MODELING STUDENT ACHIEVEMENT USING LINEAR AND NONLINEAR MODELS

RYAN LAWSON (WITH DR. JEBESSA MIJENA)

Abstract. Linear regression models are widely used to make predictions based on data. In this paper, we use linear as well as nonlinear models to fit and predict standardized test results of various high schools in Georgia, using the average teacher salary, student-to-classroom ratio, and student-to-computer ratio as predictors. We develop and test predictor models using polynomial, spline, stepwise, ridge, and LASSO regression methods. We compared the performance of these models with respect to mean square error of test data for each standardized test. We find that stepwise regression works quite well for most of the subject tests, but in the case of 9th Grade Literature & Composition, the nonlinear LASSO model gives us less overall mean square error on test data set.

Submitted to the Faculty of the Department of Mathematics in partial fulfillment of the requirements for the degree of
Bachelor of Science

Georgia College & State University
2014

Date: 22 November 2014.

1. Introduction

Educational institutions typically use standardized tests to determine how much students have learned. In Georgia, certain high school courses use a standardized End-of-Course Test to fulfill this purpose.

Our goal in this study is to build a predictive model that allows us to estimate student achievement on a standardized test as determined by three factors: the school's student-classroom ratio, student-computer ratio, and the average teacher salary in the district.

2. Background

It is necessary to introduce some concepts from an area of statistics called **regression analysis**.

2.1. Introducing Regression.

Regression analysis is used to predict one or more variables of a dataset in terms of the other(s). For example, say we want to predict a student's GPA based on the hours of sleep they get each night. We have a dataset containing a **predictor variable** (hours of sleep) and a **response variable** (GPA). The general form of the model is given by

$$y = E(y) + \epsilon$$

where y is the predictor, $E(y)$ is a response equation, and ϵ is an error term (which we seek to minimize). The general form of the response equation $E(y)$ is determined by the type of regression we are using.

2.2. Types of Regression Models.

Multiple Linear Regression: Here

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_n x_n + \beta_{n+1} x_i x_j$$

where y is our response variable, $x_1, ..., x_n$ are our predictor variables, i and j are integers with $i \neq j$, and $\beta_0, \beta_1, ..., \beta_{n+1}$ are coefficients (or weights) that are determined using the Method of Least Squares. The last term $\beta_{n+1} x_i x_j$ is called an **interaction term** and is used whenever we want to account for variables that are highly correlated.

Polynomial Regression: Here

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_1 x_1^2 + ... + \beta_1 x_1^p + \beta_2 x_2 + \beta_2 x_2^2 + ... + \beta_n x_n^s$$

This model is still linear in parameter β_i's, and can be used for scatter plots that show data following a more curved path.

Spline (or Piecewise) Regression: In smoothing spline regression, we separate our data into n intervals and we generate a line of best fit for the data in each interval. The result is a piecewise-defined function that fits our data with n degrees of freedom. The points at which the data is split are called **knot values**. The equation for three intervals is given by

$$E(y) = \beta_0 + \beta_1 x_1 + \beta_2 (x_1 - k_1) x_2 + \beta_3 (x_1 - k_2) x_3$$

where k_1 and k_2 are knot values, x_1 is our independent variable, $k_1 < k_2$, and x_2 and x_3 are defined as follows:

$$x_2 = \begin{cases} 1 & \text{if } x_1 > k_1 \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad x_3 = \begin{cases} 1 & \text{if } x_1 > k_2 \\ 0 & \text{otherwise} \end{cases}$$

We can extend this equation to more intervals, or even incorporate nonlinear terms. The point of spline regression is to compensate for shifts in a regression model, or any intervals of discontinuity in the scatter plot.

3. Data Collection

We collected certain data from public high schools in Georgia. All of these datasets are available online. Refer to the References page for specific links to the datasets.

3.1. Gathering Data.

The variables used in our model are as follows:

- **EOCT Score Reports:** We obtained data on Georgia End-of-Course Tests for each subject area (English, math, science, and social science), further broken down into demographics, e.g. students who were reported being white, black, and/or economically disadvantaged. The values in the score reports represent the percentage of students who met or exceeded the state's benchmark for each test.
- **Student-Computer Ratio:** We wanted to consider the effect of technology on student success. Each year, school districts report their schools' technology inventory to the state department. This data is, again, available on the GaDOE website. Part of the inventory seeks to find the student-computer ratio. The term "modern computer" is defined by the state and must meet certain specifications (processor speed, memory capacity, etc.). In any case, the ratio obtained gives us an indication of how much technology is used in the classroom, and how much access a student has to that technology. A sample report is included in the appendix.
- **Student-Classroom Ratio:** This ratio was calculated using the technology inventory reports. The reports include a count of the classrooms and the student population. We find the ratio by dividing the population by the number of classrooms.
- **Average Teacher Salary:** Salaries of public servants are a matter of public record. There has been some discussion of "pay for performance" initiatives, whereby teachers would be paid more, or perhaps even penalized, based on how well their students perform (or how poorly). In light of these discussions, we decided to include average teacher salary as a possible factor in student achievement.

3.2. Cleaning Methods.

The first thing we had to do after gathering our data was to put it in a form that we could feasibly use (particularly with R). Essentially we had to take these distinct spreadsheets (or, in the case of the technology reports, PDF files) and merge them into one dataset. There were a few challenges that we faced in this endeavor that warrant mention, but the most prolific challenge was

filling in the massive holes in the score reports. Holes occurred when data was not reported to the state or if there were too few test-takers in a particular demographic (for instance, we wanted to include Latino students as another demographic, but were unable to do so because there were too many holes in the dataset).

We decided to clean this data using a four-step process:

(1) If a row (school) contained too many missing parts (particularly in an "all" category), we dropped the row. We started with 490 schools in our dataset, and we ended up dropping around 100 of them. We tried to limit the number of schools we dropped. One exception is the Atlanta Public School System. We dropped that entire district due to both massive holes, and allegation of cheating (or test tampering) on standardized tests from previous years.
(2) If a test score was missing for an individual school in a district with more than one high school, we replaced the missing value with the average of the other schools in the district.
(3) If a test score was missing for an individual school in a district where it is the only high school, we replaced the missing value with the state average.
(4) If we were missing a technology report from a certain school, we replaced the missing value with data from the previous year (the 2011-2012 school year).

This is a "before" snapshot of the dataset:

AVG.SALARY	SC.RATIO	STUDENT.PO	CLASSROOM	AMLIT.ALL	AMLIT.BLACI	AMLIT.HISP	AMLIT.WHIT	AMLIT.DISAD	9LIT.ALL	9LIT.BLACK	9LIT.HISP	9LIT.WHITE	9LIT.DISAD	ECON.ALL
49625.87	2.29	959	55	91	87.5	Too Few Stu	93.3	87.6	83	75	80.8	85.5	77.9	87.2
49625.87	0.51	22	4	No Data	No Data	No Data	No Data	No Data	Too Few Stu	Too Few Stu	Too Few Stu	No Data	Too Few Stu	No Data
48694.83	2.17	423	22	81.6	53.3	74.2	94	81.6	68.1	57.7	62.9	76.9	68.1	67.6
51725.58	2.42	511	29	88.7	85.7	Too Few Stu	88.4	84.1	88.4	76.9	Too Few Stu	92.5	84.2	80.1
22004.05	1.5	337	31	62.5	Too Few Stu	No Data	Too Few Stu	62.5	66.7	Too Few Stu	No Data	Too Few Stu	66.7	26.3
46100.78	1.48	1319	82	72.5	67.8	Too Few Stu	84.4	68.2	73.9	69.4	Too Few Stu	84.2	72.1	73.3
42951.8	2.16	830	47	88.7	Too Few Stu	Too Few Stu	90.1	85.4	87	Too Few Stu	Too Few Stu	88	85.1	66.9
46868.67	3.09	1646	127	91.8	88.7	75.5	95	87.1	86.3	78.1	91.1	86.3	82	85.9
46868.67	0.5	101	12	Too Few Stu	Too Few Stu	No Data	Too Few Stu	Too Few Stu	59.1	Too Few Stu	Too Few Stu	68.8	61.1	48.1
46868.67	2.97	1751	135	91.1	83.6	81.1	94.9	89.3	83.7	74.2	78.9	85.5	78.8	83.6
46785.74	2.43	952	60	84.4	Too Few Stu	Too Few Stu	85.1	80.8	85.5	Too Few Stu	Too Few Stu	86	83.6	82.1
46785.74	2.64	1407	86	84.9	78.9	86.8	87.2	80.3	80.4	75.7	72.2	83.9	73.7	73.5
46785.74	2.97	1743	97	91.6	84.4	86.4	93.2	87.5	81.9	74.5	90.3	81.7	75.6	77.9
45877.25	1.81	769	52	87.6	86	Too Few Stu	89.7	89.3	80.8	72.2	81.3	88.9	75.9	88
47522.24	No Data	817	57	86.4	79.3	Too Few Stu	88.2	81.6	79.9	65.4	66.7	83.1	71.5	68.7
47522.24	No Data	77	5	Too Few Stu	Too Few Stu	Too Few Stu	Too Few Stu	Too Few Stu	46.7	Too Few Stu	Too Few Stu	40.9	40.7	45
43115.08	1.61	1158	64	85.9	77.8	Too Few Stu	95.2	77.8	75.3	69.5	Too Few Stu	86.7	66.3	76.1
43115.08	1.4	1035	64	87.3	85.9	Too Few Stu	97	84.1	72.6	69.5	Too Few Stu	88.9	67.2	70.7
43115.08	0.48	1198	61	76.7	75	Too Few Stu	Too Few Stu	72.9	59.5	56.7	Too Few Stu	92	56.9	54.2
43115.08	1.8	1126	55	83.3	74.8	Too Few Stu	91.2	74.2	77.5	73.5	66.7	83.3	70.7	59
43115.08	1.16	736	43	72.4	72.2	No Data	No Data	72.6	61.5	61	Too Few Stu	Too Few Stu	60.8	45.2
43115.08	1.26	318	24	81.4	81.6	Too Few Stu	Too Few Stu	80.9	69.2	68.4	Too Few Stu	Too Few Stu	65.9	94.3
43115.08	1.3	969	53	69.5	69	Too Few Stu	Too Few Stu	68.8	54.3	53.6	No Data	Too Few Stu	55.7	40.8
43115.08	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
49451.97	2.01	686	38	91.5	84.4	Too Few Stu	95.1	88.6	84.6	69	Too Few Stu	88.3	77.5	79.5
49451.97	0.85	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data	No Data
45982.11	1.85	927	50	90.4	Too Few Stu	Too Few Stu	92	89.8	84.3	Too Few Stu	Too Few Stu	84.5	80.5	64.4
44785.58	1.31	570	40	85.6	83.1	Too Few Stu	92.9	83.3	84.1	78.1	73.3	95.1	78.5	73.9
44785.58	0.7	64	10	Too Few Stu	Too Few Stu	No Data	No Data	Too Few Stu	33.3	Too Few Stu	Too Few Stu	Too Few Stu	Too Few Stu	Too Few Stu
51150.93	3.15	1667	79	93	86.4	90	95.2	87.2	93.3	88.6	82.1	95.8	90.2	92.7
51150.93	1.13	486	44	87.2	76.7	Too Few Stu	90.2	83.6	86	70	Too Few Stu	90.8	80.2	60.6
50187.79	1.78	908	50	94.9	Too Few Stu	Too Few Stu	94.6	90.7	88.2	88.9	Too Few Stu	87.5	82.6	77.8
50187.79	1.54	1391	71	83.2	77.1	Too Few Stu	92.8	78.7	83.1	73.9	94.1	92.8	77.3	78.2
50187.79	0.97	No Data	No Data	86.5	81.8	No Data	89.3	82.4	80.7	62.5	Too Few Stu	91.2	73.8	76.2
48613.17	1.35	1213	74	85.2	81.8	No Data	93.9	85.2	80.5	76.4	Too Few Stu	88.8	80.5	66.4

And this is an "after" snapshot of the dataset:

