College Success: Linear Regression

Description:

This data set, "College Success", provides high school grades, SAT scores, and Grade Point Average of 224 university students.

Variables:

- **id -** Participant ID.
- **gpa -** Grade Point Average (GPA) after three semester in college.
- **hsm -** Average high-school grade in mathematics.
- **hss -** Average high-school grade in science.
- **hse -** Average high-school grade in English.
- **satm -** SAT score for mathematics.
- **satv -** SAT score for verbal knowledge.
- **sex -** Gender (*labels not available)*

This example JASP file demonstrates the use of linear regression. Specifically, we will examine which variables best predict GPA. First, we will fit a model predicting GPA by high school grades. Then, we will use a model that predicts GPA by SAT scores. Finally, we wil fit a model that uses both high shool grades and SAT scores to predict GPA.

Reference:

Moore, D. S., McCabe, G. P., and Craig, B. A. (2012). *Introduction to the Practice of Statistics* (7th ed.). New York: Freeman.

Campbell, P. F. and McCabe, G. P. (1984). Predicting the success of freshmen in a computer science major. *Communications of the ACM*, 27: 1108–1113.

Descriptive Statistics

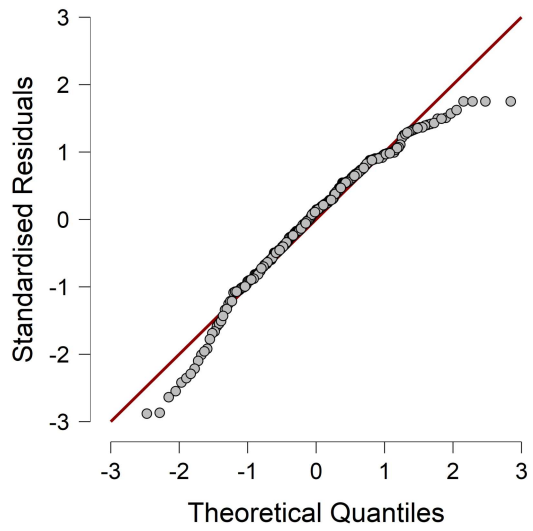
Before conducting the analysis, it is good to produce some descriptive statistics.

Descriptive Statistics

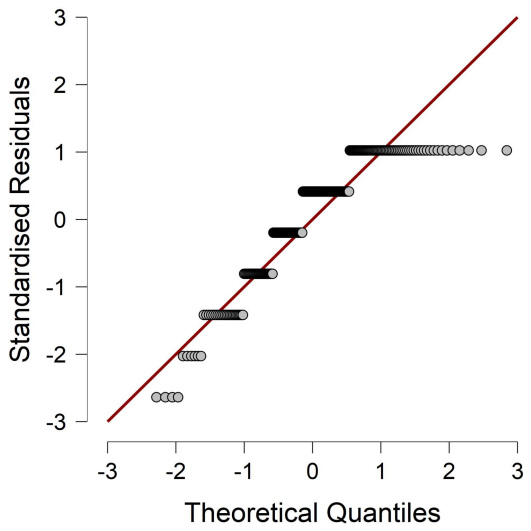
	gpa	hsm	hss	hse	satm	satv
Valid	224	224	224	224	224	224
Missing	0	0	0	0	0	0
Mean	2.635	8.321	8.089	8.094	595.286	504.549
Std. Deviation	0.779	1.639	1.700	1.508	86.401	92.610
Minimum	0.120	2.000	3.000	3.000	300.000	285.000
Maximum	4.000	10.000	10.000	10.000	800.000	760.000

