Statistical Case Study #1

Analysis of Correlates of School Performance

Douglas Locke

March 10, 2018. 8:30 pm

**Author’s Note:**

This paper is 28 pages long. Most of the explanation is in the Appendix, as I wrote that first. I wrote a shorter summary paper but I think the Appendix contains all the additional explanation sought for the assignment.

**Memo & Introduction:**

The purpose of this study is to understand the predictors of school level variation in academic performance within California elementary schools. 400 elementary schools have been chosen at random for inclusion in the study.

**A (Brief) Survey of Related Educational Studies & Literature:**

A number of academic papers & books were researched prior to review of the data sets. Some of these papers did not specifically focusing on elementary schools, nor contained broad state-level samples, however they did provide some domain immersion experience and understanding of the types of demographic factors that may affect student academic performance.

In “The Congressionally Mandated Study of Education Growth and Opportunity: First Year Report on Language Minority and Limited English Proficient Students” (Moss, Puma, 1995) suggests “limited English proficiency students.. are particularly disadvantaged…. They come from poor families, and live in urban communities with high concentrations of poverty… their parents rarely speak English at home.”

In “Parent Involvement in Early Intervention for Disadvantaged Children: Does it Matter?” (Miedal, Reynolds, 1999) found that parental involvement was significantly related to higher reading achievement.

In the book “Growing Up With a Single Parent, What Hurts, What Helps” (McLanahan, Sandefur, 1994) the authors suggest using national surveys and multi-decade research that children whose parents live apart are nearly twice as likely to drop out of high school. While not specifically related to elementary school, still this is an interesting finding.

In “Teacher Variables as Predictors of Academic Achievement of Primary School Pupils Mathematics” (Tella, 2008) found that teacher’s self-efficacy and interest were important variables in predicting strong math outcomes in students. They found that attitude, qualifications, and Experience were **not** significantly correlated with student math achievement.

Many other papers I found related similar social type themes; children who exhibit self-regulating behavior, maintain social involvement with peers & teachers.

**Summary of Assumptions:**

Given the above studies & articles as background, it appeared that parental involvement would be a very important factor. Since it is very difficult to measure “parental involvement” other factors may act as proxies. “Household Income” can in a way act as a proxy for parental involvement. For example, a family with a high income may in effect be able to afford to spend more time out of work with their children and encourage their learning and growth. This is true when one parent works, and the other parent is able to stay at home with the children. High income is usually correlated with higher education levels, which in turn may these parents may also place an importance on learning and education in their childrens lives.

**Summary of Data Acquisition & Preparation:**

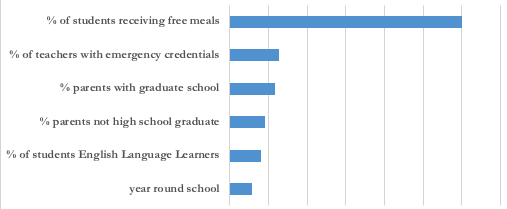
The data set provided identifies each observation at a school level. Academic performance was the choice for the dependent variable. The data set contained proxies for household income, parental education levels, as well as variables like average class size and credentialing of teachers. The variables were rank ordered for entry into the model. Some information that I would have like to have tested but was not available in the model included things like % of parents married, % of parents living in the same home, and early student literacy rate were not available. This may explain at least some of the missing signal the final model of this study did not capture.

From the initial data set of 400 schools, 379 were chosen for the modeling because of data completeness.

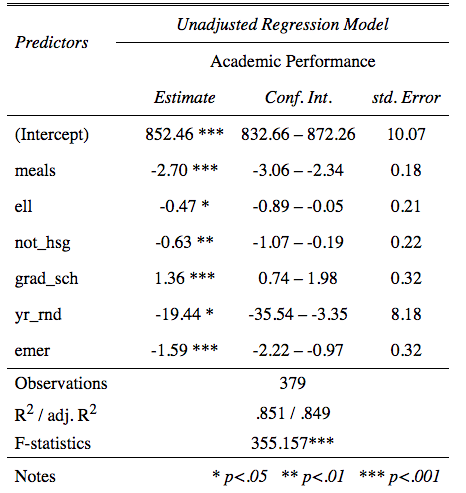
**Summary of Model:**

The final model contains 6 independent variables. The model relies heavily upon “% of students receiving free meals” to drive its prediction power.

Relative Variable Importance (using Standardized Betas)



The model is a linear regression model, it is statistically significant, and it can explain 85% of the variance in academic performance. There is very little multi-collinearity present. Several schools were flagged for further investigation as they seemed to be outliers that exert undue influence in the model. The appendix contains the details of how many and how to find them in the data. However these were not present frequently enough to be of concern to the integrity of the model.



For full details of the model, how it was created, and explanations of the model statistics, see the appendix.

**Future:**

Future data collection may increase the focus on at home social behaviors to further add accuracy to the prediction.

**APPENDIX:**

**Forming Assumptions:**

Following Step 1:

*Without looking at the data, record expectations: what factors are likely to explain school performance (make a ‘wish list’ of independent variables)?*

After completing some literature review, the following rank-ordered feature list was constructed:

1. % Parents married indicator
2. % Parents living in same home indicator
3. Household Income
4. % English Spoken At Home
5. Mother education level
6. % Student literacy rate in first grade
7. Avg Class Size
8. % Student Absenteeism rate
9. % Parent incarceration rate
10. % Parental alcohol/drug abuse
11. % Student suspension rate

*Step 2: Reconcile “wish list” with available data. Take note of variables that you can’t measure because they aren’t available (to gauge omitted variable bias). List those variables here.*

Variables not in the provided data set:

* Parents Married
* Parents living in same home
* Parent Incarceration Rate
* % Student Literacy rate in first grade
* % Parental alcohol/drug abuse

Similar variables in data set

|  |  |
| --- | --- |
| **My Assumption List** | **Similar Variables in Available Data Set** |
| % Parents married indicator | None |
| % Parents living in same home indicator | None |
| Household Income | % students receiving free meals, free meals in 3 categories |
| % English Spoken At Home | % english language learners |
| Mother education level | 5 variables on parent education level |
| % Student literacy rate in first grade | None |
| Class Size | Avg class size |
| % Student Absenteeism rate | None |
| Parent incarceration rate | None |
| % Parental alcohol/drug abuse | None |
| % Student suspension rate | None |

**Available Data Set Description:**

|  |  |
| --- | --- |
| **Variable Name** | **Variable Label** |
| snum | school number |
| dnum | district number |
| acadperf | schoolwide academic performance |
| meals | % of students receiving free meals: this is a proxy for low income school districts |
| ell | % english language learners |
| yr\_rnd | year round school (dummy coded). Schools that have year-round schedules do so primarily to maximize school building use, as they cut out the off period of summer school. This happens more often in urban, overcrowded areas. There is some debate about their effectiveness as discussed here: http://en.wikipedia.org/wiki/Year-round\_school\_in\_the\_United\_States |
| mobility | % 1st year in school |
| acs | avg class size (but note that there was a cap on class sizes in CA during this time so the range is not large) |
| not\_hsg | % parent not hs grad |
| hsg | % parent hs grad |
| some\_col | %parent some college |
| col\_grad | %parent college grad |
| grad\_sch | %parent grad school |
| full | % teachers with full credentials |
| emer | % teachers with emergency credential. Teachers with emergency credentials complete their graduate work while they are teaching. They have less training/experience and often work in more distressed neighborhoods where there are shortages of qualified teachers. |
| enroll | number of students |
| mealcat | Percentage free meals in 3 categories |

**Approach to Variable Testing:**

*Step 3: Create a list of the variables in your wish list that are available in the data (or have close proxies). These are your candidate independent variables.*

The variables I wish to test first are:

* % students receiving free meals ( as a proxy for income)
* % English language learners
* The 5 variables on parent education level (as a possible distant proxy for parental involvement; my assumption is that higher educated parents spend more time with their children, lower educated parents may work more hours and spend less time with their children)

The other variables that were not part of my assumption list but seemed interesting (in an intuition sense) to also test are:

* Year round (as another proxy for income)
* Average Class Size (with the assumption that small class sizes = better academic outcomes)

Given my research, I am less optimistic about the prediction power of:

* % teachers with full credentials
* % teachers with emergency credential

My own assumption is that teacher credentialing is not as important as teacher enthusiasm and dedication, but those are difficult to observe and measure.