AVG.SALARY	TECH.RATIO	CR.RATIO	STUDENT.PO	CLASSROOM	AMLIT	X9LIT	ECON	USH	MATH1	MATH2	PHYS	BIO	MATH	ENGLISH
49625.87	2.29	17.44	959	55	91	83	87.2	66.3	64.5	61.9	74	71.4	63.2	87
48694.83	2.17	19.23	423	22	81.6	68.1	67.6	57.8	50.5	43.3	77.3	52.9	46.9	74.9
51725.58	2.42	17.62	511	29	88.7	88.4	80.1	51.7	84.6	69.2	87.2	83.2	76.9	88.6
46100.78	1.48	16.09	1319	82	72.5	73.9	73.3	60.7	50.4	39.5	67.1	63.6	45	73.2
42951.8	2.16	17.66	830	47	88.7	87	66.9	77.1	74.2	54.7	89.7	82.3	64.5	87.9
46868.67	3.09	12.96	1646	127	91.8	86.3	85.9	68.2	69.1	58.9	75.9	79	64	89.1
46868.67	2.97	12.97	1751	135	91.1	83.7	83.6	84.5	52.5	52.9	77.1	83	52.7	87.4
46785.74	2.43	15.87	952	60	84.4	85.5	82.1	70.5	57.9	48.3	76.8	70.3	53.1	85
46785.74	2.64	16.36	1407	86	84.9	80.4	73.5	63.7	47.7	56.6	67.5	71.9	52.2	82.7
46785.74	2.97	17.97	1743	97	91.6	81.9	77.9	81.3	51	56.5	78	78.1	53.8	86.8
45877.25	1.81	14.79	769	52	87.6	80.8	88	65	67.5	61.8	83.4	81.5	64.7	84.2
47522.24	10.21	14.33	817	57	86.4	79.9	68.7	55.5	65.9	53	67.8	68.8	59.5	83.2
43115.08	1.61	18.09	1158	64	85.9	75.3	76.1	62.6	45.1	46.9	70.5	63.3	46	80.6
43115.08	1.4	16.17	1035	64	87.3	72.6	70.7	46.3	41.1	35.3	70.5	54.5	38.2	80
43115.08	0.48	19.64	1198	61	76.7	59.5	54.2	40.8	35	26.8	70.5	42.1	30.9	68.1
43115.08	1.8	20.47	1126	55	83.3	77.5	59	58.5	43.8	38.1	70.5	59.5	41	80.4
43115.08	1.16	17.12	736	43	72.4	61.5	45.2	32.9	32.4	41.3	70.5	35	36.9	67
43115.08	1.26	13.25	318	24	81.4	69.2	94.3	51.7	36.8	41.2	70.5	50	39	75.3
43115.08	1.3	18.28	969	53	69.5	54.3	40.8	22.2	23.4	12.4	70.5	39.4	17.9	61.9
49451.97	2.01	18.05	686	38	91.5	84.6	79.5	54.5	84.9	74.5	87.5	77.1	79.7	88.1
45982.11	1.85	18.54	927	50	90.4	84.3	64.4	67.7	62.6	34.1	89.8	81.2	48.4	87.4
44785.58	1.31	14.25	570	40	85.6	84.1	73.9	60.3	71.3	52.3	86.6	69.6	61.8	84.9
51150.93	3.15	21.1	1667	79	93	93.3	92.7	87.6	79.1	82.2	85	85.8	80.7	93.2
51150.93	1.13	11.05	486	44	87.2	86	60.6	61.2	62.9	44.5	81	71.7	53.7	86.6
50187.79	1.78	18.16	908	50	94.9	88.2	77.8	73.4	70.2	70	79.7	81.5	70.1	91.6
50187.79	1.54	19.59	1391	71	83.2	83.1	78.2	57	55.5	57.4	78.7	66.4	56.5	83.2
50187.79	1.8	13.33	480	36	86.5	80.7	76.2	58	73.1	63.1	82.5	74	68.1	83.6
48613.17	1.35	16.39	1213	74	85.2	80.5	66.4	60.7	58.2	65.6	76.3	68.2	61.9	82.9
41024.87	2.82	16.58	978	59	86.8	84.7	66.2	70.9	60.6	47.4	70.5	74.4	54	85.8
40905.94	1.88	19.13	306	16	91.9	91.9	35.6	52.3	65.8	44.4	89.5	60.5	55.1	91.9
50921.93	2.35	17.01	2807	165	93.5	88.6	83.5	78.4	82.4	74.4	83	80.8	78.4	91.1
36939.67	1.03	13.13	525	40	90.2	71.7	74.1	79.1	56.1	55	65.5	69.7	55.6	81
46811.69	1.31	15.8	632	40	87.8	75.7	59	56.5	61	76.2	77.5	65.7	68.6	81.8
46811.69	0.75	16.22	373	23	82.5	81.5	70.7	65.7	75.9	74	80	81.5	75	82

3.3. **Final Draft.** The final draft of the dataset contained the following variables:

- **Predictors**
  - `AVG.SALARY`: Average Teacher Salary
  - `TECH.RATIO`: Student-Computer Ratio
  - `CR.RATIO`: Student-Classroom Ratio
- **Responses**
  - `X9LIT`: Scores for 9th Grade Literature and Composition
  - `AMLIT`: Scores for American Literature and Composition
  - `MATH1`: Scores for Mathematics I
  - `MATH2`: Scores for Mathematics II
  - `BIO`: Scores for Biology
  - `PHYS`: Scores for Physical Science
  - `USH`: Scores for United States History
  - `ECON`: Scores for Economics

4. Exploratory Analysis

Visual representations of the data are provided in the appendix. The table below summarizes our numerical measures.

Predictor/Response	**Minimum**	**Median**	**Mean**	**Maximum**	**SD (σ)**
AVG.SALARY	22,504	47,518	47,733	57,278	4169.858
TECH.RATIO	0.000	2.080	2.223	10.210	1.099391
CR.RATIO	6.22	16.79	16.49	39.76	3.45684
X9LIT	42.90	84.65	83.95	100.00	9.685169
AMLIT	47.20	89.55	88.46	100.00	7.618085
MATH1	7.40	66.70	65.36	100.00	17.50276
MATH2	10.50	58.50	59.18	100.00	19.6544
BIO	18.90	74.00	72.58	100.00	15.12471
PHYS	34.40	77.75	77.82	100.00	14.50959
USH	13.00	68.30	67.13	100.00	17.1142
ECON	26.90	77.70	75.47	100.00	14.89702

5. Model Testing

We tested each of our regression models using **cross validation**, which is done by randomly splitting our dataset into two groups: a training section and a testing section. Our training section (which contained 248 observations) was used to fit the model. The testing section (containing 122 observations) was used to test the models. We used the Root Mean Squared Error to determine how effective our models were at predicting responses. Specifically,

$$\text{RMSE} = \sqrt{\frac{\sum(y-\hat{y})^2}{122}}$$

where y is our actual response, $\hat{y}$ is the predicted response.

6. Regression Models

During our analysis we considered four types of models, explained below:

Note: We define the following denotations for the purposes of simplicity.

- A = AVG.SALARY
- T = TECH.RATIO
- C = CR.RATIO

6.1. **Stepwise Regression:** The stepwise regression process involves taking a large general model and dropping variables that do not significantly contribute to the model predictions. One variable is dropped at a time, and the remaining variables are tested using ANOVA to determine the model's overall effectiveness. Stepwise regression seeks to minimize the objective function

$$\sum_{i=1} \left(y_i - \beta_0 - \sum_{j=1} x_{ij}\beta_j \right)^2$$

We applied stepwise regression to our largest polynomial model

$$A^3 + A^2 + A + T^3 + T^2 + T + C^3 + C^2 + C + AT + AC + CT + ACT$$

and arrived at the more efficient models, given below.

$$\text{X9LIT: } 80 - .0003A + 6.14T - 1.81C + .00006AC - .28CT$$

$$\text{AMLIT: } 106 - .0005A + .78T - 2.36C + .00005AC$$

$$\text{MATH1: } 3.35 + .00094A + 1.74T + .87C$$

$$\text{MATH2: } 157 - .003A + 13T - 11C + .00027AC - .76CT$$

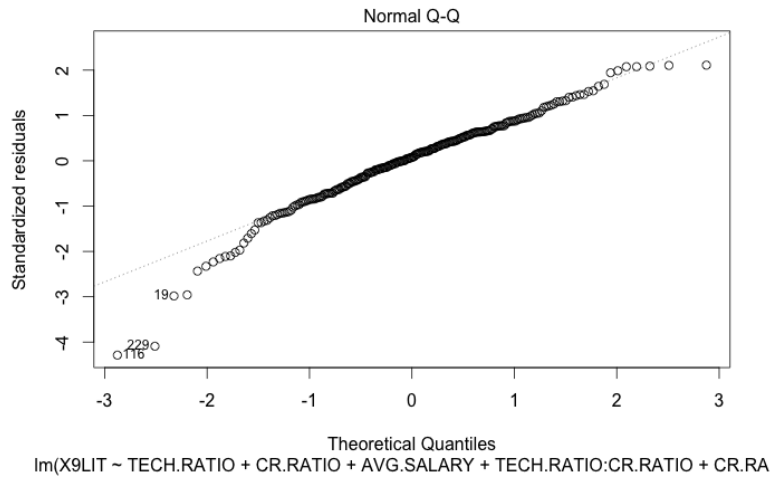
$$\text{BIO: } 105 - .001A + 11T - 5.45C + .0002AC - .58CT$$

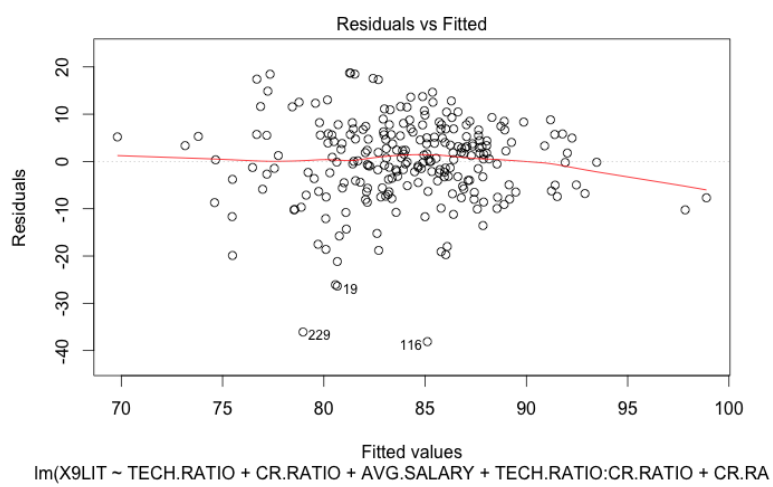
$$\text{PHYS: } 42 + .00076A$$

$$\text{USH: } 78 - .0008A + 8.69T - 4.4C + .0001AC - .42CT$$

$$\text{ECON: } 50 - .00006A + 46T - 4C + .0001AC - .0007AT - .53CT$$

We make two assumptions about our models: (1) that the error associated with each model is normally distributed with a mean $\mu = 0$, and (2) that the variance of the error remains constant. We can check our assumptions by examining the two plots on the following page.





The first plot allows us to verify our normality assumption. In particular, the more the individual plots adhere to the dotted line, the more normally the error is distributed. The second plot provides verification of our assumed constant variance. To put it simply, the straighter the red line, the more constant the variance. It should be clear that, *en masse*, our assumptions are satisfied. There are outliers which will affect our normality assumption, particularly at the lower end of the plot, but overall the assumptions hold.

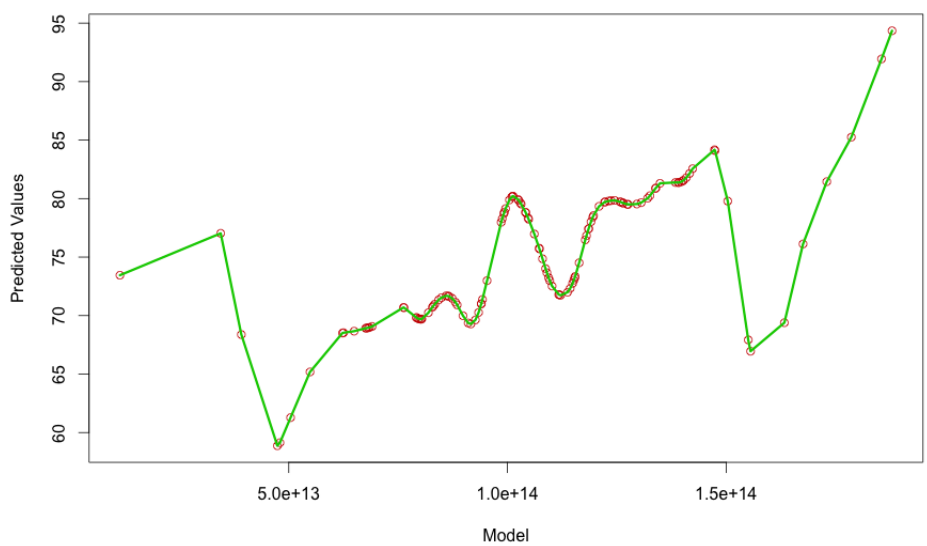
6.2. **Smoothing Splines:** Spline regression works well if you know where the knots should occur in the dataset. However, **smoothing splines** make this a nonissue by choosing a maximal set of knots. More specifically, smoothing spline regression seeks to find a function $f(x)$ that minimizes

$$\sum_{i=1}^{n}[y_i - f(x_i)]^2 + \lambda \int [f''(t)]^2 \, dt$$

where $\lambda \geq 0$ is a fixed *smoothing parameter* and $f(x)$ is twice differentiable. We allow R to generate our function numerically using ten-fold cross-validation. For smoothing splines, we used our complete model again:

$$A^3 + A^2 + A + T^3 + T^2 + T + C^3 + C^2 + C + AT + AC + CT + ACT$$

We can see our functions by plotting the smooth spline model. The model for Economics is given on the following page. Further models are provided in the appendix.



One consequence of smoothing splines is **overfitting**. Our model may work well for the training data, but then may perform poorly on the test data. If overfitting occurs, this means our model is not generalizable, and therefore not useful outside of the context of our dataset.

6.3. **Ridge Regression: Multicollinearity** is a condition wherein there is a correlation between two or more independent variables. This translates to instability and bias in our β coefficients. To compensate, we can use ridge regression. Ridge regression introduces a penalty on the size of our β coefficients. Specifically, our predictor $\hat{\beta}$ is estimated by finding the value of β that minimizes

$$\sum_{i=1} \left(y_i - \beta_0 - \sum_{j=1} x_{ij}\beta_j \right)^2 + \lambda \sum_{j=1} \beta_j^2$$

where λ is our penalty factor. There are analytical methods for finding λ, but R is equipped to find an optimal choice for λ numerically (again, using cross-validation).

