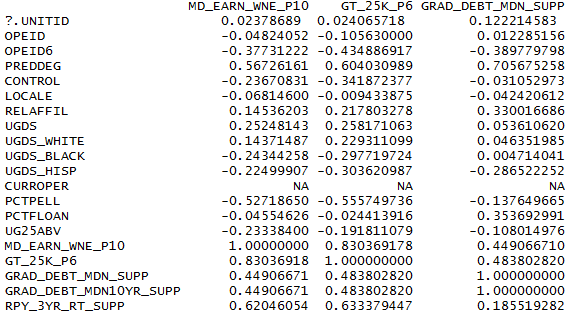
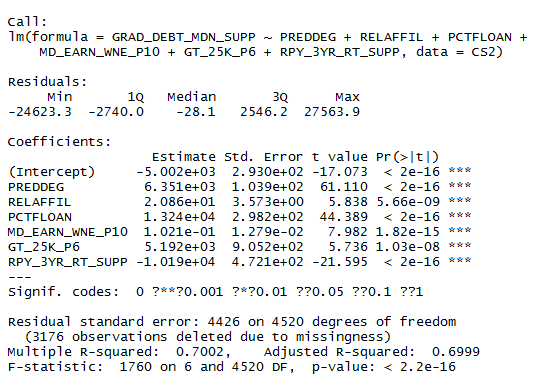
Chaoyi Huang

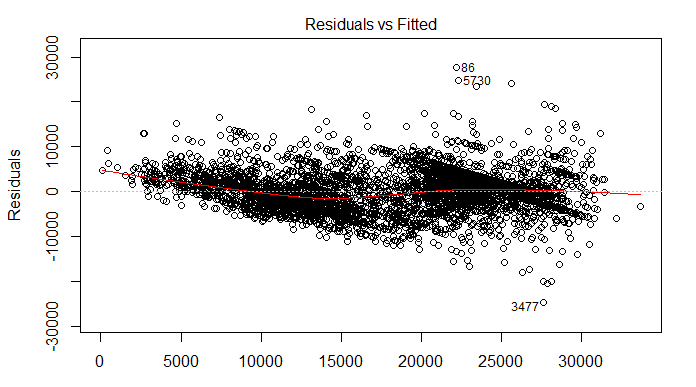
Assignment 2

* Since there are many missing values in this dataset, I replaced all the values 0 and 0.0000 with NA
* Because the percentage of missing values in certain columns are way higher than being able to analysis, so I filter the dataset by deleting those columns with over 70% percent of missing value
* I only used numeric columns in this analysis
* After correlation, I found out there are only couple variables associate with GRAD\_DEBT\_MDN\_SUPP (I make this variable as the median graduate debt) are PREDDEG, RELAFFIL, PCTFLOAN, MD\_EARN\_WNE\_P10, GT\_25K\_P6 and RPY\_3YR\_RT\_SUPP

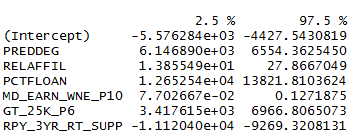


* After built the model, we can tell the model is sufficient because the P-value in each variable are lower than 0.05, as follow:





* **The regression check is as follow:**



**It turns out**

**1. MD\_EARN\_WNE\_P10 (mean earnings of students working and not enrolled 10 years after entry) would not effects the debt when the rest of variables keep the same.**

**2. RPY\_3YR\_RT\_SUPP (3-year repayment rate, suppressed for n=30) would not effects the debt. It actually means the data after debt, so I think these two variables are supposed to be removed from the variables.**

**3. I think this dataset contains too many missing values, I do not think the result I gained was sufficient unless I replace those NAs value with either the median or mean in each column in order to fill up the dataset, but in this way the dataset won’t be accurate any more.**