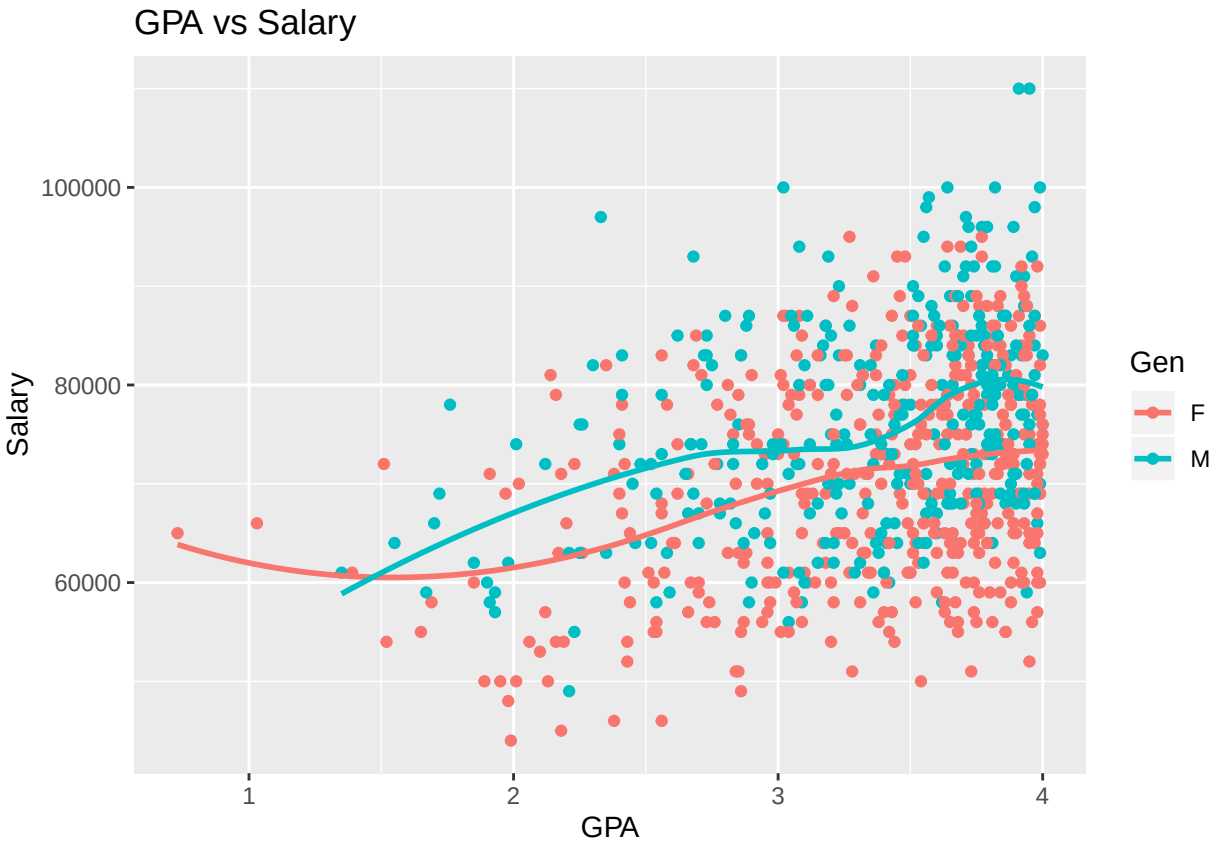
Value of College Education

Travis Andersen

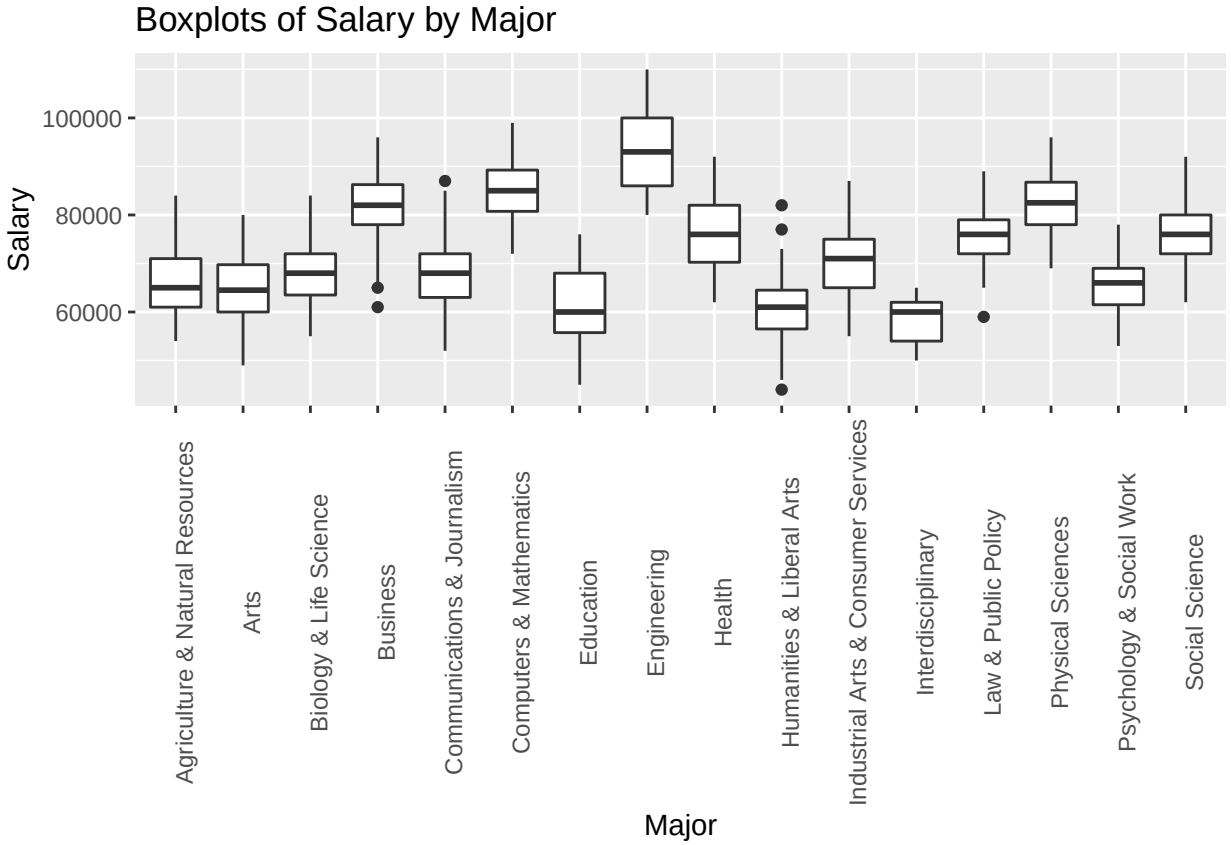
1/21/2020

1.

Following is a scatterplot of GPA and salary with gender being designated with color.



From the graph above, it seems that as GPA increases, salary increases. The correlation between GPA and salary is 0.3387. There also seems to be evidence for males being payed more than females. Below are side-by-side boxplots which compare salaries across different majors.



From these boxplots it appears that there is evidence for certain majors being payed more than others.

2.

$$\overline{y} \sim MVN(\mathbf{X}\overline{\beta}, \sigma^2\mathbf{I})$$

$\overline{\beta}$ is a vector containing a coefficient for each major except one, a coefficient for GPA, a coefficient for gender, and a coefficient for the intercept. σ^2 is the variance about the best-fit line. By fitting our model, we obtain estimates of these model parameters. To evaluate if there is a difference in majors, we test to see if all of the coefficients corresponding to major are equal to zero. To evaluate if there is a difference in gender, we test to see if $\beta_{\text{male}} = 0$.

3.