Q-Q Plots

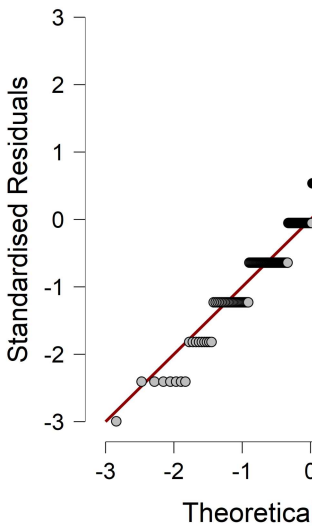
gpa



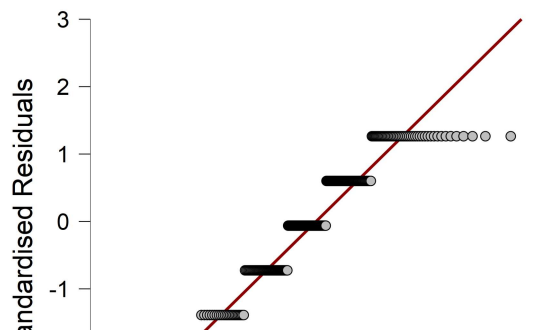
hsm



hss

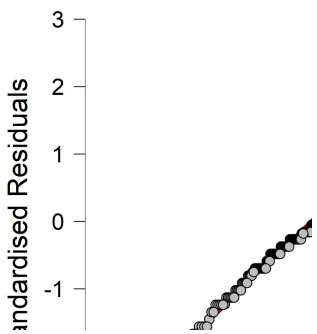


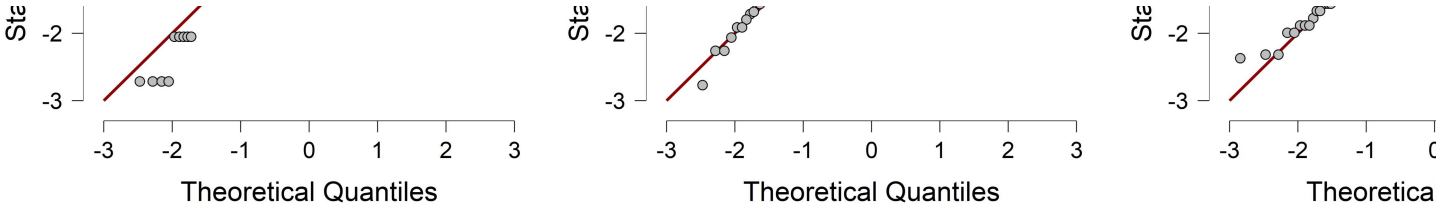
hse



satm

satv





Correlation Matrix

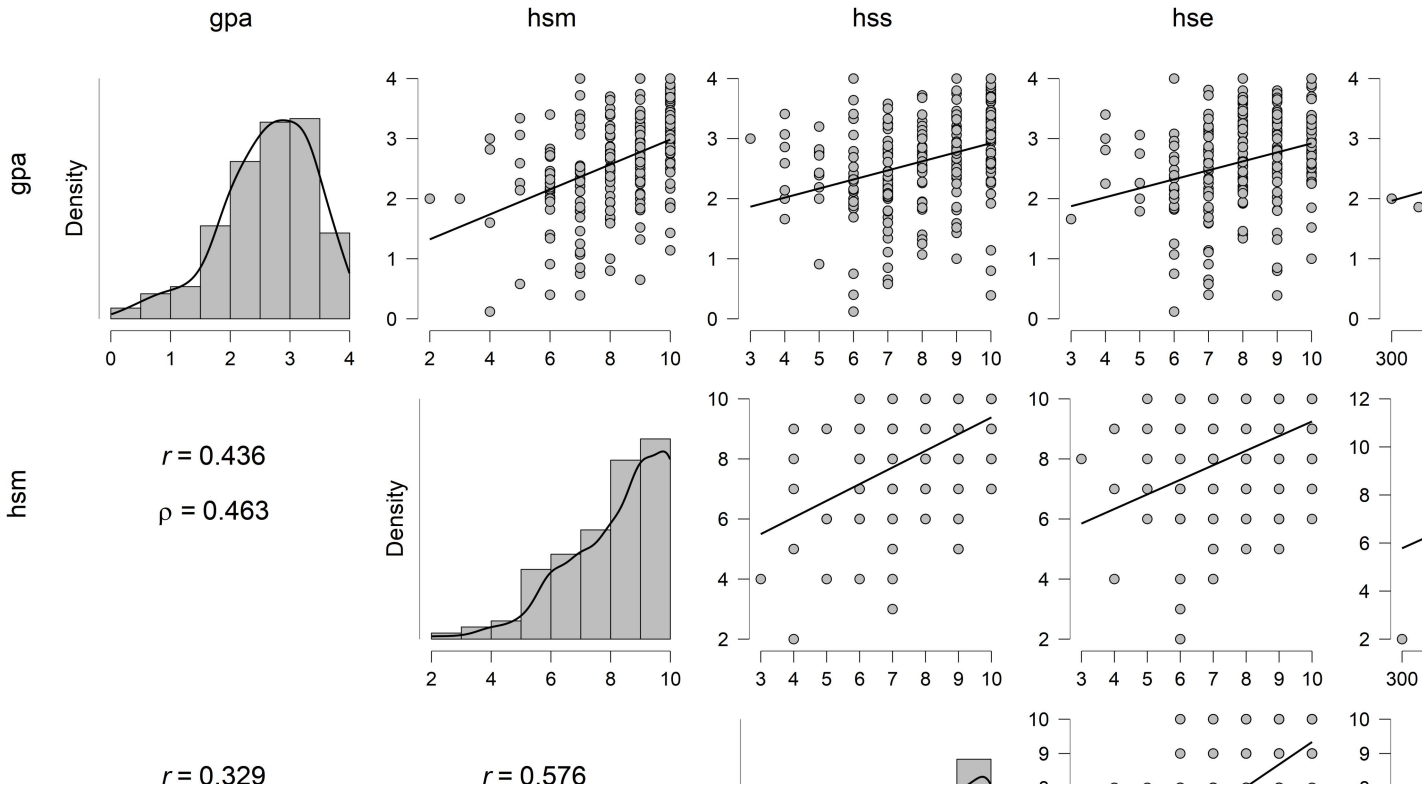
Exploratory correlation analyses reveal that the all variables are positively correlated.