Our models using ridge regression are given by

$$\text{X9LIT: } 55 + .8T + .12C + .0004A - .006CT + .00002AT + 000006AC$$

$$\text{AMLIT: } 71 + .2T + .04C + .0003A + .008CT + .000008AT + .000004AC$$

$$\text{MATH1: } 19 + .9T + .33C + .0006A - .017CT + .00002AT + .00001AC$$

$$\text{MATH2: } -3 + .6T + .006C + .001A - .1CT + .00004AT + .00002AC$$

$$\text{BIO: } 24 + 1.3T + .1C + .0007A - .05CT + .00003AT + .00001AC$$

$$\text{PHYS: } 59 + .08T - .0004C + .0003A + .000007CT + .000007AT + .000003AC$$

$$\text{USH: } 13 + T + .2C + .0008CA - .03CT + .00003AT + .00001AC$$

$$\text{ECON: } 48 + .8T + .2C + .0004A + .001CT + .00001AT + .000006AC$$

6.4. **LASSO Regression:** One final regression method, called LASSO (Least Absolute Shrinkage and Selection Operator) Regression, involves penalizing the absolute values of the β coefficients. Specifically, LASSO Regression minimizes

$$\sum_{i=1}\left(y_i - \beta_0 - \sum_{j=1} x_{ij}\beta_j\right)^2 + \lambda \sum_{j=1} |\beta_j|$$

It is similar to ridge regression, with one additional feature: variable selection. In other words, LASSO regression does allow certain variables to be penalized completely and set to zero (0). The LASSO models are given below.

$$\text{X9LIT: } 48 + .0006A + 1.5T + .38C$$

$$\text{AMLIT: } 66 + .0004A + .69T + .25C$$

$$\text{MATH1: } 58 + .0009A + 1.6T + .84C$$

$$\text{MATH2: } -20 + .001A + .71T + .73C$$

$$\text{BIO: } 12 + .001A + 2.1T + .58C$$

$$\text{PHYS: } 13 + .0007A + .25T + .026C$$

$$\text{USH: } -1.2 + .001A + 1.9T + .77C$$

$$\text{ECON: } 37 + .0006A + 1.6T + .46C$$

7. Summary

The following table lists the Root Mean Squared Error associated with each regression model. The lowest RMSE for each test is in **boldface**.

EXAM/SUBJECT	**LASSO**	**Ridge**	**Spline**	**Stepwise**
9th Grade Lit/Comp	**8.75900175**	8.76300823	9.76229093	8.9350864
American Literature	8.43414292	8.48333596	9.12699524	**6.39328181**
Mathematics I	17.1487619	17.2158415	20.0041873	**15.7984536**
Mathematics II	19.494727	19.4980139	22.3881131	**16.7411321**
Biology	14.004759	13.9821969	16.2234089	**13.5401353**
Physical Science	15.2079303	15.2711218	17.1844594	**13.5420342**
US History	17.0895619	17.0999318	19.0581139	**14.8808439**
Economics	15.0960979	15.1768749	15.8498572	**13.6579309**

8. Conclusions

First we want to highlight some of the models.

- The stepwise model for Physical Science

$$\text{PHYS: } 42 + .00076A$$

  seems to indicate that the only contributing factor to higher student achievement in Physical Science is the average teacher salary. Surely science teachers find this particularly inspiring. Also, the stepwise model generates the least amount of error.
- The LASSO model for Ninth Grade Literature & Composition

$$\text{X9LIT: } 48 + .0006A + 1.5T + .38C$$

  generates less error than the stepwise model. A possible reason could be multicollinearity in the dataset. Also, the difference between the ridge regression error and the LASSO error is approximately 0.004 points. Again, multicollinearity could be the reason.
- The difference between the Biology stepwise model

$$\text{BIO: } 105 - .001A + 11T - 5.45C + .0002AC - .58CT$$

  and the ridge model

$$\text{BIO: } 24 + 1.3T + .1C + .0007A - .05CT + .00003AT + .00001AC$$

  is approximately half a point.

Smoothing spline regression performed relatively poorly throughout each of our models. A possible reason could be overfitting.

Intuitively, we would expect an increase in class sizes or technology ratios to lead to a decrease in test scores (i.e. a classroom with 20 students should, in theory, perform better than a classroom with 60 students). This is overwhelmingly untrue for most of our models. However, some degree of common sense should be utilized.

Lastly, it's important to note that some error is to be expected as a result of our data cleaning methods. Every time we dropped a row or made a substitution, we were increasing our error.

9. Further Research

One idea for further research include using other regression models. In regards to this paper, a good idea for further research would be to attain a more complete dataset. Cleaning our dataset increased our error, so having data without so many holes would be beneficial. It would also be interesting to introduce more variables. We collected scores for certain demographics (white students, black students, Latino/a students, and those deemed economically disadvantaged). We decided not to include this analysis due to time constraints and overwhelming incompleteness in the data that was reported from the Georgia Department of Education, but again, it would be interesting to see if and how race affects test scores. Lastly, and pertaining to our discussion of classroom sizes in the Results section, it would be interesting to attempt to find an optimal classroom size.

10. References

(1) Mendenhall, W. and Sinich, T. *A Second Course in Statistics: Regression Analysis* 7th Edition © 2012 by Pearson
(2) Hastie et al, *Elements of Statistical Learning* 2nd Edition by Springer

11. Data Sources

(1) GaDOE Technology Inventory Reports
`http://www.gadoe.org/Technology-Services/Instructional-Technology/Pages/Dashboard.aspx`
(2) Teacher Salary Reports
`http://open.georgia.gov/sta`
(3) EOCT Score Reports
`http://www.gadoe.org/CCRPI`

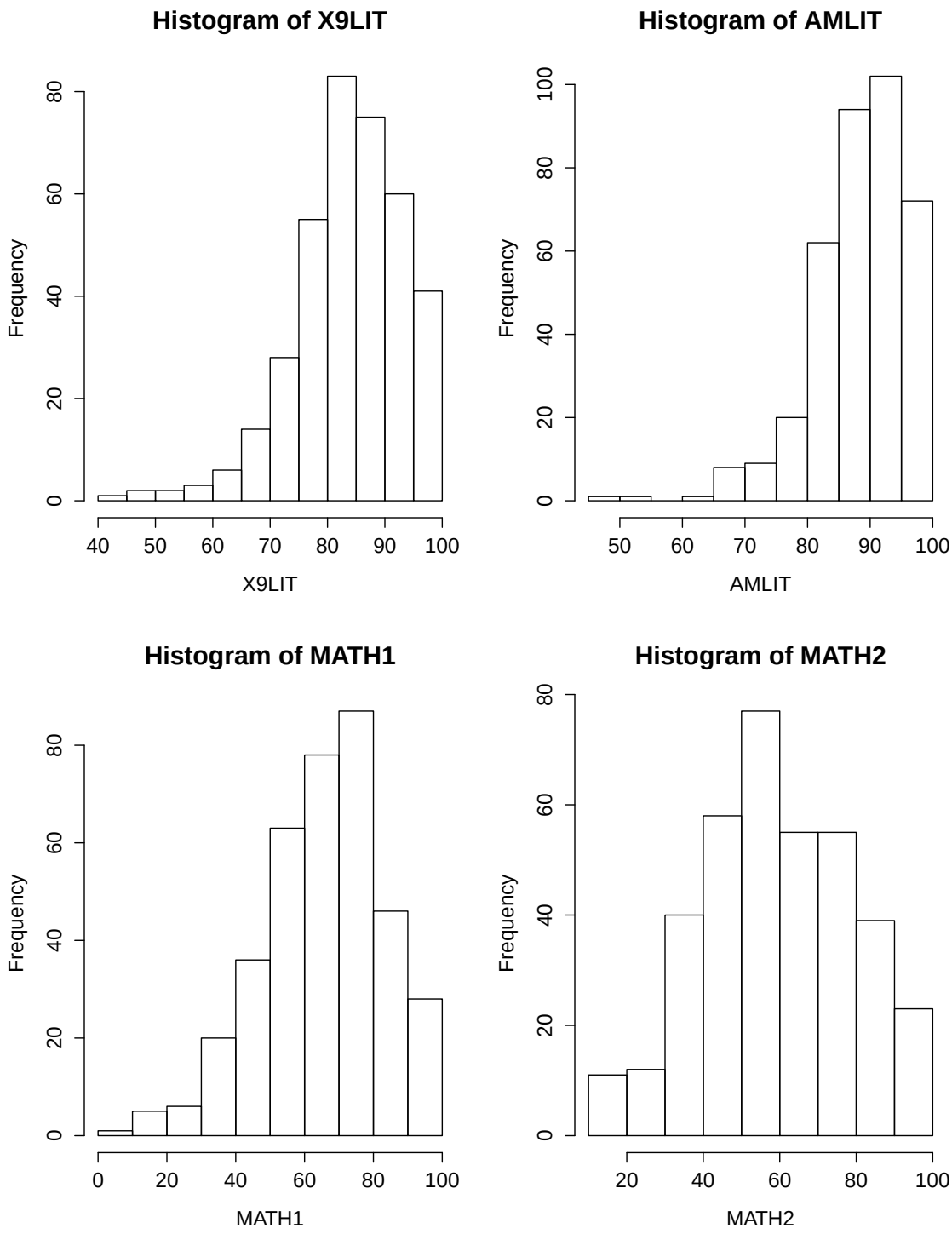
12. Acknowledgements

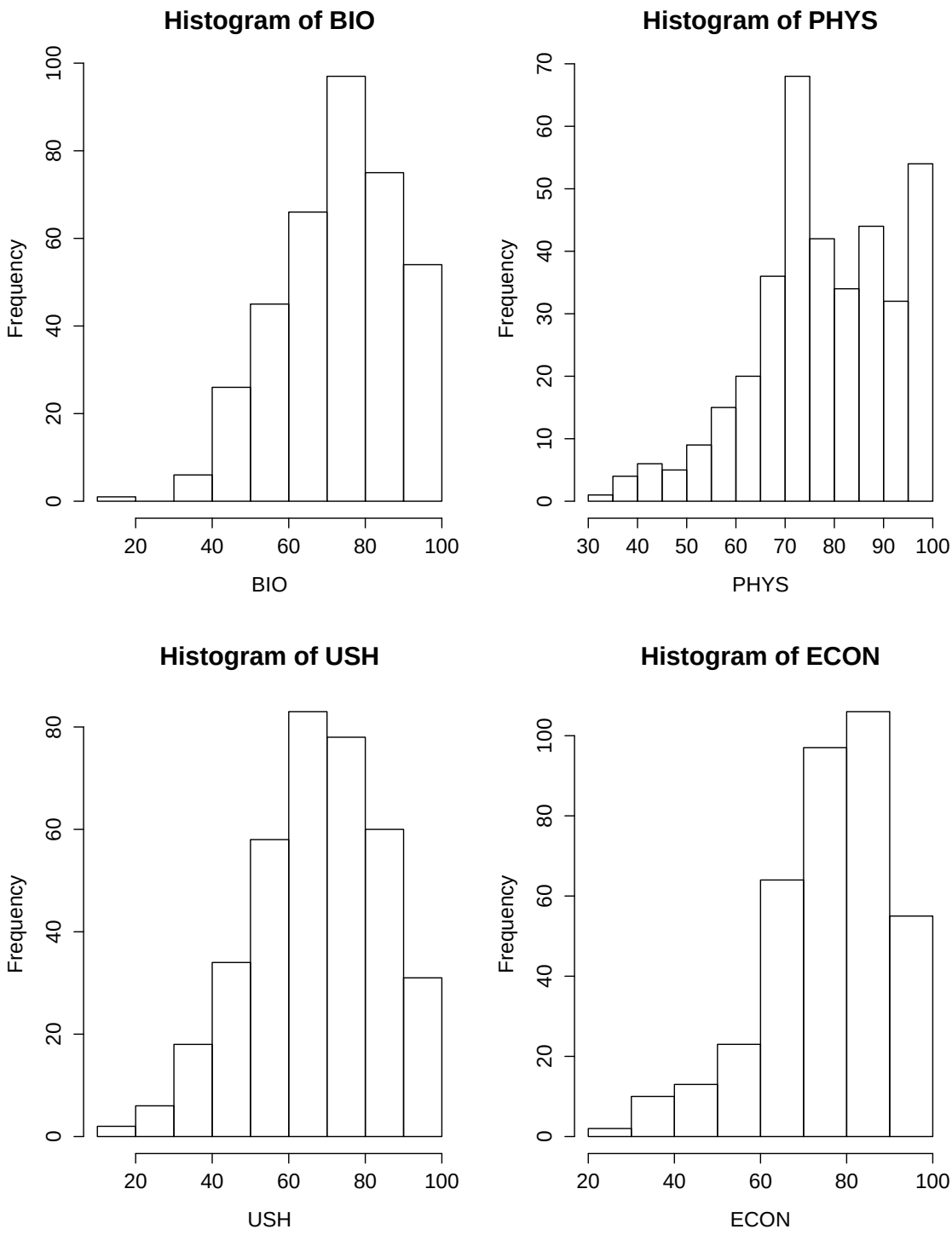
I have many people to thank for their assistance and support with this research endeavor. First I want to thank my advisor, Dr. Jebessa Mijena, without whose patience and guiding hand this project could never have been completed. I was also blessed with many other sources of support, and while the list is by no means exhaustive, I do wish to thank the following persons whose help was indispensable: Mrs. Jeanne Haslam, Dr. Kelli Brown, Dr. Bradley Koch, Mr. Hance Patrick, Dr. James Carlisle, Mr. Mike Augustine, Dr. Robert Blumenthal, Mr. Marc Cardinalli, and Dr. Michelle Masters.

