Statistical Case Study #1
Analysis of Correlates of School Performance

Douglas Locke
March 8, 2018

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.

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))
> 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
```

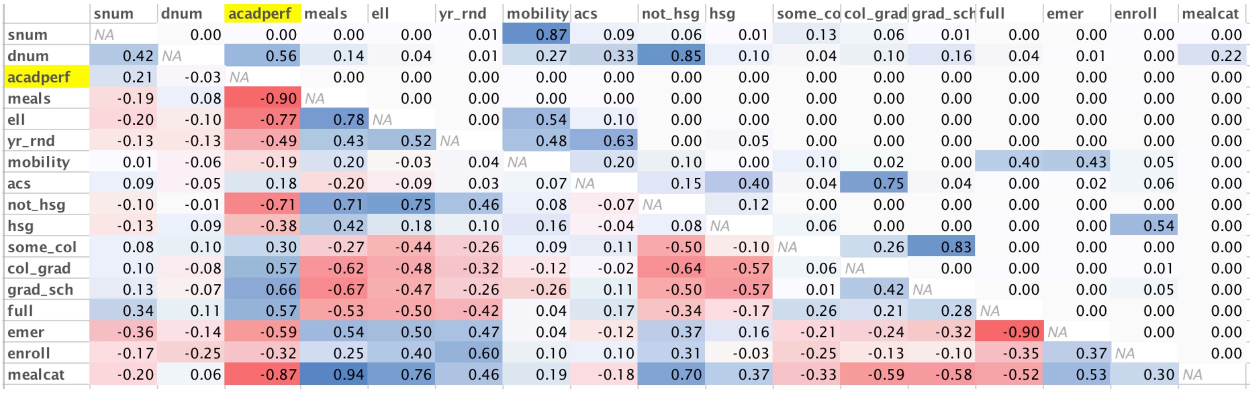
We are now left with 379 observations. Furthermore, because 19 observations out of 400 is ~ 4.7% or < 5% of our total data set, we will keep the parent education as a potential independent variable.

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.



	snum	dnum	acadperf	meals	ell	yr_rnd	mobility	acs	not_hsg	hsg	some_co	col_grad	grad_sch	full	emer	enroll	mealcat
snum	NA	0.00	0.00	0.00	0.00	0.01	0.87	0.09	0.06	0.01	0.13	0.06	0.01	0.00	0.00	0.00	0.00
dnum	0.42	NA	0.56	0.14	0.04	0.01	0.27	0.33	0.85	0.10	0.04	0.10	0.16	0.04	0.01	0.00	0.22
acadperf	0.21	-0.03	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
meals	-0.19	0.08	-0.90	NA	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
ell	-0.20	-0.10	-0.77	0.78	NA	0.00	0.54	0.10	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
yr_rnd	-0.13	-0.13	-0.49	0.43	0.52	NA	0.48	0.63	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00
mobility	0.01	-0.06	-0.19	0.20	-0.03	0.04	NA	0.20	0.10	0.00	0.10	0.02	0.00	0.40	0.43	0.05	0.00
acs	0.09	-0.05	0.18	-0.20	-0.09	0.03	0.07	NA	0.15	0.40	0.04	0.75	0.04	0.00	0.02	0.06	0.00
not_hsg	-0.10	-0.01	-0.71	0.71	0.75	0.46	0.08	-0.07	NA	0.12	0.00	0.00	0.00	0.00	0.00	0.00	0.00
hsg	-0.13	0.09	-0.38	0.42	0.18	0.10	0.16	-0.04	0.08	NA	0.06	0.00	0.00	0.00	0.00	0.54	0.00
some_col	0.08	0.10	0.30	-0.27	-0.44	-0.26	0.09	0.11	-0.50	-0.10	NA	0.26	0.83	0.00	0.00	0.00	0.00
col_grad	0.10	-0.08	0.57	-0.62	-0.48	-0.32	-0.12	-0.02	-0.64	-0.57	0.06	NA	0.00	0.00	0.00	0.01	0.00
grad_sch	0.13	-0.07	0.66	-0.67	-0.47	-0.26	-0.26	0.11	-0.50	-0.57	0.01	0.42	NA	0.00	0.00	0.05	0.00
full	0.34	0.11	0.57	-0.53	-0.50	-0.42	0.04	0.17	-0.34	-0.17	0.26	0.21	0.28	NA	0.00	0.00	0.00
emer	-0.36	-0.14	-0.59	0.54	0.50	0.47	0.04	-0.12	0.37	0.16	-0.21	-0.24	-0.32	-0.90	NA	0.00	0.00
enroll	-0.17	-0.25	-0.32	0.25	0.40	0.60	0.10	0.10	0.31	-0.03	-0.25	-0.13	-0.10	-0.35	0.37	NA	0.00
mealcat	-0.20	0.06	-0.87	0.94	0.76	0.46	0.19	-0.18	0.70	0.37	-0.33	-0.59	-0.58	-0.52	0.53	0.30	NA

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	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)	Remove in first model run	
Some_col	-0.30	Weak (Negative)	Remove in first model run	