**Perform Data Checks:**

*Step 4: Perform basic checks of the candidate variables. Do you have any missing value or out of range data problems? (if so, what did you do to resolve them, if anything?).*

> describe(calschool)

vars n mean sd median trimmed mad min max range skew kurtosis se

snum 1 400 2866.81 1543.81 3007.5 2880.86 1894.02 58 6072 6014 -0.01 -1.03 77.19

dnum 2 400 457.74 184.82 401.0 468.53 284.66 41 796 755 -0.35 -0.78 9.24

acadperf 3 400 647.62 142.25 643.0 645.79 177.17 369 940 571 0.10 -1.13 7.11

meals 4 400 60.31 31.91 67.5 62.18 37.81 0 100 100 -0.41 -1.20 1.60

ell 5 400 31.45 24.84 25.0 29.39 28.17 0 91 91 0.57 -0.87 1.24

yr\_rnd 6 400 0.23 0.42 0.0 0.16 0.00 0 1 1 1.28 -0.37 0.02

mobility 7 399 18.25 7.48 17.0 17.66 5.93 2 47 45 0.83 1.14 0.37

acs 8 398 19.16 1.37 19.0 19.21 1.48 14 25 11 -0.23 1.64 0.07

not\_hsg 9 400 21.25 20.68 14.0 18.65 19.27 0 100 100 0.99 0.44 1.03

hsg 10 400 26.02 16.33 26.0 25.29 13.34 0 100 100 0.95 3.08 0.82

some\_col 11 400 19.71 11.34 19.0 19.65 11.86 0 67 67 0.25 0.13 0.57

col\_grad 12 400 19.70 16.47 16.0 18.12 16.31 0 100 100 1.47 4.32 0.82

grad\_sch 13 400 8.64 12.13 4.0 5.85 5.93 0 67 67 2.16 4.72 0.61

full 14 400 84.55 14.95 88.0 86.60 14.83 37 100 63 -0.97 0.17 0.75

emer 15 400 12.66 11.75 10.0 11.14 10.38 0 59 59 1.06 0.76 0.59

enroll 16 400 483.46 226.45 435.0 459.41 202.37 130 1570 1440 1.34 3.02 11.32

mealcat 17 400 2.02 0.82 2.0 2.02 1.48 1 3 2 -0.03 -1.51 0.04

Variables mobility and acs have a very small amount of missing values. They are removed thusly:

calschooldist2=na.omit(calschooldist)

describe(calschooldist2)

vars n mean sd median trimmed mad min max range skew kurtosis se

snum 1 398 2869.53 1539.25 3007.5 2881.53 1892.54 58 6072 6014 -0.01 -1.02 77.16

dnum 2 398 457.71 184.90 401.0 468.07 284.66 41 796 755 -0.35 -0.78 9.27

acadperf 3 398 648.47 142.08 643.0 646.85 176.43 369 940 571 0.09 -1.13 7.12

meals 4 398 60.16 31.91 67.0 61.96 38.55 0 100 100 -0.41 -1.21 1.60

ell 5 398 31.29 24.80 25.0 29.22 28.17 0 91 91 0.59 -0.84 1.24

yr\_rnd 6 398 0.23 0.42 0.0 0.17 0.00 0 1 1 1.27 -0.39 0.02

mobility 7 398 18.26 7.49 17.0 17.67 5.93 2 47 45 0.83 1.13 0.38

acs 8 398 19.16 1.37 19.0 19.21 1.48 14 25 11 -0.23 1.64 0.07

not\_hsg 9 398 21.19 20.70 14.0 18.59 19.27 0 100 100 0.99 0.45 1.04

hsg 10 398 25.99 16.37 26.0 25.24 13.34 0 100 100 0.95 3.06 0.82

some\_col 11 398 19.71 11.36 19.0 19.64 11.86 0 67 67 0.25 0.12 0.57

col\_grad 12 398 19.74 16.50 16.0 18.18 16.31 0 100 100 1.47 4.29 0.83

grad\_sch 13 398 8.66 12.16 4.0 5.92 5.93 0 67 67 2.15 4.68 0.61

full 14 398 84.63 14.86 88.0 86.63 14.83 37 100 63 -0.97 0.18 0.74

emer 15 398 12.62 11.67 10.0 11.14 10.38 0 59 59 1.06 0.80 0.58

enroll 16 398 483.21 226.98 433.0 459.19 203.12 130 1570 1440 1.34 3.00 11.38

mealcat 17 398 2.01 0.82 2.0 2.01 1.48 1 3 2 -0.02 -1.51 0.04

Removing 2 observations that had missing values for acs & mobility reduced the data set to 398 variables.

However, another item to test for is the presence of missing data that may be encoded improperly. For example, across the 5 parent education variables, it would be impossible for all 5 variables to actually contain 0 values for one school. One cannot be both a high school graduate and simultaneously not a high school graduate. This would indicate the data was not actually collected and/or recorded. We can test for this using the following:

> calschool\_missing\_teacher\_credentials <- subset(calschool,(hsg==0 & not\_hsg==0 & some\_col==0 & col\_grad==0 & grad\_sch==0))

> describe(calschool\_missing\_teacher\_credentials)

vars n mean sd median trimmed mad min max range skew kurtosis se

snum 1 19 3326.32 1804.08 3258 3336.18 1573.04 413 6072 5659 -0.05 -1.23 413.89

dnum 2 19 307.74 232.85 259 301.00 210.53 41 689 648 0.45 -1.38 53.42

acadperf 3 19 665.11 135.24 655 661.94 173.46 490 894 404 0.26 -1.46 31.03

meals 4 19 51.16 34.29 67 51.24 38.55 4 97 93 -0.22 -1.75 7.87

ell 5 19 28.05 24.11 21 26.41 25.20 0 84 84 0.68 -0.67 5.53

yr\_rnd 6 19 0.00 0.00 0 0.00 0.00 0 0 0 NaN NaN 0.00

mobility 7 19 18.47 7.67 17 18.29 4.45 4 36 32 0.60 0.01 1.76

acs 8 19 19.21 1.96 20 19.24 1.48 15 23 8 -0.32 -0.33 0.45

not\_hsg 9 19 0.00 0.00 0 0.00 0.00 0 0 0 NaN NaN 0.00

hsg 10 19 0.00 0.00 0 0.00 0.00 0 0 0 NaN NaN 0.00

some\_col 11 19 0.00 0.00 0 0.00 0.00 0 0 0 NaN NaN 0.00

col\_grad 12 19 0.00 0.00 0 0.00 0.00 0 0 0 NaN NaN 0.00

grad\_sch 13 19 0.00 0.00 0 0.00 0.00 0 0 0 NaN NaN 0.00

full 14 19 86.74 16.44 94 88.47 8.90 44 100 56 -1.16 0.12 3.77

emer 15 19 13.21 16.60 6 11.94 8.90 0 48 48 0.97 -0.60 3.81

enroll 16 19 422.47 130.61 404 416.53 90.44 198 748 550 0.64 0.24 29.96

mealcat 17 19 1.74 0.73 2 1.71 1.48 1 3 2 0.40 -1.18 0.17

We can see 19 observations in the data set have 0 values across the variables for the parent education levels. As this these variables are included in the list of variables for testing in the model, these 19 observations are removed thusly:

> calschooldist3 <- subset(calschooldist2,!(hsg==0 & not\_hsg==0 & some\_col==0 & col\_grad==0 & grad\_sch==0))

Our final data set for modeling is described thusly:

> describe(calschooldist3)

vars n mean sd median trimmed mad min max range skew kurtosis se

snum 1 379 2846.63 1523.94 3004 2863.49 1872.52 58 6068 6010 -0.02 -1.05 78.28

dnum 2 379 465.23 179.27 401 476.20 284.66 41 796 755 -0.35 -0.74 9.21

acadperf 3 379 647.64 142.53 643 646.04 176.43 369 940 571 0.08 -1.13 7.32

meals 4 379 60.61 31.77 67 62.49 38.55 0 100 100 -0.41 -1.19 1.63

ell 5 379 31.46 24.85 25 29.43 28.17 0 91 91 0.58 -0.86 1.28

yr\_rnd 6 379 0.24 0.43 0 0.18 0.00 0 1 1 1.20 -0.57 0.02

mobility 7 379 18.25 7.49 17 17.66 5.93 2 47 45 0.84 1.17 0.38

acs 8 379 19.16 1.34 19 19.20 1.48 14 25 11 -0.22 1.75 0.07

not\_hsg 9 379 22.25 20.64 16 19.83 20.76 0 100 100 0.95 0.39 1.06

hsg 10 379 27.30 15.67 27 26.51 11.86 0 100 100 1.12 3.78 0.81

some\_col 11 379 20.70 10.72 20 20.50 10.38 0 67 67 0.34 0.37 0.55

col\_grad 12 379 20.73 16.29 18 19.19 16.31 0 100 100 1.51 4.56 0.84

grad\_sch 13 379 9.10 12.30 4 6.40 5.93 0 67 67 2.10 4.38 0.63

full 14 379 84.53 14.79 88 86.49 14.83 37 100 63 -0.96 0.17 0.76

emer 15 379 12.59 11.39 10 11.22 10.38 0 59 59 1.04 0.82 0.59

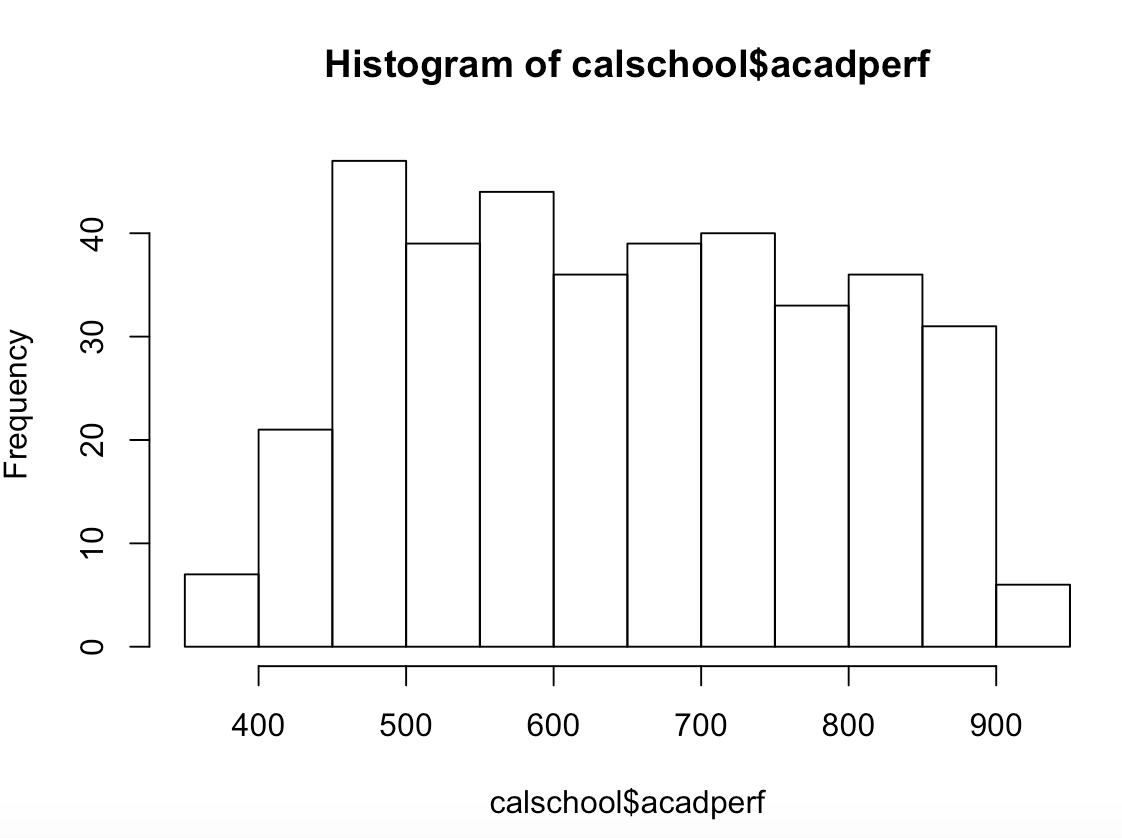
enroll 16 379 486.25 230.44 438 461.90 212.01 130 1570 1440 1.31 2.85 11.84

mealcat 17 379 2.02 0.82 2 2.03 1.48 1 3 2 -0.04 -1.52 0.04

We are now left with 379 observations. Furthermore, because 19 observations out of 400 is ~ 4.7%, or < 5% of our total data set, we will keep the parent education as a potential independent variable.