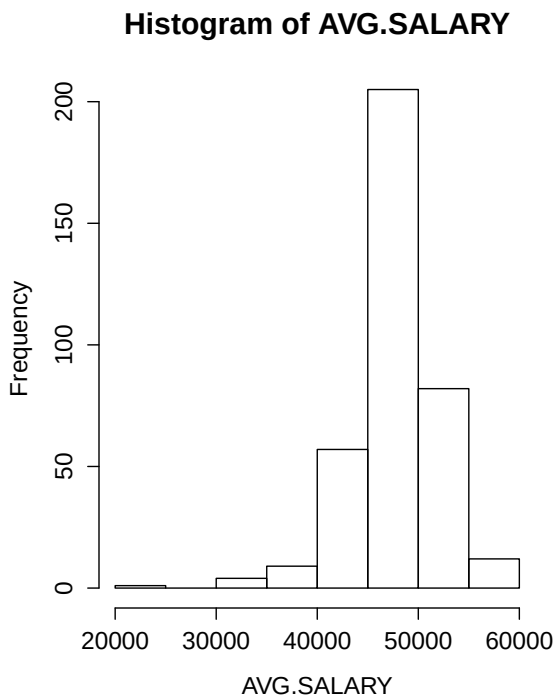
13. Appendix

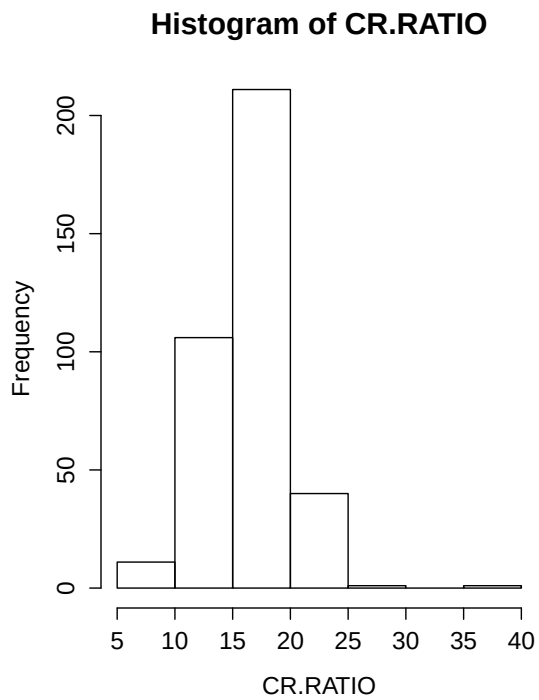
Plots and R Code are attached. The compiled dataset is available upon request (`austin.lawson@bobcats.gcsu.edu`).

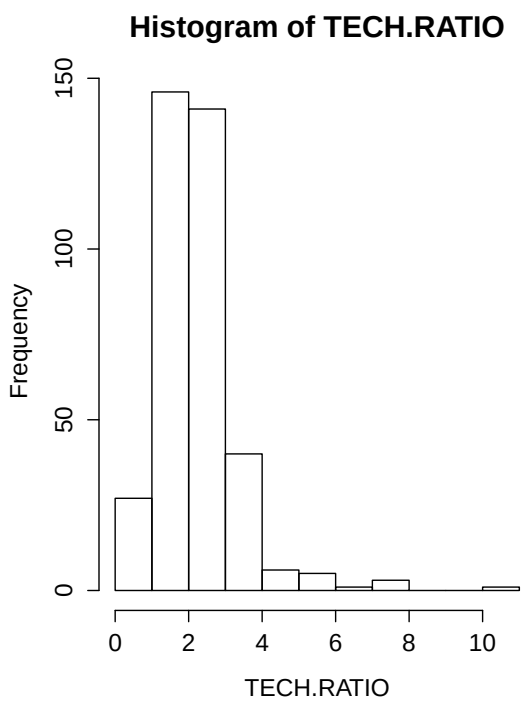
Graphical Analysis


Histogram of X9LIT
Frequency
0
20
40
60
80
40
50
60
70
80
90
100
X9LIT
Histogram of AMLIT
Frequency
0
20
40
60
80
100
50
60
70
80
90
100
AMLIT
Histogram of MATH1
Frequency
0
20
40
60
80
0
20
40
60
80
100
MATH1
Histogram of MATH2
Frequency
0
20
40
60
80
20
40
60
80
100
MATH2

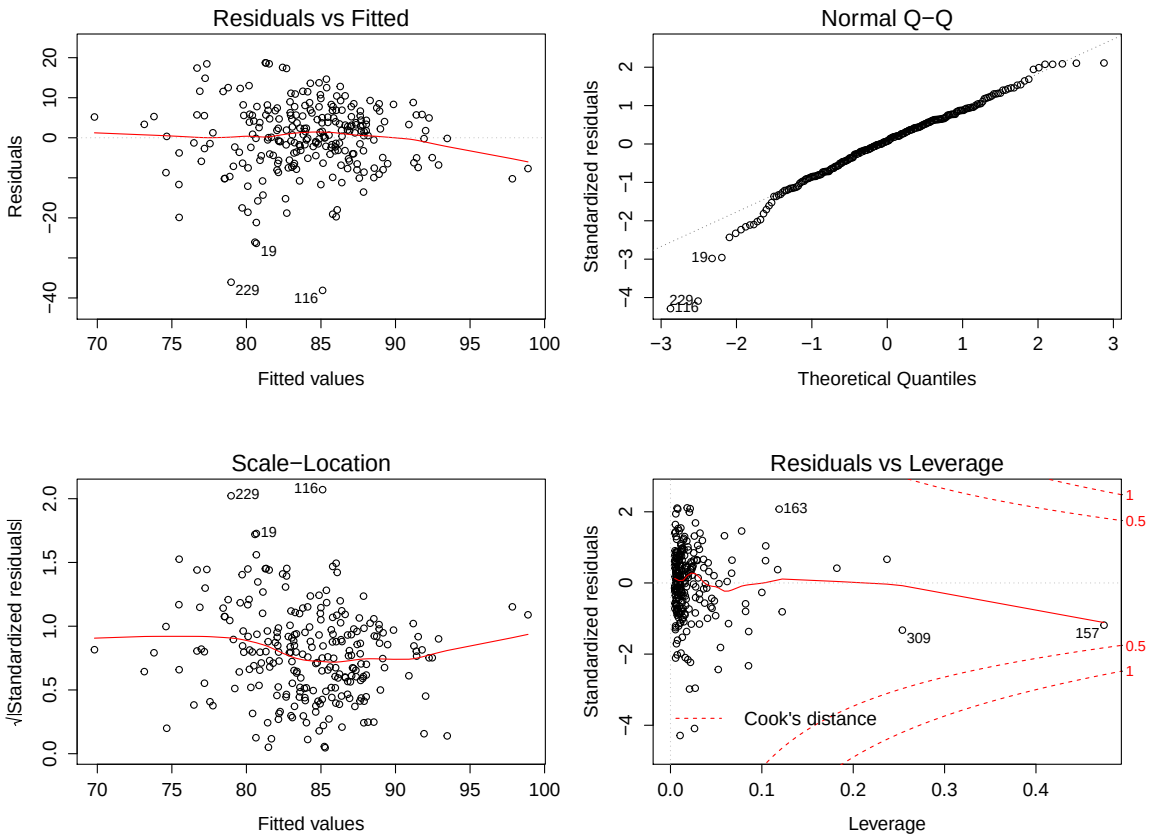

Histogram of BIO
Frequency
0
20
40
60
80
100
20
40
60
80
100
BIO
Histogram of PHYS
Frequency
0
10
20
30
40
50
60
70
30
40
50
60
70
80
90
100
PHYS
Histogram of USH
Frequency
0
20
40
60
80
20
40
60
80
100
USH
Histogram of ECON
Frequency
0
20
40
60
80
100
20
40
60
80
100
ECON


Histogram of AVG.SALARY
Frequency
0
50
100
150
200
20000
30000
40000
50000
60000
AVG.SALARY