Table 1: Parameter estimates

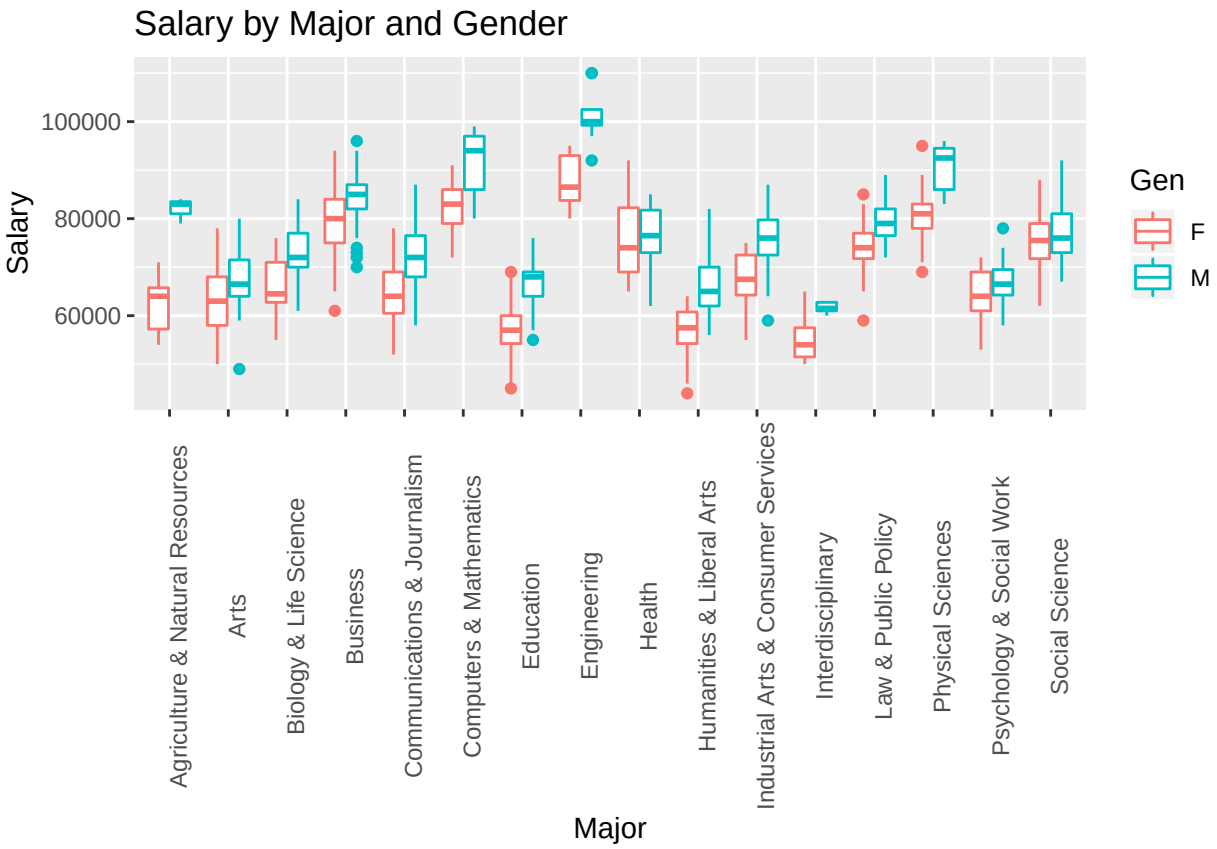
	Estimate for Beta
(Intercept)	46672.99
MajorCategoryArts	-2551.64
MajorCategoryBiology & Life Science	769.13
MajorCategoryBusiness	14282.15
MajorCategoryCommunications & Journalism	114.60
MajorCategoryComputers & Mathematics	17936.91

	Estimate for Beta
MajorCategoryEducation	-5894.85
MajorCategoryEngineering	24406.23
MajorCategoryHealth	8670.16
MajorCategoryHumanities & Liberal Arts	-5972.59
MajorCategoryIndustrial Arts & Consumer Services	2823.53
MajorCategoryInterdisciplinary	-7397.00
MajorCategoryLaw & Public Policy	7664.85
MajorCategoryPhysical Sciences	17118.28
MajorCategoryPsychology & Social Work	-1979.70
MajorCategorySocial Science	7923.38
GenM	5931.63
GPA	5488.74

If someone was a male instead of a female, their expected salary would increase by 5931.6270, holding all else constant. As GPA increases by one point, expected salary increases by 5488.7368, holding all else constant.

The estimate for residual variance is 29226668.9006777. The r-squared for the model is 0.7637.