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

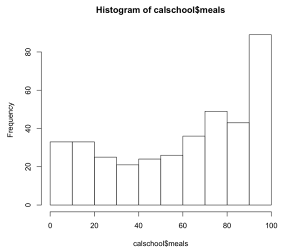
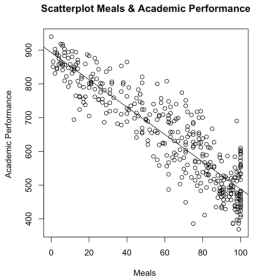
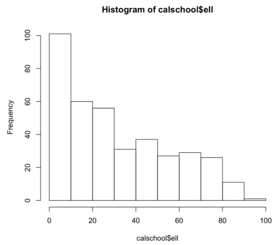
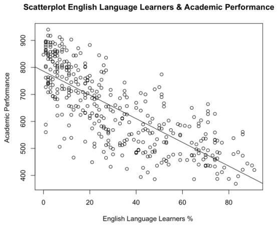
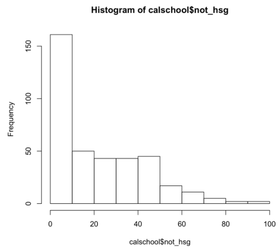
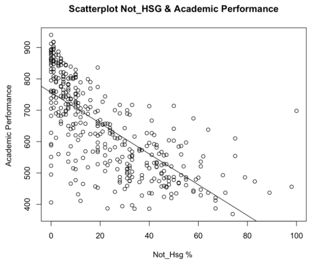
emer	-0.59	Moderate (Negative)	Possible add	
full	0.57	Moderate (Positive)	Possible add	
enroll	-0.32	Weak (Negative)	Do not add to first model run	

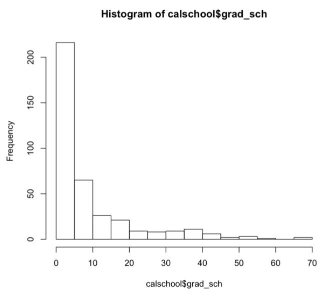
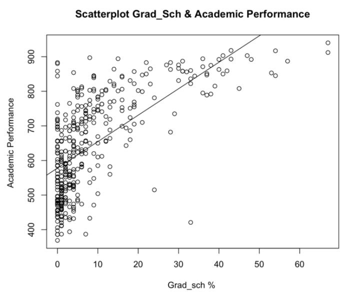
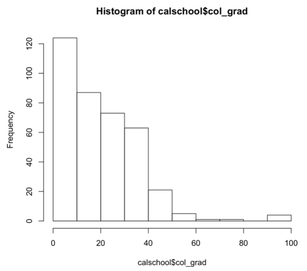
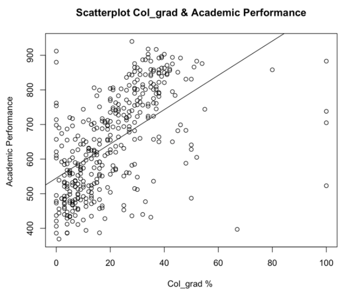
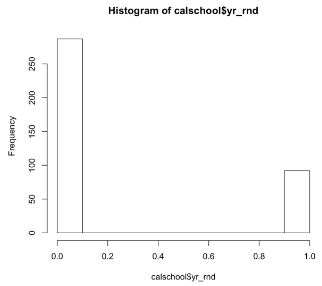
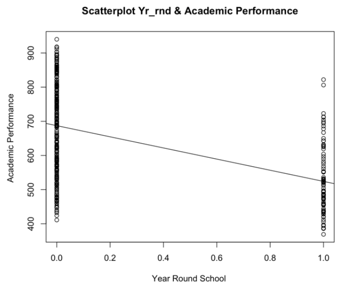
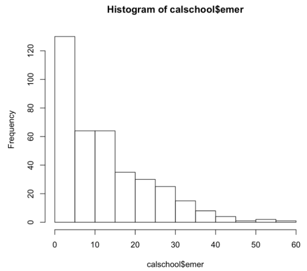
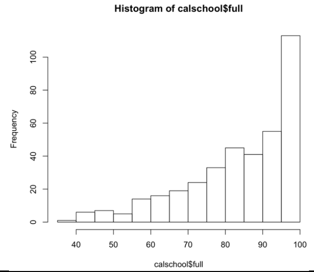
Step 5: What did your check of the correlation matrix find? Did you add any variables to the end of you list based on it? Does it look like you need to worry about multicollinearity?

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.

First Bi-Variate Model:

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

Before creating the regression model, a quick check of our first 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
```

Interpretation:

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

Step 9: Sequentially build up the model adding variables in the order you specified (don’t check reg. assumptions at each stage)

Add variables one by one. As you add variables:

- Drop variables that are insignificant unless strong theoretical reason to keep.
- If an insignificant variable makes existing variable insignificant just drop the new one.
- If the new variable is significant but adding it makes and old variable insignificant, keep both. Theory led you to think the other important, so keep it.
- Keep track of variables which are not significant. This is important to document.

Briefly document what you kept and what you dropped.

You do NOT!! Need to check assumptions for each variable you add..only do this for the bivariate model and your final model. The one exception relates to multicollinearity. It can be useful to check for multi-collinearity as you add variables.

Step 10: Recheck model assumptions, for your final model. The final model is the one you should write about.

Discuss your final model, review the coefficient table in detail, and the other key statistics (Bs, Rsq,T stats,Fstats,StandardizedBs etc). 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?)