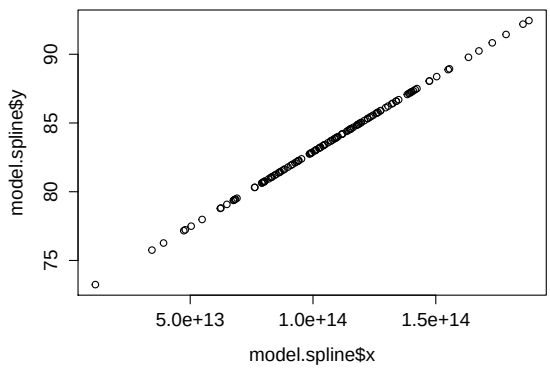

Histogram of CR.RATIO
Frequency
0
50
100
150
200
5
10
15
20
25
30
35
40
CR.RATIO

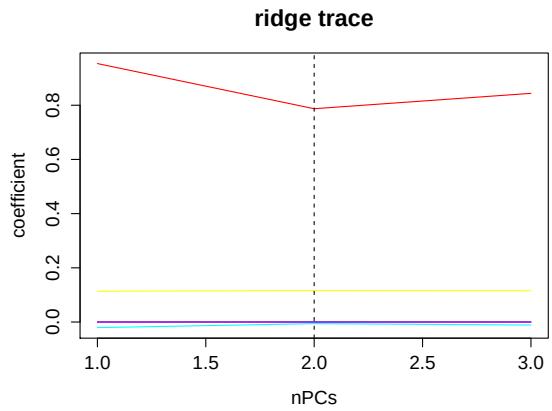

Histogram of TECH.RATIO
Frequency
0
50
100
150
0
2
4
6
8
10
TECH.RATIO

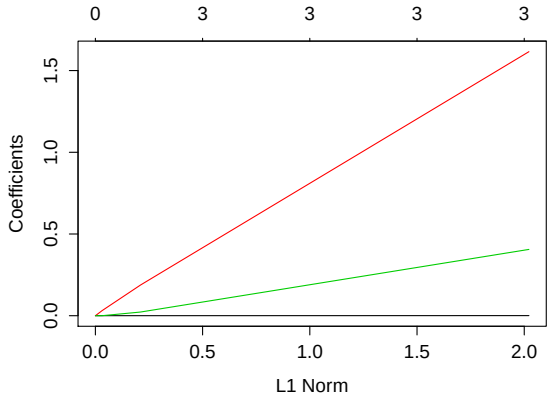
Ninth Grade Literature & Composition



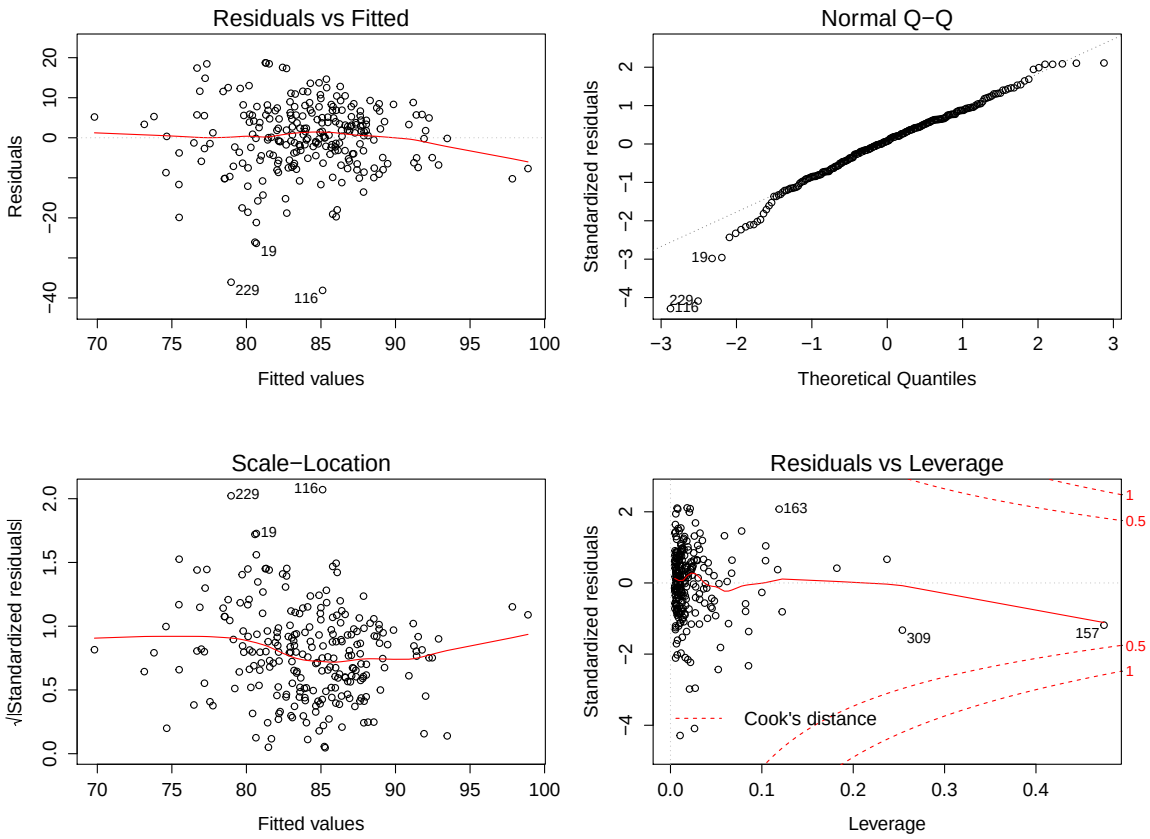
Ninth Grade Literature & Composition



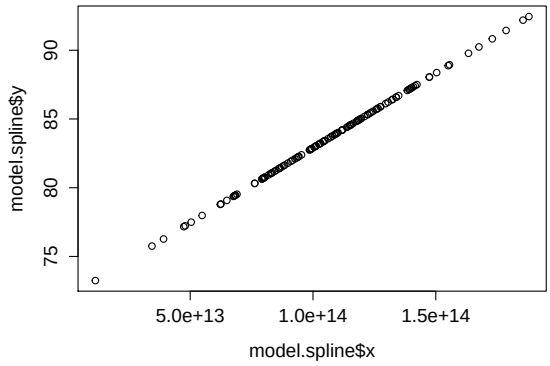


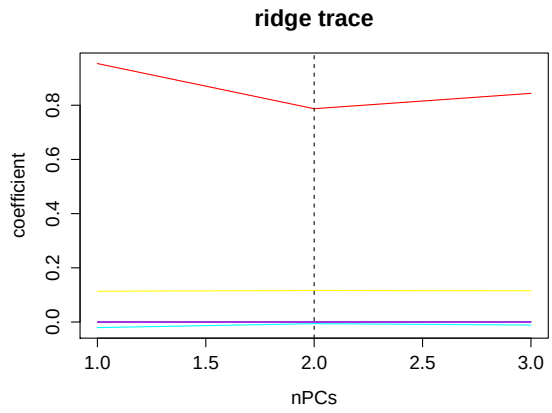


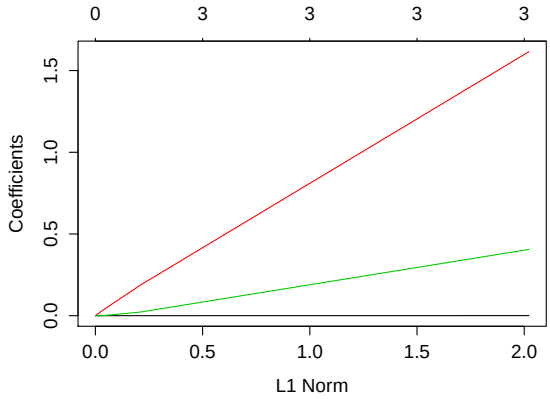
American Literature & Composition



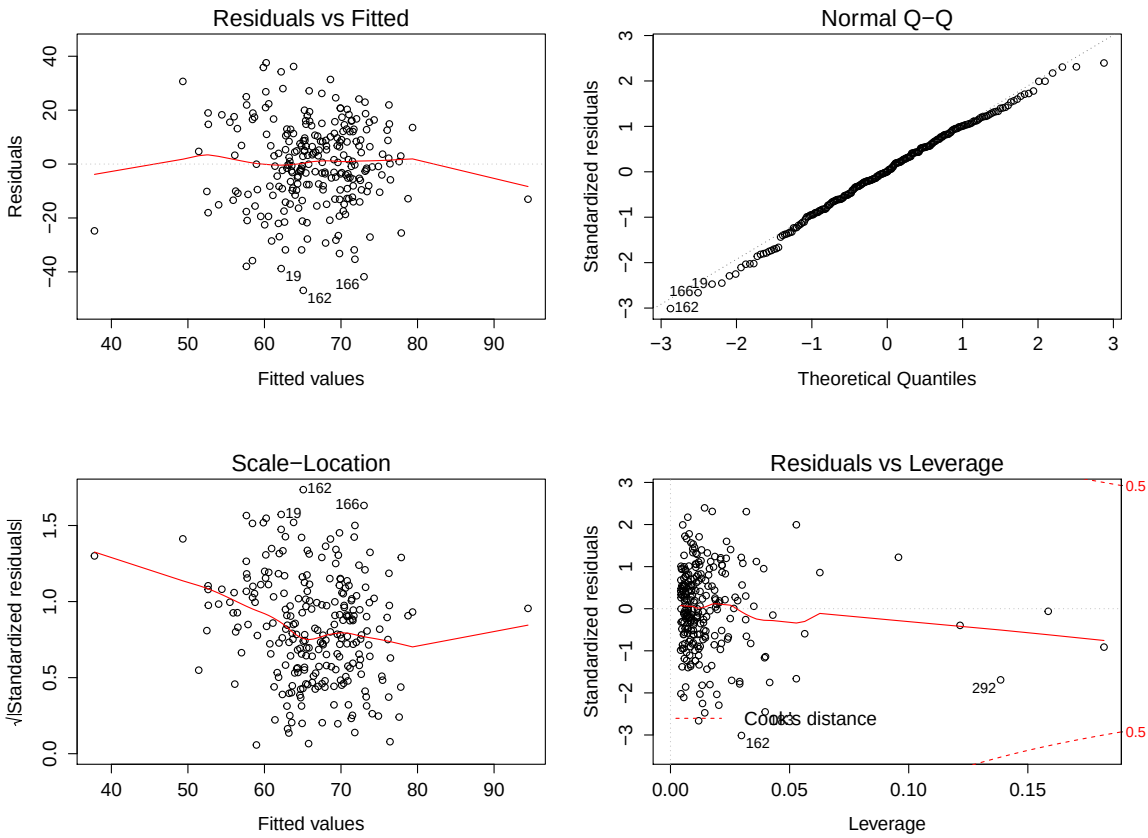
American Literature & Composition



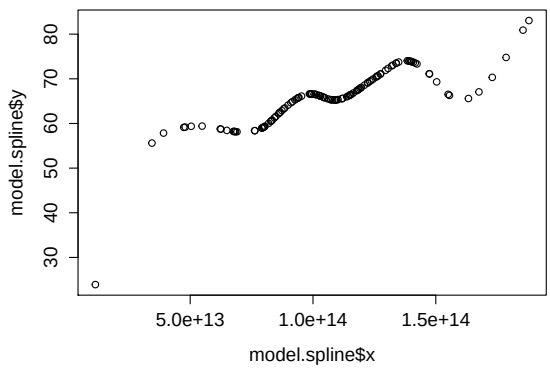


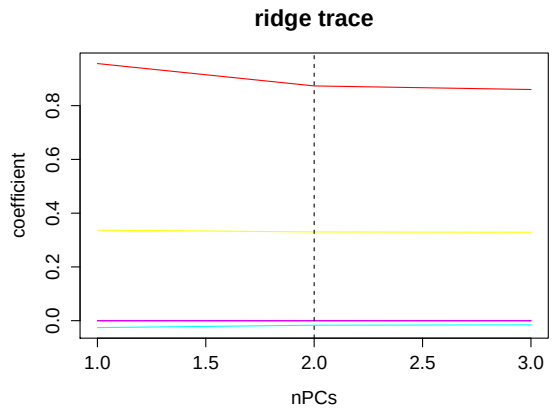


