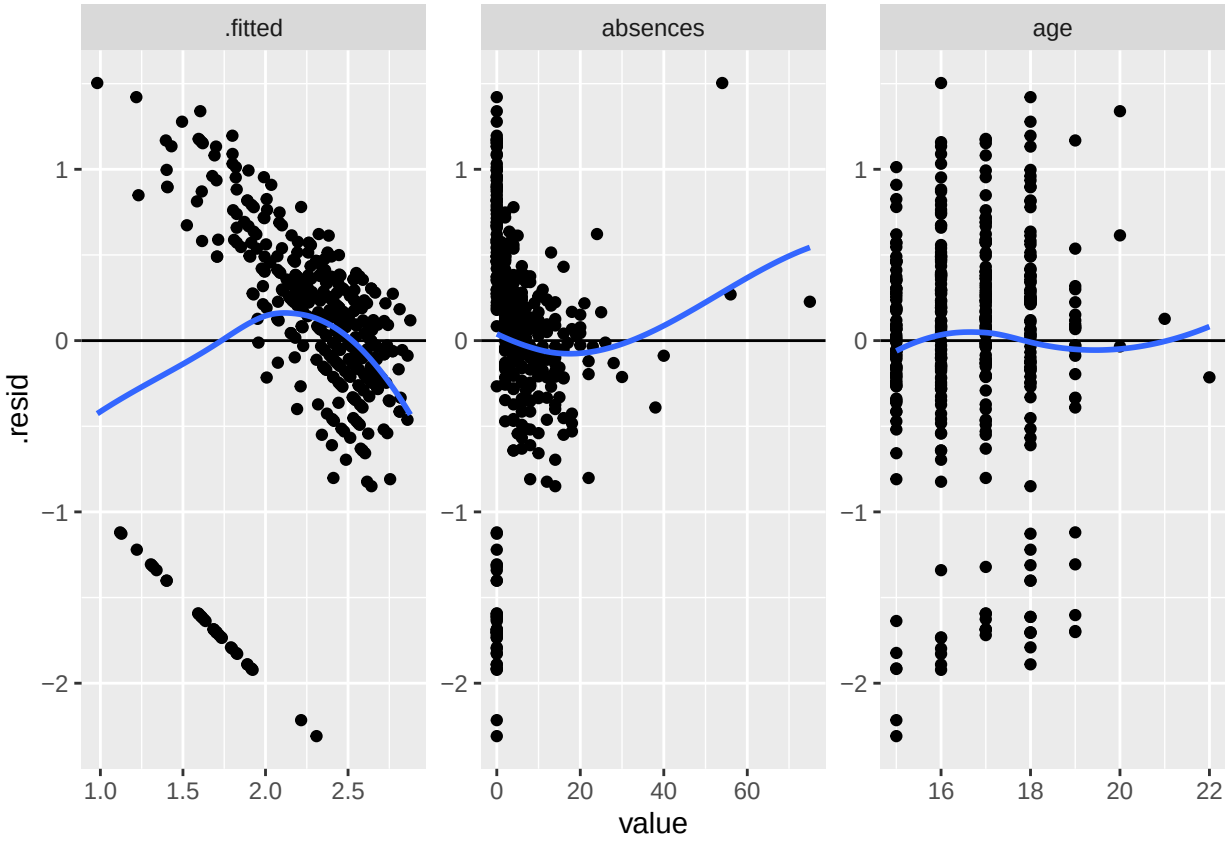
2023-12-15

Our data consisted of 395 observations from a math course. We took this data and fitted a linear model with our response variable being the student's final grade. We ran hypothesis tests to further analyze the data. From this further analysis we ran several tests to see if our model had violated our assumptions of linearity, constant variance, and normality. If there were any explanatory variables that were highly correlated. We then decided to fit a few variables with quadratic, log terms, and interaction variables. The statistical model ... was used for variable selection. We accessed the fit of our model and decided on the model:

```
model_chosen <- lm( log1p(G3) ~ sqrt(absences) + log1p(age) + romantic +
absences:romantic + absences:age, data = student_mat)
```