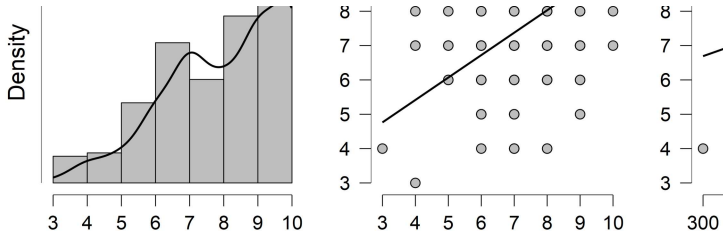
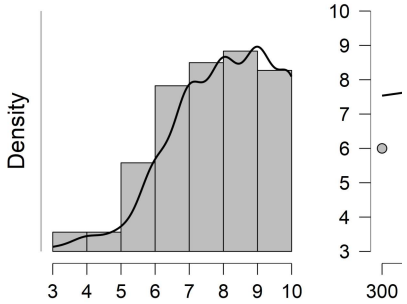
Correlation Table

		gpa	hsm	hss	hse	satm	satv
gpa	Pearson's r	—					
	p-value	—					
	Spearman's rho	—					
	p-value	—					
hsm	Pearson's r	0.436***	—				
	p-value	< .001	—				
	Spearman's rho	0.463***	—				
	p-value	< .001	—				
hss	Pearson's r	0.329***	0.576***	—			
	p-value	< .001	< .001	—			
	Spearman's rho	0.395***	0.587***	—			
	p-value	< .001	< .001	—			
hse	Pearson's r	0.289***	0.447***	0.579***	—		
	p-value	< .001	< .001	< .001	—		
	Spearman's rho	0.316***	0.483***	0.598***	—		
	p-value	< .001	< .001	< .001	—		
satm	Pearson's r	0.252***	0.454***	0.240***	0.108	—	
	p-value	< .001	< .001	< .001	0.106	—	
	Spearman's rho	0.272***	0.446***	0.237***	0.110	—	
	p-value	< .001	< .001	< .001	0.102	—	
satv	Pearson's r	0.114	0.221***	0.262***	0.244***	0.464***	—
	p-value	0.087	< .001	< .001	< .001	< .001	—
	Spearman's rho	0.155*	0.218**	0.271***	0.274***	0.403***	—
	p-value	0.021	0.001	< .001	< .001	< .001	—

* p < .05, ** p < .01, *** p < .001

Correlation Plot



hss		ρ = 0.395	ρ = 0.587	
hse	*r* = 0.289 ρ = 0.316	*r* = 0.447 ρ = 0.483	*r* = 0.579 ρ = 0.598	
satm	*r* = 0.252 ρ = 0.272	*r* = 0.454 ρ = 0.446	*r* = 0.240 ρ = 0.237	*r* = 0.108 ρ = 0.110 
satv	*r* = 0.114 ρ = 0.155	*r* = 0.221 ρ = 0.218	*r* = 0.262 ρ = 0.271	*r* = 0.244 ρ = 0.274

The dependent variable GPA has a slight negative skew. This tells us that we should examine the distribution of the residuals carefully.The independent variables show a high skew, especially the three High-School grades. The correlations between all predictors suggest that we should be mindful of multicollinearity.