Target Variable DataCheck:

hist(calschool$acadperf)



In reviewing the target variable, we do not know the unit of measure, or know with complete certainty that higher values indicate stronger academic performance. However, for purpose of this study, it is assumed that higher values are a test score where higher scores indicate higher levels of academic performance.

**Perform Data Correlation Check:**

The Data Correlation check will check for both multicollinearity and identify variables that we may wish to add to the model.

> calschool <- calschooldist3

> cor.prob <- function (X, dfr = nrow(X) - 2) {

+ R <- cor(X, use="pairwise.complete.obs")

+ above <- row(R) < col(R)

+ r2 <- R[above]^2

+ Fstat <- r2 \* dfr/(1 - r2)

+ R[above] <- 1 - pf(Fstat, 1, dfr)

+ R[row(R) == col(R)] <- NA

+ R

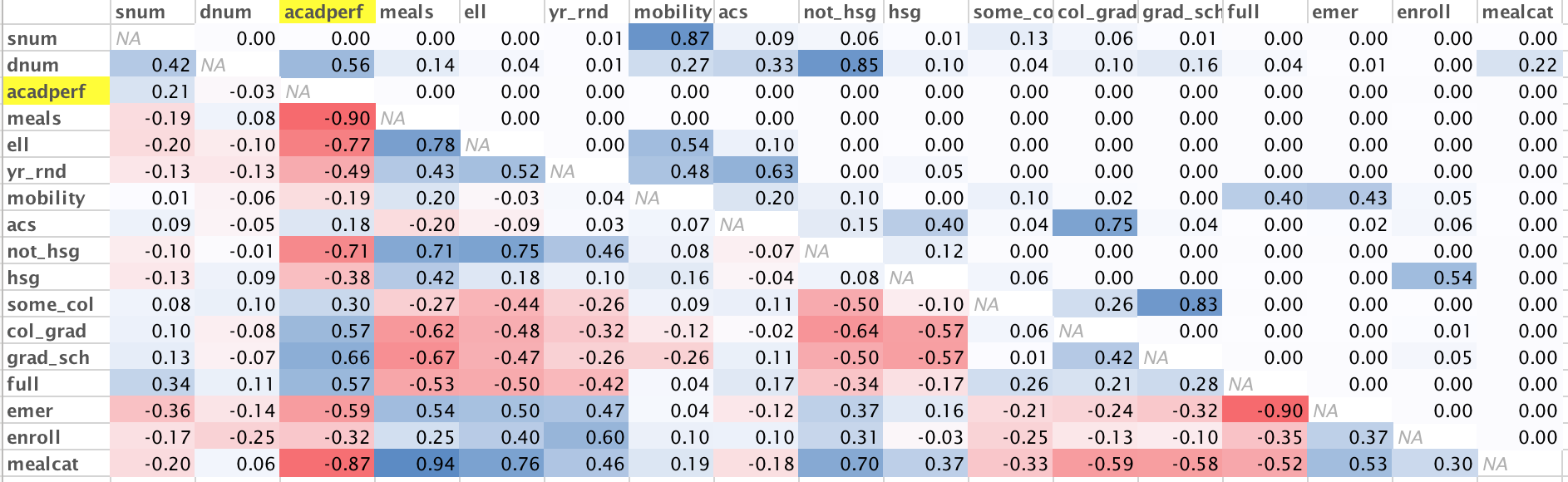
+ }

> correlation\_table\_calschool <- cor.prob(calschool)

> View(correlation\_table\_calschool)

For readability, the result was pasted into Excel.

The dependent variable is highlighted in Yellow. A low (red) to high (blue) color scheme was applied to making reading the table easier.



If we read the row for variable “acadeperf” we see almost all p-values are <.05 , meaning there is a small probably there is no relationship with most all the variables (I’m excluding snum and dnum as they are identifiers, categorical, and thus should not be subject to a quantitative correlation analysis).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Original Variables Chosen For Testing | Correlation Matrix Results With the Target Variable | Interpreted correlation result with the target variable, acadperf | Include in model? | Possible Multi-collinearity? |
| meals | -0.90 | Strong (Negative) | Keep in model | mealcat |
| ell | -0.77 | Strong (Negative) | Keep in model |  |
| Not\_hsg | -0.71 | Strong (Negative) | Keep in model |  |
| hsg | -0.38 | Weak (Negative) | Low priority; enter last if at all |  |
| Some\_col | -0.30 | Weak (Negative) | Low priority; enter last if at all |  |
| Col\_grad | 0.57 | Moderate (Positive) | Keep |  |
| Grad\_sch | 0.66 | Strong (Positive) | Keep |  |
| Original Possible Variables for Inclusion |  |  |  |  |
| Yr\_rnd | -0.49 | Moderate (Negative) | Keep |  |
| acs | 0.18 | No relationship | Remove |  |
| Newly discovered Possible Variables for Inclusion |  |  |  |  |
| mealcat | -0.87 | Strong (Negative) | Do not add; multi-collinearity | meals |
| emer | -0.59 | Moderate (Negative) | Possible add |  |
| full | 0.57 | Moderate (Positive) | Possible add |  |
| enroll | -0.32 | Weak (Negative) | Low priority; enter last if at all |  |

*Step 5: What did your check of the correlation matrix find? Did you add any variables to the end of you list based on it? Does it look like you need to worry about multicollinearity?*

Multi-collinearity problems (any time independent variables correlate > .9)

Meals & Mealcat correlate @ .9. One of these variables should be eliminated from the final model due to multi-collinearity.

**Variables Chosen for Model Build:**

*Step 6: Write down the order of entry based on your best guess given your knowledge of field (protection against specification error) . If you added any variables based on the correlation analysis, add them to the end of your list. They should be given lowest priority since prior expectations did not suggest their importance.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Entry | Variable | Note | Histogram | Scatterplot with Target Variable |
| 1 | Meals | Original |  |  |
| 2 | Ell | Original |  |  |
| 3 | Not\_hsg | Original |  |  |
| 4 | Grad\_Sch | Original |  |  |
| 5 | Col\_grad | Original |  |  |
| 6 | Yr\_rnd | Added due to correlation coefficient |  |  |
| 7 | Emer | Added due to correlation coefficient |  |  |
| 8 | Full | Added due to correlation coefficient |  |  |

Weak variables hsg, some\_col, enroll could be added after the first model build.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Entry | Variable | Note | Histogram | Scatterplot with Target Variable |
| 9 | Hsg | Weak variable, possible addition |  |  |
| 10 | Some\_col | Weak variable, possible addition |  |  |
| 11 | enroll | Weak variable, possible addition |  |  |

Code to generate Scatterplots, each with a simple linear regression line added:

dev.off()

plot(calschool$grad\_sch, calschool$acadperf, main="Scatterplot Grad\_Sch & Academic Performance",

xlab="Grad\_sch % ", ylab="Academic Performance ")

fitline <- lm(calschool$acadperf ~ calschool$grad\_sch)

abline(fitline)

plot(calschool$col\_grad, calschool$acadperf, main="Scatterplot Col\_grad & Academic Performance",

xlab="Col\_grad % ", ylab="Academic Performance ")

fitline <- lm(calschool$acadperf ~ calschool$col\_grad)

abline(fitline)

plot(calschool$yr\_rnd, calschool$acadperf, main="Scatterplot Yr\_rnd & Academic Performance",

xlab="Year Round School ", ylab="Academic Performance ")

fitline <- lm(calschool$acadperf ~ calschool$yr\_rnd)

abline(fitline)

plot(calschool$emer, calschool$acadperf, main="Scatterplot Emer & Academic Performance",

xlab="Emergency Teacher Credential % ", ylab="Academic Performance ")

fitline <- lm(calschool$acadperf ~ calschool$emer)

abline(fitline)

plot(calschool$full, calschool$acadperf, main="Scatterplot Full & Academic Performance",

xlab="Full Teacher Credential % ", ylab="Academic Performance ")

fitline <- lm(calschool$acadperf ~ calschool$full)

abline(fitline)

plot(calschool$hsg, calschool$acadperf, main="Scatterplot Hsg & Academic Performance",

xlab="HSG % ", ylab="Academic Performance ")

fitline <- lm(calschool$acadperf ~ calschool$hsg)

abline(fitline)

plot(calschool$some\_col, calschool$acadperf, main="Scatterplot Some\_col & Academic Performance",

xlab="Some\_col % ", ylab="Academic Performance ")

fitline <- lm(calschool$acadperf ~ calschool$some\_col)

abline(fitline)

plot(calschool$enroll, calschool$acadperf, main="Scatterplot enroll & Academic Performance",

xlab="Enroll % ", ylab="Academic Performance ")

fitline <- lm(calschool$acadperf ~ calschool$enroll)

abline(fitline)