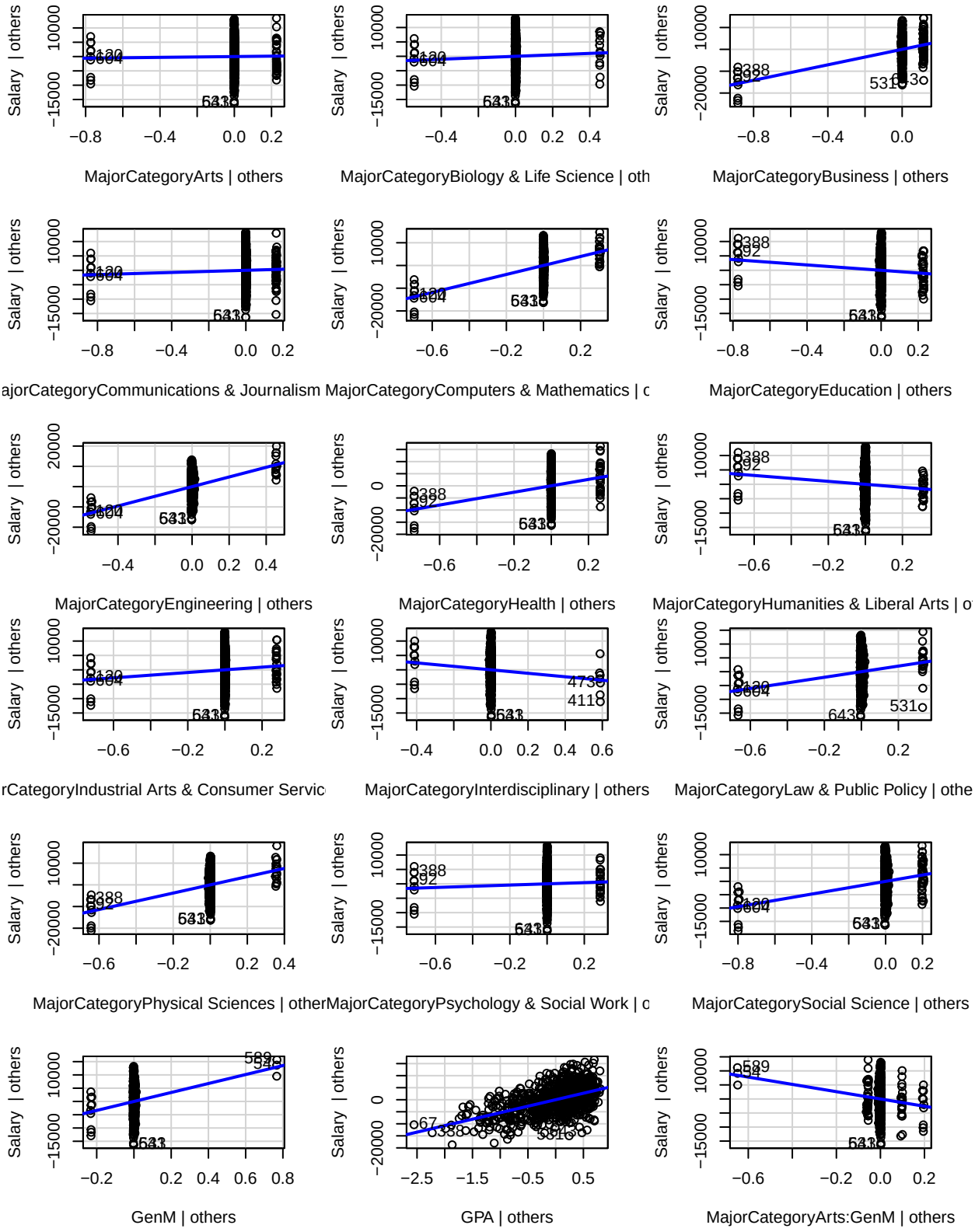
4.