Linear Regression

Regression using the high school grades:

Model Summary

Model	R	R^2	Adjusted R^2	RMSE
1	0.452	0.205	0.194	0.700

The model explains about 20% of the variance in GPA.

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
1	Regression	27.712	3	9.237	18.861	< .001
	Residual	107.750	220	0.490		
	Total	135.463	223			

The explained variance of the model is statistically significant - we may reject the simple model including only a grand mean.

Coefficients

Model	Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
						Tolerance	VIF

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
							Tolerance	VIF
1	(Intercept)	0.590	0.294		2.005	0.046		
	hsm	0.169	0.035	0.354	4.749	< .001	0.649	1.540
	hss	0.034	0.038	0.075	0.914	0.362	0.539	1.855
	hse	0.045	0.039	0.087	1.166	0.245	0.645	1.550

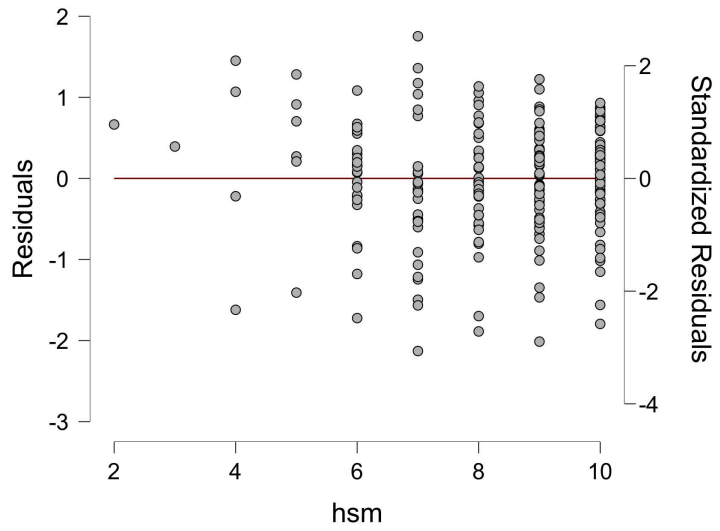
Out of the three high-school grades, only hsm is a significant predictor of GPA. The VIF statistics suggest that there is no problem with multicollinearity (VIF scores < 2).

Collinearity Diagnostics

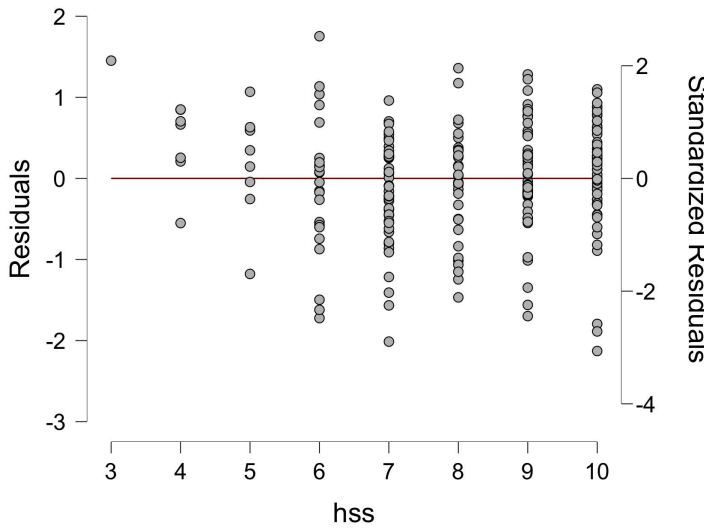
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions			
				intercept	hsm	hss	hse
1	1	3.945	1.000	0.002	0.002	0.001	0.001
	2	0.022	13.509	0.638	0.102	0.358	0.022
	3	0.019	14.240	0.044	0.634	0.098	0.411
	4	0.014	17.042	0.316	0.263	0.543	0.566