Code to generate Histograms

hist(calschool$meals)

hist(calschool$ell)

hist(calschool$not\_hsg)

hist(calschool$grad\_sch)

hist(calschool$col\_grad)

hist(calschool$yr\_rnd)

hist(calschool$emer)

hist(calschool$full)

hist(calschool$hsg)

hist(calschool$some\_col)

hist(calschool$enroll)

**First Bi-Variate Model:**

*Step 7: Add your first independent variable. Show your bivariate model. Did it accord with your expectations?*

Linear Regression model, using % of students receiving free meals as the first independent variable, and academic performance as the dependent, target variable:

> regression\_1 <- lm(acadperf ~ meals, data = calschool)

> summary(regression\_1)

Call:

lm(formula = acadperf ~ meals, data = calschool)

Residuals:

Min 1Q Median 3Q Max

-203.418 -41.130 -5.338 43.563 193.649

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 892.89393 6.83001 130.73 <0.0000000000000002 \*\*\*

meals -4.04635 0.09983 -40.53 <0.0000000000000002 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 61.66 on 377 degrees of freedom

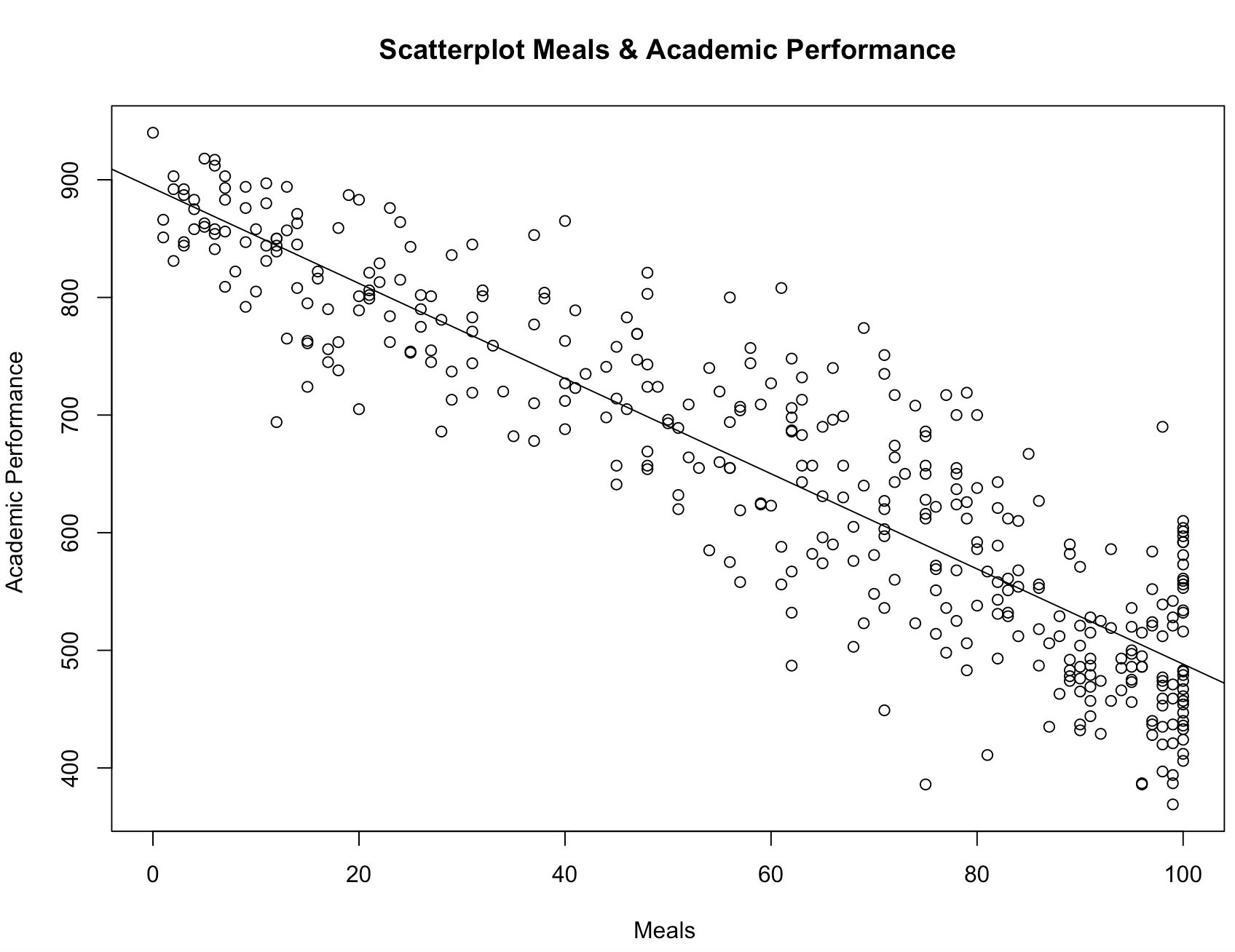
Multiple R-squared: 0.8133, Adjusted R-squared: 0.8128

F-statistic: 1643 on 1 and 377 DF, p-value: < 0.00000000000000022

>plot(calschool$meals, calschool$acadperf, main="Scatterplot Meals & Academic Performance",

xlab="Meals ", ylab="Academic Performance ")

>abline(regression\_1)



Interpretation:

The expectation was that % of students receiving meals, as a proxy for income, would be a key variable to explain academic performance. This was certainly the case, with a strong r-squared (.813) right away.

The intercept tells us that in the absence of students receiving free meals, we can expect academic performance units to equal 892.90. We know the minimum value in the modeling data set = 0, so having no free meals is not a hypothetical but a real-world situation.

The coefficient -4.404 means for every 1% increase in the % of students receiving free meals, we can expect academic performance to drop -4.04 units, while everything else is constant.

The low P value for meals tells us there is statistical significance with the meals variable.

The 3 stars indicates there is a < 1% chance the meals effect on the model is due to randomness.

R-squared of .813 indicated the model accounts for 81.3% of the variance in academic performance. This seems excellent for the first variable tested. However, a model with a high R squared and only one independent variable is attributing its signal to only one factor.

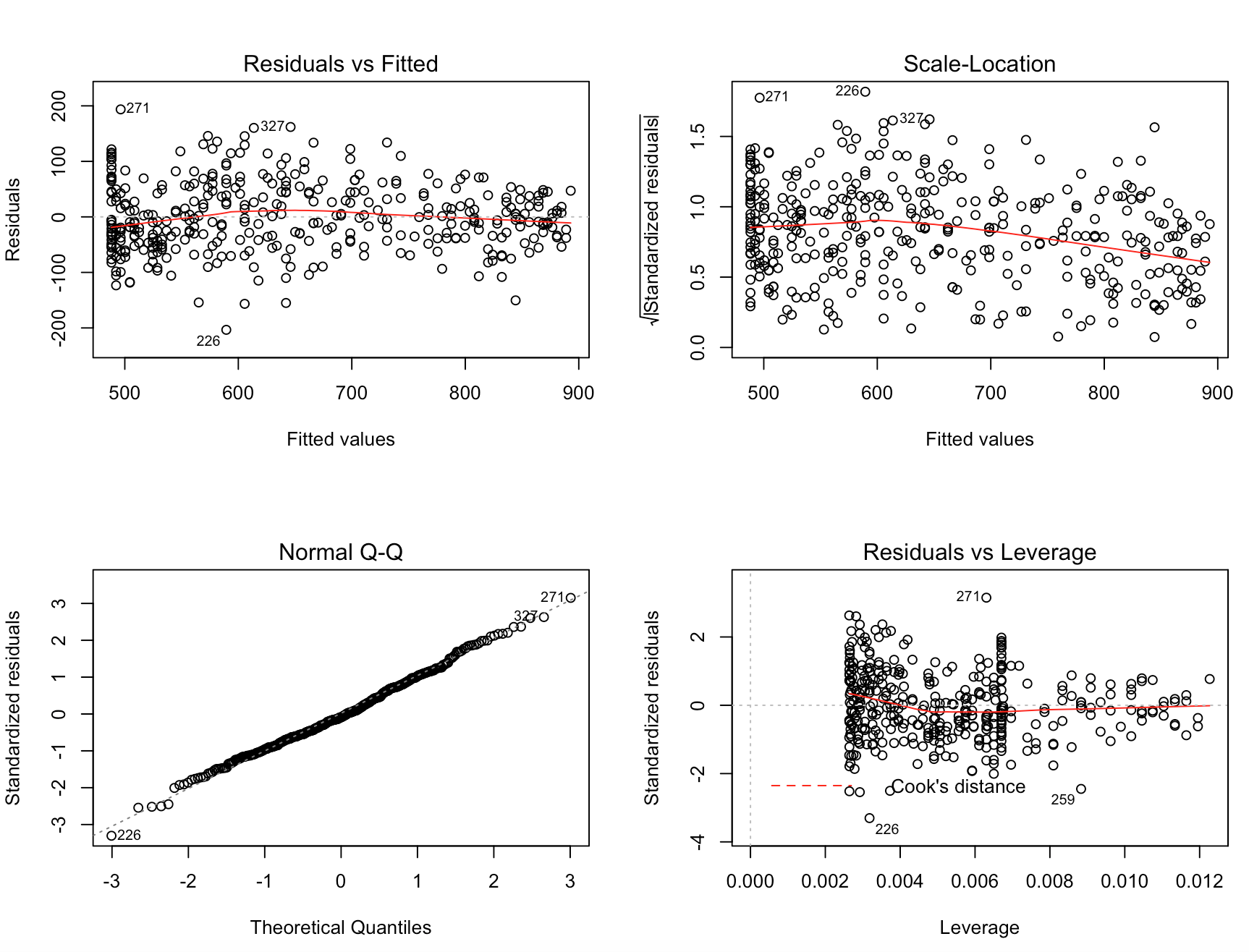
Residual standard error 61.66 is the average error in predicting academic performance from the meals variable. This gives us a sense, that for any given California elementary school, we have an average of how many units of academic performance the model may be in error.