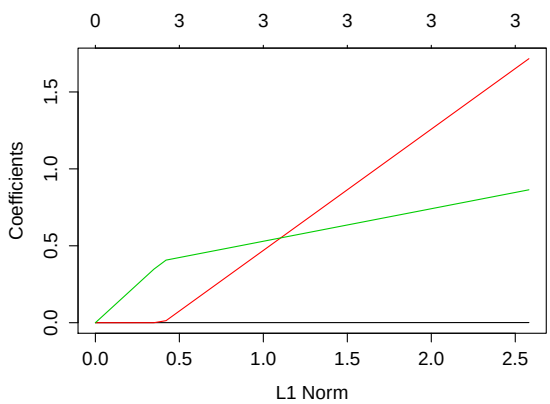
Mathematics I



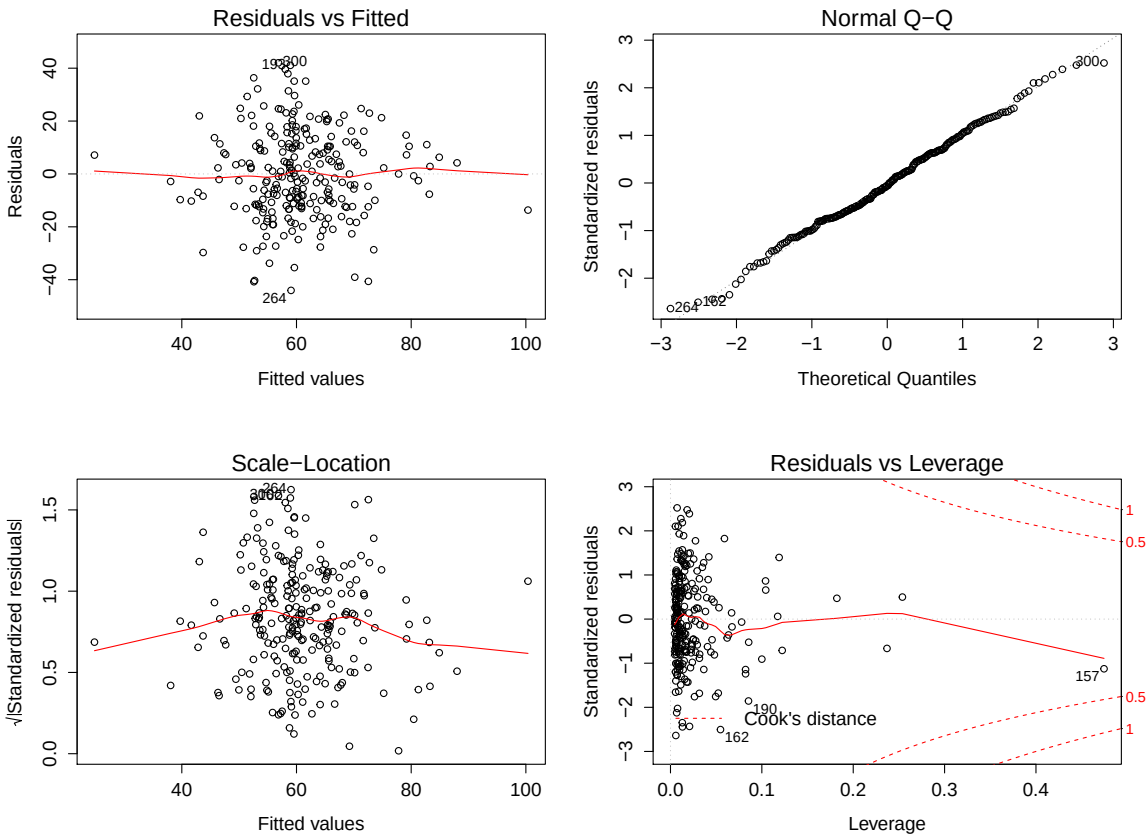
Mathematics I



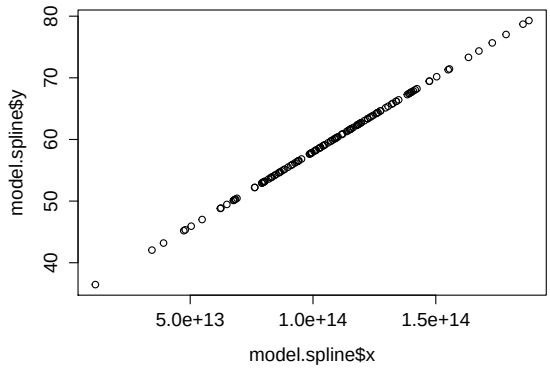


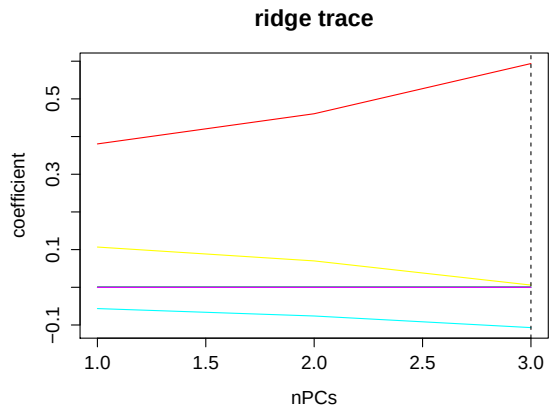


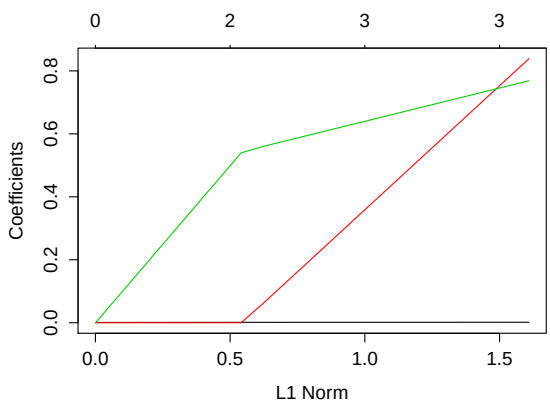
Mathematics II



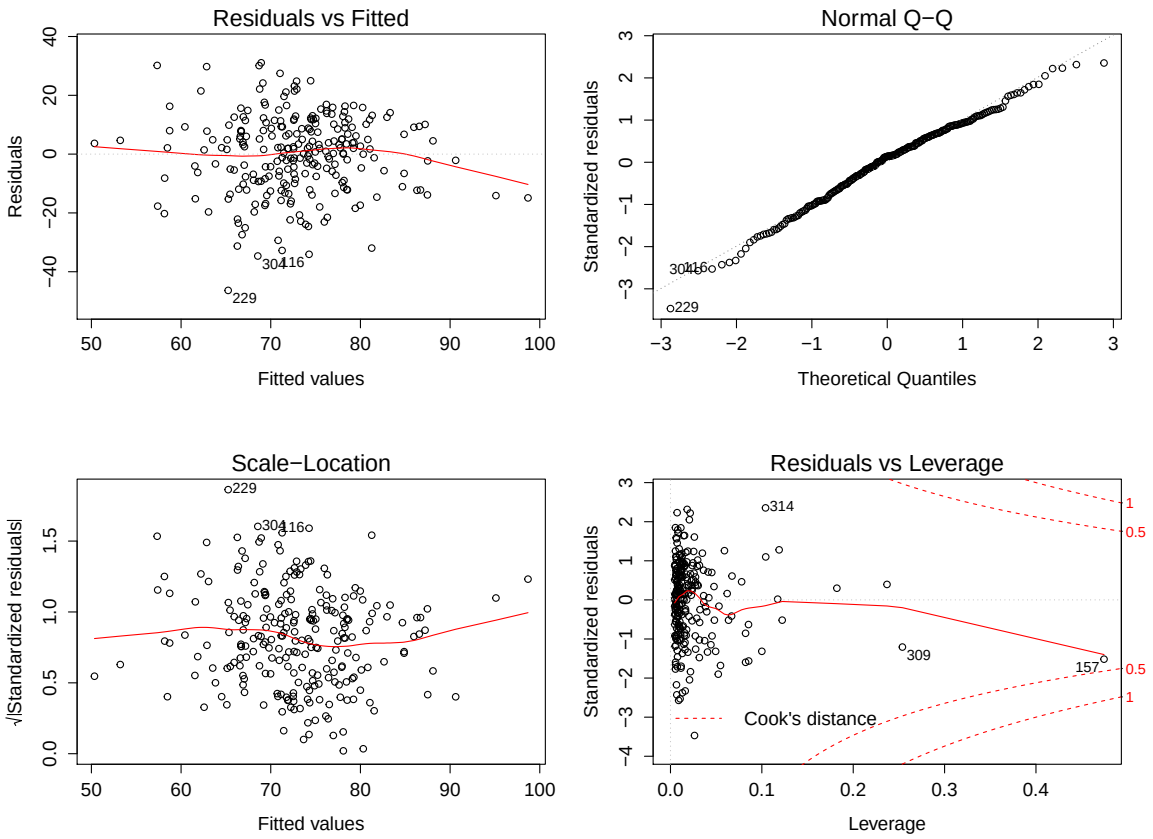
Mathematics II



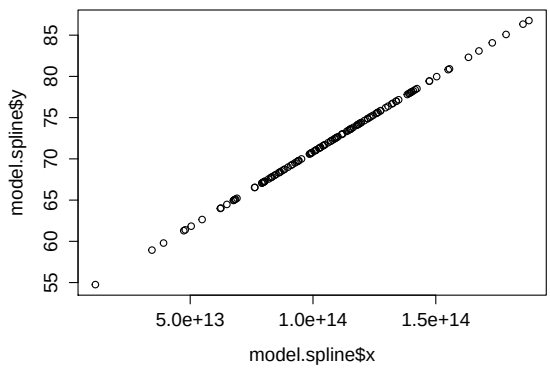


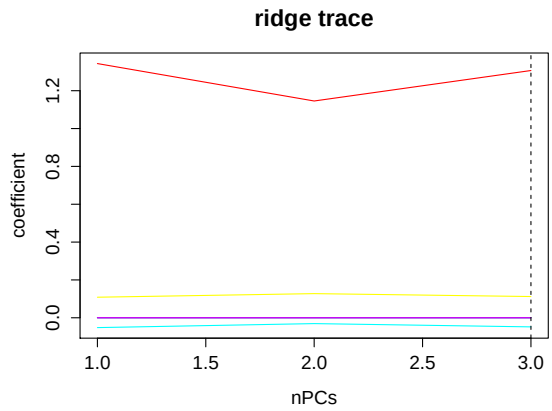


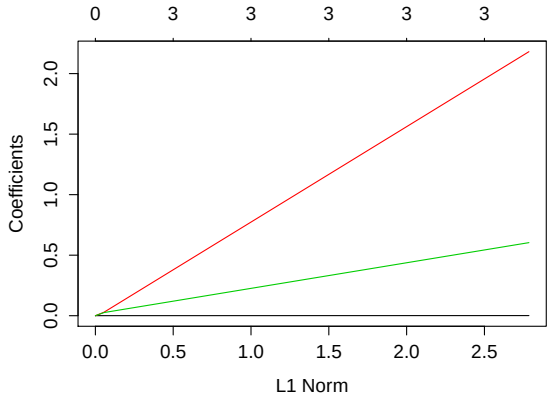
Biology


Residuals vs Fitted
Residuals
Fitted values
229
304
416
Normal Q−Q
Standardized residuals
Theoretical Quantiles
229
304
416
Scale−Location
√|Standardized residuals|
Fitted values
229
304
416
Residuals vs Leverage
Standardized residuals
Leverage
314
309
157
Cook's distance
1
0.5
0.5
1