Residuals vs. Covariates

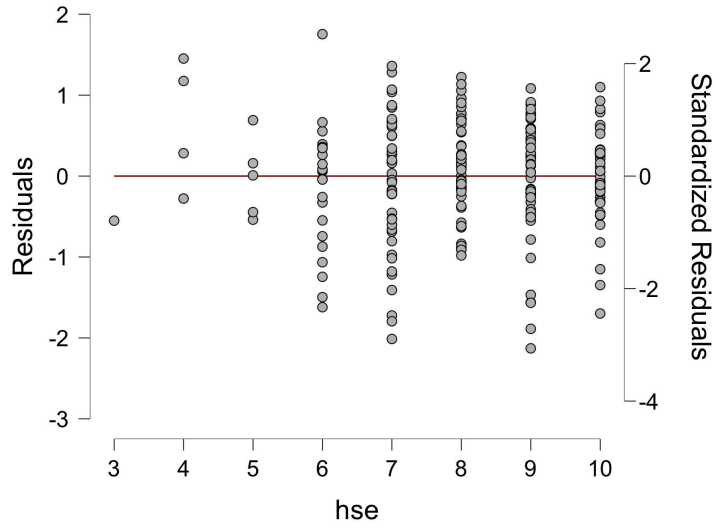
Residuals vs. hsm



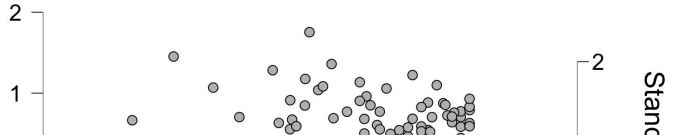
Residuals vs. hss

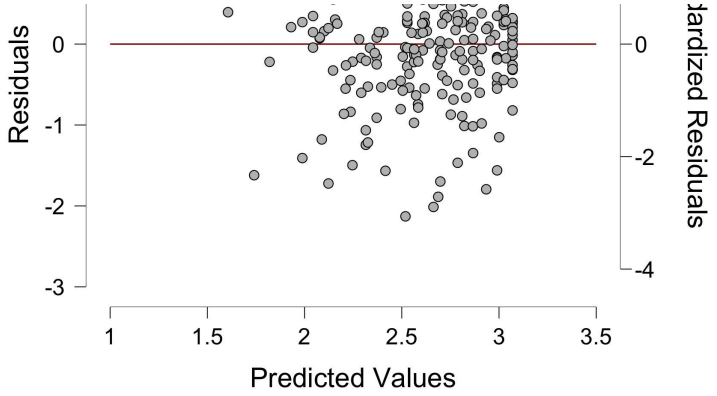


Residuals vs. hse



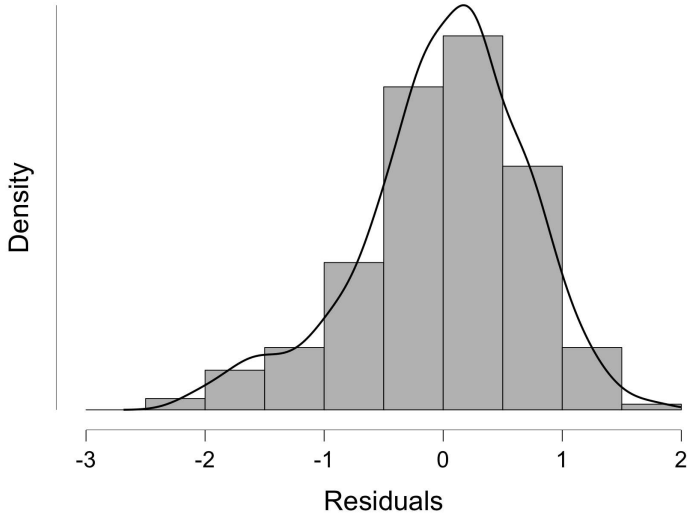
Residuals vs. Predicted



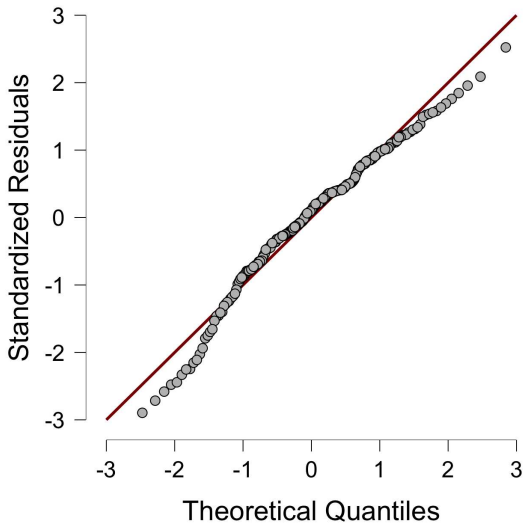


The residuals do not seem to have a problematic relationship to the predictor variables or to the predicted values. This suggests that the model is not misspecified.