The F-statistic P-value, being extremely close to 0 (and certainly below < .05), tells us the model itself has statistical significance.

**Bi-Variate: Regression Violations Check**

layout(matrix(c(1,2,3,4),2,2))

plot(regression\_1)



*Step 8: Check for regression violations for this bivariate mode. Did you find any major violations?*

Graph 1 (upper left) : Residuals vs. Fitted : We expect this graph to look like a random array of points around zero. Since nothing fans out, we can say there is no heteroscedasticity in the data. Since nothing looks curved, we can assume the data does not violate linearity.

Graph 2 (lower left) : Q-Q : This graph shows deviations from normality. The closer the points to the line, the more normal the distribution. Since the points are very close to the line, we can say the distribution of residuals is normal.

Graph 3 (upper-right): Scale-location tells us if residuals are spread equally along the range of predictors. We expect the residuals to be spread randomly, which is the case here.

Graph 4 (Residuals vs. Leverage) : This graph helps us to identify influential observations. If a case has a high Cook distance score, the case is influential to the regression result. Observation 226 and 259 look to be outside the Cook’s distance line. It may be worth it to investigate if there is something unusual about these schools, potentially remove from the data set, and re-run the model to see if there is an effect on r-squared.

|  |  |
| --- | --- |
| hist(scale(regression\_1$residuals)) | The residuals here look normally distributed. This tells us the prediction reliability of the model is pretty good. The model is NOT being influenced by a small number of cases, and predicts well across the range of values of the meals variable.  The scatterplot of meals with academic performance looked like a few schools were clustered around the high range of free meals. But their residuals seems to be reasonably normally distributed. |

**Casewise Diagnostics for Bi-Variate Model**

Hist(regression\_1$standardized.residuals )

mean(regression\_1$standardized.residuals )

[1] -0.0001365947

The mean of the residuals is practically zero, a sign the residuals are normally distributed, a sign of model goodness of fit.

regression\_1$standardized.residuals <- rstandard(regression\_1)

regression\_1$large\_residual <- regression\_1$standardized.residuals >2 | regression\_1$standardized.residuals < -2

sum(regression\_1$large\_residual)

[1] 16

There are 16 residuals with absolute values > 2. 16 out of 379 total observations is ~ 4.2%. Since we expect 95% of our data set’s residuals to be within +-2, we are within reasonable expectations.

> regression\_1$very\_large\_residual <- regression\_1$standardized.residuals >3 | regression\_1$standardized.residuals < -3

> sum(regression\_1$very\_large\_residual)

[1] 2

There are 2 residuals with absolute values > 3. These should be investigated. To find which cases should be investigated, we look at the observations here and discover observations 226 and 271 should be followed up for special investigation.

> regression\_1$very\_large\_residual

4 5 6 7 8 9 10 11 12 13 14 15 16

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

17 18 19 20 21 22 23 24 26 30 32 33 34

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

35 37 38 39 40 41 42 43 44 45 46 47 48

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

49 50 51 52 53 54 55 56 57 58 59 60 61

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

62 63 64 65 66 67 69 70 71 72 73 74 75

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

76 77 78 79 80 81 82 83 84 86 87 88 89

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

91 93 94 95 96 97 98 99 100 101 102 103 104

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

105 106 107 108 109 110 111 112 114 115 116 117 118

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

119 120 121 122 123 124 125 126 127 128 129 130 131

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

132 133 134 135 136 137 138 139 140 141 142 143 144

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

145 146 147 148 149 150 151 152 153 154 155 156 157

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

158 159 160 161 162 163 164 165 166 167 168 169 170

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

171 172 173 174 175 176 177 178 179 180 181 182 183

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

184 185 186 187 188 189 190 191 192 193 194 195 196

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

197 198 199 200 201 202 203 204 205 206 207 208 209

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

210 211 212 213 214 215 216 217 218 220 221 222 223

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

224 225 226 227 228 229 230 231 232 233 234 235 236

FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

239 240 241 242 243 244 245 246 247 248 249 250 251

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

252 253 254 255 256 257 258 259 260 261 262 263 264

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

265 266 267 268 269 270 271 272 273 274 275 276 277

FALSE FALSE FALSE FALSE FALSE FALSE TRUE FALSE FALSE FALSE FALSE FALSE FALSE

278 279 280 281 282 283 284 285 286 287 288 289 290

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

291 292 293 294 295 296 297 298 299 300 301 303 304

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

305 306 307 308 309 310 311 312 313 314 315 316 317

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

318 319 320 321 322 323 324 325 326 327 328 329 330

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

331 332 333 334 335 336 337 338 339 340 341 342 343

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

344 345 346 347 348 349 350 351 352 353 354 355 356

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

357 358 359 360 361 362 363 364 365 366 367 368 369

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

370 371 372 374 377 378 379 380 381 382 383 384 385

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

386 387 388 389 390 391 392 393 394 395 396 397 398

FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE FALSE

399 400

FALSE FALSE

> dwt(regression\_1)

lag Autocorrelation D-W Statistic p-value

1 0.2603539 1.477875 0

Alternative hypothesis: rho != 0

We expect that for any two observations the residual terms should be uncorrelated (or independent) (Field). Durbin-Watson statistic tests for serial correlations between errors. Or in other words, it tests for adjacent residuals to see if they are correlated. A value of 2 of this test indicates no correlation. Since our value is 1.47, we can say there is some positive correlation. Since the value is NOT less than 1 or greater than 3, we can say there is little cause for concern.

**Building the Model:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Model # | Variables | Adj. R-Squared | Adj. R-Squared Change | F-Stat P Value | Notes | Multli-collinearity check | Decision |
| 1 | meals | .812 |  | < 0.000 | Initial model |  |  |
| 2 | Meals, ell | .823 | +.011 | < 0.000 | Meals \*\*\*, ell \*\*\* | VIF < 3 on all | Keep all |
| 3 | Meals, ell, not\_hsg | .826 | +.003 | < 0.000 | Meals \*\*\*, ell \*\*, not\_hsg \*\* | VIF < 3.2 on all | Not much help but significant, so keep all |
| 4 | Meals, ell, not\_hsg, grad\_sch | .832 | +.006 | < 0.000 | Meals \*\*\*, ell \*\*, not\_hsg \*, grad\_sch \*\*\* | VIF < 3.8 on all | Keep all |
| 5 | Meals, ell, not\_hsg, grad\_sch, col\_grd | .834 | +.002 | < 0.000 | Meals \*\*\*, ell \*\*, not\_hsg \*, grad\_sch \*\*\*, col\_grad | VIF < 4.2 on all | Col\_grad p = .62 (dropping col\_grad) |
| 6 | Meals, ell, not\_hsg, grad\_sch, yr\_rnd | .839 | +.007 (vs model 4) | < 0.000 | Meals \*\*\*, ell \*\*, not\_hsg \*, grad\_sch \*\*\*, yr\_rnd \*\*\* | VIF < 3.8 on all | Keep all |
| **7**  **(Final Model)** | **Meals, ell, not\_hsg, grad\_sch, yr\_rnd, emer** | **.849** | **+.010** | **< 0.000** | **Meals \*\*\*, ell \*, not\_hsg \*\*, grad\_sch \*\*\*, yr\_rnd \*, emer \*\*\*** | **VIF < 4.1 on all** | **Keep all** |
| 8 | Meals, ell, not\_hsg, grad\_sch, yr\_rnd, emer, full | .852 | +.003 | < 0.000 | Meals \*\*\*, ell \*, not\_hsg \*\*, grad\_sch \*\*\*, yr\_rnd \*, emer (p = .21), full (p=.10) | VIF  Emer = 5.8  Full = 5.6 | Drop full, keep emer |
| 9 (weak variable add) | Meals, ell, not\_hsg, grad\_sch, yr\_rnd, emer, hsg | .848 | -.001 (vs model 7) | < 0.000 | Meals \*\*\*, ell \*, not\_hsg \*\*, grad\_sch \*\*\*, yr\_rnd \*, emer \*\*\*, hsg (p=.51) | VIF < 4.5 on all | Drop hsg |
| 10 (weak variable add) | Meals, ell, not\_hsg, grad\_sch, yr\_rnd, emer, some\_col | .848 | -.001 (vs model 7) | < 0.000 | Meals \*\*\*, ell \*, not\_hsg \*\*, grad\_sch \*\*\*, yr\_rnd \*, emer \*\*\*, some\_col (p=.37) | VIF < 4.1 on all | Drop some\_col |
| 11 (weak variable add) | Meals, ell, not\_hsg, grad\_sch, yr\_rnd, emer, enroll | .849 | +.001 (vs model 7) | < 0.000 | Meals \*\*\*, ell \*, not\_hsg \*\*, grad\_sch \*\*\*, yr\_rnd (p=.11), emer \*\*\*, enroll (p=.26) | VIF <4.2 on all | Drop enroll |

Model 7 is chosen as the “final” model. All significant variables that were added after meals were kept in the model, despite very minor gains in adjusted R-squared. As we will see, this allowed for the model to better attribute its prediction power across several independent variables.