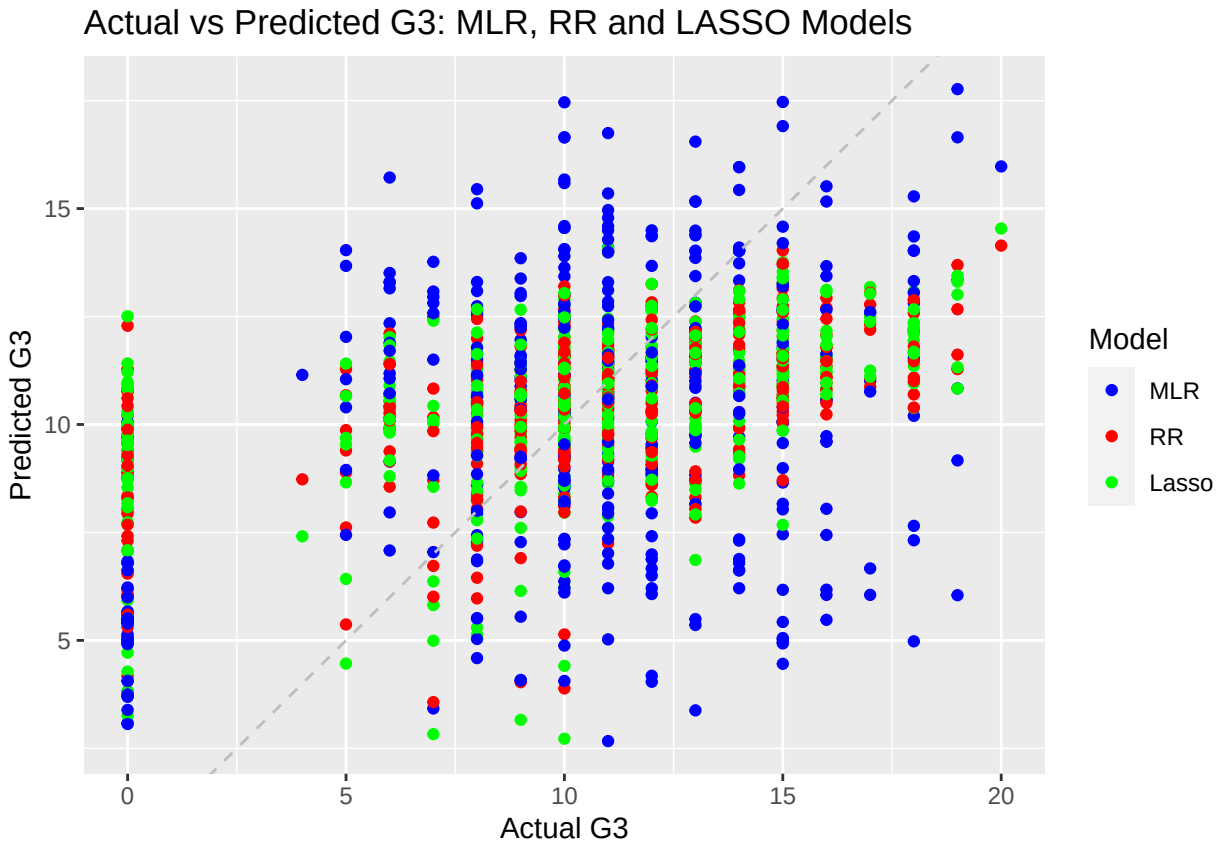
We came to a conclusion on this model by running multiple forward and backward step selection processes. As well as multiple trial and error transformation of variables.

Furthermore, one important change we made was the removal of the G1 and G2 variables from the data to make sure they wouldn't influence our models as our variable of interest is G3, the student's final grade, while G1 and G2 are grades for the students' during period 1 and 2 respectively.



Here we have a plot of the residuals vs fittee values and residuals vs absences and age, our predictor variables we saw as being highly significant. The residual vs fitted is centered around zero but shows a downward trend indicating non-normality.

Furthermore, after running a Ridge and Lasso regression model, we found that there was not enough evidence to support there being an issue with multicollinearity with the predictors used.



In this plot, we can see that there's not much difference between the MLR, RR and Lasso models. Points are spread pretty evenly around the line of perfect prediction (the gray dashed line) .

All in all, The MLR model displayed moderate predictive power, with an R-squared value indicating that a significant portion of the variance in final grades was captured by the predictors used.

Though not exceedingly high, the R-squared value was substantial enough to consider the model informative. The Ridge and Lasso regressions, which are often employed to refine predictions when multicollinearity is present or when feature selection is necessary, did not demonstrate a significant improvement over the MLR model.

This finding suggests that the predictors chosen did not suffer significantly from multicollinearity and that the additional complexity introduced by these methods may not be warranted in this context.

Upon applying the Box-Cox transformation to address potential non-normality of residuals in the MLR model, the transformation confirmed a log-like transformation (close to $= 0$) as appropriate. However, this transformation led to a slight decrease in the explained variance of the model. This result underlines the trade-off between improving the statistical assumptions of the model and maximizing the explained variance.

Across the models, several variables showed significance,Across the models, several predictors consistently appeared to be significant:

Age: Older age was associated with lower grades, which could suggest a variety of social and educational dynamics at play.

Absences: More absences were correlated with higher grades only when square root transformed, which could indicate a non-linear relationship or reflect other underlying factors.

Mother's Education (Medu): Higher maternal education levels were positively associated with student grades, highlighting the role of parental education in student academic performance.

Romantic Involvement: Being in a romantic relationship was generally associated with lower grades, which aligns with the notion that such engagements could distract from academic focus.

Going Out (goout): Frequent outings had a negative impact on grades, possibly indicating a diversion from study time.

Final Thoughts

The MLR model provided a solid baseline for understanding the relationships between predictors and student grades. The predictors identified are in line with educational theories and empirical observations, suggesting that the model has captured meaningful relationships.

While our model is not perfect—no model ever is—it has successfully identified key predictors of student academic performance. The findings from this analysis can inform interventions and policies aimed at improving student outcomes. For instance, they suggest focusing on attendance, parental engagement, and managing distractions such as romantic relationships and frequent social outings.