Residuals Histogram



Q-Q Plot Standardized Residuals



The histogram of residals and a Q-Q plot shows that the residuals are slightly negatively skewed.

Linear Regression

Now, we include also the SAT scores. Specifically, we include the high school grades in the 'null model'. Then, we add the SAT scores to the model to test whether SAT scores contribute to the prediction of GPA over and above the high school grades.

Model Summary

Model Summary

Model	R	R^2	Adjusted R^2	RMSE	R^2 Change	F Change	df1	df2	p

Model	R	R²	Adjusted R²	RMSE	R² Change	F Change	df1	df2	p
0	0.452	0.205	0.194	0.700	0.205	18.861	3	220	< .001
1	0.460	0.211	0.193	0.700	0.007	0.950	2	218	0.388

Note. Null model includes hsm, hss, hse

The 'null model' is the same as we inspected in the first step and explains 20% of the variance of GPA. After adding the SAT scores to the prediction, the model explains only 1% of the variance more than the model with only the high school grades. The F-test of change is not statistically significant, indicating that we cannot reject the hypothesis that SAT scores do not add any new information in predicting GPA compared to the high school grades.

ANOVA

Model		Sum of Squares	df	Mean Square	F	p
0	Regression	27.712	3	9.237	18.861	< .001
	Residual	107.750	220	0.490		
	Total	135.463	223			
1	Regression	28.644	5	5.729	11.691	< .001
	Residual	106.819	218	0.490		
	Total	135.463	223			

Note. Null model includes hsm, hss, hse

Coefficients

Model		Unstandardized	Standard Error	Standardized	t	p	Collinearity Statistics	
							Tolerance	VIF
0	(Intercept)	0.590	0.294		2.005	0.046		
	hsm	0.169	0.035	0.354	4.749	< .001	0.649	1.540
	hss	0.034	0.038	0.075	0.914	0.362	0.539	1.855
	hse	0.045	0.039	0.087	1.166	0.245	0.645	1.550
1	(Intercept)	0.327	0.400		0.817	0.415		
	hsm	0.146	0.039	0.307	3.718	< .001	0.531	1.884
	hss	0.036	0.038	0.078	0.950	0.343	0.532	1.878
	hse	0.055	0.040	0.107	1.397	0.164	0.617	1.620
	satm	9.436e-4	6.857e-4	0.105	1.376	0.170	0.626	1.597
	satv	-4.078e-4	5.919e-4	-0.048	-0.689	0.492	0.731	1.367

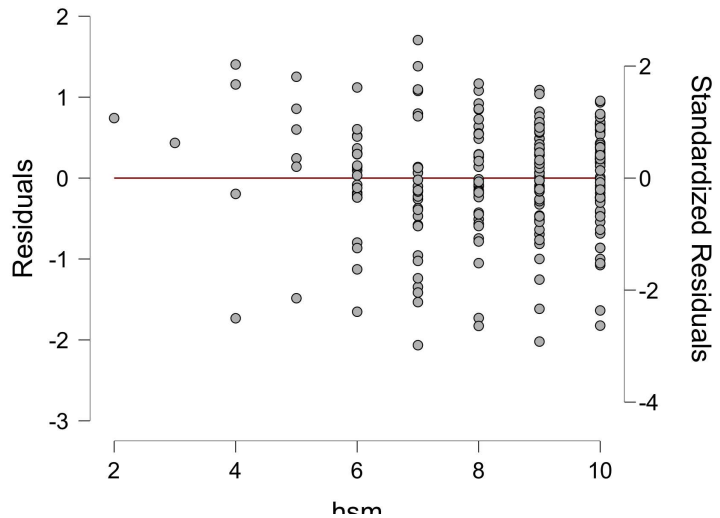
The table of coefficients shows that there is only one significant predictor: the high school grade in mathematics (hsm).