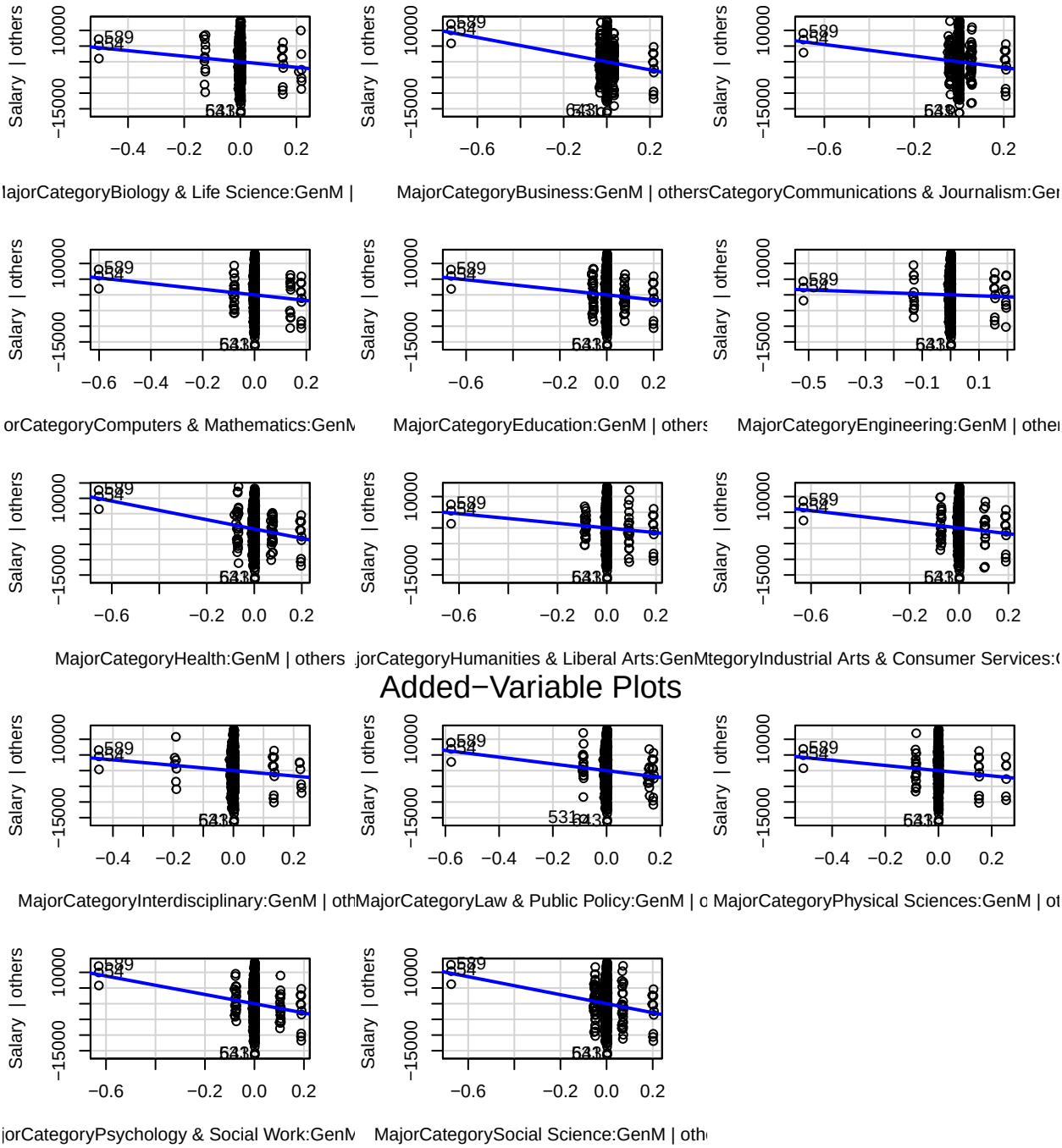


From the above boxplots, it seems that the effect gender has on salary depends on the major. For example, in the Agriculture and Natural Resources field the gap between male's and female's salary is much greater than in many other fields. To test this hypothesis, we create two models, one with an interaction term and the other without, and run an F test. Our null hypothesis is that the two models are the same, and the

alternative hypothesis is that the two models are different. The F-statistic from the test is 4.3595, and the p-value is 0.000000072. Since the p-value is so small, we reject the null hypothesis and conclude that there is a significant interaction between gender and major.