**Final Model (Model 7)**

summary(regression\_7)

Call:

lm(formula = acadperf ~ meals + ell + not\_hsg + grad\_sch + yr\_rnd +

emer, data = calschool)

Residuals:

Min 1Q Median 3Q Max

-165.536 -36.287 -2.599 34.064 186.206

Coefficients:

Estimate Std. Error t value Pr(>|t|)

(Intercept) 852.4592 10.0679 84.671 < 0.0000000000000002 \*\*\*

meals -2.7000 0.1807 -14.942 < 0.0000000000000002 \*\*\*

ell -0.4694 0.2132 -2.202 0.02828 \*

not\_hsg -0.6290 0.2240 -2.808 0.00525 \*\*

grad\_sch 1.3594 0.3161 4.301 0.000021768 \*\*\*

yr\_rnd -19.4450 8.1841 -2.376 0.01801 \*

emer -1.5938 0.3162 -5.040 0.000000728 \*\*\*

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 55.39 on 372 degrees of freedom

Multiple R-squared: 0.8514, Adjusted R-squared: 0.849

F-statistic: 355.2 on 6 and 372 DF, p-value: < 0.00000000000000022

Model Interpretation:

This model’s r-squared is .851. So 85.1% of the variation in academic performance can be explained by the six independent variables selected.

In this model the intercept of 852 would tell us that if all variables are set to 0, the academic performance would equal to 852.45.

Coefficients:

The meals coefficient -2.7 means for every 1% increase in the % of students receiving free meals, we can expect academic performance to drop -2.7 units, while everything else is constant.

The ell coefficient -0.46 means for every 1% increase in the % of students who are English Language Learners, we can expect academic performance to drop -0.46 units, while everything else is constant.

The not\_hsg coefficient -0.629 means for every unit increase in the % of parents without a high school degree, we can expect academic performance to drop -0.62 units, while everything else is constant.

The grad\_sch coefficient 1.35 means for every 1% increase in the % of parents with graduate school, we can expect academic performance to increase 1.35 units, while everything else is constant.

The yr\_rnd coefficient -19.44 means that year round schools would affect academic performance -19.44 units lower than non year round schools, while everything else is constant.

The emer coefficient -1.59 means for every unit increase in teachers with emergcy credentials, we can expect academic performance to decrease 1.59 units, while everything else is constant.

The very low P value for meals, grad\_sch, and emer tells us there is statistical significance with these variables in this model. There is a < 0.1% chance these variables effect on the dependent variable is due to randomness.

The low P value for not\_hsg tells us there is statistical significance with this variable, such that there is a < 1% chance these variables effect on the dependent variable is due to randomness.

The low P value for ell and yr\_rnd tells us there is statistical significance with these variables, such that there is a < 5 % chance these variables effect on the dependent variable is due to randomness.

Residual standard error 55.39 is the average error in predicting academic performance from the independent variables. This gives us a sense, that for any given California elementary school, we have an average of how many units of academic performance the model may be in error.

The F-statistic P-value, being extremely close to 0 (and certainly below < .05), tells us the model itself has statistical significance.

> vif(regression\_7)

meals ell not\_hsg grad\_sch yr\_rnd emer

4.059612 3.456349 2.634059 1.861337 1.520854 1.599188

These values all being below 10 do not indicate serious multi-collinearity problems. Meals has moderate multi-collinearity with some other variables, such as ELL. So we expect while there is some over-lap in these variables, they both are significant and add improvement (albeit minor) to r-squared.

Final Model: Variable Importance

> lm.beta(regression\_7)

meals ell not\_hsg grad\_sch yr\_rnd emer

-0.60178154 -0.08182595 -0.09109124 0.11728490 -0.05856758 -0.12739736

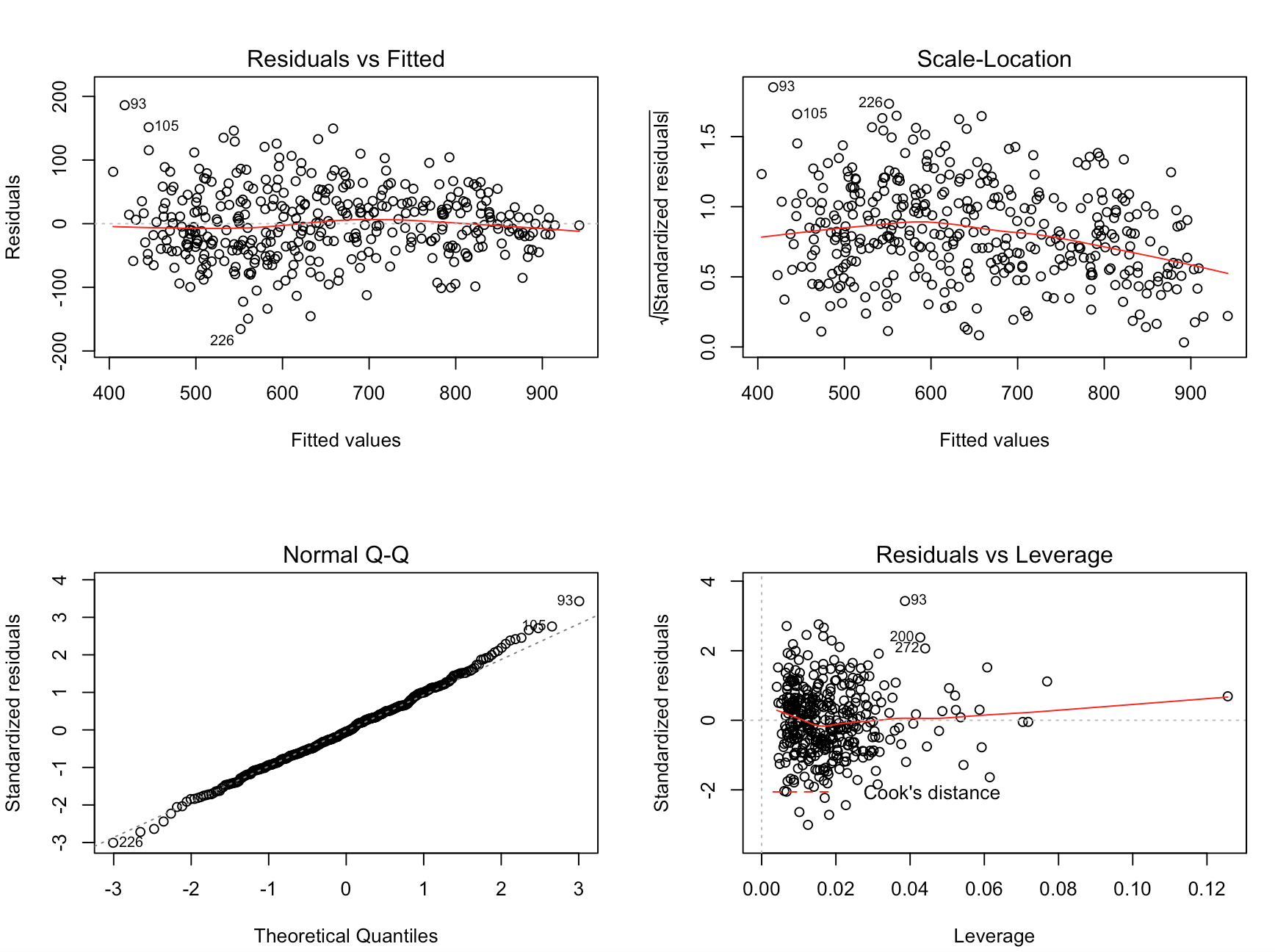
The standardized betas tell us the number of standard deviations by which the outcome will change as a result of one standard deviation change in the predictor. They tell us a lot about the importance of the variables In the model. Meals seems to dominate the model by a large margin. % Emergency credentialed teachers is next, followed by % parents with grad school, % parents who did not graduate high school, % English language learners, final the year round schools indicator.

> hist(scale(regression\_7$residuals))

|  |  |
| --- | --- |
|  | The residuals here look normally distributed. This tells us the prediction reliability of the model is pretty good. The model is NOT being influenced by a small number of cases, and predicts well across the range of values of the independent variables. |

layout(matrix(c(1,2,3,4),2,2))

plot(regression\_7)



Graph 1 (upper left) : Residuals vs. Fitted : We expect this graph to look like a random array of points around zero. Since nothing fans out, we can say there is no heteroscedasticity in the data. Since nothing looks curved, we can assume the data does not violate linearity.

Graph 2 (lower left) : Q-Q : This graph shows deviations from normality. The closer the points to the line, the more normal the distribution. Since the points are very close to the line, we can the distribution of residuals is normal. There are a few cases above the line in the upper right quadrant but it doesn’t look like very many.

Graph 3 (upper-right): Scale-location tells us if residuals are spread equally along the range of predictors. We expect the residuals to be spread randomly, which is the case here.

Graph 4 (Residuals vs. Leverage) : This graph helps us to identify influential observations. If a case has a high Cook distance score, the case is influential to the regression result. There are a handful of cases below the Cook’s distance line. It may be worth it to investigate if there is something unusual about these schools, potentially remove from the data set, and re-run the model to see if there is an effect on r-squared.

**Final Model : Casewise Diagnostics**

> regression\_7$standardized.residuals <- rstandard(regression\_7)

> regression\_7$large\_residual <- regression\_7$standardized.residuals >2 | regression\_7$standardized.residuals < -2

> sum(regression\_7$large\_residual)

[1] 18

There are 18 residuals with absolute values > 2. 18+ out of 379 total observations is ~ 4.7%. Since we expect 95% of our data set’s residuals to be within +-2, we are within reasonable expectations.

> regression\_7$large\_residual <- regression\_7$standardized.residuals >3 | regression\_7$standardized.residuals < -3

> sum(regression\_7$large\_residual)

[1] 2

There are 2 residuals with absolute values > 3.

> regression\_7$very\_large\_residual

These cases can be found via the same method above and should be flagged for investigation.

Also, briefly discuss if the final model satisfied regression assumptions overall. If not, what are some options for improving the model fit?