Collinearity Diagnostics

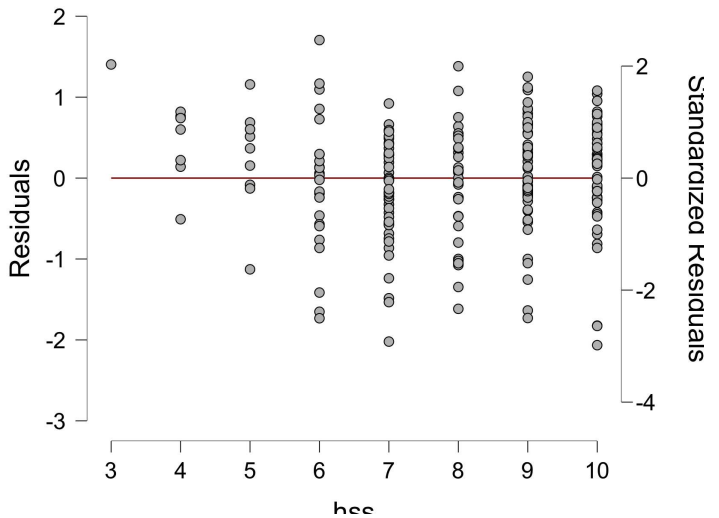
Model	Dimension	Eigenvalue	Condition Index	Variance Proportions					
				intercept	hsm	hss	hse	satm	satv
0	1	3.945	1.000	0.002	0.002	0.001	0.001		
	2	0.022	13.509	0.638	0.102	0.358	0.022		
	3	0.019	14.240	0.044	0.634	0.098	0.411		
	4	0.014	17.042	0.316	0.263	0.543	0.566		
1	1	5.902	1.000	0.000	0.001	0.001	0.001	0.000	0.001
	2	0.038	12.468	0.019	0.039	0.162	0.076	0.053	0.189
	3	0.023	16.082	0.001	0.376	0.010	0.247	0.077	0.148
	4	0.017	18.680	0.191	0.000	0.408	0.268	0.006	0.320
	5	0.013	21.356	0.260	0.359	0.413	0.186	0.038	0.229
	6	0.008	27.710	0.529	0.225	0.005	0.222	0.826	0.113