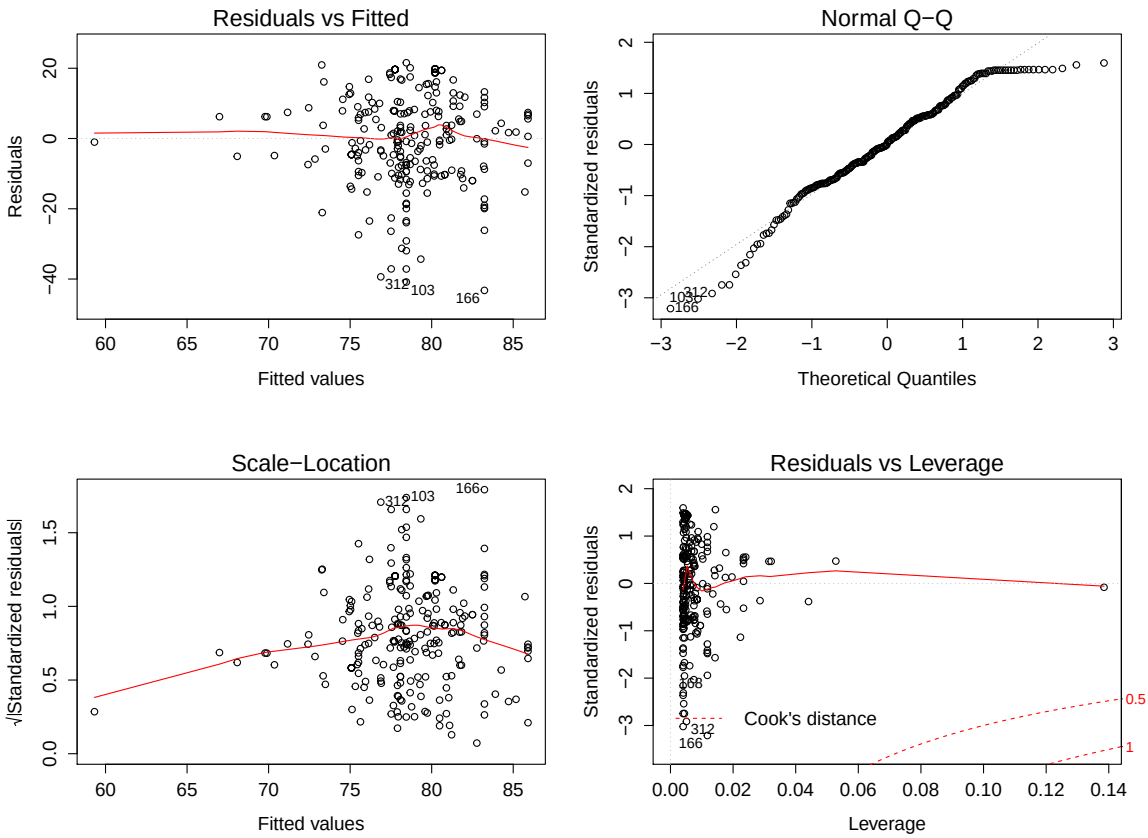
Biology


model.spline$y
55
60
65
70
75
80
85
5.0e+13
1.0e+14
1.5e+14
model.spline$x

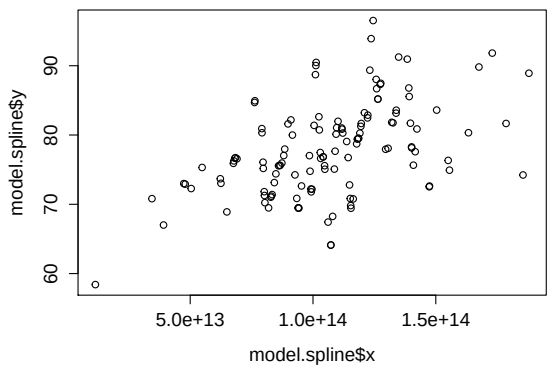

ridge trace
coefficient
0.0
0.4
0.8
1.2
1.0
1.5
2.0
2.5
3.0
nPCs

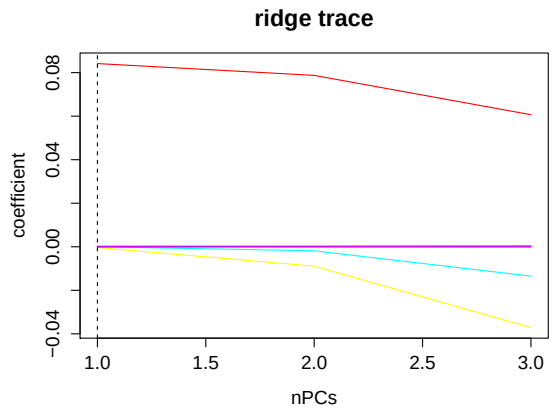

0
3
3
3
3
3
Coefficients
0.0
0.5
1.0
1.5
2.0
0.0
0.5
1.0
1.5
2.0
2.5
L1 Norm

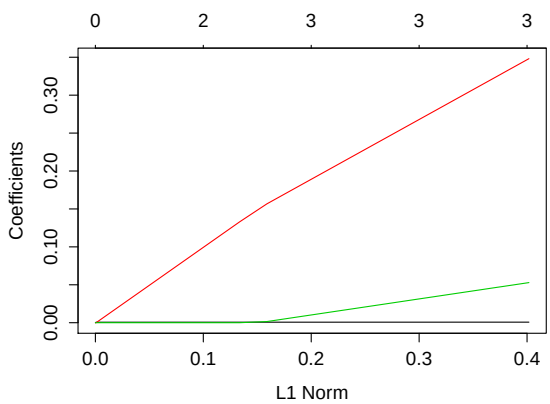
Physical Science



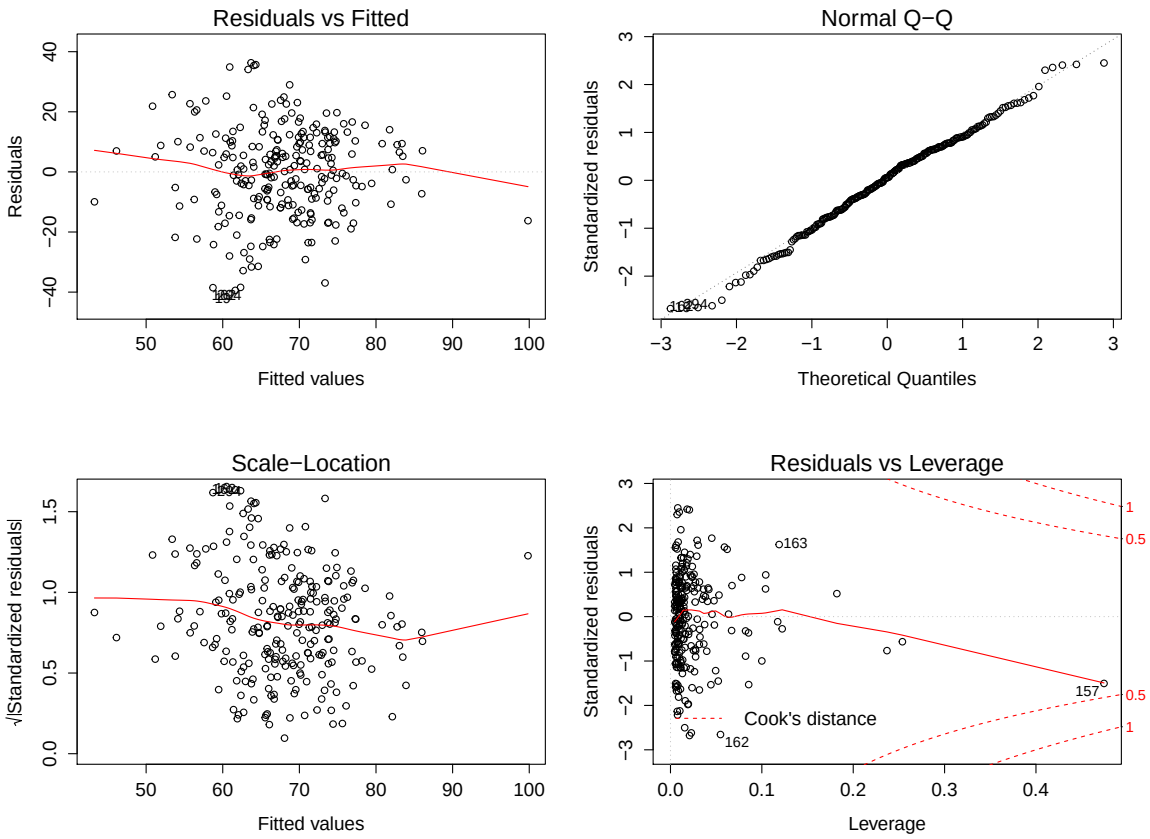
Physical Science



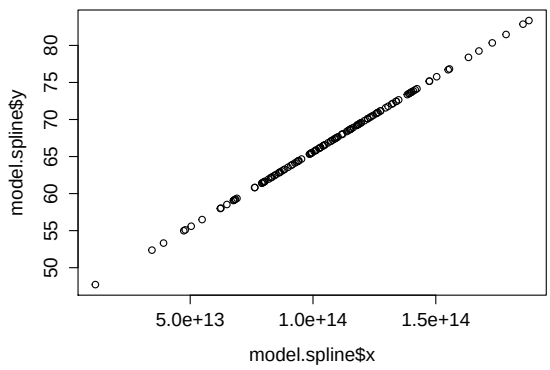


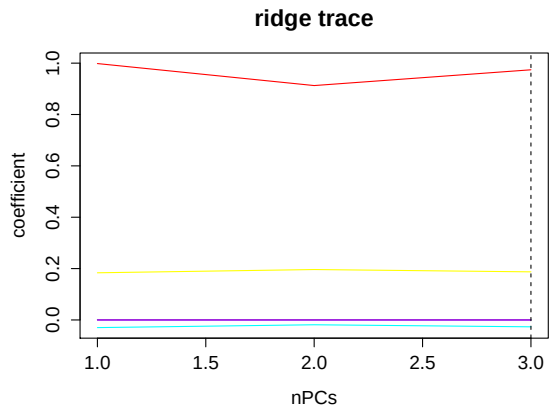


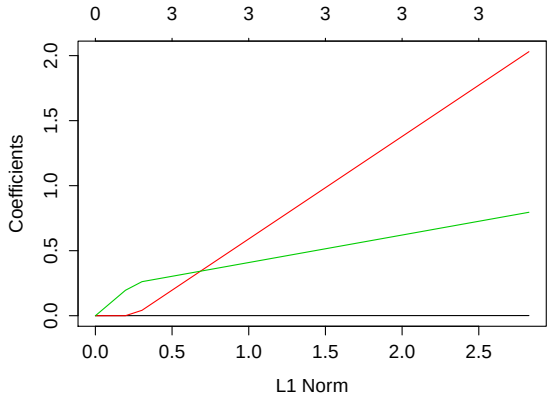
United States History



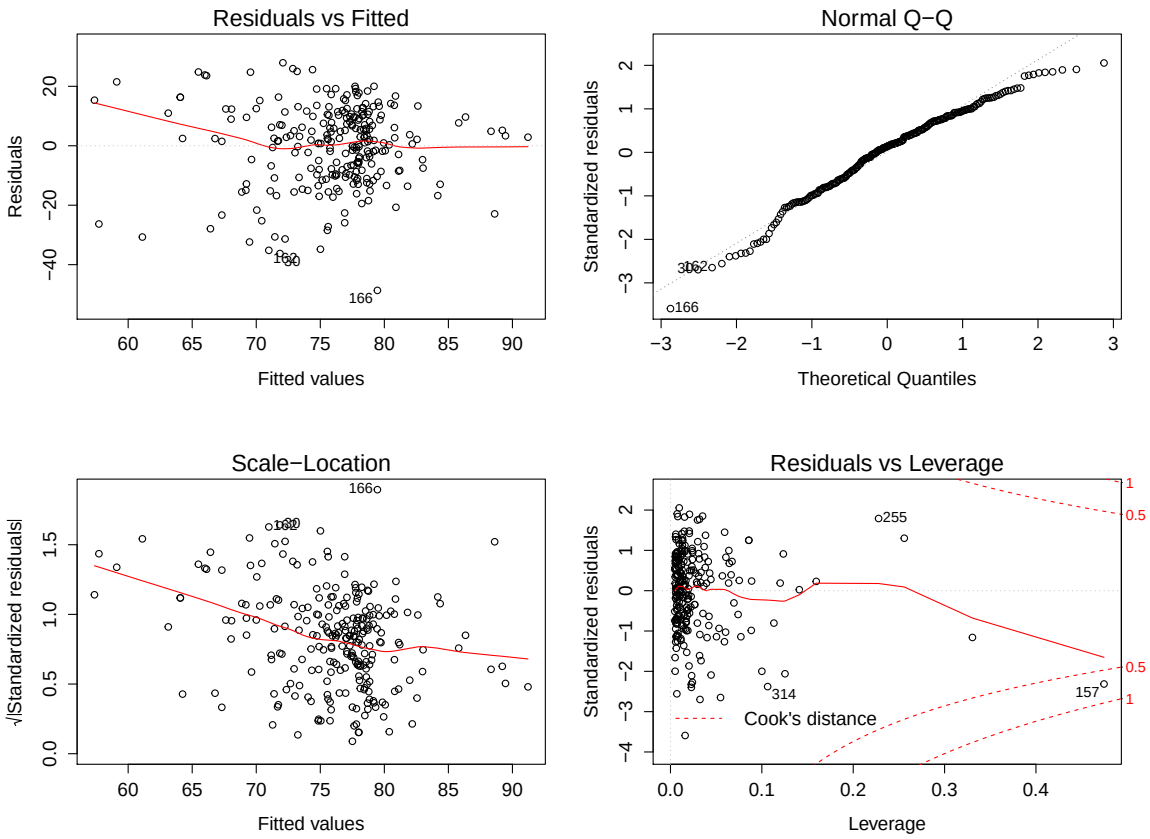
United States History



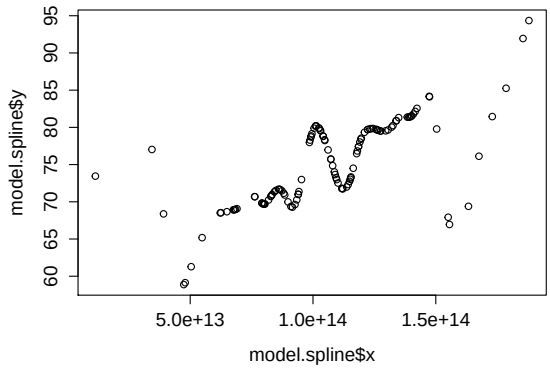


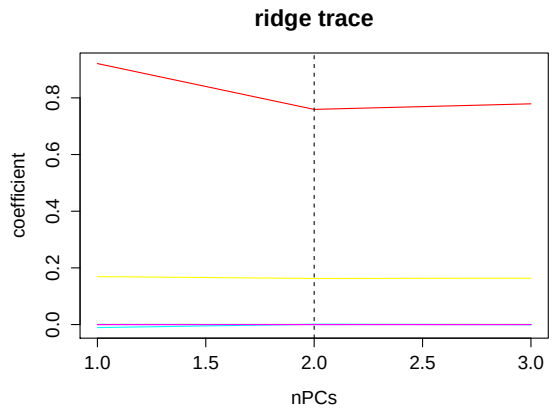


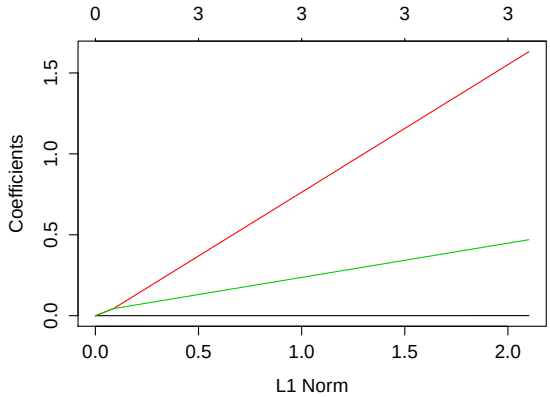
Economics


Residuals vs Fitted
Residuals
Fitted values
Normal Q−Q
Standardized residuals
Theoretical Quantiles
Scale−Location
√|Standardized residuals|
Fitted values
Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance

Economics


model.spline$y
60
65
70
75
80
85
90
95
5.0e+13
1.0e+14
1.5e+14
model.spline$x


ridge trace
coefficient
0.0
0.2
0.4
0.6
0.8
1.0
1.5
2.0
2.5
3.0
nPCs


0
3
3
3
3
Coefficients
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
L1 Norm