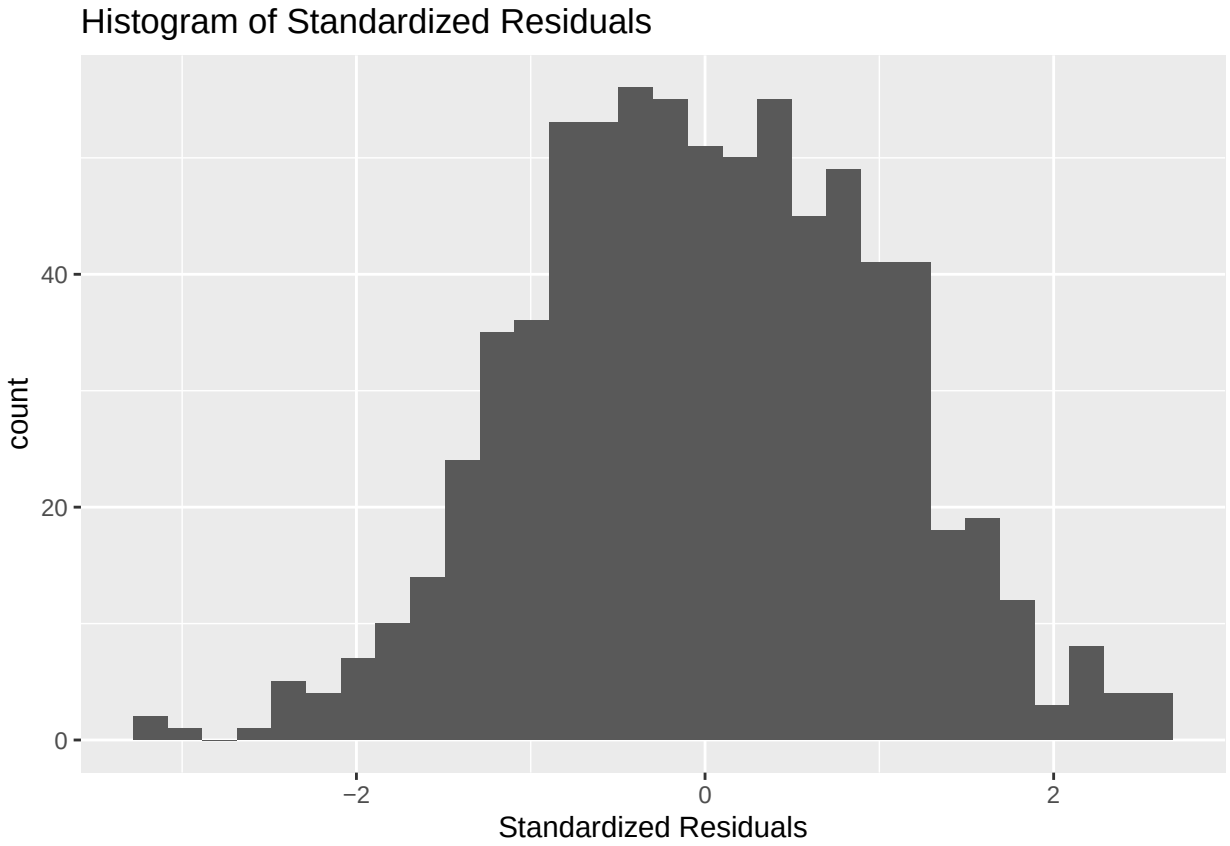
5.


Salary | others
MajorCategoryArts | others
MajorCategoryBiology & Life Science | oth
MajorCategoryBusiness | others
ajorCategoryCommunications & Journalism
MajorCategoryComputers & Mathematics | c
MajorCategoryEducation | others
MajorCategoryEngineering | others
MajorCategoryHealth | others
MajorCategoryHumanities & Liberal Arts | o
rCategoryIndustrial Arts & Consumer Servic
MajorCategoryInterdisciplinary | others
MajorCategoryLaw & Public Policy | othe
MajorCategoryPhysical Sciences | other
MajorCategoryPsychology & Social Work | c
MajorCategorySocial Science | others
GenM | others
GPA | others
MajorCategoryArts:GenM | others