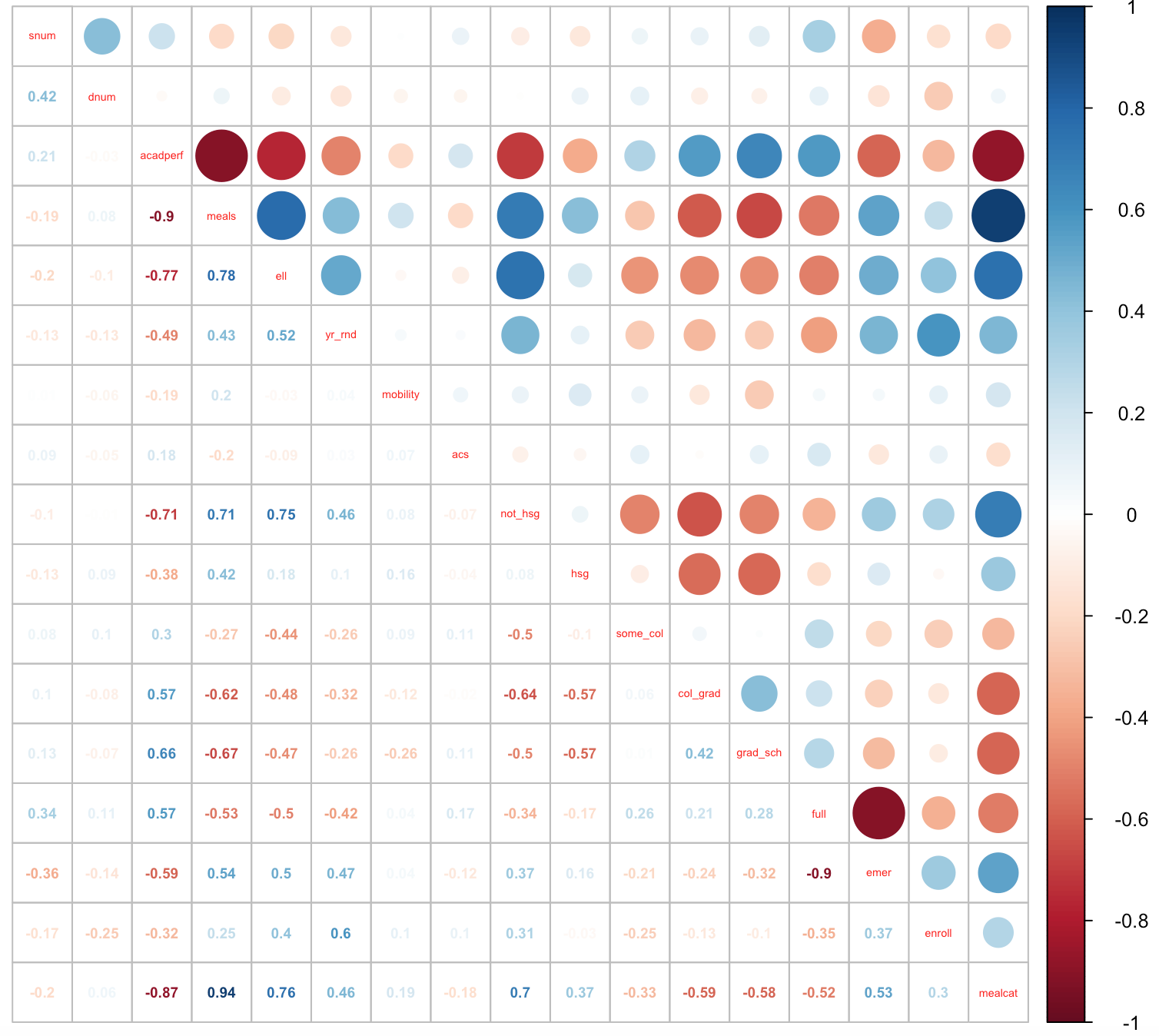
Review the distance measures and influence statistics that Field discusses for the final model (Cooks Distance), etc. What do they suggest?)

Extentsions: Visualization

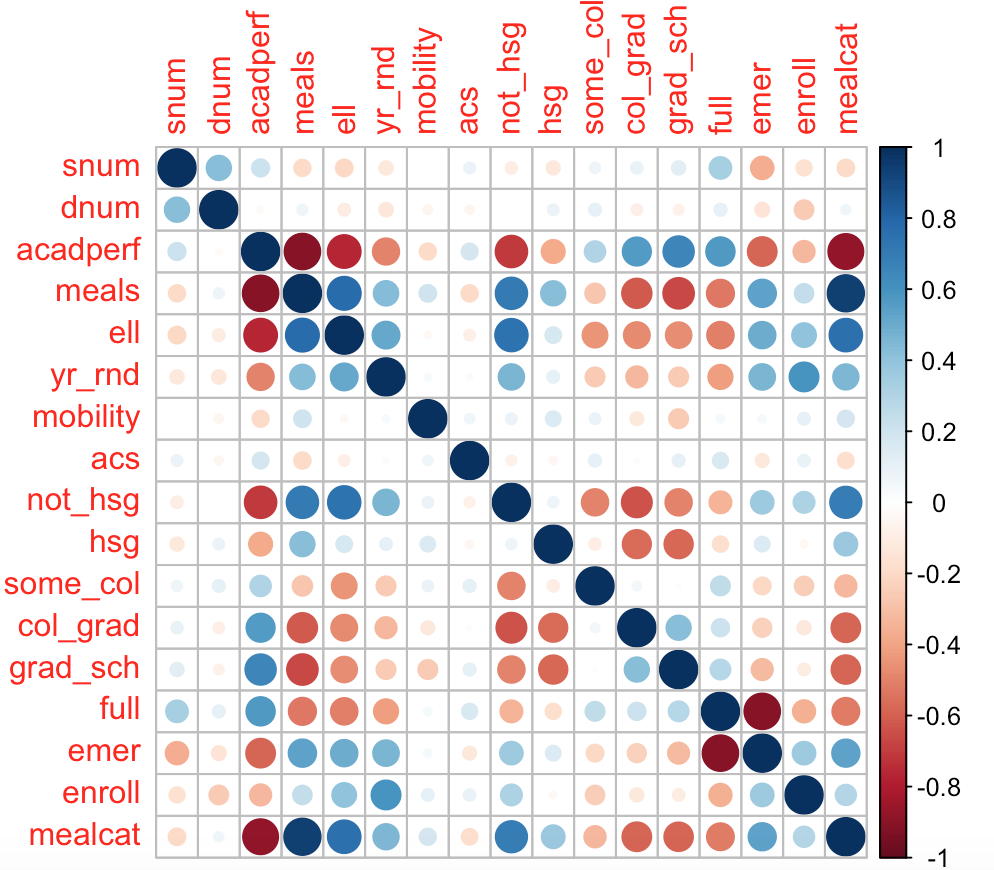
corrmatrix <- cor(calschool, use = "complete.obs")

dev.off()

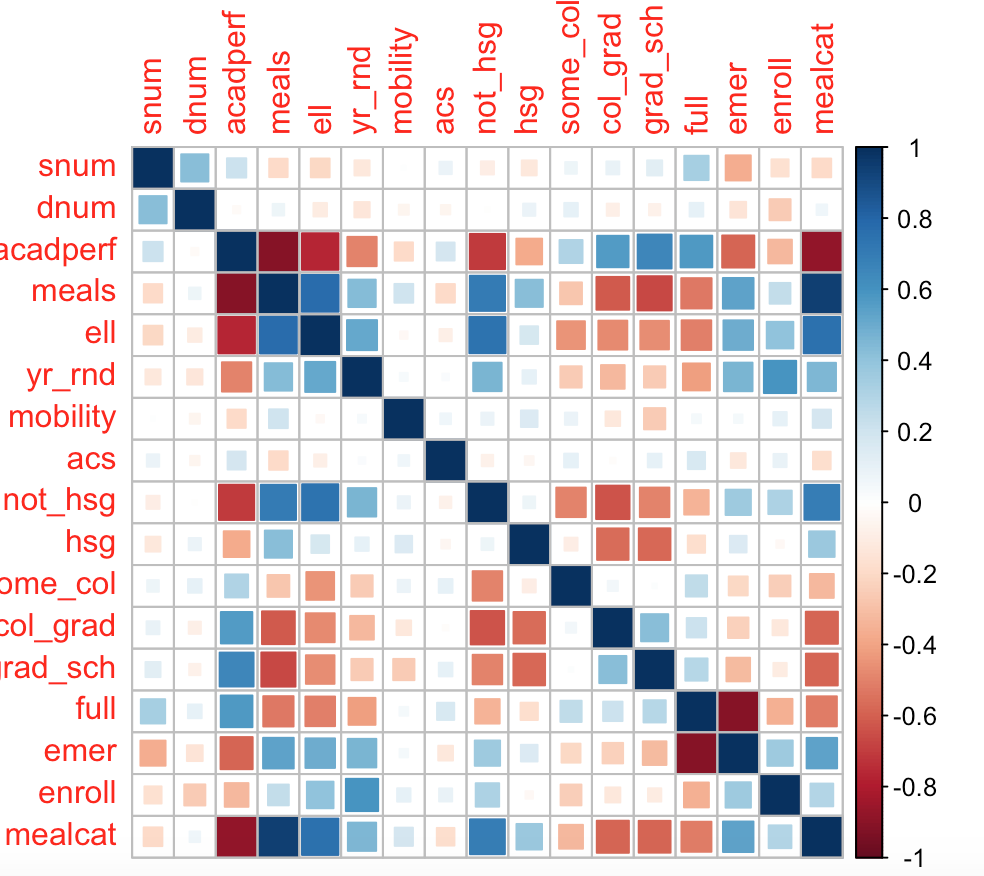
corrplot.mixed(corrmatrix, number.cex = 0.6, tl.cex = 0.45)



corrplot(corrmatrix, method="circle")



corrplot(corrmatrix, type="square")



I like ths simplicity of the square matrix. However I would potentially re-write the package such that the diagonal heavy blue squares were removed and or replaced with a light grey. Right now they add too much visual weight. The correlations are repeated in both horizontal and vertical directions, but I like the intuitive simplicity.