R Code

```
# Ninth Grade Literature & Composition

# Generating random subsets (test and train) of our population (data)
# for Cross-Validation.
set.seed(100)
data <- read.csv(file="dataCompiledDraft6.csv", header=T);
indexes <- sample(1:nrow(data), size=122, replace=FALSE)
test <- data[indexes,]
dim(test)  # 6 11
train <- data[-indexes,]
dim(train) # 26 11

# Load Packages
library(stats)
library(MASS)
library(ridge)
library(glmnet)
library(lars)

# Stepwise Regression   ### WORKS ###
model <-
lm(X9LIT~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+
              CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
              AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+
              TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,
data=train)
model.step <- stepAIC(model, direction="both")
pred <- predict(model.step,newdata=test,interval="confidence", se.fit=TRUE)
rmse.step <- pred$residual.scale

# Smooth Splines        ### WORKS ###
attach(train)
model.spline <-
smooth.spline(TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+

CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+

AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,train$X9LIT)
detach(train)
dim(test)
predicted <- predict(model.spline,newdata=test)
predicted <- as.data.frame(predicted)
predicted <- predicted[1:122,]
predicted <- predicted[,2]
predicted <- as.matrix(predicted)
observed <- as.matrix(test$X9LIT)
rmse.spline <- sqrt(sum((observed-predicted)^2)/122)

# Ridge Regression      ### WORKS ###
model.ridge <-
linearRidge(X9LIT~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+


CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
```

```
AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,data=train)
predicted <- predict(model.ridge, newdata=test)
rmse.ridge <- sqrt(sum((observed-predicted)^2)/122)

# Lasso Regression
x <- as.matrix(train)
x <- x[,1:3]
y <- train$X9LIT
y <- as.matrix(y)
model.lasso <- glmnet(x,y,family="gaussian",alpha=1)
newx <- as.matrix(test)
newx <- newx[,1:3]
predicted <- predict(model.lasso,newx)
predicted <- predicted[,56]
predicted <- as.matrix(predicted)
rmse.lasso <- sqrt(sum((observed-predicted)^2)/122)

# RMSE Summary
rmse.X9LIT <- c(rmse.lasso,rmse.ridge,rmse.spline,rmse.step)
rmse.X9LIT

# End of File
```

```
# American Literature & Composition

# Generating random subsets (test and train) of our population (data)
# for Cross-Validation.set.seed(100)
data <- read.csv(file="dataCompiledDraft6.csv", header=T);
indexes <- sample(1:nrow(data), size=122, replace=FALSE)
test <- data[indexes,]
dim(test)  # 6 11
train <- data[-indexes,]
dim(train) # 26 11

# Load Packages
library(stats)
library(MASS)
library(ridge)
library(glmnet)
library(lars)

# Stepwise Regression   ### WORKS ###
model <-
lm(AMLIT~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+
              CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
              AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+
              TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,
data=train)
model.step <- stepAIC(model, direction="both")
pred <- predict(model.step,newdata=test,interval="confidence", se.fit=TRUE)
rmse.step <- pred$residual.scale

# Smooth Splines        ### WORKS ###
attach(train)
model.spline <-
smooth.spline(TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+

CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+

AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,train$AMLIT)
detach(train)
dim(test)
predicted <- predict(model.spline,newdata=test)
predicted <- as.data.frame(predicted)
predicted <- predicted[1:122,]
predicted <- predicted[,2]
predicted <- as.matrix(predicted)
observed <- as.matrix(test$AMLIT)
rmse.spline <- sqrt(sum((observed-predicted)^2)/122)

# Ridge Regression      ### WORKS ###
model.ridge <-
linearRidge(AMLIT~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+


CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
```

```
AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,data=train)
predicted <- predict(model.ridge, newdata=test)
rmse.ridge <- sqrt(sum((observed-predicted)^2)/122)

# Lasso Regression
x <- as.matrix(train)
x <- x[,1:3]
y <- train$AMLIT
y <- as.matrix(y)
model.lasso <- glmnet(x,y,family="gaussian",alpha=1)
newx <- as.matrix(test)
newx <- newx[,1:3]
predicted <- predict(model.lasso,newx)
predicted <- predicted[,56]
predicted <- as.matrix(predicted)
rmse.lasso <- sqrt(sum((observed-predicted)^2)/122)

# RMSE Summary
rmse.AMLIT <- c(rmse.lasso,rmse.ridge,rmse.spline,rmse.step)
rmse.AMLIT

# End of File
```

```
# Mathematics I

# Generating random subsets (test and train) of our population (data)
# for Cross-Validation.set.seed(100)
data <- read.csv(file="dataCompiledDraft6.csv", header=T);
indexes <- sample(1:nrow(data), size=122, replace=FALSE)
test <- data[indexes,]
dim(test)  # 6 11
train <- data[-indexes,]
dim(train) # 26 11

# Load Packages
library(stats)
library(MASS)
library(ridge)
library(glmnet)
library(lars)

# Stepwise Regression   ### WORKS ###
model <-
lm(MATH1~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+
              CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
              AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+
              TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,
data=train)
model.step <- stepAIC(model, direction="both")
pred <- predict(model.step,newdata=test,interval="confidence", se.fit=TRUE)
rmse.step <- pred$residual.scale

# Smooth Splines        ### WORKS ###
attach(train)
model.spline <-
smooth.spline(TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+

CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+

AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,train$MATH1)
detach(train)
dim(test)
predicted <- predict(model.spline,newdata=test)
predicted <- as.data.frame(predicted)
predicted <- predicted[1:122,]
predicted <- predicted[,2]
predicted <- as.matrix(predicted)
observed <- as.matrix(test$MATH1)
rmse.spline <- sqrt(sum((observed-predicted)^2)/122)

# Ridge Regression      ### WORKS ###
model.ridge <-
linearRidge(MATH1~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+


CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
```

```
AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,data=train)
predicted <- predict(model.ridge, newdata=test)
rmse.ridge <- sqrt(sum((observed-predicted)^2)/122)

# Lasso Regression
x <- as.matrix(train)
x <- x[,1:3]
y <- train$MATH1
y <- as.matrix(y)
model.lasso <- glmnet(x,y,family="gaussian",alpha=1)
newx <- as.matrix(test)
newx <- newx[,1:3]
predicted <- predict(model.lasso,newx)
predicted <- predicted[,56]
predicted <- as.matrix(predicted)
rmse.lasso <- sqrt(sum((observed-predicted)^2)/122)

# RMSE Summary
rmse.MATH1 <- c(rmse.lasso,rmse.ridge,rmse.spline,rmse.step)
rmse.MATH1

# End of File
```

```
# Mathematics II

# Generating random subsets (test and train) of our population (data)
# for Cross-Validation.set.seed(100)
data <- read.csv(file="dataCompiledDraft6.csv", header=T);
indexes <- sample(1:nrow(data), size=122, replace=FALSE)
test <- data[indexes,]
dim(test)  # 6 11
train <- data[-indexes,]
dim(train) # 26 11

# Load Packages
library(stats)
library(MASS)
library(ridge)
library(glmnet)
library(lars)

# Stepwise Regression   ### WORKS ###
model <-
lm(MATH2~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+
              CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
              AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+
              TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,
data=train)
model.step <- stepAIC(model, direction="both")
pred <- predict(model.step,newdata=test,interval="confidence", se.fit=TRUE)
rmse.step <- pred$residual.scale

# Smooth Splines        ### WORKS ###
attach(train)
model.spline <-
smooth.spline(TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+

CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+

AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,train$MATH2)
detach(train)
dim(test)
predicted <- predict(model.spline,newdata=test)
predicted <- as.data.frame(predicted)
predicted <- predicted[1:122,]
predicted <- predicted[,2]
predicted <- as.matrix(predicted)
observed <- as.matrix(test$MATH2)
rmse.spline <- sqrt(sum((observed-predicted)^2)/122)

# Ridge Regression      ### WORKS ###
model.ridge <-
linearRidge(MATH2~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+


CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
```

```
AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,data=train)
predicted <- predict(model.ridge, newdata=test)
rmse.ridge <- sqrt(sum((observed-predicted)^2)/122)

# Lasso Regression
x <- as.matrix(train)
x <- x[,1:3]
y <- train$MATH2
y <- as.matrix(y)
model.lasso <- glmnet(x,y,family="gaussian",alpha=1)
newx <- as.matrix(test)
newx <- newx[,1:3]
predicted <- predict(model.lasso,newx)
predicted <- predicted[,56]
predicted <- as.matrix(predicted)
rmse.lasso <- sqrt(sum((observed-predicted)^2)/122)

# RMSE Summary
rmse.MATH2 <- c(rmse.lasso,rmse.ridge,rmse.spline,rmse.step)
rmse.MATH2

# End of File
```

```
# Biology

# Generating random subsets (test and train) of our population (data)
# for Cross-Validation.set.seed(100)
data <- read.csv(file="dataCompiledDraft6.csv", header=T);
indexes <- sample(1:nrow(data), size=122, replace=FALSE)
test <- data[indexes,]
dim(test)  # 6 11
train <- data[-indexes,]
dim(train) # 26 11

# Load Packages
library(stats)
library(MASS)
library(ridge)
library(glmnet)
library(lars)

# Stepwise Regression   ### WORKS ###
model <-
lm(BIO~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+
              CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
              AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+
              TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,
data=train)
model.step <- stepAIC(model, direction="both")
pred <- predict(model.step,newdata=test,interval="confidence", se.fit=TRUE)
rmse.step <- pred$residual.scale

# Smooth Splines        ### WORKS ###
attach(train)
model.spline <-
smooth.spline(TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+

CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+

AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,train$BIO)
detach(train)
dim(test)
predicted <- predict(model.spline,newdata=test)
predicted <- as.data.frame(predicted)
predicted <- predicted[1:122,]
predicted <- predicted[,2]
predicted <- as.matrix(predicted)
observed <- as.matrix(test$BIO)
rmse.spline <- sqrt(sum((observed-predicted)^2)/122)

# Ridge Regression      ### WORKS ###
model.ridge <-
linearRidge(BIO~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+


CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
```

```
AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,data=train)
predicted <- predict(model.ridge, newdata=test)
rmse.ridge <- sqrt(sum((observed-predicted)^2)/122)

# Lasso Regression
x <- as.matrix(train)
x <- x[,1:3]
y <- train$BIO
y <- as.matrix(y)
model.lasso <- glmnet(x,y,family="gaussian",alpha=1)
newx <- as.matrix(test)
newx <- newx[,1:3]
predicted <- predict(model.lasso,newx)
predicted <- predicted[,56]
predicted <- as.matrix(predicted)
rmse.lasso <- sqrt(sum((observed-predicted)^2)/122)

# RMSE Summary
rmse.BIO <- c(rmse.lasso,rmse.ridge,rmse.spline,rmse.step)
rmse.BIO

# End of File
```

```
# Physical Science

# Generating random subsets (test and train) of our population (data)
# for Cross-Validation.set.seed(100)
data <- read.csv(file="dataCompiledDraft6.csv", header=T);
indexes <- sample(1:nrow(data), size=122, replace=FALSE)
test <- data[indexes,]
dim(test)  # 6 11
train <- data[-indexes,]
dim(train) # 26 11

# Load Packages
library(stats)
library(MASS)
library(ridge)
library(glmnet)
library(lars)

# Stepwise Regression   ### WORKS ###
model <-
lm(PHYS~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+
              CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
              AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+
              TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,
data=train)
model.step <- stepAIC(model, direction="both")
pred <- predict(model.step,newdata=test,interval="confidence", se.fit=TRUE)
rmse.step <- pred$residual.scale

# Smooth Splines        ### WORKS ###
attach(train)
model.spline <-
smooth.spline(TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+

CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+

AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,train$PHYS)
detach(train)
dim(test)
predicted <- predict(model.spline,newdata=test)
predicted <- as.data.frame(predicted)
predicted <- predicted[1:122,]
predicted <- predicted[,2]
predicted <- as.matrix(predicted)
observed <- as.matrix(test$PHYS)
rmse.spline <- sqrt(sum((observed-predicted)^2)/122)

# Ridge Regression      ### WORKS ###
model.ridge <-
linearRidge(PHYS~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+


CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
```

```
AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,data=train)
predicted <- predict(model.ridge, newdata=test)
rmse.ridge <- sqrt(sum((observed-predicted)^2)/122)

# Lasso Regression
x <- as.matrix(train)
x <- x[,1:3]
y <- train$PHYS
y <- as.matrix(y)
model.lasso <- glmnet(x,y,family="gaussian",alpha=1)
newx <- as.matrix(test)
newx <- newx[,1:3]
predicted <- predict(model.lasso,newx)
predicted <- predicted[,56]
predicted <- as.matrix(predicted)
rmse.lasso <- sqrt(sum((observed-predicted)^2)/122)

# RMSE Summary
rmse.PHYS <- c(rmse.lasso,rmse.ridge,rmse.spline,rmse.step)
rmse.PHYS

# End of File
```

```
# United States History

# Generating random subsets (test and train) of our population (data)
# for Cross-Validation.set.seed(100)
data <- read.csv(file="dataCompiledDraft6.csv", header=T);
indexes <- sample(1:nrow(data), size=122, replace=FALSE)
test <- data[indexes,]
dim(test)  # 6 11
train <- data[-indexes,]
dim(train) # 26 11

# Load Packages
library(stats)
library(MASS)
library(ridge)
library(glmnet)
library(lars)

# Stepwise Regression   ### WORKS ###
model <-
lm(USH~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+
              CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
              AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+
              TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,
data=train)
model.step <- stepAIC(model, direction="both")
pred <- predict(model.step,newdata=test,interval="confidence", se.fit=TRUE)
rmse.step <- pred$residual.scale

# Smooth Splines        ### WORKS ###
attach(train)
model.spline <-
smooth.spline(TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+

CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+

AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,train$USH)
detach(train)
dim(test)
predicted <- predict(model.spline,newdata=test)
predicted <- as.data.frame(predicted)
predicted <- predicted[1:122,]
predicted <- predicted[,2]
predicted <- as.matrix(predicted)
observed <- as.matrix(test$USH)
rmse.spline <- sqrt(sum((observed-predicted)^2)/122)

# Ridge Regression      ### WORKS ###
model.ridge <-
linearRidge(USH~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+


CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
```

```
AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,data=train)
predicted <- predict(model.ridge, newdata=test)
rmse.ridge <- sqrt(sum((observed-predicted)^2)/122)

# Lasso Regression
x <- as.matrix(train)
x <- x[,1:3]
y <- train$USH
y <- as.matrix(y)
model.lasso <- glmnet(x,y,family="gaussian",alpha=1)
newx <- as.matrix(test)
newx <- newx[,1:3]
predicted <- predict(model.lasso,newx)
predicted <- predicted[,56]
predicted <- as.matrix(predicted)
rmse.lasso <- sqrt(sum((observed-predicted)^2)/122)

# RMSE Summary
rmse.USH <- c(rmse.lasso,rmse.ridge,rmse.spline,rmse.step)
rmse.USH

# End of File
```

```
# Economics

# Generating random subsets (test and train) of our population (data)
# for Cross-Validation.set.seed(100)
data <- read.csv(file="dataCompiledDraft6.csv", header=T);
indexes <- sample(1:nrow(data), size=122, replace=FALSE)
test <- data[indexes,]
dim(test)  # 6 11
train <- data[-indexes,]
dim(train) # 26 11

# Load Packages
library(stats)
library(MASS)
library(ridge)
library(glmnet)
library(lars)

# Stepwise Regression   ### WORKS ###
model <-
lm(ECON~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+
              CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
              AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+
              TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,
data=train)
model.step <- stepAIC(model, direction="both")
pred <- predict(model.step,newdata=test,interval="confidence", se.fit=TRUE)
rmse.step <- pred$residual.scale

# Smooth Splines        ### WORKS ###
attach(train)
model.spline <-
smooth.spline(TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+

CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+

AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,train$ECON)
detach(train)
dim(test)
predicted <- predict(model.spline,newdata=test)
predicted <- as.data.frame(predicted)
predicted <- predicted[1:122,]
predicted <- predicted[,2]
predicted <- as.matrix(predicted)
observed <- as.matrix(test$ECON)
rmse.spline <- sqrt(sum((observed-predicted)^2)/122)

# Ridge Regression      ### WORKS ###
model.ridge <-
linearRidge(ECON~TECH.RATIO*TECH.RATIO*TECH.RATIO+TECH.RATIO*TECH.RATIO+TECH.RATIO+


CR.RATIO*CR.RATIO*CR.RATIO+CR.RATIO*CR.RATIO+CR.RATIO+
```

```
AVG.SALARY*AVG.SALARY*AVG.SALARY+AVG.SALARY*AVG.SALARY+AVG.SALARY+

TECH.RATIO*CR.RATIO+TECH.RATIO*AVG.SALARY+CR.RATIO*AVG.SALARY,data=train)
predicted <- predict(model.ridge, newdata=test)
rmse.ridge <- sqrt(sum((observed-predicted)^2)/122)

# Lasso Regression
x <- as.matrix(train)
x <- x[,1:3]
y <- train$ECON
y <- as.matrix(y)
model.lasso <- glmnet(x,y,family="gaussian",alpha=1)
newx <- as.matrix(test)
newx <- newx[,1:3]
predicted <- predict(model.lasso,newx)
predicted <- predicted[,56]
predicted <- as.matrix(predicted)
rmse.lasso <- sqrt(sum((observed-predicted)^2)/122)

# RMSE Summary
rmse.ECON <- c(rmse.lasso,rmse.ridge,rmse.spline,rmse.step)
rmse.ECON

# End of File
```