Residuals vs. Covariates

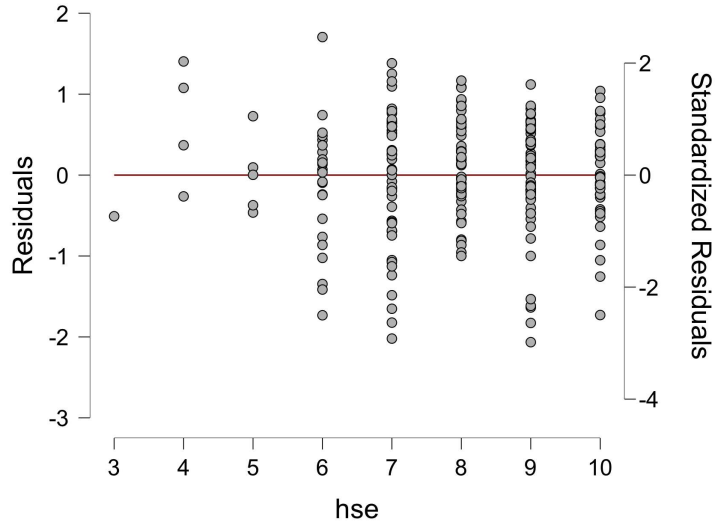
Residuals vs. hsm



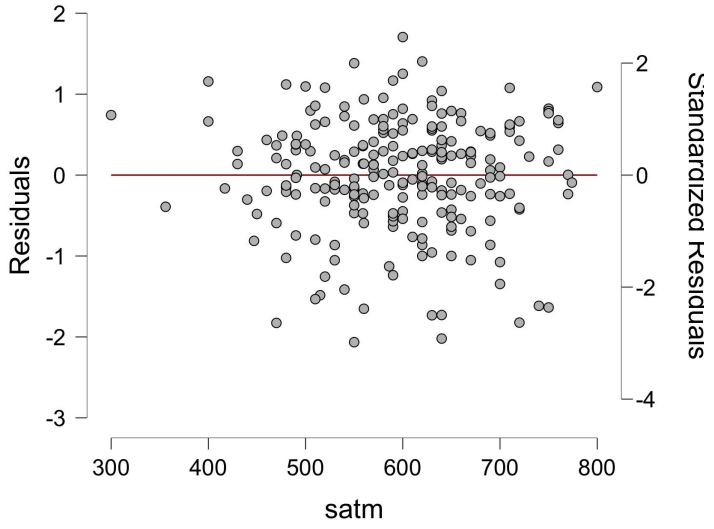
Residuals vs. hss



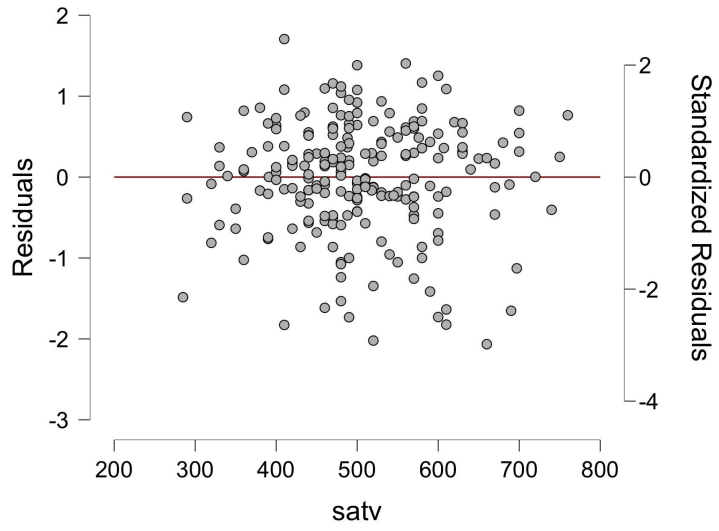
Residuals vs. hse



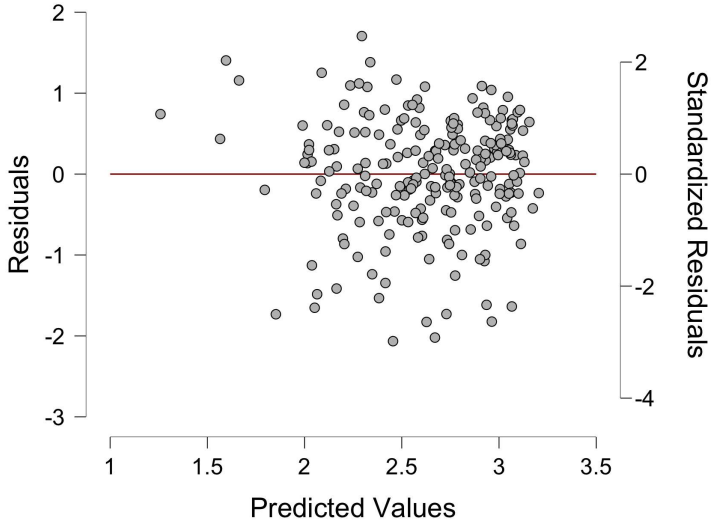
Residuals vs. satm



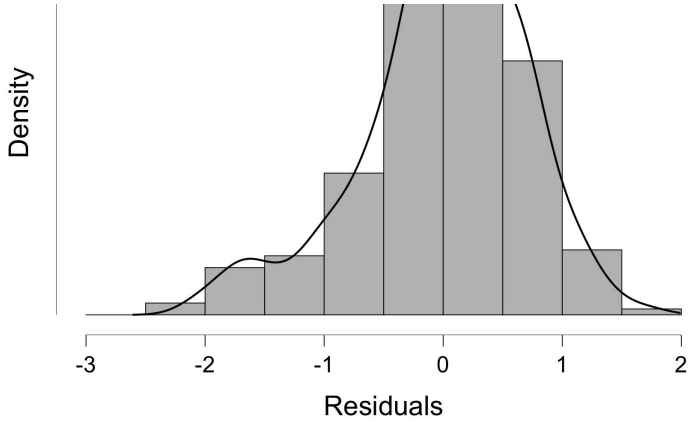
Residuals vs. satv



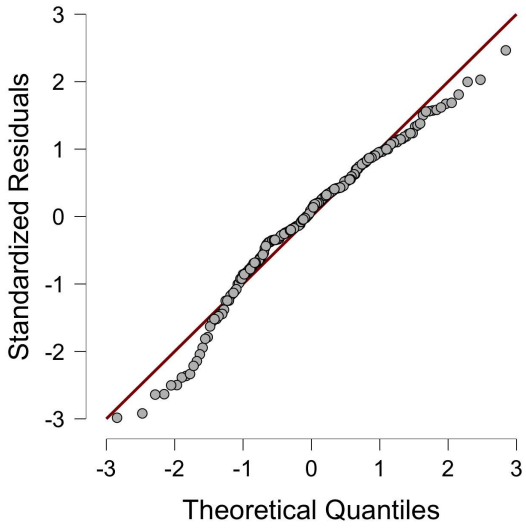
Residuals vs. Predicted



Residuals Histogram



Q-Q Plot Standardized Residuals



Again, the plots do not suggest a problem with heteroskedasticity or bias despite the residuals are slightly negatively skewed.