The first matrix at the top is more space saving and less redundant information, however it is slightly less intuitive, because you have to work in two directions to make all the comparisons, and the comparisons are not always directly adjacent.

install.packages("stargazer")

install.packages("lmtest")

install.packages("sjPlot")

library(stargazer)

library(lmtest)

library(sjPlot)

> stargazer(regression\_1, regression\_7, title="Regression Results", dep.var.labels=c("Academic Performance"), type="text")

Regression Results

=======================================================================

Dependent variable:

---------------------------------------------------

Academic Performance

(1) (2)

-----------------------------------------------------------------------

meals -4.046\*\*\* -2.700\*\*\*

(0.100) (0.181)

ell -0.469\*\*

(0.213)

not\_hsg -0.629\*\*\*

(0.224)

grad\_sch 1.359\*\*\*

(0.316)

yr\_rnd -19.445\*\*

(8.184)

emer -1.594\*\*\*

(0.316)

Constant 892.894\*\*\* 852.459\*\*\*

(6.830) (10.068)

-----------------------------------------------------------------------

Observations 379 379

R2 0.813 0.851

Adjusted R2 0.813 0.849

Residual Std. Error 61.662 (df = 377) 55.391 (df = 372)

F Statistic 1,642.732\*\*\* (df = 1; 377) 355.157\*\*\* (df = 6; 372)

=======================================================================

Note: \*p<0.1; \*\*p<0.05; \*\*\*p<0.01

The side by side result more clearly shows how the Betas have changed and allows the close comparison of the model summary statistics.

sjt.lm(regression\_1,

show.header = TRUE,

p.numeric = FALSE,

show.se = TRUE,

show.fstat = TRUE,

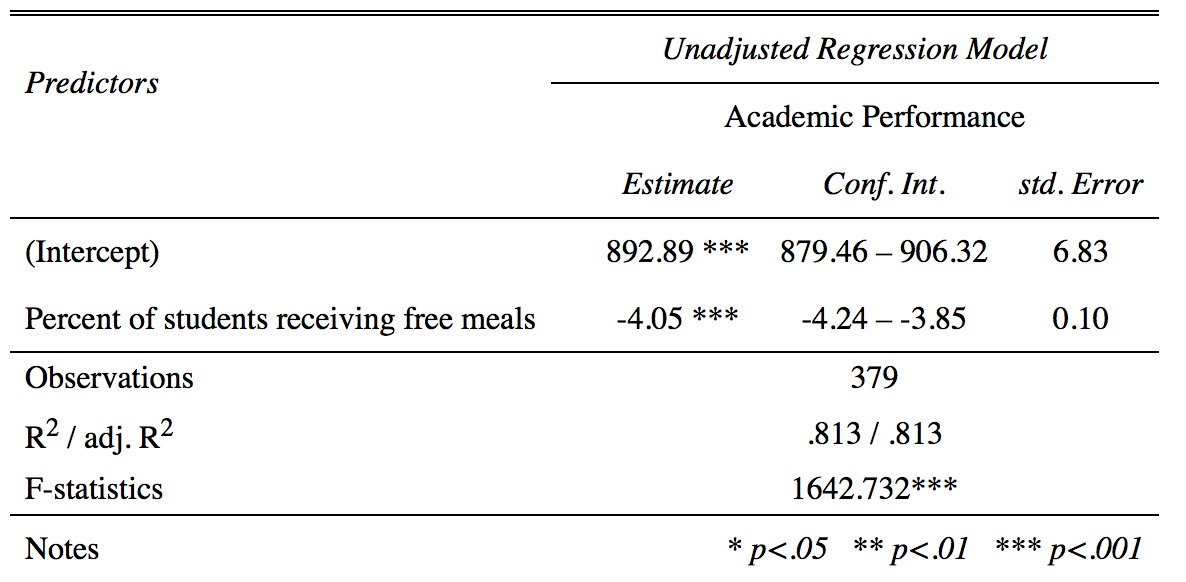
string.est = "Estimate",

string.ci = "Conf. Int.",

string.dv = "Unadjusted Regression Model",

depvar.labels = c("Academic Performance"),

pred.labels = c("Percent of students receiving free meals"))



sjt.lm(regression\_7,

show.header = TRUE,

p.numeric = FALSE,

show.se = TRUE,

show.fstat = TRUE,

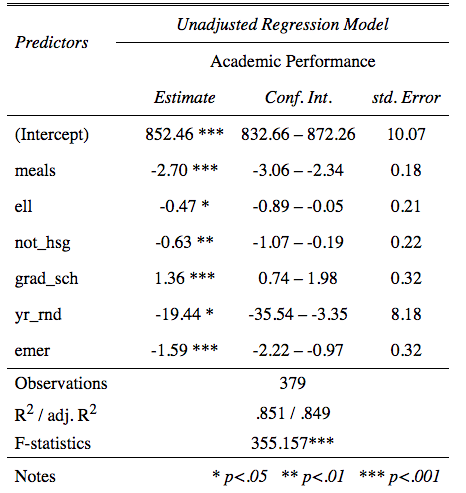
string.est = "Estimate",

string.ci = "Conf. Int.",

string.dv = "Unadjusted Regression Model",

depvar.labels = c("Academic Performance"),

pred.labels = c("Percent of students receiving free meals"))



sjt.lm(regression\_1, regression\_7,

show.header = TRUE,

p.numeric = FALSE,

show.se = TRUE,

digits.se = 3,

show.fstat = TRUE,

group.pred = FALSE,

string.est = "Estimate",

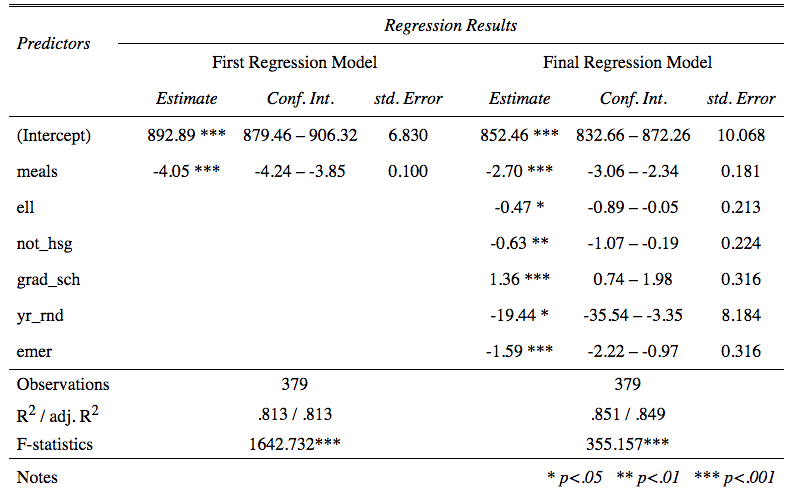
string.ci = "Conf. Int.",

string.dv = "Regression Results",

depvar.labels = c("First Regression Model", "Final Regression Model"),

pred.labels = c("% Free Meals", "% English Learners",

"% teachers with full credentials"))



First, we convert a continuous variable into categorical variables.

Take the value of ell at 75% (3rd quantile) as a cutoff.

quantile(calschool$ell, 0.75, na.rm = TRUE)

75%

50

Set value of new dummy variable : if the value off ell is less than 50 than 0, or 1 if it is above.

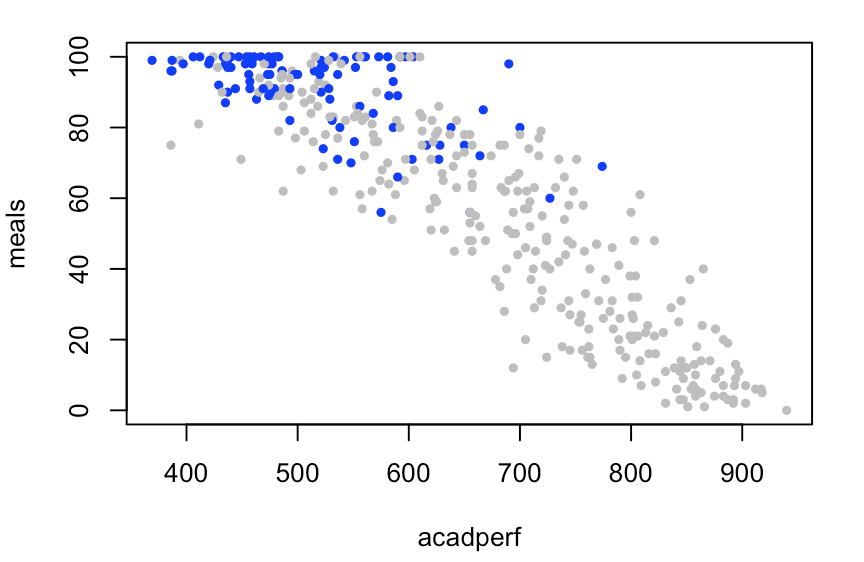
calschool$ell\_cat[calschool$ell < 50] <- 0

calschool$ell\_cat[calschool$ell >= 50] <- 1

calschool$ell\_cat <- factor(calschool$ell\_cat)

palette(c("grey","blue"))

with(calschool, plot(acadperf, meals, pch=19, col=ell\_cat, cex=0.6))



Essentially, the new ell\_cat variable identifies which schools have a large (upper quartile) number of English language learners. By highlight these in blue, we essentially introduce a 3rd variable into the scatterplot. We can more easily see that both high percentage of free meals and English language learners have strong associations with lower levels of academic performance.