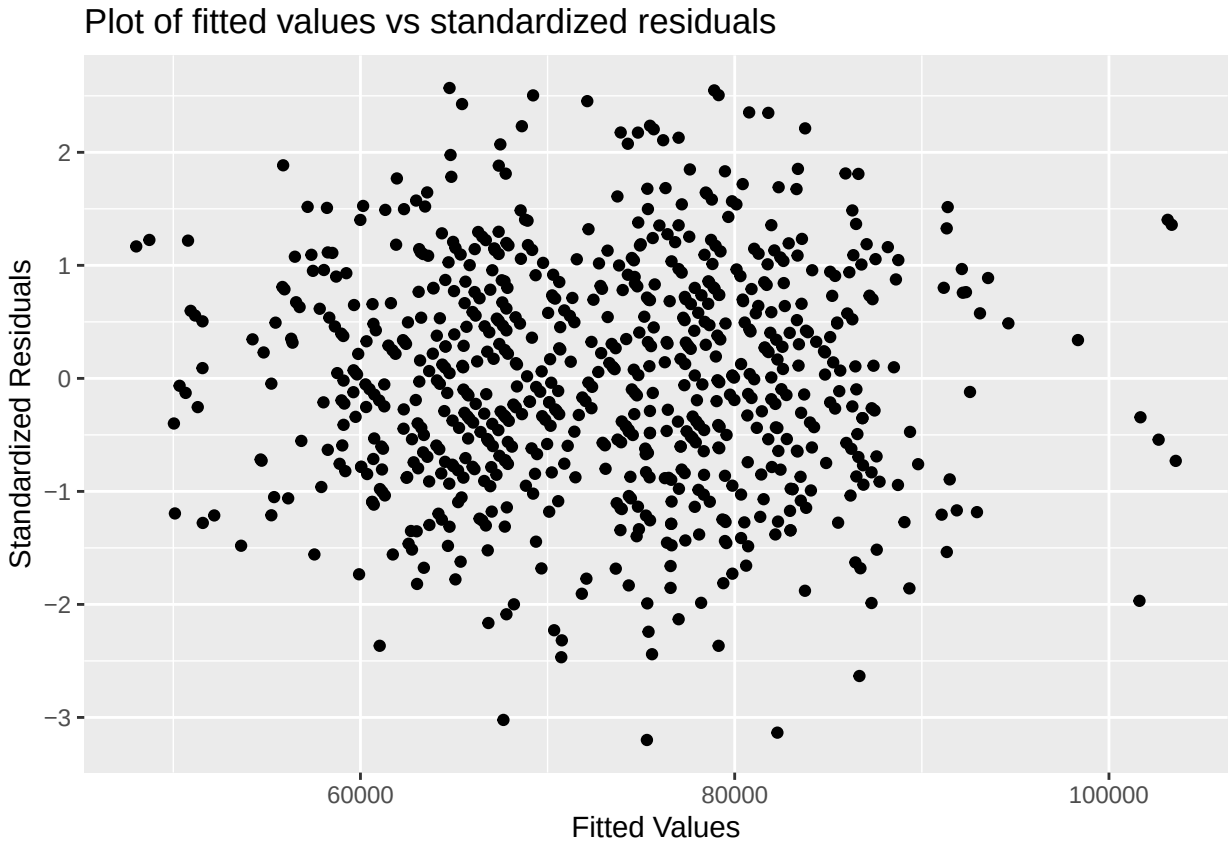
From the added variable plots we can see that GPA and Salary are linearly related, and that all other variables are categorical. Therefore, the assumption of linearity is met.

It is possible that the salary of one person would affect the salary of another, but for the type of data we have in this dataset, it is safe to assume independence.




```
## Warning in ks.test(MASS::stdres(lm), "pnorm"): ties should not be present for
## the Kolmogorov-Smirnov test
```

From the above histogram, we can see that the residuals follow a normal distribution. A ks-test returns a p-value of 0.7421. Therefore, the assumption of normality is met.



From the plot above it seems that the variance in residuals does not change with the value of fitted values, and a Breusch-Pagan test returns a p-value of 0.6075, so the assumption of equal variance is met.

6.

Table 2: 97% Confidence Interval for a few parameters

	Lower	Upper
GPA	4646.39	6129.76
GenM	9395.57	24387.63
MajorCategoryArts	-3377.77	4805.19

We are 97% confident that as GPA increases by one, holding all else constant, the expected salary will increase by between 4646.39 and 6129.76. We are 97% confident that if someone becomes male, holding all else constant, their expected salary will increase by between 9395.57 and 24387.63. We are 97% confident that if someone changed their major from Agriculture and Natural Resources to Arts, holding all else constant, their expected salary will increase by between -3377.77 and 4805.19.

7.

We want to know if women earn less than men in the Computers and Mathematics major category. Our null hypothesis is that there is no difference in salary between men and women in the Computers and

Mathematics major category. The alternative hypothesis is that women earn less than men in the Computers and Mathematics major category. The p-value is 0.0184, so we conclude that women earn less than men in the Computers and Mathematics major category. We are 95% confident that men earn between 460.95 and 14532.29 more than women in the Computers and Mathematics major category.

8.

Table 3: Prediction interval for my salary

Point	Lower	Upper
93162.19	82502.98	103821.4

My salary is predicted to be $93,162.19. We are 95% confident that my salary will be between $82,502.98 and $103,821.40.

9.

In order to evaluate how well our model predicts, we ran a leave-out-one cross validation of our data. The mean bias was 0.10, so our model neither over nor under predicted much. The coverage was 0.9497, so the prediction intervals work as they should. The mean root predictive mean square error was 4358.08, which is much less than the standard deviation of salary, which was 10,996.17. The mean prediction width was 21,013.43, which is much less than the range of salary, which was 66,000. Therfore, our model does quite well at predicting.